R01 - Simple linear regression

STAT 5870 (Engineering)
Iowa State University

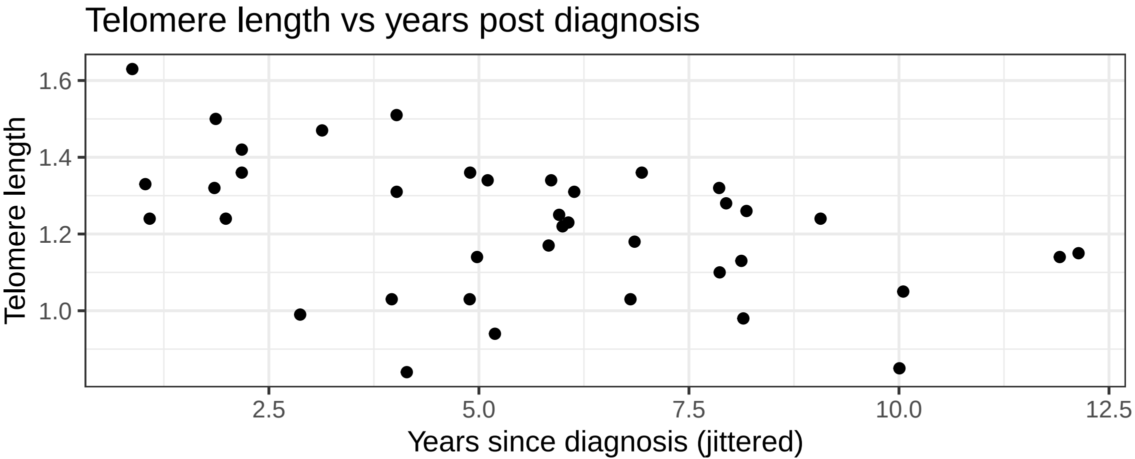
October 31, 2024

Telomere length

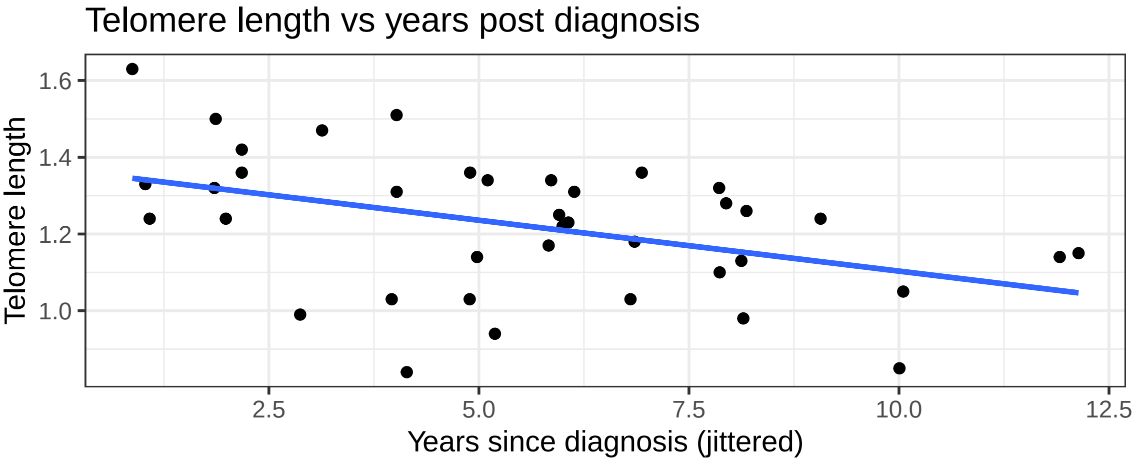
http://www.pnas.org/content/101/49/17312

> *People who are stressed over long periods tend to look haggard, and it is commonly thought that psychological stress leads to premature aging [as measured by decreased telomere length]*
>
> *...*
>
> *examine the importance of ... caregiving stress (...number of years since a child's diagnosis [of a chronic disease]) [on telomere length]*
>
> *...*
>
> *Telomere length values were measured from DNA by a quantitative PCR assay that determines the relative ratio of telomere repeat copy number to single-copy gene copy number (T/S ratio) in experimental samples as compared with a reference DNA sample.*

Data


Telomere length vs years post diagnosis
Telomere length
1.6
1.4
1.2
1.0
2.5
5.0
7.5
10.0
12.5
Years since diagnosis (jittered)

Data with regression line



Simple Linear Regression

The simple linear regression model is

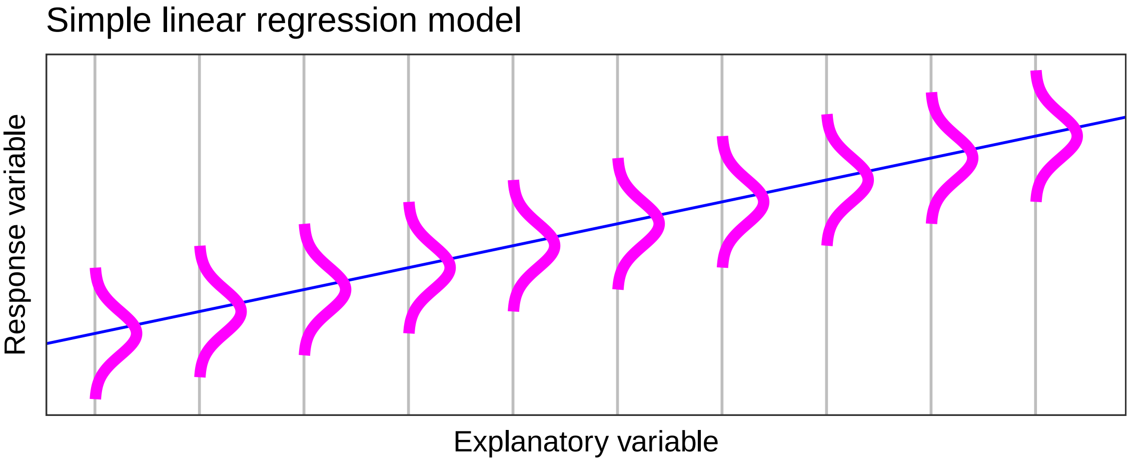
$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

where Y_i and X_i are the response and explanatory variable, respectively, for individual i.

Terminology (all of these are equivalent):

response	explanatory
outcome	covariate
dependent	independent
endogenous	exogenous

Simple linear regression - visualized



Parameter interpretation

Recall:

$$E[Y_i|X_i = x] = \beta_0 + \beta_1 x \qquad Var[Y_i|X_i = x] = \sigma^2$$

- If $X_i = 0$, then $E[Y_i|X_i = 0] = \beta_0$.

  β_0 is the expected response when the explanatory variable is zero.

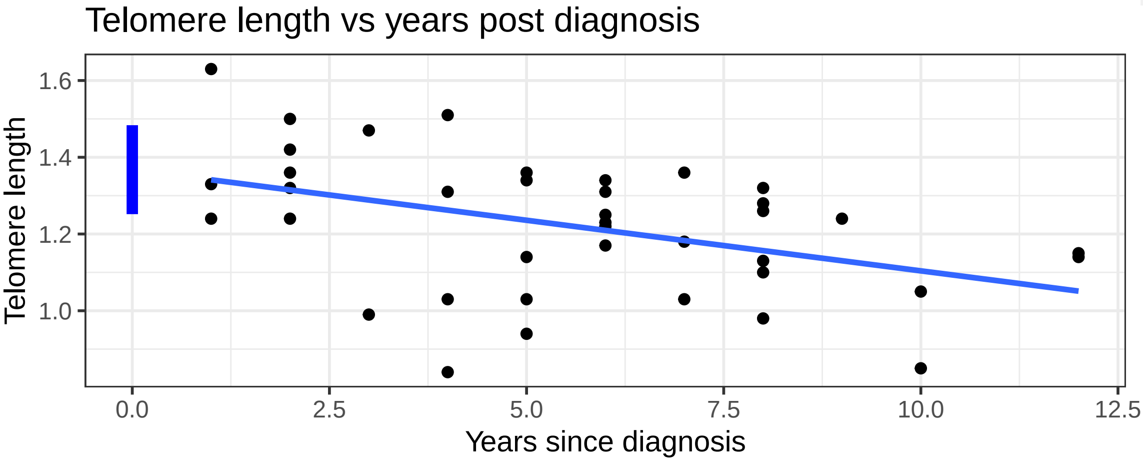
- If X_i increases from x to $x + 1$, then

$$\begin{array}{rl} E[Y_i|X_i = x + 1] & = \beta_0 + \beta_1 x + \beta_1 \\ - E[Y_i|X_i = x \qquad\;\;] & = \beta_0 + \beta_1 x \\ \hline & = \qquad\qquad\quad \beta_1 \end{array}$$

  β_1 is the expected increase in the response for each unit increase in the explanatory variable.

- σ is the standard deviation of the response for a fixed value of the explanatory variable.

Simple linear regression - visualized



Errors v residuals

Remove the mean:

$$Y_i = \beta_0 + \beta_1 X_i + e_i \qquad e_i \overset{iid}{\sim} N(0, \sigma^2)$$

So the error is

$$e_i = Y_i - (\beta_0 + \beta_1 X_i)$$

which we approximate by the residual

$$r_i = \hat{e}_i = Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i)$$

These residuals we will use for a number of purposes including

- assessing model assumptions,
- identifying outliers, and
- estimating error variance.

Estimators

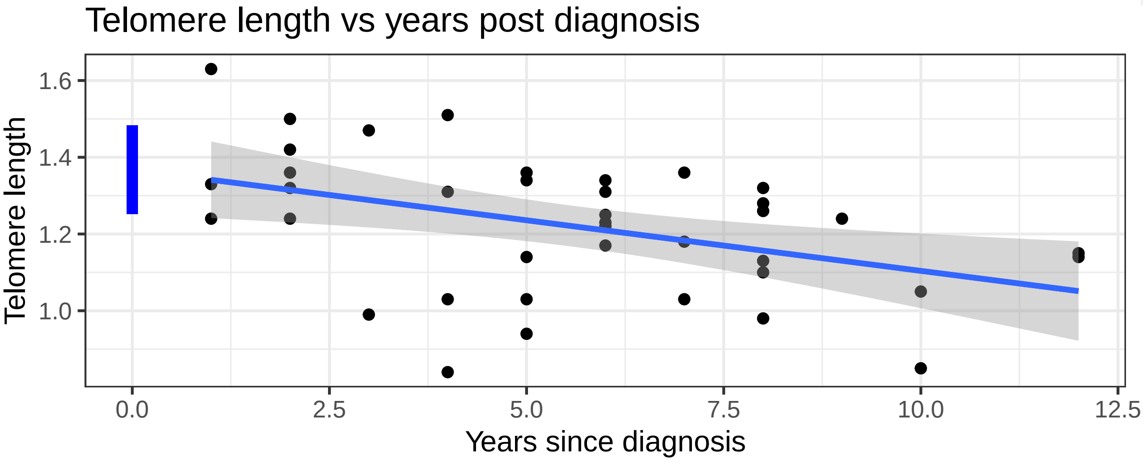
The least squares (minimize $\sum_{i=1}^n r_i^2$), maximum likelihood, and Bayesian estimators (prior $1/\sigma^2$) are

$$\begin{aligned}
\hat\beta_1 &= SXY/SXX \\
\hat\beta_0 &= \overline{Y} - \hat\beta_1 \overline{X} \\
\hat\sigma^2 &= SSE/(n-2) \qquad df = n-2
\end{aligned}$$

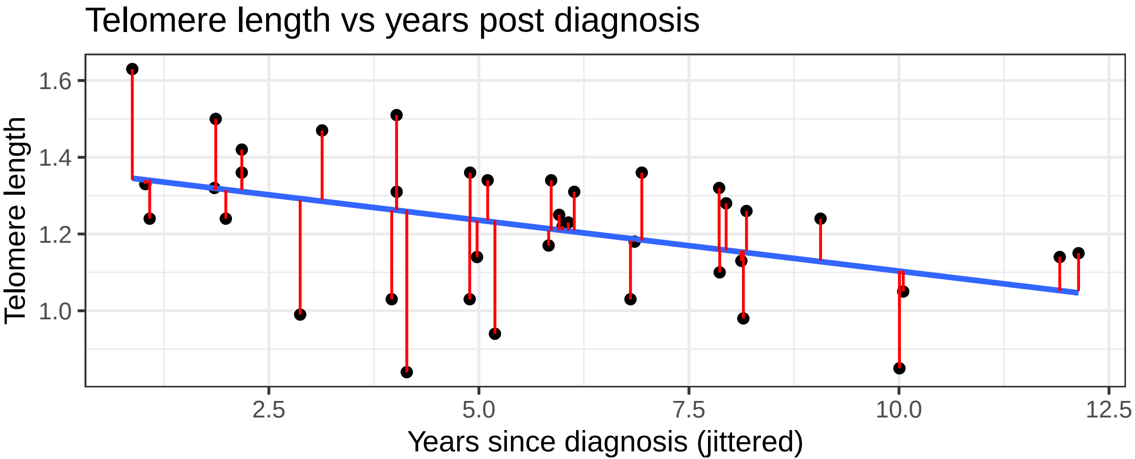
$$\begin{aligned}
\overline{X} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n X_i \\
\overline{Y} &= \tfrac{1}{n}\textstyle\sum_{i=1}^n Y_i
\end{aligned}$$

$$\begin{aligned}
SXY &= \textstyle\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y}) \\
SXX &= \textstyle\sum_{i=1}^n (X_i - \overline{X})(X_i - \overline{X}) = \sum_{i=1}^n (X_i - \overline{X})^2 \\
SSE &= \textstyle\sum_{i=1}^n r_i^2
\end{aligned}$$

Residuals


Telomere length vs years post diagnosis
Telomere length
1.6
1.4
1.2
1.0
0.0
2.5
5.0
7.5
10.0
12.5
Years since diagnosis

Residuals


Telomere length vs years post diagnosis
Telomere length
1.6
1.4
1.2
1.0
2.5
5.0
7.5
10.0
12.5
Years since diagnosis (jittered)

How certain are we about $\hat{\beta}_0$ and $\hat{\beta}_1$?

We quantify this uncertainty using their standard errors (or posterior scale parameters):

$$\begin{aligned} SE(\hat{\beta}_0) &= \hat{\sigma}\sqrt{\tfrac{1}{n} + \tfrac{\overline{X}^2}{(n-1)s_X^2}} & df &= n-2 \\ SE(\hat{\beta}_1) &= \hat{\sigma}\sqrt{\tfrac{1}{(n-1)s_X^2}} & df &= n-2 \end{aligned}$$

$$\begin{aligned} s_X^2 &= SXX/(n-1) \\ s_Y^2 &= SYY/(n-1) \\ SYY &= \textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2 \end{aligned}$$

$$\begin{aligned} r_{XY} &= \tfrac{SXY/(n-1)}{s_X s_Y} && \text{correlation coefficient} \\ R^2 &= r_{XY}^2 = \tfrac{SST-SSE}{SST} && \text{coefficient of determination} \\ SST &= SYY = \textstyle\sum_{i=1}^n (Y_i - \overline{Y})^2 \end{aligned}$$

The coefficient of determination (R^2) is the proportion of the total response variation explained by the model.

Default Bayesian analysis of the simple linear regression model

If we assume the default prior $p(\beta_0, \beta_1, \sigma^2) \propto 1/\sigma^2$, then the marginal posteriors for the mean parameters are

$$\beta_j|y \sim t_{n-2}(\hat{\beta}_j, SE(\hat{\beta}_j)^2).$$

We can construct a $100(1-a)\%$ two-sided credible interval for β_j via

$$\hat{\beta}_j \pm t_{n-2,1-a/2} SE(\hat{\beta}_j)$$

where $P(T_{n-2} < t_{n-2,1-a/2}) = 1 - a/2$ for $T_{n-2} \sim t_{n-2}$.

We can compute posterior probabilities via

$$P(\beta_j > b_j|y) = P\left(T_{n-2} > \frac{b_j - \hat{\beta}_j}{SE(\hat{\beta}_j)}\right) \quad \text{or} \quad P(\beta_j < b_j|y) = P\left(T_{n-2} < \frac{b_j - \hat{\beta}_j}{SE(\hat{\beta}_j)}\right)$$

often $b_j = 0$.

p-values and confidence interval

We can construct a $100(1-a)\%$ two-sided confidence interval for β_j via

$$\hat{\beta}_j \pm t_{n-2,1-a/2} SE(\hat{\beta}_j).$$

We can compute one-sided p-values,
$H_0 : \beta_j \geq b_j$ vs $H_A : \beta_j < b_j$ has

$$p\text{-value} = P\left(T_{n-2} < \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right)$$

and $H_0 : \beta_j \leq b_j$ vs $H_A : \beta_j > b_j$ has

$$p\text{-value} = P\left(T_{n-2} > \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right)$$

or two-sided p-values for $H_0 : \beta_j = b_j$ vs $H_A : \beta_j \neq b_j$:

$$= 2 \times \min\left\{P\left(T_{n-2} > \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right), P\left(T_{n-2} < \frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right)\right\} = 2 \times P\left(T_{n-2} < -\left|\frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right|\right)$$

software default is usually $b_j = 0$.

Calculations “by hand” in R

```
n    = nrow(Telomeres)
Xbar = mean(Telomeres$years)
Ybar = mean(Telomeres$telomere.length)
s_X  = sd(Telomeres$years)
s_Y  = sd(Telomeres$telomere.length)
r_XY = cor(Telomeres$telomere.length, Telomeres$years)

SXX = (n-1)*s_X^2
SYY = (n-1)*s_Y^2
SXY = (n-1)*s_X*s_Y*r_XY

beta1 = SXY/SXX
beta0 = Ybar - beta1 * Xbar

R2  = r_XY^2
SSE = SYY*(1-R2)

sigma2 = SSE/(n-2)
sigma  = sqrt(sigma2)

SE_beta0 = sigma*sqrt(1/n + Xbar^2/((n-1)*s_X^2))
SE_beta1 = sigma*sqrt(           1/((n-1)*s_X^2))
```

Calculations “by hand” in R (continued)

```
# 95% CI for beta0
beta0 + c(-1, 1) * qt(.975, df = n-2) * SE_beta0

[1] 1.251761 1.483603

# 95% CI for beta1
beta1 + c(-1, 1) * qt(.975, df = n-2) * SE_beta1

[1] -0.044785794 -0.007962836

# pvalue for H0: beta0 <= 0 and P(beta0 > 0 | y)
pt(beta0 / SE_beta0, df = n - 2)

[1] 1

# pvalue for H0: beta1 <= 0 and P(beta1 > 0 | y)
pt(beta1 / SE_beta1, df = n - 2)

[1] 0.003102353

# pvalue for H0: beta1 = 0
2 * pt(-abs(beta1 / SE_beta1), df = n - 2)

[1] 0.006204706
```

Calculations by hand

$$
\begin{aligned}
SXX &= (n-1)s_x^2 = (39-1) \times 2.9354274^2 = 327.4358974 \\
SYY &= (n-1)s_Y^2 = (39-1) \times 0.1797731^2 = 1.2280974 \\
SXY &= (n-1)s_X s_Y r_{XY} = (39-1) \times 2.9354274 \times 0.1797731 \times -0.4306534 = -8.6358974 \\
\hat{\beta}_1 &= SXY/SXX = -8.6358974/327.4358974 = -0.0263743 \\
\hat{\beta}_0 &= \overline{Y} - \hat{\beta}_1 \overline{X} = 1.2202564 - (-0.0263743) \times 5.5897436 = 1.3676821 \\
R^2 &= r_{XY}^2 = (-0.4306534)^2 = 0.1854624 \\
SSE &= SYY(1-R^2) = 1.2280974(1-0.1854624) = 1.0003316 \\
\hat{\sigma}^2 &= SSE/(n-2) = 1.0003316/(39-2) = 0.027036 \\
\hat{\sigma} &= \sqrt{\hat{\sigma}^2} = \sqrt{0.027036} = 0.1644262 \\
SE(\hat{\beta}_0) &= \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\overline{X}^2}{(n-1)s_x^2}} = 0.1644262\sqrt{\frac{1}{39} + \frac{5.5897436^2}{(39-1)*2.9354274^2}} = 0.0572111 \\
SE(\hat{\beta}_1) &= \hat{\sigma}\sqrt{\frac{1}{(n-1)s_x^2}} = 0.1644262\sqrt{\frac{1}{(39-1)*2.9354274^2}} = 0.0090867 \\
p_{H_A:\beta_0 \neq 0} &= 2P\left(T_{n-2} < -\left|\frac{\hat{\beta}_0}{SE(\hat{\beta}_0)}\right|\right) = 2P(t_{37} < -23.9058799) = 4.2740348 \times 10^{-24} \\
p_{H_A:\beta_1 \neq 0} &= 2P\left(T_{n-2} < -\left|\frac{\hat{\beta}_1}{SE(\hat{\beta}_1)}\right|\right) = 2P(t_{37} < -2.9025065) = 0.0062047 \\
CI_{95\%\ \beta_0} &= \hat{\beta}_0 \pm t_{n-2,1-a/2} SE(\hat{\beta}_0) \\
&= 1.3676821 \pm 2.0261925 \times 0.0572111 = (1.2517613, 1.4836028) \\
CI_{95\%\ \beta_1} &= \hat{\beta}_1 \pm t_{n-2,1-a/2} SE(\hat{\beta}_1) \\
&= -0.0263743 \pm 2.0261925 \times 0.0090867 = (-0.0447858, -0.0079628)
\end{aligned}
$$

Regression in R

```
m = lm(telomere.length ~ years, Telomeres)
summary(m)


Call:
lm(formula = telomere.length ~ years, data = Telomeres)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.42218 -0.08537  0.02056  0.10738  0.28869 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept)  1.367682   0.057211  23.906   <2e-16 ***
years       -0.026374   0.009087  -2.903   0.0062 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1644 on 37 degrees of freedom
Multiple R-squared:  0.1855,Adjusted R-squared:  0.1634 
F-statistic: 8.425 on 1 and 37 DF,  p-value: 0.006205


confint(m)


                  2.5 %       97.5 %
(Intercept)  1.25176134  1.483602799
years       -0.04478579 -0.007962836
```

Conclusion

Telomere ratio at the time of diagnosis of a child's chronic illness is estimated to be 1.37 with a 95% credible interval of (1.25, 1.48). For each year since diagnosis, the telomere ratio decreases on average by 0.026 with a 95% credible interval of (0.008, 0.045) . The proportion of variability in telomere length described by a linear regression on years since diagnosis is 18.5%.

http://www.pnas.org/content/101/49/17312

> *The correlation between chronicity of caregiving and mean telomere length is* -0.445 *(P <0.01). [*$R^2 = 0.198$ *was shown in the plot.]*

Remark I'm guessing our analysis and that reported in the paper don't match exactly due to a discrepancy in the data.

Summary

- The simple linear regression model is

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

  where Y_i and X_i are the response and explanatory variable, respectively, for individual i.

- Know how to use R to obtain $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\sigma}^2$, R^2, p-values, CIs, etc.
- Interpret regression output:
  - β_0 is the expected value for the response when the explanatory variable is 0.
  - β_1 is the expected increase in the response for each unit increase in the explanatory variable.
  - σ is the standard deviation of responses around their mean.
  - R^2 is the proportion of the total variation of the response variable explained by the model.

R01a - Simple linear regression:

Choosing explanatory variables

STAT 5870 (Engineering)
Iowa State University

October 30, 2024

Simple linear regression

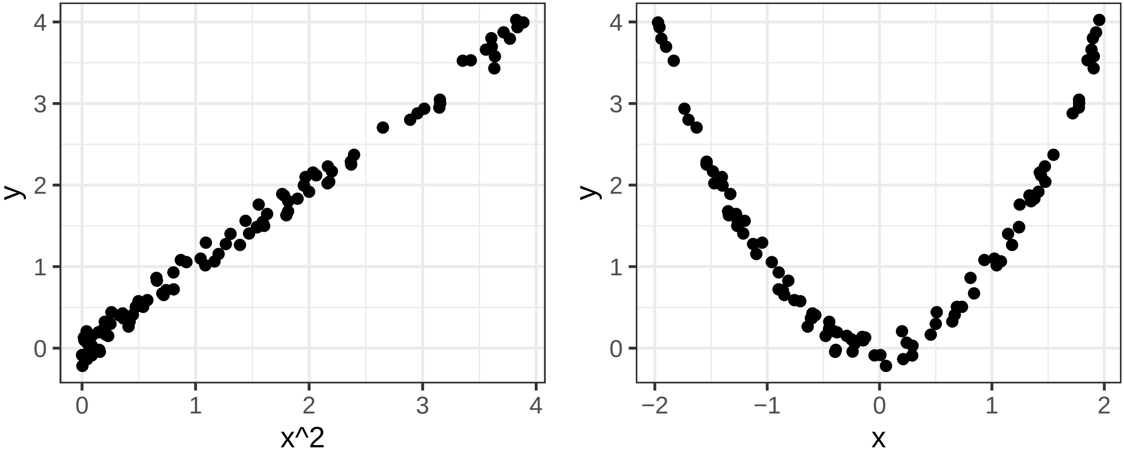
Let

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 f(X_i), \sigma^2).$$

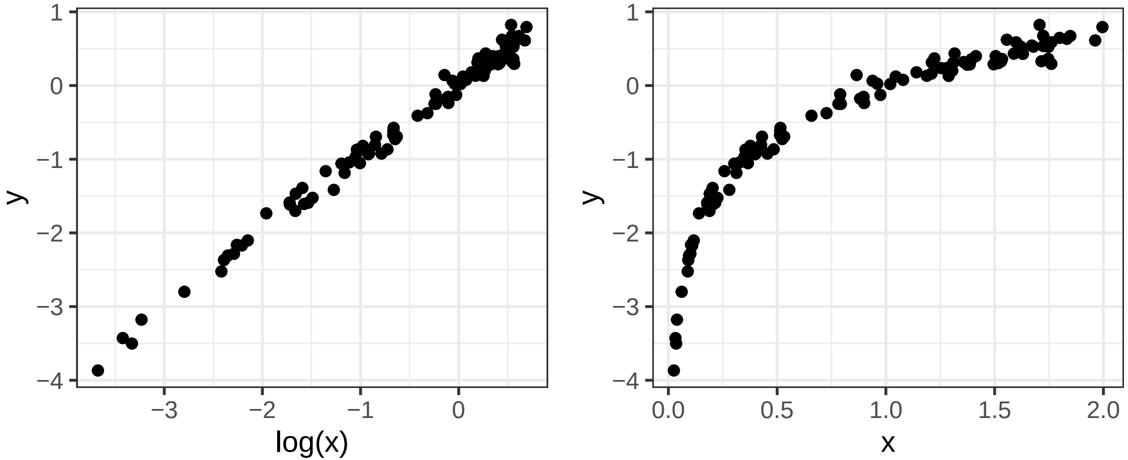
Possible choices for f:

- binary: $f(x) = I(x < c)$
- quadratic: $f(x) = x^2$
- logarithmic: $f(x) = \log(x)$
- centered: $f(x) = x - m$
- scaled: $f(x) = x/s$

Quadratic relationship



Logarithmic relationship



Shifting the intercept

The intercept is the expected response when the explanatory variable is zero. If we use

$$f(x) = x - m,$$

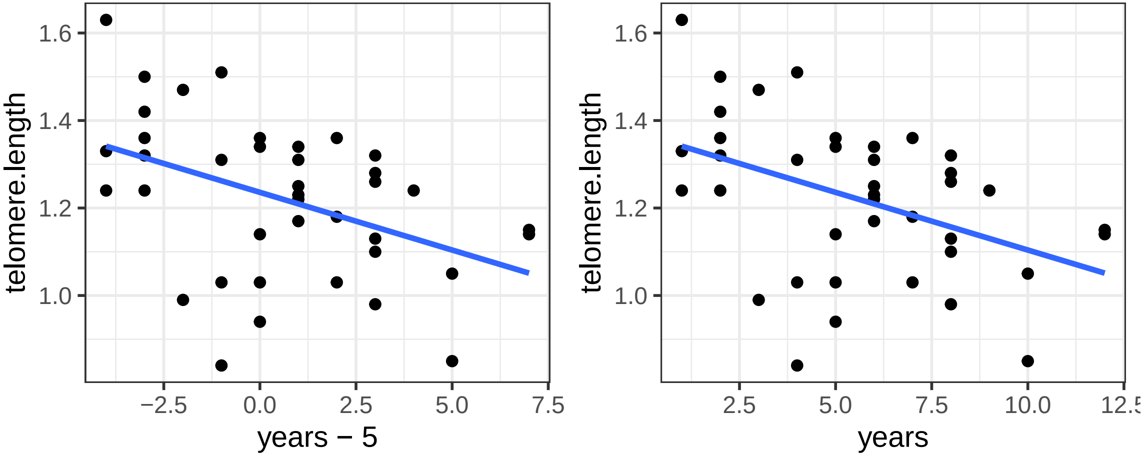
then the new intercept is the expected response when the explanatory variable is m.

$$E[Y|X = x] = \beta_0 + \beta_1(x - m) = \tilde{\beta}_0 + \tilde{\beta}_1 x$$

so our new parameters for the mean are

- slope $\tilde{\beta}_1 = \beta_1$ (unchanged) but
- intercept $\tilde{\beta}_0 = (\beta_0 - m\beta_1)$.

Telomere data



Telomere data: shifting the intercept

```
m0 = lm(telomere.length ~   years   , abd::Telomeres)
m4 = lm(telomere.length ~ I(years-5), abd::Telomeres)

coef(m0)

(Intercept)       years
 1.36768207 -0.02637431

coef(m4)

 (Intercept) I(years - 5)
  1.23581049  -0.02637431

confint(m0)

                  2.5 %       97.5 %
(Intercept)  1.25176134  1.483602799
years       -0.04478579 -0.007962836

confint(m4)

                   2.5 %       97.5 %
(Intercept)   1.18136856  1.290252429
I(years - 5) -0.04478579 -0.007962836
```

Rescaling the slope

The slope is the expected increase in the response when the explanatory variable increases by 1. If we use

$$f(x) = x/s,$$

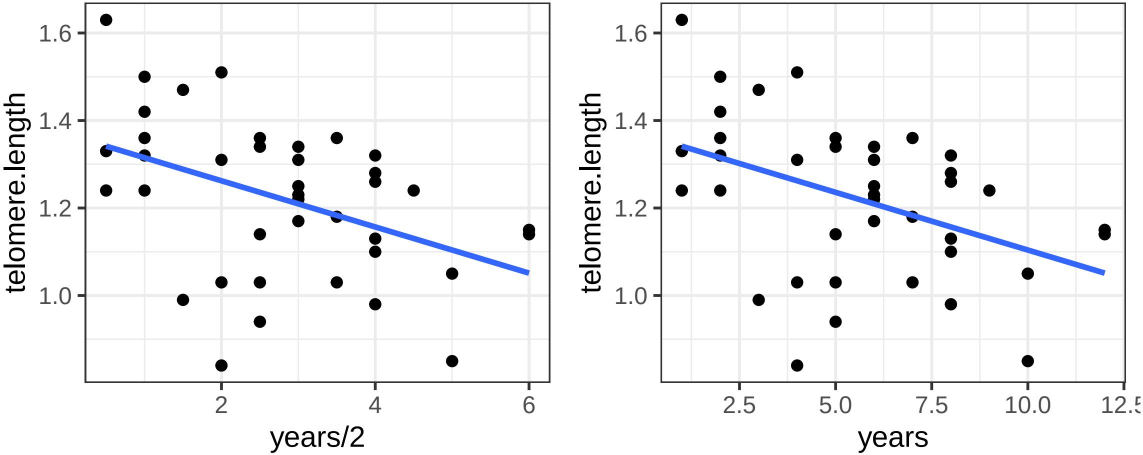
then the new slope is the expected increase in the response when the explanatory variable increases by s.

$$E[Y|X = x] = \beta_0 + \beta_1(x/s) = \tilde{\beta}_0 + \tilde{\beta}_1 x$$

so our new parameters are

- intercept $\tilde{\beta}_0 = \beta_0$ (unchanged) but
- slope $\tilde{\beta}_1 = \beta_1/s$.

Telomere data: rescaling the slope



Telomere data: rescaling the slope

```
m0 = lm(telomere.length ~   years   , abd::Telomeres)
m4 = lm(telomere.length ~ I(years/2), abd::Telomeres)

coef(m0)

(Intercept)       years
 1.36768207 -0.02637431

coef(m4)

(Intercept)  I(years/2)
 1.36768207 -0.05274863

confint(m0)

                  2.5 %       97.5 %
(Intercept)  1.25176134  1.483602799
years       -0.04478579 -0.007962836

confint(m4)

                  2.5 %      97.5 %
(Intercept)  1.25176134  1.48360280
I(years/2)  -0.08957159 -0.01592567
```

Summary

Let

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 f(X_i), \sigma^2).$$

Choose f based on

- Scientific understanding
- Interpretability
- Diagnostics

R01b - Simple linear regression

Uncertainty and prediction intervals

STAT 5870 (Engineering)
Iowa State University

October 29, 2024

Uncertainty when explanatory variable is zero

Let

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2),$$

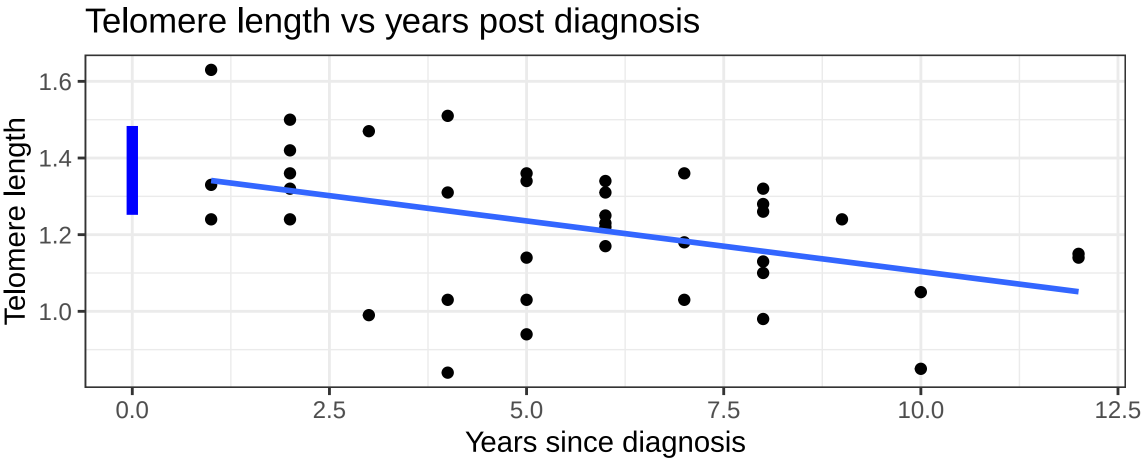
then

$$E[Y_i | X_i = 0] = \beta_0$$

and a $100(1-a)\%$ credible/confidence interval is

$$\hat{\beta}_0 \pm t_{n-2,1-a/2}\, \hat{\sigma} \sqrt{\frac{1}{n} + \frac{\overline{x}^2}{(n-1)s_x^2}}.$$

Telomere data: uncertainty



Uncertainty when explanatory variable is x

Let

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2),$$

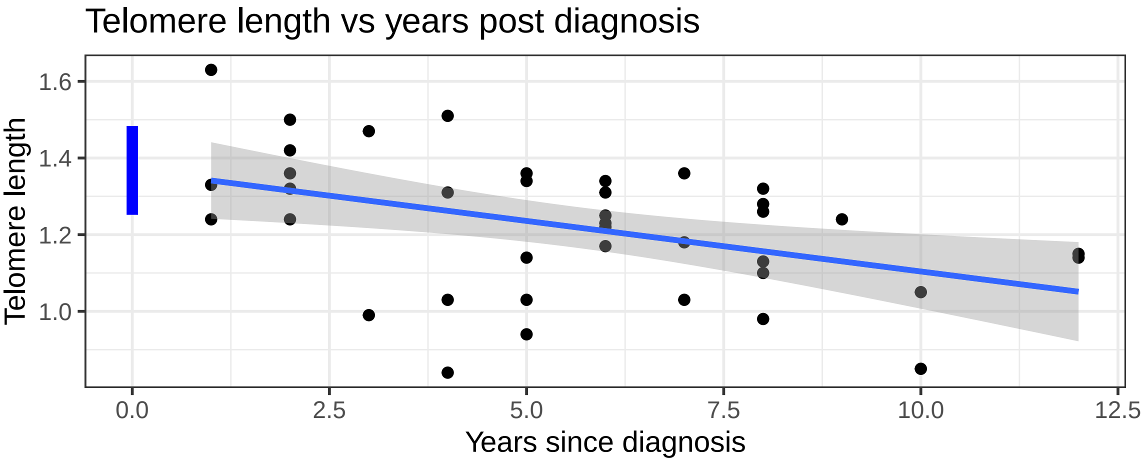
then

$$E[Y_i | X_i = x] = \beta_0 + \beta_1 x$$

and a $100(1-a)\%$ credible/confidence interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2,1-a/2}\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(\overline{x} - x)^2}{(n-1)s_x^2}}.$$

Telomere data: uncertainty



Prediction intervals

Let

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2),$$

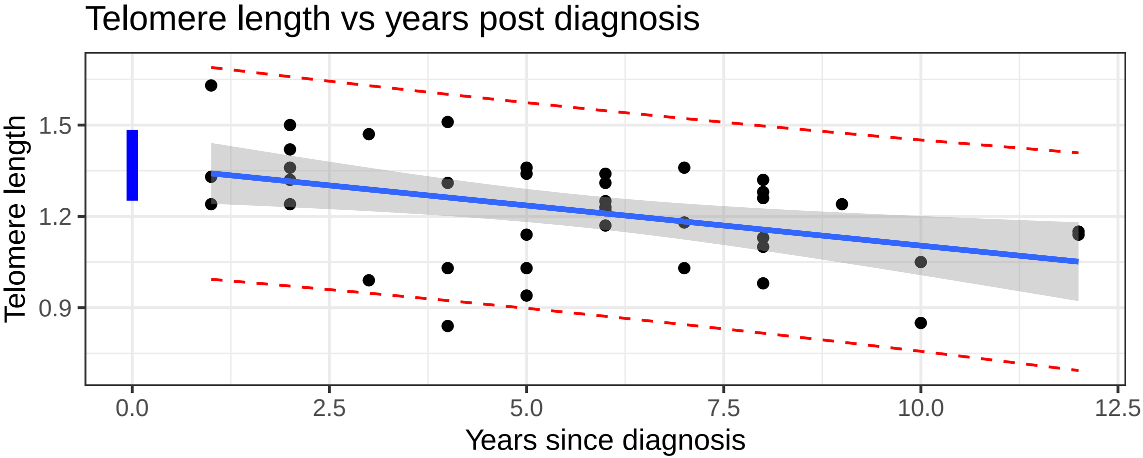
then

$$E[Y_i | X_i = x] = \beta_0 + \beta_1 x$$

and a $100(1-a)\%$ prediction interval is

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2,1-a/2}\, \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x)^2}{(n-1)s_x^2}}.$$

Telomere data: prediction intervals



Summary

Two main types of uncertainty intervals:

- where is the line?

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2,1-a/2}\,\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(\overline{x} - x)^2}{(n-1)s_x^2}}$$

- where will a new data point fall?

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{n-2,1-a/2}\,\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(\overline{x} - x)^2}{(n-1)s_x^2}}$$

Both intervals are confidence and credible intervals.

R02 - Regression diagnostics

STAT 5870 (Engineering)
Iowa State University

November 4, 2024

All models are wrong!

George Box (Empirical Model-Building and Response Surfaces, 1987):

> *All models are wrong, but some are useful.*

```
http://stats.stackexchange.com/questions/57407/what-is-the-meaning-of-all-models-are-wrong-but-some-are-useful
```

> *"All models are wrong" that is, every model is wrong because it is a simplification of reality. Some models, especially in the "hard" sciences, are only a little wrong. They ignore things like friction or the gravitational effect of tiny bodies. Other models are a lot wrong - they ignore bigger things.*
>
> *"But some are useful" - simplifications of reality can be quite useful. They can help us explain, predict and understand the universe and all its various components.*
>
> *This isn't just true in statistics! Maps are a type of model; they are wrong. But good maps are very useful.*

Simple Linear Regression

The simple linear regression model is

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

this can be rewritten as

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Key assumptions are:

- The errors are
  - normally distributed,
  - have constant variance, and
  - are independent of each other.
- There is a linear relationship between the expected response and the explanatory variables.

Multiple Regression

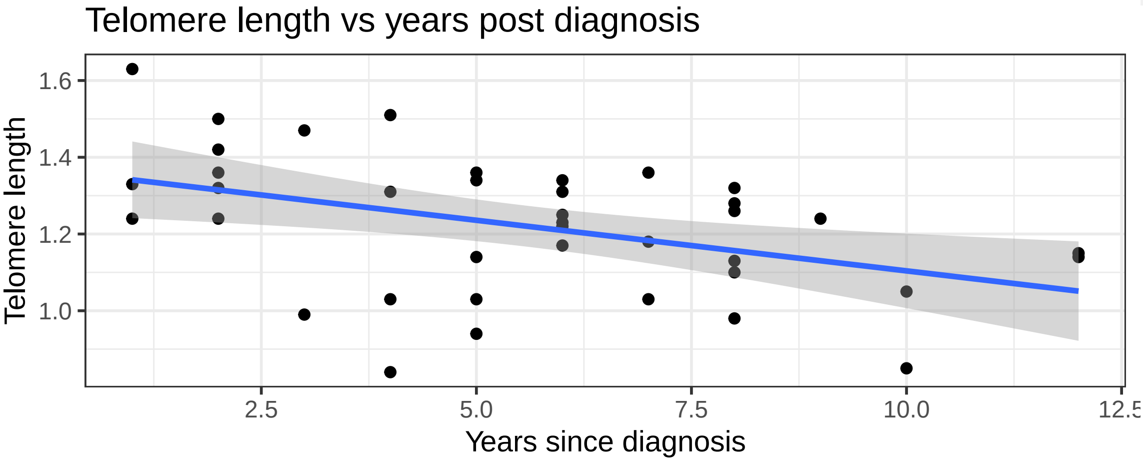
The multiple regression model is

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p} + \epsilon_i \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2).$$

Key assumptions are:

- The errors are
  - normally distributed,
  - have constant variance, and
  - are independent of each other.
- There is a specific relationship between the expected response and the explanatory variables.

Telomere data

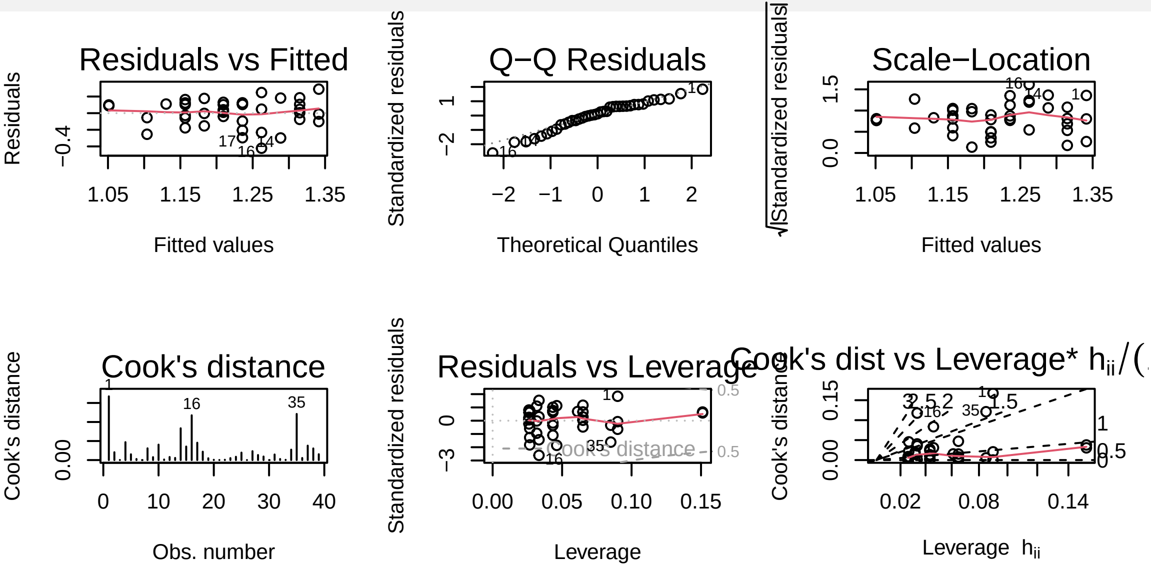


Case statistics

To evaluate these assumptions, we will calculate a variety of case statistics:

- Leverage
- Fitted values
- Residuals
  - Standardized residuals
  - Studentized residuals
- Cook's distance

Default diagnostic plots in R



Leverage

The leverage ($0 \leq h_i \leq 1$) of an observation i is a measure of how far away that observation's explanatory variable value is from the other observations. Larger leverage indicates a larger potential influence of a single observation on the regression model.

In simple linear regression,

$$h_i = \frac{1}{n} + \frac{(\overline{x} - x_i)^2}{(n-1)s_X^2}$$

which is involved in the standard error for the line for a location x_i.

The variability in the residuals is a function of the leverage, i.e.

$$Var[r_i] = \sigma^2(1 - h_i)$$

Telomere data

```
m <- lm(telomere.length~years, Telomeres)

cbind(Telomeres, leverage = hatvalues(m)) %>%
  select(years, leverage) %>%
  unique() %>%
  arrange(-years)
```

```
   years   leverage
37    12 0.15113547
35    10 0.08504307
39     9 0.06115897
27     8 0.04338293
25     7 0.03171496
20     6 0.02615505
12     5 0.02670321
10     4 0.03335944
8      3 0.04612373
4      2 0.06499608
1      1 0.08997651
2      1 0.08997651
```

Residuals and Fitted values

A regression model can be expressed as

$$Y_i = \mu_i + \epsilon_i = \beta_0 + \beta_1 X_i + \epsilon_i, \quad \epsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

A fitted value $\hat{Y}_i$ for an observation i is

$$\hat{Y}_i = \hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$$

and the residual is

$$r_i \quad = Y_i - \hat{Y}_i = \hat{e}_i$$

Standardized residuals

Often we will standardize residuals, i.e.

$$\frac{r_i}{\sqrt{\widehat{Var}[r_i]}} = \frac{r_i}{\hat{\sigma}\sqrt{1-h_i}}$$

If $|r_i|$ is large, it will have a large impact on $\hat{\sigma}^2 = \sum_{i=1}^n r_i^2/(n-2)$. Thus, we can calculate an externally studentized residual

$$\frac{r_i}{\hat{\sigma}_{(i)}\sqrt{1-h_i}}$$

where $\hat{\sigma}^2_{(i)} = \sum_{j\neq i} r_j^2/(n-3)$.

Both of these residuals can be compared to a standard normal distribution.

Telomere data: residuals

```
   years telomere.length   leverage     residual standardized studentized
1      1            1.63 0.08997651  0.288692247   1.84050794  1.90475158
2      1            1.24 0.08997651 -0.101307753  -0.64587021 -0.64070443
3      1            1.33 0.08997651 -0.011307753  -0.07209064 -0.07111476
4      2            1.50 0.06499608  0.185066562   1.16399233  1.16977226
5      2            1.42 0.06499608  0.105066562   0.66082533  0.65571510
6      2            1.36 0.06499608  0.045066562   0.28345009  0.27989750
7      2            1.32 0.06499608  0.005066562   0.03186659  0.03143344
8      3            1.47 0.04612373  0.181440877   1.12984272  1.13420749
9      2            1.24 0.06499608 -0.074933438  -0.47130041 -0.46628962
10     4            1.51 0.03335944  0.247815192   1.53293696  1.56251168
11     4            1.31 0.03335944  0.047815192   0.29577555  0.29209673
12     5            1.36 0.02670321  0.124189507   0.76558098  0.76121769
13     5            1.34 0.02670321  0.104189507   0.64228860  0.63711129
14     3            0.99 0.04612373 -0.298559123  -1.85914473 -1.92601533
15     4            1.03 0.03335944 -0.232184808  -1.43625042 -1.45793267
16     4            0.84 0.03335944 -0.422184808  -2.61155376 -2.85227987
17     5            0.94 0.02670321 -0.295810493  -1.82355895 -1.88546999
18     5            1.03 0.02670321 -0.205810493  -1.26874325 -1.27962563
19     5            1.14 0.02670321 -0.095810493  -0.59063518 -0.58536500
20     6            1.17 0.02615505 -0.039436179  -0.24304058 -0.23992534
21     6            1.23 0.02615505  0.020563821   0.12673244  0.12503525
22     6            1.25 0.02615505  0.040563821   0.24999011  0.24679724
23     6            1.31 0.02615505  0.100563821   0.61976313  0.61452870
24     6            1.34 0.02615505  0.130563821   0.80464964  0.80073848
25     7            1.36 0.03171496  0.176938136   1.09357535  1.09656310
26     6            1.22 0.02615505  0.010563821   0.06510360  0.06422148
27     8            1.32 0.04338293  0.163312451   1.01549809  1.01593894
28     8            1.28 0.04338293  0.123312451   0.76677288  0.76242192
29     8            1.26 0.04338293  0.103312451   0.64241028  0.63723335
```

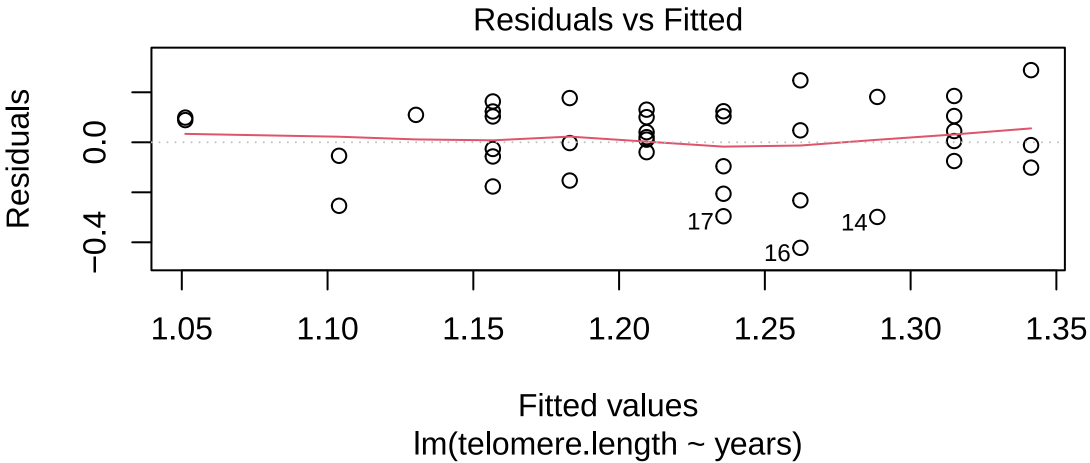
Cook's distance

The Cook's distance for an observation i ($d_i > 0$) is a measure of how much the regression parameter estimates change when that observation is included versus when it is excluded.

Operationally, we might be concerned when d_i is

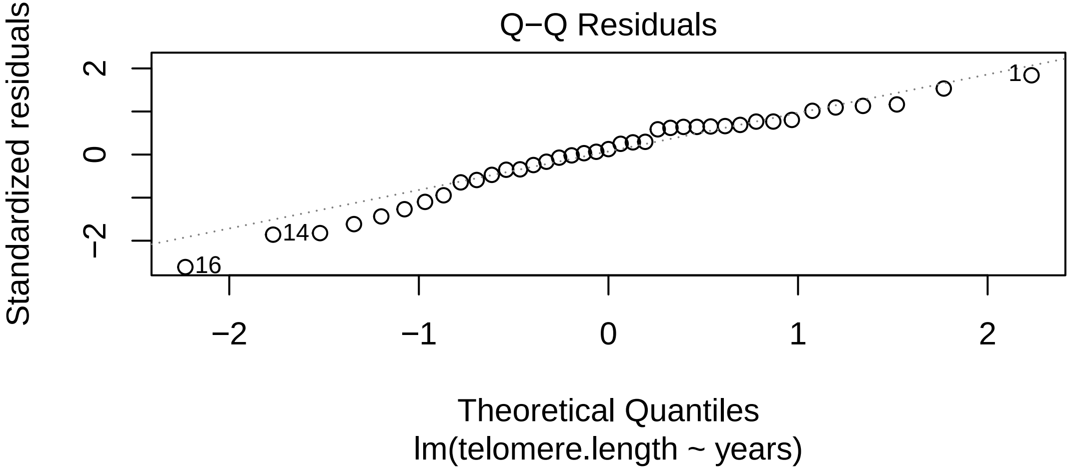
- larger than 1 or
- larger then 4/n.

Residuals vs fitted values



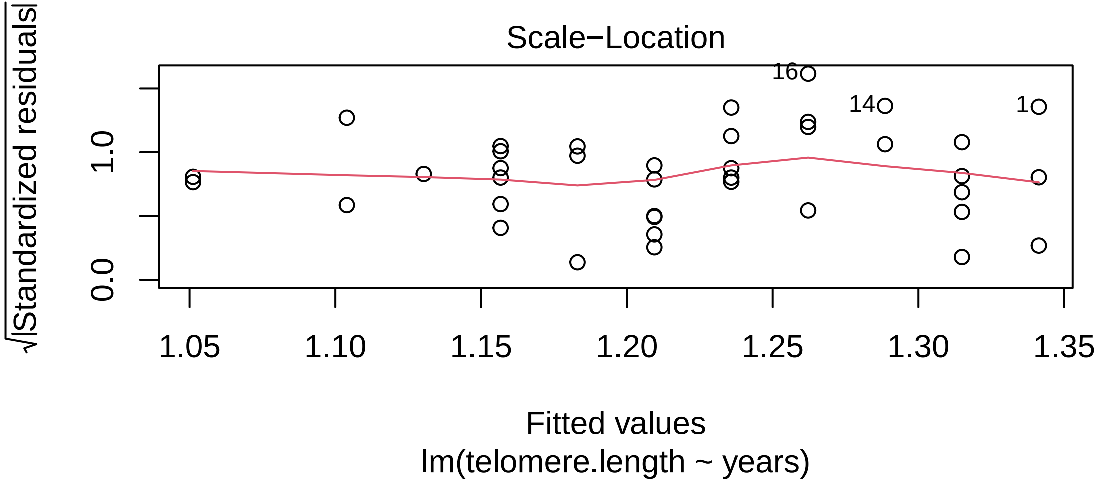
Assumption	Violation
Linearity	Curvature
Constant variance	Funnel shape

QQ-plot



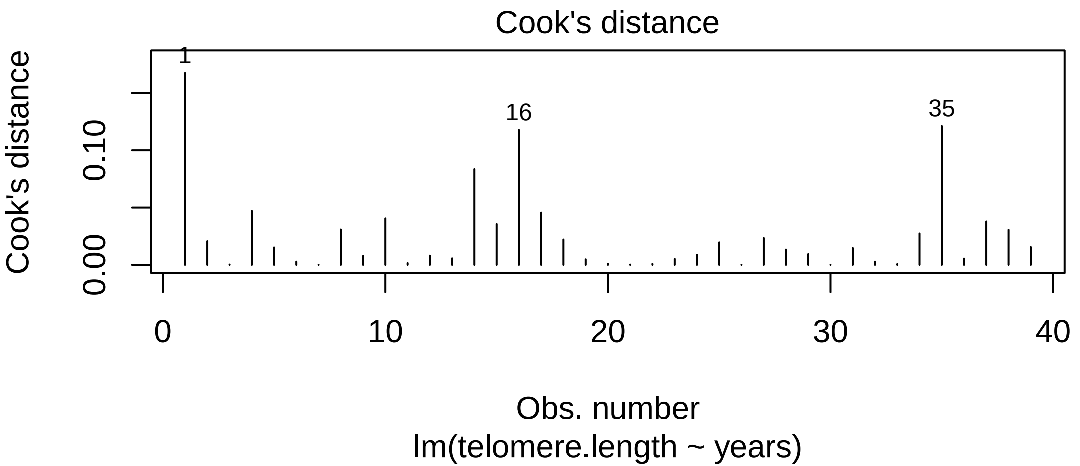
Assumption	Violation
Normality	Points don't generally fall along the line

Absolute standardized residuals vs fitted values



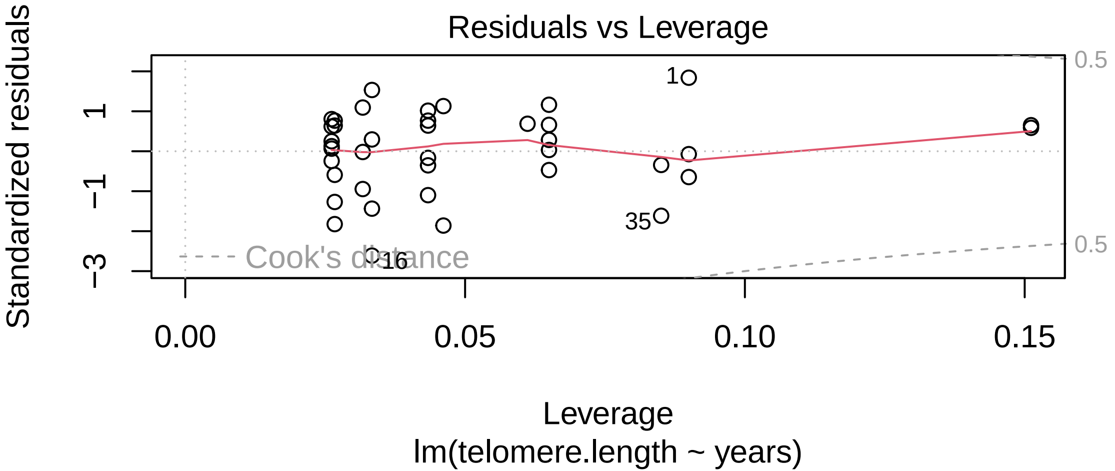
Assumption	Violation
Constant variance	Increasing (or decreasing) trend

Cook's distance



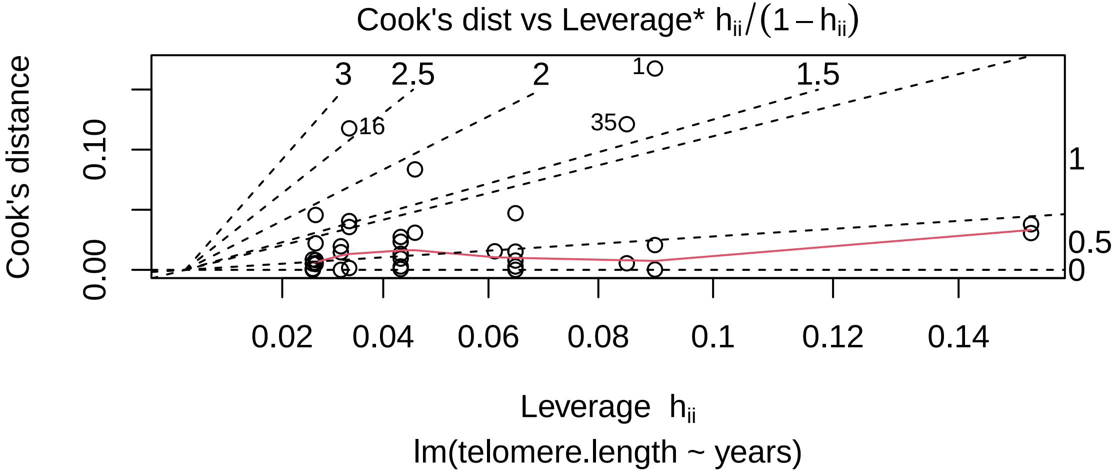
Outlier	Violation
Influential observation	Cook's distance larger than (1 or 4/n)

Residuals vs leverage



Outlier	Violation
Influential observation	Points outside red dashed lines

Cooks' distance vs leverage



This plot is pretty confusing.

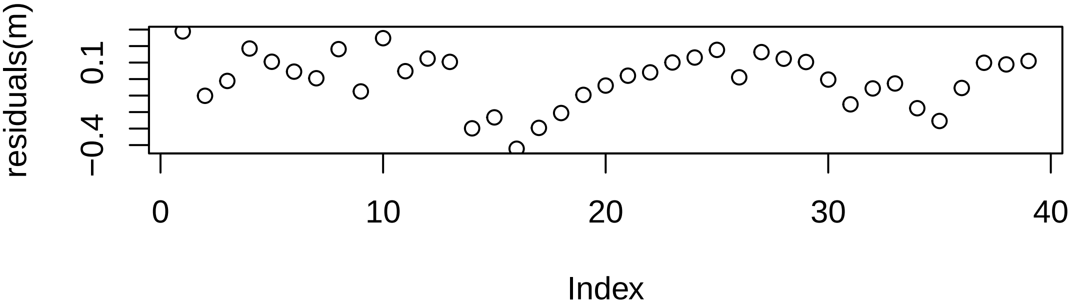
Additional plots

Default plots do not assess all model assumptions.

Two additional suggested plots:

- Residuals vs row number
- Residuals vs (each) explanatory variable

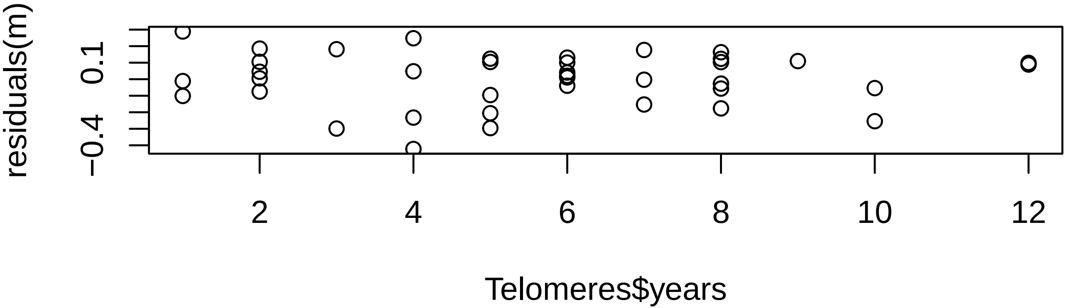
Plot residuals vs row number (index)



```
plot(residuals(m))
```

Assumption	Violation
Independence	A pattern suggests temporal correlation (or an artefact)

Residual vs explanatory variable

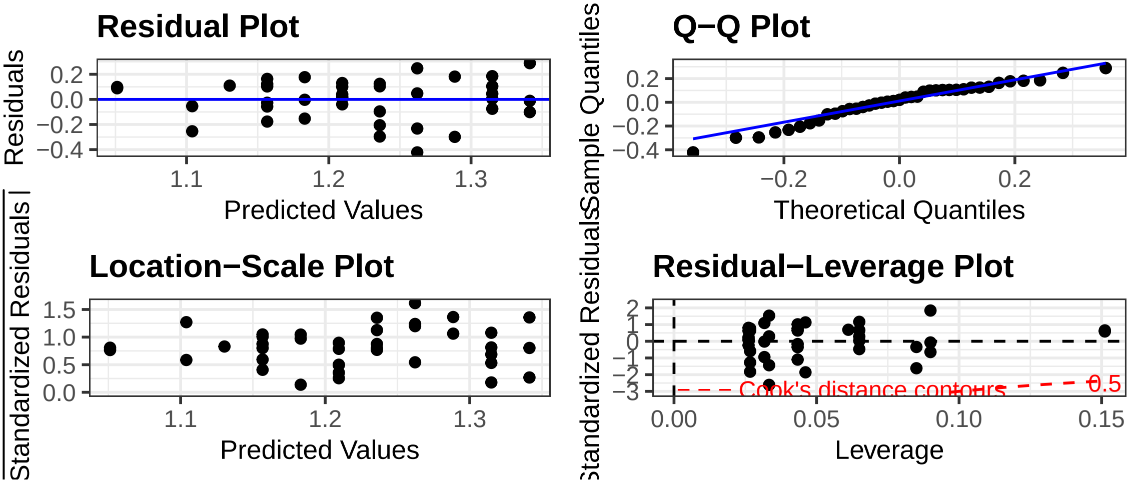


```
plot(Telomeres$years, residuals(m))
```

Assumption	Violation
Linearity	A pattern suggests non-linearity

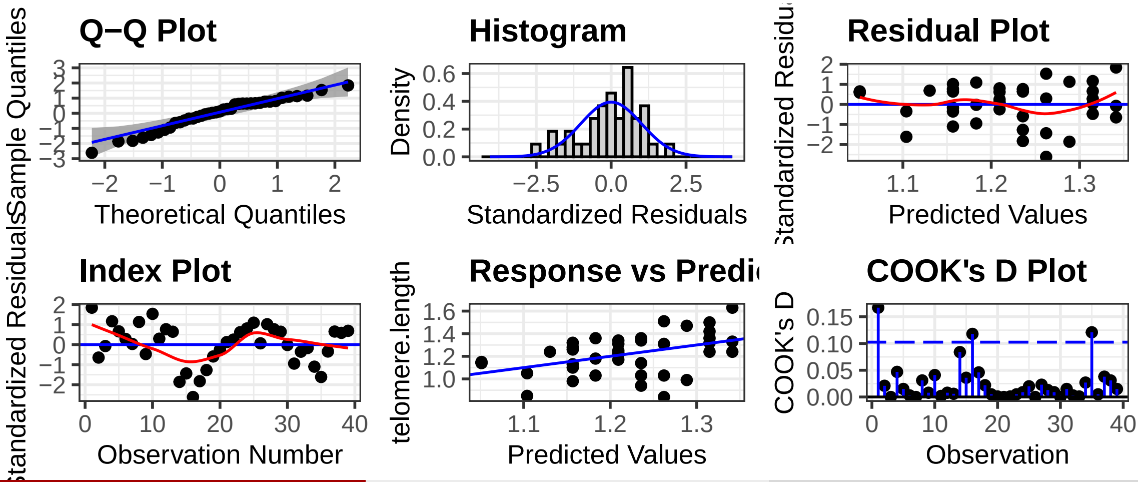
ggResidpanel: R default

```
resid_panel(m, plots = "R")
```



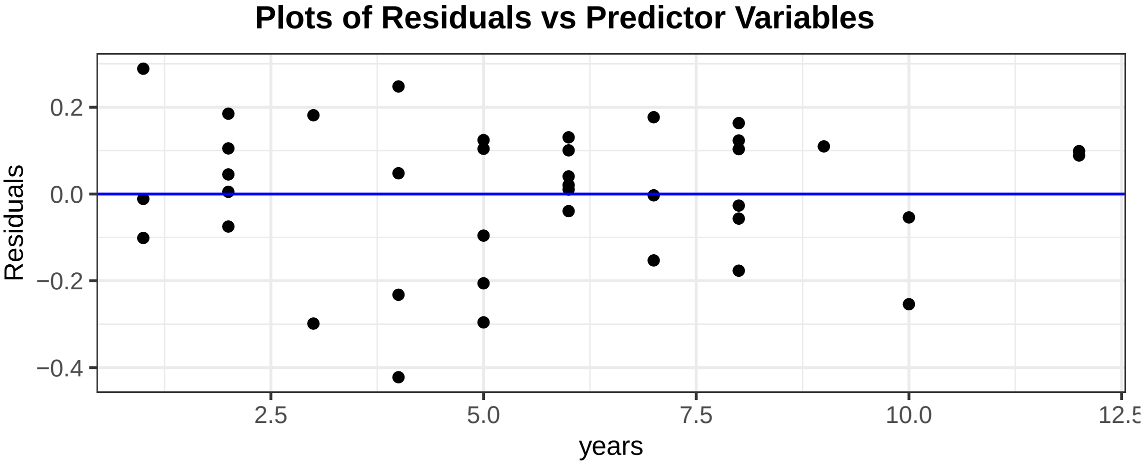
ggResidpanel: R all plots

```
resid_panel(m, plots = c("qq", "hist", "resid", "index", "yvp", "cookd"),
            bins = 30, smoother = TRUE, qqbands = TRUE,
            type = "standardized") # what I was calling studentized
```



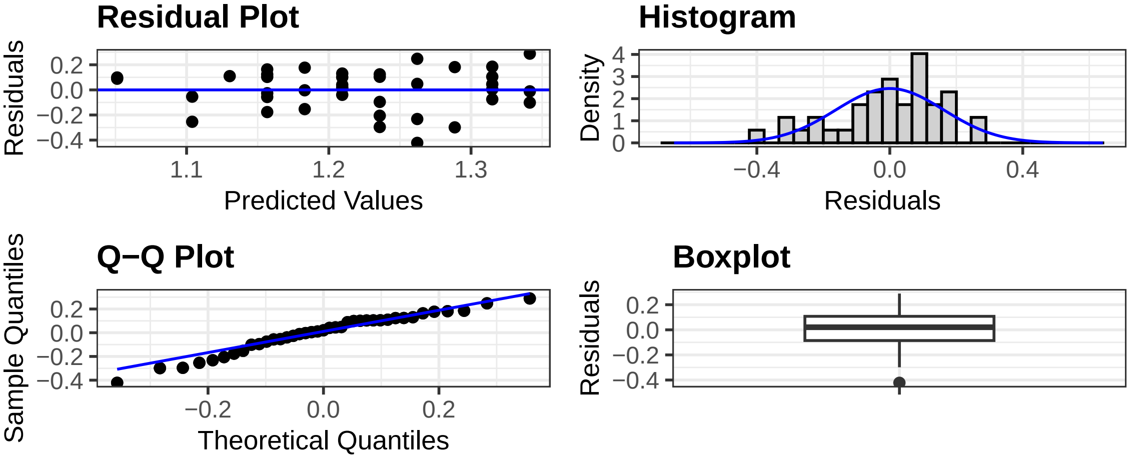
ggResidpanel: R explanatory

```r
resid_xpanel(m)
```



ggResidpanel: SAS

```
resid_panel(m, plots = "SAS")
```



Summary

Case statistics:

- Fitted values
- Leverage
- Residuals
  - Standardized residuals
  - Studentized residuals
- Cook's distance

Model assumptions:

- Normality
- Constant variance
- Independence
- Linearity

R03 - Regression: using logarithms

STAT 5870 (Engineering)
Iowa State University

November 8, 2024

Parameter interpretation in regression

If

$$E[Y|X] = \beta_0 + \beta_1 X,$$

then

- β_0 is the expected response when X is zero and
- $d\beta_1$ is the expected (additive) increase in the response for a d unit (additive) increase in the explanatory variable.

For the following discussion,

- Y is always going to be the original response and
- X is always going to be the original explanatory variable.

Corn yield example

Suppose

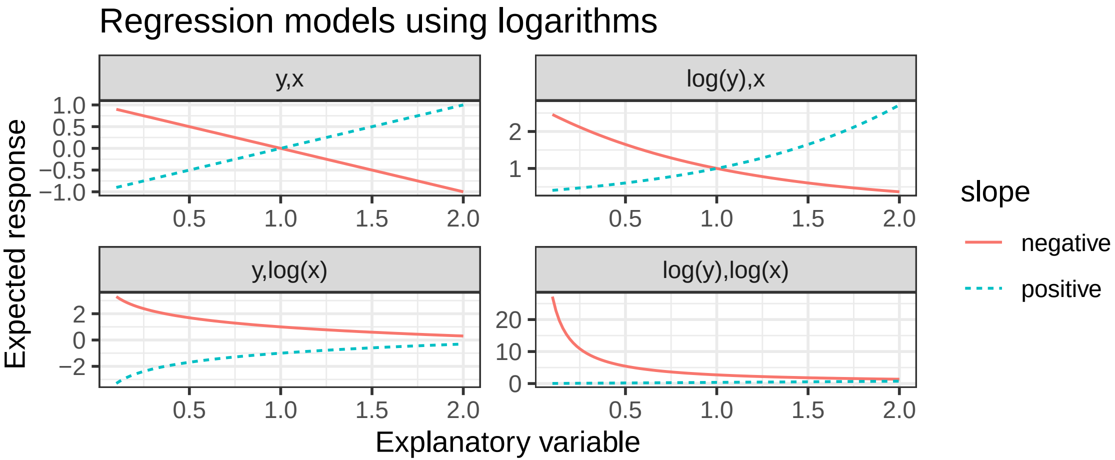
- Y is corn yield (bushels/acre)
- X is fertilizer level in lbs/acre

Then, if

$$E[Y|X] = \beta_0 + \beta_1 X$$

- β_0 is the expected corn yield (bushels/acre) when fertilizer level is zero and
- $d\beta_1$ is the expected increase in corn yield (bushels/acre) when fertilizer is increased by d lbs/acre.

Regression with logarithms (plotted on the original scale)



Response is logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 X,$$

then we have

$$\text{Median}[Y|X] = e^{\beta_0+\beta_1 X} = e^{\beta_0}e^{\beta_1 X}$$

then

- e^{β_0} is the median of Y when X is zero
- $e^{d\beta_1}$ is the multiplicative increase in the median of Y for a d unit (additive) increase in the explanatory variable.

Response is logged

Let be Y is corn yield (bushels/acre) and X is fertilizer level in lbs/acre.
If we assume

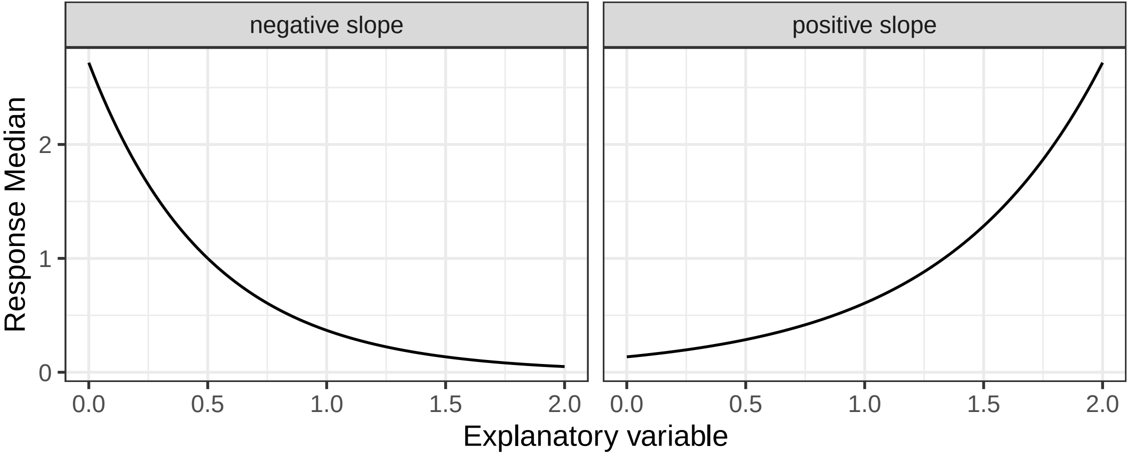
$$E[\log(Y)|X] = \beta_0 + \beta_1 X$$

then

$$\text{Median}[Y|X] = e^{\beta_0} e^{\beta_1 X}$$

- e^{β_0} is the median corn yield (bushels/acre) when fertilizer level is 0 (lbs/acre) and
- $e^{d\beta_1}$ is the multiplicative increase in median corn yield (bushels/acre) when fertilizer is increased by d lbs/acre.

Response is logged



Explanatory variable is logged

If

$$E[Y|X] = \beta_0 + \beta_1 \log(X),$$

then,

- β_0 is the expected response when X is 1 and
- $\beta_1 \log(d)$ is the expected (additive) increase in the response when X increases multiplicatively by d,e.g.
  - $\beta_1 \log(2)$ is the expected (additive) increase in the response for each doubling of X or
  - $\beta_1 \log(10)$ is the expected (additive) increase in the response for each ten-fold increase in X.

Explanatory variable is logged

Suppose

- Y is corn yield (bushels/acre)
- X is fertilizer level in lbs/acre

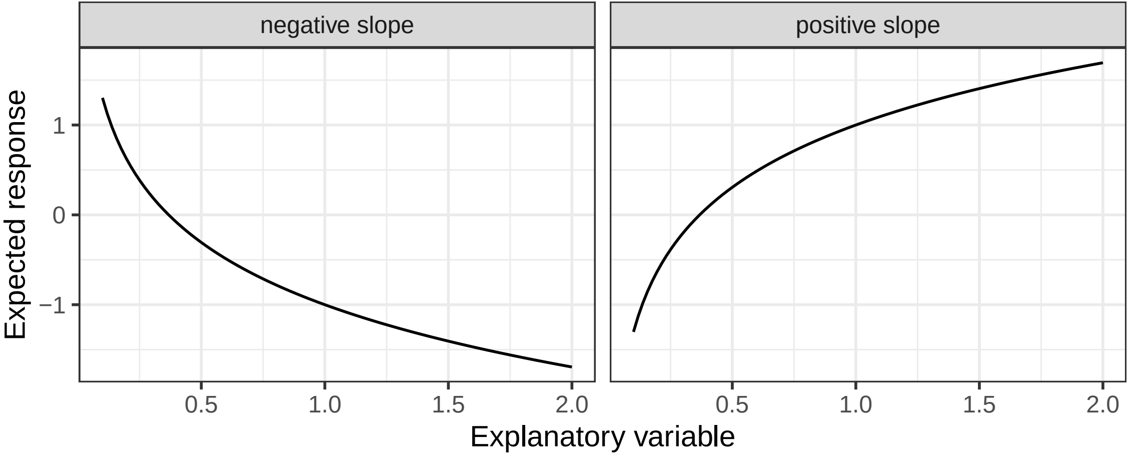
If

$$E[Y|X] = \beta_0 + \beta_1 \log(X)$$

then

- β_0 is the expected corn yield (bushels/acre) when fertilizer level is 1 lb/acre and
- $\beta_1 \log(2)$ is the expected (additive) increase in corn yield when fertilizer level is doubled.

Explanatory variable is logged



Both response and explanatory variable are logged

If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X),$$

then

$$\text{Median}[Y|X] = e^{\beta_0} X^{\beta_1},$$

and thus

- e^{β_0} is the median of Y when X is 1 and
- d^{β_1} is the multiplicative increase in the median of the response when X increases multiplicatively by d, e.g.
  - 2^{β_1} is the multiplicative increase in the median of the response for each doubling of X or
  - 10^{β_1} is the multiplicative increase in the median of the response for each ten-fold increase in X.

Both response and explanatory variables are logged

Suppose

- Y is corn yield (bushels/acre)
- X is fertilizer level in lbs/acre

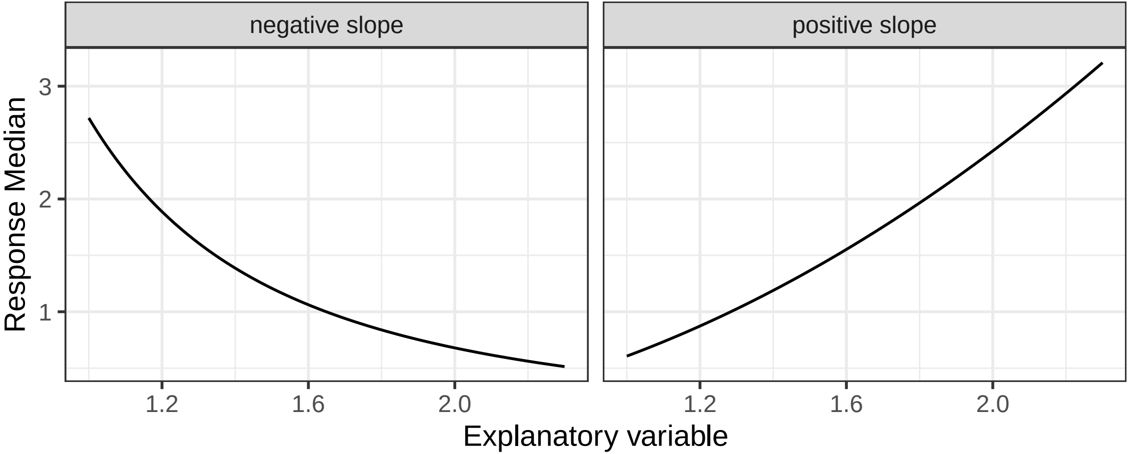
If

$$E[\log(Y)|X] = \beta_0 + \beta_1 \log(X) \quad \text{or} \quad \text{Median}[Y|X] = e^{\beta_0} e^{\beta_1 \log(X)} = e^{\beta_0} X^{\beta_1},$$

then

- e^{β_0} is the median corn yield (bushels/acre) at 1 lb/acre of fertilizer and
- 2^{β_1} is the multiplicative increase in median corn yield (bushels/acre) when fertilizer is doubled.

Both response and explanatory variables are logged



Why use logarithms

The most common transformation of either the response or explanatory variable(s) is to take logarithms because

- linearity will often then be approximately true,
- the variance will likely be approximately constant,
- influence of some observations may decrease, and
- there is a (relatively) convenient interpretation.

Summary of interpretations when using logarithms

- When using the log of the response,
  - β_0 determines the median response
  - β_1 determines the multiplicative increase in the median response
- When using the log of the explanatory variable (X),
  - β_0 determines the response when $X = 1$
  - β_1 determines the increase in the response when there is a multiplicative increase in X

Constructing credible intervals

Recall the model

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2).$$

Let (L, U) be a $100(1-a)\%$ credible interval for β.

For ease of interpretation, it is often convenient to calculate functions of β, e.g.

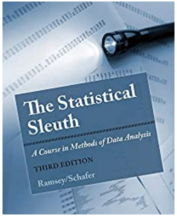
$$f(\beta) = d\beta \qquad \text{and} \qquad f(\beta) = e^{\beta}.$$

A $100(1-a)\%$ credible interval for $f(\beta)$ (when f is monotonic) is

$$(f(L), f(U)).$$

Breakdown times

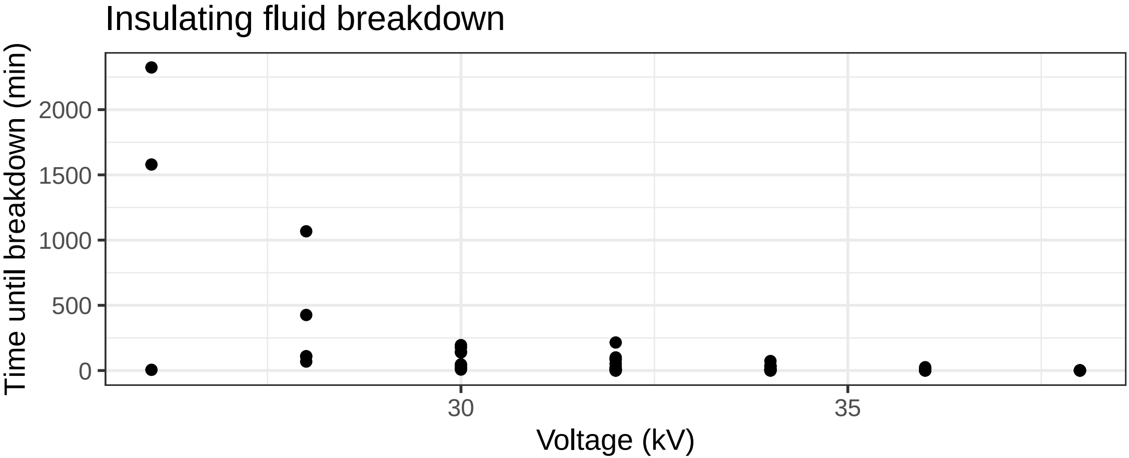
In an industrial laboratory, under uniform conditions, batches of electrical insulating fluid were subjected to constant voltages (kV) until the insulating property of the fluids broke down. Seven different voltage levels were studied and the measured responses were the times (minutes) until breakdown.



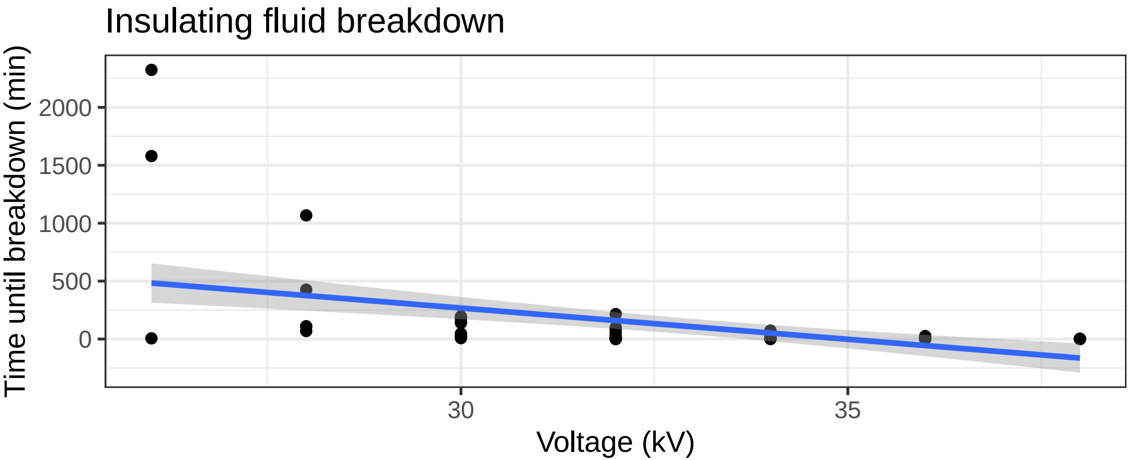
```
summary(Sleuth3::case0802)
```

```
      Time              Voltage          Group   
 Min.   :   0.090   Min.   :26.00   Group1: 3  
 1st Qu.:   1.617   1st Qu.:31.50   Group2: 5  
 Median :   6.925   Median :34.00   Group3:11  
 Mean   :  98.558   Mean   :33.13   Group4:15  
 3rd Qu.:  38.383   3rd Qu.:36.00   Group5:19  
 Max.   :2323.700   Max.   :38.00   Group6:15  
                                    Group7: 8  
```

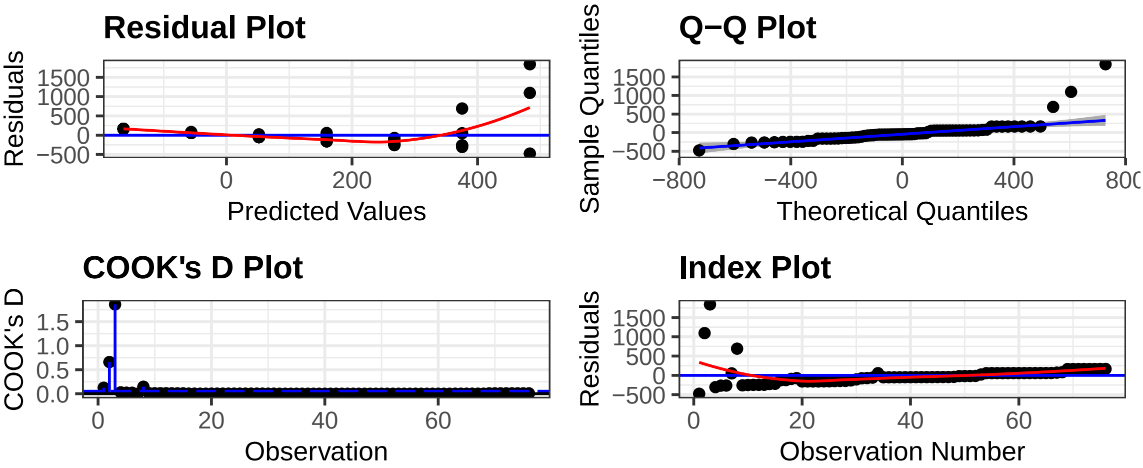
Insulating fluid breakdown



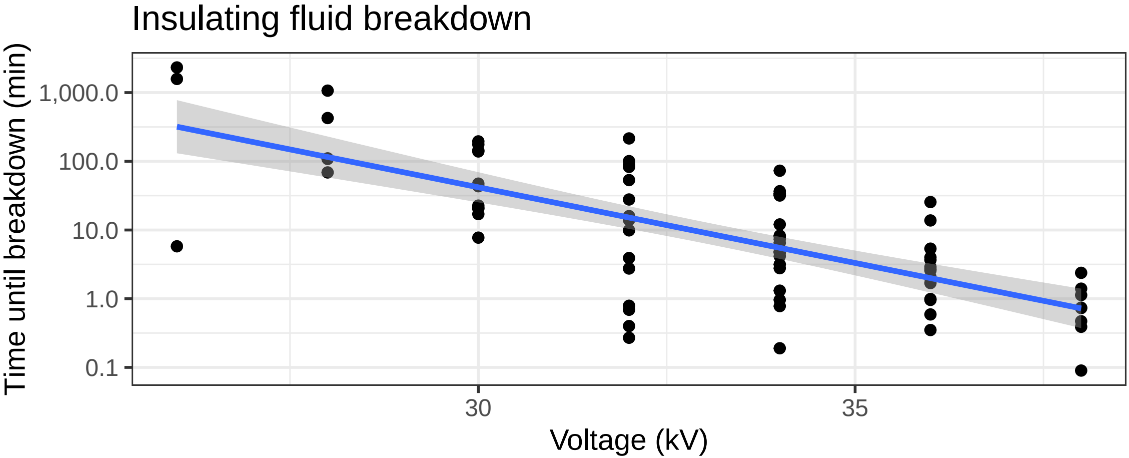
Insulating fluid breakdown



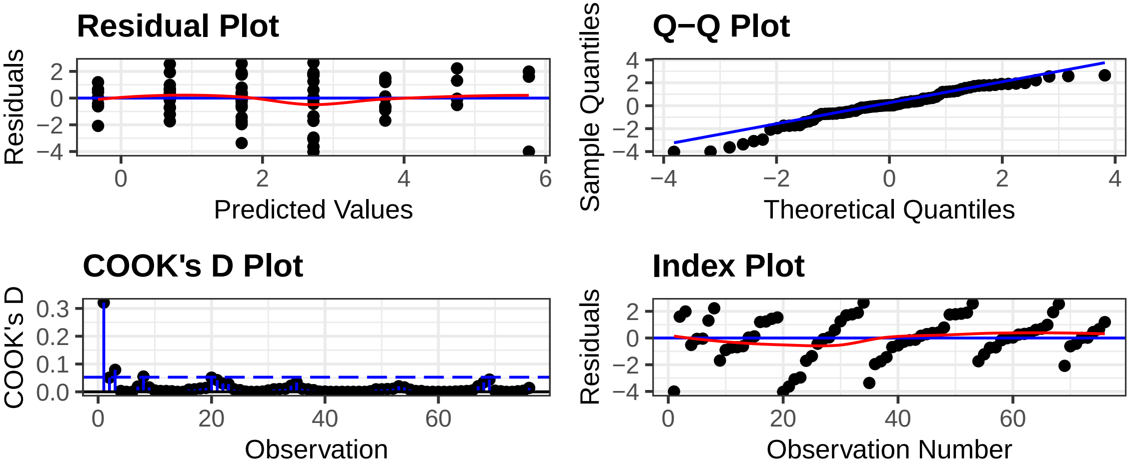
Run the regression and look at diagnostics



Logarithm of time (response)



Logarithm of time (response): residuals



Summary

```
m <- lm(log(Time) ~ I(Voltage-30), Sleuth3::case0802)
exp(m$coefficients)

    (Intercept) I(Voltage - 30)
       41.86752         0.60208

exp(confint(m))

                     2.5 %     97.5 %
(Intercept)     25.2582342 69.3987157
I(Voltage - 30)  0.5370152  0.6750281
```

- At 30 kV, the median breakdown time is estimated to be 42 minutes with a 95% credible interval of (25, 69).
- Each 1 kV increase in voltage was associated with a 40% (32%, 46%) reduction in median breakdown time.

R04 - Regression with Categorical Explanatory Variables

STAT 5870 (Engineering)
Iowa State University

November 11, 2024

Binary explanatory variable

Recall the simple linear regression model

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2).$$

If we have a binary explanatory variable, i.e. the explanatory variable only has two levels say level A and level B, we can code it as

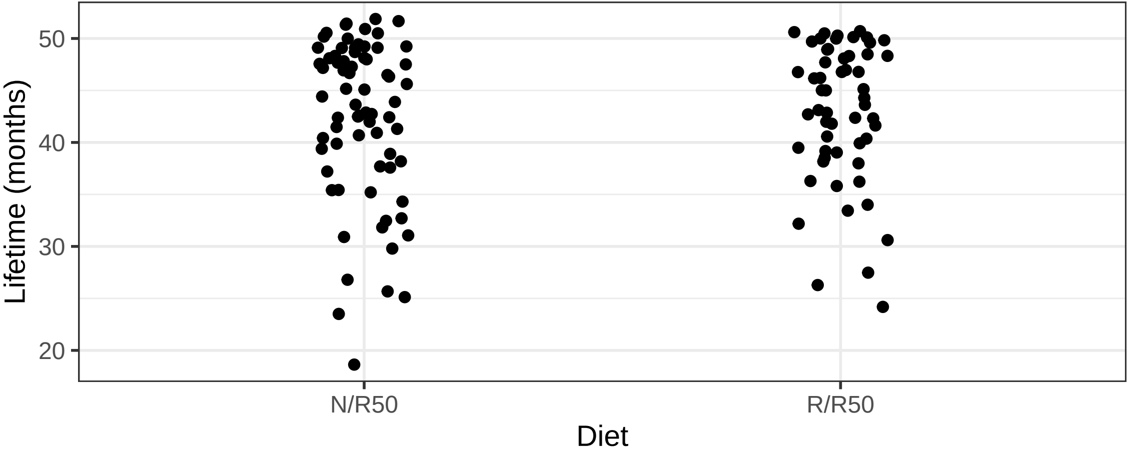
$$X_i = \mathrm{I}(\text{observation } i \text{ is level A})$$

where $\mathrm{I}(statement)$ is an indicator function that is 1 when $statement$ is true and 0 otherwise. Then

- β_0 is the expected response for level B,
- $\beta_0 + \beta_1$ is the expected response for level A, and
- β_1 is the expected difference in response (level A minus level B).

Mice lifetimes

```
Sleuth3::case0501
```



Regression model for mice lifetimes

Let

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_i, \sigma^2)$$

where Y_i is the lifetime of the ith mouse and

$$X_i = \mathrm{I}(Diet_i = \text{N/R50})$$

then

$$\begin{aligned} E[\text{Lifetime}|\text{R/R50}] &= E[Y_i|X_i = 0] &= \beta_0 \\ E[\text{Lifetime}|\text{N/R50}] &= E[Y_i|X_i = 1] &= \beta_0 + \beta_1 \end{aligned}$$

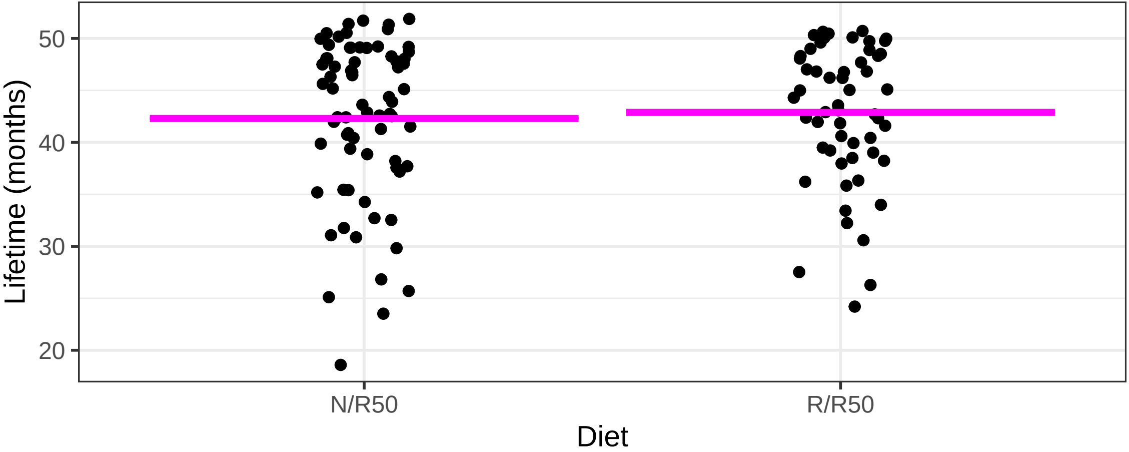
and

$$\begin{aligned} &E[\text{Lifetime difference}] \\ &\quad = E[\text{Lifetime}|\text{N/R50}] - E[\text{Lifetime}|\text{R/R50}] \\ &\quad = (\beta_0 + \beta_1) - \beta_0 = \beta_1. \end{aligned}$$

R code

```
case0501$X <- ifelse(case0501$Diet == "N/R50", 1, 0)
(m <- lm(Lifetime ~ X, data = case0501, subset = Diet %in% c("R/R50","N/R50")))

Call:
lm(formula = Lifetime ~ X, data = case0501, subset = Diet %in%
    c("R/R50", "N/R50"))

Coefficients:
(Intercept)            X
    42.8857      -0.5885

confint(m)

                2.5 %    97.5 %
(Intercept) 40.952257 44.819172
X           -3.174405  1.997342

predict(m, data.frame(X=1), interval = "confidence") # Expected lifetime on N/R50

       fit      lwr     upr
1 42.29718 40.58007 44.0143
```

Mice lifetimes



Equivalence to a two-sample t-test

Recall that our two-sample t-test had the model

$$Y_{ij} \overset{ind}{\sim} N(\mu_j, \sigma^2)$$

for groups $j = 0, 1$. This is equivalent to our current regression model where

$$\begin{aligned} \mu_0 &= \beta_0 \\ \mu_1 &= \beta_0 + \beta_1 \end{aligned}$$

assuming

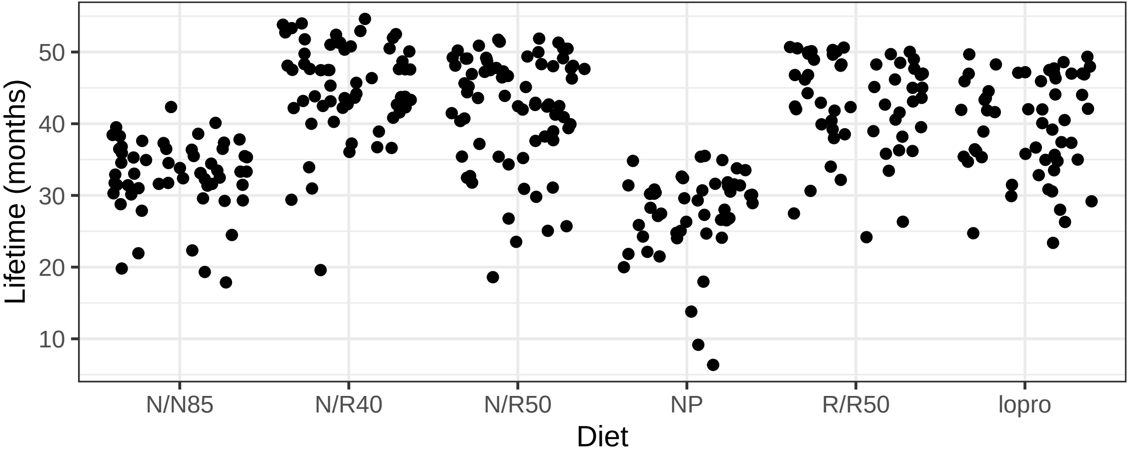
- μ_0 represents the mean for the R/R50 group and
- μ_1 represents the mean for N/R50 group.

When the models are effectively the same, but have different parameters we say the model is reparameterized.

Equivalence

```
summary(m)$coefficients[2,4] # p-value

[1] 0.6531748

confint(m)

                2.5 %    97.5 %
(Intercept) 40.952257 44.819172
X           -3.174405  1.997342

t.test(Lifetime ~ Diet, data = case0501, subset = Diet %in% c("R/R50","N/R50"), var.equal=TRUE)

Two Sample t-test

data:  Lifetime by Diet
t = -0.45044, df = 125, p-value = 0.6532
alternative hypothesis: true difference in means between group N/R50 and group R/R50 is not equal to 0
95 percent confidence interval:
 -3.174405  1.997342
sample estimates:
mean in group N/R50 mean in group R/R50
           42.29718            42.88571
```

Using a categorical variable as an explanatory variable.



Regression with a categorical variable

1. Choose one of the levels as the reference level, e.g. N/N85
2. Construct dummy variables using indicator functions, i.e.

$$\mathrm{I}(A) = \begin{cases} 1 & A \text{ is TRUE} \\ 0 & A \text{ is FALSE} \end{cases}$$

   for the other levels, e.g.

$$\begin{aligned} X_{i,1} &= \mathrm{I}(\text{diet for observation } i \text{ is N/R40}) \\ X_{i,2} &= \mathrm{I}(\text{diet for observation } i \text{ is N/R50}) \\ X_{i,3} &= \mathrm{I}(\text{diet for observation } i \text{ is NP}) \\ X_{i,4} &= \mathrm{I}(\text{diet for observation } i \text{ is R/R50}) \\ X_{i,5} &= \mathrm{I}(\text{diet for observation } i \text{ is lopro}) \end{aligned}$$

3. Estimate the parameters of a multiple regression model using these dummy variables.

Regression model

Our regression model becomes

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,3} + \beta_4 X_{i,4} + \beta_5 X_{i,5}, \sigma^2)$$

where

- β_0 is the expected lifetime for the N/N85 group
- $\beta_0 + \beta_1$ is the expected lifetime for the N/R40 group
- $\beta_0 + \beta_2$ is the expected lifetime for the N/R50 group
- $\beta_0 + \beta_3$ is the expected lifetime for the NP group
- $\beta_0 + \beta_4$ is the expected lifetime for the R/R50 group
- $\beta_0 + \beta_5$ is the expected lifetime for the lopro group

and thus β_p for $p > 0$ is the difference in expected lifetimes between one group and a reference group.

R code

```
case0501 <- case0501 |>
  mutate(X1 = Diet == "N/R40",
         X2 = Diet == "N/R50",
         X3 = Diet == "NP",
         X4 = Diet == "R/R50",
         X5 = Diet == "lopro")

m <- lm(Lifetime ~ X1 + X2 + X3 + X4 + X5, data = case0501)
m


Call:
lm(formula = Lifetime ~ X1 + X2 + X3 + X4 + X5, data = case0501)

Coefficients:
(Intercept)       X1TRUE       X2TRUE       X3TRUE       X4TRUE       X5TRUE
     32.691       12.425        9.606       -5.289       10.194        6.994

confint(m)


                2.5 %    97.5 %
(Intercept) 30.951394 34.431062
X1TRUE       9.995893 14.854984
X2TRUE       7.269897 11.942013
X3TRUE      -7.848142 -2.730232
X4TRUE       7.723030 12.665943
X5TRUE       4.523030  9.465943
```

R code (cont.)

```
summary(m)

Call:
lm(formula = Lifetime ~ X1 + X2 + X3 + X4 + X5, data = case0501)

Residuals:
     Min       1Q   Median       3Q      Max
-25.5167  -3.3857   0.8143   5.1833  10.0143

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  32.6912     0.8846  36.958  < 2e-16 ***
X1TRUE       12.4254     1.2352  10.059  < 2e-16 ***
X2TRUE        9.6060     1.1877   8.088 1.06e-14 ***
X3TRUE       -5.2892     1.3010  -4.065 5.95e-05 ***
X4TRUE       10.1945     1.2565   8.113 8.88e-15 ***
X5TRUE        6.9945     1.2565   5.567 5.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.678 on 343 degrees of freedom
Multiple R-squared:  0.4543,Adjusted R-squared:  0.4463
F-statistic:  57.1 on 5 and 343 DF,  p-value: < 2.2e-16
```

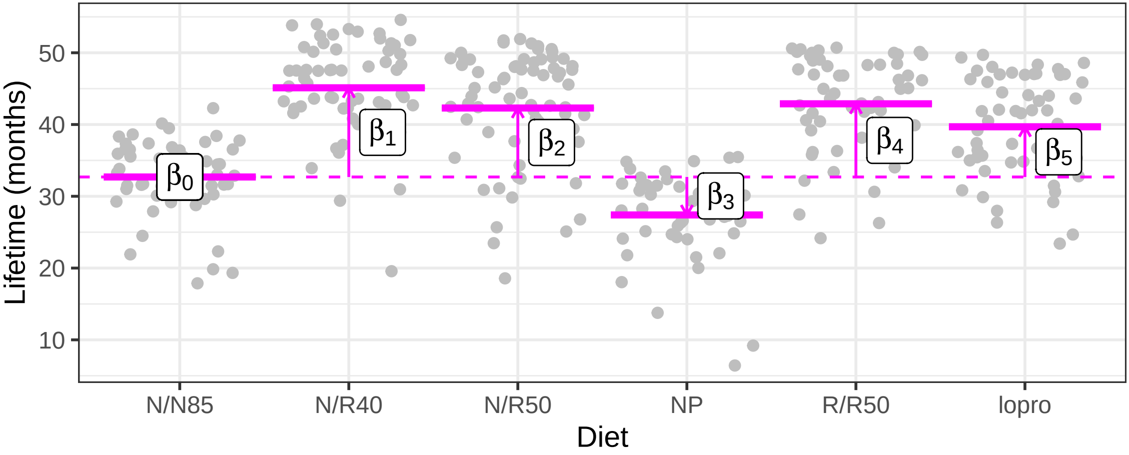
Interpretation

- $\beta_0 = E[Y_i|\text{reference level}]$, i.e. expected response for the reference level
  Note: the only way $X_{i,1} = \cdots = X_{i,p} = 0$ is if all indicators are zero, i.e. at the reference level.
- $\beta_p, p > 0$: expected change in the response moving from the reference level to the level associated with the p^{th} dummy variable
  Note: the only way for $X_{i,p}$ to increase by one is if initially $X_{i,1} = \cdots = X_{i,p} = 0$ and now $X_{i,p} = 1$

For example,

- The expected lifetime for mice on the N/N85 diet is 32.7 (31.0,34.4) months.
- The expected increase in lifetime for mice on the N/R40 diet compared to the N/N85 diet is 12.4 (10.0,14.9) months.
- The model explains 45% of the variability in mice lifetimes.

Using a categorical variable as an explanatory variable.



Equivalence to multiple group model

Recall that we had a multiple group model

$$Y_{ij} \overset{ind}{\sim} N(\mu_j, \sigma^2)$$

for groups $j = 0, 1, 2, \ldots, 5$.

Our regression model is a reparameterization of the multiple group model:

$$\begin{array}{lll} N/N85: & \mu_0 & = \beta_0 \\ N/R40: & \mu_1 & = \beta_0 + \beta_1 \\ N/R50: & \mu_2 & = \beta_0 + \beta_2 \\ NP: & \mu_3 & = \beta_0 + \beta_3 \\ R/R50: & \mu_4 & = \beta_0 + \beta_4 \\ lopro: & \mu_5 & = \beta_0 + \beta_5 \end{array}$$

assuming the groups are labeled appropriately.

Summary

1. Choose one of the levels as the reference level.
2. Construct dummy variables using indicator functions for all other levels, e.g.

$$X_i = \mathrm{I}(\text{observation } i \text{ is } <\text{some non-reference level}>).$$

3. Estimate the parameters of a multiple regression model using these dummy variables.

R05 - Multiple Regression

STAT 5870 (Engineering)
Iowa State University

November 22, 2024

Multiple regression

Recall the simple linear regression model is

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2), \quad \mu_i = \beta_0 + \beta_1 X_i$$

The multiple regression model has mean

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

where for observation i

- Y_i is the response and
- $X_{i,p}$ is the p^{th} explanatory variable.

Explanatory variables

There is a lot of flexibility in the mean

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

as there are many possibilities for the explanatory variables $X_{i,1}, \ldots, X_{i,p}$:

- Functions ($f(X)$)
- Dummy variables for categorical variables ($X_1 = \mathrm{I}()$)
- Higher order terms (X^2)
- Additional explanatory variables (X_1, X_2)
- Interactions ($X_1 X_2$)
  - Continuous-continuous
  - Continuous-categorical
  - Categorical-categorical

Parameter interpretation

Model:

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}, \sigma^2)$$

The interpretation is

- β_0 is the expected value of the response Y_i when all explanatory variables are zero.
- β_p, $p \neq 0$ is the expected increase in the response for a one-unit increase in the p^{th} explanatory variable when all other explanatory variables are held constant.
- R^2 is the proportion of the variability in the response explained by the model

Parameter estimation and inference

Let

$$y = X\beta + \epsilon$$

where

- $y = (y_1, \ldots, y_n)^\top$
- X is $n \times p$ with ith row $X_i = (1, X_{i,1}, \ldots, X_{i,p})$
- $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$
- $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^\top$

Then we have

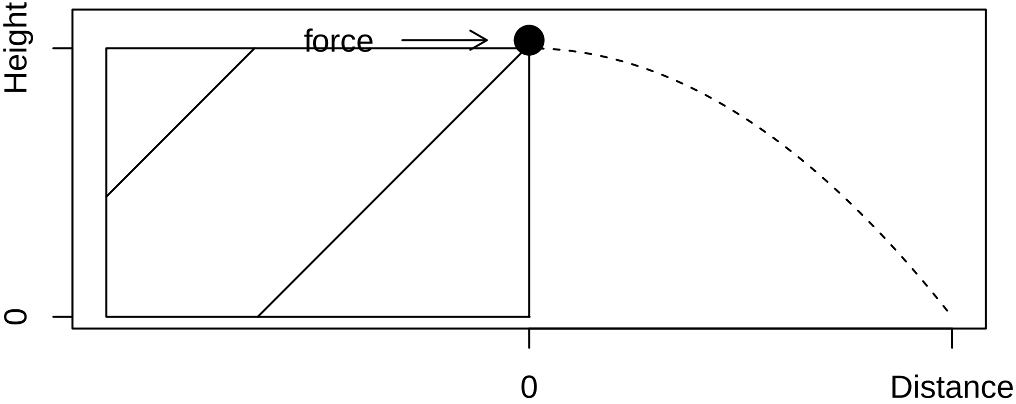
$$\begin{aligned} \hat{\beta} &= (X^\top X)^{-1} X^\top y \\ Var(\hat{\beta}) &= \sigma^2 (X^\top X)^{-1} \\ r &= y - X\hat{\beta} \\ \hat{\sigma}^2 &= \frac{1}{n-(p+1)} r^\top r \end{aligned}$$

Confidence/credible intervals and (two-sided) p-values are constructed using

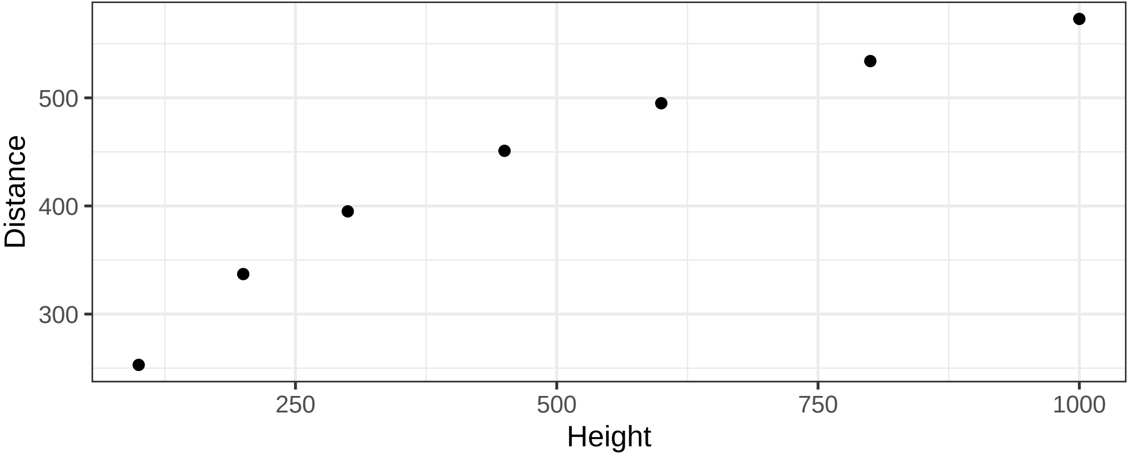
$$\hat{\beta}_j \pm t_{n-(p+1),1-a/2} SE(\hat{\beta}_j) \quad \text{and} \quad \text{pvalue} = 2P\left(T_{n-(p+1)} > \left|\frac{\hat{\beta}_j - b_j}{SE(\hat{\beta}_j)}\right|\right)$$

where $T_{n-(p+1)} \sim t_{n-(p+1)}$ and $SE(\hat{\beta}_j)$ is the jth diagonal element of $\hat{\sigma}^2 (X^\top X)^{-1}$.

Galileo experiment



Galileo data (`Sleuth3::case1001`)



Higher order terms (X^2)

Let

- Y_i be the distance for the i^{th} run of the experiment and
- H_i be the height for the i^{th} run of the experiment.

Simple linear regression assumes

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 H_i \qquad\qquad\qquad\qquad , \sigma^2)$$

The quadratic multiple regression assumes

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 H_i + \beta_2 H_i^2 \qquad\qquad , \sigma^2)$$

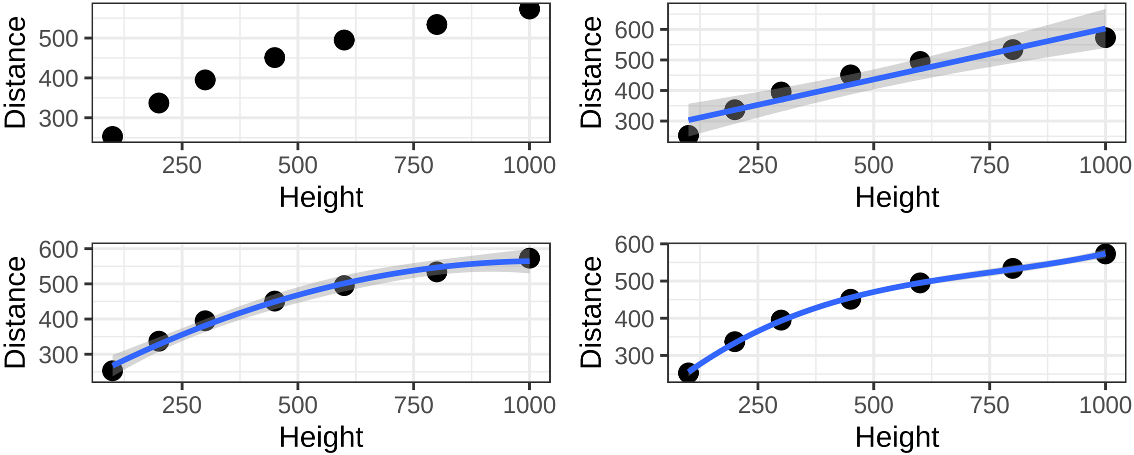
The cubic multiple regression assumes

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 H_i + \beta_2 H_i^2 + \beta_3 H_i^3, \sigma^2)$$

R code and output

```
# Construct the variables by hand
m1 = lm(Distance ~ Height,                                  case1001)
m2 = lm(Distance ~ Height + I(Height^2),                    case1001)
m3 = lm(Distance ~ Height + I(Height^2) + I(Height^3), case1001)

coefficients(m1)

(Intercept)      Height
 269.712458    0.333337

coefficients(m2)

  (Intercept)        Height   I(Height^2)
 1.999128e+02  7.083225e-01 -3.436937e-04

coefficients(m3)

  (Intercept)        Height   I(Height^2)   I(Height^3)
 1.557755e+02  1.115298e+00 -1.244943e-03  5.477104e-07
```

Galileo experiment (Sleuth3::case1001)



Longnose Dace Abundance

From `http://udel.edu/~mcdonald/statmultreg.html`:

> *I extracted some data from the Maryland Biological Stream Survey. ... The [response] variable is the number of Longnose Dace ... per 75-meter section of [a] stream. The [explanatory] variables are ... the maximum depth (in cm) of the 75-meter segment of stream; nitrate concentration (mg/liter) ....*

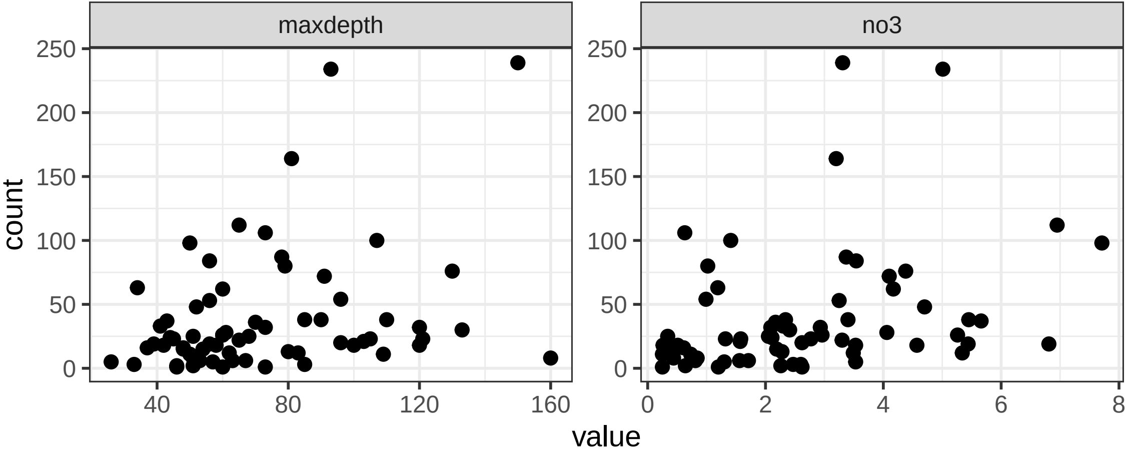
Consider the model

$$Y_i \stackrel{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}, \sigma^2)$$

where

- Y_i: count of Longnose Dace in stream i
- $X_{i,1}$: maximum depth (in cm) of stream i
- $X_{i,2}$: nitrate concentration (mg/liter) of stream i

Exploratory



R code and output

```
m <- lm(count ~ maxdepth + no3, longnosedace)
summary(m)


Call:
lm(formula = count ~ maxdepth + no3, data = longnosedace)

Residuals:
    Min      1Q  Median      3Q     Max 
-55.060 -27.704  -8.679  11.794 165.310 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -17.5550    15.9586  -1.100  0.27544   
maxdepth      0.4811     0.1811   2.656  0.00997 **
no3           8.2847     2.9566   2.802  0.00671 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.39 on 64 degrees of freedom
Multiple R-squared:  0.1936,Adjusted R-squared:  0.1684 
F-statistic: 7.682 on 2 and 64 DF,  p-value: 0.001022
```

Interpretation

- Intercept (β_0): The expected count of Longnose Dace when maximum depth and nitrate concentration are both zero is -18.
- Coefficient for maxdepth (β_1): Holding nitrate concentration constant, each cm increase in maximum depth is associated with an additional 0.48 Longnose Dace counted on average.
- Coefficient for no3 (β_2): Holding maximum depth constant, each mg/liter increase in nitrate concentration is associated with an addition 8.3 Longnose Dace counted on average.
- Coefficient of determination (R^2): The model explains 19% of the variability in the count of Longnose Dace.

Interactions

Why an interaction?

> *Two explanatory variables are said to interact if the effect that one of them has on the mean response depends on the value of the other.*

For example,

- Longnose dace count: The effect of nitrate (no3) on longnose dace count depends on the maxdepth. (Continuous-continuous)
- Energy expenditure: The effect of mass depends on the species type. (Continuous-categorical)
- Crop yield: the effect of tillage method depends on the fertilizer brand (Categorical-categorical)

Continuous-continuous interaction

For observation i, let

- Y_i be the response
- $X_{i,1}$ be the first explanatory variable and
- $X_{i,2}$ be the second explanatory variable.

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2}.$$

The mean with the interaction is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \beta_3 X_{i,1} X_{i,2}.$$

Intepretation - main effects only

Let $X_{i,1} = x_1$ and $X_{i,2} = x_2$, then we can rewrite the line (μ) as

$$\mu = (\beta_0 + \beta_2 x_2) + \beta_1 x_1$$

which indicates that the intercept of the line for x_1 depends on the value of x_2.

Similarly,

$$\mu = (\beta_0 + \beta_1 x_1) + \beta_2 x_2$$

which indicates that the intercept of the line for x_2 depends on the value of x_1.

Intepretation - with an interaction

Let $X_{i,1} = x_1$ and $X_{i,2} = x_2$, then we can rewrite the mean (μ) as

$$\mu = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

which indicates that both the intercept and slope for x_1 depend on the value of x_2.

Similarly,

$$\mu = (\beta_0 + \beta_1 x_1) + (\beta_2 + \beta_3 x_1) x_2$$

which indicates that both the intercept and slope for x_2 depend on the value of x_1.

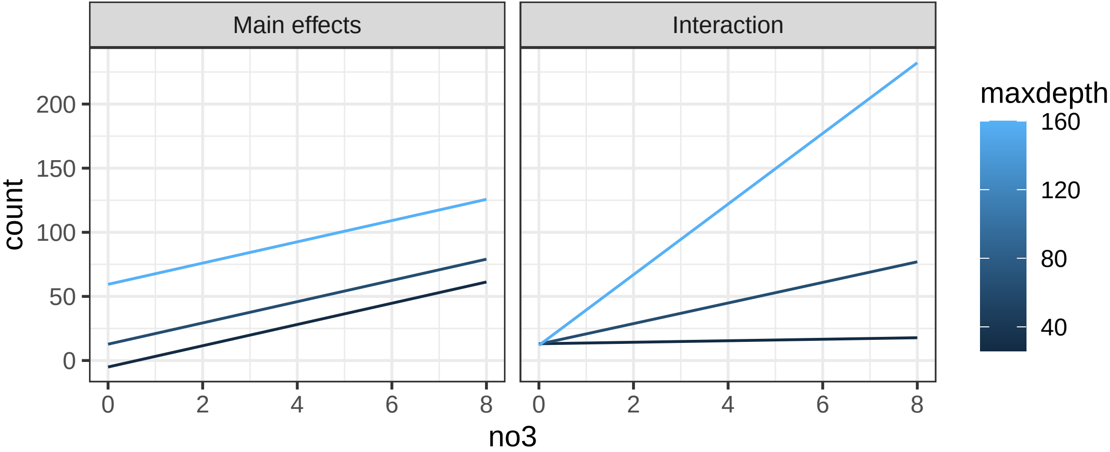
R code and output - main effects only

```
Call:
lm(formula = count ~ no3 + maxdepth, data = longnosedace)

Residuals:
    Min      1Q  Median      3Q     Max 
-55.060 -27.704  -8.679  11.794 165.310 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) -17.5550    15.9586  -1.100  0.27544   
no3           8.2847     2.9566   2.802  0.00671 **
maxdepth      0.4811     0.1811   2.656  0.00997 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 43.39 on 64 degrees of freedom
Multiple R-squared:  0.1936,Adjusted R-squared:  0.1684 
F-statistic: 7.682 on 2 and 64 DF,  p-value: 0.001022
```

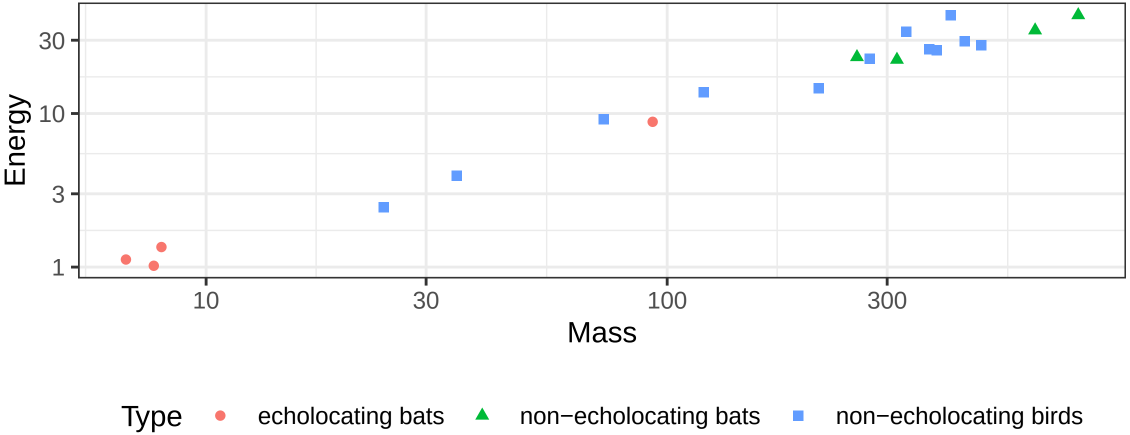
R code and output - with an interaction

```
Call:
lm(formula = count ~ no3 * maxdepth, data = longnosedace)

Residuals:
    Min      1Q  Median      3Q     Max 
-65.111 -21.399  -9.562   5.953 151.071 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)  13.321043  23.455710   0.568   0.5721  
no3          -4.646272   7.856932  -0.591   0.5564  
maxdepth     -0.009338   0.329180  -0.028   0.9775  
no3:maxdepth  0.201219   0.113576   1.772   0.0813 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 42.68 on 63 degrees of freedom
Multiple R-squared:  0.2319,Adjusted R-squared:  0.1953 
F-statistic: 6.339 on 3 and 63 DF,  p-value: 0.0007966
```

Visualizing the model



In-flight energy expenditure (Sleuth3::case1002)



Continuous-categorical interaction

Let category A be the reference level. For observation i, let

- Y_i be the response
- $X_{i,1}$ be the continuous explanatory variable,
- B_i be a dummy variable for category B, and
- C_i be a dummy variable for category C.

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i.$$

The mean with the interaction is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i + \beta_4 X_{i,1} B_i + \beta_5 X_{i,1} C_i.$$

Interpretation for the main effect model

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i.$$

For each category, the line is

Category	Line (μ)		
A	β_0	$+$	$\beta_1 X$
B	$(\beta_0 + \beta_2)$	$+$	$\beta_1 X$
C	$(\beta_0 + \beta_3)$	$+$	$\beta_1 X$

Each category has a different intercept, but a common slope.

Interpretation for the model with an interaction

The model with an interaction is

$$\mu_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 B_i + \beta_3 C_i + \beta_4 X_{i,1} B_i + \beta_5 X_{i,1} C_i$$

For each category, the line is

Category	Line (μ)	
A	β_0	$+\ \beta_1 \qquad X$
B	$(\beta_0 + \beta_2)$	$+(\beta_1 + \beta_4)X$
C	$(\beta_0 + \beta_3)$	$+(\beta_1 + \beta_5)X$

Each category has its own intercept and its own slope.

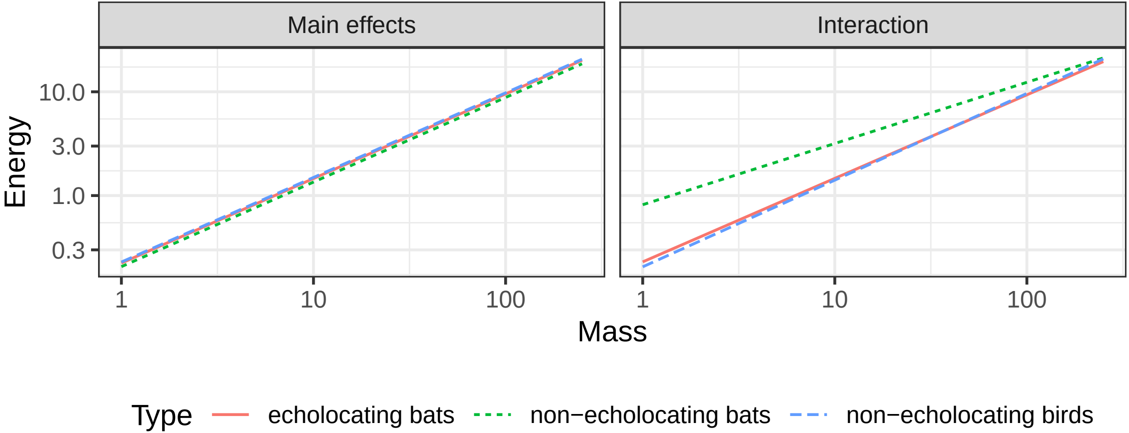
R code and output - main effects only

```
summary(mM <- lm(log(Energy) ~ log(Mass) + Type, case1002))

Call:
lm(formula = log(Energy) ~ log(Mass) + Type, data = case1002)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.23224 -0.12199 -0.03637  0.12574  0.34457 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                -1.49770    0.14987  -9.993 2.77e-08 ***
log(Mass)                   0.81496    0.04454  18.297 3.76e-12 ***
Typenon-echolocating bats  -0.07866    0.20268  -0.388    0.703    
Typenon-echolocating birds  0.02360    0.15760   0.150    0.883    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.186 on 16 degrees of freedom
Multiple R-squared:  0.9815,Adjusted R-squared:  0.9781 
F-statistic: 283.6 on 3 and 16 DF,  p-value: 4.464e-14
```

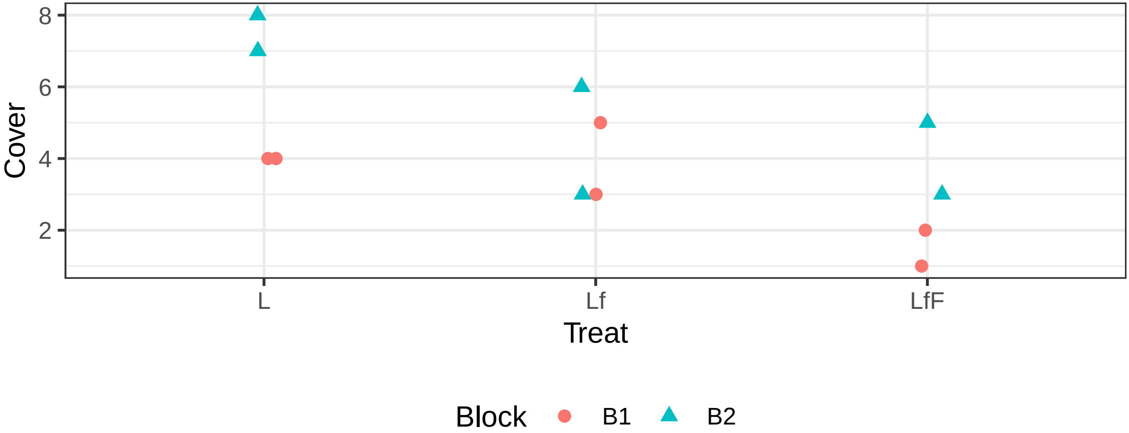
R code and output - with an interaction

```
summary(mI <- lm(log(Energy) ~ log(Mass) * Type, case1002))

Call:
lm(formula = log(Energy) ~ log(Mass) * Type, data = case1002)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.25152 -0.12643 -0.00954  0.08124  0.32840 

Coefficients:
                                     Estimate Std. Error t value Pr(>|t|)    
(Intercept)                          -1.47052    0.24767  -5.937 3.63e-05 ***
log(Mass)                             0.80466    0.08668   9.283 2.33e-07 ***
Typenon-echolocating bats             1.26807    1.28542   0.987    0.341    
Typenon-echolocating birds           -0.11032    0.38474  -0.287    0.779    
log(Mass):Typenon-echolocating bats  -0.21487    0.22362  -0.961    0.353    
log(Mass):Typenon-echolocating birds  0.03071    0.10283   0.299    0.770    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1899 on 14 degrees of freedom
Multiple R-squared:  0.9832,Adjusted R-squared:  0.9771 
F-statistic: 163.4 on 5 and 14 DF,  p-value: 6.696e-12
```

Visualizing the models



Seaweed regeneration (Sleuth3::case1301 subset)



Categorical-categorical

Let category A and type 0 be the reference level. For observation i, let

- Y_i be the response,
- 1_i be a dummy variable for type 1,
- B_i be a dummy variable for category B, and
- C_i be a dummy variable for category C.

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 1_i + \beta_2 B_i + \beta_3 C_i.$$

The mean with an interaction is

$$\mu_i = \beta_0 + \beta_1 1_i + \beta_2 B_i + \beta_3 C_i + \beta_4 1_i B_i + \beta_5 1_i C_i.$$

Interpretation for the main effects model

The mean containing only main effects is

$$\mu_i = \beta_0 + \beta_1 1_i + \beta_2 B_i + \beta_3 C_i.$$

The means in the main effect model are

	Category		
Type	A	B	C
0	β_0	$\beta_0 \quad + \beta_2$	$\beta_0 \quad + \beta_3$
1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2$	$\beta_0 + \beta_1 + \beta_3$

Interpretation for the model with an interaction

The mean with an interaction is

$$\mu_i = \beta_0 + \beta_1 1_i + \beta_2 B_i + \beta_3 C_i + \beta_4 1_i B_i + \beta_5 1_i C_i.$$

The means are

	Category		
Type	A	B	C
0	β_0	$\beta_0 \quad + \beta_2$	$\beta_0 \quad + \beta_3$
1	$\beta_0 + \beta_1$	$\beta_0 + \beta_1 + \beta_2 + \beta_4$	$\beta_0 + \beta_1 + \beta_3 + \beta_5$

This is equivalent to a cell-means model where each combination has its own mean.

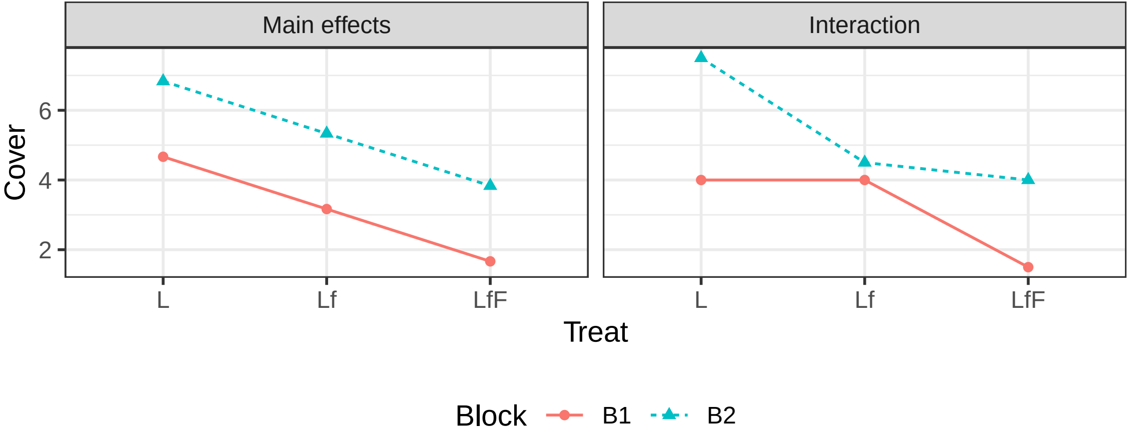
R code and output - main effects only

```
Call:
lm(formula = Cover ~ Block + Treat, data = case1301_subset)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3333 -0.6667  0.0000  0.7917  1.8333 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   4.6667     0.7683   6.074 0.000298 ***
BlockB2       2.1667     0.7683   2.820 0.022491 *  
TreatLf      -1.5000     0.9410  -1.594 0.149578    
TreatLfF     -3.0000     0.9410  -3.188 0.012838 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.331 on 8 degrees of freedom
Multiple R-squared:  0.6937,Adjusted R-squared:  0.5788 
F-statistic: 6.039 on 3 and 8 DF,  p-value: 0.01881
```

R code and output - with an interaction

```
Call:
lm(formula = Cover ~ Block * Treat, data = case1301_subset)

Residuals:
   Min     1Q Median     3Q    Max 
-1.500 -0.625  0.000  0.625  1.500 

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)   
(Intercept)       4.000e+00  8.898e-01   4.496  0.00412 **
BlockB2           3.500e+00  1.258e+00   2.782  0.03193 * 
TreatLf          -4.441e-16  1.258e+00   0.000  1.00000   
TreatLfF         -2.500e+00  1.258e+00  -1.987  0.09413 . 
BlockB2:TreatLf  -3.000e+00  1.780e+00  -1.686  0.14280   
BlockB2:TreatLfF -1.000e+00  1.780e+00  -0.562  0.59450   
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.258 on 6 degrees of freedom
Multiple R-squared:  0.7946,Adjusted R-squared:  0.6234 
F-statistic: 4.642 on 5 and 6 DF,  p-value: 0.04429
```

Visualizing the models



When to include interaction terms

From The Statistical Sleuth (3rd ed) page 250:

- when a question of interest pertains to an interaction
- when good reason exists to suspect an interaction or
- when interactions are proposed as a more general model for the purpose of examining the goodness of fit of a model without interaction.

Multiple regression explanatory variables

The possibilities for explanatory variables are

- Higher order terms (X^2)
- Additional explanatory variables (X_1 and X_2)
- Dummy variables for categorical variables ($X_1 = \mathrm{I}()$)
- Interactions (X_1X_2)
  - Continuous-continuous
  - Continuous-categorical
  - Categorical-categorical

We can also combine these explanatory variables, e.g.

- including higher order terms for continuous variables along with dummy variables for categorical variables and
- including higher order interactions ($X_1X_2X_3$).

R06 - ANOVA and F-tests

STAT 5870 (Engineering)
Iowa State University

November 18, 2024

One-way ANOVA model/assumptions

The one-way ANOVA (ANalysis Of VAriance) model is

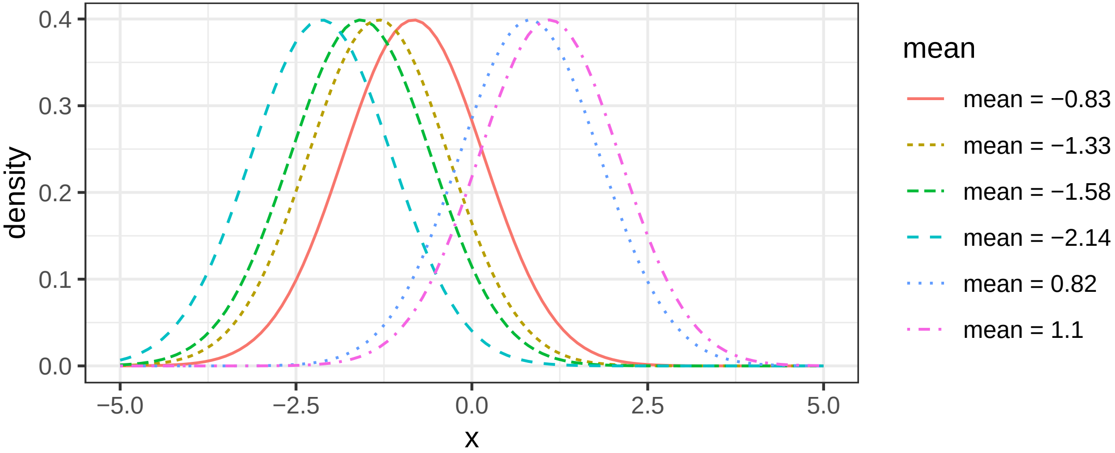
$$Y_{ig} \overset{ind}{\sim} N\left(\mu_g, \sigma^2\right) \quad \text{or} \quad Y_{ig} = \mu_g + \epsilon_{ig},\ \epsilon_{ig} \overset{iid}{\sim} N(0, \sigma^2)$$

for $g = 1, \ldots, G$ and $i = 1, \ldots, n_g$.

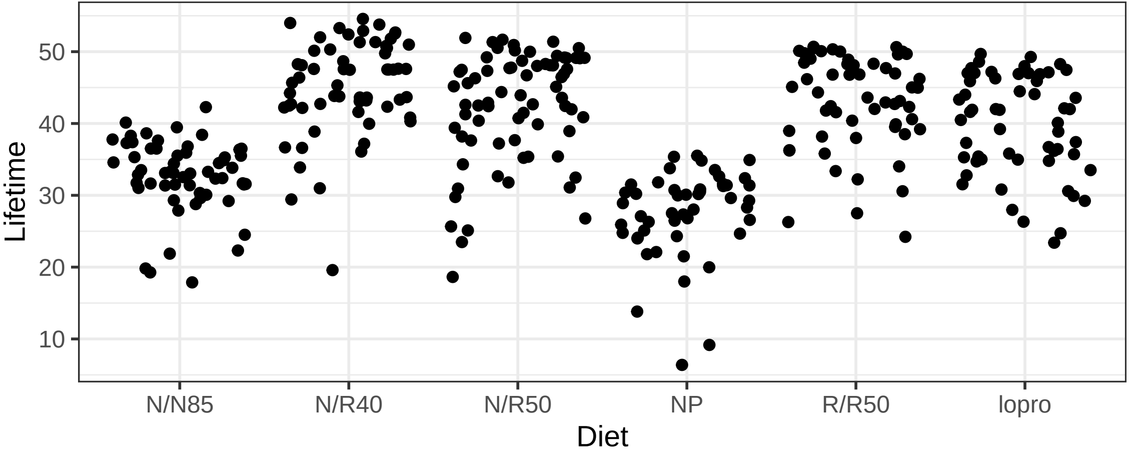
Assumptions:

- Errors
  - are normally distributed.
  - have a common variance.
  - are independent.
- Each group has its own mean.

ANOVA assumptions graphically



Consider the mice data set



One-way ANOVA F-test

Are any of the means different?

Hypotheses in English:

H_0: all the means are the same

H_1: at least one of the means is different

Statistical hypotheses:

$$\begin{array}{rlr} H_0: & \mu_g = \mu \text{ for all } g & Y_{ig} \stackrel{iid}{\sim} N(\mu, \sigma^2) \\ H_1: & \mu_g \neq \mu_{g'} \text{ for some } g \text{ and } g' & Y_{ig} \stackrel{ind}{\sim} N\left(\mu_g, \sigma^2\right) \end{array}$$

An ANOVA table organizes the relevant quantities for this test and computes the pvalue.

ANOVA table

A start of an ANOVA table:

Source of variation	Sum of squares	d.f.	Mean square
Factor A (Between groups)	$SSA = \sum_{g=1}^{G} n_g \left(\overline{Y}_g - \overline{Y}\right)^2$	$G-1$	$\frac{SSA}{G-1}$
Error (Within groups)	$SSE = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(Y_{ig} - \overline{Y}_g\right)^2$	$n-G$	$\frac{SSE}{n-G}$ $(= \hat{\sigma}^2)$
Total	$SST = \sum_{g=1}^{G} \sum_{i=1}^{n_g} \left(Y_{ig} - \overline{Y}\right)^2$	$n-1$	

where

- G is the number of groups,
- n_g is the number of observations in group g,
- $n = \sum_{g=1}^{G} n_g$ (total observations),
- $\overline{Y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} Y_{ig}$ (average in group g),
- and $\overline{Y} = \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{n_g} Y_{ig}$ (overall average).

ANOVA table

An easier to remember ANOVA table:

Source of variation	Sum of squares	df	Mean square	F-statistic	p-value
Factor A (between groups)	SSA	$G-1$	MSA $=$ SSA$/G-1$	MSA$/$MSE	(see below)
Error (within groups)	SSE	$n-G$	MSE $=$ SSE$/n-G$		
Total	SST $=$ SSA$+$SSE	$n-1$			

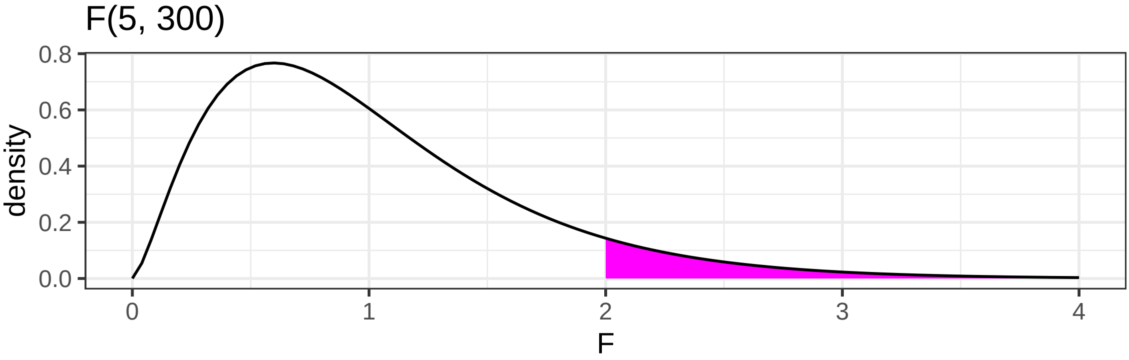
Under H_0 $(\mu_g = \mu)$,

- the quantity MSA$/$MSE has an F-distribution with $G-1$ numerator and $n-G$ denominator degrees of freedom,
- larger values of MSA$/$MSE indicate evidence against H_0, and
- the p-value is determined by $P(F_{G-1,n-G} > MSA/MSE)$.

F-distribution

F-distribution has two parameters:

- numerator degrees of freedom (ndf)
- denominator degrees of freedom (ddf)



One-way ANOVA F-test (by hand)

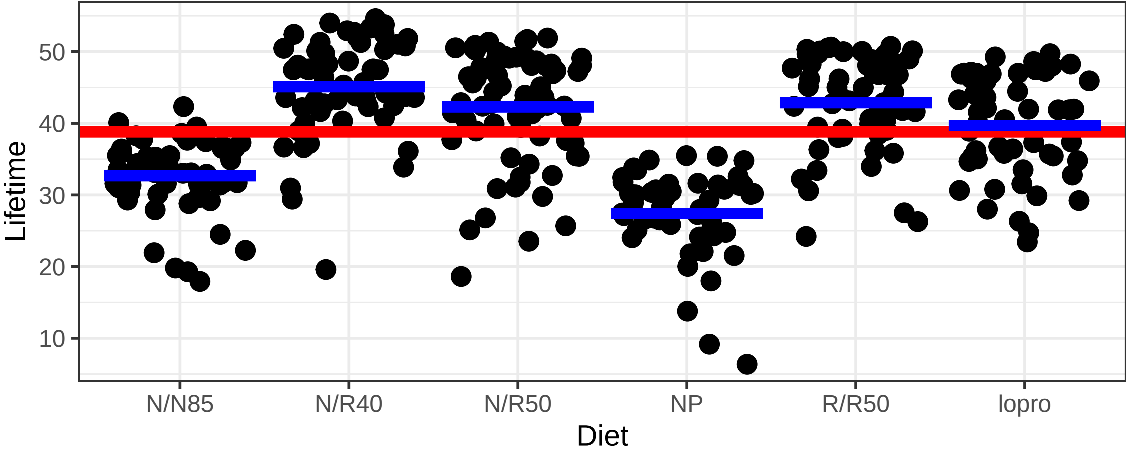
```
# A tibble: 7 x 4
  Diet        n  mean    sd
  <chr> <int> <dbl> <dbl>
1 N/N85    57  32.7  5.13
2 N/R40    60  45.1  6.70
3 N/R50    71  42.3  7.77
4 NP       49  27.4  6.13
5 R/R50    56  42.9  6.68
6 lopro    56  39.7  6.99
7 Total   349  38.8  8.97
```

So

$$\begin{aligned}
SSA = \quad & 57 \times (32.7 - 38.8)^2 + 60 \times (45.1 - 38.8)^2 + 71 \times (42.3 - 38.8)^2 + 49 \times (27.4 - 38.8)^2 \\
& +56 \times (42.9 - 38.8)^2 + 56 \times (39.7 - 38.8)^2 = 12734 \\
SST = \quad & (349 - 1) \times 8.97^2 = 28000 \\
SSE = \quad & SST - SSA = 28000 - 12734 = 15266 \\
G - 1 = \quad & 5 \\
n - G = \quad & 349 - 6 = 343 \\
n - 1 = \quad & 348 \\
MSA = \quad & SSA/G - 1 = 12734/5 = 2547 \\
MSE = \quad & SSE/n - G = 15266/343 = 44.5 = \hat{\sigma}^2 \\
F = \quad & MSA/MSE = 2547/44.5 = 57.2 \\
p = \quad & P(F_{5,343} > 57.2) < 0.0001
\end{aligned}$$

F statistic is off by 0.1 relative to the table later, because of rounding of 8.97. The real SST is 28031 which would be the F statistic of 57.1.

Graphical comparison



R code and output for one-way ANOVA

```
m <- lm(Lifetime ~ Diet, case0501)
anova(m)

Analysis of Variance Table

Response: Lifetime
           Df Sum Sq Mean Sq F value    Pr(>F)
Diet        5  12734  2546.8  57.104 < 2.2e-16 ***
Residuals 343  15297    44.6
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is evidence against the null model $Y_{ig} \overset{ind}{\sim} N(\mu, \sigma^2)$, i.e. our data seem incompatible with this model.

General F-tests

The one-way ANOVA F-test is an example of a general hypothesis testing framework that uses F-tests. This framework can be used to test

- composite alternative hypotheses or, equivalently,
- a full vs a reduced model.

The general idea is to balance the amount of variability remaining when moving from the reduced model to the full model measured using the sums of squared errors (SSEs) relative to the amount of complexity, i.e. parameters, added to the model.

Testing full vs reduced models

If $Y_{ig} \overset{ind}{\sim} N(\mu_g, \sigma^2)$ for $g = 1, \ldots, G$ and we want to test the hypotheses

- $H_0 : \mu_g = \mu$ for all g
- $H_1 : \mu_g \neq \mu_{g'}$ for some g and g'

think about this as two models:

- $H_0 : Y_{ig} \overset{ind}{\sim} N(\mu, \sigma^2)$ (reduced)
- $H_1 : Y_{ig} \overset{ind}{\sim} N(\mu_g, \sigma^2)$ (full)

We can use an F-test to calculate a p-value for tests of this type.

Nested models: full vs reduced

Two models are nested if the reduced model is a special case of the full model.

For example, consider the full model

$$Y_{ig} \overset{ind}{\sim} N(\mu_g, \sigma^2).$$

One special case of this model occurs when $\mu_g = \mu$ and thus

$$Y_{ig} \overset{ind}{\sim} N(\mu, \sigma^2).$$

is a reduced model and these two models are nested.

Calculating the sum of squared residuals (errors)

Model	Full	Reduced
Assumption	$H_1 : Y_{ig} \overset{ind}{\sim} N\left(\mu_g, \sigma^2\right)$	$H_0 : Y_{ig} \overset{iid}{\sim} N(\mu, \sigma^2)$
Mean	$\hat{\mu}_g = \overline{Y}_g = \frac{1}{n_g} \sum_{i=1}^{n_g} Y_{ig}$	$\hat{\mu} = \overline{Y} = \frac{1}{n} \sum_{g=1}^{G} \sum_{i=1}^{n_g} Y_{ig}$
Residual	$r_{ig} = Y_{ig} - \hat{\mu}_g = Y_{ig} - \overline{Y}_g$	$r_{ig} = Y_{ig} - \hat{\mu} = Y_{ig} - \overline{Y}$
SSE	$\sum_{g=1}^{G} \sum_{i=1}^{n_g} r_{ig}^2$	$\sum_{g=1}^{G} \sum_{i=1}^{n_g} r_{ig}^2$

General F-tests

Do the following

1. Calculate

   Extra sum of squares =
   Residual sum of squares (reduced) - Residual sum of squares (full)

2. Calculate

   Extra degrees of freedom =
   # of mean parameters (full) - # of mean parameters (reduced)

3. Calculate F-statistics

$$\text{F} = \frac{\text{Extra sum of squares / Extra degrees of freedom}}{\text{Estimated residual variance in full model } (\hat{\sigma}^2)}$$

4. A pvalue is $P(F_{\text{ndf,ddf}} > \text{F})$
   - numerator degrees of freedom (ndf) = Extra degrees of freedom
   - denominator degrees of freedom (ddf): df associated with $\hat{\sigma}^2$

Mice lifetimes

Consider the hypothesis that mice on all diets have a common mean lifetime except NP.

Let

$$Y_{ig} \overset{ind}{\sim} N(\mu_g, \sigma^2)$$

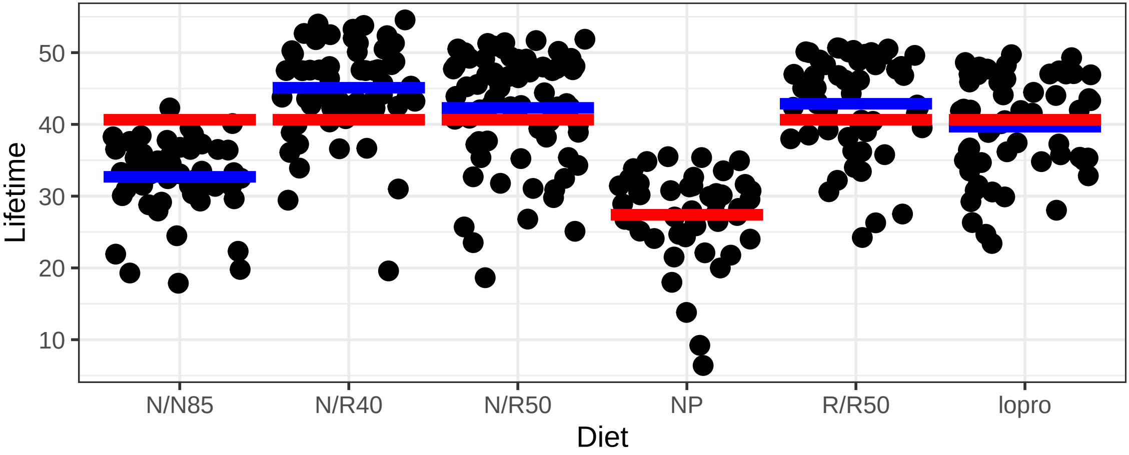
with $g = 1$ being the NP group then the hypotheses are

- $H_0 : \mu_g = \mu$ for $g \neq 1$
- $H_1 : \mu_g \neq \mu_{g'}$ for some $g, g' = 2, \ldots, 6$

As models:

- $H_0 : Y_{i1} \overset{iid}{\sim} N(\mu_1, \sigma^2)$ and $Y_{ig} \overset{iid}{\sim} N(\mu, \sigma^2)$ for $g \neq 1$
- $H_1 : Y_{ig} \overset{ind}{\sim} N(\mu_g, \sigma^2)$

As a picture



Making R do the calculations

```
case0501$NP = factor(case0501$Diet == "NP")

modR = lm(Lifetime ~ NP,   case0501) # (R)educed model
modF = lm(Lifetime ~ Diet, case0501) # (F)ull    model
anova(modR,modF)

Analysis of Variance Table

Model 1: Lifetime ~ NP
Model 2: Lifetime ~ Diet
  Res.Df   RSS Df Sum of Sq     F    Pr(>F)
1    347 20630
2    343 15297  4    5332.2 29.89 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Lack-of-fit F-test for linearity

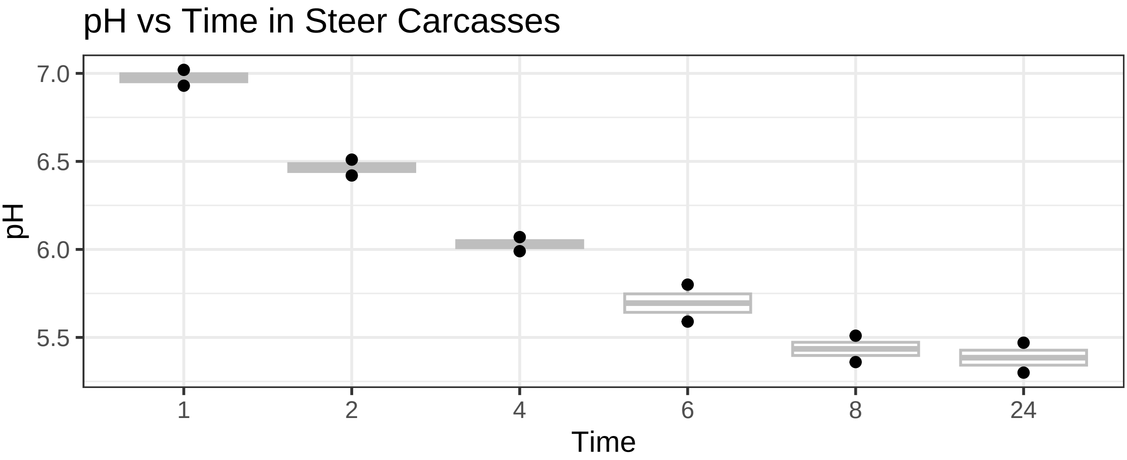
Let Y_{ig} be the i^{th} observation from the g^{th} group where the group is defined by those observations having the same explanatory variable value (X_g).

Two models:

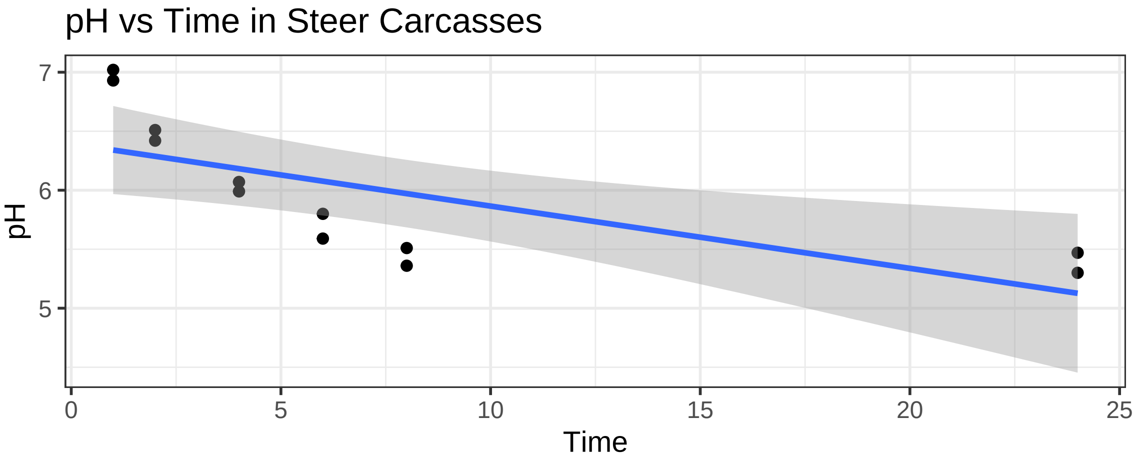
ANOVA:	$Y_{ig} \overset{ind}{\sim} N(\mu_g, \sigma^2)$	(full)
Regression:	$Y_{ig} \overset{ind}{\sim} N(\beta_0 + \beta_1 X_g, \sigma^2)$	(reduced)

- Regression model is reduced:
  - ANOVA has G parameters for the mean
  - Regression has 2 parameters for the mean
  - Set $\mu_g = \beta_0 + \beta_1 X_g$.
- Small pvalues indicate a lack-of-fit, i.e. the regression (reduced) model is not adequate.
- Lack-of-fit F-test requires multiple observations at a few X_g values!

pH vs Time - ANOVA



pH vs Time - Regression



Lack-of-fit F-test in R

```
# Use as.factor to turn a continuous variable into a categorical variable
m_anova = lm(pH ~ as.factor(Time), Sleuth3::ex0816)
m_reg   = lm(pH ~           Time , Sleuth3::ex0816)
anova(m_reg, m_anova)

Analysis of Variance Table

Model 1: pH ~ Time
Model 2: pH ~ as.factor(Time)
  Res.Df     RSS Df Sum of Sq      F    Pr(>F)
1     10 1.97289
2      6 0.05905  4    1.9138 48.616 0.0001048 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

There is evidence the data are incompatible with the null hypothesis that states the means of each group fall along a line.

Summary

- Use F-tests for comparison of full vs reduced model
    - One-way ANOVA F-test
    - General F-tests
    - Lack-of-fit F-tests

Think about F-tests as comparing models.

R06a - Interpreting Regression p-values as Posterior Probabilities

STAT 5870 (Engineering)
Iowa State University

November 22, 2024

Regression p-values

Recall the regression model

$$Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2), \qquad \mu_i = \beta_0 + \beta_1 X_{i,1} + \cdots + \beta_p X_{i,p}$$

A common hypothesis test is

$$H_0 : \beta_j = 0 \qquad \text{versus} \qquad H_A : \beta_j \neq 0$$

which has

$$p\text{-value} = 2P\left(T > |t|\right)$$

where $T \sim t_{n-(p+1)}$ and $t = \hat{\beta}_j / SE(\beta_j)$.

Example Regression Output

```
Call:
lm(formula = Speed ~ Conditions * log(NetToWinner), data = Sleuth3::ex0920)

Residuals:
     Min       1Q   Median       3Q      Max 
-1.50551 -0.32127 -0.00219  0.35201  1.13026 

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)    
(Intercept)                     33.23367    0.34584  96.095  < 2e-16 ***
ConditionsSlow                  -2.04517    0.72404  -2.825   0.0056 ** 
log(NetToWinner)                 0.27830    0.02942   9.458 5.88e-16 ***
ConditionsSlow:log(NetToWinner)  0.08664    0.06583   1.316   0.1908    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4978 on 112 degrees of freedom
Multiple R-squared:  0.7015,Adjusted R-squared:  0.6935 
F-statistic: 87.75 on 3 and 112 DF,  p-value: < 2.2e-16
```

Bayesian Posterior Probabilities

With prior $p(\beta, \sigma^2) \propto 1/\sigma^2$, we have

$$\beta_j | y \sim t_{n-(p+1)} \left(\hat{\beta}_j, SE(\beta_j)^2 \right).$$

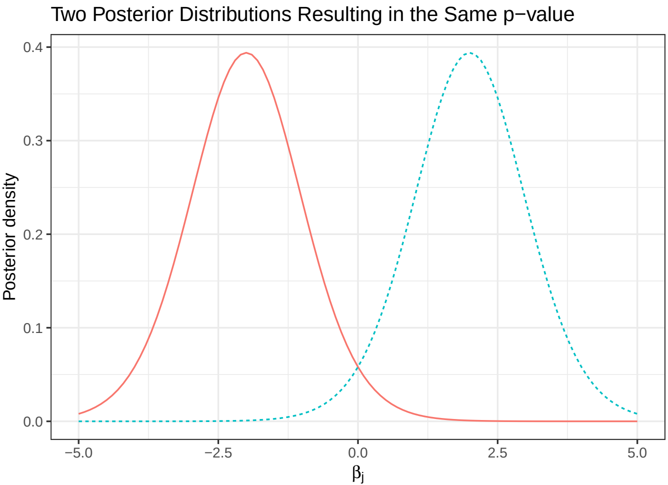
Thus

$$P\left(\beta_j > 0 | \, y\right) = P\left(\frac{\beta_j - \hat{\beta}_j}{SE(\beta_j)} > \frac{0 - \hat{\beta}_j}{SE(\beta_j)} \middle| y \right) = P\left(T > -t\right)$$

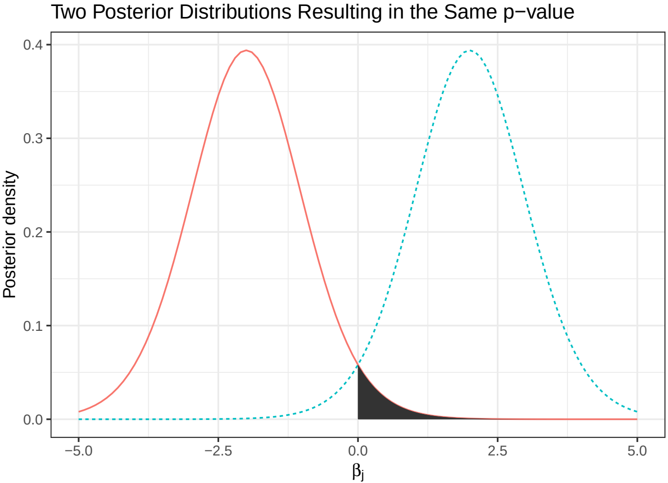
which is very close to

$$p\text{-value} = 2P\left(T > |t|\right).$$

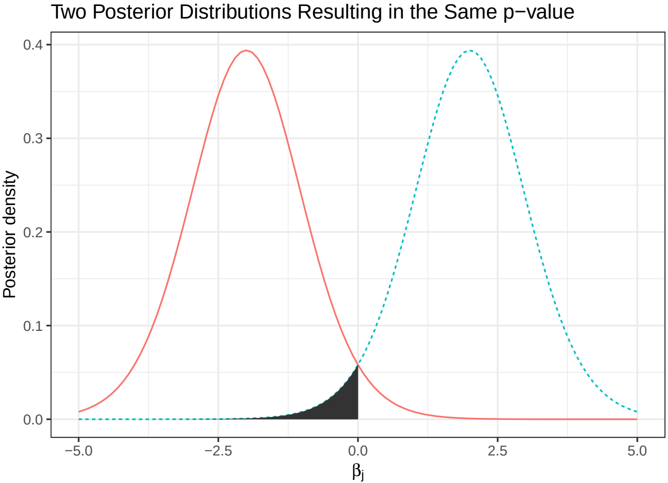
Visualizing Posterior Distribution



Visualizing Posterior Distribution



Visualizing Posterior Distribution



Interpreting Regression p-values as Posterior Probabilities

Suppose we have a p-value for $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$. Then

- If $\hat{\beta}_j < 0$, then
$$P(\beta_j > 0|y) = p\text{-value}/2.$$
- If $\hat{\beta}_j > 0$, then
$$P(\beta_j < 0|y) = p\text{-value}/2.$$

Alternatively,

- If $\hat{\beta}_j < 0$, then
$$P(\beta_j < 0|y) = 1 - p\text{-value}/2.$$
- If $\hat{\beta}_j > 0$, then
$$P(\beta_j > 0|y) = 1 - p\text{-value}/2.$$

Example Interpretation

```
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                      33.23       0.35   96.09     0.00
ConditionsSlow                   -2.05       0.72   -2.82     0.01
log(NetToWinner)                  0.28       0.03    9.46     0.00
ConditionsSlow:log(NetToWinner)   0.09       0.07    1.32     0.19
```

Intercept	$P(\beta_0 > 0 \vert y) \approx 1$
ConditionsSlow	$P(\beta_1 < 0 \vert y) \approx 0.99$
log(NetToWinner)	$P(\beta_2 > 0 \vert y) \approx 1$
ConditionsSlow:log(NetToWinner)	$P(\beta_3 > 0 \vert y) \approx 0.90$

Summary

Suppose we have a regression p-value for $H_0 : \beta_j = 0$ vs $H_A : \beta_j \neq 0$. Then

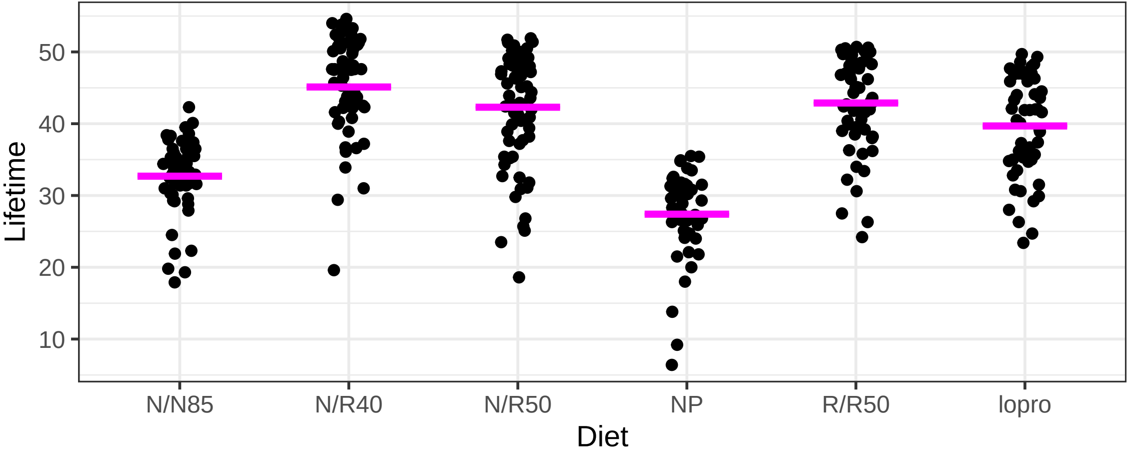
- If $\hat{\beta}_j < 0$, then
$$P(\beta_j < 0|y) = 1 - p\text{-value}/2.$$
- If $\hat{\beta}_j > 0$, then
$$P(\beta_j > 0|y) = 1 - p\text{-value}/2.$$

R07 - Contrasts

STAT 5870 (Engineering)
Iowa State University

November 22, 2024

Diet Effect on Mice Lifetimes



ANOVA and Regression Models

ANOVA model:

$$Y_{ij} \overset{ind}{\sim} N(\mu_j, \sigma^2)$$

with Y_{ij} being the lifetime for the ith mouse on the jth diet for $j = 0, 1, 2, 3, 4, 5$.

Regression model:

$$Y_i \overset{ind}{\sim} N(\beta_0 + \beta_1 X_{i,1} + \ldots + \beta_p X_{i,p}, \sigma^2)$$

where Y_i is the lifetime for the ith mouse and $X_{i,j}$ is an indicator for the ith mouse being on the jth diet.

Reparameterized model since

$$\mu_0 = \beta_0 \quad \text{and} \quad \mu_j = \beta_0 + \beta_j$$

for $j > 0$.

Scientific questions

Here are a few example scientific questions:

1. What is the effect of pre-wean calorie restriction on mean lifetimes?
2. What is the difference in mean lifetimes between mice on a 40 kcal diet compared to those on a 50 kcal diet?
3. What is the effect of high calorie vs low calorie diets on mean lifetimes?

We can compute contrasts:

$$\begin{aligned}
\gamma_1 = & \quad \mu_{R/R50} - \mu_{N/R50} \\
\gamma_2 = & \quad \mu_{N/R40} - \tfrac{1}{2}(\mu_{N/R50} + \mu_{R/R50}) \\
\gamma_3 = & \quad \tfrac{1}{4}(\mu_{N/R50} + \mu_{R/R50} + \mu_{N/R40} + \mu_{lopro}) \\
& \quad -\tfrac{1}{2}(\mu_{NP} + \mu_{N/N85})
\end{aligned}$$

Contrasts

A linear combination of group means has the form

$$\gamma = C_1\mu_1 + C_2\mu_2 + \ldots + C_J\mu_J$$

where C_j are known coefficients and μ_j are the unknown population means.

A linear combination with $C_1 + C_2 + \cdots + C_J = 0$ is a contrast.

Contrast interpretation is usually best if $|C_1| + |C_2| + \cdots + |C_J| = 2$, i.e. the positive coefficients sum to 1 and the negative coefficients sum to -1.

Inference on Contrasts

Contrast

$$\gamma = C_1\mu_1 + C_2\mu_2 + \cdots + C_J\mu_J \quad \text{with} \quad \hat{\gamma} = C_1\overline{Y}_1 + C_2\overline{Y}_2 + \cdots + C_J\overline{Y}_J$$

with standard error

$$SE(\hat{\gamma}) = \hat{\sigma}\sqrt{\frac{C_1^2}{n_1} + \frac{C_2^2}{n_2} + \cdots + \frac{C_J^2}{n_J}}.$$

p-values for $H_0 : \gamma = g_0$ vs $H_A : \gamma \neq g_0$ and posterior probabilities (i.e. $2P(\gamma > 0|y)$ or $2P(\gamma < 0|y)$):

$$t = \frac{g - g_0}{SE(g)}, \quad p = 2P(T_{n-J} < -|t|).$$

Two-sided equal-tail $100(1-\alpha)\%$ confidence/credible intervals:

$$g \pm t_{n-J,1-\alpha/2}SE(g).$$

Contrasts for mice lifetime dataset

For these contrasts:

1. Difference in mean lifetimes for N/R50 v R/R50 diet
2. Difference in mean lifetimes for N/R40 v N/R50 and R/R50 combined
3. Difference in mean lifetimes for high calorie (NP and N/N85) diets v low calorie (others) diets

$H_0 : \gamma = 0 \qquad H_A : \gamma \neq 0 :$

$$\begin{aligned}
\gamma_1 = &\quad \mu_{R/R50} - \mu_{N/R50} \\
\gamma_2 = &\quad \mu_{N/R40} - \tfrac{1}{2}(\mu_{N/R50} + \mu_{R/R50}) \\
\gamma_3 = &\quad \tfrac{1}{4}(\mu_{N/R50} + \mu_{R/R50} + \mu_{N/R40} + \mu_{lopro}) \\
&\quad -\tfrac{1}{2}(\mu_{NP} + \mu_{N/N85})
\end{aligned}$$

	N/N85	N/R40	N/R50	NP	R/R50	lopro
early rest - none @ 50kcal	0.00	0.00	-1.00	0.00	1.00	0.00
40kcal/week - 50kcal/week	0.00	1.00	-0.50	0.00	-0.50	0.00
lo cal - hi cal	-0.50	0.25	0.25	-0.50	0.25	0.25

Fit the Multiple Regression Model

```
m <- lm(Lifetime ~ Diet, data = Sleuth3::case0501)
summary(m)


Call:
lm(formula = Lifetime ~ Diet, data = Sleuth3::case0501)

Residuals:
     Min       1Q   Median       3Q      Max 
-25.5167  -3.3857   0.8143   5.1833  10.0143 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  32.6912     0.8846  36.958  < 2e-16 ***
DietN/R40    12.4254     1.2352  10.059  < 2e-16 ***
DietN/R50     9.6060     1.1877   8.088 1.06e-14 ***
DietNP       -5.2892     1.3010  -4.065 5.95e-05 ***
DietR/R50    10.1945     1.2565   8.113 8.88e-15 ***
Dietlopro     6.9945     1.2565   5.567 5.25e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.678 on 343 degrees of freedom
Multiple R-squared:  0.4543,Adjusted R-squared:  0.4463 
F-statistic:  57.1 on 5 and 343 DF,  p-value: < 2.2e-16
```

Estimate Group Means

```
library("emmeans")
em <- emmeans(m, ~ Diet)
em

 Diet  emmean    SE  df lower.CL upper.CL
 N/N85   32.7 0.885 343     31.0     34.4
 N/R40   45.1 0.862 343     43.4     46.8
 N/R50   42.3 0.793 343     40.7     43.9
 NP      27.4 0.954 343     25.5     29.3
 R/R50   42.9 0.892 343     41.1     44.6
 lopro   39.7 0.892 343     37.9     41.4

Confidence level used: 0.95
```

```
K_list

$`early rest - none @ 50kcal`
[1]  0  0 -1  0  1  0

$`40kcal/week - 50kcal/week`
[1]  0.0  1.0 -0.5  0.0 -0.5  0.0

$`lo cal - hi cal`
[1] -0.50  0.25  0.25 -0.50  0.25  0.25
```

```
co <- contrast(em, K_list)

# p-values (and posterior tail probabilities)
co

 contrast                   estimate   SE  df t.ratio p.value
 early rest - none @ 50kcal    0.589 1.19 343   0.493  0.6223
 40kcal/week - 50kcal/week     2.525 1.05 343   2.408  0.0166
 lo cal - hi cal              12.450 0.78 343  15.961  <.0001

# confidence/credible intervals
confint(co)

 contrast                   estimate   SE  df lower.CL upper.CL
 early rest - none @ 50kcal    0.589 1.19 343   -1.759     2.94
 40kcal/week - 50kcal/week     2.525 1.05 343    0.463     4.59
 lo cal - hi cal              12.450 0.78 343   10.915    13.98

Confidence level used: 0.95
```

Summary

- Contrasts are linear combinations of means where the coefficients sum to zero
- t-test tools are used to calculate pvalues and confidence intervals

Sulfur effect on scab disease in potatoes

> *The experiment was conducted to investigate the effect of sulfur on controlling scab disease in potatoes. There were seven treatments: control, plus spring and fall application of 300, 600, 1200 lbs/acre of sulfur. The response variable was percentage of the potato surface area covered with scab averaged over 100 random selected potatoes. A completely randomized design was used with 8 replications of the control and 4 replications of the other treatments.*

Cochran and Cox. (1957) Experimental Design (2nd ed). pg96 and Agron. J. 80:712-718 (1988)

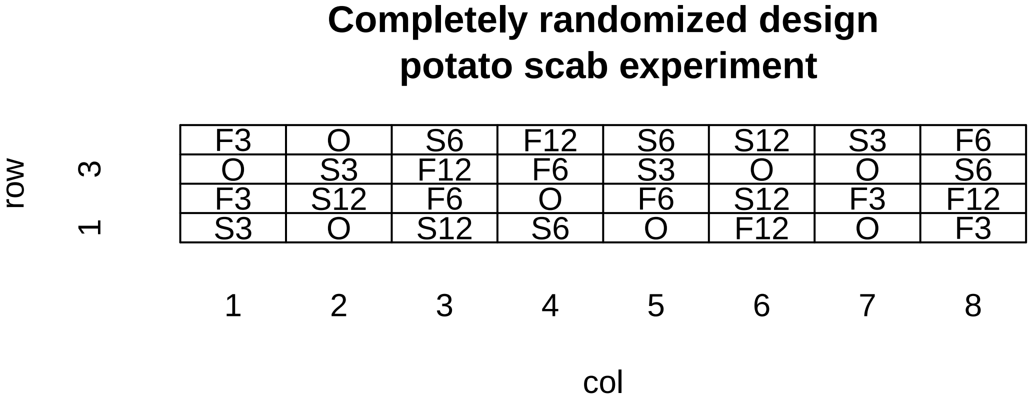
Scientific questions:

- Does sulfur have any impact at all?
- What is the difference between spring and fall application of sulfur?
- What is the effect of increased sulfur application?

Data

```
   inf trt row col sulfur    application treatment
1    9  F3   4   1    300           fall        F3
2   12   0   4   2      0 not applicable         0
3   18  S6   4   3    600         spring        S6
4   10 F12   4   4   1200           fall       F12
5   24  S6   4   5    600         spring        S6
6   17 S12   4   6   1200         spring       S12
7   30  S3   4   7    300         spring        S3
8   16  F6   4   8    600           fall        F6
9   10   0   3   1      0 not applicable         0
10   7  S3   3   2    300         spring        S3
11   4 F12   3   3   1200           fall       F12
12  10  F6   3   4    600           fall        F6
13  21  S3   3   5    300         spring        S3
14  24   0   3   6      0 not applicable         0
15  29   0   3   7      0 not applicable         0
16  12  S6   3   8    600         spring        S6
17   9  F3   2   1    300           fall        F3
18   7 S12   2   2   1200         spring       S12
19  18  F6   2   3    600           fall        F6
20  30   0   2   4      0 not applicable         0
21  18  F6   2   5    600           fall        F6
22  16 S12   2   6   1200         spring       S12
23  16  F3   2   7    300           fall        F3
24   4 F12   2   8   1200           fall       F12
25   9  S3   1   1    300         spring        S3
26  18   0   1   2      0 not applicable         0
27  17 S12   1   3   1200         spring       S12
```

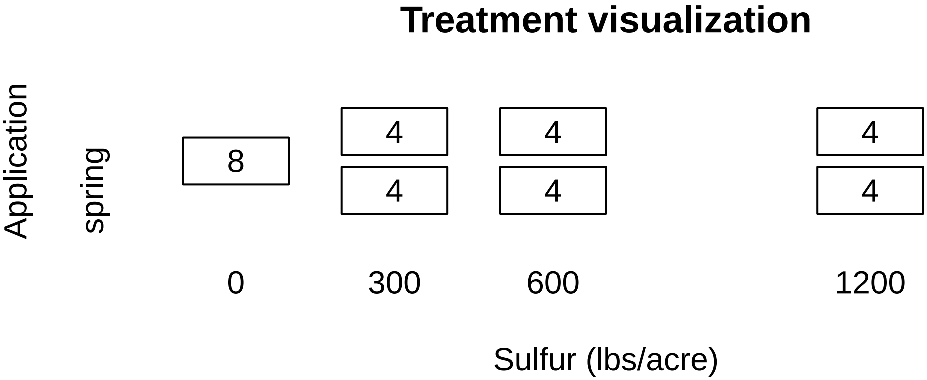
Design



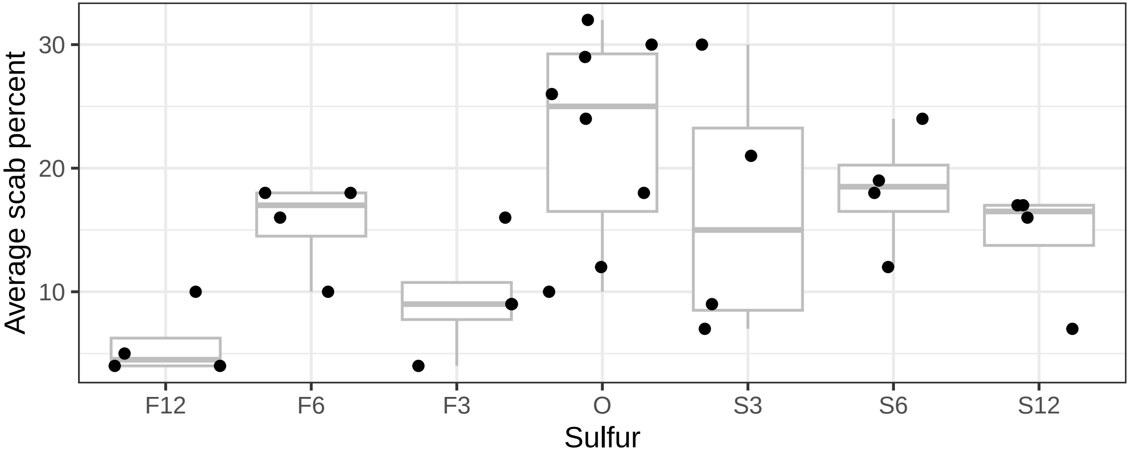
Completely randomized design
potato scab experiment

	1	2	3	4	5	6	7	8
	F3	O	S6	F12	S6	S12	S3	F6
3	O	S3	F12	F6	S3	O	O	S6
	F3	S12	F6	O	F6	S12	F3	F12
1	S3	O	S12	S6	O	F12	O	F3

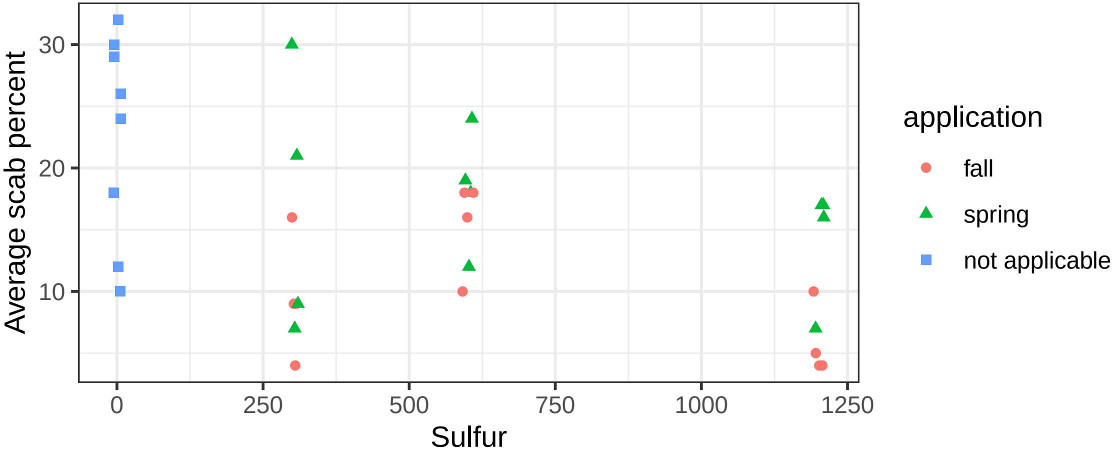
Design


Treatment visualization
Application
spring
8
4
4
4
4
4
4
0
300
600
1200
Sulfur (lbs/acre)

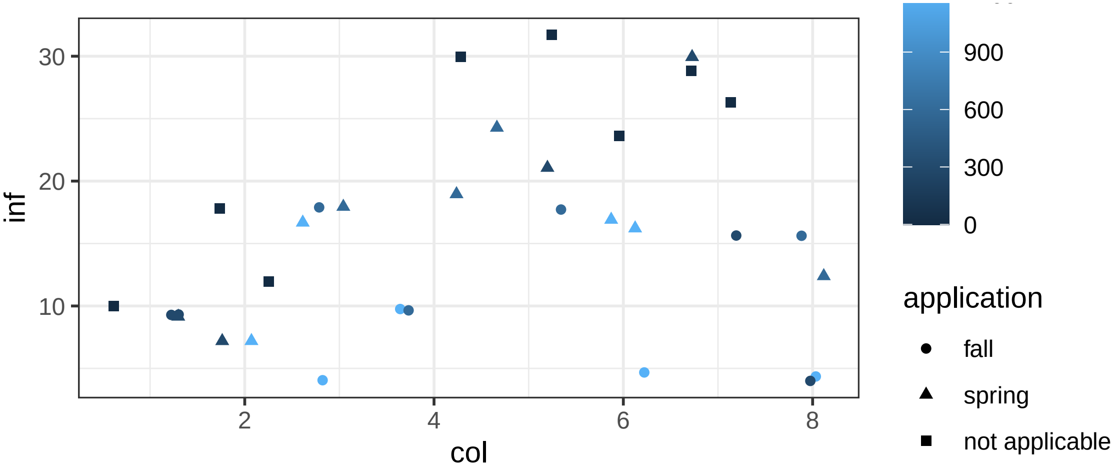
Data



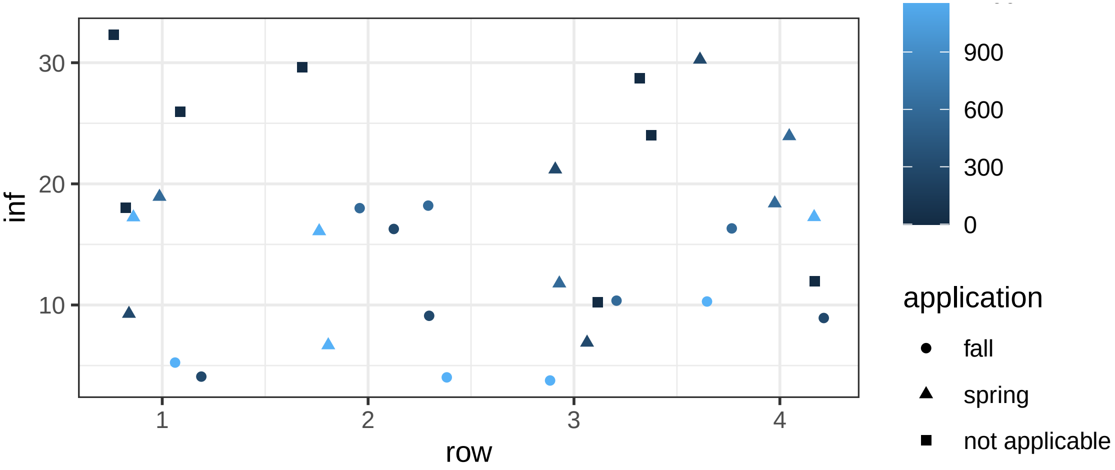
Data


Average scab percent
30
20
10
0
250
500
750
1000
1250
Sulfur
application
● fall
▲ spring
■ not applicable

Data


30
20
10
inf
2
4
6
8
col
900
600
300
0
application
● fall
▲ spring
■ not applicable

Data


30
20
10
inf
1
2
3
4
row
900
600
300
0
application
fall
spring
not applicable

Model

Y_{ij}: avg % of surface area covered with scab for plot i in treatment j for $j = 1, \ldots, 7$.

Assume $Y_{ij} \overset{ind}{\sim} N(\mu_j, \sigma^2)$.

Hypotheses:

- Difference amongst any means:
  One-way ANOVA F-test
- *Any effect*:
  Contrast: control vs sulfur
- *Fall vs spring*:
  Contrast: fall vs spring applications
- *Sulfur level*:
  Contrast: linear trend

Contrasts

- *Sulfur effect*: Any sulfur vs none

$$\begin{aligned}\gamma &= \tfrac{1}{6}(\mu_{F12} + \mu_{F6} + \mu_{F3} + \mu_{S3} + \mu_{S6} + \mu_{S12}) - \mu_O \\ &= \tfrac{1}{6}(\mu_{F12} + \mu_{F6} + \mu_{F3} + \mu_{S3} + \mu_{S6} + \mu_{S12} - 6\mu_O)\end{aligned}$$

- *Fall vs spring*: Contrast comparing fall vs spring applications

$$\begin{aligned}\gamma &= \tfrac{1}{3}(\mu_{F12} + \mu_{F6} + \mu_{F3}) + 0\mu_O - \tfrac{1}{3}(\mu_{S3} + \mu_{S6} + \mu_{S12}) \\ &= \tfrac{1}{3}\left[1\mu_{F12} + 1\mu_{F6} + 1\mu_{F3} + 0\mu_O - 1\mu_{S3} - 1\mu_{S6} - 1\mu_{S12}\right]\end{aligned}$$

Contrasts (cont.)

- Sulfur linear trend
  - The group sulfur levels (X_j) are 12, 6, 3, 0, 3, 6, and 12 (100 lbs/acre)
  - and a linear trend contrast is $X_j - \overline{X}$

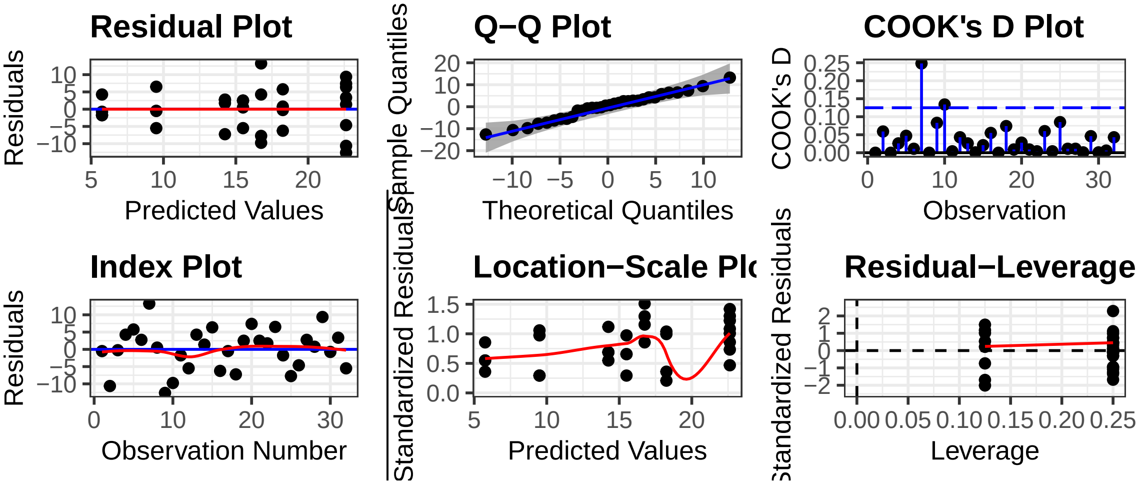
X_i	12	6	3	0	3	6	12
$X_i - \overline{X}$	6	0	-3	-6	-3	0	6

$$\gamma \quad = 6\mu_{F12} + 0\mu_{F6} - 3\mu_{F3} - 6\mu_O - 3\mu_{S3} + 0\mu_{S6} + 6\mu_{S12}$$

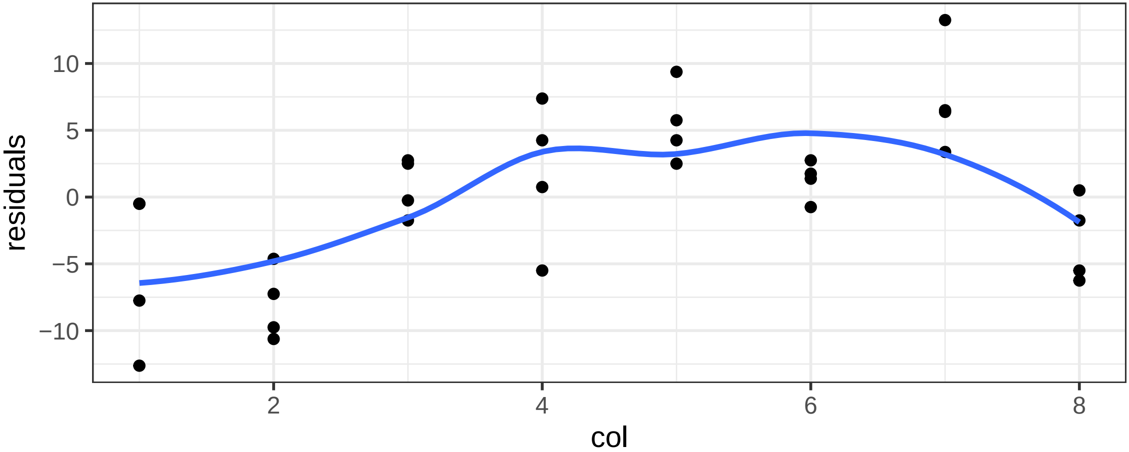
Trt	F12	F6	F3	O	S3	S6	S12	Div
Sulfur v control	1	1	1	-6	1	1	1	6
Fall v Spring	1	1	1	0	-1	-1	-1	3
Linear Trend	6	0	-3	-6	-3	0	6	1

```r
K <-
#                                      F12 F6 F3  0 S3 S6 S12
              list("sulfur - control" = c( 1, 1, 1,-6, 1, 1,  1)/6,
                   "fall - spring"    = c( 1, 1, 1, 0,-1,-1, -1)/3,
                   "linear trend"     = c( 6, 0,-3,-6,-3, 0,  6)/1)

m <- lm(inf ~ treatment, data = d)
anova(m)
```

```
Analysis of Variance Table

Response: inf
          Df  Sum Sq Mean Sq F value  Pr(>F)
treatment  6  972.34 162.057  3.6081 0.01026 *
Residuals 25 1122.88  44.915
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residual Plot
Residuals
10
5
0
−5
−10
5
10
15
20
Predicted Values
Q−Q Plot
Sample Quantiles
20
10
0
−10
−20
−10
−5
0
5
10
Theoretical Quantiles
COOK's D Plot
COOK's D
0.25
0.20
0.15
0.10
0.05
0.00
0
10
20
30
Observation
Index Plot
Residuals
10
5
0
−5
−10
0
10
20
30
Observation Number
Location−Scale Plo
Standardized Residuals
1.5
1.0
0.5
0.0
5
10
15
20
Predicted Values
Residual−Leverage
Standardized Residuals
2
1
0
−1
−2
0.00
0.05
0.10
0.15
0.20
0.25
Leverage

```
em <- emmeans(m, ~treatment); em

 treatment emmean   SE df lower.CL upper.CL
 F12         5.75 3.35 25    -1.15     12.7
 F6         15.50 3.35 25     8.60     22.4
 F3          9.50 3.35 25     2.60     16.4
 O          22.62 2.37 25    17.74     27.5
 S3         16.75 3.35 25     9.85     23.7
 S6         18.25 3.35 25    11.35     25.2
 S12        14.25 3.35 25     7.35     21.2

Confidence level used: 0.95

co <- contrast(em, K)
confint(co)

 contrast         estimate    SE df lower.CL upper.CL
 sulfur - control    -9.29  2.74 25    -14.9   -3.657
 fall - spring       -6.17  2.74 25    -11.8   -0.532
 linear trend       -94.50 34.82 25   -166.2  -22.779

Confidence level used: 0.95
```


10
5
0
−5
−10
residuals
2
4
6
8
col

Summary

For this particular data analysis

- Significant differences in means between the groups (ANOVA $F_{6,25} = 3.61$ p=0.01)
- Having sulfur was associated with a reduction in scab % of 9 (4,15) compared to no sulfur
- Fall application reduced scab % by 6 (0.5,12) compared to spring application
- Linear trend in sulfur was significant (p=0.01)

- Concerned about spatial correlation among columns
- Consider a logarithm of the response
  - CI for F12 (-1.2, 12.7)
  - Non-constant variance (residuals vs predicted, sulfur, application)

R08 - Experimental design

STAT 5870 (Engineering)
Iowa State University

November 22, 2024

Random samples and random treatment assignment

Recall that the objective of data analysis is often to make an inference about a population based on a sample. For the inference to be statistically valid, we need a random sample from the population.

In order to make a causal statment, the levels of the explanatory variables need to be randomly assigned to the experimental units.

- random assignment → randomized experiment
- non-random assignment → observational study

Data collection

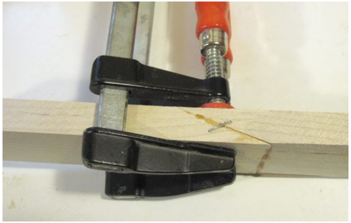
	Treatment randomly assigned?	
	No	Yes
Sample	Observational study	Randomized experiment
Not random	No inference to population No cause-and-effect	No inference to population Yes cause-and-effect
Random	Yes inference to population No cause-and-effect	Yes inference to population Yes cause-and-effect

Strength of wood glue

You are interested in testing two different wood glues:

- Gorilla Wood Glue
- Titebond 1413 Wood Glue

On a scarf joint:



So you collect up some wood, glue the pieces together, and determine the weight required to break the joint. (Lots of details are missing.)

Inspiration: `https://woodgears.ca/joint_strength/glue.html`

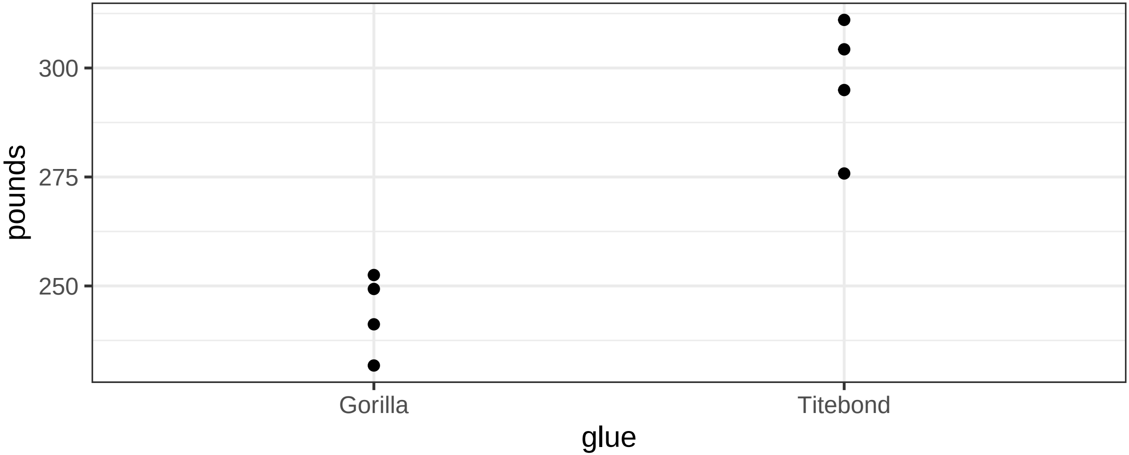
Completely Randomized Design (CRD)

Suppose I have 8 pieces of wood laying around. I cut each piece and randomly use either Gorilla or Titebond glue to recombine the pieces. I do the randomization in such a way that I have exactly 4 Gorilla and 4 Titebond results, e.g.

```
# A tibble: 8 x 2
  woodID glue
  <chr>  <chr>
1 wood1  Gorilla
2 wood2  Titebond
3 wood3  Gorilla
4 wood4  Titebond
5 wood5  Titebond
6 wood6  Gorilla
7 wood7  Titebond
8 wood8  Gorilla
```

This is called a completely randomized design (CRD). Because all treatment (combinations) have the same number of replicates, the design is balanced. Because all treatment (combinations) are repeated, the design is replicated.

Visualize the data



Model

Let

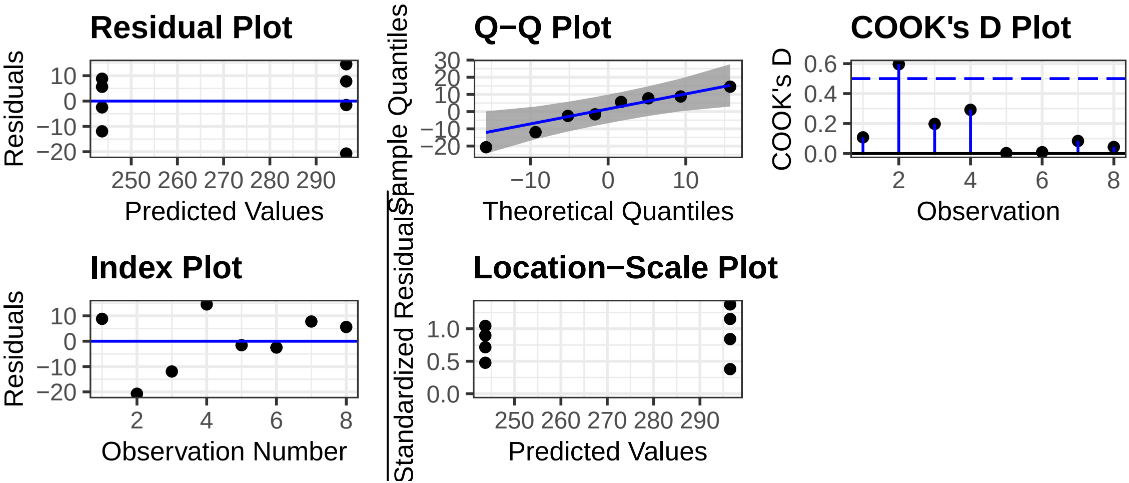
- P_w be the weight (pounds) needed to break wood w,
- T_w be an indicator that the Titebond glue was used on wood w, i.e.

$$T_w = \mathrm{I}(\text{glue}_w = \text{Titebond}).$$

Then a regression model for these data is

$$P_w \overset{ind}{\sim} N(\beta_0 + \beta_1 T_w, \sigma^2).$$

Check model assumptions



Obtain statistics

```
coefficients(m)
```

```
 (Intercept) glueTitebond 
    243.6971      52.8206 
```

```
summary(m)$r.squared
```

```
[1] 0.8531122
```

```
confint(m)
```

```
                 2.5 %    97.5 %
(Intercept)  228.21529 259.17885
glueTitebond  30.92606  74.71514
```

```
emmeans(m, ~glue)
```

```
 glue     emmean   SE df lower.CL upper.CL
 Gorilla     244 6.33  6      228      259
 Titebond    297 6.33  6      281      312

Confidence level used: 0.95 
```

Interpret results

A randomized experiment was designed to evaluate the effectiveness of Gorilla and Titebond in preventing failures in scarf joints cut at a 20 degree angle through 1" × 2" spruce with 4 replicates for each glue type. The mean break weight (lbs) was 244 with a 95% CI of (228,259) for Gorilla and 297 (281,312) for Titebond. Titebond glue caused an increase in break weight of 53 (31,75) lbs compared to Gorilla Glue. This difference accounted for 85 % of the variability in break weight.

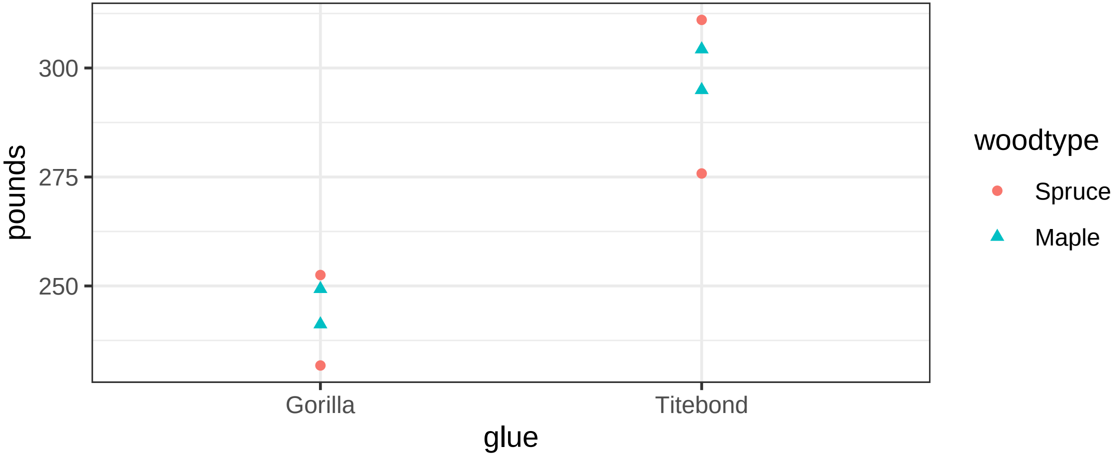
Randomized complete block design (RCBD)

Suppose the wood actually came from two different types: Maple and Spruce. And perhaps you have reason to believe the glue will work differently depending on the type of wood. In this case, you would want to block by wood type and perform the randomization within each block, i.e.

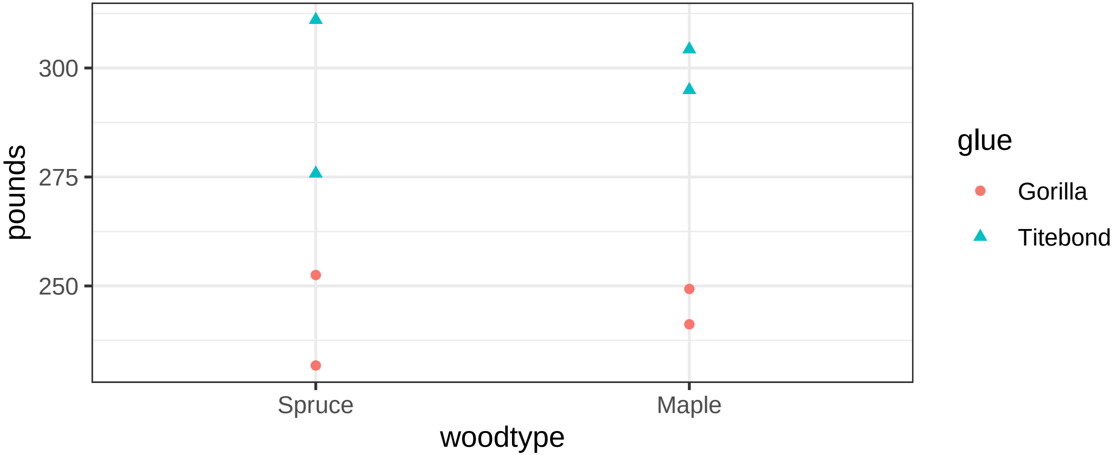
```
# A tibble: 8 x 3
  woodID woodtype glue
  <chr>  <fct>    <chr>
1 wood1  Spruce   Gorilla
2 wood2  Spruce   Titebond
3 wood3  Spruce   Gorilla
4 wood4  Spruce   Titebond
5 wood5  Maple    Titebond
6 wood6  Maple    Gorilla
7 wood7  Maple    Titebond
8 wood8  Maple    Gorilla
```

This is called a randomized complete block design (RCBD). If all treatment combinations exist, then the design is complete. If a treatment combination is missing, then the design is incomplete.

Visualize the data



Visualize the data - a more direct comparison



Main effects model

Let

- P_w be the weight (pounds) needed to break wood w
- T_w be an indicator that Titebond glue was used on wood w, and
- M_w be an indicator that wood w was Maple.

Then a main effects model for these data is

$$P_w \overset{ind}{\sim} N(\beta_0 + \beta_1 T_w + \beta_2 M_w, \sigma^2)$$

Perform analysis

```
Call:
lm(formula = pounds ~ glue + woodtype, data = d)

Residuals:
      1       2       3       4       5       6       7       8 
 11.146 -18.384  -9.611  16.849  -3.902  -4.822   5.437   3.286 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)    241.366      8.294  29.100 8.98e-07 ***
glueTitebond    52.821      9.578   5.515  0.00268 ** 
woodtypeMaple    4.662      9.578   0.487  0.64702    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 13.54 on 5 degrees of freedom
Multiple R-squared:  0.8598,Adjusted R-squared:  0.8037 
F-statistic: 15.33 on 2 and 5 DF,  p-value: 0.007365
                  2.5 %    97.5 %
(Intercept)   220.04467 262.68760
glueTitebond   28.20070  77.44051
woodtypeMaple -19.95804  29.28177
```

Replication

Since there are more than one observation for each woodtype-glue combination, the design is replicated:

```
d |> group_by(woodtype, glue) |> summarize(n = n())

# A tibble: 4 x 3
# Groups:   woodtype [2]
  woodtype glue         n
  <fct>    <chr>    <int>
1 Spruce   Gorilla      2
2 Spruce   Titebond     2
3 Maple    Gorilla      2
4 Maple    Titebond     2
```

When the design is replicated, we can consider assessing an interaction.

Interaction model

Let

- P_w be the weight (pounds) needed to break wood w
- T_w be an indicator that Titebond glue was used on wood w, and
- M_w be an indicator that wood w was Maple.

Then a model with the interaction for these data is

$$P_w \overset{ind}{\sim} N(\beta_0 + \beta_1 T_w + \beta_2 M_w + \beta_3 T_w M_w, \sigma^2)$$

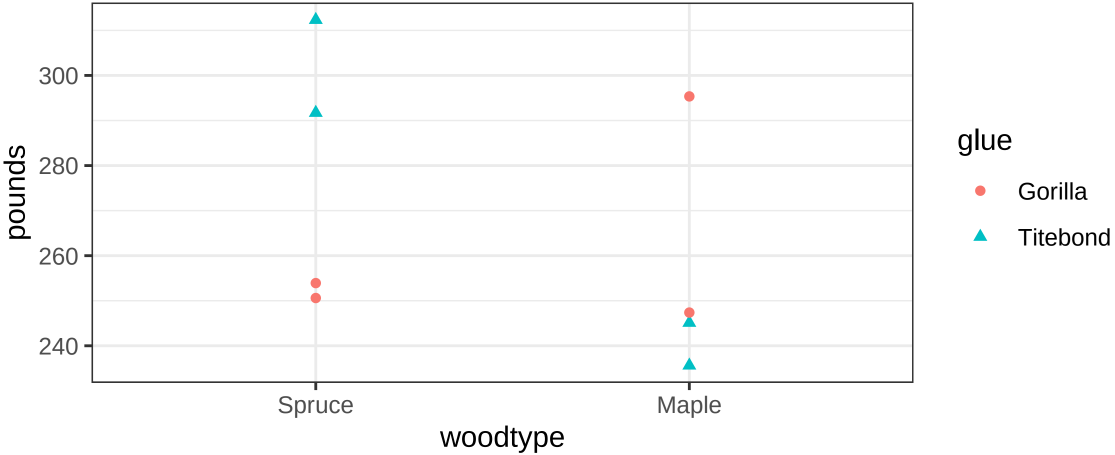
Assessing an interaction using a t-test

```
Call:
lm(formula = pounds ~ glue * woodtype, data = d)

Residuals:
      1       2       3       4       5       6       7       8 
 10.379 -17.616 -10.379  17.616  -4.670  -4.054   4.670   4.054 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                 242.134     10.680  22.671 2.24e-05 ***
glueTitebond                 51.285     15.104   3.395   0.0274 *  
woodtypeMaple                 3.127     15.104   0.207   0.8461    
glueTitebond:woodtypeMaple    3.070     21.361   0.144   0.8927    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 15.1 on 4 degrees of freedom
Multiple R-squared:  0.8605,Adjusted R-squared:  0.7558 
F-statistic: 8.223 on 3 and 4 DF,  p-value: 0.03475
```

Assessing an interaction using an F-test

```
anova(m)

Analysis of Variance Table

Response: pounds
              Df Sum Sq Mean Sq F value   Pr(>F)
glue           1 5580.0  5580.0 24.4582 0.007786 **
woodtype       1   43.5    43.5  0.1905 0.685012
glue:woodtype  1    4.7     4.7  0.0207 0.892654
Residuals      4  912.6   228.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

drop1(m, test='F')

Single term deletions

Model:
pounds ~ glue * woodtype
              Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                     912.58 45.895
glue:woodtype  1     4.714 917.30 43.936  0.0207 0.8927
```

What if this had been your data?



Assessing an interaction using a t-test

```
Call:
lm(formula = pounds ~ glue * woodtype, data = d)

Residuals:
      1       2       3       4       5       6       7       8 
  1.657  -1.657 -10.312  10.312  -4.741  23.986   4.741 -23.986 

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)    
(Intercept)                  252.26      13.29  18.976 4.54e-05 ***
glueTitebond                  49.76      18.80   2.647   0.0572 .  
woodtypeMaple                 19.10      18.80   1.016   0.3670    
glueTitebond:woodtypeMaple   -80.76      26.59  -3.038   0.0385 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.8 on 4 degrees of freedom
Multiple R-squared:  0.7544,Adjusted R-squared:  0.5702 
F-statistic: 4.095 on 3 and 4 DF,  p-value: 0.1034
```

Unreplicated study

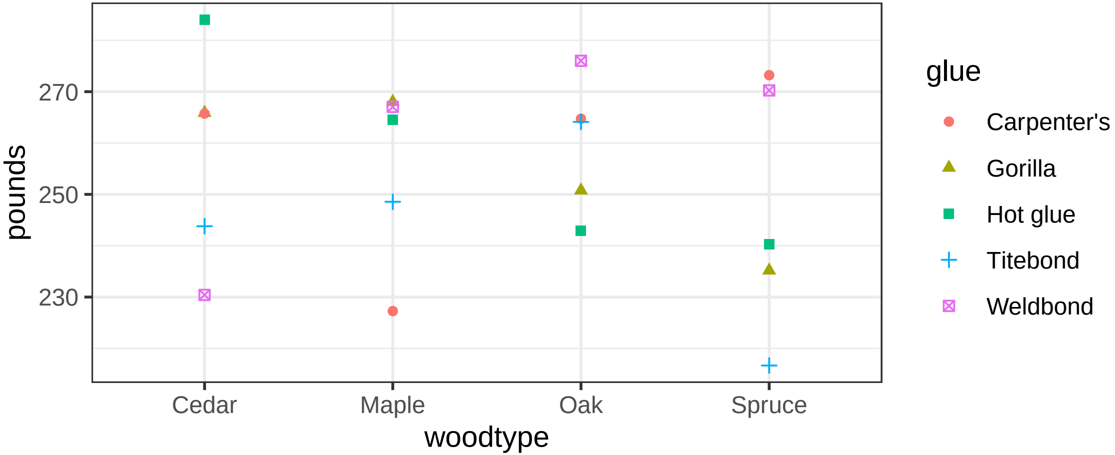
Suppose you now have

- 5 glue choices
- 4 different types of wood with
- 5 samples of each type of wood.

Thus you can only run each glue choice once on each type of wood.

Then you can run an unreplicated RCBD.

Visualize



pounds
270
250
230
Cedar
Maple
Oak
Spruce
woodtype
glue
Carpenter's
Gorilla
Hot glue
Titebond
Weldbond

Fit the main effects (or additive) model

```
m <- lm(pounds ~ glue + woodtype, data = d)
anova(m)

Analysis of Variance Table

Response: pounds
          Df Sum Sq Mean Sq F value Pr(>F)
glue       4  754.3  188.58  0.4332 0.7822
woodtype   3  465.1  155.04  0.3562 0.7857
Residuals 12 5223.7  435.31
```

Fit the main effects (or additive) model

```
Call:
lm(formula = pounds ~ glue + woodtype, data = d)

Residuals:
    Min      1Q  Median      3Q     Max 
-33.498 -10.327   5.084  10.989  23.325 

Coefficients:
               Estimate Std. Error t value Pr(>|t|)    
(Intercept)    260.7220    13.1956  19.758 1.61e-10 ***
glueGorilla     -2.7764    14.7531  -0.188    0.854    
glueHot glue     0.2159    14.7531   0.015    0.989    
glueTitebond   -14.4517    14.7531  -0.980    0.347    
glueWeldbond     3.1903    14.7531   0.216    0.832    
woodtypeMaple   -2.8726    13.1956  -0.218    0.831    
woodtypeOak      1.7564    13.1956   0.133    0.896    
woodtypeSpruce -10.8349    13.1956  -0.821    0.428    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.86 on 12 degrees of freedom
Multiple R-squared:  0.1893,Adjusted R-squared:  -0.2837 
F-statistic: 0.4002 on 7 and 12 DF,  p-value: 0.8845
```

Fit the full (with interaction) model

```
Warning in anova.lm(m):  ANOVA F-tests on an essentially perfect fit are unreliable


Analysis of Variance Table

Response: pounds
              Df Sum Sq Mean Sq F value Pr(>F)
glue           4  754.3  188.58     NaN    NaN
woodtype       3  465.1  155.04     NaN    NaN
glue:woodtype 12 5223.7  435.31     NaN    NaN
Residuals      0    0.0     NaN
```

Fit the full (with interaction) model

```
Call:
lm(formula = pounds ~ glue * woodtype, data = d)

Residuals:
ALL 20 residuals are 0: no residual degrees of freedom!

Coefficients:
                            Estimate Std. Error t value Pr(>|t|)
(Intercept)                 265.7301        NaN     NaN      NaN
glueGorilla                   0.1451        NaN     NaN      NaN
glueHot glue                 18.2476        NaN     NaN      NaN
glueTitebond                -21.9394        NaN     NaN      NaN
glueWeldbond                -35.3158        NaN     NaN      NaN
woodtypeMaple               -38.4658        NaN     NaN      NaN
woodtypeOak                  -1.0001        NaN     NaN      NaN
woodtypeSpruce                7.4822        NaN     NaN      NaN
glueGorilla:woodtypeMaple    40.6031        NaN     NaN      NaN
glueHot glue:woodtypeMaple   19.0424        NaN     NaN      NaN
glueTitebond:woodtypeMaple   43.2335        NaN     NaN      NaN
glueWeldbond:woodtypeMaple   75.0869        NaN     NaN      NaN
glueGorilla:woodtypeOak     -14.1101        NaN     NaN      NaN
glueHot glue:woodtypeOak    -40.0202        NaN     NaN      NaN
glueTitebond:woodtypeOak     21.3197        NaN     NaN      NaN
glueWeldbond:woodtypeOak     46.5929        NaN     NaN      NaN
glueGorilla:woodtypeSpruce  -38.1789        NaN     NaN      NaN
glueHot glue:woodtypeSpruce -51.1490        NaN     NaN      NaN
glueTitebond:woodtypeSpruce -34.6024        NaN     NaN      NaN
glueWeldbond:woodtypeSpruce  32.3448        NaN     NaN      NaN
```

Summary

- Designs:
  - Completely randomized design (CRD)
  - Randomized complete block design (RCBD)
- Deviations
  - Unbalanced
  - Incomplete
  - Unreplicated

R09 - Analysis of Experiments with Two Factors

Two-way ANOVA and Contrasts

STAT 5870 (Engineering)
Iowa State University

November 22, 2024

Two factors

Consider the question of the affect of variety and density on yield under various experimental designs:

- Balanced, complete design
- Unbalanced, complete
- Incomplete

We will also consider the problem of finding the density that maximizes yield.

Data

An experiment was run on tomato plants to determine the effect of

- 3 different varieties (A,B,C) and
- 4 different planting densities (10,20,30,40)

on yield.

A balanced completely randomized design (CRD) with replication was used.

- complete: each treatment (variety × density) is represented
- balanced: each treatment has the same number of replicates
- randomized: treatment was randomly assigned to the plot
- replication: each treatment is represented more than once

This is also referred to as a full factorial or fully crossed design.

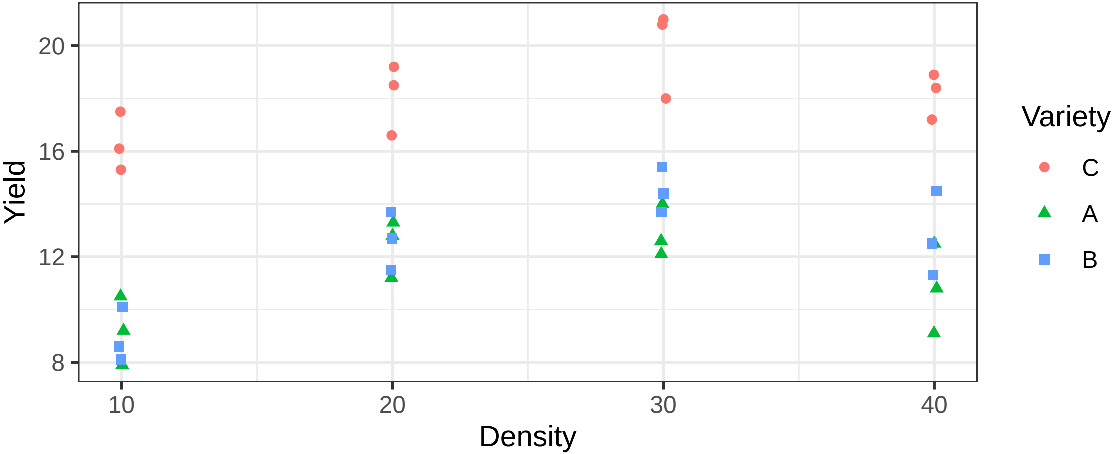
Hypotheses

- How does variety affect mean yield?
  - How is the mean yield for variety A different from B on average?
  - How is the mean yield for variety A different from B at a particular value for density?
- How does density affect mean yield?
  - How is the mean yield for density 10 different from density 20 on average?
  - How is the mean yield for density 10 different from density 20 at a particular value for variety?
- How does density affect yield differently for each variety?

For all of these questions, we want to know

- is there any effect and
- if yes, what is the magnitude and direction of the effect.

Confidence/credible intervals can answer these questions.


Yield
20
16
12
8
10
20
30
40
Density
Variety
C
A
B

Summary statistics

```
# A tibble: 12 x 5
# Groups:   Variety [3]
   Variety Density     n  mean    sd
   <fct>     <int> <int> <dbl> <dbl>
 1 C            10     3 16.3  1.11
 2 C            20     3 18.1  1.35
 3 C            30     3 19.9  1.68
 4 C            40     3 18.2  0.874
 5 A            10     3  9.2  1.3
 6 A            20     3 12.4  1.10
 7 A            30     3 12.9  0.985
 8 A            40     3 10.8  1.7
 9 B            10     3  8.93 1.04
10 B            20     3 12.6  1.10
11 B            30     3 14.5  0.854
12 B            40     3 12.8  1.62
```

Two-way ANOVA

- Setup: Two categorical explanatory variables with I and J levels respectively
- Model:

$$Y_{ijk} \overset{ind}{\sim} N(\mu_{ij}, \sigma^2)$$

where Y_{ijk} is the

- kth observation at the
- ith level of variable 1 (variety) with $i = 1, \ldots, \mathrm{I}$ and the
- jth level of variable 2 (density) with $j = 1, \ldots, \mathrm{J}$.

Consider the models:

- Additive/Main effects: $\mu_{ij} = \mu + \nu_i + \delta_j$
- Cell-means: $\mu_{ij} = \mu + \nu_i + \delta_j + \gamma_{ij}$

	10	20	30	40
A	μ_{11}	μ_{12}	μ_{13}	μ_{14}
B	μ_{21}	μ_{22}	μ_{23}	μ_{24}
C	μ_{31}	μ_{32}	μ_{33}	μ_{34}

As a regression model

1. Assign a reference level for both variety (C) and density (40).
2. Let V_i and D_i be the variety and density for observation i.
3. Build indicator variables, e.g. $\mathrm{I}(V_i = A)$ and $\mathrm{I}(D_i = 10)$.
4. The additive/main effects model:

$$\begin{aligned}\mu_i = \quad & \beta_0 \\ & +\beta_1 \mathrm{I}(V_i = A) + \beta_2 \mathrm{I}(V_i = B) \\ & +\beta_3 \mathrm{I}(D_i = 10) + \beta_4 \mathrm{I}(D_i = 20) + \beta_5 \mathrm{I}(D_i = 30).\end{aligned}$$

5. The cell-means model:

$$\begin{aligned}\mu_i = \quad & \beta_0 \\ & +\beta_1 \mathrm{I}(V_i = A) + \beta_2 \mathrm{I}(V_i = B) \\ & +\beta_3 \mathrm{I}(D_i = 10) + \beta_4 \mathrm{I}(D_i = 20) + \beta_5 \mathrm{I}(D_i = 30) \\ \\ & +\beta_6 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 10) + \beta\ _7 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 20) + \beta\ _8 \mathrm{I}(V_i = A)\mathrm{I}(D_i = 30) \\ & +\beta_9 \mathrm{I}(V_i = B)\mathrm{I}(D_i = 10) + \beta_{10} \mathrm{I}(V_i = B)\mathrm{I}(D_i = 20) + \beta_{11} \mathrm{I}(V_i = B)\mathrm{I}(D_i = 30)\end{aligned}$$

ANOVA Table

ANOVA Table - Additive/Main Effects model

Source	SS	df	MS	F
Factor A	SSA	(I-1)	SSA/(I-1)	MSA/MSE
Factor B	SSB	(J-1)	SSB/(J-1)	MSB/MSE
Error	SSE	n-I-J+1	SSE/(n-I-J+1)	
Total	SST	n-1		

ANOVA Table - Cell-means model

Source	SS	df	MS	
Factor A	SSA	I-1	SSA/(I-1)	MSA/MSE
Factor B	SSB	J-1	SSB/(J-1)	MSB/MSE
Interaction AB	SSAB	(I-1)(J-1)	SSAB /(I-1)(J-1)	MSAB/MSE
Error	SSE	n-IJ	SSE/(n-IJ)	
Total	SST	n-1		

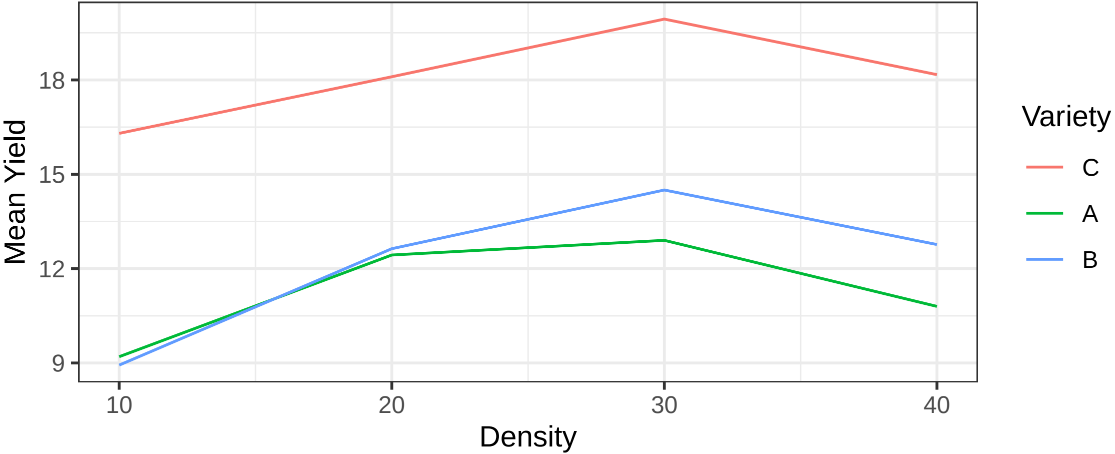
```
tomato$Density = factor(tomato$Density)
m = lm(Yield~Variety+Density, tomato)
drop1(m, test="F")

Single term deletions

Model:
Yield ~ Variety + Density
        Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                46.07 20.880
Variety  2    327.60 373.67 92.235 106.659 2.313e-14 ***
Density  3     86.69 132.76 52.980  18.816 4.690e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

m = lm(Yield~Variety*Density, tomato)
drop1(m, scope = ~Variety+Density+Variety:Density, test="F")

Single term deletions

Model:
Yield ~ Variety * Density
                Df Sum of Sq     RSS    AIC F value    Pr(>F)
<none>                        38.040 25.984
Variety          2   104.749 142.789 69.603 33.0438 1.278e-07 ***
Density          3    19.809  57.849 35.076  4.1660   0.01648 *
Variety:Density  6     8.032  46.072 20.880  0.8445   0.54836
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Additive vs cell-means

Opinions differ on whether to use an additive vs a cell-means model when the interaction is not significant. Remember that an insignificant test does not prove that there is no interaction.

	Additive	Cell-means
Interpretation	Direct	More complicated
Estimate of σ^2	Biased	Unbiased

We will continue using the cell-means model to answer the scientific questions of interest.


Mean Yield
18
15
12
9
10
20
30
40
Density
Variety
C
A
B

Two-way ANOVA in R

```
tomato$Density = factor(tomato$Density)
m = lm(Yield~Variety*Density, tomato)
anova(m)
```

```
Analysis of Variance Table

Response: Yield
                Df Sum Sq Mean Sq  F value    Pr(>F)
Variety          2 327.60 163.799 103.3430 1.608e-12 ***
Density          3  86.69  28.896  18.2306 2.212e-06 ***
Variety:Density  6   8.03   1.339   0.8445    0.5484
Residuals       24  38.04   1.585
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variety comparison

```
library(emmeans)
emmeans(m, pairwise˜Variety)

$emmeans
 Variety emmean    SE df lower.CL upper.CL
 C         18.1 0.363 24     17.4     18.9
 A         11.3 0.363 24     10.6     12.1
 B         12.2 0.363 24     11.5     13.0

Results are averaged over the levels of: Density
Confidence level used: 0.95

$contrasts
 contrast estimate    SE df t.ratio p.value
 C - A       6.792 0.514 24  13.214  <.0001
 C - B       5.917 0.514 24  11.512  <.0001
 A - B      -0.875 0.514 24  -1.702  0.2249

Results are averaged over the levels of: Density
P value adjustment: tukey method for comparing a family of 3 estimates
```

Density comparison

```
emmeans(m, pairwise~Density)

$emmeans
 Density emmean   SE df lower.CL upper.CL
 10        11.5 0.42 24     10.6     12.3
 20        14.4 0.42 24     13.5     15.3
 30        15.8 0.42 24     14.9     16.6
 40        13.9 0.42 24     13.0     14.8

Results are averaged over the levels of: Variety
Confidence level used: 0.95

$contrasts
 contrast              estimate    SE df t.ratio p.value
 Density10 - Density20   -2.911 0.593 24  -4.905  0.0003
 Density10 - Density30   -4.300 0.593 24  -7.245  <.0001
 Density10 - Density40   -2.433 0.593 24  -4.100  0.0022
 Density20 - Density30   -1.389 0.593 24  -2.340  0.1169
 Density20 - Density40    0.478 0.593 24   0.805  0.8514
 Density30 - Density40    1.867 0.593 24   3.145  0.0213

Results are averaged over the levels of: Variety
P value adjustment: tukey method for comparing a family of 4 estimates
```

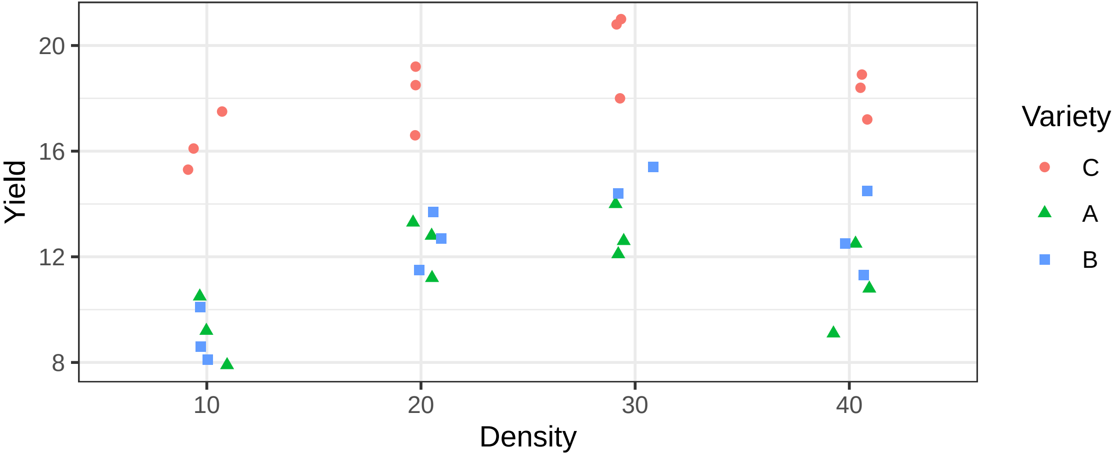
```
emmeans(m, pairwise~Variety*Density)

$emmeans
 Variety Density emmean    SE df lower.CL upper.CL
 C       10       16.30 0.727 24    14.80     17.8
 A       10        9.20 0.727 24     7.70     10.7
 B       10        8.93 0.727 24     7.43     10.4
 C       20       18.10 0.727 24    16.60     19.6
 A       20       12.43 0.727 24    10.93     13.9
 B       20       12.63 0.727 24    11.13     14.1
 C       30       19.93 0.727 24    18.43     21.4
 A       30       12.90 0.727 24    11.40     14.4
 B       30       14.50 0.727 24    13.00     16.0
 C       40       18.17 0.727 24    16.67     19.7
 A       40       10.80 0.727 24     9.30     12.3
 B       40       12.77 0.727 24    11.27     14.3

Confidence level used: 0.95

$contrasts
 contrast                  estimate   SE df t.ratio p.value
 C Density10 - A Density10   7.1000 1.03 24   6.907  <.0001
 C Density10 - B Density10   7.3667 1.03 24   7.166  <.0001
 C Density10 - C Density20  -1.8000 1.03 24  -1.751  0.8276
 C Density10 - A Density20   3.8667 1.03 24   3.762  0.0356
 C Density10 - B Density20   3.6667 1.03 24   3.567  0.0543
 C Density10 - C Density30  -3.6333 1.03 24  -3.535  0.0582
 C Density10 - A Density30   3.4000 1.03 24   3.308  0.0932
 C Density10 - B Density30   1.8000 1.03 24   1.751  0.8276
 C Density10 - C Density40  -1.8667 1.03 24  -1.816  0.7947
 C Density10 - A Density40   5.5000 1.03 24   5.350  0.0008
 C Density10 - B Density40   3.5333 1.03 24   3.437  0.0714
```

Summary

- Use emmeans to answer questions of scientific interest.
- Check model assumptions
- Consider alternative models, e.g. treating density as continuous

Unbalanced design

Suppose for some reason that a variety B, density 30 sample was contaminated. Although you started with a balanced design, the data is now unbalanced. Fortunately, we can still use the tools we have used previously.


Yield
20
16
12
8
10
20
30
40
Density
Variety
● C
▲ A
■ B

Summary statistics

```
# A tibble: 12 x 5
# Groups:   Variety [3]
   Variety Density     n  mean    sd
   <fct>   <fct>   <int> <dbl> <dbl>
 1 C       10          3 16.3  1.11
 2 C       20          3 18.1  1.35
 3 C       30          3 19.9  1.68
 4 C       40          3 18.2  0.874
 5 A       10          3  9.2  1.3
 6 A       20          3 12.4  1.10
 7 A       30          3 12.9  0.985
 8 A       40          3 10.8  1.7
 9 B       10          3  8.93 1.04
10 B       20          3 12.6  1.10
11 B       30          2 14.9  0.707
12 B       40          3 12.8  1.62
```

Two-way ANOVA in R

```
m = lm(Yield~Variety*Density, tomato_unbalanced)
anova(m)
```

```
Analysis of Variance Table

Response: Yield
                Df Sum Sq Mean Sq F value    Pr(>F)    
Variety          2 329.99 164.994 102.343 3.552e-12 ***
Density          3  84.45  28.150  17.461 3.947e-06 ***
Variety:Density  6   8.80   1.467   0.910    0.5052    
Residuals       23  37.08   1.612                      
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Variety comparison

```
emmeans(m, pairwise~Variety)

$emmeans
 Variety emmean    SE df lower.CL upper.CL
 C         18.1 0.367 23     17.4     18.9
 A         11.3 0.367 23     10.6     12.1
 B         12.3 0.389 23     11.5     13.1

Results are averaged over the levels of: Density
Confidence level used: 0.95

$contrasts
 contrast estimate    SE df t.ratio p.value
 C - A       6.792 0.518 23  13.102  <.0001
 C - B       5.817 0.534 23  10.886  <.0001
 A - B      -0.975 0.534 23  -1.825  0.1839

Results are averaged over the levels of: Density
P value adjustment: tukey method for comparing a family of 3 estimates
```

Density comparison

```
emmeans(m, pairwise~Density)

$emmeans
 Density emmean    SE df lower.CL upper.CL
 10        11.5 0.423 23     10.6     12.4
 20        14.4 0.423 23     13.5     15.3
 30        15.9 0.457 23     15.0     16.9
 40        13.9 0.423 23     13.0     14.8

Results are averaged over the levels of: Variety
Confidence level used: 0.95

$contrasts
 contrast              estimate    SE df t.ratio p.value
 Density10 - Density20   -2.911 0.599 23  -4.864  0.0004
 Density10 - Density30   -4.433 0.623 23  -7.116  <.0001
 Density10 - Density40   -2.433 0.599 23  -4.065  0.0025
 Density20 - Density30   -1.522 0.623 23  -2.443  0.0967
 Density20 - Density40    0.478 0.599 23   0.798  0.8545
 Density30 - Density40    2.000 0.623 23   3.210  0.0189

Results are averaged over the levels of: Variety
P value adjustment: tukey method for comparing a family of 4 estimates
```

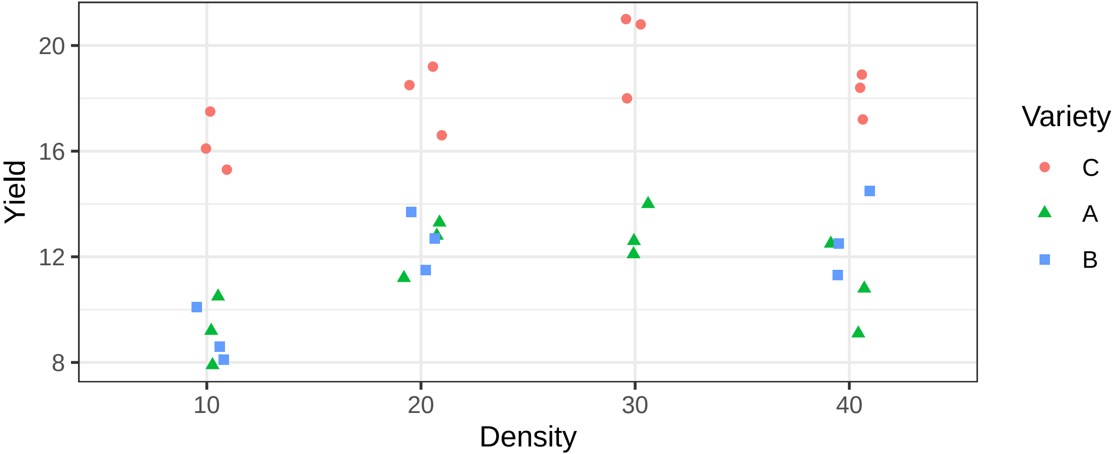
```
emmeans(m, pairwise~Variety*Density)

$emmeans
 Variety Density emmean    SE df lower.CL upper.CL
 C       10       16.30 0.733 23    14.78     17.8
 A       10        9.20 0.733 23     7.68     10.7
 B       10        8.93 0.733 23     7.42     10.4
 C       20       18.10 0.733 23    16.58     19.6
 A       20       12.43 0.733 23    10.92     13.9
 B       20       12.63 0.733 23    11.12     14.1
 C       30       19.93 0.733 23    18.42     21.4
 A       30       12.90 0.733 23    11.38     14.4
 B       30       14.90 0.898 23    13.04     16.8
 C       40       18.17 0.733 23    16.65     19.7
 A       40       10.80 0.733 23     9.28     12.3
 B       40       12.77 0.733 23    11.25     14.3

Confidence level used: 0.95

$contrasts
 contrast                  estimate   SE df t.ratio p.value
 C Density10 - A Density10   7.1000 1.04 23   6.849  <.0001
 C Density10 - B Density10   7.3667 1.04 23   7.106  <.0001
 C Density10 - C Density20  -1.8000 1.04 23  -1.736  0.8341
 C Density10 - A Density20   3.8667 1.04 23   3.730  0.0396
 C Density10 - B Density20   3.6667 1.04 23   3.537  0.0597
 C Density10 - C Density30  -3.6333 1.04 23  -3.505  0.0638
 C Density10 - A Density30   3.4000 1.04 23   3.280  0.1008
 C Density10 - B Density30   1.4000 1.16 23   1.208  0.9828
 C Density10 - C Density40  -1.8667 1.04 23  -1.801  0.8022
 C Density10 - A Density40   5.5000 1.04 23   5.305  0.0011
 C Density10 - B Density40   3.5333 1.04 23   3.408  0.0778
```

Unbalanced Summary

The analysis can be completed just like the balanced design using `emmeans` to answer scientific questions of interest.

Incomplete design

Suppose none of the samples from variety B, density 30 were obtained. Now the analysis becomes more complicated.


Yield
20
16
12
8
10
20
30
40
Density
Variety
C
A
B

Summary statistics

```
# A tibble: 11 x 5
# Groups:   Variety [3]
   Variety Density     n  mean    sd
   <fct>   <fct>   <int> <dbl> <dbl>
 1 C       10          3 16.3  1.11
 2 C       20          3 18.1  1.35
 3 C       30          3 19.9  1.68
 4 C       40          3 18.2  0.874
 5 A       10          3  9.2  1.3
 6 A       20          3 12.4  1.10
 7 A       30          3 12.9  0.985
 8 A       40          3 10.8  1.7
 9 B       10          3  8.93 1.04
10 B       20          3 12.6  1.10
11 B       40          3 12.8  1.62
```

Treat as a One-way ANOVA

When the design is incomplete, use a one-way ANOVA combined with contrasts to answer questions of interest. For example, to compare the average difference between B and C, we want to only compare at densities 10, 20, and 40.

	10	20	30	40
A	μ_{11}	μ_{12}	μ_{13}	μ_{14}
B	μ_{21}	μ_{22}		μ_{24}
C	μ_{31}	μ_{32}	μ_{33}	μ_{34}

Thus, the contrast is

$$\begin{aligned}\gamma &= \tfrac{1}{3}(\mu_{31}+\mu_{32}+\mu_{34}) - \tfrac{1}{3}(\mu_{21}+\mu_{22}+\mu_{24}) \\ &= \tfrac{1}{3}(\mu_{31}+\mu_{32}+\mu_{34}-\mu_{21}-\mu_{22}-\mu_{24})\end{aligned}$$

The Regression model

The regression model here considers variety-density combination as a single explanatory variable with 11 levels: A10, A20, A30, A40, B10, B20, B40, C10, C20, C30, and C40. Let C40 be the reference level. For observation i, let

- Y_i be the yield
- V_i be the variety
- D_i be the density

The model is then $Y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$ and

$$\begin{aligned}
\mu_i \quad &= \beta_0 \\
&+\beta_1 \mathrm{I}(V_i = A, D_i = 10) + \beta_2 \mathrm{I}(V_i = A, D_i = 20) + \beta_3 \mathrm{I}(V_i = A, D_i = 30) \; +\beta_4 \mathrm{I}(V_i = A, D_i = 40) \\
&+\beta_5 \mathrm{I}(V_i = B, D_i = 10) + \beta_6 \mathrm{I}(V_i = B, D_i = 20) \qquad\qquad\qquad\qquad\quad +\beta_7 \mathrm{I}(V_i = B, D_i = 40) \\
&+\beta_8 \mathrm{I}(V_i = C, D_i = 10) + \beta_9 \mathrm{I}(V_i = C, D_i = 20) + \beta_{10} \mathrm{I}(V_i = C, D_i = 30)
\end{aligned}$$

Two-way ANOVA in R

```
m <- lm(Yield ~ Variety*Density, data=tomato_incomplete)
anova(m)


Analysis of Variance Table

Response: Yield
                Df Sum Sq Mean Sq F value    Pr(>F)
Variety          2 347.38 173.691 104.462 5.868e-12 ***
Density          3  66.65  22.218  13.362 3.514e-05 ***
Variety:Density  5   7.06   1.412   0.849      0.53
Residuals       22  36.58   1.663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

How can you tell the design is not complete?

One-way ANOVA in R

```
m = lm(Yield~Variety:Density, tomato_incomplete)
anova(m)

Analysis of Variance Table

Response: Yield
                Df Sum Sq Mean Sq F value    Pr(>F)
Variety:Density 10 421.09  42.109  25.326 8.563e-10 ***
Residuals       22  36.58   1.663
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Contrasts

```
m = lm(Yield ~ VarietyDensity, tomato_incomplete)
em <- emmeans(m, ~ VarietyDensity)
contrast(em, method = list(
#         A10 A20 A30 A40 B10 B20     B40 C10 C20 C30 C40
"C-B" = c(  0,  0,  0,  0, -1, -1,     -1,  1,  1,  0,  1)/3,
"C-A" = c( -1, -1, -1, -1,  0,  0,      0,  1,  1,  1,  1)/4,
"B-A" = c( -1, -1,  0, -1,  1,  1,      1,  0,  0,  0,  0)/3)) |>
  confint()

 contrast estimate    SE df lower.CL upper.CL
 C-B         6.078 0.608 22    4.817     7.34
 C-A         6.792 0.526 22    5.700     7.88
 B-A         0.633 0.608 22   -0.627     1.89

Confidence level used: 0.95
```

```
m = lm(Yield~Variety:Density, tomato_incomplete)
emmeans(m, pairwise~Variety:Density) # We could have used the VarietyDensity model, but this looks nicer

$emmeans
 Variety Density emmean    SE df lower.CL upper.CL
 C       10       16.30 0.744 22    14.76     17.8
 A       10        9.20 0.744 22     7.66     10.7
 B       10        8.93 0.744 22     7.39     10.5
 C       20       18.10 0.744 22    16.56     19.6
 A       20       12.43 0.744 22    10.89     14.0
 B       20       12.63 0.744 22    11.09     14.2
 C       30       19.93 0.744 22    18.39     21.5
 A       30       12.90 0.744 22    11.36     14.4
 B       30      nonEst    NA NA       NA       NA
 C       40       18.17 0.744 22    16.62     19.7
 A       40       10.80 0.744 22     9.26     12.3
 B       40       12.77 0.744 22    11.22     14.3

Confidence level used: 0.95

$contrasts
 contrast                  estimate   SE df t.ratio p.value
 C Density10 - A Density10   7.1000 1.05 22   6.744  <.0001
 C Density10 - B Density10   7.3667 1.05 22   6.997  <.0001
 C Density10 - C Density20  -1.8000 1.05 22  -1.710  0.8157
 C Density10 - A Density20   3.8667 1.05 22   3.673  0.0407
 C Density10 - B Density20   3.6667 1.05 22   3.483  0.0606
 C Density10 - C Density30  -3.6333 1.05 22  -3.451  0.0646
 C Density10 - A Density30   3.4000 1.05 22   3.229  0.1007
 C Density10 - B Density30   nonEst   NA NA      NA      NA
 C Density10 - C Density40  -1.8667 1.05 22  -1.773  0.7829
 C Density10 - A Density40   5.5000 1.05 22   5.224  0.0012
```

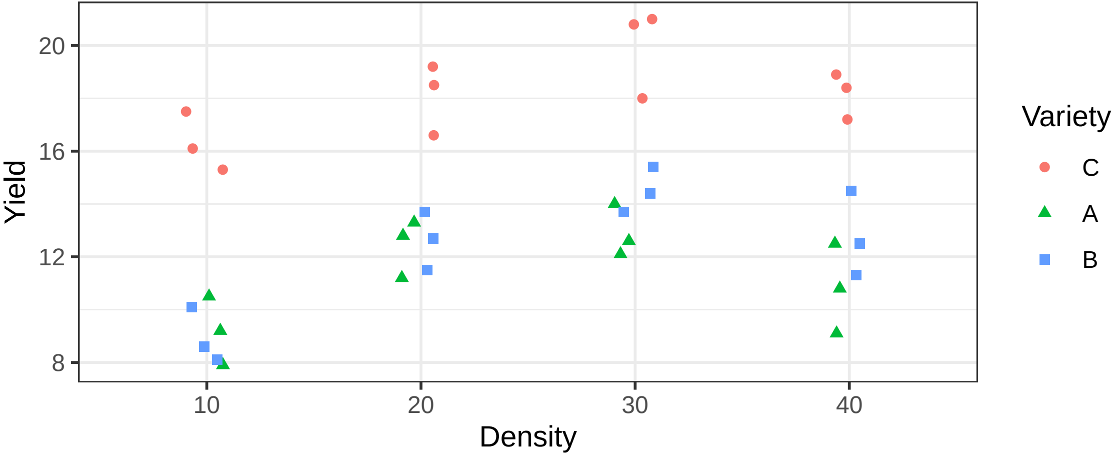
Summary

When dealing with an incomplete design, it is often easier to treat the analysis as a one-way ANOVA and use contrasts to answer scientific questions of interest.

Optimal yield

Now suppose you have the same data set, but your scientific question is different. Specifically, you are interested in choosing a variety-density combination that provides the optimal yield.

You can use the ANOVA analysis to choose from amongst the 3 varieties and one of the 4 densities, but there is no reason to believe that the optimal density will be one of those 4.


Yield
20
16
12
8
10
20
30
40
Density
Variety
● C
▲ A
■ B

Modeling

Considering a single variety, if we assume a linear relationship between Yield (Y_i) and Density (D_i) then the maximum Yield will occur at either $-\infty$ or $+\infty$ which is unreasonable. The easiest way to have a maximum (or minimum) is to assume a quadratic relationship, e.g.

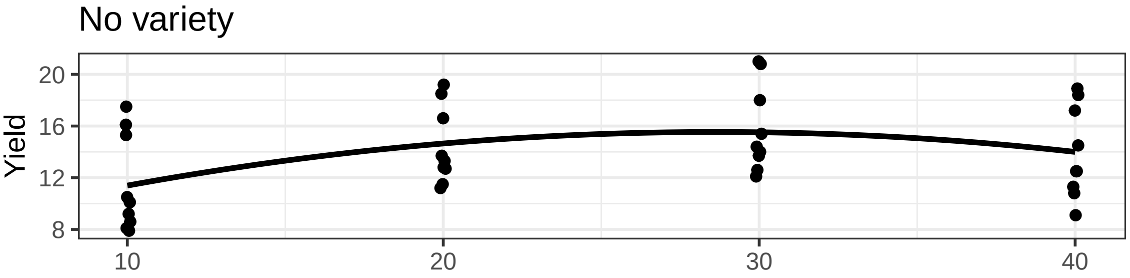
$$E[Y_i] = \mu_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$$

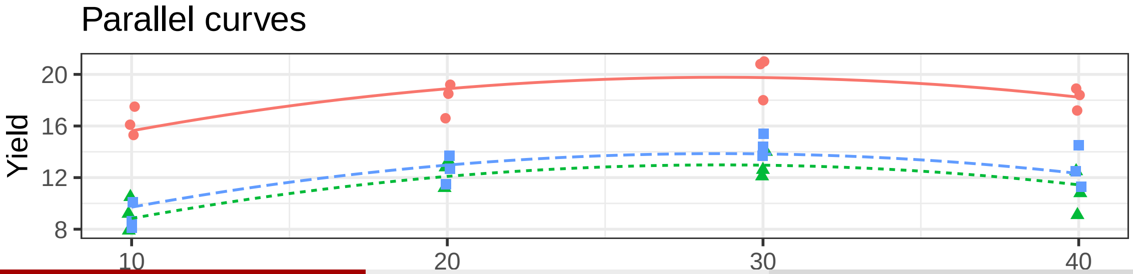
Now we can incorporate Variety (V_i) in many ways. Two options are parallel curves or completely independent curves.
Parallel curves:

$$\begin{aligned} \mu_i = \quad & \beta_0 + \beta_1 D_i + \beta_2 D_i^2 \\ & + \beta_3 \mathrm{I}(V_i = A) + \beta_4 \mathrm{I}(V_i = B) \end{aligned}$$

Independent curves:

$$\begin{aligned} \mu_i = \quad & \beta_0 + \beta_1 D_i + \beta_2 D_i^2 \\ & + \beta_3 \mathrm{I}(V_i = A) + \beta_4 \mathrm{I}(V_i = B) \\ & + \beta_5 \mathrm{I}(V_i = A) D_i + \beta_6 \mathrm{I}(V_i = B) D_i \\ & + \beta_7 \mathrm{I}(V_i = A) D_i^2 + \beta_8 \mathrm{I}(V_i = B) D_i^2 \end{aligned}$$


No variety
Yield
20
16
12
8
10
20
30
40


Parallel curves
Yield
20
16
12
8
10
20
30
40

Finding the maximum

For a particular variety, there will be an equation like

$$E[Y_i] = \mu_i = \beta_0 + \beta_1 D_i + \beta_2 D_i^2$$

where these β_1 and β_2 need not correspond to any particular β_1 and β_2 we have discussed thus far.

If $\beta_2 < 0$, then the quadratic curve has a maximum and it occurs at $-\beta_1/2\beta_2$.

No variety

```
Call:
lm(formula = Yield ~ Density + I(Density^2), data = tomato)

Residuals:
   Min     1Q Median     3Q    Max 
-4.898 -2.721 -1.320  3.364  6.109 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)  
(Intercept)   5.744444   3.128242   1.836   0.0753 .
Density       0.684111   0.285384   2.397   0.0223 *
I(Density^2) -0.011944   0.005618  -2.126   0.0411 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.371 on 33 degrees of freedom
Multiple R-squared:  0.1854,Adjusted R-squared:  0.136 
F-statistic: 3.755 on 2 and 33 DF,  p-value: 0.03395
```

Parallel curves

```
Call:
lm(formula = Yield ~ Density + I(Density^2) + Variety, data = tomato)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.3422 -0.9039  0.1744  0.8082  2.1828 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)   9.980556   1.184193   8.428 1.61e-09 ***
Density       0.684111   0.104707   6.534 2.71e-07 ***
I(Density^2) -0.011944   0.002061  -5.794 2.21e-06 ***
VarietyA     -6.791667   0.504942 -13.450 1.76e-14 ***
VarietyB     -5.916667   0.504942 -11.718 6.39e-13 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.237 on 31 degrees of freedom
Multiple R-squared:  0.897,Adjusted R-squared:  0.8837 
F-statistic: 67.48 on 4 and 31 DF,  p-value: 7.469e-15
```

Independent curves

```
Call:
lm(formula = Yield ~ Density * Variety + I(Density^2) * Variety,
    data = tomato)

Residuals:
     Min       1Q   Median       3Q      Max
-2.04500 -0.82125 -0.01417  0.94000  1.71000

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           11.808333   1.968364   5.999 2.12e-06 ***
Density                0.520167   0.179570   2.897  0.00739 **
VarietyA              -8.458333   2.783687  -3.039  0.00523 **
VarietyB              -9.733333   2.783687  -3.497  0.00165 **
I(Density^2)          -0.008917   0.003535  -2.522  0.01787 *
Density:VarietyA       0.199167   0.253951   0.784  0.43971
Density:VarietyB       0.292667   0.253951   1.152  0.25924
VarietyA:I(Density^2) -0.004417   0.005000  -0.883  0.38482
VarietyB:I(Density^2) -0.004667   0.005000  -0.933  0.35889
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.225 on 27 degrees of freedom
Multiple R-squared:  0.912,Adjusted R-squared:  0.886
F-statistic: 34.99 on 8 and 27 DF,  p-value: 2.678e-12
```

Completely randomized design (CRD)

This semester, we have assumed a completely randomized design. As an example, consider 36 plots and we are randomly assigning our variety-density combinations to the plots such that we have 3 reps of each combination. The result may look something like this

A20	A10	A20	B10	B10	A30
C10	C30	C30	C10	C20	A10
B30	B10	B20	B30	B40	B40
C40	B20	A10	C20	B30	A40
C30	B40	A30	C40	B20	C40
C10	C20	A40	A30	A20	A40

Complete randomized block design (RBD)

A randomized block design is appropriate when there is a nuisance factor that you want to control for. In our example, imagine you had 12 plots at 3 different locations and you expect these locations would have impact on yield. A randomized block design might look like this.

B10	B40
C30	A30
C40	C10
A20	B20
B30	A40
A10	C20

Block 1

C20	B40
C30	A30
C10	B10
A10	A20
B20	C40
B30	A40

Block 2

A20	B30
C10	A30
A10	C30
B20	C40
B40	A40
C20	B10

Block 3

RBD Analysis

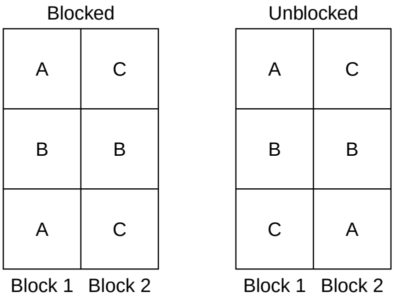
Generally, you will want to model a randomized block design using an additive model for the treatment and blocking factor. If you have the replication, you should test for an interaction. Let's compute the degrees of freedom for the ANOVA tables for this current design considering the variety-density combination as the treatment.

V+D+B		T+B		Cell-means	
Factor	df	Factor	df	Factor	df
Variety	2				
Density	3	Treatment	11	Treatment	11
Block	2	Block	2	Block	2
				Treatment x Block	22
Error	28	Error	22	Error	0
Total	35	Total	35	Total	35

The cell-means model does not have enough degrees of freedom to estimate the interaction because there is no replication of the treatment within a block.

Why block?

Consider a simple experiment with 2 blocks each with 3 experimental units and 3 treatments (A, B, C).



Blocked

Block 1	Block 2
A	C
B	B
A	C

Unblocked

Block 1	Block 2
A	C
B	B
C	A

Let's consider 3 possible analyses:

- Blocked experiment using an additive model for treatment and block (RBD)
- Unblocked experiment using only treatment (CRD)
- Unblocked experiment using an additive model for treatment and block

Why block?

Now suppose, the true model is

$$\mu_{ij} = \mu + T_i + B_j$$

where $T_1 = T_2 = T_3$ and $B_1 = 0$ and $B_2 = \delta$.

In the Blocked experiment using an additive model for treatment and block, the expected treatment differences to all be zero.

In the Unblocked design using only treatment, the expected difference between treatments is

$$\mu_C - \mu_B = \delta \qquad \text{and} \qquad \mu_C - \mu_A = \delta/2.$$

In the Unblocked design using an additive model for treatment and block, we would have an unbalanced design and it would be impossible to compare B and C.

Summary

Block what you can control; randomize what you cannot.