M7S3 - Regression Thoughts

Professor Jarad Niemi

STAT 226 - Iowa State University

November 27, 2018

Outline

- Regression thoughts
    - Properties
        - Coefficient of determination (r^2) is amount of variation explained
        - Not reversible
        - Always through $(\overline{x}, \overline{y})$
        - Residuals sum to zero
        - Residual plots
        - Leverage and influence
    - Cautions
        - Extrapolation
        - Correlation does not imply causation
        - Lurking variables (Simpson's Paradox)
        - Correlations based on average data

Simple linear regression

For a collection of observations (x_i, y_i) for $i = 1, \ldots, n$, we can fit a regression line

$$y_i = b_0 + b_1 x_i + e_i$$

where

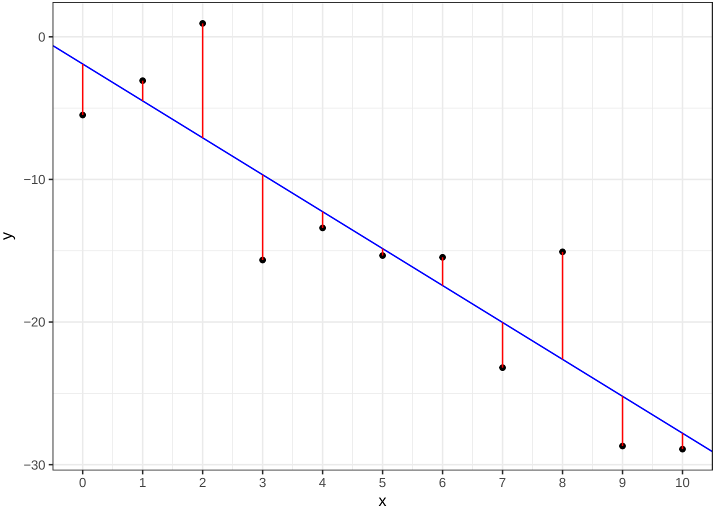
- b_0 is the sample intercept,
- b_1 is the sample slope, and
- e_i is the residual for individual i

by minimizing the sum of squared residuals

$$\sum_{i=1}^{n} e_i^2 \qquad \text{where} \qquad e_i = y_i - \hat{y}_i = y_i - (b_0 + b_1 x_i)$$

and $\hat{y}_i$ is the predicted value for individual i.

Simple linear regression graphically



Coefficient of determination

The sample correlation r measures the direction and strength of the linear relationship between x and y.

Definition

The coefficient of determination

$$r^2 = 1 - \frac{\sum_{i=1}^n e_i^2}{\sum_{i=1}^n (y_i - \overline{y})^2}$$

measures the amount of variability in y that can be explained by the linear relationship between x and y.

Example

The correlation between weekly sales amount and weekly radio ads is 0.98. The coefficient of variation is $r^2 \approx 0.96$. Thus about 96% of the variability in weekly sales amount can be explained by the linear relationsihp with weekly radio ads.

If you are only told r^2, you cannot determine the direction of the relationship.

Symmetric

Correlation is symmetric, the correlation of x with y is the same as the correlation of y with x.

```
cor(x,y)

[1] -0.8866024

cor(y,x)

[1] -0.8866024
```

Thus the coefficient of determination is symmetric.

Equation not reversible

The regression line is

$$y = b_0 + b_1 x$$

but the opposite regression line is not

$$x = -\frac{b_0}{b_1} + \frac{1}{b_1} y.$$

```
regress(y,x)


(Intercept)           x
  -1.904408   -2.589194


-b0/b1; 1/b1


[1] -0.7355215
[1] -0.3862206


regress(x,y)


(Intercept)           x
  0.4915144  -0.3035940
```

Always through $(\overline{x}, \overline{y})$

Recall that knowing any two points is enough to determine a straight line. It can be proved that the regression line always passes through the point $(\overline{x}, \overline{y})$.

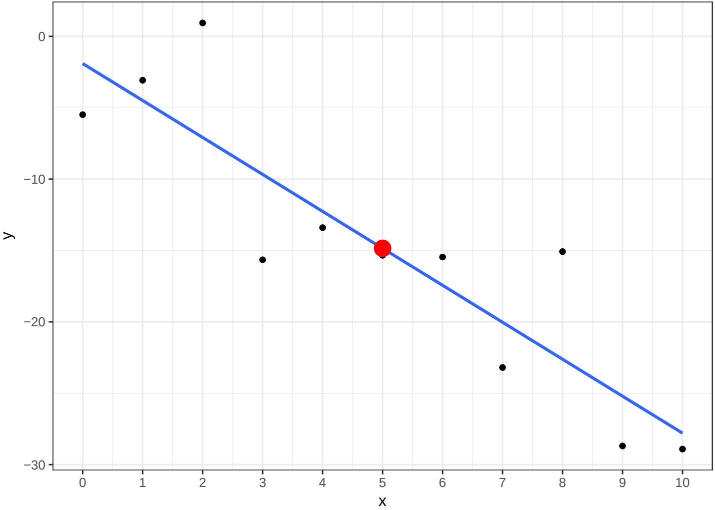
Suppose you know that $\overline{x} = 5$, $\overline{y} = -15$, and the y-intercept is -2. What is the slope?

$$\overline{y} = b_0 + b_1\overline{x} \implies b_1 = (\overline{y} - b_0)/\overline{x}$$

So the slope is

```
(ybar-b0)/xbar

[1] -2.6
```


0
−10
−20
−30
y
0
1
2
3
4
5
6
7
8
9
10
x

Residuals sum to zero

When the regression includes an intercept (b_0), it can be proved that the residuals sum to zero, i.e.

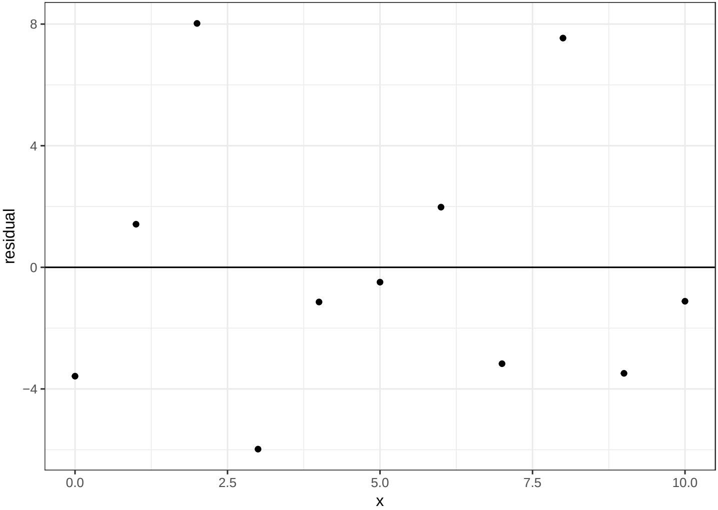
$$\sum_{i=1}^{n} e_i = 0.$$

We will often look at residual plots:

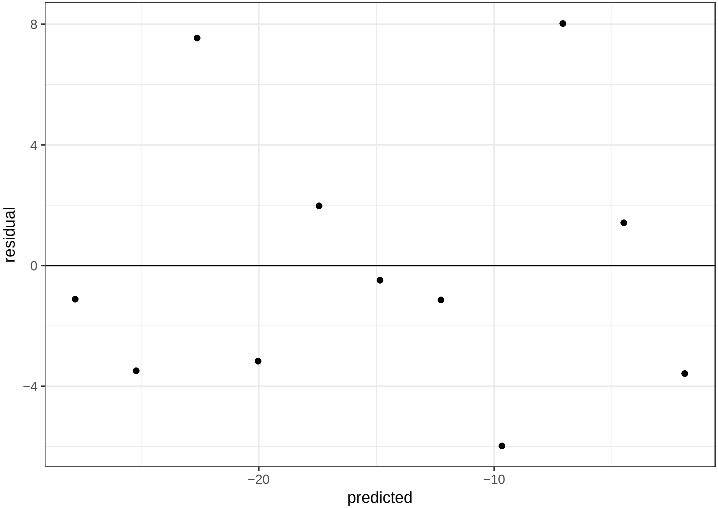
- Residuals vs explanatory variable
- Residuals vs predicted value

These will be centered on 0 due to the result above.

Residual vs explanatory variable



Residual vs predicted

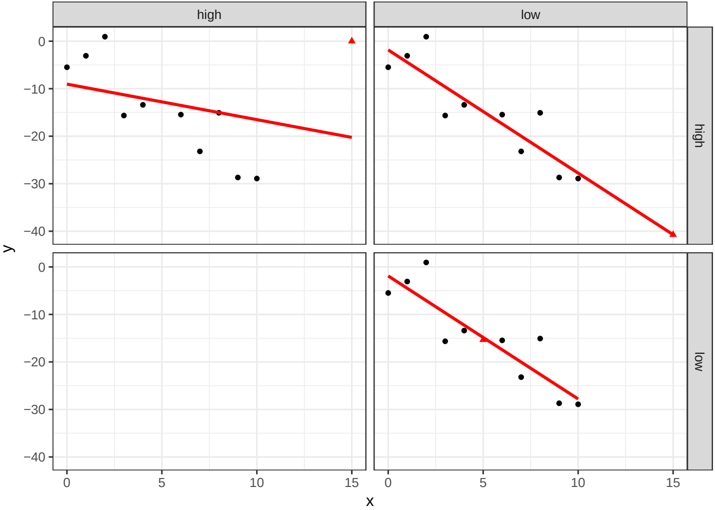


Leverage and influence

Definition

An individual has high leverage if its explanatory variable value is far from the explanatory variable values of the other observations. An individual with high leverage will be an outlier in the x direction. An individual has high influence if its inclusion dramatically affects the fitted regression line. Only individuals with high leverage can have high influence.

Leverage and influence



Correlation does not imply causation

You have all likely heard the addage

correlation does not imply causation.

If two variables have a correlation that is close to -1 or 1, the two variables are highly correlated. This does not mean that one variable causes the other.

Spurious correlations:
`http://www.tylervigen.com/spurious-correlations`

Correlation does not imply causation (cont.)

From https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5402407/:

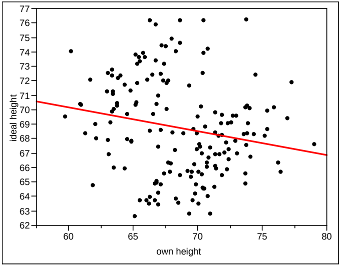
> *My attention was drawn to the recent article by Song at al. entitled "How jet lag impairs Major League Baseball performance" (1), not only by its slightly unusual subject but more importantly because I wondered how one could ever actually prove the effect of jet lag on baseball performance.*
>
> *...Although I do not dispute the large amount of work involved and would be well-nigh incapable of judging the validity of the analyses performed, I must admit that I was taken aback by the way Song et al. (1) systematically present the correlations they identify as direct proof of causality between jet lag and the affected variables. It is actually quite remarkable to me that the word "correlation" does not appear even once in the paper, when this is actually what the authors have been looking at and, in my opinion, to be scientifically accurate, the title of the article should really read: "How jet lag correlates with impairments in Major League Baseball performance."*
>
> *...this tendency to amalgamate correlation with causality is apparently extremely frequent in this field of investigation. But given the broad readership of PNAS and the subject of this article, I feel that it is likely to be relayed by the press and to attract the attention of many people, both scientists and nonscientists.*
>
> *Considering the current tendency to misinterpret scientific data, via the misuse of statistics in particular, I feel that a journal such as PNAS should aim to educate by example, and thus ought to enforce more rigor in the presentation of scientific articles regarding the difference between correlations and proven causality.*

Lurking variables

Definition

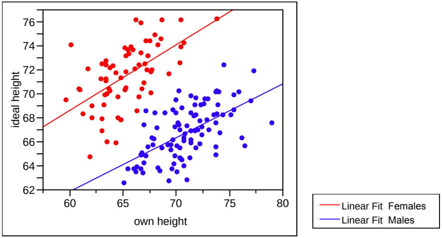
A lurking variable is a variable that has an important effect on the relationship of the variables under investigation, but that is not included in the variables being studied.

What is the relationship between a person's height and their ideal partner's height?



Ideal partner height

In this example, gender is a lurking variable:



```
Linear Fit  Females
predicted ideal height  = 35.798818 + 0.5469203 own height

Linear Fit  Males
predicted ideal height  = 34.971329 + 0.4484906 own height
```

This phenomenon is called Simpson's Paradox.

Correlations based on average data

Correlations based on average data are often much higher (closer to -1 or 1) than correlations based on individual data. This occurs because the averages smooth out the variability between individuals.

Extrapolation

Definition

Extrapolation occurs when making predictions for explanatory variable values below the sample minimum or above the sample maximum of the explanatory variable.

Regression assumes a linear pattern between the response variable and the explanatory variable. Even if this linear assumption is correct for a range of explanatory variable values, there is no reason to expect that this will continue beyond that range.