Name _____

Spring 2017 **STAT 401** **Final exam**
(100 points)

Instructions:

- Full credit will be given only if you show your work.
- The questions are not necessarily ordered from easiest to hardest.
- You are allowed to use any resource except aid from another individual.
- Aid from another individual will automatically earn you a 0.
- Feel free to tear off the last page. There is no need to turn it in.

One-way ANOVA

Suppose you fit two regression models: an intercept-only model and a model with a categorical variable named "Var". The table below provides an estimate for the error variance and its degrees of freedom.

Model	df	$\hat{\sigma}$
Intercept-only	20	3
Intercept with Var	14	2

Use this information to answer the following questions.

1. How many levels of the categorical variable "Var" are there? (1 pts)

   Answer: Since we have added 6 (=20-14) βs (indicator variables) to the model and there must be a reference level, then there are 7 levels to this categorical variable.

2. How many total observations are there? (1 pts)

   Answer: Since the intercept-only model has 20 degrees of freedom and this model only has a single β, we have 21 observations.

3. If the design is balanced, how many replicates are there for each level of the categorical variable "Var"? (2 pts)

   Answer: Since there are 21 observations and 7 levels, we have 3 observations per level in a balanced design.

4. Fill out this one-way ANOVA table below (12 pts)

	SS	df	MS	F	p
Var					
Error					
Total					

   Answer:

	SS	df	MS	F	p
Var	124	6	21	5.25	0.005
Error	56	14	4		
Total	180	20			

5. Interpret this p-value. (4 pts)

   Answer: This p-value indicates the data are incompatible with the null hypothesis model. The null hypothesis model is a normal model with a common mean (intercept-only, i.e. does not include Var) and that assumes the errors are independent, normally distributed, and have a constant variance.

Regression diagnostics

The file `diagnostics.csv` contains a set of 5 response variables (`y1`, `y2`, `y3`, `y4`, and `y5`) and a common explanatory variable x. Consider simple linear regression models for each of the five response variables separately. One of the five response variables meets all model assumptions while each of the other four violates exactly one model assumption. For each response, 1) identify the model assumption violation (if any) and 2) describe how you know that assumption is violated, e.g. what diagnostic plot is informative and what does it look like. (4 pts each)

`y1` Answer: Normality of errors is violated. The qq-plot shows heavy tails.

`y2` Answer: Independence. A plot of residuals vs row number indicates a pattern to the residuals.

`y3` Answer: Constant variance. The residuals vs fitted values plot shows a funnel pattern and the scale-location plot shows increasing residuals as a function of fitted values.

`y4` Answer: No assumptions are violated. All diagnostic plots look reasonable as do residuals vs row number and response vs explanatory value plots.

`y5` Answer: Linearity. Curvature in the residuals vs fitted values plot and response vs explanatory values plot.

Wool

For the following questions, please refer to the "Wool - R Code" page. If you need any background information, please see `?warpbreaks` in `R`.

Write down the model that was used in this analysis. Make sure to define any notation you introduce. (20 pts)

Answer: Define the following notation

- Y_i is the number of breaks for loom i
- W_i is the type of wool for loom i (A or B)
- T_i is the level of tension for loom i (L, M, and H)

The model is

$$\begin{aligned} Y_i &\overset{ind}{\sim} N(\mu_i, \sigma^2) \\ \mu_i &= \beta_0 + \beta_1 \mathrm{I}(W_i = B) + \beta_2 \mathrm{I}(T_i = M) + \beta_3 \mathrm{I}(T_i = H) \end{aligned}$$

Wool (continued)

Provide an interpretation for the following quantities (4 pts each):

- 39.278

  Answer: The is the estimated mean number of breaks for wool A and tension L.

- 11.62

  Answer: The is the estimated error standard deviation, i.e. $\hat{\sigma}$ on the previous page.

- (-17.77790,-2.2221006)

  Answer: This is a 95% confidence/credible inteval for the different in mean number of breaks between tension M and tension L (for both types of wool). Both endpoints of the interval are less than 0 indicating there is statistically significant evidence that tension M has fewer breaks than tension L.

- 26.38889

  Answer: This is the estimated mean number of breaks for tension M averaged over the levels of wool. More formally, it is the average of the means of the number of breaks under tension M for wool A and wool B.

- 4.722222

  Answer: This is the estimated difference in means between tension M and tension H averaged across wool types A and B.

Donation

For the following questions, please use the `donation.csv` data file. These data are filtered version of the data used in the Data Mining competition (see `https://archive.ics.uci.edu/ml/machine-learning-databases/kddcup98-mld/kddcup98.html`) where the filtering only includes Iowa. In addition, only three variables remain: the donation amount from the last targetted mailing (TARGET_D), the type of neighborhood the donor lives in (DOMAIN), and a measure of the donor's wealth (WEALTH2). Fit a linear regression model using log(TARGET_D+1) as the response and DOMAIN and WEALTH2 as the explanatory variables treating WEALTH2 as continuous.

1. Write the R code you used to fit this model. (4 pts)

   Answer:

```
d <- readr::read_csv("donation.csv") # Not necessary for answer
m <- lm(log(TARGET_D+1) ~ WEALTH2 + DOMAIN, data = d)
```

2. Provide an estimate for the multiplicative effect of a one-unit increase in WEALTH2 level on the median TARGET_D. (4 pts)

   Answer:

```
exp(coef(m)[2])

##   WEALTH2
## 0.9945059
```

3. Provide a 95% credible interval for a contrast estimate to compare mean log(TARGET_D+1) for rural (R1,R2, and R3) vs city (C1, C2, and C3) domains averaged over WEALTH2. (8 pts)

   Answer:

```
library("emmeans")
em <- emmeans(m, ~DOMAIN)
#                                          C1 C2 C3 R1 R2 R3 S1 S2 S3 T1 T2 T3
co <- contrast(em, list(`Rural-City`= c(-1,-1,-1, 1, 1, 1, 0, 0, 0, 0, 0, 0)/3))
confint(co)[,5:6]

##     lower.CL   upper.CL
## 1 -0.2246866 0.01965251
```

4. Why might this model not be appropriate for these data? (4 pts)

   Answer: The vast majority of the data have zero for the non-logged response. Also, we may want to treat WEALTH2 as categorical.

(intentionally left blank)

Wool - R Code

```
library("emmeans")
m <- lm(breaks ~ wool + tension, data = warpbreaks)
summary(m)
##
## Call:
## lm(formula = breaks ~ wool + tension, data = warpbreaks)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.500  -8.083  -2.139   6.472  30.722
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   39.278      3.162  12.423  < 2e-16 ***
## woolB         -5.778      3.162  -1.827 0.073614 .
## tensionM     -10.000      3.872  -2.582 0.012787 *
## tensionH     -14.722      3.872  -3.802 0.000391 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.62 on 50 degrees of freedom
## Multiple R-squared:  0.2691,Adjusted R-squared:  0.2253
## F-statistic: 6.138 on 3 and 50 DF,  p-value: 0.00123
confint(m)
##                 2.5 %     97.5 %
## (Intercept)  32.92715 45.6284061
## woolB       -12.12841  0.5728505
## tensionM    -17.77790 -2.2221006
## tensionH    -22.50012 -6.9443228
(em <- emmeans(m, ~tension))
##  tension emmean   SE df lower.CL upper.CL
##  L         36.4 2.74 50     30.9     41.9
##  M         26.4 2.74 50     20.9     31.9
##  H         21.7 2.74 50     16.2     27.2
##
## Results are averaged over the levels of: wool
## Confidence level used: 0.95
co <- contrast(em, "pairwise")
confint(co)
##  contrast estimate   SE df lower.CL upper.CL
##  L - M       10.00 3.87 50    0.647     19.4
##  L - H       14.72 3.87 50    5.369     24.1
##  M - H        4.72 3.87 50   -4.631     14.1
##
## Results are averaged over the levels of: wool
## Confidence level used: 0.95
## Conf-level adjustment: tukey method for comparing a family of 3 estimates
```