Final Exam Practice Problems

Choose the appropriate method:

1. Researchers want to see if there is a relationship between movement speed (in km/sec) and distance to other galaxies (in light years).
2. Do people from different parts of the country have different taste preferences in craft beer? To answer this, we collect random samples from each of 4 different regions of the country (Northeast, South, Midwest, and West) and ask participants what their favorite style of beer is (IPA, brown, wheat, stout, porter, pale ale, other).
3. A group of researchers in Washington D.C. collect information from different polling agencies to determine if they are being biased in what they are reporting. They collect President Obama's approval ratings on a scale from 1-100 for 10 weeks across 3 different agencies (Gallup, Fox, CNN) to see if one of the agencies had a different mean approval rating.
4. A group collected a random sample of 2023 individuals from the population of Greenland residents to see if there was a relationship between eye color (blue, brown, green, other) and hair color (brown, black, blonde, other).

For each of the following, select the best answer. [1 pt each]

5. Suppose that we are comparing four models. According to the AICs below, which model is preferred?

Model	AIC
Model 1	642
Model 2	799
Model 3	-1000
Model 4	92

(a) Model 1

(b) Model 2

(c) Model 3

(d) Model 4

6. Suppose that we have a 95% confidence interval of $(1, 2)$ for β_1. What does it mean to be 95% confident?

(a) There is a 95% chance that β_1 is between 1 and 2.

(b) We are 95% confident that b_1 is between 1 and 2.

(c) We are 95% confident that $\hat{\beta_1}$ is between 1 and 2

(d) 95% of the time, a confidence interval constructed in this manner will contain β_1.

7. Suppose we have a quantitative response variable, a quantitative explanatory variable, and 3 groups. What is the first step when setting up a MLR in this example?

(a) Fit the interaction model to see if multiple slopes are needed.

(b) Fit the additive model and see if groups have different intercepts.

(c) Fit the One Way ANOVA model using groups as the explanatory variable.

(d) Fit the SLR using the quantitative variable only.

8. When checking the regression assumptions, we typically make the following plots: Residuals vs Fitted values, Normal Q-Q, Scale-Location, Residuals vs Leverage. Using these plots, which assumptions can we check? **Circle all that apply**. [2 pts]

   (a) Constant variance.

   (b) Independence.

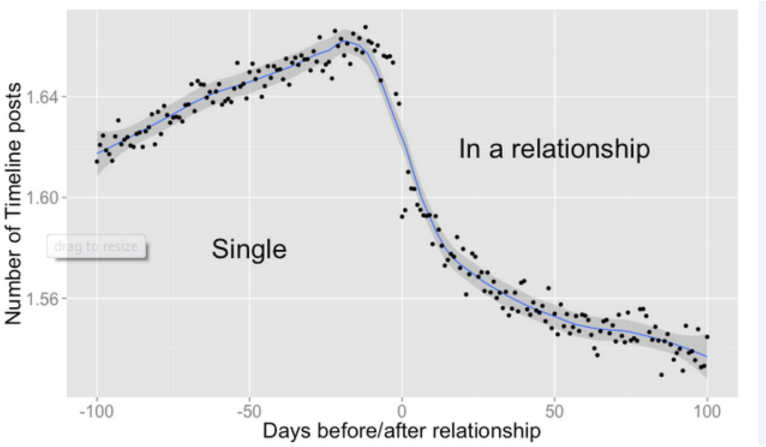
   (c) Normality

   (d) Linearity.

   (e) Randomization/Representation

   (f) No influential observations

9. The plot below shows the number of Facebook posts exchanged between two people who are about to become a couple. Based on this plot, which of the regression assumptions is **most clearly violated**?



   (a) Constant Variance

   (b) Linearity

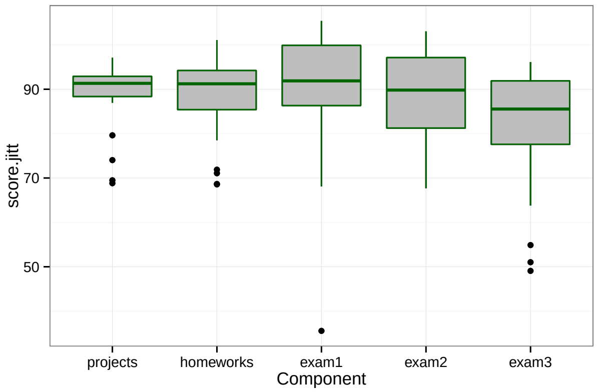
   (c) Normality

   (d) No Influential Points

   (e) Multicollinearity

Provide the appropriate answer to the following questions.

10. In this class, there are several major components which contribute to student's grades: Projects, Homeworks, and Exams. We try and make all of these components equally challenging so that we can get a good idea of a student's ability when we assign final grades. Boxplots of grades for each component from last fall are plotted below. Let's check to see if students performed differently on one component than on the others.



(a) What are the appropriate null and alternative hypotheses?
H_0 :

H_A :

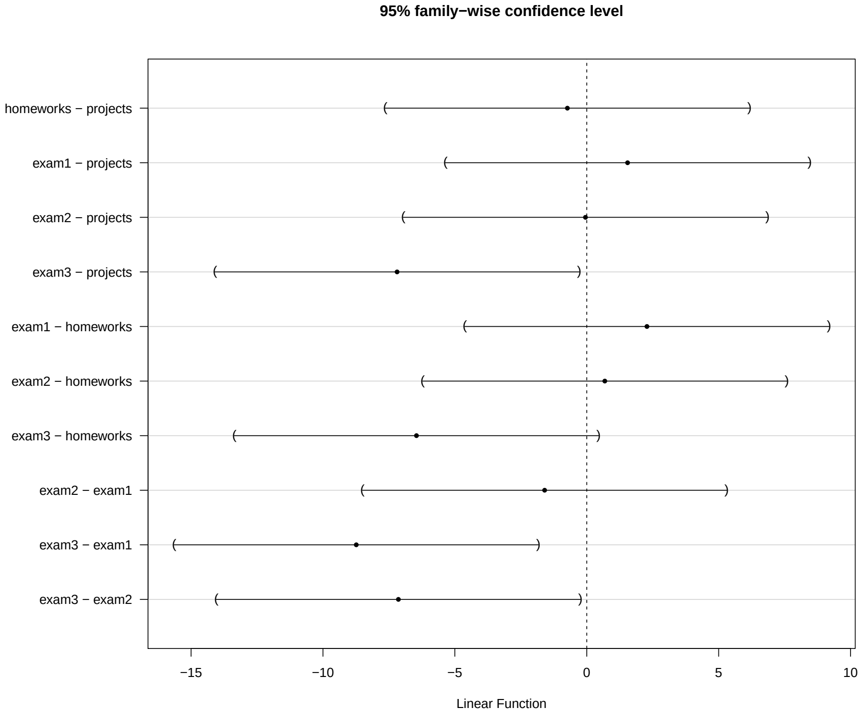
(b) Based on the ANOVA output below, identify the test statistic and state its distribution under the null hypothesis

```
#> Analysis of Variance Table
#>
#> Response: Score
#>            Df Sum Sq Mean Sq F value Pr(>F)
#> Component   4   1623     406     3.7 0.0065
#> Residuals 170  18646     110
```

(c) At the $\alpha = 0.05$ significance level, should we reject or fail to reject the null hypothesis?

(d) Write a conclusion for your decision **in the context of the problem**

(e) Next, let's perform a multiple comparison. Which pairs of components were detected to be different from the others?



11. Let's take a look at the distribution of ages in social networks. We took a random sample of Montanans and asked them their most frequented social network from the list of Reddit, Tumblr, Twitter, Pinterest, Facebook, and LinkedIn. We also recorded the participants age and grouped them by age category (< 24, $25 - 34$, $35 - 44$, > 45).

(a) State the appropriate null and alternative hypotheses of these data. [2pts]
H_0 :

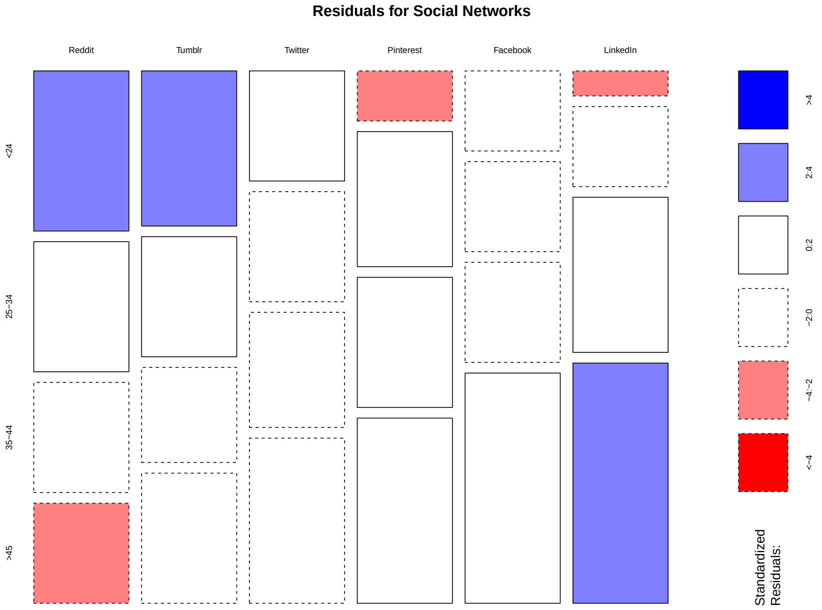
H_A :

(b) Here is the R output from a Chi-Squared test. What is the value of the test statistic? [1 pt]

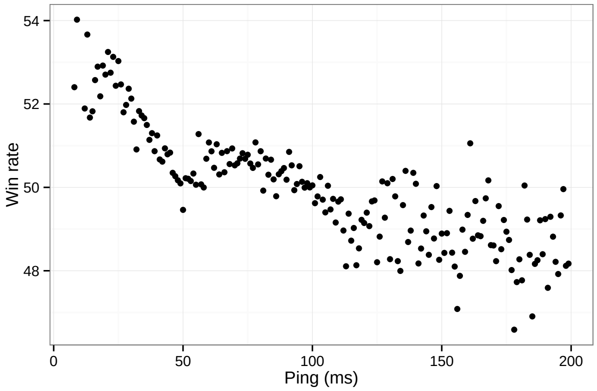
```
Pearson's Chi-squared test

data:  Social Network data
X-squared = 57.1883, df = --, p-value = 7.624e-07
```

(c) What is the distribution of the test statistic under the null hypothesis? [2 pts]

(d) Here is a plot of the residuals. Which networks are preferred by noticeably **more** younger users (i.e. users under 24) than would be expected under the null hypothesis? [2 pts]



12. League of Legends is an online video game. We're going to take a look at the win rate of League of Legends players explained by the ping time (or lag) on their computers. Here is a plot of the relationship for all games with less than 200 ping.



The regression model will be $Rate_i = \beta_0 + \beta_1 * Ping_i + \epsilon_i$. Here is a summary of the fitted model:

```
Call:
lm(formula = Rate ~ Ping, data = lol2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 52.166179   0.115781   450.6   <2e-16
```

```
Ping        -0.021143   0.000981   -21.6   <2e-16

Residual standard error: 0.742 on 188 degrees of freedom
Multiple R-squared:  0.712,  Adjusted R-squared:  0.71
F-statistic:  464 on 1 and 188 DF,  p-value: <2e-16
```

(a) Report the estimated regression model. [2 pts]

(b) Estimate the mean win rate for a person who has has 100 ping. [1 pt]

(c) Interpret the coefficient on ping **in the context of the problem**. [2 pts]

(d) Another variable we could consider is the number of games a user has played (because people with more experience should be better). Will the R^2 (also known as the coefficient of determination and as Multiple R^2) increase or decrease if we add games played to the model? [1 pt]