STAT 217 FINAL EXAM NAME: KEY

Multiple Choice: Circle the SINGLE best answer to each question, 3 pts each.
Show your work for all other questions. You are allowed a calculator (no cell phones) and two two-sided sheets (8.5 by 11) of notes prepared by you. This is not group work. There are 80 total points.

1. For each of the following problems, state whether a One-Way ANOVA (OWA), Two-Way ANOVA (TWA), Chi-Squared Test of Independence (TOI), Chi-Squared Test of Homogeneity (TOH), Simple Linear Regression (SLR) or Multiple Linear Regression (MLR) is most appropriate. [1 pt each]

   (a) TOI Medical researchers followed 6272 Swedish men for 30 years to see if there was any association between the amount of fish in their diet and prostate cancer. They recorded whether the individual was diagnosed with prostate cancer or not and the amount of fish in their diet (never, sometimes, most of the time, all the time).

   (b) SLR Researchers are interested in whether the age of terrestrial impact craters (years) is associated with the diameter of the crater (meters).

   (c) TWA To determine which combination of water (hot, warm, cold) and soap (anti-bacterial, anti-bacterial spray) are most effective on eliminating bacteria (parts per cubic centimeter), researchers randomly assign each of these 6 treatments to 8 people each and record the number of bacteria on their hands after the washing.

   (d) OWA To see whether restricting dietary intake increases life expectancy, female mice were randomly assigned to one of 6 different diet treatments and recorded the lifetime of each mouse (in months).

   (e) MLR A group of animal scientists want to examine if the relationship between an animals average brain weight (grams) and their average body weight (kilograms) is different across 5 different species of animals. To investigate, they collect the brain weight and body weight of 20 animals from each of the 5 species.

2. In multiple linear regression, multicollinearity occurs when

   (a) The response variable is correlated with one of the predictor variables.

   (b) The response variable is not correlated with any of the predictor variables.

   (c) There is a high correlation between the residuals and the fitted values from the model.

   (d) There is a high correlation between the predictor variables.

3. If we add an extra predictor variable in a multiple linear regression model, the coefficient of determination (R^2) will ALWAYS!

   (a) Increase.

   (b) Decrease.

   (c) Stay the same.

   (d) No way to tell before fitting.

4. When doing model selection using R^2_{adj} and AIC, we have learned in this class to pick the model with the _____ R^2_{adj} and the _____ AIC.

   (a) Larger, smaller.

   (b) Larger, larger.

   (c) Smaller, larger.

   (d) Smaller, smaller.

5. Researchers who are interested in the relationship between number of beers drank and blood alcohol content (BAC) perform a randomized experiment to see if increases in alcohol consumption causes a change in BAC. If they construct a 93% prediction interval for a new participant who has drank 3 beers, it will be _____ than a confidence interval for the mean BAC for all participants who have drank 3 beers. 93%

   (a) Narrower.
   (b) Wider.
   (c) The same length.
   (d) Either narrower or wider.

6. In a simple linear regression model between two quantitative variables, the criteria used in this class to determine the estimated intercept and slope is called the

   (a) Coefficient of determination.
   (b) The parametric approach.
   (c) The least-squares criteria.
   (d) The independence method.

7. In multiple linear regression models, which diagnostic plot is used to examine for violations in linearity?

   (a) Residuals versus fitted. Look for curvature!
   (b) Normal QQ.
   (c) Scale-location.
   (d) Residuals versus leverage.

8. Researchers studying the relationship between the size of a house (1000 ft^2) and the cost of the house (1000$) collect this information from a random sample of 1064 houses in Saratoga, New York. The results of the simple linear regression model predicting cost of the home from its size found that slope coefficient was positive and significant at $\alpha = 0.01$. What is the scope of inference for this problem?

   (a) Larger homes are associated with higher costs in the sample taken.
   (b) Homes that were larger caused them to cost more in Saratoga homes.
   (c) Homes that were larger caused them to cost more in the sample taken.
   (d) Larger homes were associated with higher costs in Saratoga homes.

No random assignment of size or cost.

9. To examine which factors influence supermodel salaries the most, a talent agency records the salary (thousands of dollars a year), age(years), years of experience (years), and a beauty score assigned by a professional model photographer for 231 models in their company database. The multiple linear regression model that they will be working with is $log\,(salary) = \beta_0 + \beta_1 Age + \beta_2 Years + \beta_3 Beauty + \epsilon$. The output for the linear morel is shown below.

```
lm(formula = log(salary) ~ age + years + beauty, data = super)
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -11.75598    2.37557  -4.949 1.46e-06
age           1.05379    0.20322   5.185 4.77e-07
years        -1.17822    0.30560  -3.855  0.00015
beauty       -0.01025    0.02194  -0.467  0.64072

Residual standard error: 2.098 on 227 degrees of freedom
Multiple R-squared:  0.1661,  Adjusted R-squared:  0.1551
F-statistic: 15.07 on 3 and 227 DF,  p-value: 5.593e-09
```

(a) What is the estimated regression equation for this model? [2pts]

$\log(\text{salary})_i = -11.75598 + 1.054\,age_i - 1.178\,years_i - 0.0103\,beauty_i$

(b) Interpret the coefficient for years in *the context of the problem.* [3 pts]

-1.17 ; =0.3 → Increasing years of experience by 1 year, we expect a decrease of 70% in mean salary.

1−0.3=0.7 or 70%

(c) What are the null and alternative hypothesis that are being tested in the age line? [2 pts]

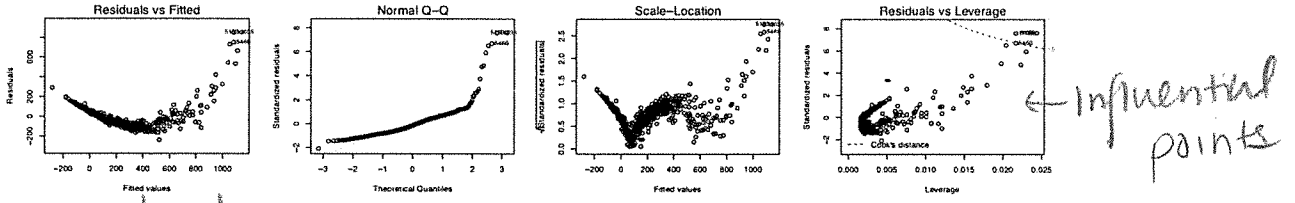
- H_0 : β_1 or $\beta_{age} = 0$, given years and beauty are in the model
- H_A β_1 or $\beta_{age} \neq 0$, " "

(d) What is the conclusion to the hypothesis test in the previous question *in the context of the problem*? [3 pts.]

There is very strong evidence against the null, so we reject it, and conclude that there is a linear relationship between age and salary (after accounting for years of experience and beauty score).

10. In 2005, to investigate, among other things, the relationship between fish length and fish weight in Montana rivers, researchers collected information on the length and weight of fish caught in four Montana rivers. Interest lies in estimating the weight of fish (grams) by using the length (mm). Diagnostic plots for the model with untransformed response and explanatory variables are shown on the top row. The diagnostic plots for the model with a log transformation on the response AND the explanatory variables' are shown on the bottom row.

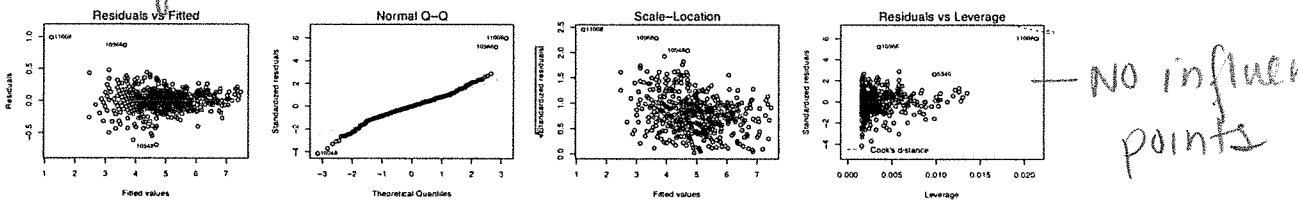
SLR!



← Influential points

curvature!
Increasing variance!

Figure 1: Diagnostic plots for untransformed Model



NO influential points

Figure 2: Diagnostic plots for log-transformed model

(a) The log transformation was appropriate to consider for this example because (**circle all that apply**) [5 pts]

can't have a fish weight of 0!

i. The response variable contained all positive, non-zero values.
ii. There appeared to be a violation of independence in the observations.
iii. Non-constant variance was detected in the residuals versus fitted plot for the original data.
iv. After transformation, there are no influential observations present in the residuals versus leverage plot.
v. The relationship appears to be non-linear between fish length and weight.

(b) Below is the output for the linear regression model predicting log-weight for fish from log-length.

log(weight) = −10.92964 + 2.89904 log(length)

```
lm(formula = log(weight) ~ log(length), data = fish)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -10.92964    0.10411  -105.0   <2e-16
log(length)   2.89904    0.01923   150.7   <2e-16

Residual standard error: 0.166 on 626 degrees of freedom
Multiple R-squared:  0.9732,  Adjusted R-squared:  0.9731
F-statistic: 2.272e+04 on 1 and 626 DF,  p-value: < 2.2e-16
```

Provide the interpretation of b_1 on the **original scale** of the data *in context of the problem.* [4 pts]

For a doubling of fish length, we expect the mean fish weight to increase by 645.92%! (or a multiplicative increase by a factor of 7.459)

100×(7.459−1) = 645.92%

11. Environmentalists in Yellowstone national park are interested in whether or not rattlesnakes in different regions of the park have different associations between snake length and snake weight. To answer this question, the researchers collect the length (cm) and weight (mg) of 15 rattlesnakes each from three different regions (A, B, and C). Since they are interested in whether or not the relationship between length and weight is different in the three regions, they will be working with the following multiple linear regression model:

intercept adjustments

$$weight = \beta_0 + \beta_1 length + \beta_2 RegionB + \beta_3 RegionC + \beta_4 RegionB * length + \beta_5 RegionC * length + \epsilon$$

slope adjustments

(a) The ANOVA table for the fit of this model is displayed below.

```
Analysis of Variance Table

Response: weight
              Df Sum Sq Mean Sq F value    Pr(>F)
length         1  95048   95048 940.909 < 2.2e-16
region         2  50009   25005 247.530 < 2.2e-16
length:region  2   6734    3367  33.332 3.624e-09
Residuals     39   3940     101
```

What are the null and alternative hypothesis being tested in the line `length:region`? [2 pts]

- H_0: A one unit increase in length leads to the same change in weight for rattlesnakes in all regions of the park.
- H_A: A one unit increase in length is different for rattlesnakes in at least one region of the park.

(b) What is the conclusion of this test *in the context of the problem*? [3 pts]

(c) The summary of the fitted model is shown below.

```
lm(formula = weight ~ length * region)

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)    641.84288    7.56326  84.863  < 2e-16
length           1.54522    0.06931  22.296  < 2e-16
regionB          3.78212   11.34095   0.333  0.74055
regionC         11.91760   11.17070   1.067  0.29259
length:regionB   0.42163    0.11162   3.777  0.00053
length:regionC  -0.50516    0.10179  -4.963 1.41e-05

Residual standard error: 10.05 on 39 degrees of freedom
Multiple R-squared:  0.9747,  Adjusted R-squared:  0.9715
F-statistic: 300.5 on 5 and 39 DF,  p-value: < 2.2e-16
```

Based on this output, which region is estimated to have the largest increases in mean snake weight for a 1 cm increase in snake length? [1 pt]

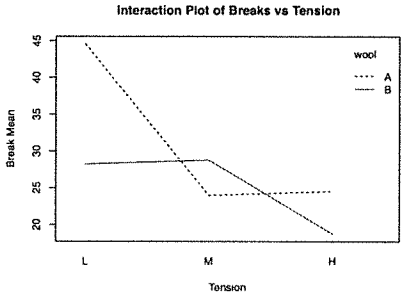
Region B

(d) Write out the estimated simple linear regression equation predicting snake weight from snake length in region C. [2 pts]

$$\widehat{weight} = (641.84288 + 11.91760) + (1.54522 - 0.50516)length$$
$$= 653.76 + 1.04\, length$$

12. A manufacturing company is interested in the breaking strength of yarn for three types of wool (A and B) and three levels of tension (L, M, H). The record the number of breaks for 54 different looms of yarn (8 looms for each of the 6 treatments).

(a) The researchers are first interested in whether or not tension and wool interact with one another, so they fit the two-way interaction ANOVA model: $y_i = \mu + \tau_j + \gamma_k + \omega_{jk} + \epsilon_i$. The interaction plot between tension and wool is shown below.



Reminder: there can be interactions with non-crossing lines!

Based on this plot, does it appear that tension and wool interact? Why or why not? [2 pts]

Yes, there are non-parallel lines!

(b) The resulting ANOVA table is partially filled in below. Fill in the **6** missing values in the table. [3 pts]

```
Analysis of Variance Table

Response: breaks
             Df Sum Sq Mean Sq F value    Pr(>F)
tension       2 2034.3 1017.13  8.498 0.0006926
wool          1 450.67 450.67   3.7653 0.0582130
tension:wool  2 1002.8  501.4   4.1891 0.0210442
Residuals    48 5745.1  119.69
total        53 9232.9
```

$= \frac{1017.13}{119.69}$

$= \frac{1002.8}{2}$

(c) What is the null and alternative hypothesis being tested in the `tension:wool` line? [2 pts]

- H_0: The relationship between type of wool and mean breaking strength is the same for all levels of tension.
- H_A: " " is different for at least one level of tension

(d) Assuming that all assumptions are met, what distribution does the test statistic in the `tension:wool` line follow under the null hypothesis? [2 pts]

$F(2, 48)$

(e) What is the decision for the hypothesis test in part c? [1 pt]

Reject, strong evidence against H_0

OWA!

13. A student performed an experiment with three different grips to see what effect it might have on the distance of backhand frisbee throws. She tried throwing the frisbee 8 times with each of the three grips (Finger out, Inverted grip, Normal) and recorded the distance of the throws (paces). The model that the student is using is the deviations one-way ANOVA model: $distance_i = \mu + \tau_j + \epsilon_i$ to predict the distance of the throw based on which grip was used.

Reference coded model!

(a) The summary of the one-way ANOVA model is displayed below.

```
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)          29.682      1.037  28.611  < 2e-16
gripInverted Grip     3.061      1.467   2.086   0.0494
gripNormal            7.282      1.467   4.963 6.54e-05

Residual standard error: 2.934 on 21 degrees of freedom
Multiple R-squared:  0.5419,  Adjusted R-squared:  0.4983
F-statistic: 12.42 on 2 and 21 DF,  p-value: 0.0002753
```

What is the baseline level for this model? [1 pt]

Finger out

(b) What is the estimated mean paces for throws with Normal grip? [1 pt]

$\hat{\mu}_{Normal} = 29.682 + 7.282 = 36.964$

(c) The ANOVA table for this model is shown below.

```
Analysis of Variance Table

Response: distance
          Df Sum Sq Mean Sq F value    Pr(>F)
grip       2 213.91  106.96  12.422 0.0002753
Residuals 21 180.82    8.61
```

What is the interpretation of the p-value for the `grip` line *in the context of the problem*? [3 pts]

The probability of observing differences in means like we did or more extreme if the true mean distances for each group were the same is 0.02%.

(d) **True or False**: The next step in the analysis would be to perform a multiple comparison procedure. [1 pt]

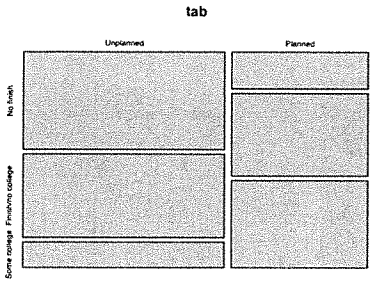
TRUE! We know at least one differs, but which one?

14. A 1954 study of a single random sample of 1438 pregnant women examined whether there is an association between a women's education level (didn't finish highschool, finished highshool but no college, some college) and whether or not her pregnancy was planned (planned, unplanned).

(a) What are the appropriate null and alternative hypothesis? [2 pts]

- H_0 : Pregnant women's education and planned pregnancies are independent.
- H_A : are associated (or dependent).

(b) The appropriate plot of the data is displayed below. The name of this type of plot is a



i. Stacked barchart.
ii. Mosaic plot.
iii. Scatterplot.
iv. Interaction plot.

(c) The expected cell counts are shown below.

	No finish	Finish/no college	Some college
Unplanned	292.62	334.34	202.04
Planned	198.38	226.66	136.96

Based on these expected cell counts, would you choose to use a parametric χ^2 test or a permutation X^2 test for the hypothesis test in part a? [1 pt]

Yes, all expected counts are > 5.

(d) Assuming the assumptions are met, what distribution does the X^2 statistic follow if the null hypothesis is true? [2 pts]

d.f. $= (2-1) \times (3-1) = 2$ a chi-square with 2 d.f.

χ^2_2

(e) The output for the parametric χ^2 statistic is shown below.

```
> chisq.test(pregnancy)
  Pearson's Chi-squared test
data:  pregnancy
X-squared = 204.4735, df = , p-value < 2.2e-16
```

Based on this output, write a conclusion to the hypothesis tested *in the context of the problem.* [3 pts]

There is strong evidence against Ho, therefore we reject it and conclude that pregnant women's education and planned pregnancies are associated.