Statistics 110, Practice Final Exam KEY

1. The scatterplot below shows the regression fit to predict Y = the typical time of a hike in the Adirondack Mountains (in New York) using X = length of the hike (in miles).

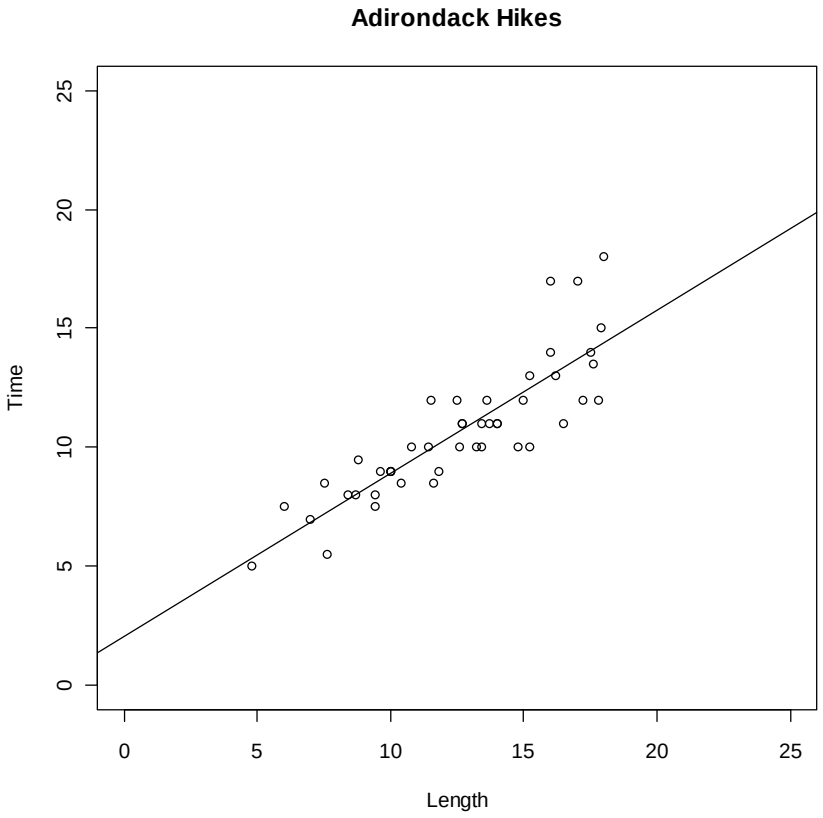
a. Add *three* new data points to this plot that would clearly have the properties listed below. Label them as A, B and C:

Point A: Influential but not an outlier
Point B: Residual near zero & large leverage
Point C: An outlier but not influential

Solution: Points A and B should be on or near the line but removed from the other points in either direction. Point C should be far above or below the line, but in the mid-range of the "length" variable.

b. Could any of the three points you added in Part (a) have been on top of each other? Explain.

Solution: Yes, Points A and B are illustrating the same thing. "Residual near zero" mean it is not an outlier, and "Large leverage" means it is influential.



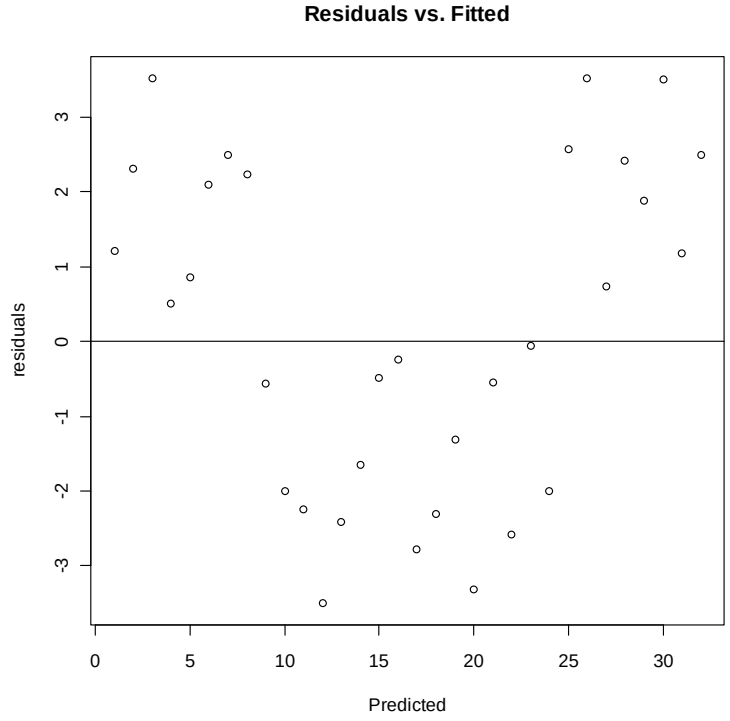
2. Why is it better to use adjusted R^2 rather than R^2 when comparing multiple regression models?

Solution: Adjusted R^2 adjusts for the number of predictors in a model. R^2 will always increase when a new variable is added but adjusted R^2 will not, and could go down. It reflects the trade-off between reducing SSE and decreasing the denominator of MSE, which could make MSE go up.

3. Explain briefly what the Variance Inflation Factor (VIF) tells us about a multiple regression model coefficient.

Solution: The VIF tells us how strongly correlated a predictor is with the other predictors in the model. Specifically, it is a measure of how well that predictor could be estimated if it were the response variable, using the other variables in the model as predictors.

4. The plot below shows residuals versus fitted values after fitting a linear regression model. Of the four conditions of linearity, equal variance, normality and independence, choose two and discuss whether or not they appear to be met based on this plot.



Solution:

Linearity is clearly a problem, as evidenced by the U-shape of the residuals.

Equal variance may be a problem because the points appear to be slightly more spread out at the bottom of the "U shape" than at the sides.

The other two conditions cannot be checked with this plot.

5. A guidebook contains information on the *Distance* (miles), average *Time* (minutes) and *Elevation* gain (feet) for a sample of 72 day hikes in Southern California. Below is some R output from a simple linear regression model to predict the *Elevation* gain using the hike *Time.*

```
Coefficients:
            Estimate  Std. Error t value Pr(>|t|)
Constant     137.24     75.12      1.83    0.072
Time         1.9195     0.625      3.07    0.003
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 336.103 on 70 degrees of freedom
Multiple R-squared: 0.119,    Adjusted R-squared: 0.106
F-statistic: 9.42 on 1 and 70 DF,  p-value: 0.003
```

a. Write a sentence that interprets the *value* of the estimated slope of the model in the context of this data situation.

Solution: The estimated slope is 1.9195, or about 1.9. It tells us that as Time goes up by 1 minute, the *predicted* or *estimated average* Elevation gain goes up by about 1.9 feet.

b. Write a sentence that interprets the value of R^2 (not adjusted) in the context of this data situation.

Solution: We can explain 11.9% of the variability in elevation gain by using time in a regression model.

c. What would you conclude about the usefulness of using hike time to predict elevation gain? Give statistical justification for your answer.

Solution: This is the equivalent of testing H_0: $\beta_1 = 0$. You can use either the t-test ($t = 3.07$, $p = 0.003$) or the F test ($F = 9.42$, $p = 0.003$). In either case, the p-value is 0.003, so we reject the null hypothesis. Therefore, hike time is useful as a predictor of elevation gain.

6. When examining case diagnostics in multiple regression, under what circumstance is it acceptable to remove a case that is clearly a Y outlier?

Solution: It is only acceptable to remove a Y outlier if it is clearly a mistake.

7. Give two circumstances in which it is acceptable to remove one or more cases that are outliers in the X variables.

Solution:

1. *One of more of the X values for the case is clearly a mistake.*

2. *If several cases have X values different from the rest of the data, and the prediction does not work well for cases with those combinations of X value, then it's likely that the cases belong to a different population and the relationship between Y and the predictor variables is not the same as it is for the main population. The cases can be removed, but the model would not apply in the future for predicting Y when X has values in that area.*

8. Suppose you have four possible predictor variables (X_1, X_2, X_3, and X_4) that could be used in a regression analysis. You run a forward selection procedure, and the variables are entered as follows:
Step 1: X_2 Step 2: X_4 Step 3: X_1 Step 4: X_3
In other words, after Step 1, the model is $Y = \beta_0 + \beta_1 X_2 + \varepsilon$
After Step 2, the model is $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \varepsilon$
And so on…
You also run an all subsets regression analysis using R^2 as the criterion for the "best" model for each possible number of predictors (1, 2, 3, 4). Would the same models result from this analysis as from the

forward stepwise procedure? In other words, would "all subsets regression" definitely identify the following as the best models for 1, 2, 3, and 4 variables? Circle Yes or No in each case.

a. β_0 + 1 variable, best model would be $Y = \beta_0 + \beta_1 X_2 + \varepsilon$ ***YES***

b. β_0 + 2 variables, best model would be $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \varepsilon$ ***NO***

c. β_0 + 3 variables, best model would be $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_1 + \varepsilon$ ***NO***

d. β_0 + 4 variables, best model would be $Y = \beta_0 + \beta_1 X_2 + \beta_2 X_4 + \beta_3 X_1 + \beta_4 X_3 + \varepsilon$ ***YES***

9. An international company is worried that employees in a certain job at its headquarters in Country A are not being given raises at the same rate as employees in the same job at its headquarters in Country B. Using a random sample of employees from each country, a regression model is fit with:

Y = employee salary
X_1 = length of time employee has worked for the company
X_2 = 1 if employee is in Country A, and 0 if employee is in Country B.

New employees, who have $X_1 = 0$, all start at the same salary, so the company is not interested in fitting a model with different intercepts, only with different slopes.

a. Write the full and reduced models for determining whether or not the slopes are different for employees in the two countries, using the variable definitions above and standard notation.

Solution:
Full model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_1 X_2 + \varepsilon$

Reduced model: $Y = \beta_0 + \beta_1 X_1 + \varepsilon$

b. For the full model, write the population model for an employee with 10 years of experience in Country A, and then write the model for an employee with 12 years of experience in Country B.

Solution: $Y = \beta_0 + \beta_1 (10) + \beta_2 (10) + \varepsilon$ and $Y = \beta_0 + \beta_1 (12) + \varepsilon$

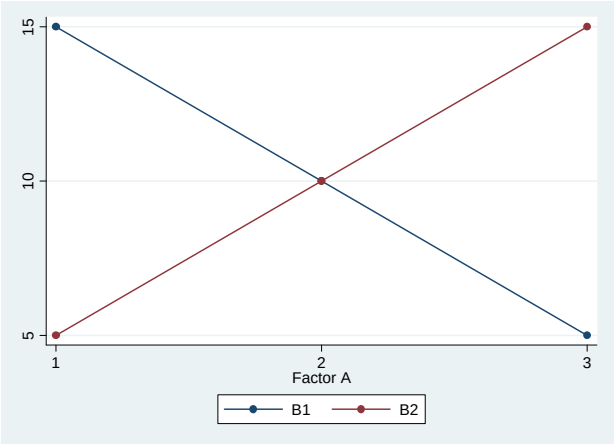
10. Consider a two-factor experiment in which one factor is "Restaurants" and five restaurants are used in the study. Give a set of circumstances under which the restaurant factor would be considered fixed, and a set of circumstances under which it would be considered random.

Solution: The factor would be considered fixed if those 5 restaurants were of interest. It would be considered random if the 5 restaurants were randomly chosen to represent a much larger collection of restaurants. For example, a company might own hundreds of restaurants, but want to try a new menu or a new method of greeting customers, etc., at a small set of them to see what works.

11. Sketch a picture of possible cell means (i.e., an interaction plot) for the following scenarios:

a. Factor A has 3 levels, Factor B has 2 levels. There is an AB interaction, but no A or B main effects.

Solution: Here's one possibility*:*



b. Factor A has 3 levels, Factor B has 2 levels. There is an effect for Factor A, but no interaction and no Factor B effect.

***Solution*:** Here is one possibility (the lines are right on top of each other for B1 and B2):

12. A study was done to see if meditation would reduce blood pressure in patients with high blood pressure. There were 100 people available for the study. Half of the patients were randomly chosen to learn meditation and told to practice it for half an hour a day. The other half was told not to alter their regular daily routine. Blood pressure measurements were taken at the beginning of the study, after 5 weeks, and after 10 weeks.

a. Is this a randomized block design? Explain why or why not.

Solution: Yes. The individuals are blocks. The reason for using them is to reduce a *known* source of variability, which is that individuals have very different blood pressure.

b. One factor in this experiment is "Meditation group." How many levels does that factor have, and what are the levels?

Solution: It has two levels – Meditation and no meditation.

c. Another factor in this experiment is "Time period." How many levels does that factor have, and what are the levels?

Solution: It has 3 levels – beginning of the study, after 5 weeks, and after 10 weeks.

d. Are the factors "Meditation group" and "Time period" crossed, or is one of them nested under the other? Explain.

Solution: They are crossed. Both groups are measured at all 3 time periods.

13. Comment briefly on the following statements:

a. In one factor ANOVA where the factor is fixed, a highly significant F statistic ($p < .001$) indicates that the K population means, μ_1 to μ_K are all different.

Solution: The statement is not true. The test only tells us that at least one of the means differs from the others.

b. In one factor ANOVA, we need to use multiple comparisons (like the Tukey procedure) because it is impossible to compare K means all at once.

Solution: Not quite right. We use multiple comparisons to control the overall chance of making an erroneous conclusion, over all of the tests or confidence intervals examined.

14. A student analyzed data for a one-way analysis of variance situation for which there were 3 levels of the factor, and 21 people measured at each level. Unfortunately, after running the analysis, the student lost the computer output. She said "All I remember is that one of the mean squares was 100 and the other one was 500, but I can't remember which was which. Oh, and I remember that the p-value for the test was about .01." Based on this information, can you construct the analysis of variance table? (I've provided headings to remind you of the table structure.) If so, fill it in. If not, explain why not. If you think you can partially fill it in, do that.

Solution:

Source	SS	df	MS	F	p-value
Model	*1000*	*3 – 1 = 2*	*500*	*5*	*.01*
Error	*6000*	*3(21 – 1) = 60*	*100*		
Total	*7000*	*62*			

Note that the larger MS must be for the model and not for error, because an F statistic less than one would not lead to a p-value of .01 in any case.