STATISTICS 608 Linear Models -EXAM I
February 18, 2014

Student's Name: _____

Student's Email Address: _____

INSTRUCTIONS FOR STUDENTS:

1. There are **9** pages including this cover page.
2. You have exactly 75 minutes to complete the exam.
3. There may be more than one correct answer; choose the best answer.
4. You will not be penalized for submitting too much detail in your answers, but you may be penalized for not providing enough detail.
5. Do not discuss or provide any information to any one concerning any of the questions on this exam or your solutions until I post the solutions next week.
6. You may use one 8.5" X 11" sheet of notes and a calculator.
7. At the end of the exam, leave your sheet of notes with your proctor along with the exam.

I attest that I spent no more than 75 minutes to complete the exam. I used only the materials described above. I did not receive assistance from anyone during the taking of this exam.

Student's Signature: _____

INSTRUCTIONS FOR PROCTOR:

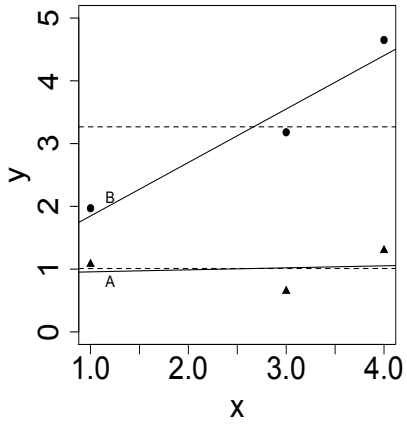
Immediately after the student completes the exam scan it to a pdf file and have student upload to Webassign.

1. I certify that the time at which the student started the exam was _____ and the time at which the student completed the exam was _____.
2. I certify that the student has followed all the **INSTRUCTIONS FOR STUDENTS** listed above.
3. I certify that the exam was scanned in to a pdf and uploaded to Webassign in my presence.
4. I certify that the student has left the exam and sheet of notes with me, to be returned to the student no less than one week after the exam or shredded.

Proctor's Signature: _____

Part I: Multiple choice

1. Below is shown a plot of two generated data sets. The least squares regression lines are plotted on the graph as solid lines; the value of $\bar{y}$ is also plotted as a dotted line for each data set. The values of x are the same for both data sets. The value of $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ is also the same for both data sets.



Which of the following is true?

(a) R^2 for data set (A) $>$ R^2 for data set (B).

(b) **R^2 for data set (A) $<$ R^2 for data set (B). Notice that Regression SS is much larger for data set (B).

(c) R^2 for data set (A) $=$ R^2 for data set (B).

(d) There is not enough information to determine which R^2 is the largest.

2. In a simple linear regression setting with the usual assumptions met, which of the following is true about a confidence interval for the mean value of y at a given value of the explanatory variable x^* vs. a prediction interval for an individual at a given value of the explanatory variable x^*?

(a) The confidence interval always requires normality of the residuals, while the prediction interval can sometimes rely on the Central Limit Theorem.

(b) The confidence interval always requires large sample sizes, while the prediction interval can be calculated for smaller n.

(c) The prediction interval and confidence interval are usually the same width.

(d) **The prediction interval is usually wider than the confidence interval.

3. Suppose we fit a simple linear regression model $y = \beta_0 + \beta_1 x + e$ to a data set. Assume all usual assumptions are met for the model. Which of the following is true about $R^2 = SSM/SST$?

   (a) **The higher the value of R^2, the smaller the residuals will be in absolute value, on average.

   (b) The higher the value of R^2, the more random the residuals are, on average.

   (c) The higher the value of R^2, the more valid the model is. Remember my rant about R^2.

   (d) The higher the value of R^2, the more normally distributed the residuals are.

   (e) All of the above are true.

4. For which of the following is an inverse response plot used?

   (a) Determining whether a polynomial model should be fit.

   (b) Determining an appropriate transformation for the explanatory variable.

   (c) **Determining an appropriate transformation for the response variable.

   (d) Determining a transformation that will stabilize variance in the response variable.Partial credit. Several people chose this; stabilizing variance certainly may also happen as a consequence of transforming variables.

   (e) Determining whether the residuals have constant variance.

5. In a linear regression model predicting y = the price of beef in cents per pound in 1930, a 95% prediction interval when x = consumption was 48 pounds per person was found to be (64.1, 74.1) cents per pound. How should we interpret this prediction interval in context? Assume all model assumptions are met.

   (a) We can be 95% confident the average price of beef when consumption was 48 pounds per person was between 64.1 and 74.1 cents per pound.

   (b) ** We expect about 95% of the observed prices for beef when consumption was 48 pounds per person to be between 64.1 and 74.1 cents per pound. Remember that prediction intervals are for individuals; confidence intervals are for parameters.

   (c) In repeated sampling, 95% of prediction intervals will capture the true population slope. This is one such interval.

   (d) In repeated sampling, 95% of the time the population mean price of beef when consumption is 48 pounds per person will fall within the calculated interval. This is one such interval.

   (e) In repeated sampling, 95% of the time the population mean price of beef when consumption is 48 pounds per person will be between 64.1 and 74.1 cents per pound.

Part II: Short Answer

6. Show that the variance stabilizing transformation for a response variable Y with mean μ and variance $1/\mu^2$ is $f(Y) = Y^2$. The variance of the transformed Y is not equal to 1; is that relevant? Why or why not?

$$\begin{aligned} \text{Var}(f(Y)) &\approx [f'(\mu)]^2 \, \text{Var}(Y) \\ &= (2\mu)^2 (1/\mu^2) \\ &= 4 \end{aligned}$$

No, it's not relevant that 4 is not equal to 1; all that matters is that the resulting variance does not involve the mean, μ, so that variance does not increase when the mean increases or decreases.

7. Define a linear regression model written in matrix notation as $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{Y}$ and $\mathbf{X}$ are the $(n \times 1)$ random response vector and the $(n \times 2)$ design matrix, respectively. The first column of $\mathbf{X}$ is the $\mathbf{1}$ vector. Also define the (2×1) vector of unknown regression parameters as $\boldsymbol{\beta} = (\beta_0 \beta_1)'$. Finally, the $(n \times 1)$ vector $\mathbf{e}$ is the vector of random errors.

   Assume $\text{Var}(\mathbf{e}) = \sigma^2 \mathbf{I}$ and $E[\mathbf{e}] = \mathbf{0}$, a vector of 0's. The vector of residuals is given by $\hat{\mathbf{e}} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, where $\mathbf{H} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$.

   (a) Show that $E[\hat{\mathbf{e}}|\mathbf{X}] = 0$.

$$\begin{aligned} E[\hat{\mathbf{e}}|\mathbf{X}] &= E[(\mathbf{I} - \mathbf{H})\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{I} - \mathbf{H})E[\mathbf{Y}|\mathbf{X}] \\ &= (\mathbf{I} - \mathbf{H})E[\mathbf{X}\boldsymbol{\beta} + \mathbf{e}|\mathbf{X}] \\ &= (\mathbf{I} - \mathbf{H})(\mathbf{X}\boldsymbol{\beta} + \mathbf{0}) \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{X}\boldsymbol{\beta} - \mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{0} \end{aligned}$$

(b) Show that $\text{Var}(\hat{\mathbf{e}}|\mathbf{X}) = \sigma^2(\mathbf{I} - \mathbf{H})$.

$$\begin{aligned}
\text{Var}(\hat{\mathbf{e}}|\mathbf{X}) &= \text{Var}((\mathbf{I} - \mathbf{H})\mathbf{Y}|\mathbf{X}) \\
&= (\mathbf{I} - \mathbf{H})\text{Var}(\mathbf{Y}|\mathbf{X})(\mathbf{I} - \mathbf{H})' \\
&= (\mathbf{I} - \mathbf{H})\text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{e}|\mathbf{X})(\mathbf{I} - \mathbf{H})' \\
&= (\mathbf{I} - \mathbf{H})\sigma^2\mathbf{I}(\mathbf{I} - \mathbf{H})' \\
&= \sigma^2(\mathbf{I} - \mathbf{H})(\mathbf{I}' - \mathbf{H}') \\
&= \sigma^2(\mathbf{I} - \mathbf{H}' - \mathbf{H} + \mathbf{H}\mathbf{H}') \\
&= \sigma^2(\mathbf{I} - \mathbf{H} - \mathbf{H} + \mathbf{H}) \\
&= \sigma^2(\mathbf{I} - \mathbf{H})
\end{aligned}$$

We know $\mathbf{H}' = \mathbf{H}$ and $\mathbf{H}\mathbf{H} = \mathbf{H}$ from homework.

(c) A student has written that:

$$\begin{aligned}
\mathbf{H} &= \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}' \\
&= \mathbf{X}(\mathbf{X}^{-1}\mathbf{X}'^{-1})\mathbf{X}' \\
&= \mathbf{X}\mathbf{X}^{-1}\mathbf{X}'^{-1}\mathbf{X}' \\
&= \mathbf{I}\mathbf{I} = \mathbf{I}.
\end{aligned}$$

But the hat matrix is not the identity matrix. What is the student's mistake here?
The student's mistake is that $\mathbf{X}$ is not square, so is not invertible.

Part III: Long Answer

8. The Wall Street Journal published data on spending on TV advertising in millions of dollars for 21 companies and the retained impressions of adults from a survey. (Participants were asked to cite a commercial they had seen for that product category in the past week. The values are millions of impressions per week.)

   (a) The first model fit to the data was $Impression = \beta_0 + \beta_1 Spend + e$ (Model 1). Output from this first model is found at the end of the exam. If the model were valid, how would we interpret the slope estimate, $\hat{\beta}_1 = 0.36$?
   If the amount spent on advertising increases by \$1 million, the model predicts that the number of impressions will increase by 0.36 million, on average.

   Notice that we don't say that anything else is held constant; that's the magic of multivariate models, to hold other variables in the model constant.

   (b) Is this first model indeed valid? List any apparent weaknesses in this first model.
   No, the first model is not valid. We have the following problems:

      i. The scatterplot of the residuals does not look like a uniformly distributed scatter of points; instead there is a distinct pattern.

      ii. The residuals do not appear to be normally distributed, according to the Q-Q plot.

      iii. The variability of the residuals does not appear to be constant, as seen on the Scale-Location plot.

      iv. There does not appear to be a linear relationship between Spend and Impressions.

      v. There appears to be a "bad" leverage point. Ford has a much higher value of Cook's distance than the other points, meaning it is both a leverage point and an outlier. Its influence on the location of the least squares regression line is much higher than the other points' influence.

      vi. The companies weren't necessarily randomly selected, so p-values based on that assumption are invalid. It is unclear what population we may make inferences to using this model.

(c) A coworker suggests deleting Ford from the model (the point that is the outlier on the Cook's Distance by Spend plot, and point 10 on the Standardized residuals by Leverage plot). Comment on the appropriateness of deleting an observation just because it does not fit the model.

An outlier may belong to the population we are interested in. If this is the case, it may indicate that we need a different model to better predict the behavior of the population. What may be only a single point in our data set could be representative of many individuals in the population with similar behavior.

(d) After transforming the predictor variable *Spend* using a 1/5 power transformation, we decide to attempt a transformation of the response variable using a Box-Cox transformation. Output is shown below. Based on this output, what transformation would you suggest for the response variable? Explain your reasoning.

```
     Est.Power Std.Err. Wald Lower Bound Wald Upper Bound
Y1     0.2436   0.2341          -0.2152           0.7024

Likelihood ratio tests about transformation parameters
                            LRT df        pval
LR test, lambda = (0) 1.131400  1 0.287476895
LR test, lambda = (1) 8.625375  1 0.003315121
```

Because 0.25 is a close fraction to 0.2436, and is also found in the confidence interval, this would be a reasonable transformation to try. The fraction $1/5 = 0.2$ is also in the confidence interval, as is $1/3$. $1/2$ is also in the interval, but isn't necessarily going to stabilize variance, as the number of impressions is actually not Poisson. The value for lambda $= 0$ is also an acceptable transformation, as indicated by the p-value, suggesting a log transformation.

(e) Regardless of your answer above, the second model fit to the data was $\log(Impression) = \beta_0 + \beta_1 \log(Spend) + e$. Output from Model 2 is shown below. Use the output to test whether there is a relationship between advertising spending and advertising impressions. Be sure to state your hypotheses clearly, give the appropriate statistics, and state your conclusion in context. Assume all relevant assumptions are met; simply state what they are.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.2999     0.4236   3.069  0.00632 **
x2            0.6135     0.1191   5.153 5.66e-05 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.581 on 19 degrees of freedom
Multiple R-squared:  0.5829,Adjusted R-squared:  0.561
F-statistic: 26.55 on 1 and 19 DF,  p-value: 5.655e-05
```

I. Hypotheses:
$H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$

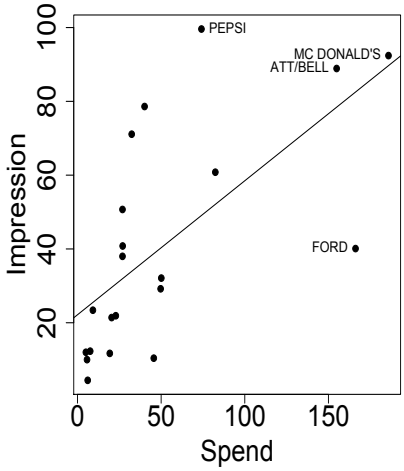
II. Assumptions: Assume met

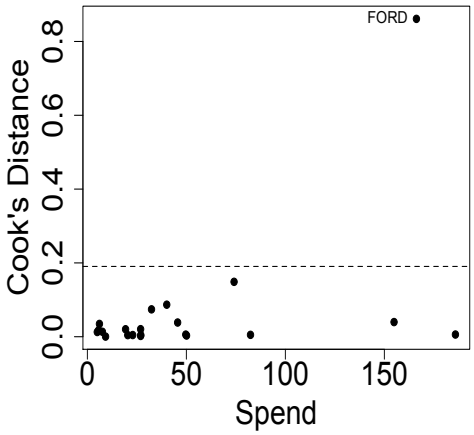
III. Test statistic and p-value:
$t = \frac{0.6135-0}{0.1191} = 5.153$, p-value $= 5.66 \times 10^{-5} \approx 0$

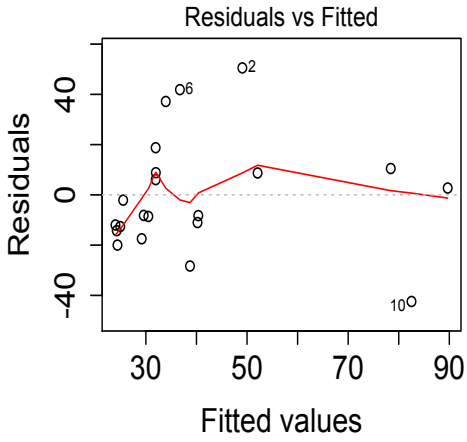
IV. Conclusion:
Because our p-value is so small, we have evidence that there is a relationship between advertising and spending.

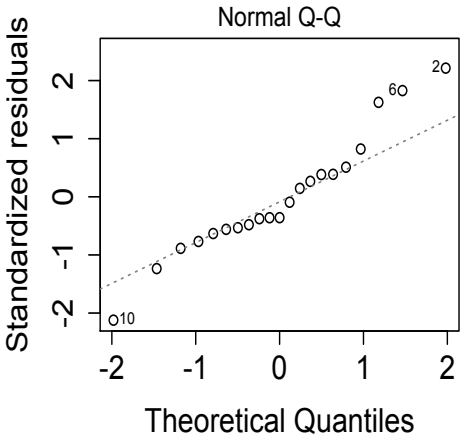
Note: we can make a one-sided conclusion from a two-sided hypothesis (but not the other way around) and say that we have evidence that as spending on advertising increases, the number of impressions increases as well. Also note we shouldn't say we have evidence that the amount of evidence is **equal** to the sample slope. If we wanted to make an inference about the amount of increase, we should use a confidence interval.

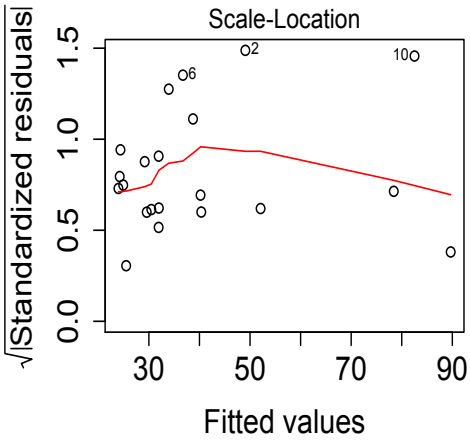
Model 1

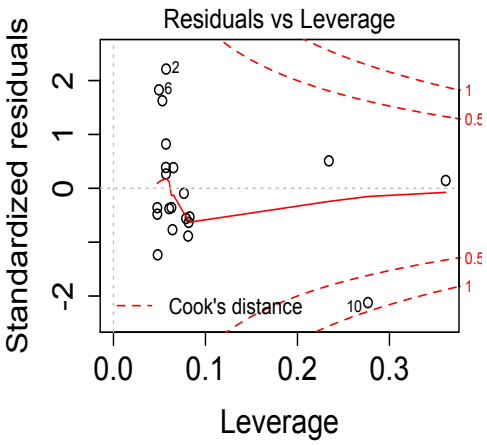

PEPSI
MC DONALD'S
ATT/BELL
FORD
Impression
Spend


FORD
Cook's Distance
Spend


Residuals vs Fitted
Residuals
Fitted values


Normal Q-Q
Standardized residuals
Theoretical Quantiles


Scale-Location
√|Standardized residuals|
Fitted values


Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage