STAT 608 - Exam I

Feb. 21, 2022

Student's Name: ____

INSTRUCTIONS:

1. There are **6** pages including this cover page.
2. You have exactly 50 minutes to complete the exam.
3. Complete the exam on this form.
4. You will not be penalized for providing too much detail in your answers, but you may be penalized for not providing enough detail.
5. You may use **one** 8.5" $\times$ 11" sheet of notes and a calculator.
6. You may choose not to scan the appendix if you make no notes on it.
7. Do not discuss or provide any information to anyone concerning any of the questions on this exam or your solutions until I post the solutions next week.

I attest that I spent no more than 50 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: ____

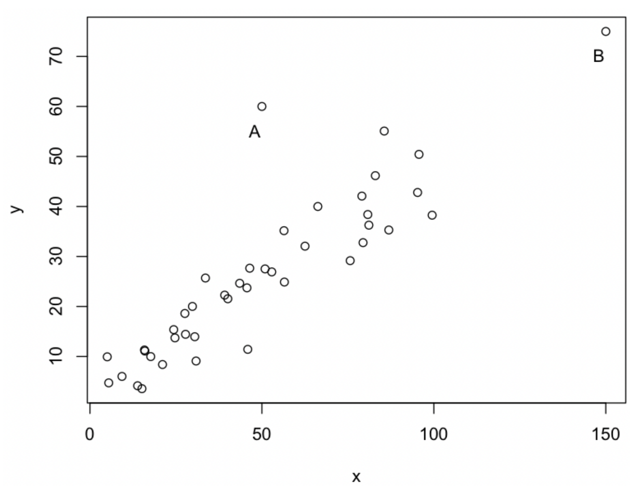
PART I: Multiple Choice (5 Points Per Question). Choose the **best** answer.

1. Which of the following is a linear model?
   (a) $E(Y|X=x) = (\beta_0 + \beta_1 x)^2$
   (b) $E(Y|X=x) = \beta_0 + \beta_1 x + \beta_2 x^2$
   (c) $E(Y|X=x) = \sqrt{\beta_0 + \beta_1 x + \beta_2 x^2}$
   (d) $E(Y|X=x) = 1/(\beta_0 + \beta_1 x)$

2. Which of the following must be true in order for a simple linear regression model to be valid?
   (a) The data must be collected across time.
   (b) The errors must have constant variance.
   (c) The errors must be normally distributed.
   (d) The relationship between x and y must be exponential.
   (e) The relationship between x and y must be quadratic.

3. Consider a dataset with response y and predictor variable x. Suppose we fit the model $\sqrt{Y} = \beta_0 + \beta_1 x + e$. A 95% confidence interval for the mean response with $x = 10$ is $(1.25, 4.15)$. The MSE of the model was 0.05. How should the confidence interval be back-transformed?
   (a) $\left(e^{1.25+0.05/2}, e^{4.15+0.05/2}\right)$
   (b) $\left(e^{1.25-0.05}, e^{4.15+0.05}\right)$
   (c) $\left(1.25^2 + 0.05, 4.15^2 + 0.05\right)$
   (d) $\left(1.25^2 - 0.05, 4.15^2 + 0.05\right)$
   (e) $\left(\frac{1}{1.25}\left(1 + \frac{0.05}{1.25}\right), \frac{1}{4.15}\left(1 + \frac{0.05}{4.15}\right)\right)$

4. The points labeled “A” and “B” in the graph below can be described using which of the following?



(a) Point A is a bad leverage point.
(b) Point A is a good leverage point.
(c) Point B is a bad leverage point.
(d) Point B is a good leverage point.

5. Which of the following is **not** a reason for transforming predictor and / or response variables?

   (a) To ensure the relationship between the predictor and response is a straight line.
   (b) To transform outliers into leverage points.
   (c) To stabilize the variance of the residuals.
   (d) To reduce the influence of outliers.
   (e) To estimate percentage effects (elasticity).

6. What should be done with an outlier?

   (a) Always remove outliers.
   (b) Only remove an outlier if there is a plausible way in which it is different from the rest of the observations.
   (c) Remove an outlier if including it gives unexpected results.
   (d) Only remove an outlier if it is also a good leverage point.

7. Suppose we have a large sample size and fit a simple linear regression model. Residuals are required to be Normally distributed for which of the following?

   (a) For prediction intervals for an individual response.
   (b) For prediction intervals for a mean response.
   (c) For confidence intervals for an individual response.
   (d) For confidence intervals for a mean response.

PART II: Short Answer (8 Points Each Part)

8. An advertising executive is interested in the number of clicks on his company's website within one hour of running television ads. He randomly chose four prime-time one-hour slots and ran either 1, 2, 3, or 4 ads in each hour. There were four responses measured, y_1, y_2, y_3, y_4, depending on the number of ads run.

   (a) For the model $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{e}$, where $\mathbf{y}$, $\boldsymbol{\beta}$, and $\mathbf{e}$ are vectors and $\mathbf{X}$ is the design matrix, write out $\mathbf{X}$ and $\boldsymbol{\beta}$. Assume there is both an intercept β_0 and a slope β_1.

(b) Interpret the model coefficients in the context of the problem (don't just say one is the intercept and the other is the slope).

(c) Use the design matrix above and the general least squares solution for the parameter estimate vector to calculate estimates for the parameters.

9. Recall the 2016 / 2020 presidential election data that we saw in class. The response variable is the percentage of Trump voters in each of the 50 states in 2020, while the explanatory variable is the percentage of Trump voters in each state in 2016. The model summary from `R` is shown below:

```
Call:
lm(formula = Trump_2020 ~ Trump_2016)

Residuals:
      Min        1Q    Median        3Q       Max
-0.159344 -0.007675  0.008245  0.019979  0.125834

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.05547    0.03605   1.539     0.13
Trump_2016   0.88598    0.07180  12.340   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.05136 on 48 degrees of freedom
Multiple R-squared:  0.7603,Adjusted R-squared:  0.7553
F-statistic: 152.3 on 1 and 48 DF,  p-value: < 2.2e-16
```

For use below, here are a few critical values:

$z_{0.025}$	$t_{0.025,45}$	$t_{0.025,46}$	$t_{0.025,47}$	$t_{0.025,48}$	$t_{0.025,49}$	$t_{0.025,50}$
1.9600	2.0141	2.0129	2.0117	2.0106	2.0096	2.0086

A few more pieces of information that you may need:

- `mean(Trump_2016) = 0.4918`
- `mean(Trump_2020) = 0.4912`
- `sum((Trump_2016 - mean(Trump_2016)) ^ 2) = 0.5117`
- `sum((Trump_2020 - mean(Trump_2020)) ^ 2) = 0.5283`

(a) Compute a 95% confidence interval for the mean percentage voting Trump in 2020 for states that voted 60% Trump in 2016.