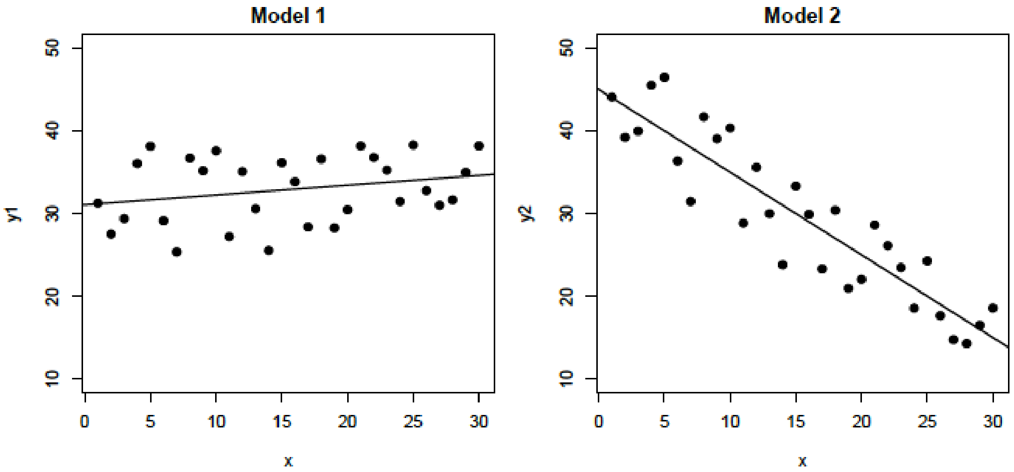
PART I: Multiple Choice (5 Points Per Question). Choose the **best** answer.

1. A statistician is comparing two proposed models. The same predictor x is being used for both response variables, y_1 and y_2. Residual sum of squares is the same for both models. The plots are drawn to scale on the same axes. Which of the following statements is true?



   (a) SSReg for Model 1 is greater than SSReg for Model 2.

   (b) **SSReg for Model 2 is greater than SSReg for Model 1. (NOTE THE DIFFERENCE BETWEEN $\hat{y}_i$ AND $\bar{y}$ IS MUCH LARGER FOR MOST POINTS ON THE SECOND GRAPH.)

   (c) The values of SSReg for both models are equal.

   (d) We know R^2 is higher for Model 1, but we can't determine whether SSReg is higher or lower for Model 2.

   (e) We know R^2 is higher for Model 2, but we can't determine whether SSReg is higher or lower for Model 1.

2. Which of the following must be true in order for a simple linear regression model to be valid?

   (a) The relationship between x and y must be exponential.

   (b) The relationship between x and y must be quadratic.

   (c) The data must be collected across time.

   (d) **The errors must have constant variance.

   (e) The errors must be normally distributed. (THIS IS NECESSARY ONLY FOR SMALL SAMPLE SIZES AND PREDICTION INTERVALS.)

3. In the Advertisements dataset from class, we looked at the relationship between x = Circulation (in millions) and y = Ad Revenue (in thousands of dollars) of various magazines. The model $\log(\text{AdRevenue}_i) = \beta_0 + \beta_1 \log(\text{Circulation}_i) + e_i$ was fit to the data. The magazine "Family Circle" had a circulation of 3.954, and a confidence interval from the output for that point was $(5.345, 5.458)$. The MSE from the model was 0.0313. How should the confidence interval be back-transformed?

(a) **$\left(e^{5.345+0.0313/2}, e^{5.458+0.0313/2}\right)$ (Remember that the inverse transformation for log is the exponential, so pick one of the first two at the very least.)

(b) $\left(e^{5.345-0.0313}, e^{5.458+0.0313}\right)$

(c) $\left(5.345^2+0.0313, 5.458^2+0.0313\right)$

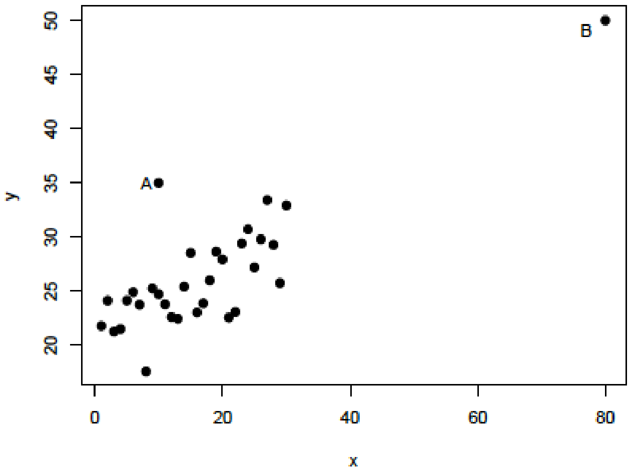
(d) $\left(5.345^2-0.0313, 5.458^2+0.0313\right)$

(e) $\left(\frac{1}{5.345}\left(1+\frac{0.0313}{5.345^2}\right), \frac{1}{5.458}\left(1+\frac{0.0313}{5.458^2}\right)\right)$

(f) $\left(\frac{1}{5.345}\left(1-\frac{0.0313}{5.345^2}\right), \frac{1}{5.458}\left(1+\frac{0.0313}{5.458^2}\right)\right)$

PART II: Multiple Select (2 Points Per Choice). Please circle the letter for **all** of the correct answers; more than one answer may be correct.

4. The points labeled "A" and "B" in the graph below can be described using which of the following? Select all that apply.



(a) Point A is a high leverage point. (Point A is quite close to the mean of x.)

(b) Point A is influential. (Point A isn't going to change the slope substantially. It might change the intercept a bit, but not the slope.)

(c) **Point B is a high leverage point. (Point B is definitely far from the mean of x.)

(d) Point B is influential. (Point B won't change the slope substantially.)

5. Which of the following are reasons for transforming predictor and / or response variables? Select all that apply.

(a) **To ensure the relationship between the predictor and response is a straight line.

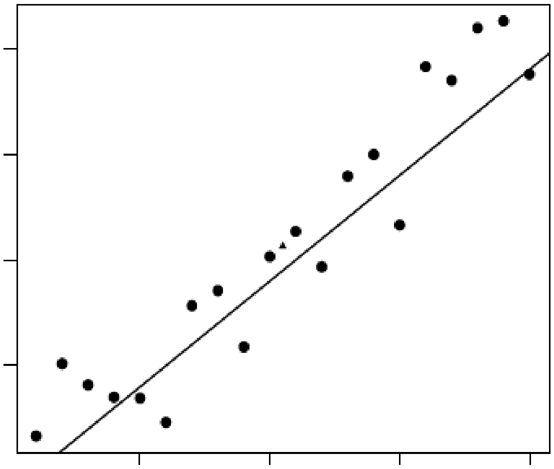
(b) **To stabilize the variance of the residuals.

(c) **To reduce the influence of outliers.

(d) **To estimate percentage effects (elasticity).

PART III: Short Answer (8 Points Each Part)

6. A model is being fit to some data as shown below; the plot is drawn to scale. The triangle-shaped point in the middle is the point $(\bar{x}, \bar{y})$. Explain to the researcher why the proposed line on the graph is not the least squares regression line.



Recall that the least squares equations resulted in a solution for the intercept as $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$, which can be rearranged as $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$. But if we notice that the equation for the least squares regression line is $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$, we see that the predicted value of y at the value $\bar{x}$ is the sample mean of y. Thus, the least squares regression line must go through the point $(\bar{x}, \bar{y})$.

7. A researcher is interested in the effect of different fertilizers on the amount of corn produced. As a preliminary study, **one** gram of Fertilizer A is added to the first two pots with corn plants, and **two** grams of Fertilizer B are added to the second two pots, for a total of four pots. (Each pot has one corn plant.) The response variable is the weight of the corn produced from the plants. Let β_A and β_B be the mean amount of corn produced per gram of Fertilizer A and Fertilizer B, respectively.

   (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 $\boldsymbol{\beta}$.

$$\mathbf{X} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 2 \\ 0 & 2 \end{bmatrix}, \; \boldsymbol{\beta} = \begin{bmatrix} \beta_A \\ \beta_B \end{bmatrix}$$

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

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 2 \\ 0 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 \\ 0 & 8 \end{bmatrix}$$

$$\left(\mathbf{X}'\mathbf{X}\right)^{-1} = \begin{bmatrix} 0.5 & 0 \\ 0 & 0.125 \end{bmatrix}$$

$$\mathbf{X}'\mathbf{y} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} y_1 + y_2 \\ 2y_3 + 2y_4 \end{bmatrix}$$

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{y} = \begin{bmatrix} 0.5\left(y_1 + y_2\right) \\ 0.25\left(y_3 + y_4\right) \end{bmatrix}$$

Notice this makes sense, since we're interested in mean growth per gram. The estimate for mean growth per gram for Fertilizer A can simply average the first two pots' growths, but since two grams of fertilizer were added to pots 3 and 4, we need to cut their average in half to get growth per gram.

PART IV: Long Answer (8 Points Each Part)

8. Education researchers are interested in predicting students who are at risk of failing secondary school in Portugese schools. The response variable of interest is the final year exam grade (*exam*, on a scale of 0-20), and possible predictors include attributes about the students' family lives. We consider two possible predictor variables: the students' current health status (*health*) and how often the student goes out with friends (*goout*), both on a scale of 1: very low to 5: very high. Some relevant R output is shown in the Appendix.

   (a) First we consider a model using health status as a predictor variable. The calculated p-value for the slope of the model is 0.224. At a significance level of 0.05, what does that tell us about using health as a predictor for final year exam performance? Assume all assumptions are met. Explain as if to someone with no statistical experience.

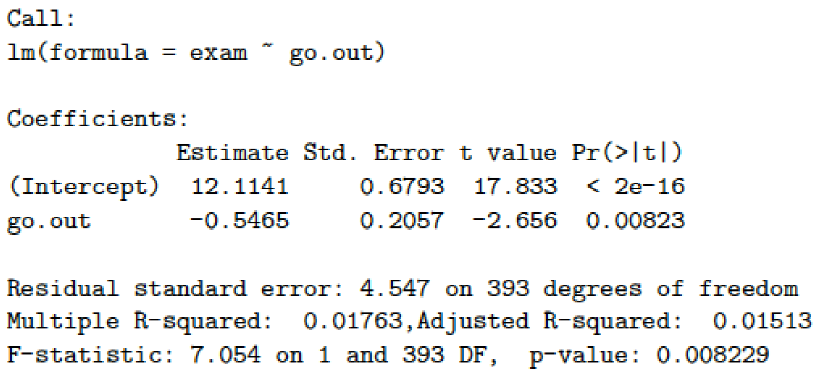
   The p-value tells us that we are quite likely to see a slope like the one in our dataset (or one more extreme) just from random chance. We don't have evidence that health and exam scores are associated with the population (be careful to fail to reject $H_0 : \beta_1 = 0$ in your wording; don't accept it!), so we wouldn't use it if we were aiming for simplicity.

   (b) Next we consider the model $exam_i = \beta_0 + \beta_1 goout_i + e_i$. Interpret the estimated slope in context, paying special attention to making sure a layman understands what the sign of the slope means.

   We observe that our estimate of the slope is $\hat{\beta}_1 = -0.5465$. This means that if the frequency of students' going out with friends increases by 1 unit (perhaps 1 time per week?), we estimate that their average exam score declines by about half a point. The negative slope indicates the decline in scores as going out increases.

Appendix

Portugese Schools Output (Going Out Model):



```
Call:
lm(formula = exam ~ go.out)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.1141     0.6793  17.833  < 2e-16
go.out       -0.5465     0.2057  -2.656  0.00823

Residual standard error: 4.547 on 393 degrees of freedom
Multiple R-squared:  0.01763,Adjusted R-squared:  0.01513
F-statistic: 7.054 on 1 and 393 DF,  p-value: 0.008229
```