STAT 7630 - Midterm Exam

Augusta University

March 5th, 2024

Unless stated otherwise, assume all error terms are distributed iid normal: $\varepsilon \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$.

1. The usual simple linear regression model is stated as

$$\begin{aligned} y_i &= \beta_0 + \beta_1 x_i + \varepsilon_i \\ \varepsilon_i &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2), i = 1, \ldots, n. \end{aligned}$$

The intercept-free model is written $y_i = \beta x_i + \varepsilon_i$, with the same distributional assumptions on the error terms. For this problem, assume σ^2 is known.

a.) Compute the MLE for β in the intercept-free model.

b.) How does the interpretation of β differ from that of β_1 in the usual SLR model? Consider whether any of the assumptions implied by the model statements are different.

c.) If it known *a priori* that a regression line must have 0 intercept, what is the advantage of using the intercept-free model over the usual SLR model?

2. Suppose we have a data set of n observations, with outcome measurements y_i, and two covariate measurements x_{i1} and x_{i2}. We first fit the model $y_i = \beta_0 + \beta_1 x_{i1} + \varepsilon_i$ and find that β_1 is significantly different than 0. Does this imply that β_1 will also be significant when we fit the model $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \varepsilon_i$ with the same data set? Explain why or why not.

3. a.) Consider the SLR model $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. Suppose β_0 is known, but β_1 and σ^2 are not. Find $\hat{\beta}_1$ and $\hat{\sigma}^2$.

   b.) Compute the bias of the MLEs $\hat{\beta}_1$ and $\hat{\sigma}^2$ you found in (a).

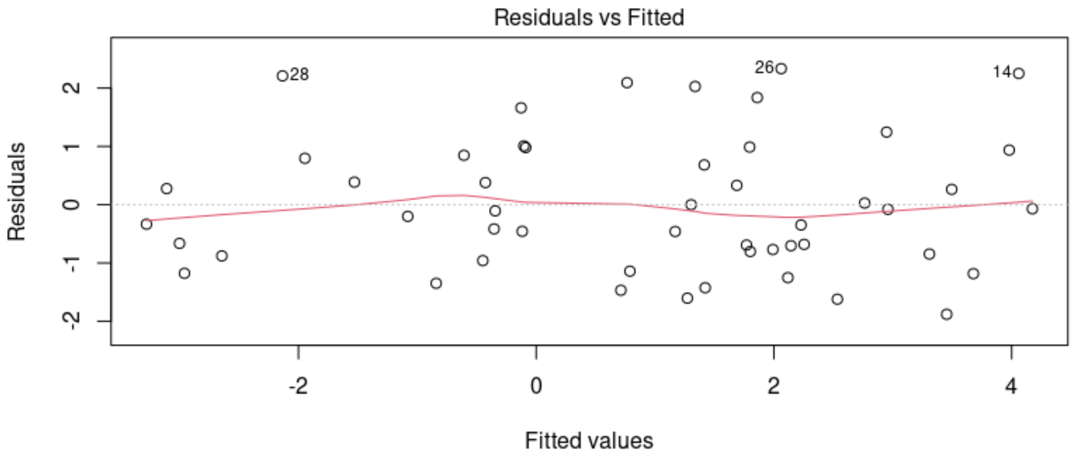
4. Briefly explain the difference between a standard normal distribution and a t-distribution. Will a 95% confidence interval from a t-distribution be wider or more narrow than an interval computed from a standard normal distribution?

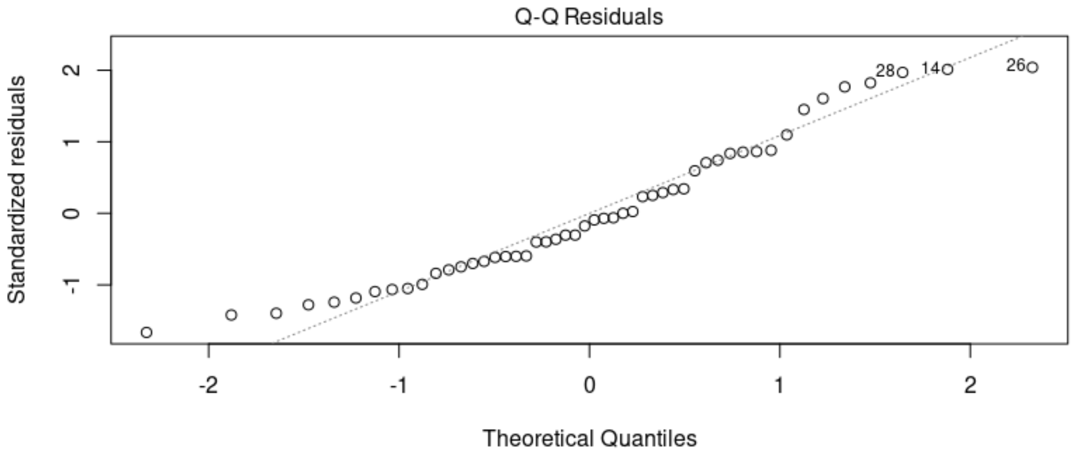
5. The matrix form of Multiple Linear Regression is written $Y = X\beta + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$, where n is the number of observations. For certain choices of covariate values, it can happen that $X^T X = (X^T X)^{-1} = I_p$; in this case, we say X is an *orthogonal design matrix*.

   a.) Simplify the expressions for $\hat{\beta}$ and $\hat{Y}$ in the case of an orthogonal design matrix.

   b.) Find the variance for $\hat{\beta}$ in the case of an orthogonal design matrix.

6. Suppose we have n observations with response vector Y and $n \times p$ covariate matrix X. Assume that the covariance matrix of the error terms is diagonal with diagonal entries $\sigma_i^2, i = 1, \ldots, n$. That is, each observation has its own error variance σ_i^2. In the case that all the σ_i^2 are known, explain how to augment the data so that the constant variance assumption of the linear regression model is satisfied.





7. For the two regression diagnostic plots shown below (residuals vs fitted, and residual QQ plot), comment on whether these plots indicate any potential violations of the linear regression model assumptions.

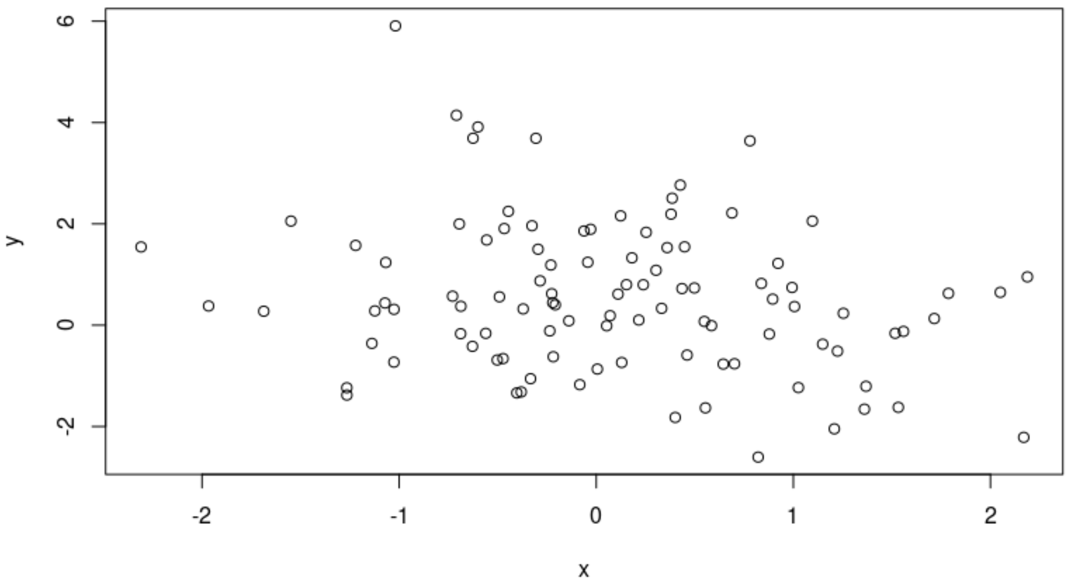


Figure 1: Problem 8: Scatterplot of the data.

8. For this problem, refer to the scatter plot of the data and the regression output in the figure.

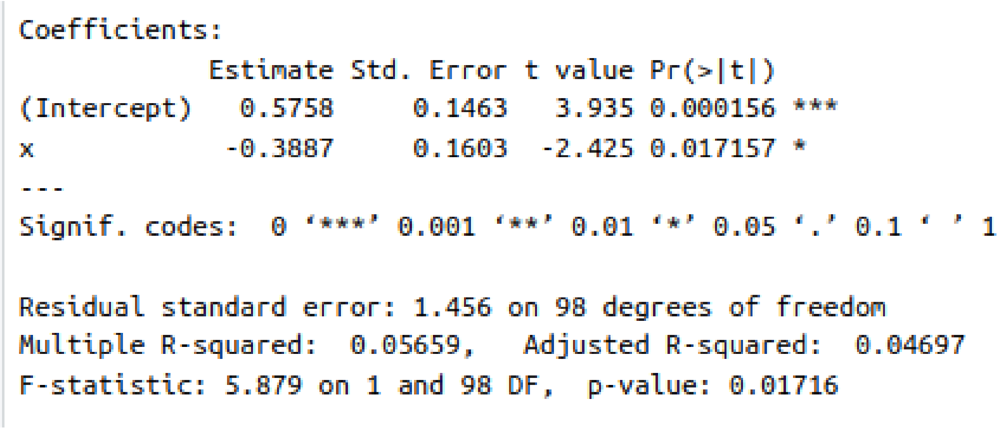
   a.) Which of the four linear regression assumptions can be evaluated from the scatter plot? Does this plot suggest those assumptions are violated? Explain.

   b.) From the R regression output, how many observations were used to fit the model? Explain how you determined this.

   c.) From the R regression output, is the hypothesis $H_0 : \beta_1 \neq 0$ significant at the 0.05 level?

   d.) On the scatter plot, draw the fit regression line (as best you can, need not be exact). On the line, identify $\hat{\beta}_0$ with a Δ. What are the coordinates of the point?

   e.) Write the formula for the confidence interval of β_1, plugging in any values that are known from the regression output. You do not need to compute or simplify.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.5758     0.1463   3.935 0.000156 ***
x            -0.3887     0.1603  -2.425 0.017157 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.456 on 98 degrees of freedom
Multiple R-squared:  0.05659,	Adjusted R-squared:  0.04697
F-statistic: 5.879 on 1 and 98 DF,  p-value: 0.01716
```

Figure 2: Problem 8: Regression output from R.

9. Consider the balanced one-way ANOVA model

$$\begin{aligned} Y_{ij} &= \mu_j + \varepsilon_{ij} \\ \varepsilon_{ij} &\overset{iid}{\sim} \mathcal{N}(0, \sigma^2), \end{aligned}$$

for $i = 1, \ldots, I; j = 1, \ldots, J$.

a.) Give a brief interpretation of the value $\mu = \frac{1}{J}\sum_{j=1}^{J} \mu_j$.

b.) Show that $SSTotal = SSW + SSB$. That is, verify the equality

$$\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{\cdot j})^2 + I\sum_{j=1}^{J}(\bar{Y}_{\cdot j} - \bar{Y}_{\cdot\cdot})^2$$

10. Explain how a confounding relationship can arise in a multiple linear regression setting. Why are unaccounted for confounders problematic in the analysis of real-world data? You may refer to a real-world example or example from the homework to aid your explanation.