L23: Homework Answer Key

Instructions: You are encouraged to collaborate with other students on the homework, but it is important that you do your own work. Before working with someone else on the assignment, you should attempt each problem on your own.

1. Give the estimated linear regression equation and the true linear regression equation. What are 3 differences between these two equations?

   Estimated linear regression equation:

   $$\hat{Y} = b_0 + b_1 X$$

   True linear regression equation:

   $$Y = \beta_0 + \beta_1 X + \epsilon$$

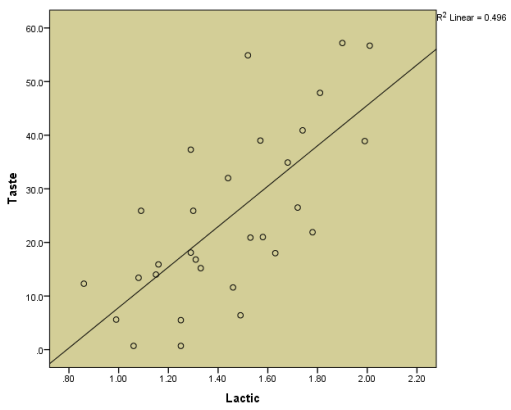
2. What are the requirements to check when doing linear regression, how do you check for them, and how can you tell that the requirements are met?

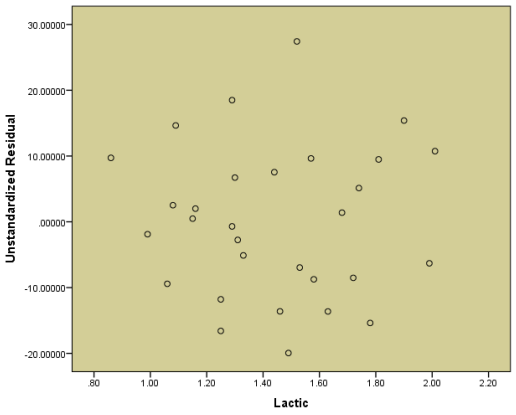
   See the wiki for a review of this important concept.

As cheese ages, various chemical processes take place that determine the taste of the final product. Concentrations of various chemicals were measured in 30 samples of mature cheddar cheese, and a subjective measure of taste was recorded for each sample. Higher values of Taste indicate a more desirable product. The variable Lactic gives the concentration of lactic acid in each sample. We want to know if there is a statistically significant linear relationship between the concentration of lactic acid in the cheese and the quality of the taste. We want to predict quality of taste given its concentration of lactic acid. Open the data file DASL-Cheese. Use this information to answer questions 3 through 12.

3. Check the requirements for simple linear regression with the variables Lactic and Taste. Complete the remainder of the problems even if the requirements are not all met.

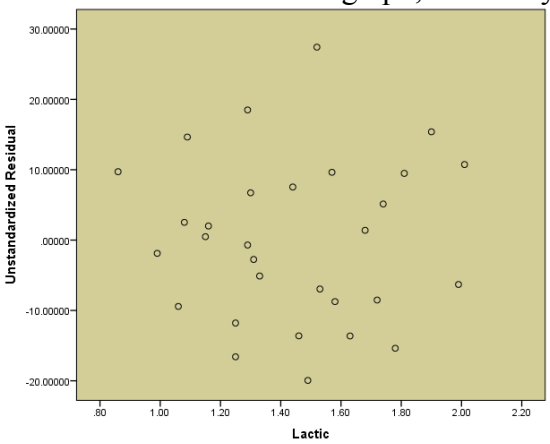
   a. Create and attach the graphs that are appropriate to check for the requirement of a linear relationship. Based on the graphs, what do you conclude?





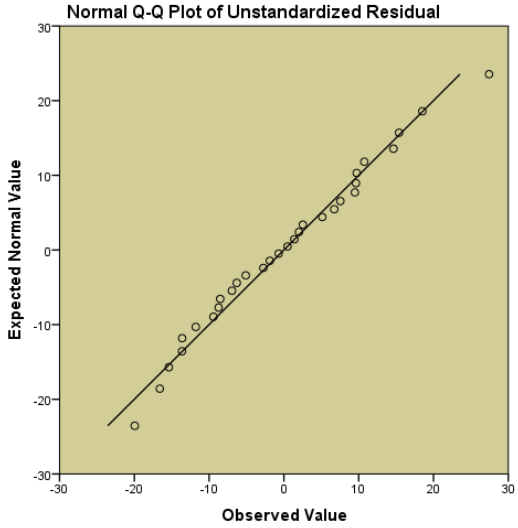
The appropriate graphs to check for a linear relationship are a scatterplot and a residual plot. The scatterplot seems to show a linear relationship and there is no pattern in the residual plot, so we can conclude that there is a linear relationship in the data.

b. Create and attach a graph that is appropriate to check for the requirement of constant variance. Based on the graph, what do you conclude?



The appropriate graph to check for constant variance is a residual plot. There is no pattern in the residual plot, so we can conclude that there is a constant variance in the data.

c. Create and attach a graph that is appropriate to check for the requirement of a normal error term. Based on the graph, what do you conclude?



The appropriate graph to check for a normal error term is a Q-Q plot of the residuals. The points in the plot are close to the line, so we can conclude that there is a normal error term in the data.

4. Compute the sample correlation coefficient (***r***).

r = 0.704

5. Find the equation of the linear regression line used to predict the quality of taste of a cheese sample given its concentration of lactic acid.
   $\hat{Y} = -29.859 + 37.720X$

6. Use software to predict the quality of taste of a cheese sample that has a concentration level of 2.11 of lactic acid.
   Y = 49.730

7. Find and interpret a 95% confidence interval for the slope of the regression line obtained when Lactic is used to predict Taste.
   (22.999, 52.441) We are 95% confident that the slope of the true linear regression line of Lactic with Taste is between 22.999 and 52.441.

Conduct a hypothesis test using this information to see if there is a statistically significant linear relationship between the concentration of lactic acid in the cheese and the quality of the taste. Use a level of significance of $\alpha = 0.05$.

8. State the correct null and alternative hypotheses.
   $H_o: \beta_1 = 0$
   $H_a: \beta_1 \neq 0$

9. Give the test statistic and its value.
   $t = 5.249$

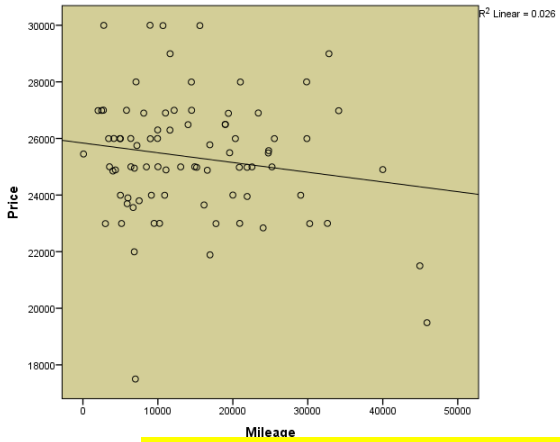
10. Calculate the P-value based on the test statistic.
    P-value is very close to 0.

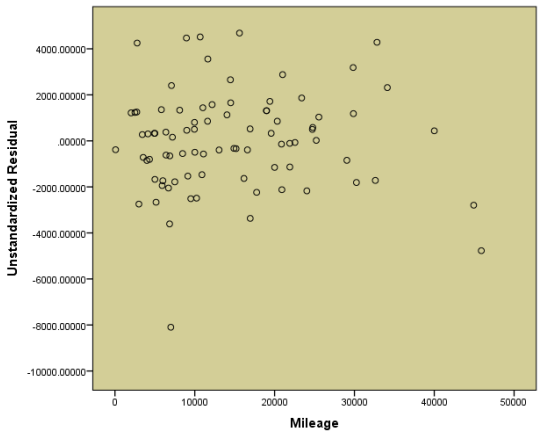
11. What decision do you make based on the P-value and the level of significance (α)?
    Reject the null hypothesis.

12. State your conclusion in an English sentence.
    There is sufficient evidence to suggest that the slope of the true linear regression line does not equal zero. We conclude that there is a linear relationship between the concentration of lactic acid in cheese and the quality of its taste.

Data was collected from a sample of used 2005 Toyota Prius, including the advertised price, the mileage, and a description of each car. You want to determine if there is a statistically significant linear relationship between the mileage of a used Prius and its price. You want to predict price based off of mileage. Open the data set ToyotaPrius2005. Use this information to answer questions 13 through 21.

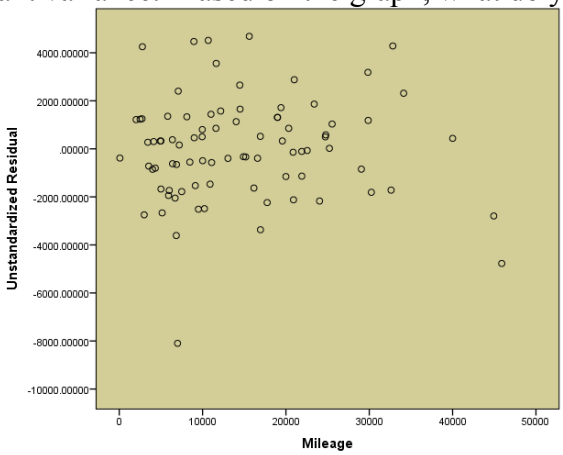
13. Check the requirements for simple linear regression with the variables Mileage and Price. Complete the remainder of the problems even if the requirements are not all met.
    a. Create and attach the graphs that are appropriate to check for the requirement of a linear relationship. Based on the graphs, what do you conclude?





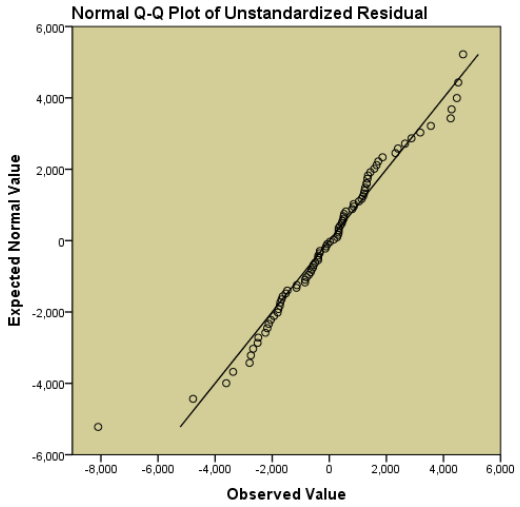
The appropriate graphs to check for a linear relationship are a scatterplot and a residual plot. The scatterplot does not seem to show a significant linear relationship, so we cannot conclude that there is a linear relationship in the data.

    b. Create and attach a graph that is appropriate to check for the requirement of constant variance. Based on the graph, what do you conclude?



The appropriate graph to check for constant variance is a residual plot. There is no pattern in the residual plot, so we can conclude that there is a constant variance in the data.

c. Create and attach a graph that is appropriate to check for the requirement of a normal error term. Based on the graph, what do you conclude?



The appropriate graph to check for a normal error term is a Q-Q plot of the residuals. The points in the plot are close to the line, so we can conclude that there is a normal error term in the data.

14. Find the equation of the linear regression line used to predict the listing price of a used Prius given its mileage.

$\widehat{Y} = 25838.626 - 0.034X$

15. Use software to predict the listing price of a used Prius with 100,000 miles.

Y = 22401.192

16. Find and interpret a 90% confidence interval for the slope of the regression line obtained when Mileage is used to predict Price.

(-0.073, 0.004) We are 90% confident that the slope of the true linear regression line of Mileage with Price is between -0.073 and 0.004.

Conduct a hypothesis test using this information to see if there is a statistically significant linear relationship between the mileage of a used Prius and its price. Use a level of significance of $\alpha = 0.05$.

17. State the correct null and alternative hypotheses.

$H_o: \beta_1 = 0$
$H_a: \beta_1 \neq 0$

18. Give the test statistic and its value.

$t = -1.476$

19. Calculate the P-value based on the test statistic.

P-value = 0.144

20. What decision do you make based on the P-value and the level of significance (α)?
    Fail to reject the null hypothesis.

21. State your conclusion in an English sentence.
    There is insufficient evidence to suggest that the slope of the true linear regression line does not equal zero. We conclude that there is not a linear relationship between the mileage of a Prius listed for sale and its price.

Research was conducted to understand whether exposure to lead in children between the ages of 1 and 3 years old affected their behavior. Researchers measured the blood lead level of each child in µmol/L. Researchers also assessed the behavior of each child using the Behavior Rating Scale (BRS). Higher scores indicate fewer behavioral problems. Imagine you are part of this research study. You want to determine if there is a statistically significant relationship between lead exposure (Lead) and children's behavior (BRS). Researchers want to predict BRS rating given blood lead level. Open the data file LeadExposureandBehavior. Use this information to answer questions 22 through 30.

22. Compute the sample correlation coefficient (***r***).
    $r = -0.181$

23. Find the equation of the linear regression line used to predict the BRS rating of a child given his or her blood lead level.
    $\hat{Y} = 62.825 - 18.236X$

24. Use software to predict the BRS rating of a child given a blood lead level of 0.75 µmol/L.
    Y = 49.148

25. Find and interpret a 95% confidence interval for the slope of the regression line obtained when Lead is used to predict BRS.
    (-41.855, 5.383) We are 95% confident that the slope of the true linear regression line of Lead with BRS is between -41.855 and 5.383.

Conduct a hypothesis test using this information to see if there is a statistically significant linear relationship between lead exposure (Lead) and children's behavior (BRS). Use a level of significance of α = 0.05.

26. State the correct null and alternative hypotheses.
    $H_o: \beta_1 = 0$
    $H_a: \beta_1 \neq 0$

27. Give the test statistic and its value.

t = -1.540

28. Calculate the P-value based on the test statistic.
    P-value = 0.128

29. What decision do you make based on the P-value and the level of significance (α)?
    Fail to reject the null hypothesis.

30. State your conclusion in an English sentence.
    There is insufficient evidence to suggest that the slope of the true linear regression line does not equal zero. We conclude that there is not a linear relationship between a child's level of lead exposure and his or her behavioral rating.

31. A scatterplot is made to test correlation between two variables. When residuals were calculated, one residual had a value of 4.5. What does this mean?
    a. There is a strong positive correlation between the two variables
    b. There is a weak positive correlation between the two variables
    c. The predicted `Y` value was 4.5 units higher than the actual `Y` value
    d. The actual `Y` value was 4.5 units higher than the predicted `Y` value