Lab 5

STA 216 F19

Suppose that we are interested in purchasing a Porsche sports car. (I mean, hey, why not?) If we can’t afford the high sticker price of a new Porsche, we might be interested in finding a used one. How much should we expect to pay? The price might depend on many factors including the model, condition, and special features of the car. The sample of Porsches in “Lab 5 – Porsche price.xlsx” came from Autotrader.com in Spring 2007. The variables include the following:

|  |  |
| --- | --- |
| SAS variable name | Description |
| Price | Price in thousands of dollars |
| Mileage | Number of miles the car has been driven in thousands |
| Age | Age of the car in years |

1. Let’s start by predicting the price of a used Porsche based on its mileage.
   1. Identify the explanatory variable and the response variable in this case.
   2. Copy and paste a scatterplot of these variables including the least squared regression line.
2. The scatterplot should show a linear relationship, so let’s apply the simple linear regression model. This model states that .
   1. In this particular data context, what are Y and X?
   2. This model implies 4 assumptions. List these four assumptions.
3. We will check the other 3 assumptions besides independence. (The latter can be assumed because it is reasonable to think the price of one car would not affect the price of another car.)
   1. Copy and paste the residual vs. predicted plot obtained from SAS.
   2. We use this plot to check two assumptions. Explain specifically what about this plot indicates whether each of the two assumptions is satisfied or violated. (Hint: in general, your interpretation of the plot should not be affected by two or three points, but rather the general patterns of the plot.)
   3. Copy and paste a normal quantile plot of the residuals from SAS.
   4. We use this plot to check the other assumption. Explain specifically what about this plot indicates whether this assumption is satisfied or violated.
   5. Write the equation of the regression line in terms of Price and Mileage, using “(hat)” where appropriate. Copy and paste the table of the SAS output that supplies the numbers used.
   6. Interpret the intercept of the regression line.
   7. Interpret the slope of the regression line.
4. Assuming this sample is representative of the population of all used Porsches for sale in 2007, we can do a hypothesis test about the slope between mileage and price in this population. For the hypothesis test of the slope – i.e. Ho: vs. Ha: ,
   1. Report the test statistic and show how it is calculated from the slope estimate and its standard error.
   2. Report the p-value.
   3. Write a conclusion for this hypothesis test.
   4. Report a 95% confidence interval for the slope, copying and pasting the SAS table that includes it.
   5. Interpret this confidence interval.
5. Another hypothesis test of the slope (the F test) is calculated from the ANOVA table.
   1. Copy and paste the ANOVA table for the simple linear regression of price by mileage.
   2. Explain where the DF values come from.
   3. The sum of squares values come from the sums of squares of three differences:   
      SS(model) is the sum of , SS(error) is the sum of , and SS(total) is the sum of .
      1. Copy and paste code that creates variables yhat\_ybar2, y\_yhat2, and y\_ybar2 that represent each of these three squared differences.
      2. Copy and paste output from PROC MEANS that shows that the sums of these variables are equal to the sum of squares given in the ANOVA table.
   4. Show how the mean square values are calculated from the sum of squares values.
   5. Show how the F statistic given as “F value” is calculated from the mean square values.
   6. Report the R-squared value and copy and paste the SAS table that gives it.
   7. Show how it is calculated from the ANOVA table.
   8. Interpret the R-squared value in the context of the problem.
   9. Report the root MSE value.
   10. Show how it is calculated from the ANOVA table.
   11. Interpret the root MSE value in the context of the model.
   12. Which parameter in the model , is the root MSE an estimate of?
   13. Report a 95% prediction interval for the price of an individual used Porsche in 2007 with 40,000 miles on it. Copy and paste the first row of the table that gives it in the SAS output.
   14. Report a 95% confidence interval for the mean price of all used Porsches in 2007 with 40,000 on them. Copy and paste the first row of the table that gives it in the SAS output.
   15. The prediction interval should be wider than the confidence interval. Why does this make sense based on what each is trying to capture?
6. Let’s consider age in comparison to mileage in regards to how well it can predict price.
   1. Report the R-squared and root MSE values when we try to predict price using age. Also, copy and paste the table from SAS that gives them.
   2. Using these values and comparing them to those from the price vs. mileage model, explain whether age or mileage is a better predictor of price.