Stats 230 - Fall 2022 Exam 1 - SLR only

Dhyey Mavani

TOTAL POINTS

39.5 / 44

QUESTION 1

1 Q1 - corr coeff **2 / 2**

✓ **- 0 pts** Correct

QUESTION 2

2 Q2 - scatterplots **1.5 / 2**

✓ **- 0.5 pts** price with mileage is nonlinear

QUESTION 3

3 Q3 - concerns **2.5 / 3**

✓ **- 0.5 pts** Run model with and without potential outliers to see if they are impacting conclusions.

QUESTION 4

4 Q4 - r from R2 output **1 / 2**

✓ **- 1 pts** Correlation coefficient (r) is negative

QUESTION 5

5 Q5 - Resids **3 / 4**

✓ **- 1 pts** Resids v. fitted looks for pattern (curve, heteroscedastic)

✓ **- 0 pts** histogram also checks errors centered at zero

QUESTION 6

6 q6 - conclude lm1 **2 / 2**

✓ **- 0 pts** Correct

QUESTION 7

7 q7 - Transform & subset code **2.5 / 3**

✓ **- 0.5 pts** log transformation to try and linearize the relationship

QUESTION 8

8 q8 scatterplot **6 / 6**

✓ **- 0 pts** Correct

QUESTION 9

9 19- slope and intercept lm2 **4 / 4**

✓ **- 0 pts** Correct

QUESTION 10

10 q10 - error lm2 **2 / 2**

✓ **- 0 pts** Correct

QUESTION 11

11 q11 - ConfInter coeff **1.5 / 2**

✓ **- 0 pts** true slope in the population (true relationship between mileage and logprice). Ours is just an estimate from one sample.

✓ **- 0.5 pts** predicting logprice from mileage

QUESTION 12

12 q12a- Predict price **2 / 2**

✓ **- 0 pts** Correct

QUESTION 13

13 q12b - prediction cautions **1.5 / 2**

✓ **- 0.5 pts** Top of observations, we limited our data to <150,000 so this may be a bit of an extrapolation.

QUESTION 14

14 q13 - ANOVA source table **2 / 2**

✓ **- 0 pts** Correct

QUESTION 15

15 q14 - lm2 or lm3 better **4 / 4**

✓ **- 0 pts** Correct

QUESTION 16

16 q15 - F test **2 / 2**

✓ **- 0 pts** Correct

Stats 230 – Intermediate Statistics Exam 1 (chapters 0 – 2) Fall 2022 Pamela B. Matheson, Ph.D.

NAME: Dhyey Dharmendrakumar Mavani

1. Show all work. You may receive partial credit for partially completed problems.

2. The exam is closed book. You may use a calculator. You may NOT use your cell phone.

3. You may not discuss the exam with anyone but Prof. Matheson. Uphold the Honor Code. Cell phones and mobile devices must be turned off and put away.

4. More space than needed is provided. You don't need to fill the space! (If you do, you are probably writing too much and taking too much time on that problem). YOU MUST ANSWER IN THE SPACE PROVIDED. If your answer appears elsewhere or drags into the next section (outside the boundaries give) I will not be able to see it to grade it.

5. If you cannot solve an earlier part of a problem, and a later part depends on that part, use a reasonable value and state you are ASSUMING that is the right value to use. This allows you to still get credit for later parts if you get stuck on something early on.

We will use a data set collected by former students. They randomly selected cars from the Cars.com website near their hometowns. Our goal was to predict price. We have two variables that could be potential predictors; they are mileage and year.

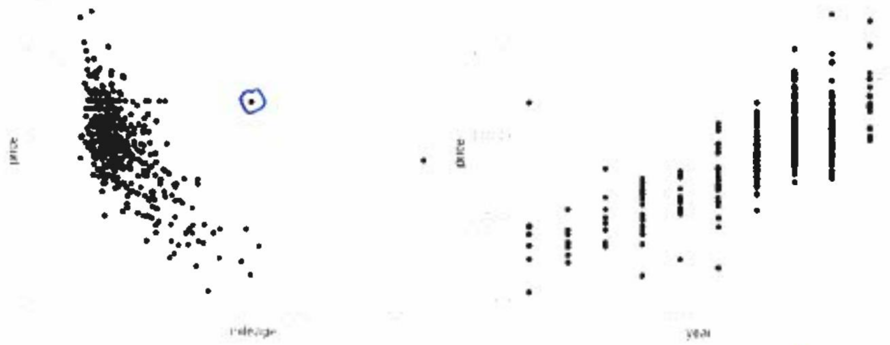
1. Based on the following output (using just what is shown here), state which variable you would use in a model to predict price and why.

```
cor(price ~year,      cars)
## [1] 0.744

cor (price ~ mileage,      cars)
## [1] -0.594
```

I would use "year" to predict price because of the higher absolute value of correlation coefficient in that case when compared to that in the case of "mileage".

2. Now you are given more information. Based on the following scatterplots, would your conclusion change? If so, why? If not, why not?



Based on the graphs, I would say that "mileage" might be a better choice since it appears to have ~~lower~~ higher R^2 values. Also, it ~~has a stronge~~ appears to have lower mean standard error of residuals.

3. Depending upon what predictor you chose, what concerns/issues do you have and what could you do to address them?

If we go with "mileage" as a predictor, we might have to deal with less or less dense data over higher mileages. Also, there might be a couple of outliers (as for example one circled in the graph), which might affect our regression analysis. Finally, there might be a better linear relationship between transformed versions of the variables such as exponential, logarithmic or square root of mileage, so we might need to look into that for better analysis.

___/___

The following is the regression output for a model predicting price from mileage as shown above in the scatterplot.

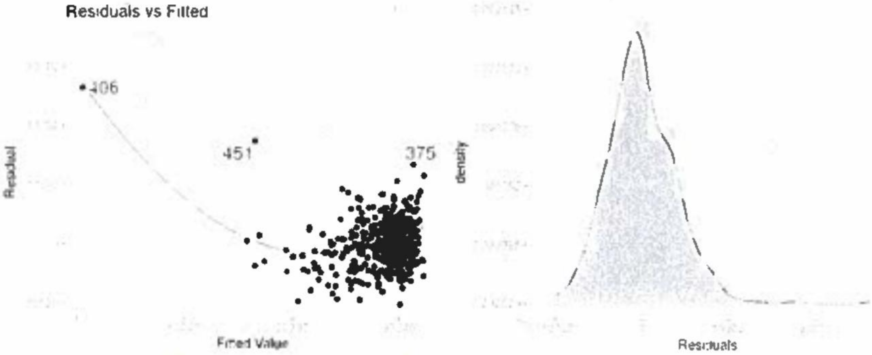
```
lm1 <- lm(price ~ mileage,       cars)
summary(lm1)

##
## Call:
## lm(formula = price ~ mileage, data = cars)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -10068  -2570   -351   2494  26422
##
## Coefficients:
##               Estimate  Std. Error t value Pr(>|t|)
## (Intercept) 32518.86012   309.27321     105   <2e-16 ***
## mileage        -0.09860     0.00617     -16   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4030 on 470 degrees of freedom
## Multiple R-squared:  0.352,  Adjusted R-squared:  0.351
## F-statistic:  256 on 1 and 470 DF,  p-value: <2e-16
```

4. Report the correlation value, r, between mileage and price based on the regression output from R above. (SHOW ALL WORK TO RECEIVE FULL CREDIT).

$r = \sqrt{0.352} \approx \boxed{0.593}$

5. What condition do each of the diagnostic plots (shown below) checking? Which diagnostic plot is the most concerning and why?



Residual vs Fitted plot checks the linearity of relationship.
The ~~P~~ residual density plot checks if the errors are nomally distribu
The Residual vs Fitted plot is most concerning since it is Ushaped, s
it might suggest non-linear relationship.

6. What do you conclude from this model (lm1)?

$\widehat{price} = 32518.86012 - 0.09860(\text{mileage})$.

We can say that the predicted price can be obtained with the above equation from least squares regression. But, it might be worthwhile to apply some transfomation to get a better linear relationship because currently the residual vs fitted plot is Ushaped, which suggests non-linearity

____/____

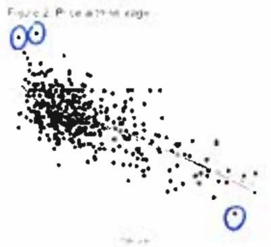
7. Two lines of R code are provided below. Briefly, explain what this code does, and why this was done in the context of lm1?

```
cars3 <- mutate (cars, logprice=log(price), logmileage=log(mileage))
cars4 <- filter(cars3, mileage<150000)
```

→ adds log(price) & log(mileage) columns in out dataset.

↳ filters the data such that we can get rid of outliers in terms of points which have more than 150000 mileage

8. A new scatterplot is created using cars4 data. Assess the relationship and talk about strength and direction of relationship, linearity, homoscedasticity and outliers.



→ We can see a ~~negativ~~ moderately strong negative linear relationship between price and mileage.

→ There are some outliers as marked in the graph, which pulls the line towards the top left hand corner and bottom right hand corner.

→ As the line is not wandering too much off the regression line, we can say that homoscedasticity is mostly valid.

A second regression model (lm2) was produced and after consultation with someone who has completed Stats 230 all assumptions were met. Here is that output.

```
lm2    lm(logprice ~ mileage,      cars4)
summary(lm2)
## Call:
## lm(formula = logprice ~ mileage, data = cars4)
##
## Residuals:
##      Min       1Q  Median       3Q      Max
## -0.5018 -0.0844  0.0036  0.0890  0.4082
##
## Coefficients:
##                  Estimate   Std. Error t value Pr(>|t|)
## (Intercept) 10.450289961  0.011614511   899.8   <2e-16 ***
## mileage     -0.000005306  0.000000257   -20.7   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.134 on 464 degrees of freedom
## Multiple R-squared:  0.479,  Adjusted R-squared:  0.478
## F-statistic:  427 on 1 and 464 DF,  p-value: <2e-16
```

9. Report and interpret the slope and y-intercept in context from the model. (state the actual values and explain what each means).

$\widehat{\log(price)} = 10.450289961 - 0.000005306$ (~~log~~(mileage))

→ y-intercept means that when mileage is 0 ~~or log(mileage) is 0~~, our predicted log(price) according to the model would be 10.450289961.

→ slope means that as we increase ~~log~~ (mileage) by 1 ~~(or multiply mileage by e)~~ our predicted value of log(price) decreases by 0.00000536.

___/___

10. What is the typical error in the model (lm2)? Make sure to interpret this in the context of the question.

Typical/Residual Standard error in the case of model lm2 is 0.134, which tells us how much on average does points wander off the least squares regression line of our model.
Based on the scale of the log(price) [which I cannot see that clear:
I would say 0.134 is ~~pret~~ significant typical error, so we might want to be cautious about that.

11. Your friend who already took STATS 230 showed you this output as well. What do the numbers next to mileage tell us?

```
confint(lm2)
##                    2.5 %      97.5 %
## (Intercept) 10.42746640 10.4731135
## mileage     -0.00000581 -0.0000048
```

It tells us that the coefficients of mileage in the lm2 model is between -0.00000581 and -0.0000048 with 95% confidence.
Or
We can say with 95% confidence that the coefficient of mileage in our model lm2's linear regression equation to predict log(pr lies between -0.00000581 and -0.0000048.

12. A friend wants to buy a used car from 2015 with 150,000 miles on it. You run the necessary commands in R (shown below).

```
predlogprice <- makeFun(lm2)
predprice150k = predlogprice (        150000)
exp(predprice150k)

##     1
## 15591
```

a. What do you advise in terms of price your model predicts they will pay?

My model predicts that they will pay a price of 15591 for a ~~car~~ usec car with 150000 miles on it as mileage.

b. What cautions might you lend to your prediction?

I would say since the least squares regression line has ~~comparitivel~~ comparatively high typical error, he should understand that this is not that accurate as it can wander up or down a bit, but the R^2 of around 48% seems promising.

____/____

Before we give up, perhaps another model might be useful. Here is some additional R output from a model (lm3) with year as a predictor.

```
lm3    lm(price ~ year,        cars2)
summary(lm3)

##
## Call:
## lm(formula = price ~ year, data = cars2)
##
## Residuals:
##     Min      1Q Median     3Q    Max
## -11669  -2050   -385   1794  12526
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4392705.8   179760.1   -24.4   <2e-16 ***
## year            2190.6       89.1    24.6   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3210 on 464 degrees of freedom
##

## F-statistic:  605 on 1 and 464 DF,  p-value: <2e-16
```

13. The following ANOVA output got mangled in the printer. Please fix by entering the proper values (*a, b, c and calculate R squared*). SHOW ALL WORK TO RECEIVE FULL CREDIT

```
anova(lm3)

## Analysis of Variance Table
##
## Response: price
##             Df     Sum Sq    Mean Sq  F value         Pr(>F)
## year         a  6250215413 6250215413     c            <2e-16 ***
## Residuals 464 4793866913        b
```

a= 1

b= $\cancel{(3210)^2 = 10304100}$ $\cancel{\frac{6250215413}{608}} = \cancel{\boxed{10330934.57}} = \frac{4793866913}{464} = \boxed{10331609.7}$

c= $\boxed{604.96} = \frac{6250215413}{10331609.73}$

R^2= $1 - \cancel{\frac{4793866913}{6250215413}} = \cancel{\boxed{0.233}} = \cancel{R^2}\ \frac{6250215413}{6250215413 + 4793866913} = \boxed{0.5659...}$ ≈ 0.57.

14. Assuming the diagnostic plots are good for lm3, which model is better (lm2 or lm3)? Why?

lm3 has better R^2 value, and thus I think lm3 is better model to predict price compared to lm3. ~~Also, the residual standard error is lesser for lm2 compared to lm3~~.

15. What does the F test tell us?

F test tells us that the pvalue is very low, which gives us indication that there is a significant relationship between the variables price and year.