IN THE NAME OF GOD

STATISTICAL INFERENCE HW#6

AMIN ASADI SARIJALOU

810196410

SPRING 1400

Contents

Problem 1

1-1

$$CI = point\ estimate \pm SE \times t^*_{df}$$

$$df = n - 2 = 24 - 2 = 22$$

$$t^*_{22,\,90\%} = 1.71$$

$$\rightarrow CI = 2.52 \pm 0.283 \times 1.71$$

$$= (2.03,\ 3)$$

1-2

a)

- Linearity
  - relationship between the explanatory and response variable should be linear
- Nearly Normal Residuals:
  - residuals should be nearly normally distributed
- Constant variability
  - variability of points around the least square line should be roughly constant

b)

$$T = \frac{b_1 - Null\ Value}{SE_{b1}} = \frac{-5 - 0}{0.52} = -9.61$$

$$\rightarrow\ pvalue \approx\ 0 \rightarrow$$

$$we\ confidently\ reject\ H_0\ in\ favor\ of\ H_a\ and\ hence\ we\ conclude\ that\ there\ is\ a\ linear\ relationship\ between\ these\ two\ variables.$$

1-3

a)

True

b)

False, it is a measure of (the strength of linear) association between two numerical variables.

Problem 2

The outlier is point: (5, 160),

because it falls away from the cloud of points.

Effects:

- The correlation coefficient (R) will increase and get closer to 1.

- b_0 (y-intercept) of the regression line will decrease and slope will increase.
- The standard deviation of the residuals will decrease.

Problem 3

3-1

No, because p-value for slope is 0.660 which is (much) greater than significance level (0.01) so we fail to reject H_0 and hence there isn't enough evidence to claim that there is a linear relationship between these variables for all players in this population.

3-2

a is the correct choice

Since $\alpha = 0.05$ so result of the p-value method will be equivalent to the result of 95% confidence interval method.

Therefore, because the Null Value = 0 falls into the confidence interval, we fail to reject H_0 and hence Nader can't conclude a linear relationship between how much toothpaste he uses and how long he brushes.

Problem 4

a)

$$\hat{y} = b_0 + b_1 x$$

$$\rightarrow \bar{y} = b_0 + b_1 \bar{x}$$

$$\rightarrow 3.9983 = 4.010 - 0.0883 \times b_1$$

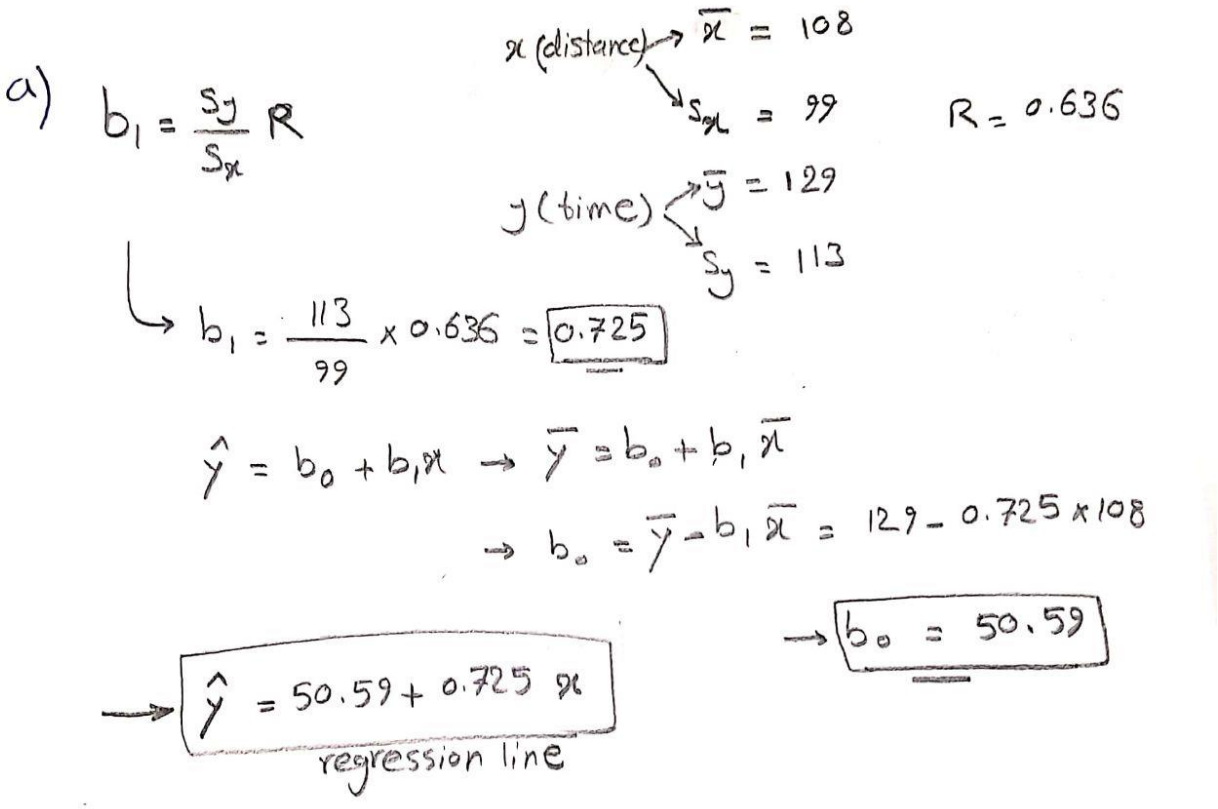
$$\rightarrow b_1 \approx 0.13$$

b)

Yes. According the given table, for a two-tailed test, the p-value is 0. Here we need a one-tailed test. Since the p-value for a one-tailed test is ½ of that of two-tailed test, and because half of the two-tailed p-value is still 0, so we have sufficient evidence to reject H_0 in favor of H_A $(= slope\ is\ positive)$ and conclude that the slope is positive.

Problem 5

a.



b.

Interpretation of Slop: For each 1 mile increase in the distance between two stops, we would expect that the time of travelling between those points, to be more on average by 0.725 minutes.

Interpretation of Intercept: For two stops having zero distance between each other, it takes on average 50.59 minutes to travel between them.

c.

$$R^2 = 0.636^2 \approx 0.40$$

It tells us that about 40% of the variability in the time of travel between stops is explained by the model (which include distance between stops as the variable).

Also 1 - 0.40 = 60% of the variability in the time of travel between stops is explained by other variables not included in the model.

d.

$$\hat{y} = b0 + b1 \times x = 50.59 + 0.725 \times 103 \approx 125.37\ minutes$$

e.

$$residual = actual - estimated = 168 - 125.07 \approx 42.62\ minutes$$

The residual is positive. This means that we have underestimated the response variable (time of travel) by about 42.62 minutes.

f.

No, we cannot. Because in fitting out model, we don't consider this much distant points. 500 miles distance, is a very long distance

because it is $\frac{500-108}{99} \approx 4$ standard deviation away from the mean of the distances which we consider.

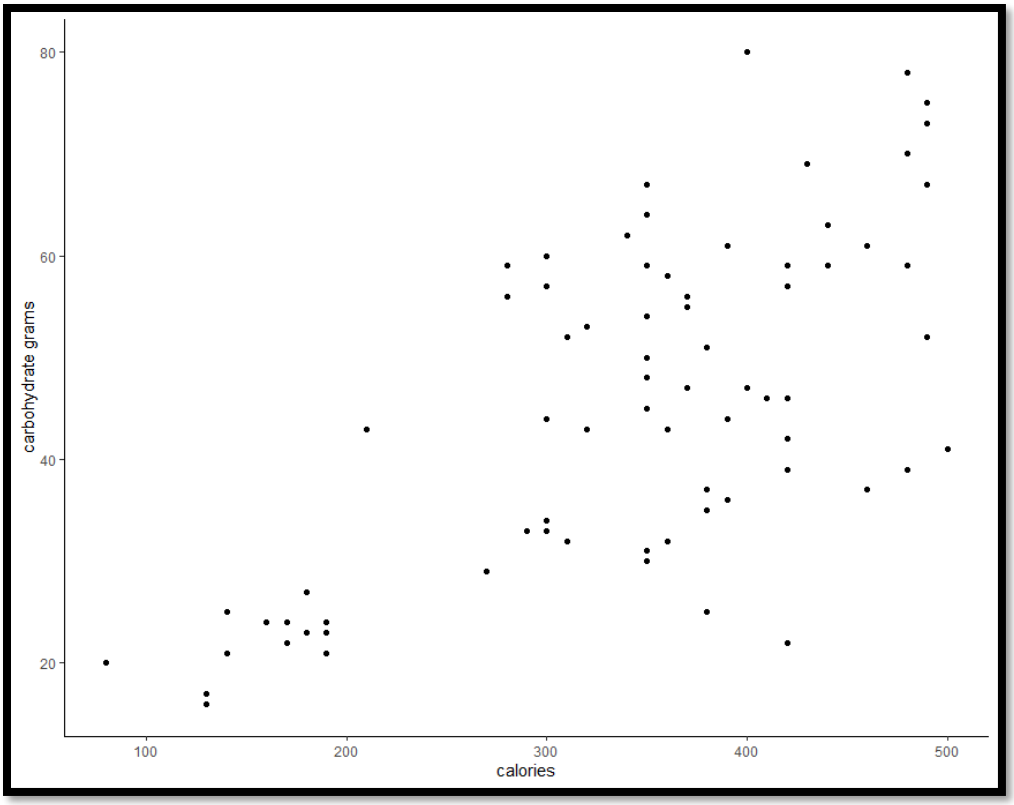
Problem 6

a)

```
# a
p_a <- ggplot(starbucks, aes(x = calories, y = carb)) +
  geom_point() +
  labs(x="calories",
       y="carbohydrate grams") +
  theme(plot.title = element_text(hjust = 0.5)) +
  theme_classic()

show(p_a)
```

The Data Points:



Scatterplot Description: I see that there seem to be a positive linear association between Carbohydrate Grams and number of Calories contained in each food item.

b)

Explanatory Variable: Calories

Response Variable: Carbohydrate Grams

We construct the problem in this way because we want to fit a model in order to predict the amount of Carbohydrate based on number of Calories.

c)

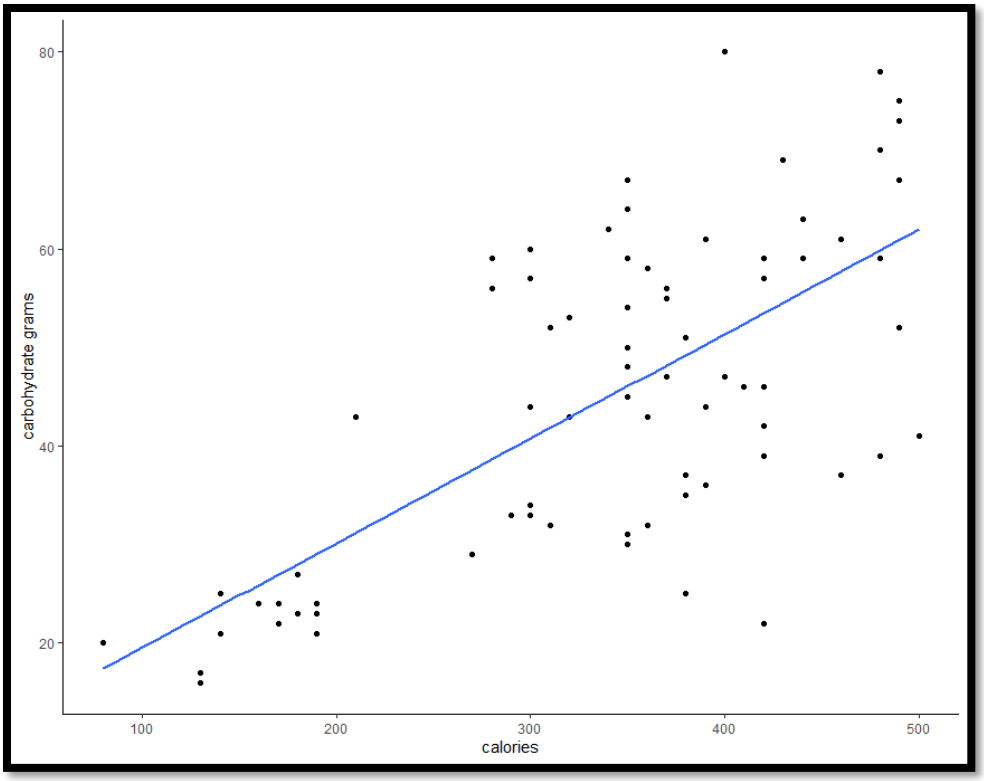
```
> lm(formula = carb ~ calories, starbucks)

Call:
lm(formula = carb ~ calories, data = starbucks)

Coefficients:
(Intercept)     calories
      8.944        0.106
```

```
p_c <- p_a + stat_smooth(method = lm, se = FALSE)
show(p_c)
```

The Fitted Regression Line:



d)

$$x = \ calories$$

$$y = \ calories$$

$$b_0 = 8.944, \quad b_1 = 0.106$$

$$\hat{y} = 0.106x + 8.944$$

Interpretation of Slop: With increasing the number of Calories by 1, we would expect that the amount of Carbohydrate increases on average by 0.106 grams.

Interpretation of Intercept: If a food has zero Calories, we would expect that it has on average 8.944 grams of Carbohydrate.

e)

```
> summary(model)

Call:
lm(formula = carb ~ calories, data = starbucks)

Residuals:
    Min      1Q  Median      3Q     Max
-31.477  -7.476  -1.029  10.127  28.644

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.94356    4.74600   1.884   0.0634 .
calories     0.10603    0.01338   7.923 1.67e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 12.29 on 75 degrees of freedom
Multiple R-squared:  0.4556,    Adjusted R-squared:  0.4484
F-statistic: 62.77 on 1 and 75 DF,  p-value: 1.673e-11
```

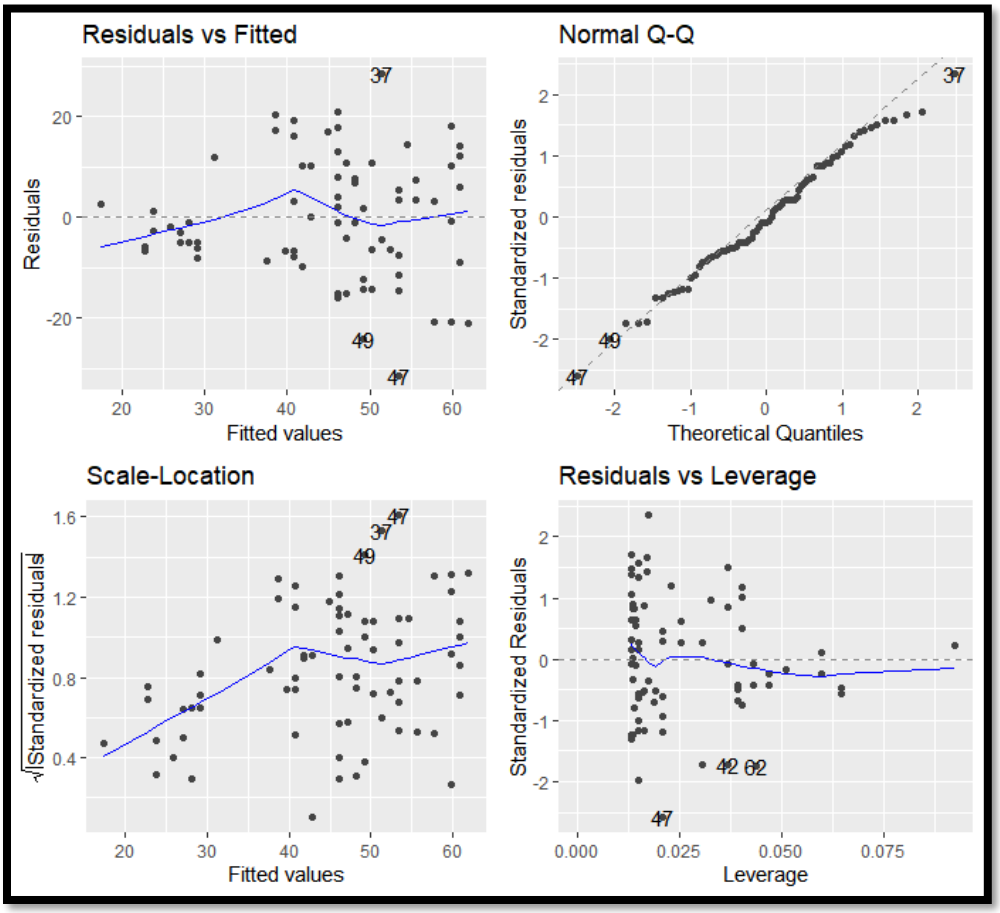
As we see in the summary of the fitted model, $R^2 \approx 0.45$ which means that about 45% of the variability of the response variable (grams of Carbohydrate) is explained by the model (which includes number of Calories as the variable).

f)

```
# c
model <- lm(formula = carb ~ calories, starbucks)
p_c <- p_a + stat_smooth(method = lm, se = FALSE)
show(p_c)

#e
summary(model)

# f
p_f <- autoplot(model)
show(p_f)
```



As we see in the Residuals Vs Fitted plot, the residuals seem to be different from the normal distribution at the tails. We can check that in the QQ-Plot beside it.

The QQ-Plot tells us that the same thing. The residuals are not normally distributed at tails.

We conclude that the fitted model is a good predictor in the middle parts, but it isn’t a good fit at the tails.