Homework 3 - Stats 230 - (MLR I: cereal, crab ships & car accidents)

Dhyey Mavani

date

PROBLEMS TO TURN IN: #3.2, #3.6, #3.7, #3.22a & d, #3.30, #3.32, Additional (see below)

Exercise 3.2

3.2 part a Show work:

SOLUTION: Predicted number of Calories = 109.3 + 1.0 * (11) - 3.7 * (1) = 116.6

3.2 part b Show work:

SOLUTION: Residual of Frosted Flakes is given by the (actual value - predicted value) = (110 - 116.6) = -6.6 Calories. Residual of -6.6 Cal in the case of Frosted Flakes means that our prediction (from the model) is 6.6 more than the actual value of Calories in Frosted Flakes.

Exercise 3.6 SOLUTION: According to our model, the value -3.7 as a coefficient of "Fiber" tells us that for every additional gram of "Fiber" in the sample of breakfast cereals at hand (with the amount of Sugar constant), will lead to a decrease of 3.7 Calories per serving.

Exercise 3.7

3.7 part a:

SOLUTION: Adjusted R^2 is always smaller than the unadjusted R^2 because adding additional parameters helps increase R^2 because it helps us predict bigger portion of the variability, but adjusted R^2 will increase by a lesser amount because it will punish (decrease) it's value for each additional variable added in our model.

(hint: look at the equation.)

3.7 part b:

SOLUTION: Not necessarily because if this additional predictor doesn't significantly increase the percentage explanability of variability, then it can keep the adjusted R^2 constant or can even make it decrease.

Exercise 3.22

3.22 part a:

SOLUTION: $R^2 = SSModel/SSTotal = 9350/17190 = 0.5439$ (approximately to 4 decimal places). This means that the model at hand helps us in explaining/ taking into account around 54% of the total variability.

3.22 part d: What does the p-value of 0.000002 mean as it relates to the F test value of 19.68?

SOLUTION: It tells us that there is a significant linear relationship between the predictor variables at hand, so we can reject the null hypothesis, which states that there is no linear relationship.

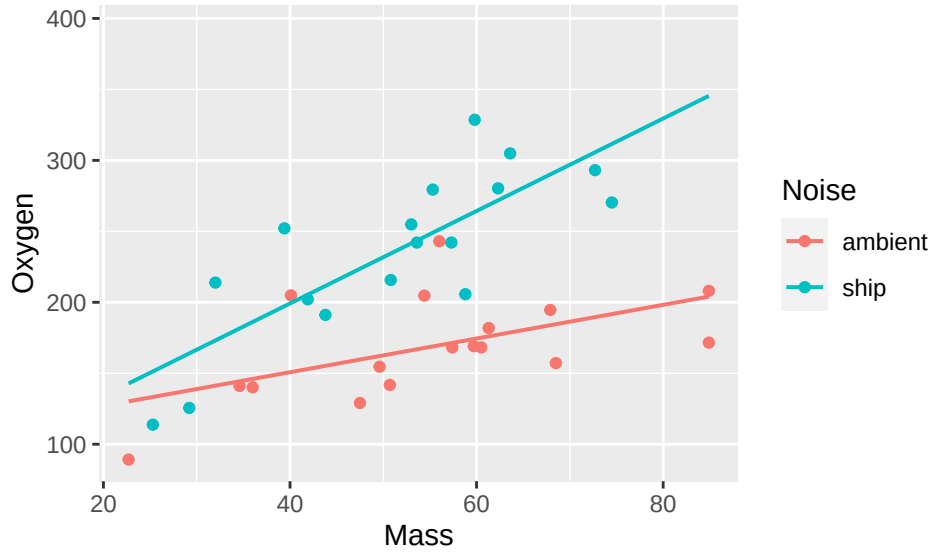
```
data(CrabShip)
```

Exercise 3.30

3.30 part a:

SOLUTION: We can see that there are different moderately linear positive relationships between Oxygen and Mass for different types of noises namely ambient and ship. For every unit increase in Mass, the Oxygen consumption increases more in the case of ambient Noise than in case of ship Noise.

```
gf_point(Oxygen ~ Mass, data = CrabShip, color = ~ Noise) %>%
   gf_lm() #to get you started
```



3.30 part b:

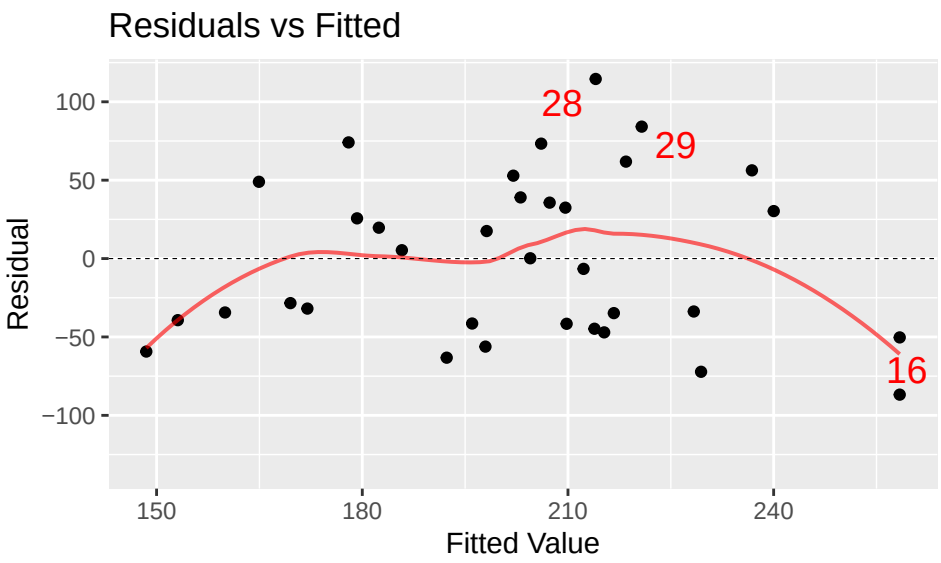
SOLUTION: If we just ignore the effect of Noise, the linear regression model between Oxygen and Mass has huge standard residual error and low R-squared value. Residuals are evenly distributed around zero, which means that there is equal variance. We can see that there is linearity of errors. Also, we can see from the Normal QQPlot that the errors are not perfectly normally distributed, which in total makes us alerted before using this a model.

```r
model1 <- lm(Oxygen ~ Mass, data = CrabShip)
msummary(model1)
```

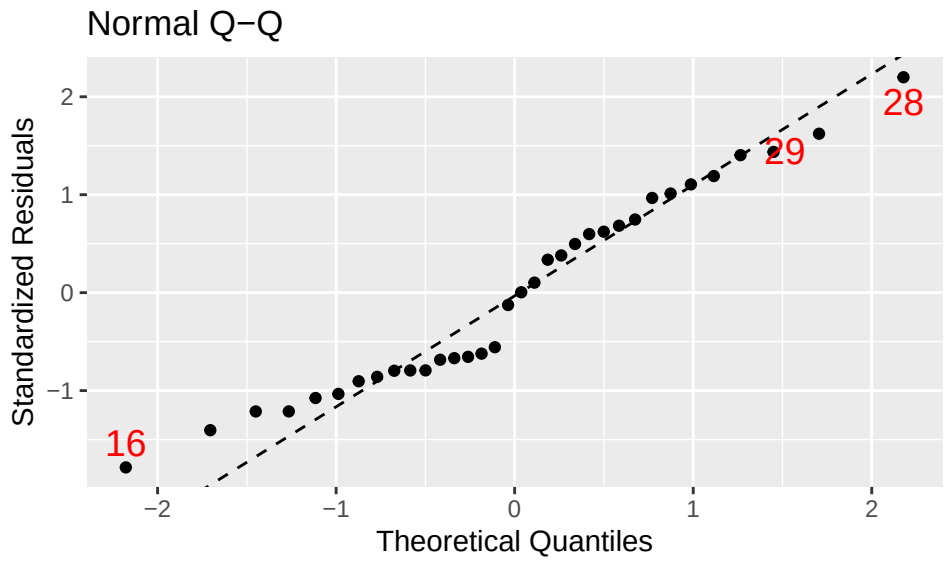
```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 108.39504   33.26499  3.2585 0.002655 **
## Mass          1.76667    0.60107  2.9392 0.006063 **
##
## Residual standard error: 53.019 on 32 degrees of freedom
## Multiple R-squared:  0.21258,    Adjusted R-squared:  0.18797
## F-statistic: 8.6389 on 1 and 32 DF,  p-value: 0.0060629
```

```r
mplot(model1, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
mplot(model1, which = 2)
```



3.30 part c:

SOLUTION: If we consider the lines parallel based on different Noise levels, then this definitely improves our model as we can see that our adjusted R-squared has risen to around 60% and our residual standard error has decreased by around 20 points. But, still I think we can improve as we can see in the first graph that the lines are not just shifted but have different slopes for different Noise types.

```
model2 <- lm(Oxygen ~ Mass + Noise, data = CrabShip)
msummary(model2)
```

```
##             Estimate Std. Error t value    Pr(>|t|)
## (Intercept) 54.42788   24.98164  2.1787     0.03708 *
## Mass         2.07337    0.42307  4.9008 0.000028537 ***
## Noiseship   75.27951   12.80002  5.8812 0.000001721 ***
##
## Residual standard error: 37.034 on 31 degrees of freedom
## Multiple R-squared:  0.62783,    Adjusted R-squared:  0.60382
## F-statistic: 26.148 on 2 and 31 DF,  p-value: 0.00000022207
```

3.30 part d:

SOLUTION: Considering the non-parallel lines case, we can see that we are getting an even better adjusted R-squared value of almost 67%. Also, we can see that the Residual standard error is reduced by 5 points.

```
model3 <- lm(Oxygen ~ Mass * Noise, data = CrabShip)
msummary(model3)
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    103.27025   29.38942  3.5139 0.001423 **
## Mass             1.18694    0.51209  2.3178 0.027462 *
## Noiseship      -34.39035   43.07815 -0.7983 0.430957
## Mass:Noiseship   2.07052    0.78264  2.6456 0.012857 *
```

```
##
## Residual standard error: 33.899 on 30 degrees of freedom
## Multiple R-squared:  0.69823,    Adjusted R-squared:  0.66806
## F-statistic: 23.138 on 3 and 30 DF,  p-value: 0.000000059425
```

3.30 part e:

SOLUTION: I think the model from part (d) is the best choice in this case due to the highest adjusted R-squared and lowest Residual standard error among all the models we made above simultaneously. The fitted prediction equation for (d) goes as follows:

$$\widehat{Oxygen} = 103.27 + 1.19(Mass) - 34.39(Noiseship) + 2.07(Mass * Noiseship)$$

Ambient Noise case:

$$\widehat{Oxygen} = 103.27 + 1.19(Mass)$$

Ship Noise case:

$$\widehat{Oxygen} = 65.88 + 3.26(Mass)$$

```
data(Speed)
```

Exercise 3.32 - StateControl = 1 means that states could change the speed limit, 0 means it was under federal control and set to 65mph on interstate highways.

3.32 part a:

SOLUTION: The slope of the least squares regression line is -0.045.

$$\widehat{FatalityRate} = 91.32 - 0.045 * (Year)$$

```
model1 <- lm(FatalityRate ~ Year, data = Speed)
msummary(model1)
```

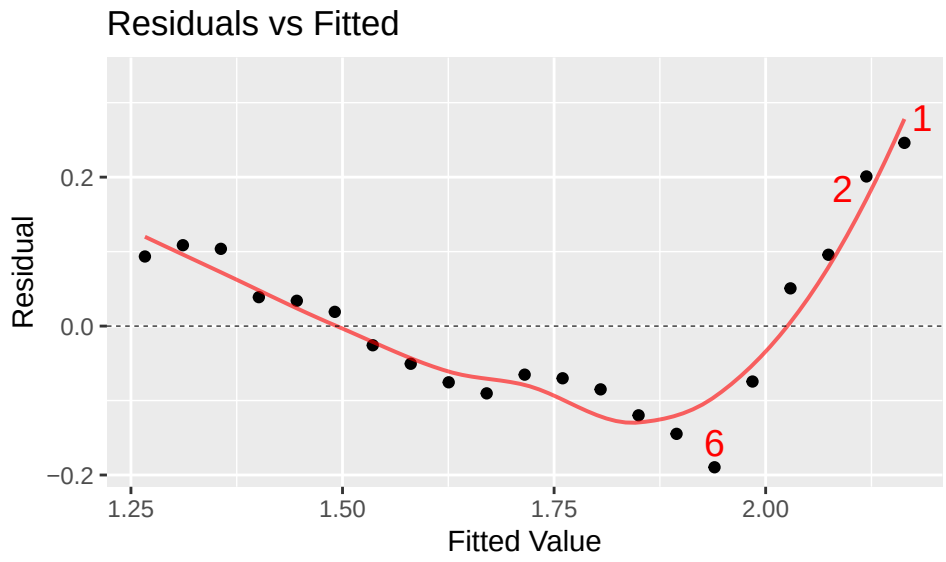
```
##               Estimate Std. Error t value       Pr(>|t|)
## (Intercept) 91.3208874  8.3742269  10.905 0.000000001281 ***
## Year        -0.0448701  0.0041934 -10.700 0.000000001750 ***
##
## Residual standard error: 0.11636 on 19 degrees of freedom
## Multiple R-squared:  0.85767,    Adjusted R-squared:  0.85018
## F-statistic: 114.49 on 1 and 19 DF,  p-value: 0.0000000017501
```

3.32 part b:

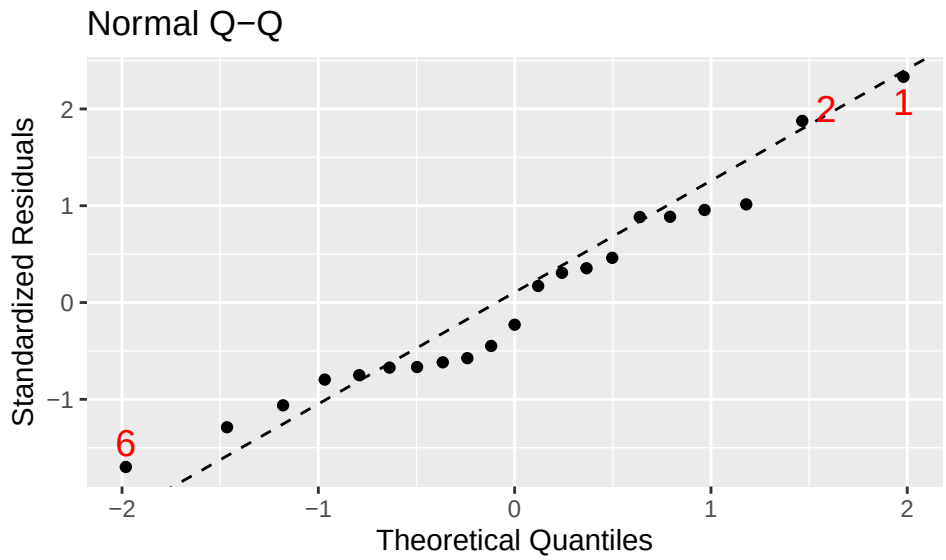
SOLUTION: The residuals vs fitted plot is U-shaped which means we should be skeptical and careful about using this model. Also, the points in the Normal QQPlot wander too much off the line, which is not a good sign especially in the terms of meeting the condition of normally distributed errors.

```
mplot(model1, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
mplot(model1, which = 2)
```



3.32 part c:

SOLUTION: We can see that there is a significant change in relationship between Fatality Rate and Year based on StateControl because the p-value of interaction term is less than 0.05.

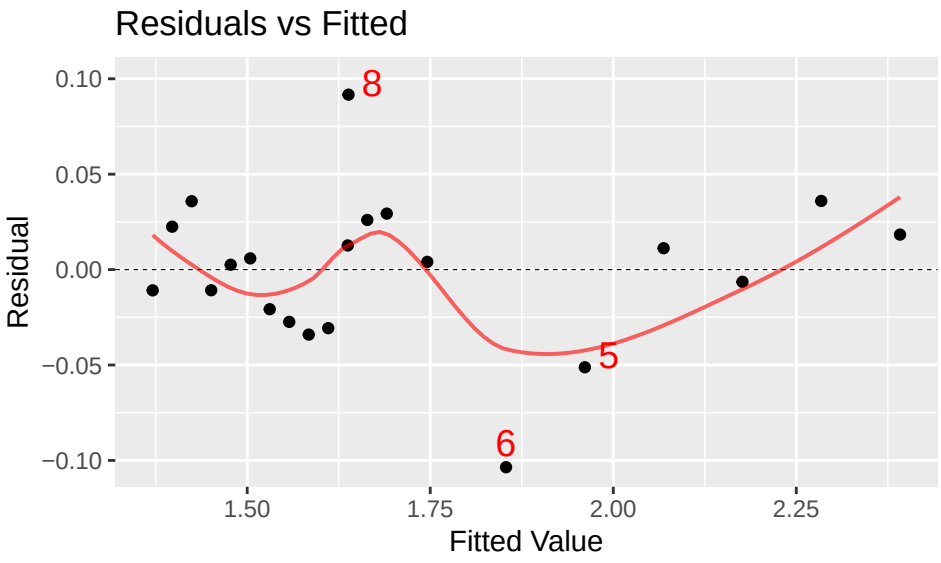
The residuals vs fitted plot is showing points distributed around 0, which means that constant variance is satisfied. Also, the points in the Normal QQPlot does not wander too much off the line, which is a good sign especially in the terms of meeting the condition of normally distributed errors.

```r
#includes a test procedure, so check conditions
model4 <- lm(FatalityRate ~ Year * StateControl, data = Speed)
msummary(model4)
```

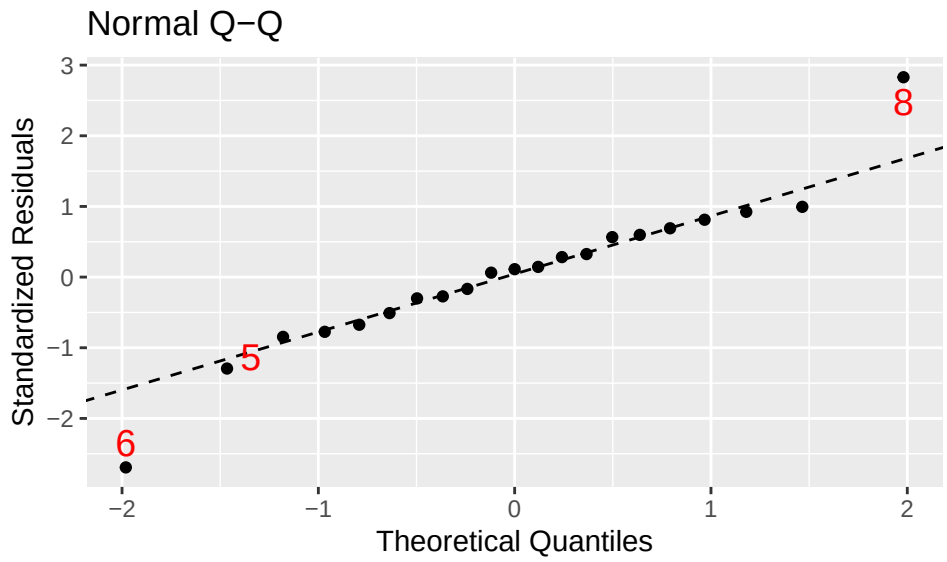
```
##                       Estimate   Std. Error t value  Pr(>|t|)
## (Intercept)        216.2307143   13.0329566  16.591 6.190e-12 ***
## Year                -0.1076190    0.0065476 -16.436 7.194e-12 ***
## StateControl      -161.3765934   14.4731016 -11.150 3.069e-09 ***
## Year:StateControl    0.0809707    0.0072639  11.147 3.082e-09 ***
##
## Residual standard error: 0.042433 on 17 degrees of freedom
## Multiple R-squared:  0.98307,    Adjusted R-squared:  0.98008
## F-statistic: 328.95 on 3 and 17 DF,  p-value: 2.9983e-15
```

```r
mplot(model4, which = 1)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```r
mplot(model4, which = 2)
```



3.32 part d:

SOLUTION:

$$\widehat{FatalityRate} = 216.231 - 0.108(Year) - 161.377(StateControl) + 0.081(Year : StateControl)$$

Before 1995:

$$\widehat{FatalityRate} = 216.231 - 0.108(Year)$$

After 1995:

$$\widehat{FatalityRate} = 54.854 - 0.027(Year)$$

Additional Problem (Not from your textbook!)

This problem is designed to give more practice with interpretations from the setting of the CrabShip problem. We'll be working with the interaction model from part d in Exercise 3.30, called model3 in the example code above. If you renamed it, use what you called it.

```
model3 <- lm(Oxygen ~ Mass * Noise, data = CrabShip)
msummary(model3)
```

```
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)    103.27025   29.38942  3.5139 0.001423 **
## Mass             1.18694    0.51209  2.3178 0.027462 *
## Noiseship      -34.39035   43.07815 -0.7983 0.430957
## Mass:Noiseship   2.07052    0.78264  2.6456 0.012857 *
##
## Residual standard error: 33.899 on 30 degrees of freedom
## Multiple R-squared:  0.69823,    Adjusted R-squared:  0.66806
## F-statistic: 23.138 on 3 and 30 DF,  p-value: 0.000000059425
```

AP1 part a: Which level of Noise is the base/reference level? How do you know?

SOLUTION: Ambient is the Base reference level of Noise. We can tell this because when Noiseship is 1, the slope will be higher. Also, Ambient comes alphabetically before ship, so R would take it as default the base case.

AP1 part b: Interpret the slope coefficient for Mass from model3.

SOLUTION: For every unit increase in Mass, the predicted value of the fatality rate rises by 1.187 units according to our interaction model.

AP1 part c: Interpret the slope coefficient for the interaction term from model3.

SOLUTION: When the Noise is of ship type then for every unit increase in Mass leads to an increase of 2.258 units instead of the increase of 1.187 units which was the case when the Noise is of ambient type

AP1 part d: We will assume appropriate conditions hold for inference (i.e. you don't need to check them here). Does the overall regression model (model3) appear to be useful for predicting Oxygen at a significance level of 0.05? Report a specific test statistic (i.e. which one), p-value, and conclusion.

SOLUTION: Since the p-value of interaction term is 0.0129 which is less than 0.05, we can say that we have sufficient evidence to suggest that there is a significant difference in slope of the regression lines in the case of different Noise types or in other words there is significant evidence to reject parallelism of lines hypothesis in the case of different types of Noise levels.

AP1 part e: One individual predictor has a large p-value for it's t-test for slope. Explain what that means in the context of the problem.

SOLUTION: Noiseship has high p-value because most of the variability which was explained by Noiseship earlier is being expressed by Mass:Noiseship in the updated model.

Note: Due to the interaction being significant, convention would still lead us to keep this term. We will discuss this more with polynomial regression.