

## DS-6030 Homework Module 2

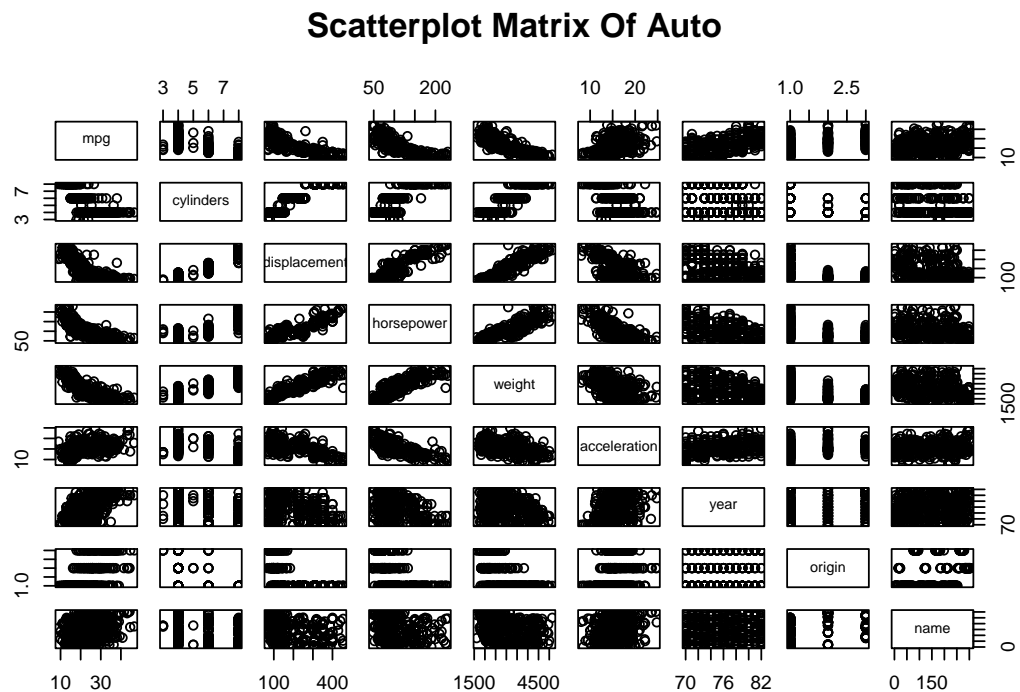
Tom Lever

05/31/2023

DS 6030 | Spring 2022 | University of Virginia

1. This question involves the use of multiple linear regression on the Auto data set.
  - (a) Produce a scatterplot matrix which includes all of the variables in the data set.

```
library(ISLR2)
pairs(Auto, main = "Scatterplot Matrix Of Auto")
```



- (b) Compute the matrix of correlations between the variables using the function `cor()`. You will need to exclude the name variable, which is qualitative.

```
library(TomLeversRPackage)
index_of_column_name <- get_column_index(Auto, "name")
data_frame_of_columns_except_name <- Auto[, -index_of_column_name]
cor(data_frame_of_columns_except_name)
```

```
#           mpg  cylinders displacement horsepower    weight
# mpg      1.000000 -0.7776175   -0.8051269  -0.7784268  -0.8322442
# cylinders -0.7776175   1.0000000    0.9508233   0.8429834   0.8975273
```

```

# displacement -0.8051269  0.9508233    1.0000000  0.8972570  0.9329944
# horsepower    -0.7784268  0.8429834    0.8972570  1.0000000  0.8645377
# weight        -0.8322442  0.8975273    0.9329944  0.8645377  1.0000000
# acceleration  0.4233285 -0.5046834   -0.5438005 -0.6891955 -0.4168392
# year          0.5805410 -0.3456474   -0.3698552 -0.4163615 -0.3091199
# origin        0.5652088 -0.5689316   -0.6145351 -0.4551715 -0.5850054
#
# acceleration    year    origin
# mpg            0.4233285 0.5805410 0.5652088
# cylinders      -0.5046834 -0.3456474 -0.5689316
# displacement  -0.5438005 -0.3698552 -0.6145351
# horsepower     -0.6891955 -0.4163615 -0.4551715
# weight        -0.4168392 -0.3091199 -0.5850054
# acceleration   1.0000000 0.2903161 0.2127458
# year           0.2903161 1.0000000 0.1815277
# origin         0.2127458 0.1815277 1.0000000

```

- (c) Use the `lm()` function to perform a multiple linear regression with `mpg` as the response and all other variables except `name` as the predictors. Use the `summary()` function to print the results. Comment on the output. For instance:

```

linear_model <- lm(
  formula = mpg ~ cylinders + displacement + horsepower + weight + acceleration + year + origin,
  data = Auto
)
summarize_linear_model(linear_model)

```

```

#
# Call:
# lm(formula = mpg ~ cylinders + displacement + horsepower + weight +
#     acceleration + year + origin, data = Auto)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -9.5903 -2.1565 -0.1169  1.8690 13.0604
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept) -17.218435   4.644294  -3.707  0.00024 ***
# cylinders    -0.493376   0.323282  -1.526  0.12780
# displacement  0.019896   0.007515   2.647  0.00844 **
# horsepower   -0.016951   0.013787  -1.230  0.21963
# weight       -0.006474   0.000652  -9.929 < 2e-16 ***
# acceleration  0.080576   0.098845   0.815  0.41548
# year         0.750773   0.050973  14.729 < 2e-16 ***
# origin       1.426141   0.278136   5.127 4.67e-07 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 3.328 on 384 degrees of freedom
# Multiple R-squared:  0.8215, Adjusted R-squared:  0.8182
# F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16
#
# E(y | x) =
#      B_0 +
#      B_cylinders * cylinders +

```

```

# B_displacement * displacement +
# B_horsepower * horsepower +
# B_weight * weight +
# B_acceleration * acceleration +
# B_year * year +
# B_origin * origin
# E(y | x) =
# -17.2184346220176 +
# -0.493376318858497 * cylinders +
# 0.0198956437420168 * displacement +
# -0.0169511442274994 * horsepower +
# -0.00647404339744045 * weight +
# 0.0805758383248626 * acceleration +
# 0.750772677950313 * year +
# 1.42614049542315 * origin
# Number of observations: 392
# Estimated variance of errors: 11.0734701313546
# Prediction R2: 0.812860206366742
# Multiple R: 0.906354277576412 Adjusted R: 0.904557223498646
# Critical value t(alpha/2 = 0.05/2, DFR = 384): 1.64883142452157
# Critical value F(alpha = 0.05, DFR = 7, DFR = 384): 2.03343852566144

```

i. Is there a relationship between the predictors and the response?

```
analyze_variance_for_one_linear_model(linear_model)
```

```

# Analysis of Variance Table
#
# Response: mpg
#
#      Df Sum Sq Mean Sq  F value    Pr(>F)
# cylinders      1 14403.1  14403.1 1300.6838 < 2.2e-16 ***
# displacement  1  1073.3   1073.3  96.9293 < 2.2e-16 ***
# horsepower     1   403.4    403.4  36.4301 3.731e-09 ***
# weight         1   975.7    975.7  88.1137 < 2.2e-16 ***
# acceleration   1     1.0     1.0   0.0872  0.7679
# year           1  2419.1   2419.1 218.4609 < 2.2e-16 ***
# origin         1   291.1    291.1  26.2912 4.666e-07 ***
# Residuals     384  4252.2    11.1
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
# DFR: 7, SSR: 19566.7809389476, MSR: 2795.25441984966
# F0: 252.42804529132, F(alpha = 0.05, DFR = 7, DFR = 384): 2.03343852566144
# p: 2.03710593075355e-139
# DFT: 391, SST: 23818.9934693878
# R2: 0.82147807648106
# Adjusted R2: 0.81822377058358
# Prediction R2: 0.812860206366742
# Number of observations: 392

```

We assume that the errors for the above linear model are random, are independent, and follow a normal distribution with mean  $E(\epsilon_i) = 0$  and variance  $Var(\epsilon_i) = \sigma^2$ . There is a relationship between the predictors and the response if at least one of the predictor variables in the set  $\{cylinders, displacement, horsepower, weight, acceleration, year, origin\}$  contributes significantly to the model. We conduct a test of the null hypothesis  $H_0 : \beta_{cylinders} = \beta_{displacement} = \beta_{horsepower} = \beta_{weight} = \beta_{acceleration} = \beta_{year} = \beta_{origin} = 0$  that all coefficients in the set  $\{\beta_{cylinders}, \beta_{displacement}, \beta_{horsepower}, \beta_{weight}, \beta_{acceleration}, \beta_{year}, \beta_{origin}\}$  are 0. The

alternate hypothesis is  $H_1 : \beta_{cylinders} \neq 0 \text{ or } \beta_{displacement} \neq 0 \text{ or } \beta_{horsepower} \neq 0 \text{ or } \beta_{weight} \neq 0 \text{ or } \beta_{acceleration} \neq 0 \text{ or } \beta_{year} \neq 0 \text{ or } \beta_{origin} \neq 0$  that at least one coefficient in the set  $\{\beta_{cylinders}, \beta_{displacement}, \beta_{horsepower}, \beta_{weight}, \beta_{acceleration}, \beta_{year}, \beta_{origin}\}$  is not 0. The alternate hypothesis is also  $H_1 : \beta_i \neq 0 \text{ for } i \in \{\beta_{cylinders}, \beta_{displacement}, \beta_{horsepower}, \beta_{weight}, \beta_{acceleration}, \beta_{year}, \beta_{origin}\}$ . If we reject the null hypothesis, at least one of the predictor variables contributes significantly to the model.

```
test_null_hypothesis_involving_MLR_coefficients(linear_model, 0.05)
```

```
# Since probability 2.03710593075355e-139 is less than significance level 0.05,
# we reject the null hypothesis.
# We have sufficient evidence to support the alternate hypothesis.
```

Alternately, since the test statistic  $F_0 = 252.428$  is greater than the critical value  $F_{\alpha=0.05, df_R=p=7, df_{Res}=n-v=392-8} = 2.033$ , we reject the null hypothesis and support the alternate hypothesis. Since we reject the null hypothesis and support the alternate hypothesis, at least one of the predictors contributes significantly to the model. Since at least one of the predictors contributes significantly to the model, there is a relationship between the predictors and the response.

ii. Which predictors appear to have a statistically significant relationship to the response?

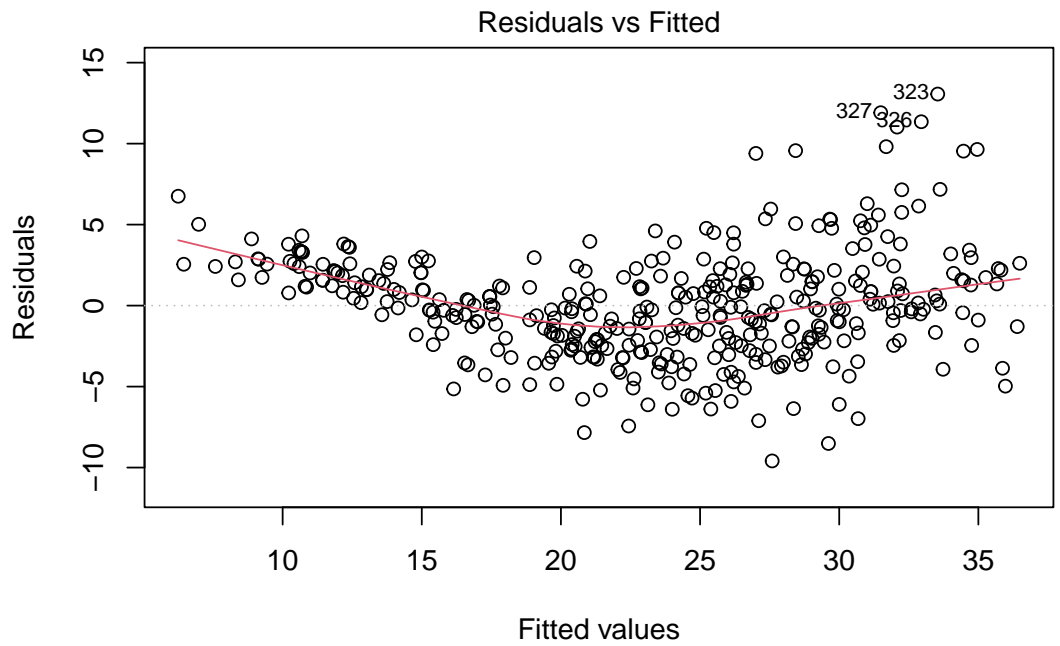
A critical value  $t_{\alpha/2=0.05, n-v=392-8} = 1.649$ . The summary for the above linear model provides test statistics for predictors. In parallel, the summary provides probabilities where each probability  $p$  is the probability that the magnitude  $|t|$  of a random test statistic is greater than the magnitude  $|t_0|$  of the appropriate test statistic. Because the magnitudes of the test statistic for displacement, weight, year, and origin are greater than the critical value, and the probabilities for these predictors are less than the significance level  $\alpha = 0.05$ , we reject null hypotheses that each individual predictor is insignificant in predicting the response in the context of the model and can be removed from the model. For these predictors we have sufficient evidence to support the alternate hypothesis that the predictor is significant in predicting the response in the context of the model and cannot be removed from the model.

iii. What does the coefficient for the year variable suggest?

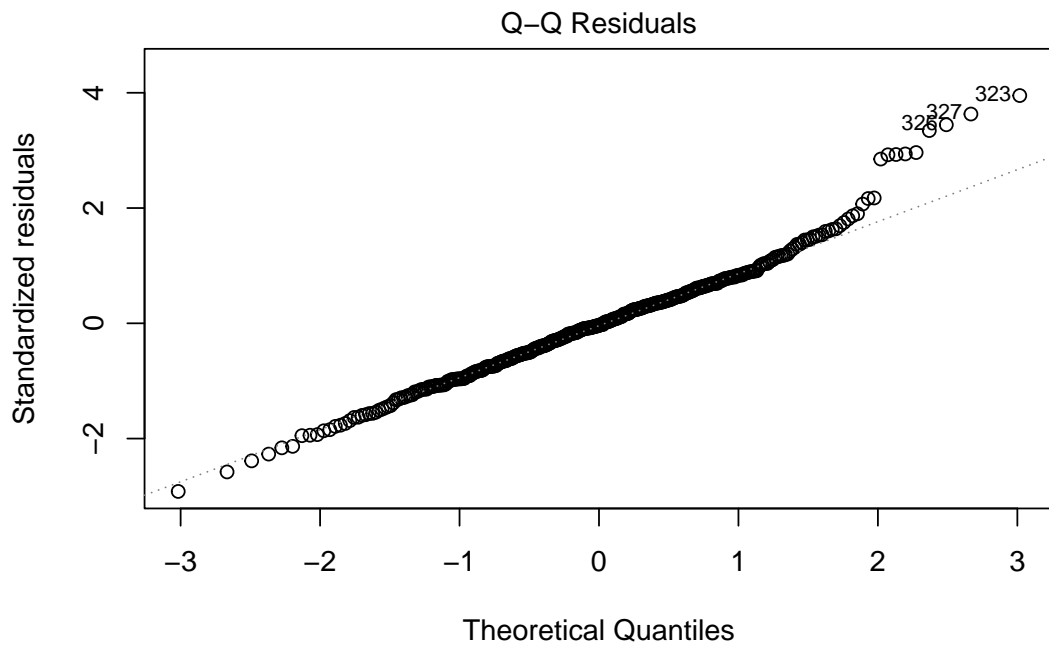
The coefficient for predictor year  $\beta_{year} = 0.751$  suggests that for every increase of 1 year, response miles per gallon *mpg* increases by 0.751, all other predictors being equal.

(d) Use the `plot()` function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?

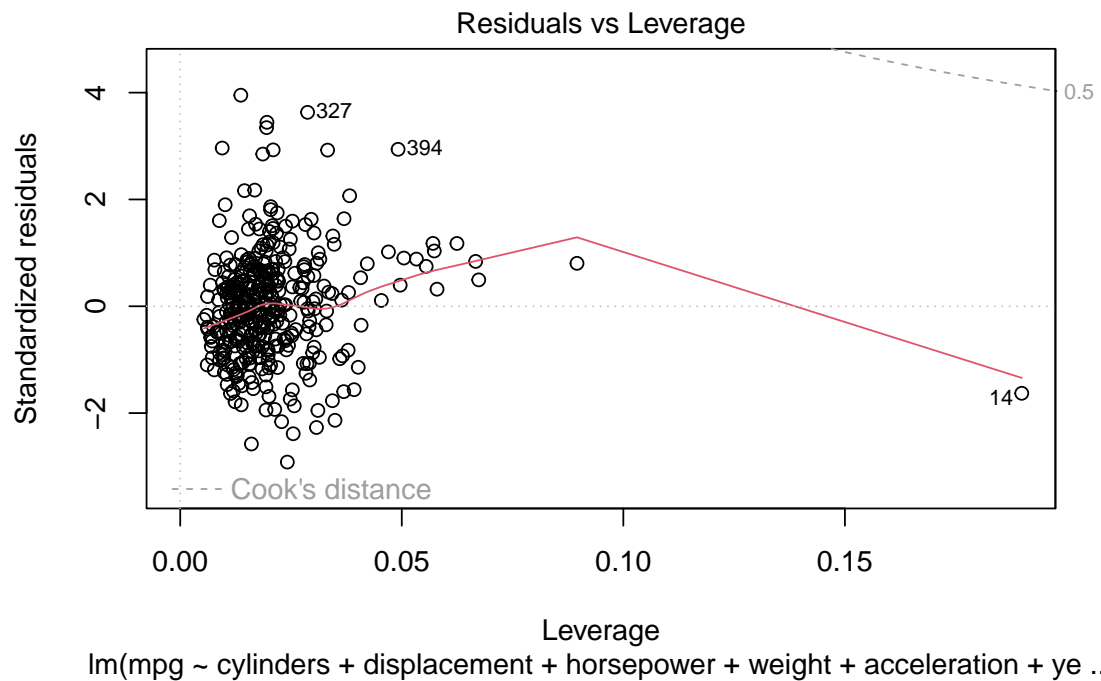
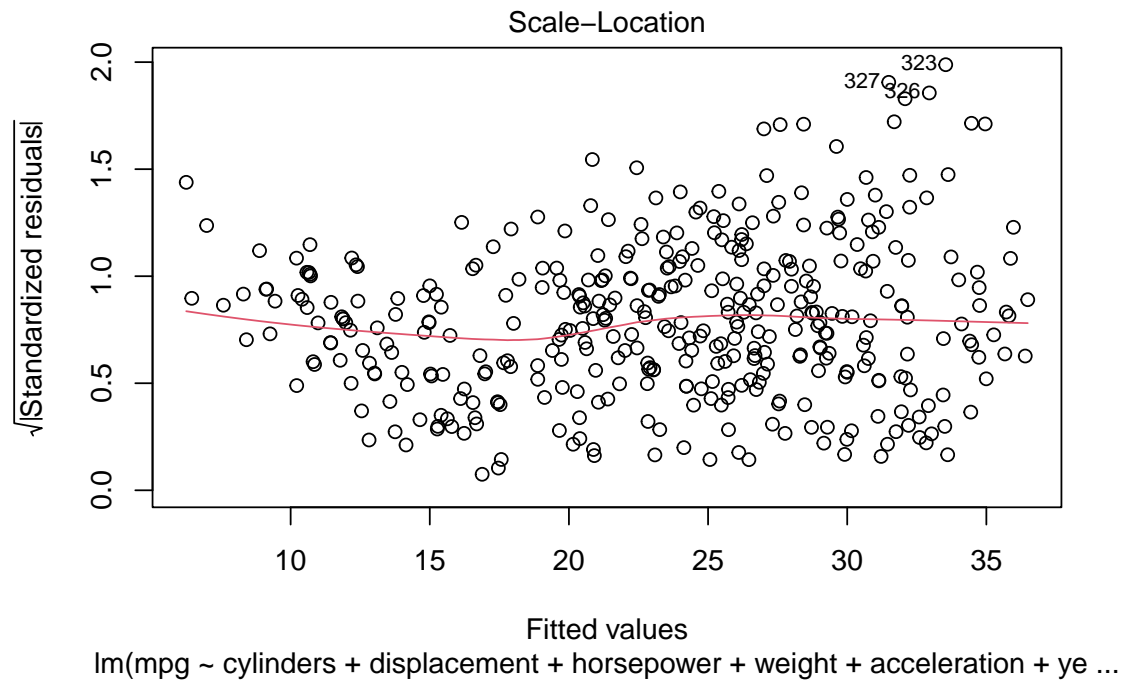
```
plot(linear_model)
```



lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...



lm(mpg ~ cylinders + displacement + horsepower + weight + acceleration + ye ...



According to <https://data.library.virginia.edu/diagnostic-plots/>, a plot of residuals vs. fitted values shows if residuals have non-linear patterns. There could be a non-linear relationship between predictor variables and an outcome variable, and the pattern could show up in this plot if the model doesn't capture the non-linear relationship. If you find equally spread residuals around

a horizontal line without distinct patterns, that is a good indication you don't have non-linear relationships."

Our plot of residuals vs. fitted values exhibits a "U" shape. According to [https://mathstat.slu.edu/~speegle/\\_book\\_summer\\_2021/SimpleReg.html](https://mathstat.slu.edu/~speegle/_book_summer_2021/SimpleReg.html), "A pattern such as a U-shape is evidence of a lurking variable that we have not taken into account. A lurking variable could be information related to the data that we did not collect, or it could be that our model should include a quadratic term."

According to <https://data.library.virginia.edu/diagnostic-plots/>, a normal Q-Q plot "shows if residuals are normally distributed. Do residuals follow a straight line well or do they deviate severely? It's good if residuals are lined well on the straight dashed line." According to <https://data.library.virginia.edu/understanding-q-q-plots/>, the QQ plot, or quantile-quantile plot, is a graphical tool to help us assess if a set of data plausibly came from some theoretical distribution such as a normal or exponential. For example, if we run a statistical analysis that assumes our residuals are normally distributed, we can use a normal QQ plot to check that assumption. It's just a visual check, not an air-tight proof, so it is somewhat subjective. But it allows us to see at-a-glance if our assumption is plausible, and if not, how the assumption is violated and what data points contribute to the violation."

"A QQ plot is a scatterplot created by plotting two sets of quantiles against one another. If both sets of quantiles came from the same distribution, we should see the points forming a line that's roughly straight."

"Now what are 'quantiles'? These are often referred to as 'percentiles'. These are points in your data below which a certain proportion of your data fall. For example, imagine the classic bell-curve standard normal distribution with a mean of 0. The 0.5 quantile, or 50th percentile, is 0. Half the data lie below 0. That's the peak of the hump in the curve. The 0.95 quantile, or 95th percentile, is about 1.64. 95 percent of the data lie below 1.64."

"So we see that quantiles are basically just your data sorted in ascending order, with various data points labelled as being the point below which a certain proportion of the data fall."

QQ plots plot standardized residuals "versus quantiles calculated from a theoretical distribution. The number of quantiles is selected to match the size of your sample data. While normal QQ plots are the ones most often used in practice due to so many statistical methods assuming normality, QQ plots can actually be created for any distribution."

"What about when points don't fall on a straight line? What can we infer about our data?" A QQ plot for which the upper right points deviate up from the line of the QQ plot indicates that your sample distribution has "heavy tails", which means that "your data has a larger number of extreme values than would be expected if they truly came from a normal distribution."

According to <https://data.library.virginia.edu/diagnostic-plots/>, a Scale / Spread - Location plot "shows if residuals are spread equally along the ranges of predictors. This is how you can check the assumption of equal variance (homoscedasticity). It's good if you see a horizontal line with equally (randomly) spread points." Residuals exhibit a "U" shape indicating that the multiple linear regression assumption that the variance of errors is constant is not met.

A plot of standardized residuals vs. leverage "helps us find influential cases (i.e., subjects) if there are any. Not all outliers are influential in linear regression analysis (whatever outliers mean). Even though data have extreme values, they might not be influential to determine a regression line. That means the results wouldn't be much different if we either include or exclude them from analysis. They follow the trend in the majority of cases and they don't really matter; they are not influential. On the other hand, some cases could be very influential even if they look to be within a reasonable range of the values. They could be extreme cases against a regression line and can alter the results if we exclude them from analysis. Another way to put it is that they don't get along with the trend in the majority of the cases."

“Unlike the other plots, this time patterns are not relevant. We watch out for outlying values at the upper right corner or at the lower right corner. Those spots are the places where cases can be influential against a regression line. Look for cases outside of the dashed lines. When cases are outside of the dashed lines (meaning they have high ‘Cook’s distance’ scores), the cases are influential to the regression results. The regression results will be altered if we exclude those cases.”

In our case where you “can barely see Cook’s distance lines” and “all cases are well inside of the Cook’s distance lines”, there are no influential cases. In a case where a point “is far beyond the Cook’s distance lines”, the point corresponds to an influential observation.

Our plot of standardized residuals vs. leverage shows a point corresponding to observation 14 with relatively high leverage.

“The four plots show potential problematic cases with the row numbers of the cases in the data set. If some cases are identified across all four plots, you might want to take a closer look at them individually. Is there anything special for the subject? Or could it be simply errors in data entry?” Observations 323, 326, and 327 are potentially problematic or outliers according to the plot of residuals vs. fitted values, the plot of standardized residuals vs. theoretical quantiles, and the Scale / Spread - Location plot. Observation 327 additionally is identified as potentially problematic or an outlier in the plot of standardized residuals vs. leverage. If we define outlier to being observation with a magnitude of standardized residual greater than 2, the plot of standardized residuals vs. leverage shows some outliers.

“In that case, you may want to go back to your theory and hypotheses. Is it really a linear relationship between the predictors and the outcome? You may want to include a quadratic term, for example. A log transformation may better represent the phenomena that you’d like to model. Or, is there any important variable that you left out from your model? Other variables you didn’t include (e.g., age or gender) may play an important role in your model and data. Or, maybe, your data were systematically biased when collecting data. You may want to redesign data collection methods.”

- (e) Use the `*` and `:` symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

Cylinders and displacement have the highest correlation at 0.951. Displacement and weight have the second highest correlation at 0.932. Let us consider a linear model involving cylinders, displacement, and a term representing interaction between cylinders and displacement. Let us consider a linear model involving displacement, weight, and a term representing interaction between displacement and weight. Let us consider a linear model involving cylinders, displacement, weight, an interaction term involving cylinders and displacement, and an interaction term involving displacement and weight.

For the linear model  $mpg = \beta_0 + \beta_1 \text{cylinders} + \beta_2 \text{displacement} + \beta_3 \text{cylinders displacement}$ , since the  $p$  value for each predictive term is less than significance level 0.05, each predictive term is statistically significant.

For the linear model  $mpg = \beta_0 + \beta_1 \text{displacement} + \beta_2 \text{weight} + \beta_3 \text{displacement weight}$ , since the  $p$  value for each predictive term is less than significance level 0.05, each predictive term is statistically significant.

For the linear model  $mpg = \beta_0 + \beta_1 \text{cylinders} + \beta_2 \text{displacement} + \beta_3 \text{weight} + \beta_4 \text{cylinders displacement} + \beta_5 \text{displacement weight}$ , since the  $p$  value for the predictive term *displacement weight* is less than significance level 0.05, this predictive term is statistically significant. Since the  $p$  value for the predictive term *cylinders displacement* is greater than significance level 0.05, each predictive term is statistically insignificant.

```
linear_model <- lm(formula = mpg ~ cylinders * displacement, data = Auto)
summarize_linear_model(linear_model)
```



```

#
# Call:
# lm(formula = mpg ~ cylinders * displacement, data = Auto)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -16.0432  -2.4308  -0.2263   2.2048  20.9051
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)    48.22040     2.34712   20.545 < 2e-16 ***
# cylinders       -2.41838     0.53456   -4.524 8.08e-06 ***
# displacement   -0.13436     0.01615   -8.321 1.50e-15 ***
# cylinders:displacement  0.01182     0.00207    5.711 2.24e-08 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 4.454 on 388 degrees of freedom
# Multiple R-squared:  0.6769, Adjusted R-squared:  0.6744
# F-statistic: 271 on 3 and 388 DF, p-value: < 2.2e-16
#
# E(y | x) =
#      B_0 +
#      B_cylinders * cylinders +
#      B_displacement * displacement +
#      B_cylinders:displacement * cylinders:displacement
# E(y | x) =
#      48.2204020561982 +
#      -2.41837563216276 * cylinders +
#      -0.134356254037509 * displacement +
#      0.011823203655611 * cylinders:displacement
# Number of observations: 392
# Estimated variance of errors: 19.8342000404213
# Prediction R2: 0.670557771549533
# Multiple R: 0.822745633347456 Adjusted R: 0.821226072133636
# Critical value t(alpha/2 = 0.05/2, DFRes = 388): 1.64879031767734
# Critical value F(alpha = 0.05, DFR = 3, DFRes = 388): 2.62790782140496

linear_model <- lm(
  formula = mpg ~ displacement + weight + displacement : weight,
  data = Auto
)
summarize_linear_model(linear_model)

#
# Call:
# lm(formula = mpg ~ displacement + weight + displacement:weight,
#     data = Auto)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -13.8664  -2.4801  -0.3355   1.8071  17.9429
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)

```

```

# (Intercept)          5.372e+01  1.940e+00  27.697 < 2e-16 ***
# displacement        -7.831e-02  1.131e-02  -6.922 1.85e-11 ***
# weight              -8.931e-03  8.474e-04 -10.539 < 2e-16 ***
# displacement:weight  1.744e-05  2.789e-06   6.253 1.06e-09 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 4.097 on 388 degrees of freedom
# Multiple R-squared:  0.7265, Adjusted R-squared:  0.7244
# F-statistic: 343.6 on 3 and 388 DF,  p-value: < 2.2e-16
#
# E(y | x) =
#   B_0 +
#   B_displacement * displacement +
#   B_weight * weight +
#   B_displacement:weight * displacement:weight
# E(y | x) =
#   53.7244024550286 +
#   -0.0783118584534707 * displacement +
#   -0.00893095090393231 * weight +
#   1.7441074399562e-05 * displacement:weight
# Number of observations: 392
# Estimated variance of errors: 16.7882883647365
# Prediction R2: 0.721774427738718
# Multiple R:  0.852365426401177    Adjusted R:  0.85112416061025
# Critical value t(alpha/2 = 0.05/2, DFRes = 388): 1.64879031767734
# Critical value F(alpha = 0.05, DFR = 3, DFRes = 388): 2.62790782140496

linear_model <- lm(
  mpg ~ cylinders * displacement + displacement * weight,
  data = Auto
)
summarize_linear_model(linear_model)

#
# Call:
# lm(formula = mpg ~ cylinders * displacement + displacement *
#   weight, data = Auto)
#
# Residuals:
#      Min       1Q   Median       3Q      Max
# -13.2934  -2.5184  -0.3476   1.8399  17.7723
#
# Coefficients:
#              Estimate Std. Error t value Pr(>|t|)
# (Intercept)    5.262e+01  2.237e+00  23.519 < 2e-16 ***
# cylinders       7.606e-01  7.669e-01   0.992  0.322
# displacement   -7.351e-02  1.669e-02  -4.403 1.38e-05 ***
# weight         -9.888e-03  1.329e-03  -7.438 6.69e-13 ***
# cylinders:displacement -2.986e-03  3.426e-03  -0.872  0.384
# displacement:weight  2.128e-05  5.002e-06   4.254 2.64e-05 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# Residual standard error: 4.103 on 386 degrees of freedom

```

```

# Multiple R-squared:  0.7272, Adjusted R-squared:  0.7237
# F-statistic: 205.8 on 5 and 386 DF,  p-value: < 2.2e-16
#
# E(y | x) =
#   B_0 +
#   B_cylinders * cylinders +
#   B_displacement * displacement +
#   B_weight * weight +
#   B_cylinders:displacement * cylinders:displacement +
#   B_displacement:weight * displacement:weight
# E(y | x) =
#   52.6234098286044 +
#   0.76064051251752 * cylinders +
#   -0.073512773408962 * displacement +
#   -0.00988816698951907 * weight +
#   -0.00298605097642129 * cylinders:displacement +
#   2.12774118909358e-05 * displacement:weight
# Number of observations: 392
# Estimated variance of errors: 16.8322704110105
# Prediction R2: 0.719690935816126
# Multiple R:  0.852774133189212    Adjusted R:  0.850699918163269
# Critical value t(alpha/2 = 0.05/2, DFRes = 386): 1.64881076434776
# Critical value F(alpha = 0.05, DFR = 5, DFRes = 386): 2.23737005310089

```

- (f) Try a few different transformations of the variables, such as  $\log(X)$ ,  $\sqrt{X}$ ,  $X^2$ . Comment on your findings.

## 14. This problem focuses on the collinearity problem.

- (a) Perform the following commands in R.

```

set.seed(1)
x1 = runif(100)
x2 = 0.5*x1 + rnorm(100)/10
y = 2 + 2*x1 + 0.3*x2 + rnorm(100)

```

The last line corresponds to creating a linear model in which  $y$  is a function of  $x_1$  and  $x_2$ . Write out the form of the linear model. What are the regression coefficients?

- (b) What is the correlation between  $x_1$  and  $x_2$ ? Create a scatterplot displaying the relationship between the variables.
- (c) Using this data, fit a least squares regression to predict  $y$  using  $x_1$  and  $x_2$ . Describe the results obtained. What are  $\hat{\beta}_0$ ,  $\hat{\beta}_1$ , and  $\hat{\beta}_2$ ? How do these relate to the true  $\beta_0$ ,  $\beta_1$ , and  $\beta_2$ ? Can you reject the null hypothesis  $H_0 : \beta_1 = 0$ ? How about the null hypothesis  $H_0 : \beta_2 = 0$ ?
- (d) Now fit a least squares regression to predict  $y$  using only  $x_1$ . Comment on your results. Can you reject the null hypothesis  $H_0 : \beta_1 = 0$ ?
- (e) Now fit a least squares regression to predict  $y$  using only  $x_2$ . Comment on your results. Can you reject the null hypothesis  $H_0 : \beta_2 = 0$ ?
- (f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.
- (g) Now suppose we obtain one additional observation, which was unfortunately mismeasured.

```

x1 <- c(x1, 0.1)
x2 <- c(x2, 0.8)

```

```
y <- c(y, 6)
```

Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

## 15. This problem involves the Boston data set, which we saw in the lab for this chapter.

We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

- For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.
- Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis  $H_0 : \beta_j = 0$ ?
- How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.
- Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor  $X$ , fit a model of the form

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \epsilon.$$