Week Two Exercises

Adeline Casali

2023-11-01

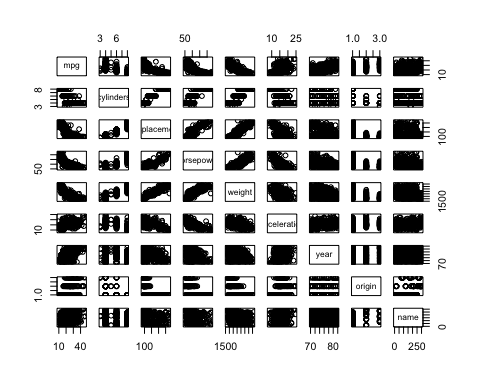
# Question 9

Loading packages that contain the dataset used

library(ISLR)

1. Produce a scatterplot matrix of all variables.

pairs(Auto)



b) Compute a matrix of correlations between all quantitative variables.

Auto\_correlation\_matrix <- cor(Auto[, -9])  
print(Auto\_correlation\_matrix)

## mpg cylinders displacement horsepower weight  
## mpg 1.0000000 -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

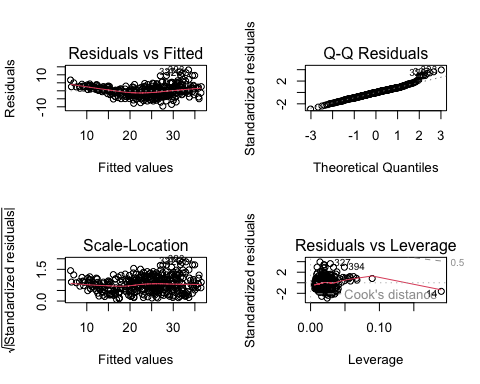
1. Perform a multiple linear regression with mpg.
2. There does seem to be a relationship between the predictors and the response. This is shown by the F-statistic’s extremely low p-value of <2.2e-16, suggesting that at least one of the predictors has a statistically significant impact on the response variable, mpg.
3. Displacement, weight, year, and origin all appear to have a statistically significant relationship to the response variable as they all have p-values less than 0.05.
4. The coefficient for the year variable, 0.750773, suggests that for each 1 year increase in the manufacturing year of the car, the expected change in mpg is an increase of approximately 0.751, holding all other predictors constant. This means that newer cars tend to have higher mpg values, and this effect is statistically significant as indicated by the p-value.

mpg\_lmodel <- lm(mpg ~ . - name, data = Auto)  
summary(mpg\_lmodel)

##   
## Call:  
## lm(formula = mpg ~ . - name, 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

1. When looking at the residuals plot, the residuals seem to be fairly evenly distributed with no clear pattern, which suggests linearity. However, the Q-Q residuals plot does show some deviation at its highest point, suggesting some skewness or outliers in the residuals. The scale-location plot shows minor funneling, which indicates some heteroscedasticity, suggesting that the variance of residuals is not constant. Finally, the residuals vs leverage plot does indicate a few observations with high leverage, with one observation having particularly high leverage.

par(mfrow=c(2,2))  
plot(mpg\_lmodel)



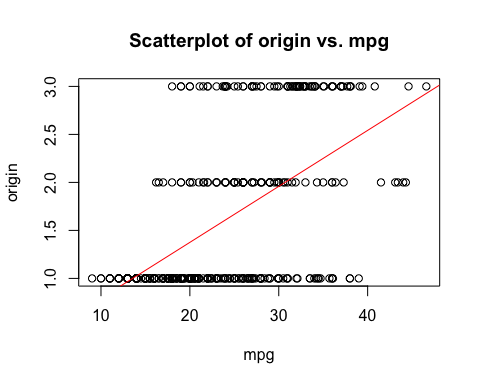
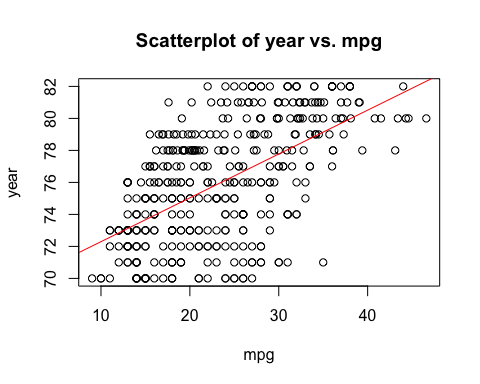
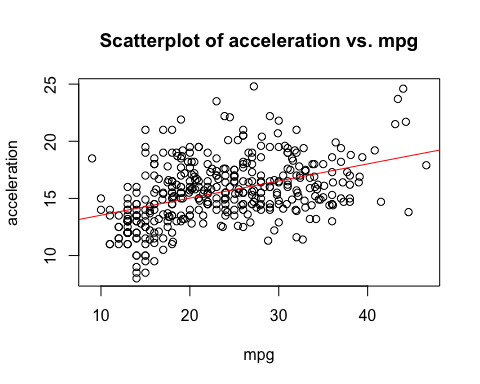
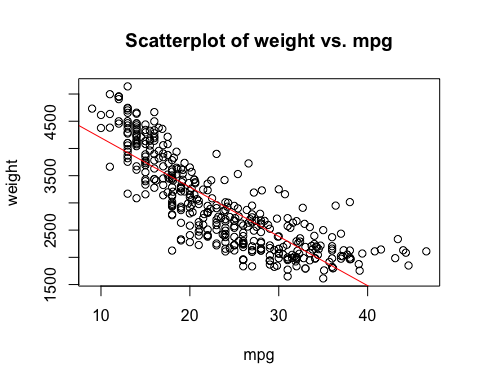
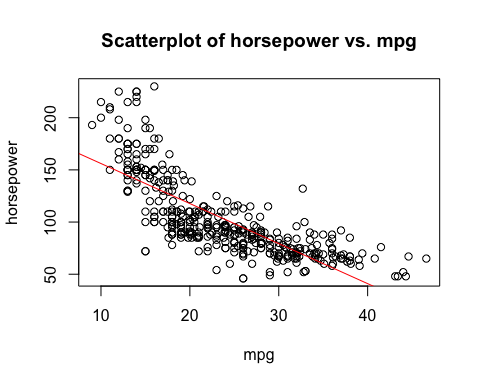
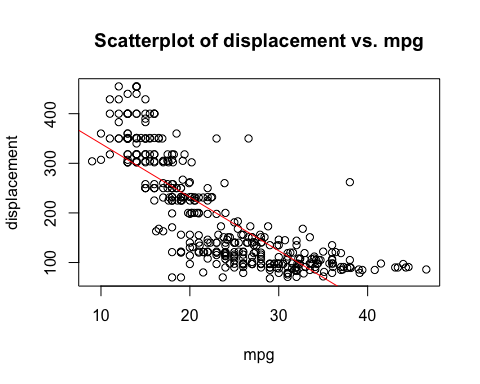
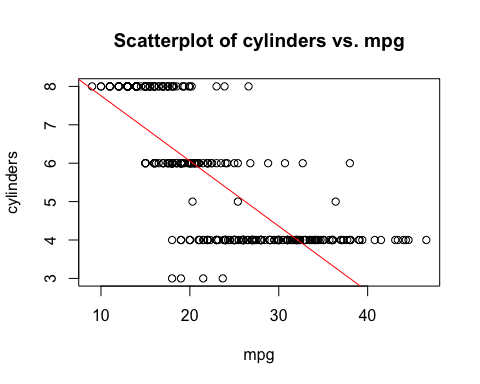
e) The interactions between displacement and year, acceleration and year, and acceleration and origin appear to be statistically significant, as indicated by their p-values of less than 0.05.

mpg\_lmodel\_interactions <- lm(mpg ~ (cylinders + displacement + horsepower + weight + acceleration + year + origin)^2, data = Auto)  
summary(mpg\_lmodel\_interactions)

##   
## Call:  
## lm(formula = mpg ~ (cylinders + displacement + horsepower + weight +   
## acceleration + year + origin)^2, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.6303 -1.4481 0.0596 1.2739 11.1386   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.548e+01 5.314e+01 0.668 0.50475   
## cylinders 6.989e+00 8.248e+00 0.847 0.39738   
## displacement -4.785e-01 1.894e-01 -2.527 0.01192 \*   
## horsepower 5.034e-01 3.470e-01 1.451 0.14769   
## weight 4.133e-03 1.759e-02 0.235 0.81442   
## acceleration -5.859e+00 2.174e+00 -2.696 0.00735 \*\*  
## year 6.974e-01 6.097e-01 1.144 0.25340   
## origin -2.090e+01 7.097e+00 -2.944 0.00345 \*\*  
## cylinders:displacement -3.383e-03 6.455e-03 -0.524 0.60051   
## cylinders:horsepower 1.161e-02 2.420e-02 0.480 0.63157   
## cylinders:weight 3.575e-04 8.955e-04 0.399 0.69000   
## cylinders:acceleration 2.779e-01 1.664e-01 1.670 0.09584 .   
## cylinders:year -1.741e-01 9.714e-02 -1.793 0.07389 .   
## cylinders:origin 4.022e-01 4.926e-01 0.816 0.41482   
## displacement:horsepower -8.491e-05 2.885e-04 -0.294 0.76867   
## displacement:weight 2.472e-05 1.470e-05 1.682 0.09342 .   
## displacement:acceleration -3.479e-03 3.342e-03 -1.041 0.29853   
## displacement:year 5.934e-03 2.391e-03 2.482 0.01352 \*   
## displacement:origin 2.398e-02 1.947e-02 1.232 0.21875   
## horsepower:weight -1.968e-05 2.924e-05 -0.673 0.50124   
## horsepower:acceleration -7.213e-03 3.719e-03 -1.939 0.05325 .   
## horsepower:year -5.838e-03 3.938e-03 -1.482 0.13916   
## horsepower:origin 2.233e-03 2.930e-02 0.076 0.93931   
## weight:acceleration 2.346e-04 2.289e-04 1.025 0.30596   
## weight:year -2.245e-04 2.127e-04 -1.056 0.29182   
## weight:origin -5.789e-04 1.591e-03 -0.364 0.71623   
## acceleration:year 5.562e-02 2.558e-02 2.174 0.03033 \*   
## acceleration:origin 4.583e-01 1.567e-01 2.926 0.00365 \*\*  
## year:origin 1.393e-01 7.399e-02 1.882 0.06062 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.695 on 363 degrees of freedom  
## Multiple R-squared: 0.8893, Adjusted R-squared: 0.8808   
## F-statistic: 104.2 on 28 and 363 DF, p-value: < 2.2e-16

1. After transformation, displacement, horsepower, and acceleration all show a statistically significant relationship with mpg, as shown by p-values less than 0.05. All three of those variables were not considered statistically significant in the original linear regression model, which indicates the impact that transformations can have.

# Make plots of all numeric variables vs mpg to check for linearity.   
variables\_to\_plot <- setdiff(names(Auto), c("name", "mpg"))  
  
for (variable in variables\_to\_plot) {  
 plot(Auto$mpg, Auto[, variable], xlab = "mpg", ylab = variable, main = paste("Scatterplot of", variable, "vs. mpg"))  
abline(lm(Auto[, variable] ~ Auto$mpg), col = "red")  
}



# Create a new linear model with transformations  
mpg\_lmodel\_transformed <- lm(mpg ~ cylinders + I(displacement^2) + log(horsepower) + weight + sqrt(acceleration) + year + origin, data = Auto)  
summary(mpg\_lmodel\_transformed)

##   
## Call:  
## lm(formula = mpg ~ cylinders + I(displacement^2) + log(horsepower) +   
## weight + sqrt(acceleration) + year + origin, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.4434 -1.9245 -0.0831 1.6375 12.2821   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.186e+01 9.474e+00 3.363 0.000848 \*\*\*  
## cylinders -6.083e-01 2.497e-01 -2.436 0.015303 \*   
## I(displacement^2) 5.536e-05 9.597e-06 5.769 1.65e-08 \*\*\*  
## log(horsepower) -9.741e+00 1.485e+00 -6.558 1.77e-10 \*\*\*  
## weight -4.590e-03 6.454e-04 -7.113 5.60e-12 \*\*\*  
## sqrt(acceleration) -1.844e+00 8.183e-01 -2.253 0.024819 \*   
## year 7.338e-01 4.693e-02 15.635 < 2e-16 \*\*\*  
## origin 1.310e+00 2.407e-01 5.440 9.49e-08 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.065 on 384 degrees of freedom  
## Multiple R-squared: 0.8486, Adjusted R-squared: 0.8458   
## F-statistic: 307.4 on 7 and 384 DF, p-value: < 2.2e-16

# Question 10

1. Producing a multiple regression model to predict Sales using Price, Urban, and US.

sales\_lmodel <- lm(Sales ~ Price + Urban + US, data = Carseats)

1. The model indicates that with every one-unit increase in price, sales are expected to decrease by 0.054459, and this effect is statistically significant. When located in an urban area, sales are expected to decrease by 0.021916, but this is not statistically significant as indicated by the high p-value. Finally, when in the US sales are expected to increase by 1.200573, and this is considered a statistically significant relationship.

summary(sales\_lmodel)

##   
## Call:  
## lm(formula = Sales ~ Price + Urban + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9206 -1.6220 -0.0564 1.5786 7.0581   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.043469 0.651012 20.036 < 2e-16 \*\*\*  
## Price -0.054459 0.005242 -10.389 < 2e-16 \*\*\*  
## UrbanYes -0.021916 0.271650 -0.081 0.936   
## USYes 1.200573 0.259042 4.635 4.86e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.472 on 396 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2335   
## F-statistic: 41.52 on 3 and 396 DF, p-value: < 2.2e-16

1. The model in equation form is written as: Sales = 13.043469 - 0.054459 \* Price - 0.021916 \* UrbanYes + 1.200573 \* USYes
2. You can reject the null hypothesis for the predictors Price and US as they have a p-value of less than 0.05, indicating a statistically significant effect on Sales.
3. Producing an updated model without the Urban predictor.

sales\_lmodel\_updated <- lm(Sales ~ Price + US, data = Carseats)  
summary(sales\_lmodel\_updated)

##   
## Call:  
## lm(formula = Sales ~ Price + US, data = Carseats)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.9269 -1.6286 -0.0574 1.5766 7.0515   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 13.03079 0.63098 20.652 < 2e-16 \*\*\*  
## Price -0.05448 0.00523 -10.416 < 2e-16 \*\*\*  
## USYes 1.19964 0.25846 4.641 4.71e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.469 on 397 degrees of freedom  
## Multiple R-squared: 0.2393, Adjusted R-squared: 0.2354   
## F-statistic: 62.43 on 2 and 397 DF, p-value: < 2.2e-16

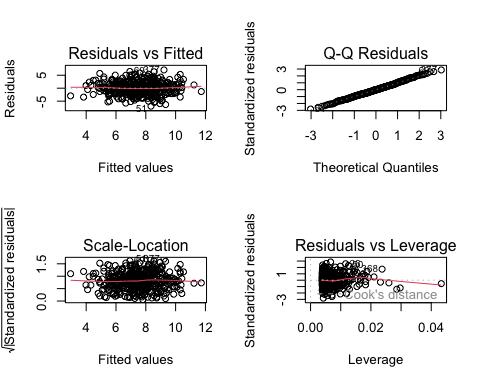
1. Both models have a highly significant F-statistic, as indicated by their low p-values, but the second model does contain a higher F-statistic value. However, when looking at the adjusted r-squared values for each model, the updated regression model has a higher value of 0.2354 compared to 0.2335 in the original model. Both models have the same multiple r-squared value of 0.2393, indicating that 23.93% of the variance in sales is explained by the predictors in this model. Finally, the residual standard error was reduced from 2.472 on 396 degrees of freedom to 2.469 on 397 degrees of freedom in the second model, suggesting a slight decrease in error.
2. The 95% confidence interval for Price is -0.06475984 to -0.04419543, and the 95% confidence interval for US is much wider at 0.69151957 to 1.70776632.

confint(sales\_lmodel\_updated)

## 2.5 % 97.5 %  
## (Intercept) 11.79032020 14.27126531  
## Price -0.06475984 -0.04419543  
## USYes 0.69151957 1.70776632

1. As seen in the plots, there does seem to be evidence of a point with high leverage, seen in the far right of the residuals vs leverage plot. However, there does not appear to be evidence of any outliers on any of the other plots.

par(mfrow = c(2, 2))  
plot(sales\_lmodel\_updated)



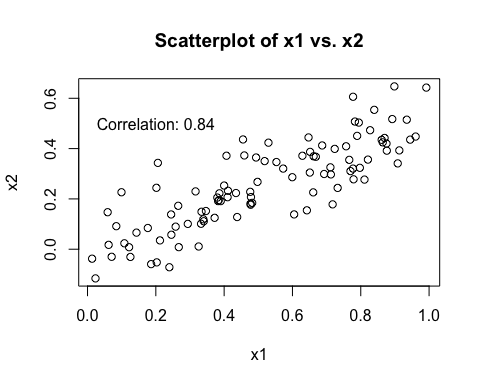
# Question 14

1. The linear model can be written as: y = B0 + B1*x1 + B2*x2 + e In this model, the regression coefficients are: B0 = 2 (intercept), B1 = 2 (coefficient for x1), and B2 = 0.3 (coefficient for x2)

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

1. The correlation between x1 and x2 is relatively high at 0.84. This is evident in the graph below.

x1\_x2\_cor <- cor(x1, x2)  
plot(x1, x2, main = "Scatterplot of x1 vs. x2", xlab = "x1", ylab = "x2")  
text(0.2, 0.5, paste("Correlation:", round(x1\_x2\_cor, 2)))

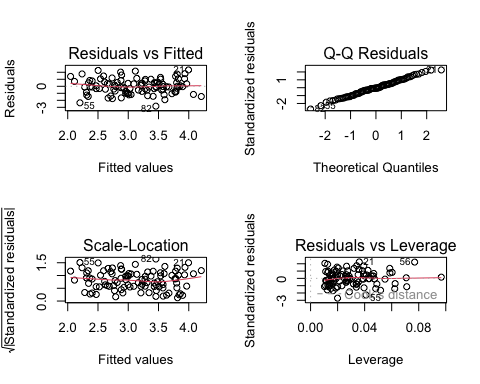


c) In this linear model, B0 = 2.13, B1 = 1.44, and B2 = 1.01. Although B0 is very close to the true B0 of 2, B1 and B2 are quite different from their true values of 2 and 0.3. According to the p-values produced by the model, we can reject the null hypothesis that B1 = 0, but we fail to reject the null hypothesis B2 = 0.

x1\_x2\_lmodel <- lm(y ~ x1 + x2)  
summary(x1\_x2\_lmodel)

##   
## Call:  
## lm(formula = y ~ x1 + x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8311 -0.7273 -0.0537 0.6338 2.3359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.1305 0.2319 9.188 7.61e-15 \*\*\*  
## x1 1.4396 0.7212 1.996 0.0487 \*   
## x2 1.0097 1.1337 0.891 0.3754   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.056 on 97 degrees of freedom  
## Multiple R-squared: 0.2088, Adjusted R-squared: 0.1925   
## F-statistic: 12.8 on 2 and 97 DF, p-value: 1.164e-05

par(mfrow = c(2, 2))  
plot(x1\_x2\_lmodel)

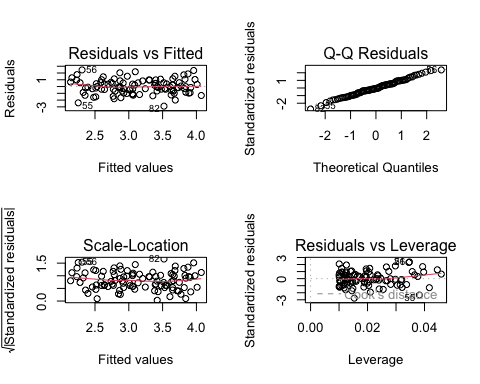


d) According to the linear model containing just x1, we can reject the null hypothesis B1 = 0 due to the extremely low p-value.

x1\_lmodel <- lm(y ~ x1)  
summary(x1\_lmodel)

##   
## Call:  
## lm(formula = y ~ x1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.89495 -0.66874 -0.07785 0.59221 2.45560   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.1124 0.2307 9.155 8.27e-15 \*\*\*  
## x1 1.9759 0.3963 4.986 2.66e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.055 on 98 degrees of freedom  
## Multiple R-squared: 0.2024, Adjusted R-squared: 0.1942   
## F-statistic: 24.86 on 1 and 98 DF, p-value: 2.661e-06

par(mfrow = c(2, 2))  
plot(x1\_lmodel)

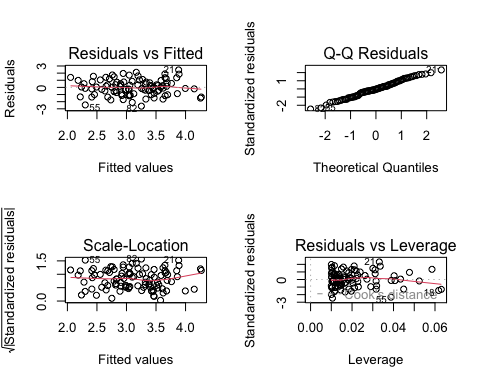


e) According to the linear model containing just x2, we can reject the null hypothesis B1 = 0 due to the extremely low p-value.

x2\_lmodel <- lm(y ~ x2)  
summary(x2\_lmodel)

##   
## Call:  
## lm(formula = y ~ x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.62687 -0.75156 -0.03598 0.72383 2.44890   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.3899 0.1949 12.26 < 2e-16 \*\*\*  
## x2 2.8996 0.6330 4.58 1.37e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.072 on 98 degrees of freedom  
## Multiple R-squared: 0.1763, Adjusted R-squared: 0.1679   
## F-statistic: 20.98 on 1 and 98 DF, p-value: 1.366e-05

par(mfrow = c(2, 2))  
plot(x2\_lmodel)



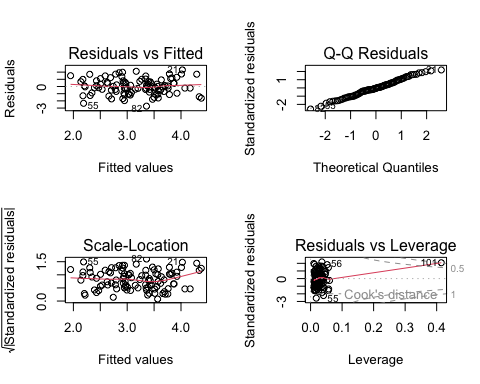
f) The results obtained from the linear models produced in (c) through (e) do contradict each other, as the linear model containing both variables shows a statistically significant relationship with x1 but not x2, and the models containing each variable individually shows a relationship with both. However, this is likely due to the collinearity between x1 and x2, as seen in the scatterplot above and the correlation value produced of 0.84. This can cause problems in the regression model because it can be difficult to separate out the responses of each individual variable on the response variable.

1. With the updated linear model containing both x1 and x2, we fail to reject the null hypothesis B1 = 0, but can reject the null hypothesis B2 = 0. This is the opposite of the original linear model. In the models containing just each individual predictor, we can still reject the null hypothesis B1 = 0 in both cases. This point appears to be both an outlier and a high leverage point as seen by the plots produced below. Each scatterplot clearly shows the point as an outlier, and the residuals vs leverage plots for each model produced show a point with high leverage that was not seen in the same plots for the models before the new observation was added. This indicates that the new observation is both an outlier and a high-leverage point.

x1 = c(x1, 0.1)  
x2 = c(x2, 0.8)  
y = c(y, 6)  
x1\_x2\_lmodel\_updated <- lm(y ~ x1 + x2)  
summary(x1\_x2\_lmodel\_updated)

##   
## Call:  
## lm(formula = y ~ x1 + x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.73348 -0.69318 -0.05263 0.66385 2.30619   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.2267 0.2314 9.624 7.91e-16 \*\*\*  
## x1 0.5394 0.5922 0.911 0.36458   
## x2 2.5146 0.8977 2.801 0.00614 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.075 on 98 degrees of freedom  
## Multiple R-squared: 0.2188, Adjusted R-squared: 0.2029   
## F-statistic: 13.72 on 2 and 98 DF, p-value: 5.564e-06

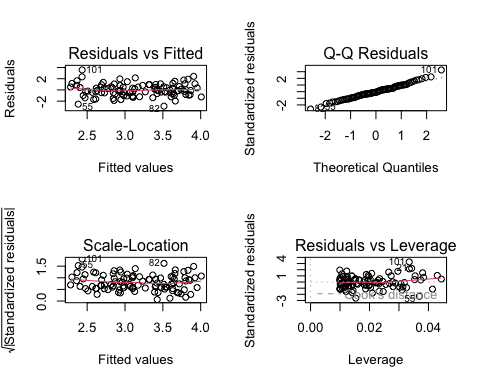
par(mfrow = c(2, 2))  
plot(x1\_x2\_lmodel\_updated)



x1\_lmodel\_updated <- lm(y ~ x1)  
summary(x1\_lmodel\_updated)

##   
## Call:  
## lm(formula = y ~ x1)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.8897 -0.6556 -0.0909 0.5682 3.5665   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.2569 0.2390 9.445 1.78e-15 \*\*\*  
## x1 1.7657 0.4124 4.282 4.29e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.111 on 99 degrees of freedom  
## Multiple R-squared: 0.1562, Adjusted R-squared: 0.1477   
## F-statistic: 18.33 on 1 and 99 DF, p-value: 4.295e-05

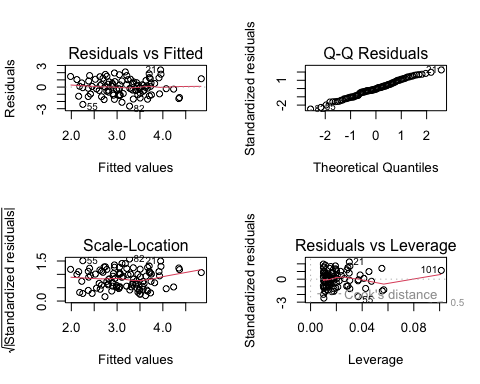
par(mfrow = c(2, 2))  
plot(x1\_lmodel\_updated)



x2\_lmodel\_updated <- lm(y ~ x2)  
summary(x2\_lmodel\_updated)

##   
## Call:  
## lm(formula = y ~ x2)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -2.64729 -0.71021 -0.06899 0.72699 2.38074   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.3451 0.1912 12.264 < 2e-16 \*\*\*  
## x2 3.1190 0.6040 5.164 1.25e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 1.074 on 99 degrees of freedom  
## Multiple R-squared: 0.2122, Adjusted R-squared: 0.2042   
## F-statistic: 26.66 on 1 and 99 DF, p-value: 1.253e-06

par(mfrow = c(2, 2))  
plot(x2\_lmodel\_updated)



plot(x1, y, main="Scatterplot of y vs. x1", xlab="x1", ylab="y")  
plot(x2, y, main="Scatterplot of y vs. x2", xlab="x2", ylab="y")  
plot(x1, x2, main="Scatterplot of x1 vs. x2", xlab="x1", ylab="x2")

