Linear Regression

Drazen, MAX, Gulkhan

**Changes to The Data**

Our group decided that the first thing we should do was change horsepower to a numeric variable versus a categorical variable. Having horsepower as a categorical variable didn’t make sense because it had 94 different values in it. That is too much variation to be a categorical variable. With the 94 different values, it is more appropriate for horsepower to be a numeric variable. Also, horsepower is measuring the output of something which would cause trouble in the model if it was a categorical variable. Before we could change horsepower to a numeric we had to take out the “?” that was presented in some rows. With that the change could easily be made.

Next our group decided to remove car name from the dataset and not use it. The “car name” variable had 305 unique names (categories). Many unique categories would not be useful when attempting to develop a model to predict mpg from a dataset with 398 observations. That is, it would be difficult to generalize from such a categorical variable because almost every mpg value is associated with a unique category. It would be difficult to find a pattern. If it was a numerical variable it would be easier for the model to find a pattern. In many cases the use of a simpler model is better.

We also removed columns origin and model year because they have no effect on predicting mpg. After these adjustments, the dataset was reduced to variables for mpg, weight, horsepower, displacement, and acceleration.

**R-Squared, Adjusted R-Squared, Formula**

**Model 1**

* *Multiple R-Squared***:** 0.7182
* *Adjusted R Squared***:** 0.7133
* y = 42.956 + -0.34 Cylinder + -0.01 Displacement + 0.009 Horsepower + -0.006 Weight+ 0.068 Acceleration

**Model 2**

* *Multiple R-Squared:* 0.6279
* *Adjusted R Squared:* 0.6266
* y = 43.396 + -3.635 Cylinder

**Model 3**

* *Multiple R-Squared:* 0.6661
* *Adjusted R Squared:* 0.6650
* y = 35.665 + -0.061 Displacement

**Model 4**

* *Multiple R-Squared:* 0.2233
* *Adjusted R Squared:* 0.2206
* y = 17.036 + 0.127 Horsepower

**Model 5**

* *Multiple R-Squared:* 0.7086
* *Adjusted R Squared:* 0.7076
* y = 46.581 + -0.008 Weight

**Model 6**

* *Multiple R-Squared:* 0.1797
* Adjusted R Squared: 0.1769
* y = 4.565 + 1.24 Acceleration

**Model 7**

* *Multiple R-Squared:* 0.6279
* *Adjusted R-Squared:* 0.6666
* y = 37.269 + -0.775 Cylinder + -0.049 Displacement

**Model 8**

* *Multiple R-Squared:* 0.6279
* *Adjusted R-Squared:* 0.6253
* y = 43.56+ -3.65 Cylinder + 0.002 Horsepower

**Model 9**

* *Multiple R-Squared:* 0.7159
* *Adjusted R-Squared:* 0.714
* y = 46.677 + -0.884 Cylinder + -0.006 Weight

**Model 10**

* *Multiple R-Squared:* 0.6279
* *Adjusted R-Squared:* 0.6253
* y = 43.621 + -3.644 Cylinder + -0.011 Acceleration

**Model 11**

* *Multiple R-Squared:* 0.6703
* *Adjusted R-Squared:* 0.6669
* y = 35.602 + -0.526 Cylinder + -0.051 Displacement + 0.013 Horsepower

**Model 12**

* *Multiple R-Squared:* 0.7173
* *Adjusted R Squared:* 0.7144
* y = 44.983 + -0.479 Cylinder + -0.011 Displacement + -0.006 Weight

**Model 13**

* *Multiple R-Squared:* 0.6721
* *Adjusted R-Squared:* 0.6687
* y = 40.905 + -0.735 Cylinder + -0.053 Displacement + -0.204 Acceleration

**Model 14**

* *Multiple R-Squared:* 0.7179
* *Adjusted R-Squared:* 0.714
* y = 43.894 + -0.327 Cylinder + -0.013 Displacement + 0.008 Horsepower + -0.006 Weight

**Model 15**

* *Multiple R-Squared:* 0.6737
* *Adjusted R-Squared:* 0.6692
* y = 39.24 + -0.471 Cylinder + -0.055 Displacement + 0.014 Horsepower + -0.209 Acceleration

**Model 16**

* *Multiple R-Squared:* 0.7173
* *Adjusted R-Squared:* 0.7134
* y = 43.811 + -0.682 Cylinder + 0.006 Horsepower + -0.006 Weight + 0.116 Acceleration

**Model 17**

* *Multiple R-Squared:* 0.7162
* *Adjusted R-Squared:* 0.7132
* y = 46.128 + -0.821 Cylinder + 0.005 Horsepower + -0.006 Weight

**Model 18**

* *Multiple R-Squared:* 0.717
* *Adjusted R-Squared:* 0.7141
* y = 44.428 + -0.75 Cylinder + -0.006 Weight + 0.114 Acceleration

**Model 19**

* *Multiple R-Squared:* 0.6693
* *Adjusted R-Squared:* 0.667
* y = 33.952 + -0.058 Displacement + 0.018 Horsepower

**Model 20**

* *Multiple R-Squared:* 0.7163
* *Adjusted R-Squared:* 0.7144
* y = 43.917 + -0.018 Displacement + -0.006 Weight

**Model 21**

* *Multiple R-Squared:* 0.6696
* *Adjusted R-Squared:* 0.6674
* y = 39.256 + -0.064 Displacement + -0.213 Acceleration

**Model 22**

* *Multiple R-Squared:* 0.7038
* *Adjusted R-Squared:* 0.7007
* y = 43.377 + -0.017 Displacement + 0.012 Horsepower + -0.006 Weight

**Model 23**

* *Multiple R-Squared:* 0.6729
* *Adjusted R-Squared:* 0.6695
* y = 37.869 + -0.062 Displacement + 0.018 Horsepower + -0.215 Acceleration

**Model 24**

* *Multiple R-Squared:* 0.7178
* *Adjusted R-Squared:* 0.7139
* y = 41.988 + -0.015 Displacement + 0.011 Horsepower + -0.006 Weight + 0.065 Acceleration

**Model 25**

* *Multiple R-Squared:* 0.7167
* *Adjusted R-Squared:* 0.7137
* y = 42.876 + -0.016 Displacement + -0.006 Weight + 0.071 Acceleration

**Model 26**

* *Multiple R-Squared:* 0.7107
* *Adjusted R-Squared:* 0.7087
* y = 45.085 + 0.014 Horsepower + -0.007 Weight

**Model 27**

* *Multiple R-Squared:* 0.3086
* *Adjusted R-Squared:* 0.3039
* y = 4.515 + 0.102 Horsepower + 0.897 Acceleration

**Model 28**

* *Multiple R-Squared:* 0.714
* *Adjusted R-Squared:* 0.711
* y = 41.64 + 0.012 Horsepower + -0.007 Weight + 0.187 Acceleration

**Model 29**

* *Multiple R-Squared:* 0.7124
* *Adjusted R-Squared:* 0.7105
* y = 42.658 + -0.007 Weight + 0.201 Acceleration

**Clean Data**

set.seed(234)  
df <- read.csv("project1.csv")  
df2 <- df  
df3 <- df2[df2$horsepower != "?",]  
df3$horsepower <- as.numeric(df3$horsepower)  
df3 <- df3[,- c(7,8,9)]  
samp <- sample(1:398, 300)  
train <- df3[samp, ]  
test <- df3[-samp, ]  
rownames(test) <- seq(length = nrow(test))  
train[!complete.cases(train),]

## mpg cylinder displacement horsepower weight acceleration  
## NA NA NA NA NA NA NA  
## NA.1 NA NA NA NA NA NA  
## NA.2 NA NA NA NA NA NA  
## NA.3 NA NA NA NA NA NA  
## NA.4 NA NA NA NA NA NA

train <- train[complete.cases(train),]

**Model 1**

model <- lm(mpg ~ ., data = train)  
summary(model)

##   
## Call:  
## lm(formula = mpg ~ ., data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.6614 -2.9376 -0.2851 2.6905 16.2482   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.9562333 2.7387379 15.685 < 2e-16 \*\*\*  
## cylinder -0.3384700 0.5077040 -0.667 0.506   
## displacement -0.0102612 0.0103563 -0.991 0.323   
## horsepower 0.0078017 0.0107885 0.723 0.470   
## weight -0.0057219 0.0008468 -6.757 7.75e-11 \*\*\*  
## acceleration 0.0682407 0.1193225 0.572 0.568   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.195 on 289 degrees of freedom  
## Multiple R-squared: 0.7182, Adjusted R-squared: 0.7133   
## F-statistic: 147.3 on 5 and 289 DF, p-value: < 2.2e-16

**Model 2**

model2 <- lm(mpg ~ cylinder , data = train)  
summary(model2)

##   
## Call:  
## lm(formula = mpg ~ cylinder, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.4914 -2.9218 -0.3173 2.6131 17.7434   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.3959 0.9290 46.72 <2e-16 \*\*\*  
## cylinder -3.6348 0.1635 -22.23 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.788 on 293 degrees of freedom  
## Multiple R-squared: 0.6279, Adjusted R-squared: 0.6266   
## F-statistic: 494.3 on 1 and 293 DF, p-value: < 2.2e-16

**Model 3**

model3 <- lm(mpg ~ displacement, data = train)  
summary(model3)

##   
## Call:  
## lm(formula = mpg ~ displacement, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.0961 -2.9706 -0.4115 2.7538 16.4796   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.364836 0.550233 64.27 <2e-16 \*\*\*  
## displacement -0.060982 0.002522 -24.18 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.536 on 293 degrees of freedom  
## Multiple R-squared: 0.6661, Adjusted R-squared: 0.665   
## F-statistic: 584.5 on 1 and 293 DF, p-value: < 2.2e-16

**Model 4**

model4 <- lm(mpg ~ horsepower, data = train)  
summary(model4)

##   
## Call:  
## lm(formula = mpg ~ horsepower, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.2496 -5.7238 -0.5536 4.8088 21.6803   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 17.03641 0.82970 20.533 <2e-16 \*\*\*  
## horsepower 0.12715 0.01385 9.178 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.918 on 293 degrees of freedom  
## Multiple R-squared: 0.2233, Adjusted R-squared: 0.2206   
## F-statistic: 84.23 on 1 and 293 DF, p-value: < 2.2e-16

**Model 5**

model5 <- lm(mpg ~ weight, data = train)  
summary(model5)

##   
## Call:  
## lm(formula = mpg ~ weight, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.1479 -2.7312 -0.4304 2.3916 16.3438   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 46.5810418 0.8921927 52.21 <2e-16 \*\*\*  
## weight -0.0077369 0.0002898 -26.69 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.237 on 293 degrees of freedom  
## Multiple R-squared: 0.7086, Adjusted R-squared: 0.7076   
## F-statistic: 712.5 on 1 and 293 DF, p-value: < 2.2e-16

**Model 6**

model6 <- lm(mpg ~ acceleration, train)  
summary(model6)

##   
## Call:  
## lm(formula = mpg ~ acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -18.473 -5.401 -1.282 4.751 22.947   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.5653 2.4231 1.884 0.0605 .   
## acceleration 1.2382 0.1545 8.012 2.69e-14 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.109 on 293 degrees of freedom  
## Multiple R-squared: 0.1797, Adjusted R-squared: 0.1769   
## F-statistic: 64.19 on 1 and 293 DF, p-value: 2.691e-14

**Model 7**

model7 <- lm(mpg ~ cylinder + displacement , data = train)  
summary(model7)

##   
## Call:  
## lm(formula = mpg ~ cylinder + displacement, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5147 -2.7835 -0.4172 2.5724 16.6442   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.268583 1.345261 27.704 < 2e-16 \*\*\*  
## cylinder -0.775319 0.500186 -1.550 0.122   
## displacement -0.048970 0.008147 -6.011 5.5e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.525 on 292 degrees of freedom  
## Multiple R-squared: 0.6688, Adjusted R-squared: 0.6666   
## F-statistic: 294.9 on 2 and 292 DF, p-value: < 2.2e-16

**Model 8**

model8 <- lm(mpg ~ cylinder + horsepower , data = train)  
summary(model8)

##   
## Call:  
## lm(formula = mpg ~ cylinder + horsepower, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.4779 -2.9534 -0.3156 2.6185 17.7359   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.560434 1.595901 27.295 <2e-16 \*\*\*  
## cylinder -3.650454 0.204875 -17.818 <2e-16 \*\*\*  
## horsepower -0.001525 0.012018 -0.127 0.899   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.796 on 292 degrees of freedom  
## Multiple R-squared: 0.6279, Adjusted R-squared: 0.6253   
## F-statistic: 246.3 on 2 and 292 DF, p-value: < 2.2e-16

**Model 9**

model9 <- lm(mpg ~ cylinder + weight, data = train)  
summary(model9)

##   
## Call:  
## lm(formula = mpg ~ cylinder + weight, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.9635 -2.6768 -0.2135 2.3541 16.4346   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 46.6767728 0.8831297 52.854 <2e-16 \*\*\*  
## cylinder -0.8841954 0.3225812 -2.741 0.0065 \*\*   
## weight -0.0061491 0.0006463 -9.514 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.191 on 292 degrees of freedom  
## Multiple R-squared: 0.7159, Adjusted R-squared: 0.714   
## F-statistic: 367.9 on 2 and 292 DF, p-value: < 2.2e-16

**Model 10**

model10 <- lm(mpg ~ cylinder + acceleration , data = train)  
summary(model10)

##   
## Call:  
## lm(formula = mpg ~ cylinder + acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.5363 -2.9156 -0.3764 2.6124 17.7574   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.62073 2.64768 16.475 <2e-16 \*\*\*  
## cylinder -3.64432 0.19434 -18.752 <2e-16 \*\*\*  
## acceleration -0.01122 0.12374 -0.091 0.928   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.796 on 292 degrees of freedom  
## Multiple R-squared: 0.6279, Adjusted R-squared: 0.6253   
## F-statistic: 246.3 on 2 and 292 DF, p-value: < 2.2e-16

**Model 11**

model11 <- lm(mpg ~ cylinder + displacement + horsepower , data = train)  
summary(model11)

##   
## Call:  
## lm(formula = mpg ~ cylinder + displacement + horsepower, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5921 -2.7804 -0.4935 2.5553 16.6648   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 35.601908 1.988973 17.900 < 2e-16 \*\*\*  
## cylinder -0.525767 0.545980 -0.963 0.336   
## displacement -0.050933 0.008324 -6.119 3.04e-09 \*\*\*  
## horsepower 0.013172 0.011583 1.137 0.256   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.522 on 291 degrees of freedom  
## Multiple R-squared: 0.6703, Adjusted R-squared: 0.6669   
## F-statistic: 197.2 on 3 and 291 DF, p-value: < 2.2e-16

**Model 12**

model12 <- lm(mpg ~ cylinder + displacement + weight, data = train)  
summary(model12)

##   
## Call:  
## lm(formula = mpg ~ cylinder + displacement + weight, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.8782 -2.7172 -0.2478 2.6511 16.3017   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 44.9832306 1.6558369 27.166 < 2e-16 \*\*\*  
## cylinder -0.4794404 0.4647885 -1.032 0.303   
## displacement -0.0111737 0.0092443 -1.209 0.228   
## weight -0.0055954 0.0007918 -7.067 1.18e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.187 on 291 degrees of freedom  
## Multiple R-squared: 0.7173, Adjusted R-squared: 0.7144   
## F-statistic: 246.2 on 3 and 291 DF, p-value: < 2.2e-16

**Model 13**

model13 <- lm(mpg ~ cylinder + displacement + acceleration, data = train)  
summary(model13)

##   
## Call:  
## lm(formula = mpg ~ cylinder + displacement + acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.2582 -2.7390 -0.4369 2.3902 16.8177   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 40.905270 2.527152 16.186 < 2e-16 \*\*\*  
## cylinder -0.734794 0.499152 -1.472 0.1421   
## displacement -0.052624 0.008401 -6.264 1.35e-09 \*\*\*  
## acceleration -0.204367 0.120374 -1.698 0.0906 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.51 on 291 degrees of freedom  
## Multiple R-squared: 0.6721, Adjusted R-squared: 0.6687   
## F-statistic: 198.8 on 3 and 291 DF, p-value: < 2.2e-16

**Model 14**

model14 <- lm(mpg ~ cylinder + displacement + horsepower + weight , data = train)  
summary(model14)

##   
## Call:  
## lm(formula = mpg ~ cylinder + displacement + horsepower + weight,   
## data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.9308 -2.8017 -0.2769 2.6852 16.3169   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.8938626 2.1912499 20.031 < 2e-16 \*\*\*  
## cylinder -0.3267243 0.5066995 -0.645 0.520   
## displacement -0.0126623 0.0094562 -1.339 0.182   
## horsepower 0.0081725 0.0107565 0.760 0.448   
## weight -0.0055553 0.0007941 -6.995 1.83e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.19 on 290 degrees of freedom  
## Multiple R-squared: 0.7179, Adjusted R-squared: 0.714   
## F-statistic: 184.5 on 4 and 290 DF, p-value: < 2.2e-16

**Model 15**

model15 <- lm(mpg ~ cylinder + displacement + horsepower + acceleration, data = train)  
summary(model15)

##   
## Call:  
## lm(formula = mpg ~ cylinder + displacement + horsepower + acceleration,   
## data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.3573 -2.7991 -0.4779 2.4081 16.8436   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.239519 2.882132 13.615 < 2e-16 \*\*\*  
## cylinder -0.471444 0.544989 -0.865 0.3877   
## displacement -0.054775 0.008584 -6.381 6.96e-10 \*\*\*  
## horsepower 0.013849 0.011550 1.199 0.2315   
## acceleration -0.209235 0.120352 -1.739 0.0832 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.507 on 290 degrees of freedom  
## Multiple R-squared: 0.6737, Adjusted R-squared: 0.6692   
## F-statistic: 149.7 on 4 and 290 DF, p-value: < 2.2e-16

**Model 16**

model16 <- lm(mpg ~ cylinder + horsepower + weight + acceleration , data = train)  
summary(model16)

##   
## Call:  
## lm(formula = mpg ~ cylinder + horsepower + weight + acceleration,   
## data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.5218 -2.8046 -0.3471 2.5488 16.2938   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.8109532 2.5992519 16.855 <2e-16 \*\*\*  
## cylinder -0.6817668 0.3710897 -1.837 0.0672 .   
## horsepower 0.0055202 0.0105396 0.524 0.6008   
## weight -0.0062556 0.0006534 -9.574 <2e-16 \*\*\*  
## acceleration 0.1161688 0.1090744 1.065 0.2877   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.195 on 290 degrees of freedom  
## Multiple R-squared: 0.7173, Adjusted R-squared: 0.7134   
## F-statistic: 183.9 on 4 and 290 DF, p-value: < 2.2e-16

**Model 17**

model17 <- lm(mpg ~ cylinder + horsepower + weight , data = train)  
summary(model17)

##   
## Call:  
## lm(formula = mpg ~ cylinder + horsepower + weight, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.0041 -2.6431 -0.3039 2.3668 16.4556   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 46.1284408 1.4220569 32.438 <2e-16 \*\*\*  
## cylinder -0.8214789 0.3472103 -2.366 0.0186 \*   
## horsepower 0.0051882 0.0105374 0.492 0.6228   
## weight -0.0061705 0.0006486 -9.513 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.196 on 291 degrees of freedom  
## Multiple R-squared: 0.7162, Adjusted R-squared: 0.7132   
## F-statistic: 244.7 on 3 and 291 DF, p-value: < 2.2e-16

**Model 18**

model18 <- lm(mpg ~ cylinder + weight + acceleration, data = train)  
summary(model18)

##   
## Call:  
## lm(formula = mpg ~ cylinder + weight + acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.4856 -2.6230 -0.2868 2.5204 16.2739   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 44.427573 2.314450 19.196 <2e-16 \*\*\*  
## cylinder -0.750471 0.346699 -2.165 0.0312 \*   
## weight -0.006232 0.000651 -9.573 <2e-16 \*\*\*  
## acceleration 0.114479 0.108891 1.051 0.2940   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.19 on 291 degrees of freedom  
## Multiple R-squared: 0.717, Adjusted R-squared: 0.7141   
## F-statistic: 245.8 on 3 and 291 DF, p-value: < 2.2e-16

**Model 19**

model19 <- lm(mpg ~ displacement + horsepower, data = train)  
summary(model19)

##   
## Call:  
## lm(formula = mpg ~ displacement + horsepower, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.3804 -2.8884 -0.3907 2.7201 16.5782   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 33.952196 1.010413 33.602 <2e-16 \*\*\*  
## displacement -0.058431 0.002945 -19.842 <2e-16 \*\*\*  
## horsepower 0.017655 0.010605 1.665 0.097 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.522 on 292 degrees of freedom  
## Multiple R-squared: 0.6693, Adjusted R-squared: 0.667   
## F-statistic: 295.4 on 2 and 292 DF, p-value: < 2.2e-16

**Model 20**

model20 <- lm(mpg ~ displacement + weight , data = train)  
summary(model20)

##   
## Call:  
## lm(formula = mpg ~ displacement + weight, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.6131 -2.7603 -0.3726 2.6143 16.1962   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 43.9169944 1.2937262 33.946 < 2e-16 \*\*\*  
## displacement -0.0180438 0.0064115 -2.814 0.00522 \*\*   
## weight -0.0056689 0.0007887 -7.188 5.52e-12 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.188 on 292 degrees of freedom  
## Multiple R-squared: 0.7163, Adjusted R-squared: 0.7144   
## F-statistic: 368.6 on 2 and 292 DF, p-value: < 2.2e-16

**Model 21**

model21 <- lm(mpg ~ displacement + acceleration, data = train)  
summary(model21)

##   
## Call:  
## lm(formula = mpg ~ displacement + acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.8933 -2.9028 -0.4458 2.5646 16.6693   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.255949 2.269737 17.295 <2e-16 \*\*\*  
## displacement -0.064133 0.003082 -20.810 <2e-16 \*\*\*  
## acceleration -0.212841 0.120476 -1.767 0.0783 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.519 on 292 degrees of freedom  
## Multiple R-squared: 0.6696, Adjusted R-squared: 0.6674   
## F-statistic: 295.9 on 2 and 292 DF, p-value: < 2.2e-16

**Model 22**

model22 <- lm(mpg ~ displacement + horsepower + weight, data = train)  
summary(model22)

##   
## Call:  
## lm(formula = mpg ~ displacement + horsepower + weight, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.7962 -2.8637 -0.3386 2.5642 16.2615   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.9148450 1.5783881 27.189 < 2e-16 \*\*\*  
## displacement -0.0171085 0.0064644 -2.647 0.00857 \*\*   
## horsepower 0.0109239 0.0098641 1.107 0.26902   
## weight -0.0055840 0.0007921 -7.050 1.31e-11 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.186 on 291 degrees of freedom  
## Multiple R-squared: 0.7175, Adjusted R-squared: 0.7146   
## F-statistic: 246.4 on 3 and 291 DF, p-value: < 2.2e-16

**Model 23**

model23 <- lm(mpg ~ displacement + horsepower + acceleration, data = train)  
summary(model23)

##   
## Call:  
## lm(formula = mpg ~ displacement + horsepower + acceleration,   
## data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -14.1900 -2.8181 -0.4396 2.5596 16.7713   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 37.868894 2.406496 15.736 <2e-16 \*\*\*  
## displacement -0.061585 0.003421 -18.000 <2e-16 \*\*\*  
## horsepower 0.017875 0.010566 1.692 0.0917 .   
## acceleration -0.215204 0.120102 -1.792 0.0742 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.505 on 291 degrees of freedom  
## Multiple R-squared: 0.6729, Adjusted R-squared: 0.6695   
## F-statistic: 199.5 on 3 and 291 DF, p-value: < 2.2e-16

**Model 24**

model24 <- lm(mpg ~ displacement + horsepower + weight + acceleration, data = train)  
summary(model24)

##   
## Call:  
## lm(formula = mpg ~ displacement + horsepower + weight + acceleration,   
## data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.5349 -2.9489 -0.3519 2.5645 16.1941   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 41.9878946 2.3196093 18.101 < 2e-16 \*\*\*  
## displacement -0.0149730 0.0075626 -1.980 0.0487 \*   
## horsepower 0.0106648 0.0098874 1.079 0.2817   
## weight -0.0057438 0.0008453 -6.795 6.16e-11 \*\*\*  
## acceleration 0.0650227 0.1191105 0.546 0.5856   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.191 on 290 degrees of freedom  
## Multiple R-squared: 0.7178, Adjusted R-squared: 0.7139   
## F-statistic: 184.4 on 4 and 290 DF, p-value: < 2.2e-16

**Model 25**

model25 <- lm(mpg ~ displacement + weight + acceleration, data = train)  
summary(model25)

##   
## Call:  
## lm(formula = mpg ~ displacement + weight + acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.3318 -2.8101 -0.3498 2.6186 16.1241   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.8760860 2.1691388 19.766 < 2e-16 \*\*\*  
## displacement -0.0156814 0.0075362 -2.081 0.0383 \*   
## weight -0.0058417 0.0008407 -6.949 2.42e-11 \*\*\*  
## acceleration 0.0711908 0.1190066 0.598 0.5502   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.193 on 291 degrees of freedom  
## Multiple R-squared: 0.7167, Adjusted R-squared: 0.7137   
## F-statistic: 245.3 on 3 and 291 DF, p-value: < 2.2e-16

**Model 26**

model26 <- lm(mpg ~ horsepower + weight, data = train)  
summary(model26)

##   
## Call:  
## lm(formula = mpg ~ horsepower + weight, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.4199 -2.9152 -0.2315 2.2259 16.4194   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.0848059 1.3625150 33.089 <2e-16 \*\*\*  
## horsepower 0.0143345 0.0098795 1.451 0.148   
## weight -0.0074848 0.0003375 -22.180 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.229 on 292 degrees of freedom  
## Multiple R-squared: 0.7107, Adjusted R-squared: 0.7087   
## F-statistic: 358.7 on 2 and 292 DF, p-value: < 2.2e-16

**Model 27**

model27 <- lm(mpg ~ horsepower + acceleration, data = train)  
summary(model27)

##   
## Call:  
## lm(formula = mpg ~ horsepower + acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -16.2780 -4.7049 -0.8519 4.0141 21.2023   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 4.51488 2.22836 2.026 0.0437 \*   
## horsepower 0.10159 0.01377 7.378 1.67e-12 \*\*\*  
## acceleration 0.89719 0.14945 6.003 5.73e-09 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 6.538 on 292 degrees of freedom  
## Multiple R-squared: 0.3086, Adjusted R-squared: 0.3039   
## F-statistic: 65.17 on 2 and 292 DF, p-value: < 2.2e-16

**Model 28**

model28 <- lm(mpg ~ horsepower + weight + acceleration, data = train)  
summary(model28)

##   
## Call:  
## lm(formula = mpg ~ horsepower + weight + acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.8034 -2.8401 -0.3815 2.4882 16.1688   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 41.6398716 2.3245143 17.913 <2e-16 \*\*\*  
## horsepower 0.0123648 0.0098993 1.249 0.213   
## weight -0.0072620 0.0003576 -20.308 <2e-16 \*\*\*  
## acceleration 0.1870072 0.1024476 1.825 0.069 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.212 on 291 degrees of freedom  
## Multiple R-squared: 0.714, Adjusted R-squared: 0.711   
## F-statistic: 242.1 on 3 and 291 DF, p-value: < 2.2e-16

**Model 29**

model29 <- lm(mpg ~ weight + acceleration, data = train)  
summary(model29)

##   
## Call:  
## lm(formula = mpg ~ weight + acceleration, data = train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.5256 -2.6818 -0.3814 2.5598 16.0857   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 42.6582393 2.1789293 19.578 <2e-16 \*\*\*  
## weight -0.0074602 0.0003208 -23.258 <2e-16 \*\*\*  
## acceleration 0.2009552 0.1019348 1.971 0.0496 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.216 on 292 degrees of freedom  
## Multiple R-squared: 0.7124, Adjusted R-squared: 0.7105   
## F-statistic: 361.7 on 2 and 292 DF, p-value: < 2.2e-16

**Best Models**

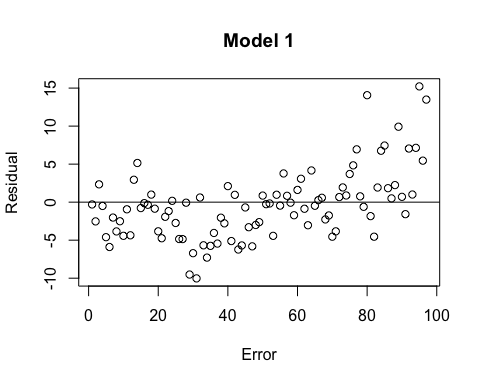
When looking over all 29 models that were developed, the multiple R-squared and the adjusted R-squared were used to select the best four models. The best four models were model 1, model 14, model 20, and model 22. All of these models had a multiple R-squared of.7160 or higher and an adjusted R-squared of .7140 or higher. Some of the other models had a high R-squared but had a lower adjusted R-squared which showed that there were some variables that were in the model that actually were not helping the model predict mpg accurately because adjusted R-squared punishes the model when there are variables in it that weren’t helpful at predicting the dependent variable.

**Plots**

In order to select the best model from the four models on the “short list” of candidate models a residual plot and a histogram was produced for each of the four models. In the histograms, the red line is the distribution of the error residuals and the blue line is the ideal normal distribution.

**Model 1 Residual Plot**

model\_pred <- predict(model, test, type = "response")  
test\_cars <- test[,1]  
error\_model <- test\_cars - model\_pred   
plot(error\_model, xlab = "Error", ylab = "Residual", main = "Model 1")  
abline(0,0)

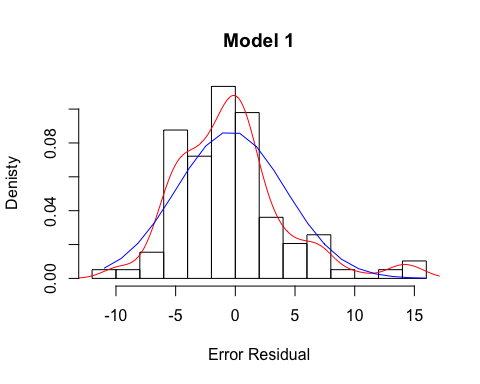


summary(error\_model)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -10.0272 -3.8352 -0.4983 -0.3770 1.0043 15.2226

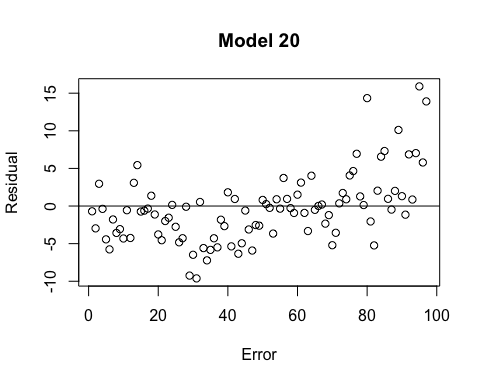
**Model 1 Histogram**

hist(error\_model, prob = T, breaks = 10, xlab = "Error Residual", ylab = "Denisty", main = "Model 1")  
lines(density(error\_model), col = "red")  
mu\_e <- mean(error\_model)  
v\_e <- var(error\_model)  
sd\_e <- sqrt(v\_e)  
x\_e <- seq(-11, 16, length = 20)  
y\_e <- dnorm(x\_e, mu\_e, sd\_e)  
lines(x\_e, y\_e, col = "blue")



**Model 20 Residual Plot**

model\_pred20 <- predict(model20, test, type = "response")  
test\_cars <- test[,1]  
error\_model20 <- test\_cars - model\_pred20  
plot(error\_model20, xlab = "Error", ylab = "Residual", main = "Model 20")  
abline(0,0)

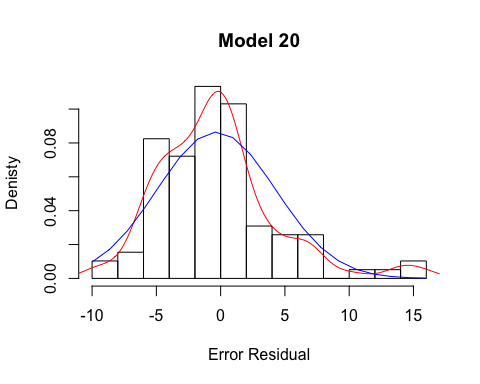


summary(error\_model20)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -9.6204 -3.5581 -0.5564 -0.3371 1.3122 15.9023

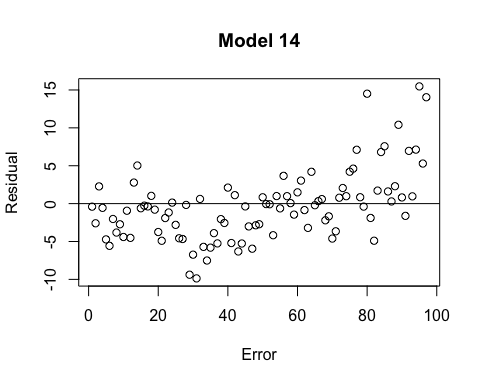
**Model 20 Histogram**

hist(error\_model20, prob = T, breaks = 10, xlab = "Error Residual", ylab = "Denisty", main = "Model 20")  
lines(density(error\_model20), col = "red")  
mu\_e <- mean(error\_model20)  
v\_e <- var(error\_model20)  
sd\_e <- sqrt(v\_e)  
x\_e <- seq(-10, 16, length = 20)  
y\_e <- dnorm(x\_e, mu\_e, sd\_e)  
lines(x\_e, y\_e, col = "blue")



**Model 14 Residual Plot**

model\_pred14 <- predict(model14, test, type = "response")  
test\_cars <- test[,1]  
error\_model14 <- test\_cars - model\_pred14   
plot(error\_model14, xlab = "Error", ylab = "Residual", main = "Model 14")  
abline(0,0)

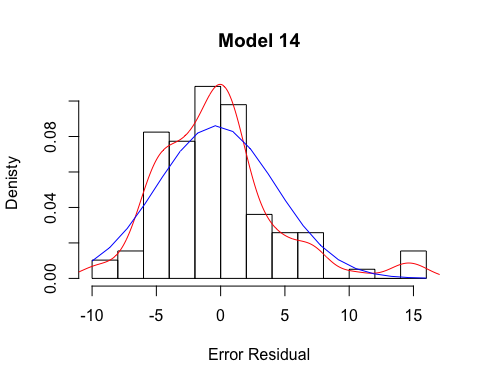


summary(error\_model14)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -9.8715 -3.6528 -0.4026 -0.3371 1.1172 15.4712

**Model 14 Histogram**

hist(error\_model14, prob = T, breaks = 10, xlab = "Error Residual", ylab = "Denisty", main = "Model 14")  
lines(density(error\_model14), col = "red")  
mu\_e <- mean(error\_model14)  
v\_e <- var(error\_model14)  
sd\_e <- sqrt(v\_e)  
x\_e <- seq(-10, 16, length = 20)  
y\_e <- dnorm(x\_e, mu\_e, sd\_e)  
lines(x\_e, y\_e, col = "blue")



**Model 22 Residual Plot**

model\_pred22 <- predict(model22, test, type = "response")  
test\_cars <- test[,1]  
error\_model22 <- test\_cars - model\_pred22  
plot(error\_model22, xlab = "Error", ylab = "Residual", main = "Model 22")  
abline(0,0)



summary(error\_model22)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## -9.8133 -3.7683 -0.4504 -0.3464 1.1842 15.5186

**Model 22 Histogram**

hist(error\_model22, prob = T, breaks = 10, xlab = "Error Residual", ylab = "Denisty", main = "Model 22")  
lines(density(error\_model22), col = "red")  
mu\_e <- mean(error\_model22)  
v\_e <- var(error\_model22)  
sd\_e <- sqrt(v\_e)  
x\_e <- seq(-10, 16, length = 20)  
y\_e <- dnorm(x\_e, mu\_e, sd\_e)  
lines(x\_e, y\_e, col = "blue")



**Comparing Best Models**

When we compared the models the first thing that we observed was that the histogram plots indicated that all four of the models are positively skewed. This is supported by the mean and median values because the mean is greater than the median for all 4 models. This is an indicator of positive skewness. Compared to the blue line in the histograms in all of the models which was the normalized line of the distribution, the red line showed that most of the data was to the left of the zero. This indicated that the models that predicted mpg values were slightly higher than the actual mpg.

Another characteristic of the four models was that the biggest bins in the histograms were the two bins to the right and left of zero. This showed that our models frequently had small prediction errors. All four of the models had a residual bin that was around 15 mpg which indicated that there may be some data in the dataset that could be outliers. We could go back and see if there needs to be some changes to the dataset to help get better predictions.

When looking at all of the residual plots we could not see much difference between all four models. In general, the residual plots supported what was seen in the histograms: that the predictions tended to have a high bias (i.e. positive residuals). Also, for all four models there were three predictions with residuals near 15. These are outliers which need to be investigated. These outliers from the test set may be out of the range of mpg that was used in the training set to train the models. One other observation that our group made from all four models in the residual plots was that as the error increased the residuals went from negative to positive.

The intercomparison of the models was very difficult because of the similarity of the plots. It was noted that model 22 produced a residual histogram in which the two biggest bars in the histogram were to the right and left of zero. This indicated that most of the predictions were around zero and were equally distributed in those two bins. This was different from the other three models in which the bin to the left of zero was the biggest.

If we had to pick a model from these four models we would select model 20 because all the variables in this model have a statistically significant correlation with mpg. This was not the case for the other three models. In model 20 weight and displacement are significantly correlated with mpg. This is based on the fact that both variables have a p-value lower than 0.05. The other three models had variables that had p-values that were higher than 0.05 and thus the null hypothesis was there is no correlation between that variable and mpg at a 95% level of confidence and the null hypothesis could not be rejected. Another reason for its selection, is that while model 20’s R-squared may not be the best, the difference between its R-squared and adjusted R-squared is the smallest. This indicated that there weren’t many wasted variables in the model because adjusted R-squared punishes the model for variables that don’t help the model. It is generally better to have a model with fewer variables if it achieves a similar level of prediction accuracy as models with more variables.