Ch3 Homework

Mac Faldet

February 24, 2019

## Chapter 3 Homework

### Question 4

The scenario: n = 100 Linear Regression Model Cubic Regression Model

1. If there is a linear relationship between X and Y, would we expect a model to perform better on training RSS?

* We would expect the cubic regressino to tighter fit our wide range of data. Thus the cubic regression would have a lower training RSS than the linear model.

1. Answer (a) using test rather than training RSS.

* Now we would expect the linear model to have a lower RSS since it would fit the inherent relaitonship between the variables best. The tight fit from the cubic function would prove to be overfitting and have a higher RSS on the test data.

1. Suppose that the true relationship between X and Y is not perfectly linear. Consider training RSS for the models. Do we know enough to expect one to perform better?

* Regardless of what the true underlying relationship is, the cubic regression model will always fit our training data with a lower RSS than the linear regression model training RSS.

1. Answer (c) using test data.

* We don’t know enough.

### Question 8

The question involves the use of simple linear regression on the Auto data set.

1. Use the lm() function to perfomr a slr with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output.

summary(Auto)

## mpg cylinders displacement horsepower   
## Min. : 9.00 Min. :3.000 Min. : 68.0 Min. : 46.0   
## 1st Qu.:17.00 1st Qu.:4.000 1st Qu.:105.0 1st Qu.: 75.0   
## Median :22.75 Median :4.000 Median :151.0 Median : 93.5   
## Mean :23.45 Mean :5.472 Mean :194.4 Mean :104.5   
## 3rd Qu.:29.00 3rd Qu.:8.000 3rd Qu.:275.8 3rd Qu.:126.0   
## Max. :46.60 Max. :8.000 Max. :455.0 Max. :230.0   
##   
## weight acceleration year origin   
## Min. :1613 Min. : 8.00 Min. :70.00 Min. :1.000   
## 1st Qu.:2225 1st Qu.:13.78 1st Qu.:73.00 1st Qu.:1.000   
## Median :2804 Median :15.50 Median :76.00 Median :1.000   
## Mean :2978 Mean :15.54 Mean :75.98 Mean :1.577   
## 3rd Qu.:3615 3rd Qu.:17.02 3rd Qu.:79.00 3rd Qu.:2.000   
## Max. :5140 Max. :24.80 Max. :82.00 Max. :3.000   
##   
## name   
## amc matador : 5   
## ford pinto : 5   
## toyota corolla : 5   
## amc gremlin : 4   
## amc hornet : 4   
## chevrolet chevette: 4   
## (Other) :365

mod1 <- lm(data = Auto, formula = mpg ~ horsepower)  
summary(mod1)

##   
## Call:  
## lm(formula = mpg ~ horsepower, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.5710 -3.2592 -0.3435 2.7630 16.9240   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 39.935861 0.717499 55.66 <2e-16 \*\*\*  
## horsepower -0.157845 0.006446 -24.49 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.906 on 390 degrees of freedom  
## Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049   
## F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

1. Given a P-value < .05 we can assert there is a significant relationship between horsepower and mpg.
2. An R^2 of .606 means over 60% of the variance in mpg is explained by horsepower.
3. The relationship between the two is negative. More horsepower, less mpg.

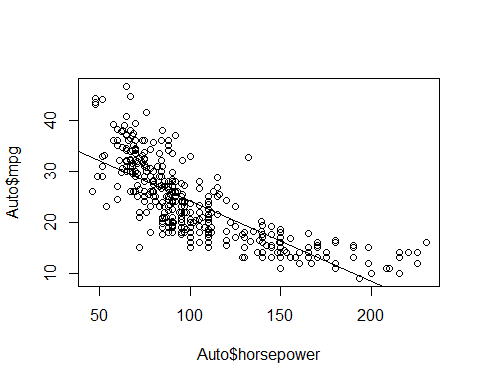
predict(mod1, data.frame(horsepower=c(98)), interval="confidence")

## fit lwr upr  
## 1 24.46708 23.97308 24.96108

predict(mod1, data.frame(horsepower=c(98)), interval="prediction")

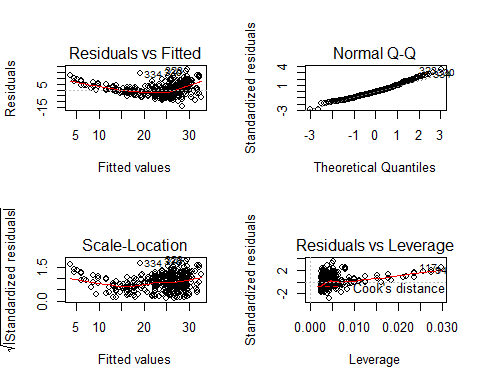
## fit lwr upr  
## 1 24.46708 14.8094 34.12476

plot(x = Auto$horsepower, y = Auto$mpg)  
abline(mod1)



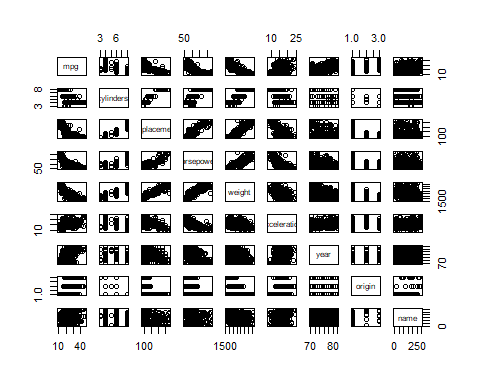
1. None have perfectly linear relationships.

par(mfrow=c(2,2))  
plot(mod1)



### Question 9

pairs(Auto)



cor(subset(Auto, select=-name))

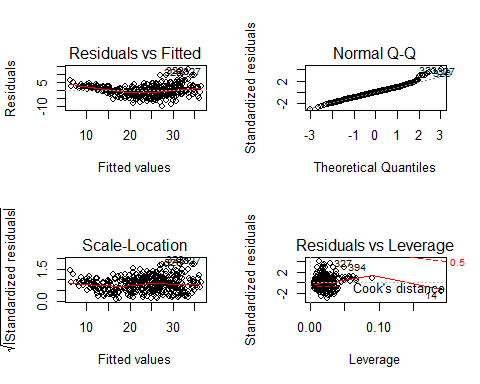
## mpg cylinders displacement horsepower weight  
## mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442  
## cylinders -0.7776175 1.0000000 0.9508233 0.8429834 0.8975273  
## displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944  
## horsepower -0.7784268 0.8429834 0.8972570 1.0000000 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

mod2 <- lm(mpg~.-name, data=Auto)  
summary(mod2)

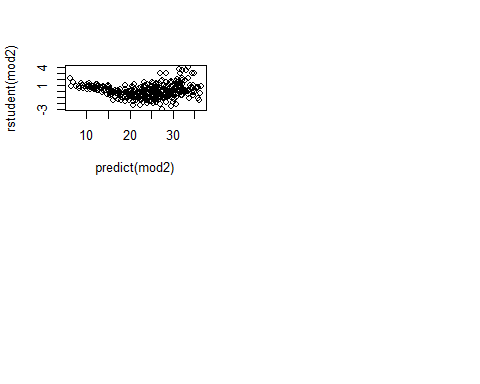
##   
## Call:  
## lm(formula = mpg ~ . - name, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5903 -2.1565 -0.1169 1.8690 13.0604   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.218435 4.644294 -3.707 0.00024 \*\*\*  
## cylinders -0.493376 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. Yes there’s a relationship. F-statistic is above 1 and the p-value is very small.
2. Look at the variable t-value for the correlation. Most have a strong correlation, but cyl, dis, hp, wt are all negative correlations.
3. Given a estimate for year of roughly .75 means that each year the relationship between horsepower and mpg is improving every year. (more fuel efficicent)

par(mfrow=c(2,2))  
plot(mod2)



plot(predict(mod2), rstudent(mod2))



We can see point 14 has an unproportional amount of leverage on the linearity of our model. The points with greater than 3 residuals is likely ot play an effect too.

mod3 <- lm(mpg ~ cylinders\*displacement + displacement\*weight, data=Auto)  
summary(mod3)

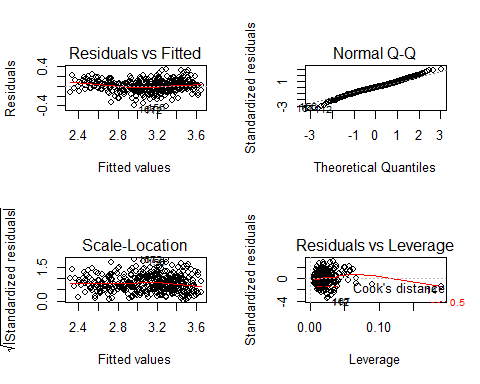
##   
## Call:  
## lm(formula = mpg ~ cylinders \* displacement + displacement \*   
## weight, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -13.2934 -2.5184 -0.3476 1.8399 17.7723   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 5.262e+01 2.237e+00 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

Given the t-statistic, we can see the interaction between dis and cyl is between -1 and 0, not significant. The interaction between dis and wt (4.25) is statistically significant.

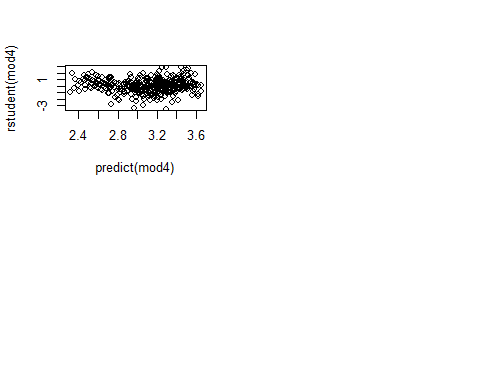
mod4 <- lm(log(mpg)~cylinders+displacement+horsepower+weight+acceleration+year+origin, data=Auto)  
summary(mod4)

##   
## Call:  
## lm(formula = log(mpg) ~ cylinders + displacement + horsepower +   
## weight + acceleration + year + origin, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.40955 -0.06533 0.00079 0.06785 0.33925   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.751e+00 1.662e-01 10.533 < 2e-16 \*\*\*  
## cylinders -2.795e-02 1.157e-02 -2.415 0.01619 \*   
## displacement 6.362e-04 2.690e-04 2.365 0.01852 \*   
## horsepower -1.475e-03 4.935e-04 -2.989 0.00298 \*\*   
## weight -2.551e-04 2.334e-05 -10.931 < 2e-16 \*\*\*  
## acceleration -1.348e-03 3.538e-03 -0.381 0.70339   
## year 2.958e-02 1.824e-03 16.211 < 2e-16 \*\*\*  
## origin 4.071e-02 9.955e-03 4.089 5.28e-05 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.1191 on 384 degrees of freedom  
## Multiple R-squared: 0.8795, Adjusted R-squared: 0.8773   
## F-statistic: 400.4 on 7 and 384 DF, p-value: < 2.2e-16

par(mfrow=c(2,2))  
plot(mod4)



plot(predict(mod4), rstudent(mod4))



Again we see that high leverage point (14).