

3. Linear Regression

Notes

Simple Linear Regression

$$Y = \beta_0 + \beta_1 X_1$$

- coefficients are minimized using ordinary least squares (OLS)
- **Important Assumptions**
 - assuming relationship between X and Y are linear
 - errors have constant variance (homoscedasticity)
 - errors are normally distributed with mean 0 (approximate)
 - errors independent of each other
- **Assessing model fit**
 - Residual Standard Error (RSE) = estimate of the SE of error term
 - * measured in units of Y
 - R^2 statistic
 - * will always increase if more variables are added

Multiple Linear Regression

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p$$

- F-statistics used for MLR from the independent variables to reject null
- qualitative predictors can be represented with dummy variables
- Polynomial regressions allow us to add non-linear relationships between predictor and response, but model overall is still linear
- potential issues (in addition to those of linear regression)
 - correlation of error terms
 - outliers
 - high-leverage point
 - collinearity/multicollinearity
 - * use variance inflation factor (VIF) to solve this

Applied

```
#create a function that loads ISLR datasets
LoadLibraries = function (){
  library(ISLR)
  library(MASS)
  print("Libraries have been loaded")
}
LoadLibraries()
```

```
## [1] "Libraries have been loaded"
```

8)

```
slr <- lm(mpg~horsepower, data = Auto)
summary(slr)
```

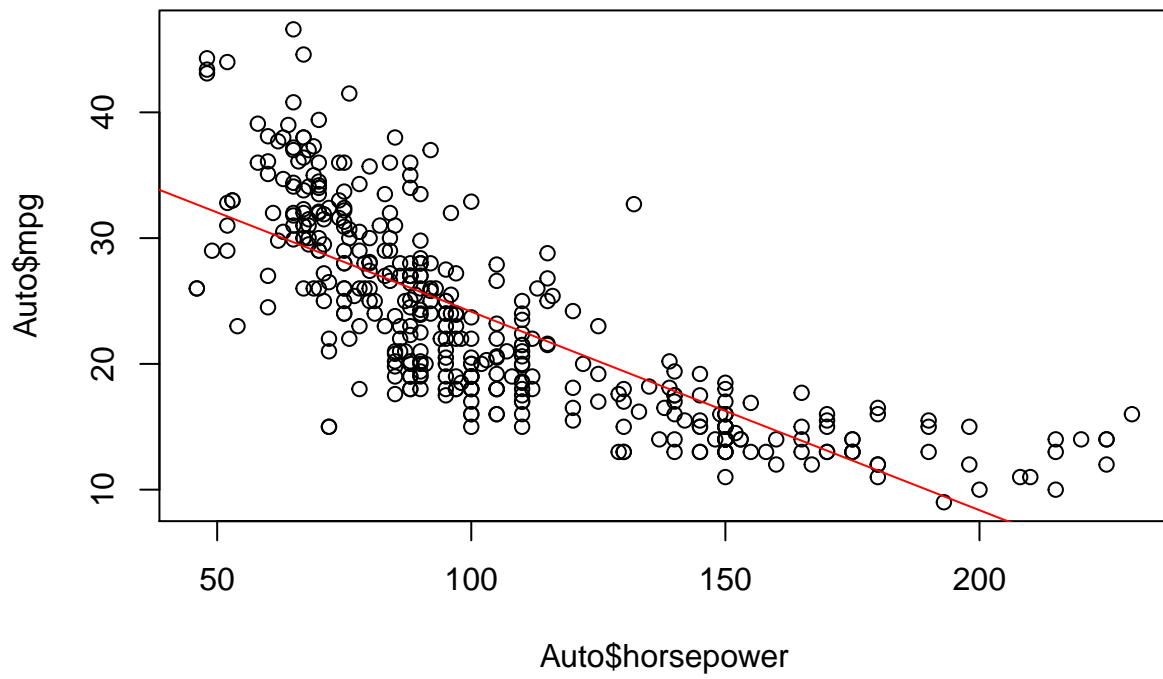
```
##
## Call:
## lm(formula = mpg ~ horsepower, data = Auto)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -13.5710  -3.2592  -0.3435   2.7630  16.9240
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  39.935861   0.717499   55.66  <2e-16 ***
## 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
```

It seems that the predictor horsepower is significant with a negative t value and very low p value

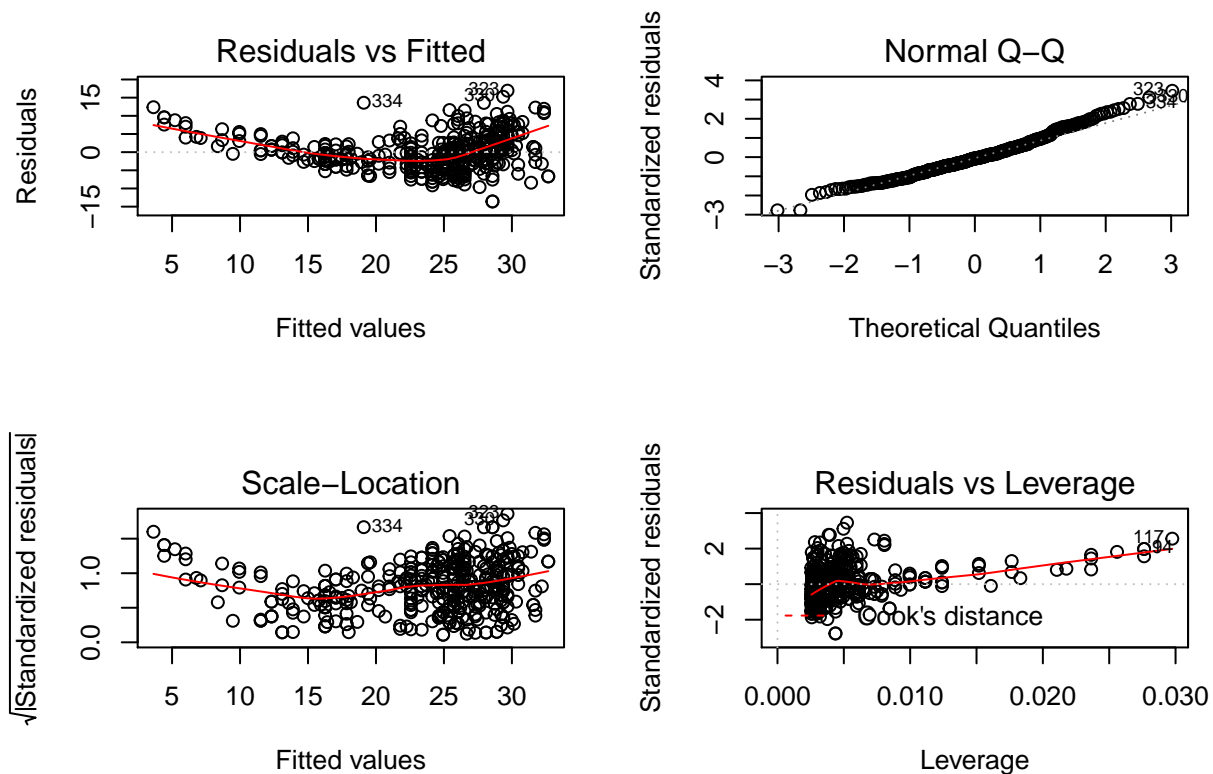
```
predict(slr, data.frame(horsepower = 98), interval = "confidence")
```

```
##          fit          lwr          upr
## 1 24.46708 23.97308 24.96108
```

```
plot(Auto$horsepower, Auto$mpg)
abline(reg = slr, col = "red") #regression line
```



```
par(mfrow=c(2,2))  
plot(slr) #diagnostics plots
```



10)

```
mlr <- lm(Sales ~ Price + Urban + US, data = Carseats)
summary(mlr)
```

```
##
## 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
```

Note that both Urban and US predictors are qualitative