Linear Regression in R

Danny Lumian, Ph.D.

NIH ODSS

Mtcars Data

```
'data.frame':   32 obs. of  11 variables:
 $ mpg : num  21 21 22.8 21.4 18.7 18.1 14.3 24.4 22.8 19.2 ...
 $ cyl : num  6 6 4 6 8 6 8 4 4 6 ...
 $ disp: num  160 160 108 258 360 ...
 $ hp  : num  110 110 93 110 175 105 245 62 95 123 ...
 $ drat: num  3.9 3.9 3.85 3.08 3.15 2.76 3.21 3.69 3.92 3.92 ...
 $ wt  : num  2.62 2.88 2.32 3.21 3.44 ...
 $ qsec: num  16.5 17 18.6 19.4 17 ...
 $ vs  : num  0 0 1 1 0 1 0 1 1 1 ...
 $ am  : num  1 1 1 0 0 0 0 0 0 0 ...
 $ gear: num  4 4 4 3 3 3 3 4 4 4 ...
 $ carb: num  4 4 1 1 2 1 4 2 2 4 ...
```

Linear Regression Assumptions

- Linear relationship.
- Multivariate normality.
- No or little multicollinearity.
- No auto-correlation.
- Homoscedasticity.

- See: http://www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/#building-a-regression-model

Train-Test Split

- trainDataIndex = createDataPartition(mtcars$mpg, p=0.7, list = FALSE)
- trainData = mtcars[trainDataIndex,]
- testData = mtcars[-trainDataIndex,]

Fit the model

- model <- lm(mpg ~ carb + qsec + wt, data = trainData)

- The model predicts miles per gallon based on number of carburetors, quarter mile time and weight.

```
Call:
lm(formula = mpg ~ carb + qsec + wt, data = trainData)

Coefficients:
(Intercept)         carb         qsec           wt
    14.7044       0.2321       1.1293      -4.8603
```

summary(model)

```
Call:
lm(formula = mpg ~ carb + qsec + wt, data = trainData)

Residuals:
    Min      1Q  Median      3Q     Max
-4.0534 -1.8717 -0.4091  1.0551  6.1684

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  14.7044    10.2974   1.428   0.1687
carb          0.2321     0.6278   0.370   0.7155
qsec          1.1293     0.5327   2.120   0.0467 *
wt           -4.8603     0.6358  -7.644 2.34e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.717 on 20 degrees of freedom
Multiple R-squared:  0.8176,    Adjusted R-squared:  0.7902
F-statistic: 29.88 on 3 and 20 DF,  p-value: 1.38e-07
```

Make predictions

- pred = predict(model, newdata = testData)
- Predictions are made on the test data set

```
   Mazda RX4 Wag          Merc 230       Merc 450SL   Toyota Corolla       Camaro Z28
        20.88040          25.71996          17.14774          28.49119          14.37202
Pontiac Firebird     Porsche 914-2     Maserati Bora
        15.73557          23.62710          15.69800
```

Compare predictions with actual values