Supplement for Lecture 14: Assessing a Multiple Regression Model

Load and Clean Data

Variables of Interest in `fatal` - *adj_fatal* = Number of Vehicle Fatalities Per 1,000 People - *youngdrivers* = Percent of Drivers 15 - 24 - *year* = Year - *unemp* = Unemployment Rate - *beertax* = Tax on Case of Beer - *miles* = Average Miles Per Driver

```
data("Fatalities") # Load Data

fatal = Fatalities[,c("fatal","pop","youngdrivers","year",
                      "unemp","beertax","miles")]
fatal$adj_fatal = (fatal$fatal/fatal$pop)*1000
fatal$youngdrivers=fatal$youngdrivers*100
fatal$year = as.numeric(fatal$year) #Need to Convert to Numeric Variable (Currently a Factor Variable)

fatal$fatal=NULL
fatal$pop=NULL

#Rearrange Variables (Put "Y" variable at Beginning)
fatal=fatal[,c(6,1:5)]

#Preview Data
head(fatal)
```

```
##   adj_fatal youngdrivers year unemp  beertax    miles
## 1  0.212836      21.1572    1  14.4 1.539379 7233.887
## 2  0.234848      21.0768    2  13.7 1.788991 7836.348
## 3  0.233643      21.1484    3  11.1 1.714286 8262.990
## 4  0.219348      21.1140    4   8.9 1.652542 8726.917
## 5  0.266914      21.3400    5   9.8 1.609907 8952.854
## 6  0.271859      21.5527    6   7.8 1.560000 9166.302
```

Fit Linear Regression Model

```
#Fit Linear Regression Model
mod = lm(adj_fatal~youngdrivers + year + unemp + beertax + miles,data=fatal)

#Summary from Model
summary(mod)
```

```
##
## Call:
## lm(formula = adj_fatal ~ youngdrivers + year + unemp + beertax +
##     miles, data = fatal)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -0.29356 -0.03185 -0.01131  0.02679  0.20298
##
## Coefficients:
##                  Estimate   Std. Error t value            Pr(>|t|)
## (Intercept)  -0.083876830  0.032434752  -2.586            0.010137 *
## youngdrivers  0.004574323  0.001386812   3.298            0.001078 **
## year          0.004616564  0.001814678   2.544            0.011414 *
## unemp         0.006396829  0.001212069   5.278 0.00000023765533377 ***
## beertax       0.022346613  0.005699494   3.921            0.000107 ***
## miles         0.000015960  0.000001904   8.381 0.00000000000000154 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04739 on 330 degrees of freedom
## Multiple R-squared:  0.3197, Adjusted R-squared:  0.3094
## F-statistic: 31.01 on 5 and 330 DF,  p-value: < 0.000000000000000022
```

```r
#Pulling Out R-squared Adjusted R-squared
sum.out = summary(mod)

sum.out$r.squared
```

```
## [1] 0.3196825
```

```r
sum.out$adj.r.squared
```

```
## [1] 0.3093747
```

```r
#Calculate R-squared by hand
cor(x=fatal$adj_fatal,y=fitted(mod))^2
```

```
## [1] 0.3196825
```

Interpretation of t-Tests: Only predictor variable where we don't have enough evidence to conclude that it's coefficient/slope is significantly different from 0 is *year*.

Interpretation of Slope for *miles*: Holding all other predictor variables (*youngdrivers, year, unemp*, and *beertax*) constant, the average number of vehicle fatalities per 1000 people in a state will increase by 0.0000015960 for every 1 unit increase in the average miles per driver in that state.

Alternative Interpretation of Slope for *miles*: Holding all other predictor variables (*youngdrivers, year, unemp*, and *beertax*) constant, the average number of vehicle fatalities per 1000 people in a state will increase by 0.0015960 if the average miles per driver in that state increased by 1,000.

ANOVA Table

```r
#Run anova() function and notice how it is broken down by predictor variable. This is called sequential
#P-value in F-Test is Actually Testing if the Previous Model is significantly different than the Previo

anova(mod)
```

```
## Analysis of Variance Table
##
## Response: adj_fatal
##               Df  Sum Sq  Mean Sq F value                Pr(>F)
## youngdrivers   1 0.05985 0.059853  26.656 0.000000420955950585 ***
## year           1 0.04080 0.040798  18.170 0.000026405251473668 ***
```

```
## unemp          1 0.03348 0.033485  14.913           0.0001355 ***
## beertax        1 0.05635 0.056347  25.095 0.000000891187452209 ***
## miles          1 0.15770 0.157701  70.234 0.000000000000001539 ***
## Residuals    330 0.74097 0.002245
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Modified ANOVA function
#anova455
anova455(mod)
```

```
## ANOVA Table
## Model: adj_fatal ~ youngdrivers + year + unemp + beertax + miles
##
##        Df  Sum Sq  Mean Sq F value                 P(>F)
## Model   5 0.34818 0.069637  31.014 < 0.00000000000000022 ***
## Error 330 0.74097 0.002245
## Total 335 1.08916
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Another Option for Getting F-statistic
mod.none=lm(adj_fatal~1,data=fatal) #Intercept Only
mod.full=lm(adj_fatal~., data=fatal) #Full Model Includes Intercept

#Table has all the same information but is presented differently.
anova(mod.none,mod.full)
```

```
## Analysis of Variance Table
##
## Model 1: adj_fatal ~ 1
## Model 2: adj_fatal ~ youngdrivers + year + unemp + beertax + miles
##   Res.Df     RSS Df Sum of Sq      F                 Pr(>F)
## 1    335 1.08916
## 2    330 0.74097  5   0.34818 31.014 < 0.00000000000000022 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

#Confidence Intervals and Prediction Intervals

```r
youngdrivers = 24
year  = 8 #Original Years 1982-1988 <=> New Years 1-7
unemp = 8.5
beertax = 1.8
miles = 8000

predict(mod,newdata=data.frame(youngdrivers,year,unemp,beertax,miles),interval="confidence")
```

```
##         fit      lwr       upr
## 1 0.2851193 0.257857 0.3123817
```

```r
predict(mod,newdata=data.frame(youngdrivers,year,unemp,beertax,miles),interval="prediction")
```

```
##         fit       lwr       upr
## 1 0.2851193 0.1879991 0.3822396
```