Supplement for Lecture 10: Partitioning Variability

Load Data

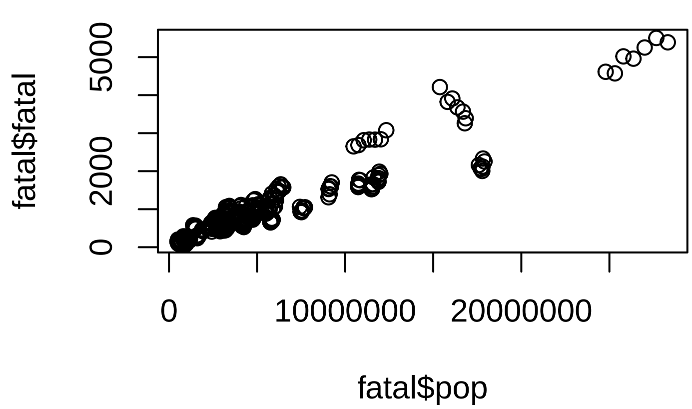
```
data("Fatalities") # Load Data

fatal = Fatalities[,c("fatal","pop","youngdrivers")]

head(fatal)
```

```
##   fatal     pop youngdrivers
## 1   839 3942002     0.211572
## 2   930 3960008     0.210768
## 3   932 3988992     0.211484
## 4   882 4021008     0.211140
## 5  1081 4049994     0.213400
## 6  1110 4082999     0.215527
```

Variables of Interest - *fatal* = Number of vehicle fatalities - *pop* = Population - *youngdrivers* = Percent of Drivers 15 - 24

Create New Variable to Adjust for Population

```
#Consider scatterplot
plot(x=fatal$pop,y=fatal$fatal)
```

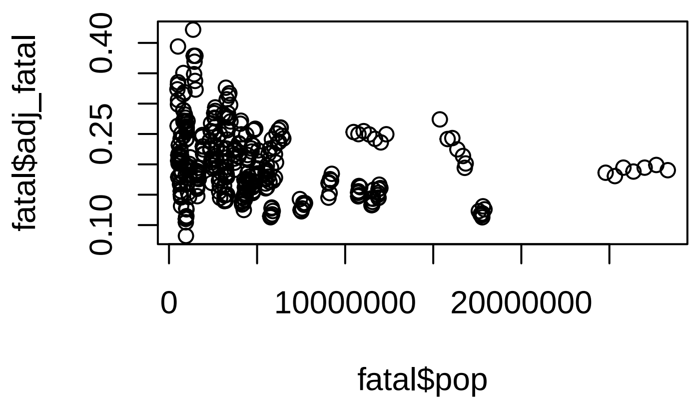



```
#Create New Variable Called adj_fatal
fatal$adj_fatal = (fatal$fatal/fatal$pop)*1000
```

```r
#Remove Original Variable
fatal$fatal = NULL

#Preview Modified Dataset
head(fatal)
```

```
##       pop youngdrivers adj_fatal
## 1 3942002     0.211572  0.212836
## 2 3960008     0.210768  0.234848
## 3 3988992     0.211484  0.233643
## 4 4021008     0.211140  0.219348
## 5 4049994     0.213400  0.266914
## 6 4082999     0.215527  0.271859
```

```r
#Consider new scatterplot
plot(x=fatal$pop,y=fatal$adj_fatal)
```



Output from Simple Linear Regression

```r
#Model for the relationship between fatalities and proportion of young drivers.

#Create new variable for youngdrivers to help interpretation of slope
fatal$yd=fatal$youngdrivers*100

mod = lm(adj_fatal~youngdrivers,data=fatal)
summary(mod)
```

```
##
## Call:
## lm(formula = adj_fatal ~ youngdrivers, data = fatal)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.119634 -0.040335 -0.007417  0.034376  0.205392
##
```

```
## Coefficients:
##               Estimate Std. Error t value   Pr(>|t|)    
## (Intercept)    0.10413    0.02287   4.552 0.00000744 ***
## youngdrivers   0.53738    0.12194   4.407 0.00001414 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.05551 on 334 degrees of freedom
## Multiple R-squared:  0.05495,    Adjusted R-squared:  0.05212 
## F-statistic: 19.42 on 1 and 334 DF,  p-value: 0.00001414
```

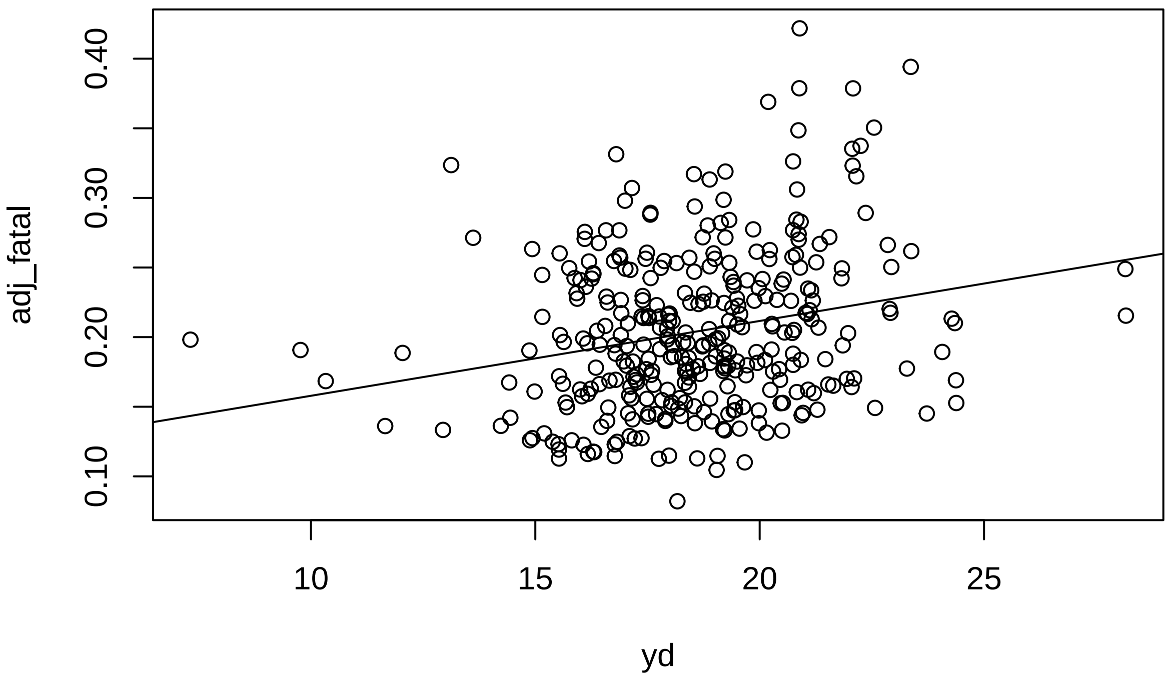
```r
mod = lm(adj_fatal~yd,data=fatal)
summary(mod)
```

```
## 
## Call:
## lm(formula = adj_fatal ~ yd, data = fatal)
## 
## Residuals:
##       Min        1Q    Median        3Q       Max 
## -0.119634 -0.040335 -0.007417  0.034376  0.205392 
## 
## Coefficients:
##              Estimate Std. Error t value   Pr(>|t|)    
## (Intercept) 0.104129   0.022873   4.552 0.00000744 ***
## yd          0.005374   0.001219   4.407 0.00001414 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.05551 on 334 degrees of freedom
## Multiple R-squared:  0.05495,    Adjusted R-squared:  0.05212 
## F-statistic: 19.42 on 1 and 334 DF,  p-value: 0.00001414
```

```r
#Manually calculate p-value using the t-distribution
2*(1-pt(4.407,334,lower.tail=T)) #Find area to right and multiply by 2
```

```
## [1] 0.00001414223
```

```r
#We have found significance. Hooray!!. Let's visualize the model.
plot(adj_fatal~yd,data=fatal)
abline(mod)
```



Focus on the "t value" and "Pr(>|t|)". These are your test statistics and p-values for testing the following hypotheses:

$$H_0 : \beta_x = 0$$
$$H_a : \beta_x \neq 0$$

In class, we focused on when $x = 1$. But we could do the same test for the intercept when $x = 0$.

Confidence Interval for the Slope (and Intercept)

```
confint(mod)
```

```
##                   2.5 %      97.5 %
## (Intercept) 0.059135422 0.149122822
## yd          0.002975191 0.007772437
```

Interpretation of the confidence interal: I am 95 percent confident, that the (average/expected/predicted) number of vehicle fatalities (per 1000) will increase by a number between 0.003 and 0.008 for every 1 percent increase in the percent of young drivers.

ANOVA

```
anova(mod)
```

```
## Analysis of Variance Table
##
## Response: adj_fatal
##            Df  Sum Sq  Mean Sq F value     Pr(>F)
## yd          1 0.05985 0.059853  19.422 0.00001414 ***
## Residuals 334 1.02930 0.003082
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#Manually find the p-value and check it matches
pf(19.422,1,334,lower.tail=FALSE) #Want the area to the right of 19.422
```

```
## [1] 0.00001413977
```

Notice how the p-value for the F-test is identical to the p-value from the t-test. Notice how this p-value is in the output for `summary()`. Also, the last row for the *Total* is not there.

```r
#Hand Calculation of SST
sum((fatal$adj_fatal-mean(fatal$adj_fatal))^2)
```

```
## [1] 1.089155
```

```r
#Notice that this equals the sum from the ANOVA table
0.0598+1.02930
```

```
## [1] 1.0891
```