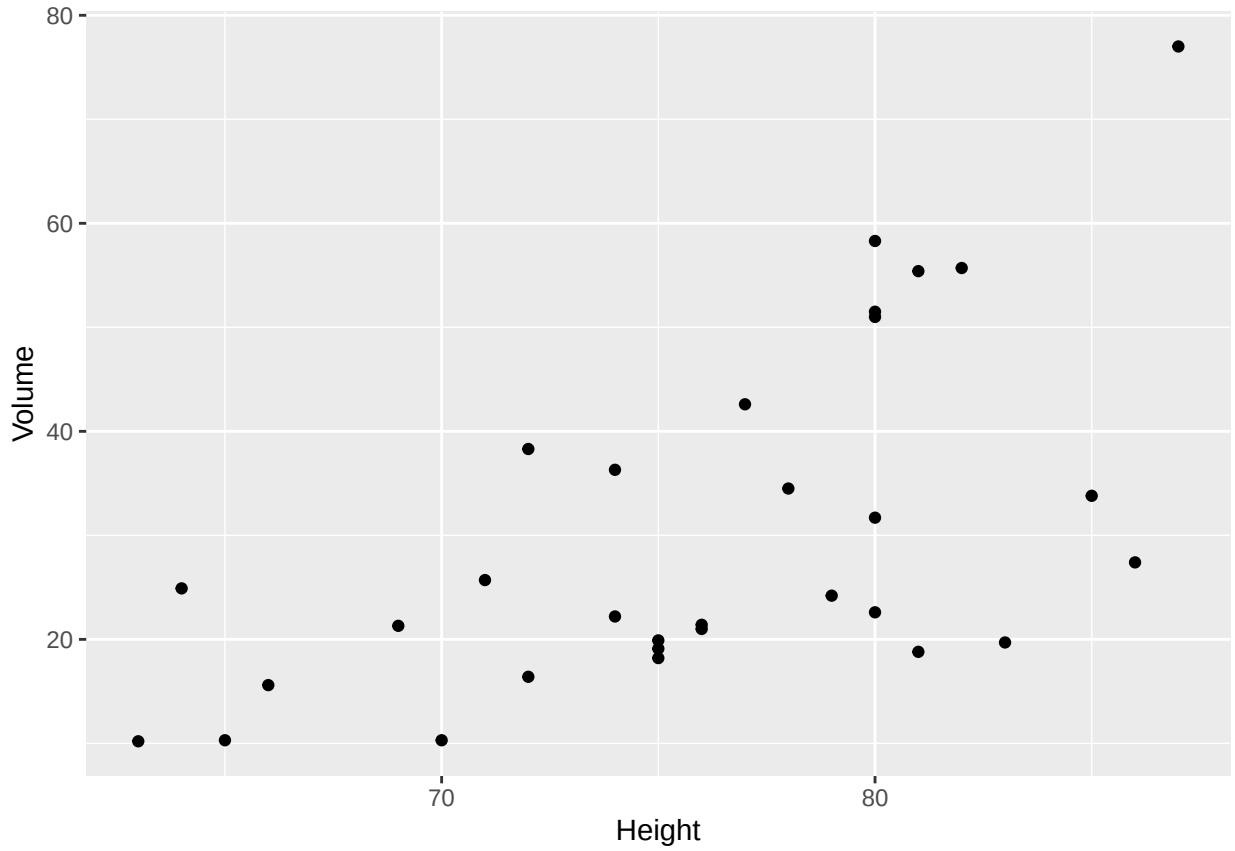
Module5

Joseph Vargovich

9/8/2020

Exercise 1 - Using the trees data frame that comes pre-installed in R, fit the regression model that uses the tree Height to explain the Volume of wood harvested from the tree.

```
data("trees")
write.csv(trees, "trees.csv")
#a. Graph the data.
graph <- ggplot(trees, aes(x=Height, y=Volume)) +
  geom_point()
graph
```




```
#b. Fit an lm model
model <- lm(Volume ~ Height, data=trees)
```

```r
#c. Print out table of coefficient names, estimated value, std erorr, and upper and lower confidence in
summary(model)$coef
```

```
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -87.12361 29.2731221 -2.976232 0.0058346689
## Height        1.54335  0.3838693  4.020509 0.0003783823
```

```r
predict(model)
```

```
##        1        2        3        4        5        6        7        8
## 20.91087 13.19412 10.10742 23.99757 37.88772 40.97442 14.73747 28.62762
##        9       10       11       12       13       14       15       16
## 36.34437 28.62762 34.80102 30.17097 30.17097 19.36752 28.62762 27.08427
##       17       18       19       20       21       22       23       24
## 44.06112 45.60447 22.45422 11.65077 33.25767 36.34437 27.08427 23.99757
##       25       26       27       28       29       30       31
## 31.71432 37.88772 39.43107 36.34437 36.34437 36.34437 47.14782
```

```r
confint(model)
```

```
##                    2.5 %     97.5 %
## (Intercept) -146.993871 -27.253357
## Height         0.758249   2.328451
```

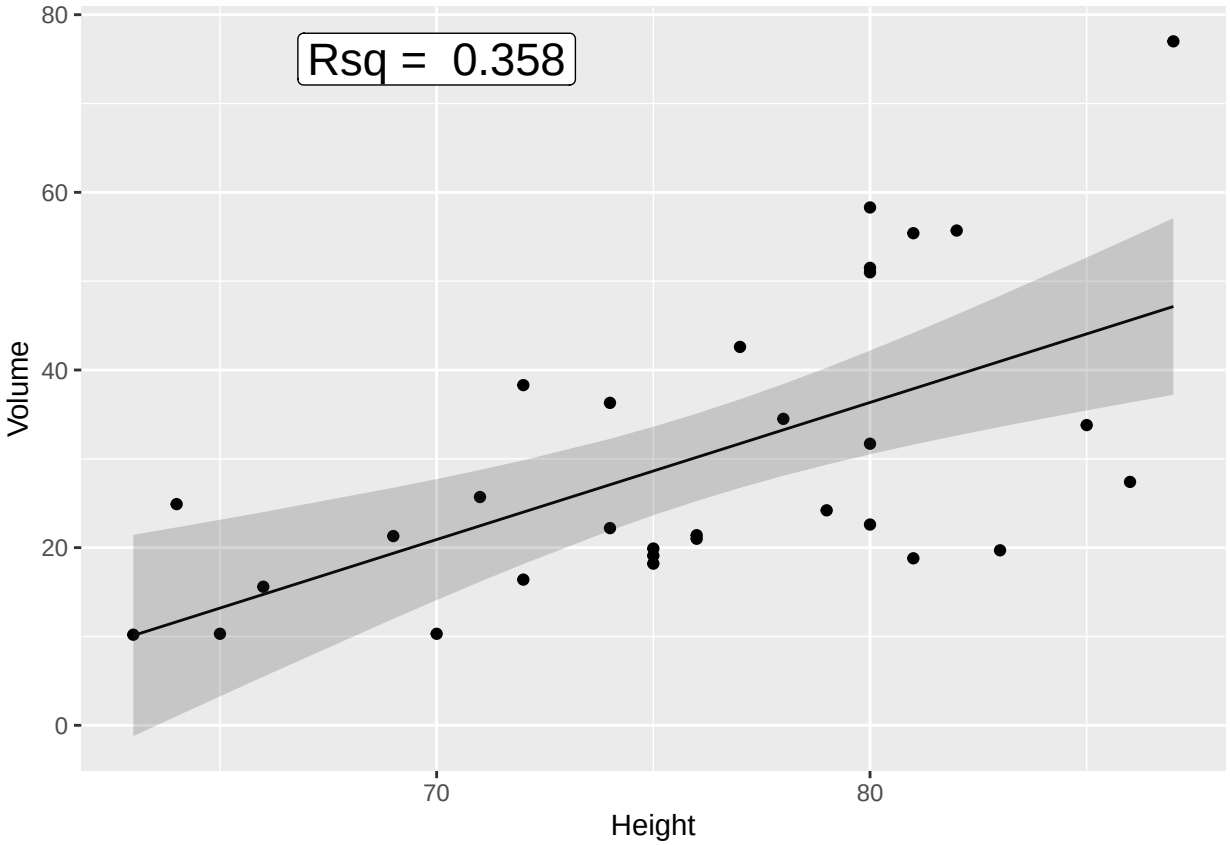
```r
#d. Add model fitted values to the trees data frame along with the regression model confidence interval
trees <- trees %>%
  select(-matches('fit'), -matches('lwr'), -matches('upr')) %>%
  cbind( predict(model, newdata=., interval='confidence') )
head(trees)
```

```
##   Girth Height Volume      fit       lwr      upr
## 1   8.3     70   10.3 20.91087 14.098550 27.72319
## 2   8.6     65   10.3 13.19412  3.254288 23.13395
## 3   8.8     63   10.2 10.10742 -1.223363 21.43821
## 4  10.5     72   16.4 23.99757 18.159758 29.83538
## 5  10.7     81   18.8 37.88772 31.592680 44.18275
## 6  10.8     83   19.7 40.97442 33.597379 48.35145
```

```r
#e. Graph the data and fitted regression line and uncertainty ribbon. Add an annotation for the rsquare
modelRSqd = summary(model)$r.squared
modelRSqd = round(modelRSqd, digits=3)

Rsq_str = paste('Rsq = ', modelRSqd)


graph2 <- ggplot(trees, aes(x=Height, y=Volume)) +
  geom_point() +
  geom_line(aes(y=fit)) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha = .2) +
  annotate('label', x=70.0, y=75, size=6,  label=Rsq_str)
graph2
```



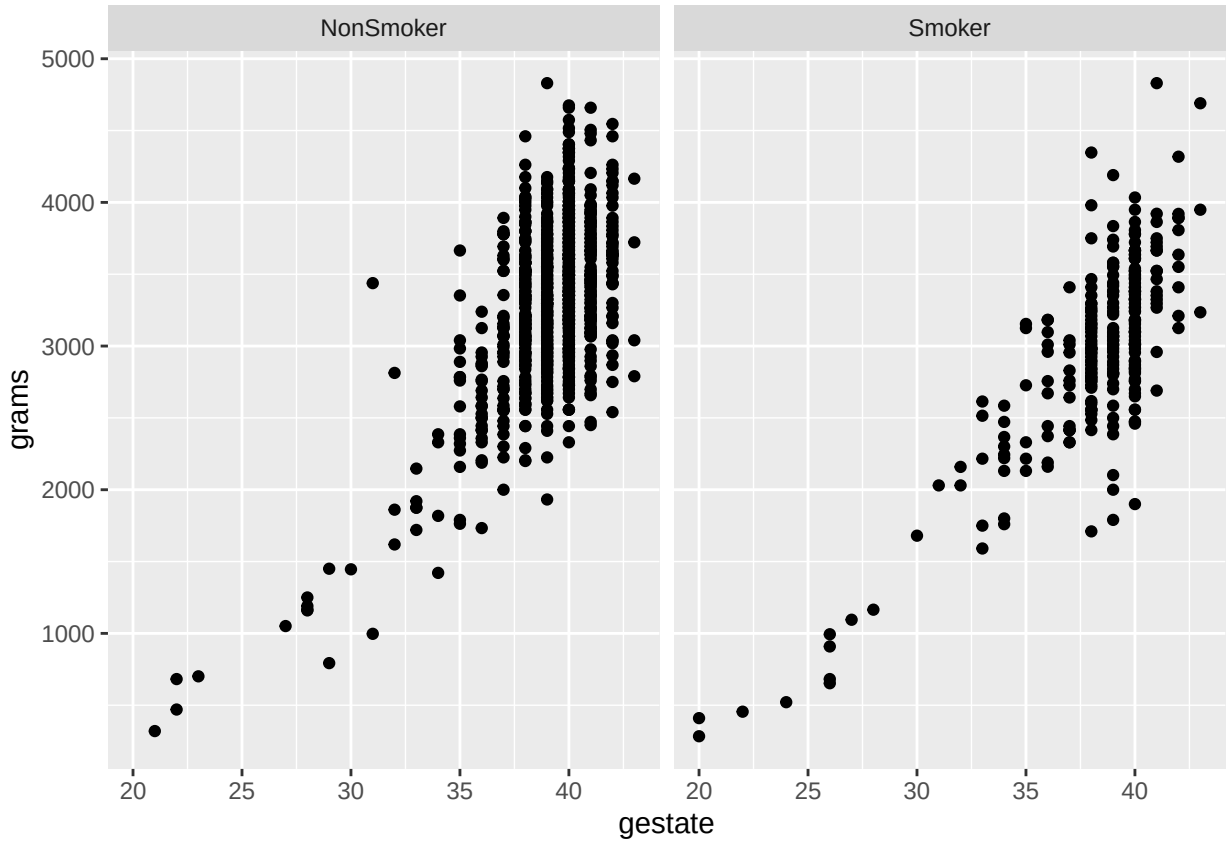
Exercise 2 - Work with the phbirths dataset from the farawaay package.

```
library(faraway)
data(phbirths)
write.csv(phbirths, "phbirths.csv")
#a. Create two scatter plots of gestational length and birthweight, one for each smoking status.

# Lets make some more desciptive names for the smoke column.
phbirths = phbirths %>%
  mutate(smoke = if_else(smoke == 'TRUE', 'Smoker', 'NonSmoker'))

#Now lets graph this modfied dataframe.
smokeGraph <- ggplot(phbirths, aes(x=gestate, y = grams)) +
  geom_point() +
  facet_grid( . ~smoke)

smokeGraph
```




```
#b. Filter results that are premature (less than 36weeks) to use only full term babies.
phbirths = phbirths %>%
  filter(gestate > 36)
head(phbirths)
```

```
##   black educ     smoke gestate grams
## 1 FALSE    0    Smoker      40  2898
## 2 FALSE    2 NonSmoker      38  3977
## 3 FALSE    2    Smoker      37  3040
## 4 FALSE    2 NonSmoker      38  3523
## 5 FALSE    5    Smoker      40  3100
## 6  TRUE    6 NonSmoker      40  3670
```

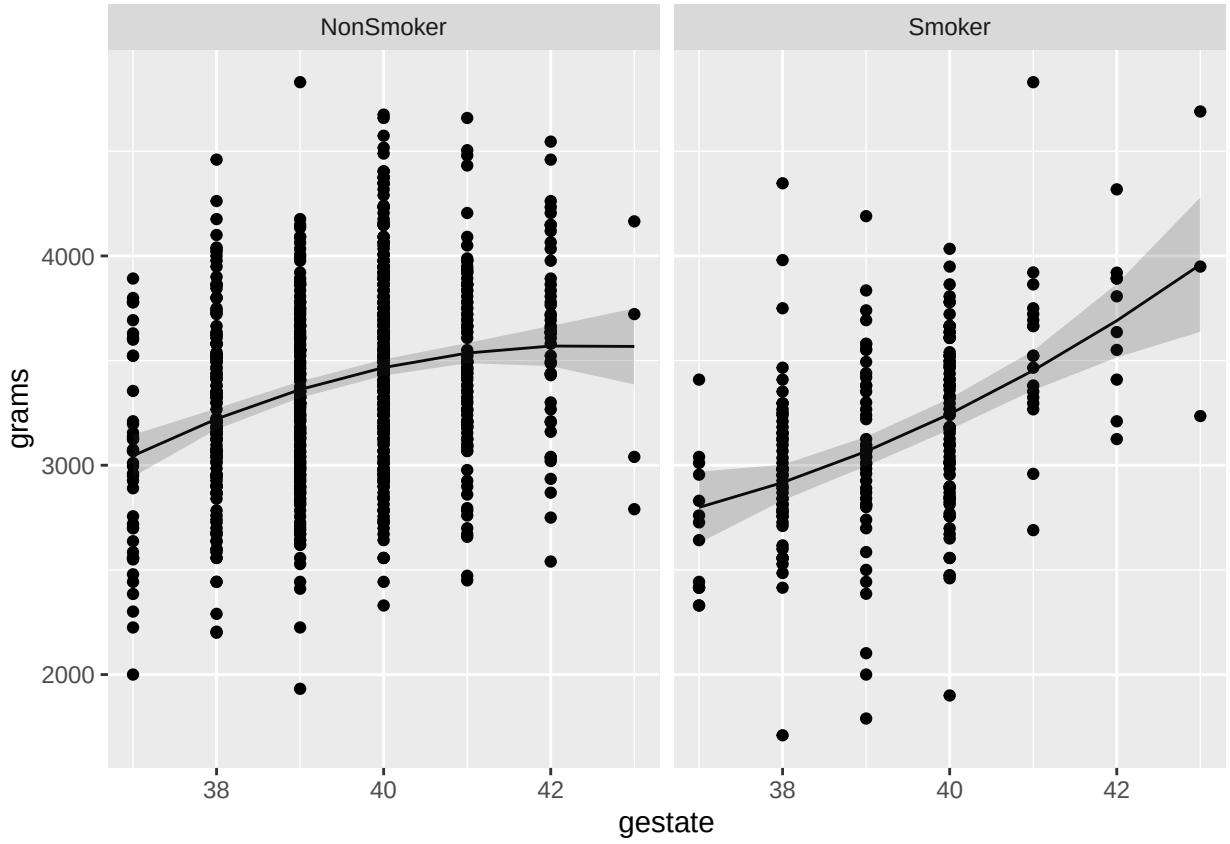
```
#c. Fit a quadratic model to this data
model <- lm(grams ~ poly(gestate,2) * smoke, data=phbirths)

#d. Add the model fit values and CI to the dataframe
phbirths <- phbirths %>%
  select(-matches('fit'), -matches('lwr'), -matches('upr')) %>%
  cbind( predict(model, newdata=., interval='confidence') )
head(phbirths)
```

```
##   black educ     smoke gestate grams      fit      lwr      upr
## 1 FALSE    0    Smoker      40  2898 3244.460 3169.988 3318.932
## 2 FALSE    2 NonSmoker      38  3977 3221.288 3171.053 3271.524
## 3 FALSE    2    Smoker      37  3040 2798.493 2628.203 2968.783
## 4 FALSE    2 NonSmoker      38  3523 3221.288 3171.053 3271.524
```

```
## 5 FALSE    5    Smoker      40  3100 3244.460 3169.988 3318.932
## 6  TRUE    6 NonSmoker      40  3670 3466.630 3427.923 3505.337
```

```
#e. Add layers to the two scatterplot graphs for the model fits and uncertainties.
smokeGraph36w <- ggplot(phbirths, aes(x=gestate, y = grams)) +
  geom_point() +
  geom_line(aes(y=fit)) +
  geom_ribbon(aes(ymin=lwr, ymax=upr), alpha = .2) +
  facet_grid(.~smoke)
smokeGraph36w
```




```
#f. Create a column for the residuals in the phbirths data set
phbirths = phbirths %>% mutate(residuals = resid(model))
head(phbirths)
```

```
##   black educ     smoke gestate grams      fit      lwr      upr residuals
## 1 FALSE    0    Smoker      40  2898 3244.460 3169.988 3318.932 -346.4599
## 2 FALSE    2 NonSmoker      38  3977 3221.288 3171.053 3271.524  755.7117
## 3 FALSE    2    Smoker      37  3040 2798.493 2628.203 2968.783  241.5072
## 4 FALSE    2 NonSmoker      38  3523 3221.288 3171.053 3271.524  301.7117
## 5 FALSE    5    Smoker      40  3100 3244.460 3169.988 3318.932 -144.4599
## 6  TRUE    6 NonSmoker      40  3670 3466.630 3427.923 3505.337  203.3700
```

```
#g. Create a histogram of the residuals
residHistogram = ggplot(phbirths, aes(x=residuals)) +
  geom_histogram()

residHistogram
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```