Assignment 5

Mandi Bluth

2024-09-16

Chapter 5

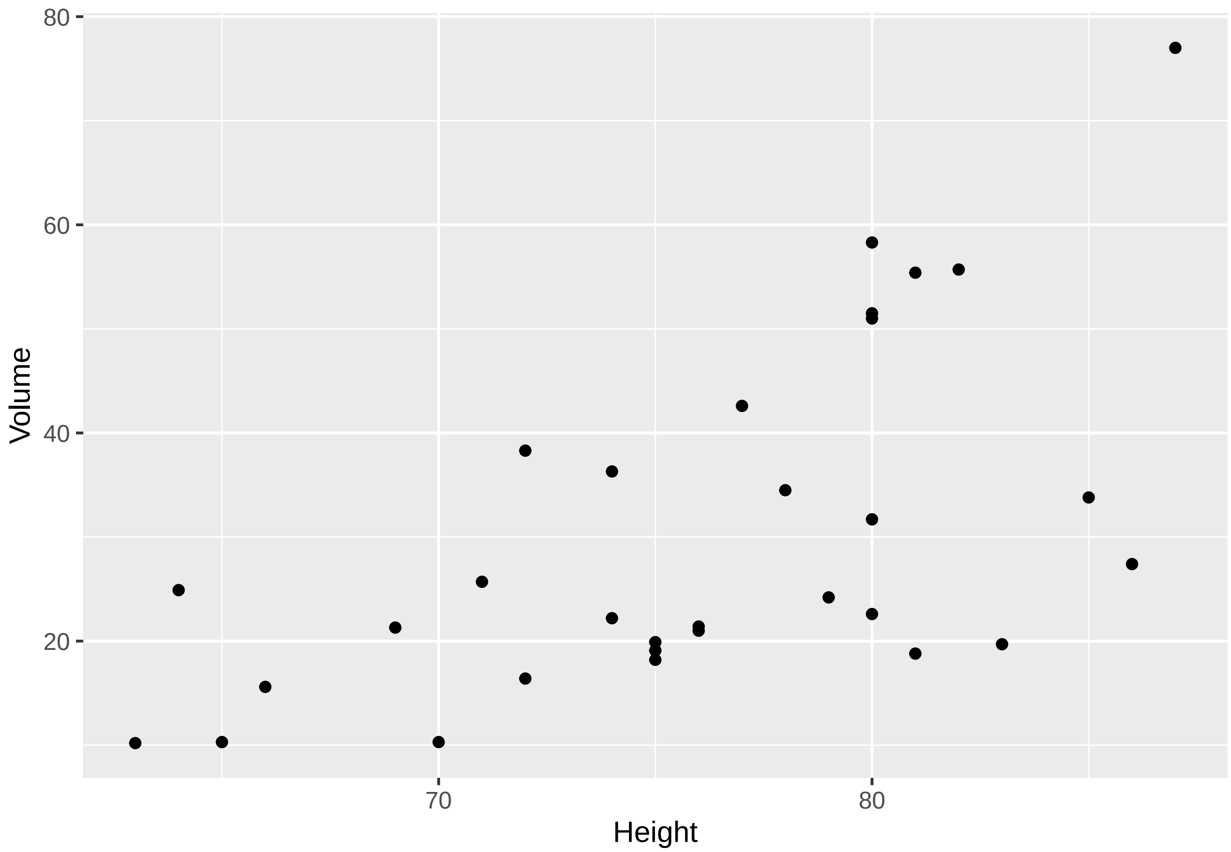
Question 1

Using the trees data frame that comes pre-installed in R, we will fit the regression model that uses the tree Height as a predictor to explain the Volume of wood harvested from the tree. We will then plot our model with some of the information of the regression model on the graph.

a) Graph the data

```
data(trees)
```

```
ggplot(trees, aes(x=Height, y=Volume))+
         geom_point()
```



b) Fit a lm model using the command model <- lm(Volume ~ Height, data=trees).

```
trees.lm <- lm(Volume ~ Height, data=trees)
```

c) Print out the table of coefficients with estimate names, estimated value, standard error, and upper and lower 95% confidence intervals.

```
summary(trees.lm)$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
```

```
confint(trees.lm)
```

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

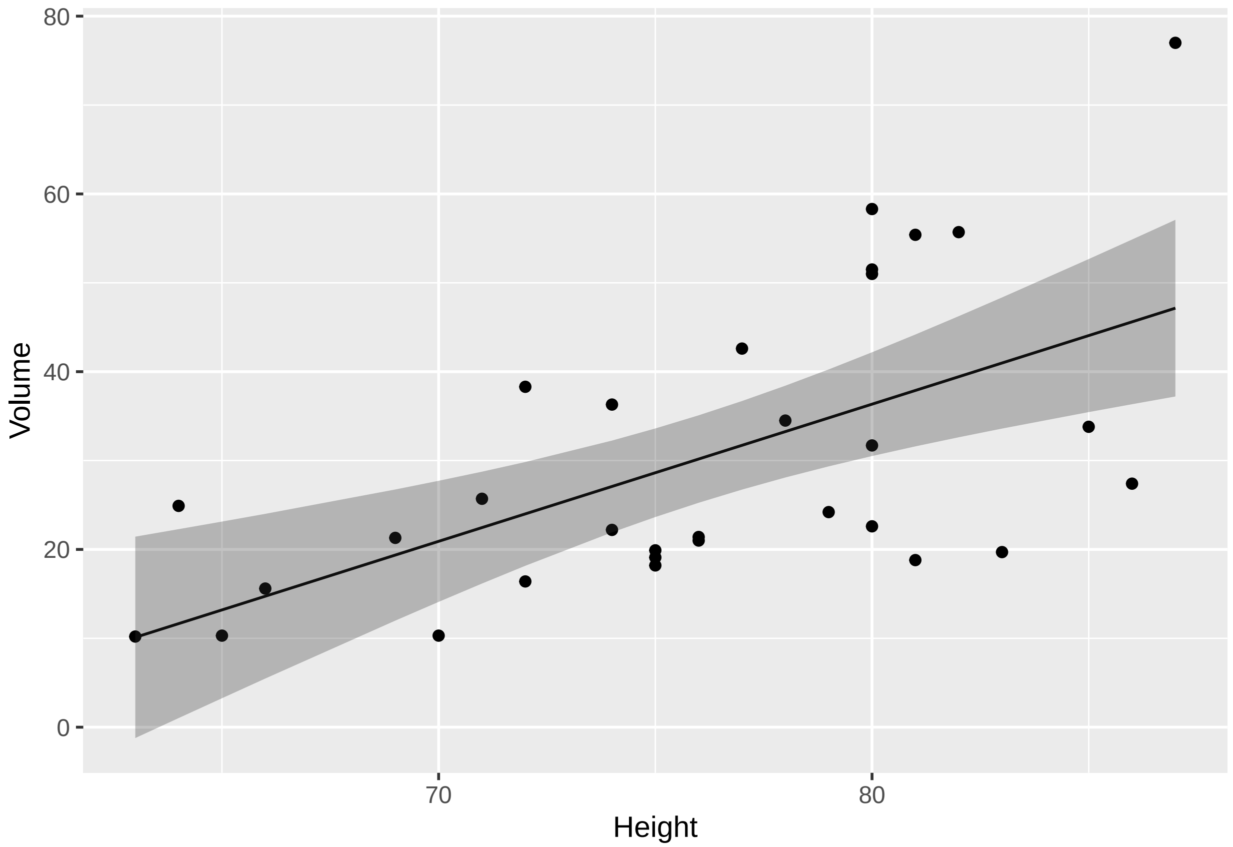
d) Add the model fitted values to the trees data frame along with the confidence interval.

```
trees <- trees %>%
  dplyr::select( -matches('fit'), -matches('lwr'), -matches('upr') ) %>%
  cbind( predict(trees.lm, 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
```

e) Graph the data including now the fitted regression line and confidence interval ribbon.

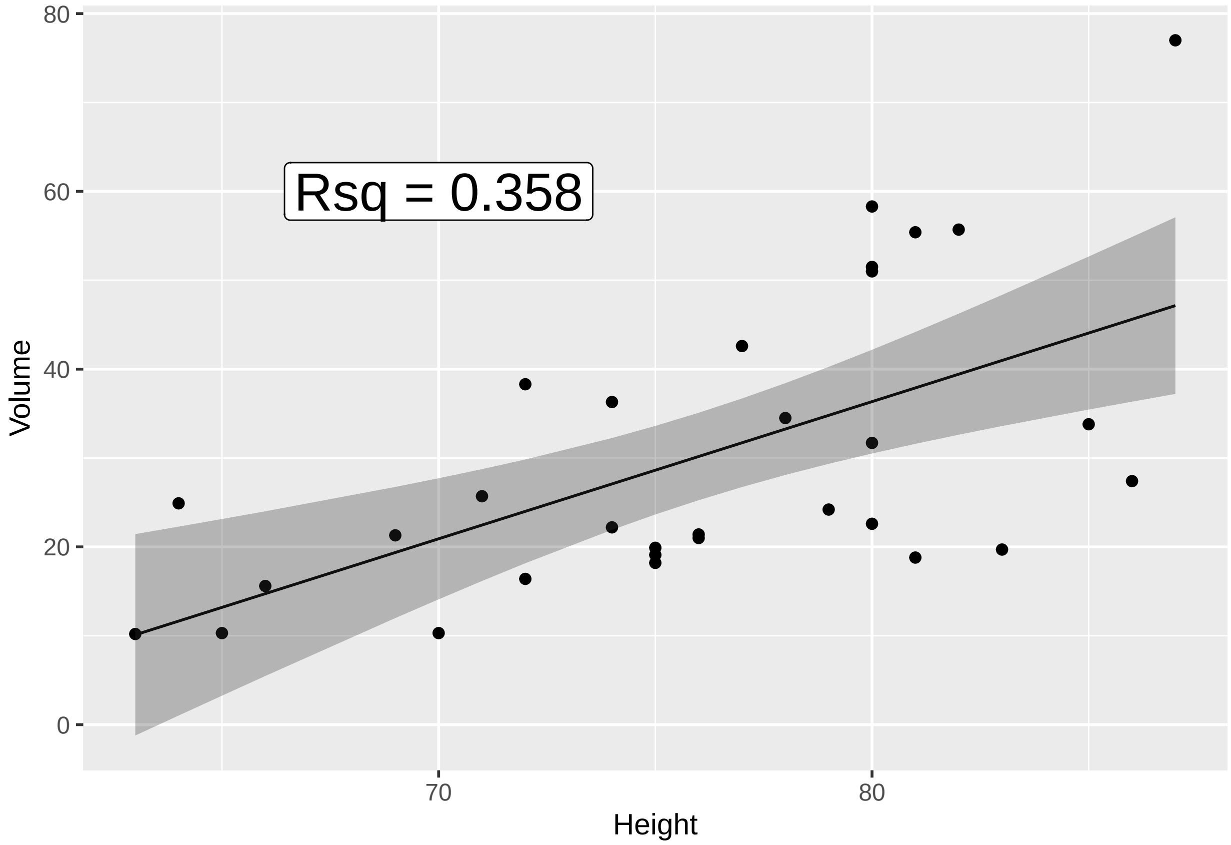
```
ggplot(trees, aes(x=Height)) +
  geom_point( aes(y=Volume) ) +
  geom_line( aes(y=fit) ) +
  geom_ribbon( aes( ymin=lwr, ymax=upr), alpha=.3 )
```



f) Add the R-squared value as an annotation to the graph.

```
Rsq_string <-
  broom::glance(trees.lm) %>%
  select(r.squared) %>%
  mutate(r.squared = round(r.squared, digits=3)) %>%
  mutate(r.squared = paste('Rsq =', r.squared)) %>%
  pull(r.squared)

ggplot(trees, aes(x=Height)) +
  geom_point( aes(y=Volume) ) +
  geom_line( aes(y=fit) ) +
  geom_ribbon( aes( ymin=lwr, ymax=upr), alpha=.3 )+
  annotate('label', label=Rsq_string, x=70, y=60, size=7)
```



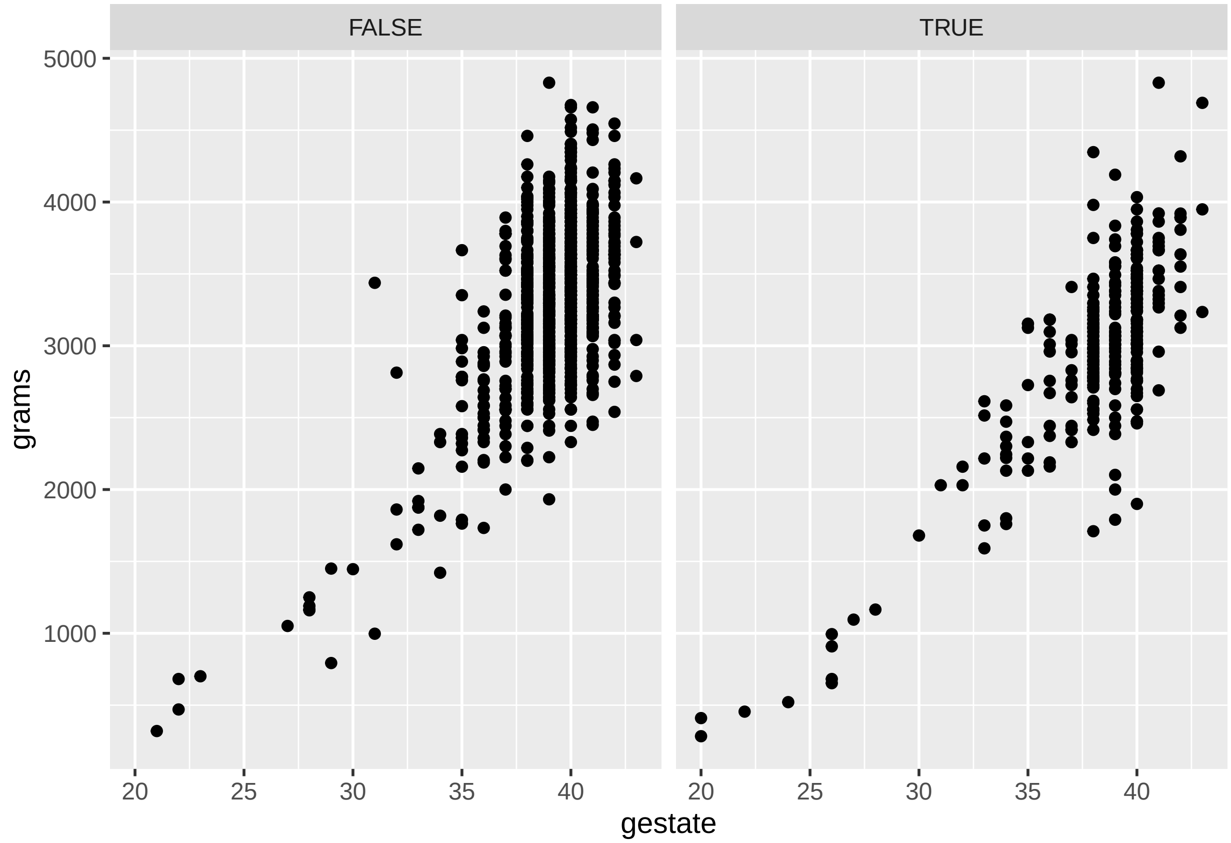
Question 2

The data set phbirths from the faraway package contains information on birth weight, gestational length, and smoking status of mother. We'll fit a quadratic model to predict infant birth weight using the gestational time.

a) Create two scatter plots of gestational length and birth weight, one for each smoking status.

```
data("phbirths", package = "faraway")
```

```
ggplot(phbirths, aes(x=gestate, y=grams))+
  geom_point()+
  facet_grid(cols=vars(smoke))
```



b) Remove all the observations that are premature (less than 36 weeks). For the remainder of the problem, only use full-term births (greater than or equal to 36 weeks).

```
phbirths2 <- phbirths %>% filter(gestate >= 36)
```

c) Fit the quadratic model

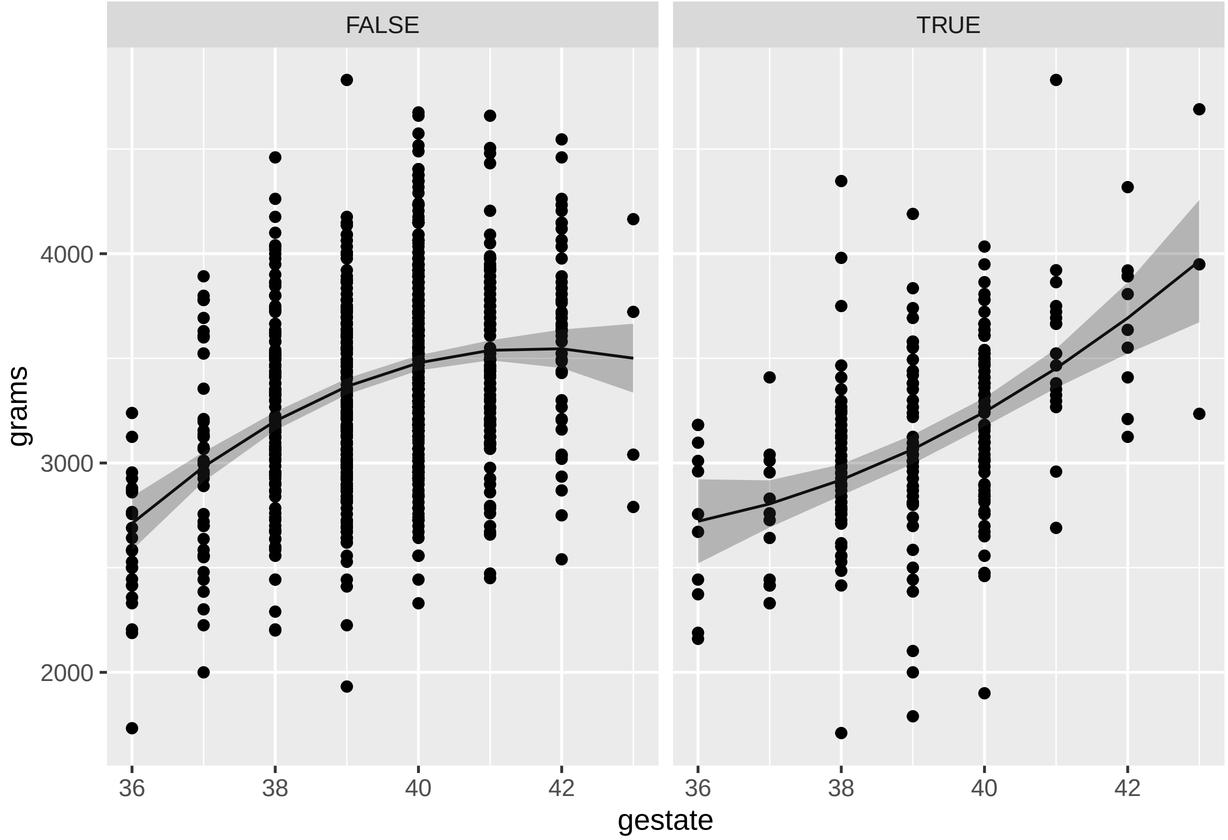
```
model <- lm(grams ~ poly(gestate,2) * smoke, data=phbirths2)
```

d) Add the model fitted values to the phbirths data frame along with the regression model confidence intervals.

```
phbirths2 <- phbirths2 %>%
  dplyr::select( -matches('fit'), -matches('lwr'), -matches('upr') ) %>%
  cbind( predict(model, newdata=., interval='confidence') )
```

e) Improve your graph from part (a) by adding layers for the model fits and confidence interval ribbon for the model fits.

```
ggplot(phbirths2, aes(x=gestate, y=grams))+
  geom_point()+
  facet_grid(cols=vars(smoke))+
  geom_line(aes(y=fit))+
  geom_ribbon( aes( ymin=lwr, ymax=upr), alpha=.3 )
```



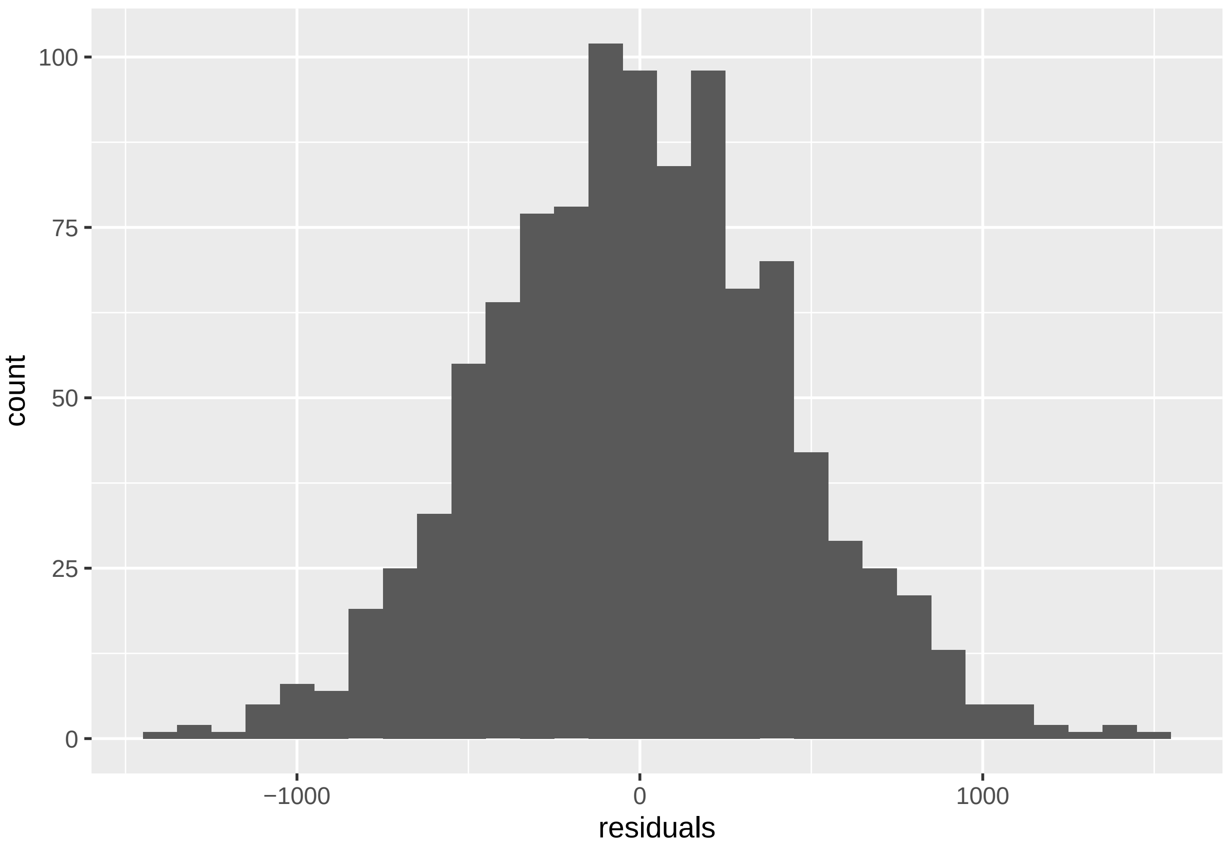
f) Create a column for the residuals in the phbirths data set using any of the following:

```
phbirths2 <- phbirths2 %>% mutate( residuals = resid(model) )
```

g) Create a histogram of the residuals.

```
ggplot(phbirths2, aes(x=residuals))+
  geom_histogram()
```

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


100
75
50
25
0
count
−1000
0
1000
residuals