ST663 - Semester 1 - Assignment 4

Sean O'Riogain (18145426)

24 November 2018

```
knitr::opts_chunk$set(echo = TRUE)
getwd()
```

```
## [1] "C:/Users/oriogain/Dropbox/Maynooth/Statistical Methods/Semester 1 - Assignment 4"
```

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------- tidyverse 1.2.1 --
```

```
## v ggplot2 3.1.0     v purrr   0.2.5
## v tibble  1.4.2     v dplyr   0.7.8
## v tidyr   0.8.2     v stringr 1.3.1
## v readr   1.1.1     v forcats 0.3.0
```

```
## -- Conflicts ------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

Question 1

We have data on ages in years from a random sample of 170 British husbands and their wives.

```
d <- read.table("Couples_Ages.txt", header=T)

str(d)
```

```
## 'data.frame':    170 obs. of  2 variables:
##  $ WifesAge   : int  43 28 30 57 52 27 52 43 23 25 ...
##  $ HusbandsAge: int  49 25 40 52 58 32 43 47 31 26 ...
```

```
head(d)
```

```
##   WifesAge HusbandsAge
## 1       43          49
## 2       28          25
## 3       30          40
## 4       57          52
## 5       52          58
## 6       27          32
```

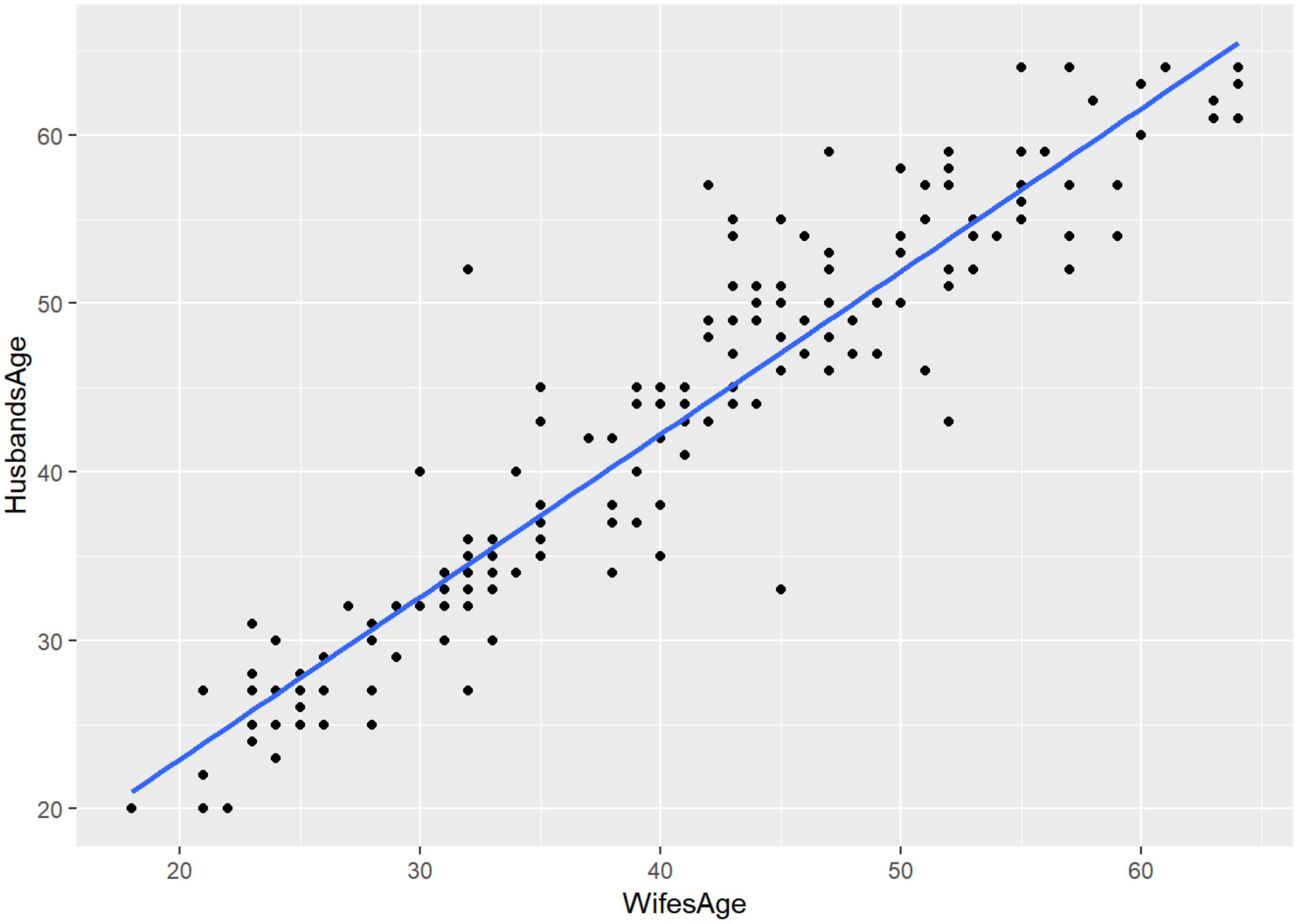
```
sum(is.na(d))           # Check for the presence of NA values in the data
```

```
## [1] 0
```

Note that this dataset contains no NA values.

a. Make a scatterplot of these data with Y=men and X=women. Superimpose the linear regresssion line.

```
ggplot (d, aes(x=WifesAge, y=HusbandsAge)) +
geom_point() +
geom_smooth(method="lm",se=F)
```



Describe the pattern in the data.

At first glance, the data broadly appears to fit a linear model having a positive slope, and with no obvious outliers.

b. Calculate the correlation coefficient.

$$r = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{(n-1)s_x s_y}$$

```
r<-sum((d$WifesAge-mean(d$WifesAge))*(d$HusbandsAge-mean(d$HusbandsAge)))/((nrow(d)-1)*sd(d$WifesAge)*s
d(d$HusbandsAge));r
```

```
## [1] 0.9385598
```

```
cor(x=d$WifesAge,y=d$HusbandsAge)
```

```
## [1] 0.9385598
```

Briefy describe what this coefficient reveals about the relationship between the ages of partners.

Because the value of r above (0.9385598) is greater than 0.9, linear association is very strong and positive for this dataset.

c. Fit the linear model relating the ages of men (Y) to those of women (X).

```
f<-lm(HusbandsAge~WifesAge, data=d);f
```

```
##
## Call:
## lm(formula = HusbandsAge ~ WifesAge, data = d)
##
## Coefficients:
## (Intercept)     WifesAge
##      3.5901       0.9667
```

What is the intercept?

```
bhat0<-round(as.numeric(f$`coefficients`[1]),4);

paste("Intercept =",bhat0)
```

```
## [1] "Intercept = 3.5901"
```

The slope?

```
bhat1<-round(as.numeric(f$`coefficients`[2]),4);

paste("Slope =",bhat1)
```

```
## [1] "Slope = 0.9667"
```

Interpret.

The above figures mean that, for this model, for two couples, where one wife is a year older than the other, almost the same age gap (0.9667) will apply to their respective husbands.

d. Deirdre is 60 years of age. Use the regression equation to predict the age of her husband.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

```
x<-60

yhat<-bhat0+(bhat1*x);yhat
```

```
## [1] 61.5921
```

```
paste("Prediction =",round(yhat,1))
```

```
## [1] "Prediction = 61.6"
```

Calculate a 95% prediction interval and interpret.

Question 2

There is some evidence that drinking a moderate amount of wine helps prevent heart attacks.

The data in wine.csv (on Moodle) gives yearly wine consumption (litres of alcohol from drinking wine, per person) and yearly deaths from heart disease (deaths per 100,000 people) in 19 developed nations.

```
wine <- read.csv("wine.csv", stringsAsFactors=F)

str(wine)
```

```
## 'data.frame':    19 obs. of  3 variables:
##  $ Country: chr  "Australia" "Austria" "Belgium" "Canada" ...
##  $ Wine   : num  2.5 3.9 2.9 2.4 2.9 0.8 9.1 0.8 0.7 7.9 ...
##  $ Deaths : int  211 167 131 19 220 297 71 211 300 107 ...
```

```
head(wine)
```

```
##     Country Wine Deaths
## 1 Australia  2.5    211
## 2   Austria  3.9    167
## 3   Belgium  2.9    131
## 4    Canada  2.4     19
## 5   Denmark  2.9    220
## 6   Finland  0.8    297
```

```
sum(is.na(wine))        # Check for the presence of NA values in the data
```

```
## [1] 0
```

Note that this dataset contains no NA values.

a. Plot the data with yearly deaths on the y-axis. Superimpose the regression line.

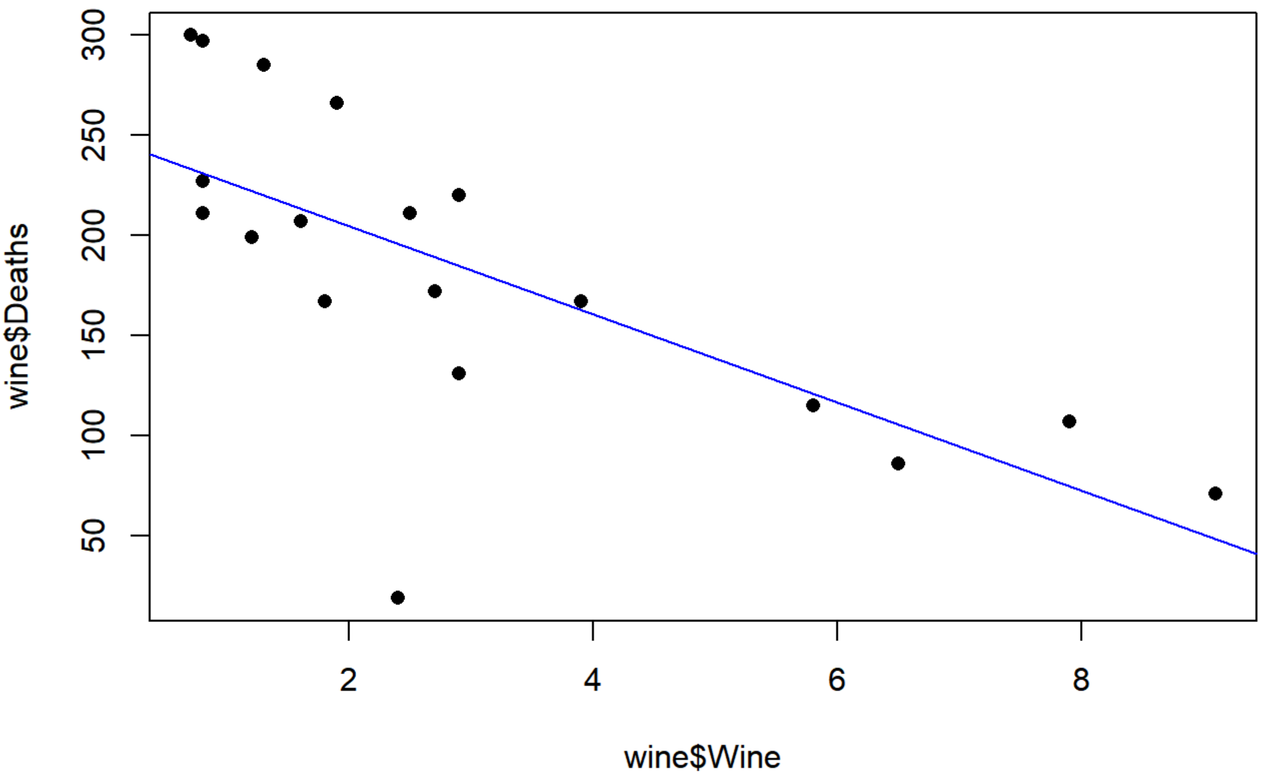
```
f<-lm(Deaths~Wine,data=wine);f
```

```
## 
## Call:
## lm(formula = Deaths ~ Wine, data = wine)
## 
## Coefficients:
## (Intercept)         Wine  
##      248.64       -22.02
```

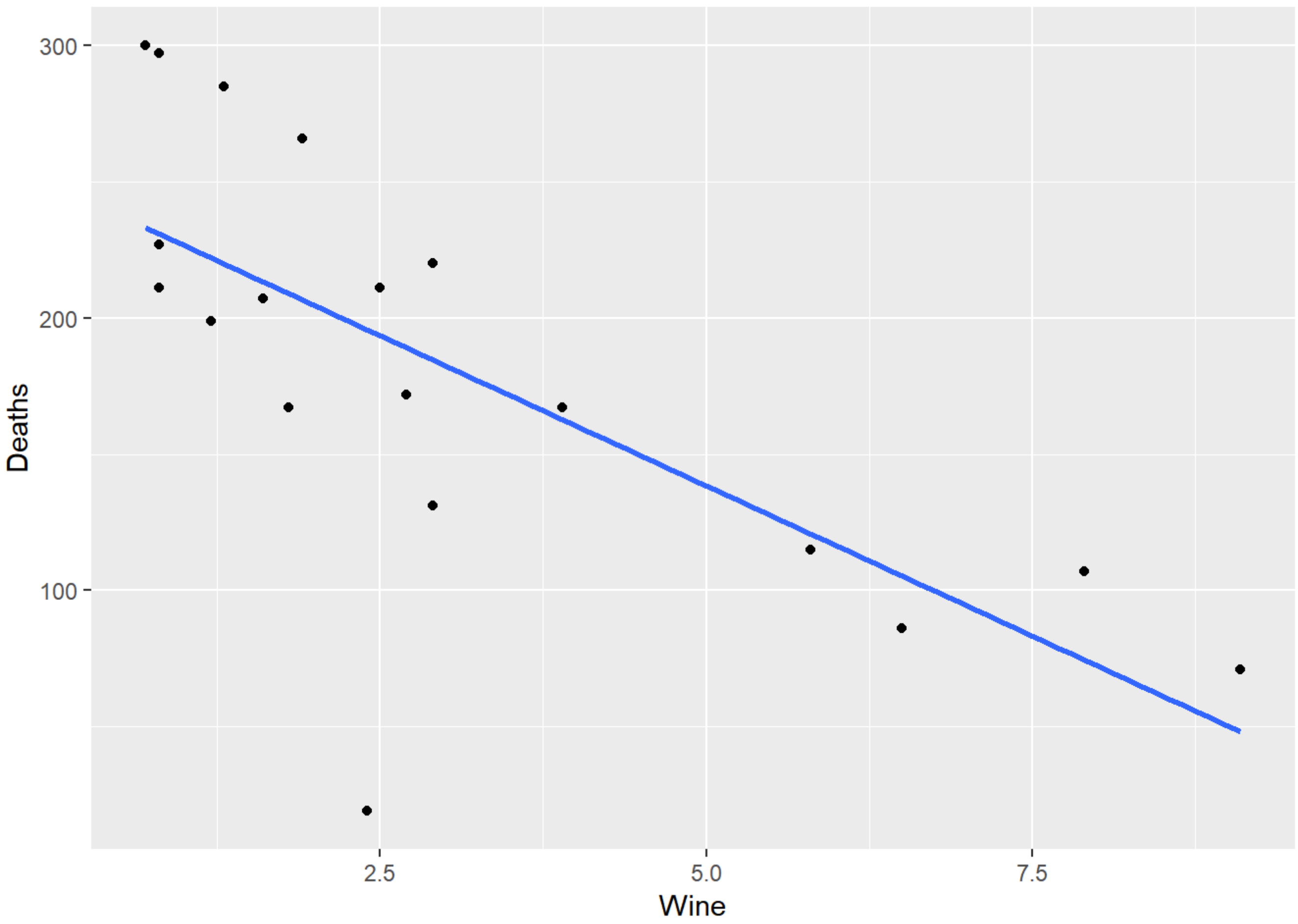
```
summary(f)
```

```
##
## Call:
## lm(formula = Deaths ~ Wine, data = wine)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -176.79  -19.77   -4.02   33.77   66.78
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  248.635     21.202  11.727 1.43e-09 ***
## Wine         -22.019      5.452  -4.039 0.000852 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 58.05 on 17 degrees of freedom
## Multiple R-squared:  0.4897, Adjusted R-squared:  0.4597
## F-statistic: 16.31 on 1 and 17 DF,  p-value: 0.0008523
```

```
plot(wine$Wine,wine$Deaths,pch=16)
abline(f,col="blue")
```



```
ggplot(aes(Wine,Deaths),data=wine) +
  geom_point() +
  geom_smooth(method="lm",se=F)
```



Describe the pattern in the data.

At first glance, apart from a single outlier, this dataset appears to conform to a broadly linear model of negative slope. The main body of the dataset appears to fan out (slightly) to the left.

b. Use the least squares regression line to estimate the effect of a 1 litre increase in wine consumption on the death rate.

```
bhat0<-as.numeric(f$coefficients[1]);bhat0
```

```
## [1] 248.6353
```

```
bhat1<-as.numeric(f$coefficients[2]);bhat1
```

```
## [1] -22.01861
```

The second figure above (i.e. the slope value) indicates that a 1-litre increase in wine consumption would result in a reduction of approximately 22 in the number of deaths.

c. Compute a 95% confidence interval for this effect.

$$SSE = \Sigma_{i=1}^{n}(\hat{y}_i - y_i)^2$$

$$s = \sqrt{SSE/(n-2)}$$

$$S_{xx} = \Sigma(x_i - \bar{x})^2$$

$$SE = s/\sqrt{Sxx}$$

$$MoE = t_{\alpha/2}(n-2) * SE$$

```
c<-0.95                                # Level of Confidence
alpha<-1-c                             # Alpha
alpha2<-alpha/2                        # Half Alpha
n<-nrow(wine)                          # Sample size
df<-n-2                                # Degrees of freedom

sse<-sum(f$residuals^2);sse            # Sum of Squares for Errs (Residuals)
```

```
## [1] 57282.75
```

```
s<-sqrt(sse/(n-2));s                   # Standard Deviation (Errs/Residuals)
```

```
## [1] 58.04803
```

```
ssx<-sum((wine$Wine-mean(wine$Wine))^2);ssx  # Sum of squared deviations of
```

```
## [1] 113.3768
```

```
                                             #    x from mean

se.bhat1<-s/sqrt(ssx);se.bhat1         # Standard error for slope
```

```
## [1] 5.451617
```

```
t<-abs(qt(alpha2,df=df));t             # t-value for 95% level of confidence
```

```
## [1] 2.109816
```

```
moe<-t*se.bhat1;moe                    # Margin of error
```

```
## [1] 11.50191
```

```
paste(c," CI for Slope = (",round(bhat1-moe,4),",",round(bhat1+moe,4),")",sep="")
```

```
## [1] "0.95 CI for Slope = (-33.5205,-10.5167)"
```

```
confint(f)                             # Compare above CI with confint o/put
```

```
##                 2.5 %   97.5 %
## (Intercept) 203.90217 293.3683
## Wine        -33.52051 -10.5167
```

Give your conclusions.

We can be 95% confident that the slope of the regression line for poulation as a whole lies somewhere in the range from -33.52 to -10.52, approximately.

d. Ireland is one of the countries in the data. What is the fitted value for Ireland?

```
paste("Fitted Value for Ireland = ",round(f$fitted.values[which(wine$Country=="Ireland")],2))
```

```
## [1] "Fitted Value for Ireland =  233.22"
```

What is the residual for Ireland?

```
paste("Residual for Ireland = ",round(f$residuals[which(wine$Country=="Ireland")],2))
```

```
## [1] "Residual for Ireland =  66.78"
```

e. Predict the yearly deaths from heart disease for a country whose yearly wine consumption is 4.5.

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

```
x0<-4.5                                      # Consumption

yhat0<-bhat0+(bhat1*x0);yhat0                # Prediction
```

```
## [1] 149.5515
```

```
paste("Prediction =",round(yhat0))
```

```
## [1] "Prediction = 150"
```

Calculate a 95% prediction interval

$$\hat{y}_0 = \hat{\beta}_0 + \hat{\beta}_1 x_0$$

$$SE_{pred}(\hat{y}_0) = s\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{Sxx}}$$

where

$$S_{xx} = \Sigma(x_i - \bar{x})^2$$

```
se<-s*sqrt(1 + (1/n) + (((x0-mean(wine$Wine))^2)/ssx))

moe<-t*se;moe
```

```
## [1] 126.7903
```

```
paste(c," PI for yhat0=", round(yhat0), " where x0=", x0, ": (",round(yhat0-moe,4),",",round(yhat0+moe,
4),")",sep="")
```

```
## [1] "0.95 PI for yhat0=150 where x0=4.5: (22.7612,276.3419)"
```

```
predict(f,data.frame(Wine=4.5),interval="prediction")
```

```
##        fit      lwr      upr
## 1 149.5515 22.76119 276.3419
```

and interpret.

We are 95% confident that an annual consumption of 4.5 litres of wine per per person in the country in question would result in a death rate in the 22.76-276.34 range per 100,000 people in that country’s population.

Question 3

For the trees data:

a. Plot Volume versus Girth and superimpose the regression line.

```
str(trees)
```

```
## 'data.frame':    31 obs. of  3 variables:
##  $ Girth : num  8.3 8.6 8.8 10.5 10.7 10.8 11 11 11.1 11.2 ...
##  $ Height: num  70 65 63 72 81 83 66 75 80 75 ...
##  $ Volume: num  10.3 10.3 10.2 16.4 18.8 19.7 15.6 18.2 22.6 19.9 ...
```

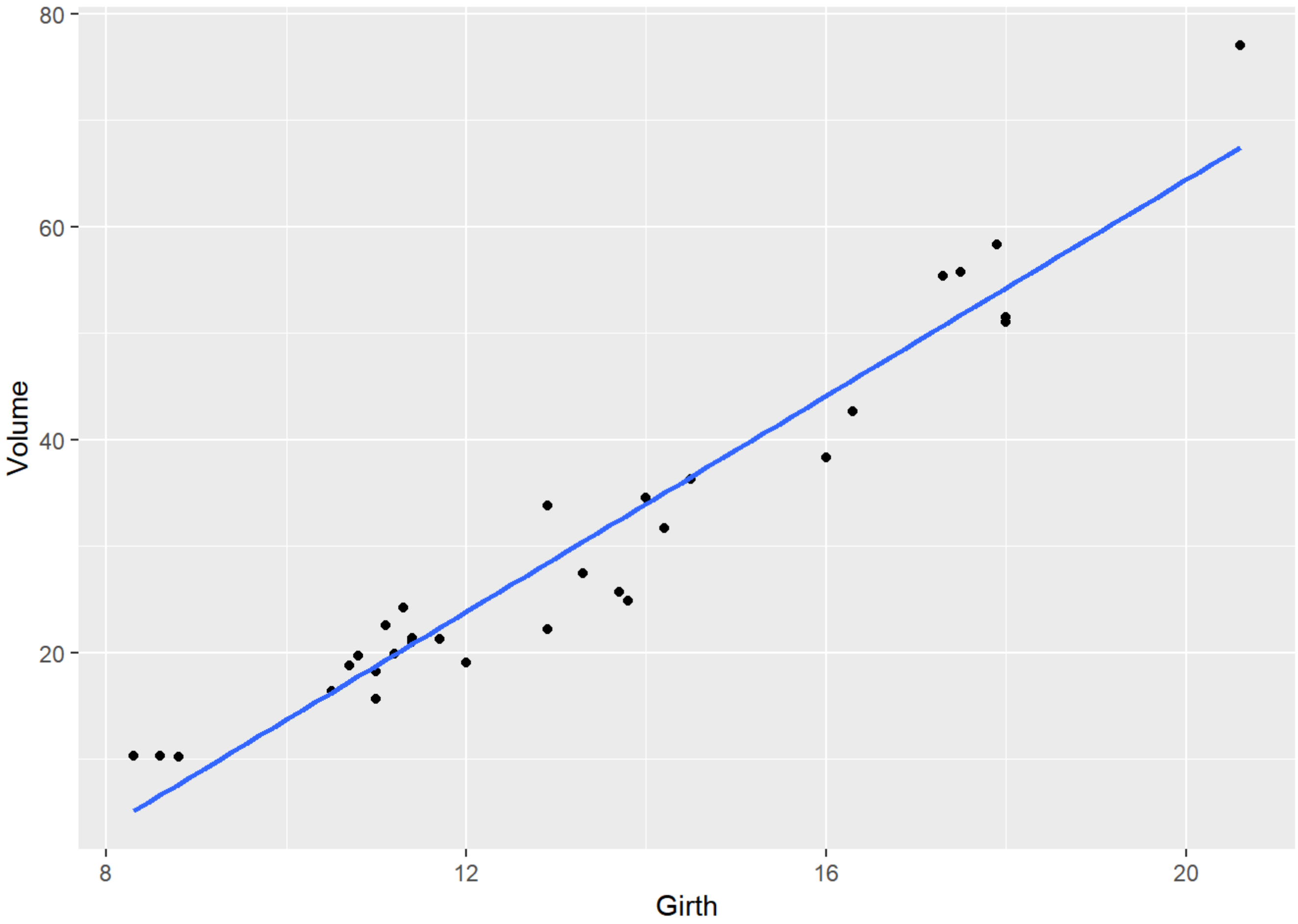
```
head(trees)
```

```
##   Girth Height Volume
## 1   8.3     70   10.3
## 2   8.6     65   10.3
## 3   8.8     63   10.2
## 4  10.5     72   16.4
## 5  10.7     81   18.8
## 6  10.8     83   19.7
```

```
sum(is.na(trees))        # Check for the presence of NA values in the data
```

```
## [1] 0
```

```
ggplot(data=trees,aes(x=Girth,y=Volume)) +
  geom_point() +
  geom_smooth(method="lm",se=F)
```



Note that this dataset does not contain NA values.

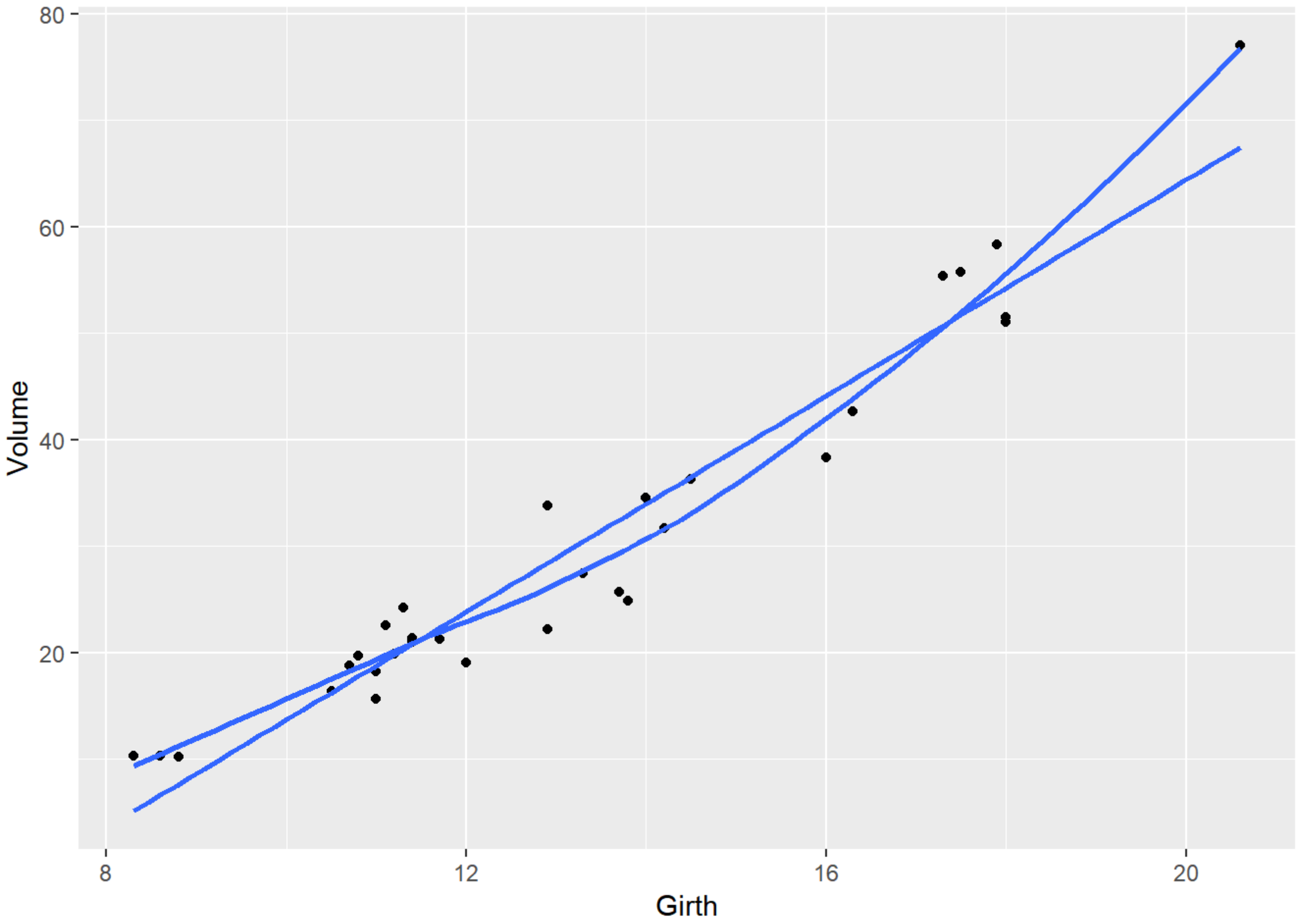
Does the relationship look linear?

At first glance, the data broadly appears to fit a linear model having a positive slope, and with no obvious outliers.

b. Now make the plot again and this time superimpose the regression line and a smooth.

```
ggplot(data=trees,aes(x=Girth,y=Volume)) +
  geom_point() +
  geom_smooth(method="lm",se=F) +
  geom_smooth(se=F)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



Does the relationship look linear?

The presence of the smooth now indicates the presence of some curvature.

c. Fit the regression plot relating Volume to Girth.

```
f<-lm(Volume~Girth,data=trees);f
```

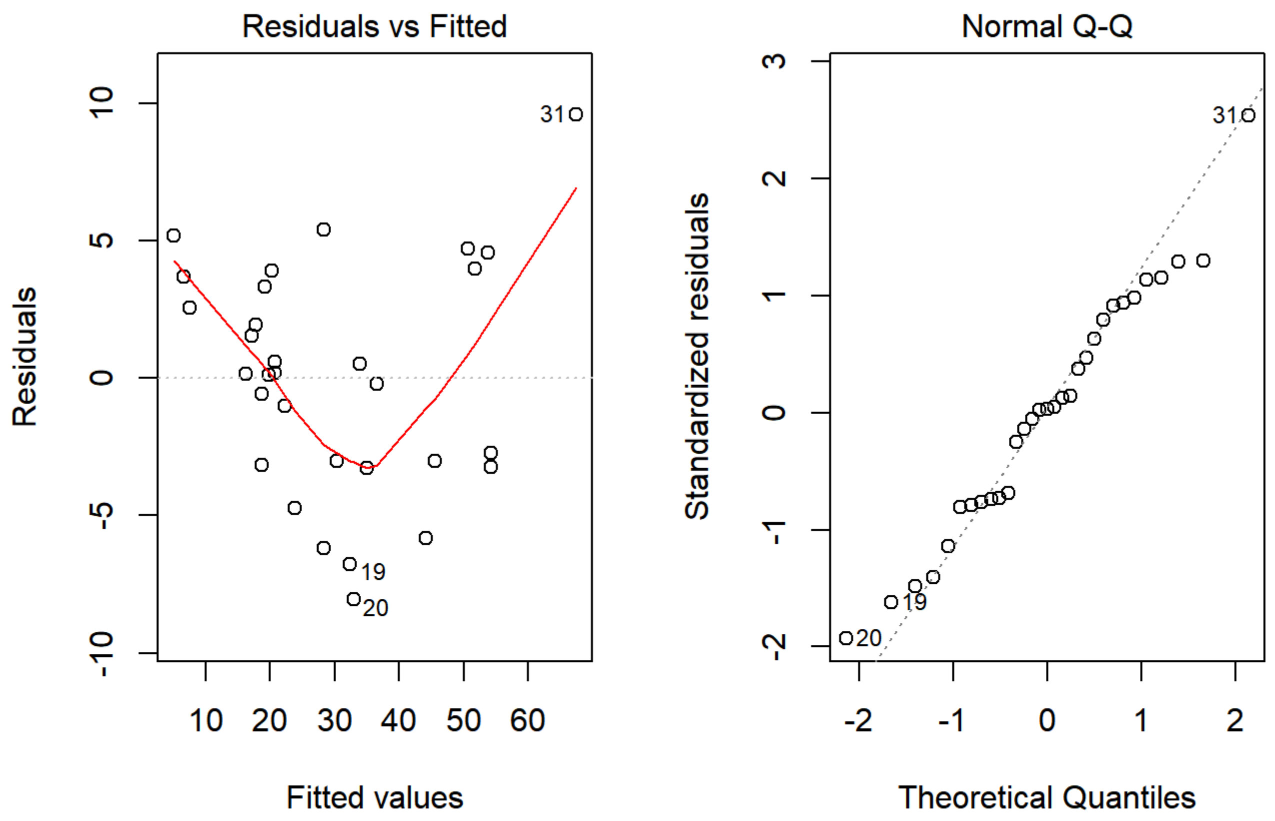
```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Coefficients:
## (Intercept)        Girth
##     -36.943        5.066
```

```
summary(f)
```

```
##
## Call:
## lm(formula = Volume ~ Girth, data = trees)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.065 -3.107  0.152  3.495  9.587
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
## Girth         5.0659     0.2474   20.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.252 on 29 degrees of freedom
## Multiple R-squared:  0.9353, Adjusted R-squared:  0.9331
## F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Construct the residuals versus fitted plot and the Normal plot of residuals.

```
par(mfrow=c(1,2))
plot(f,which=1:2)
```



In light of these plots, assess the model assumptions.

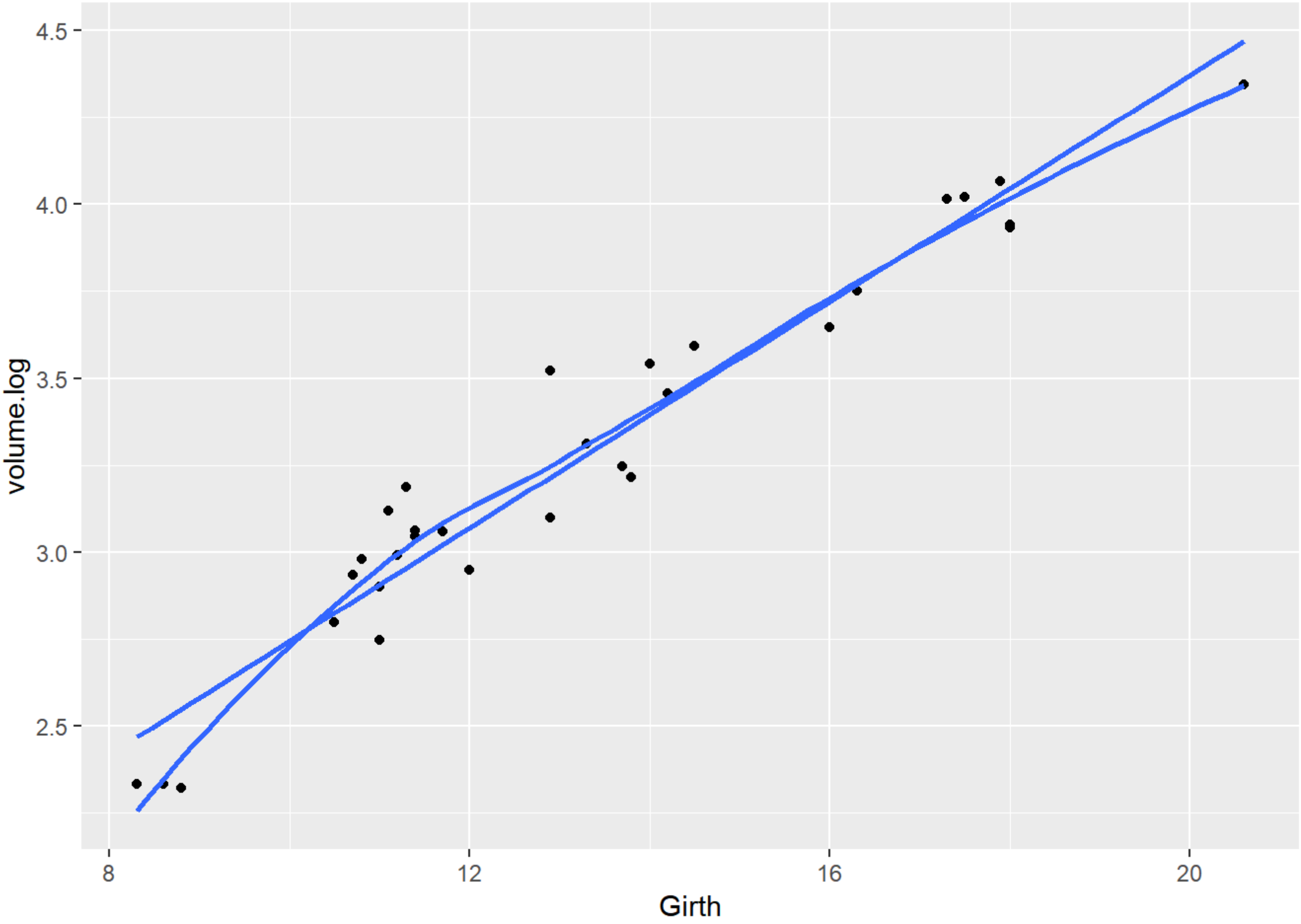
The above plots indicate that a significant amount of curvature exists in the data and that the residuals do not appear to follow a Normal distribution.

d. Construct the plot of part (b),

e. taking a log transformation of y

```
volume.log<-log(trees$Volume)
ggplot(data=trees,aes(x=Girth,y=volume.log)) +
  geom_point() +
  geom_smooth(method="lm",se=F) +
  geom_smooth(se=F)
```

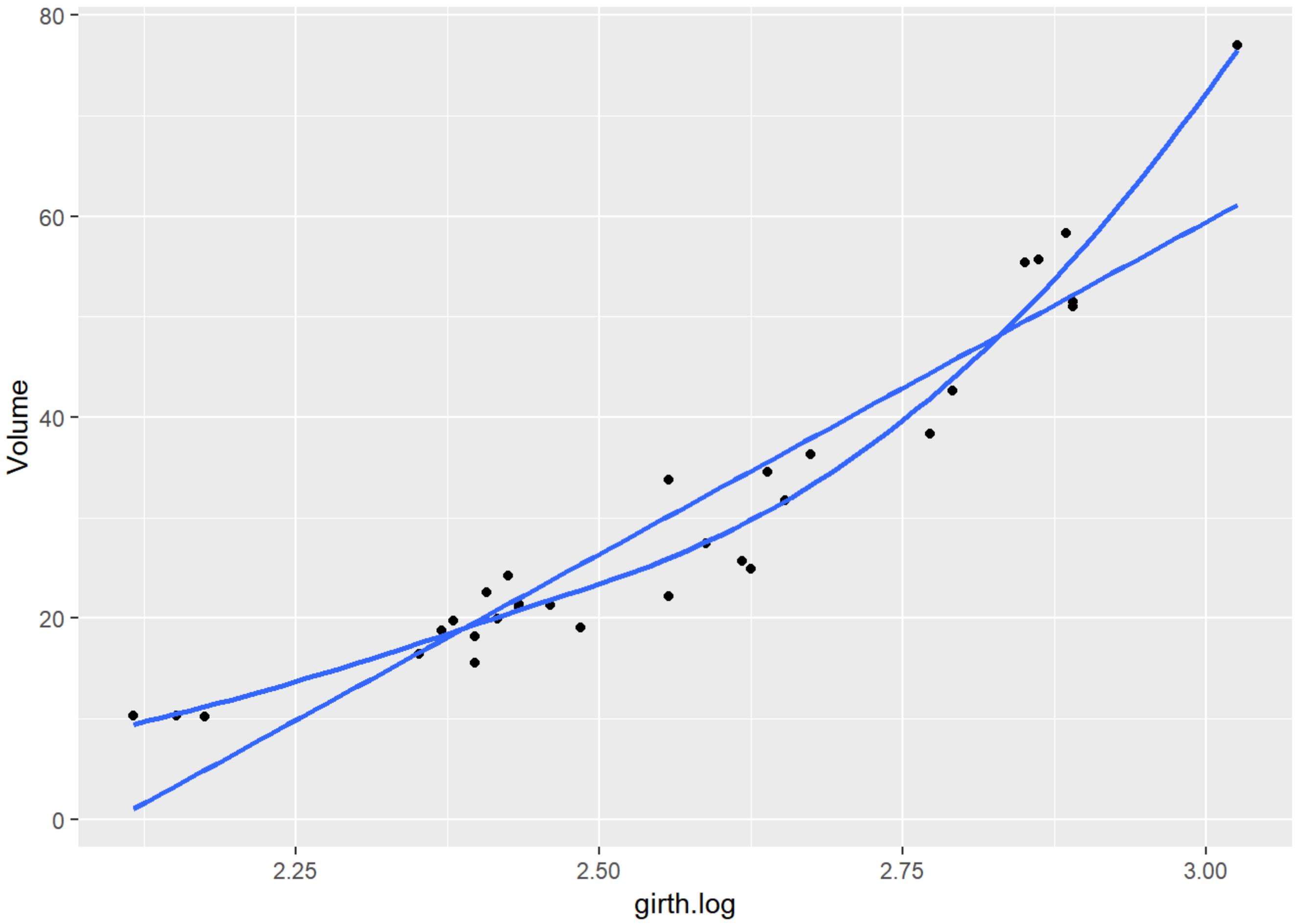
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



ii. taking a log transformattion of x and

```
girth.log<-log(trees$Girth)
ggplot(data=trees,aes(x=girth.log,y=Volume)) +
  geom_point() +
  geom_smooth(method="lm",se=F) +
  geom_smooth(se=F)
```

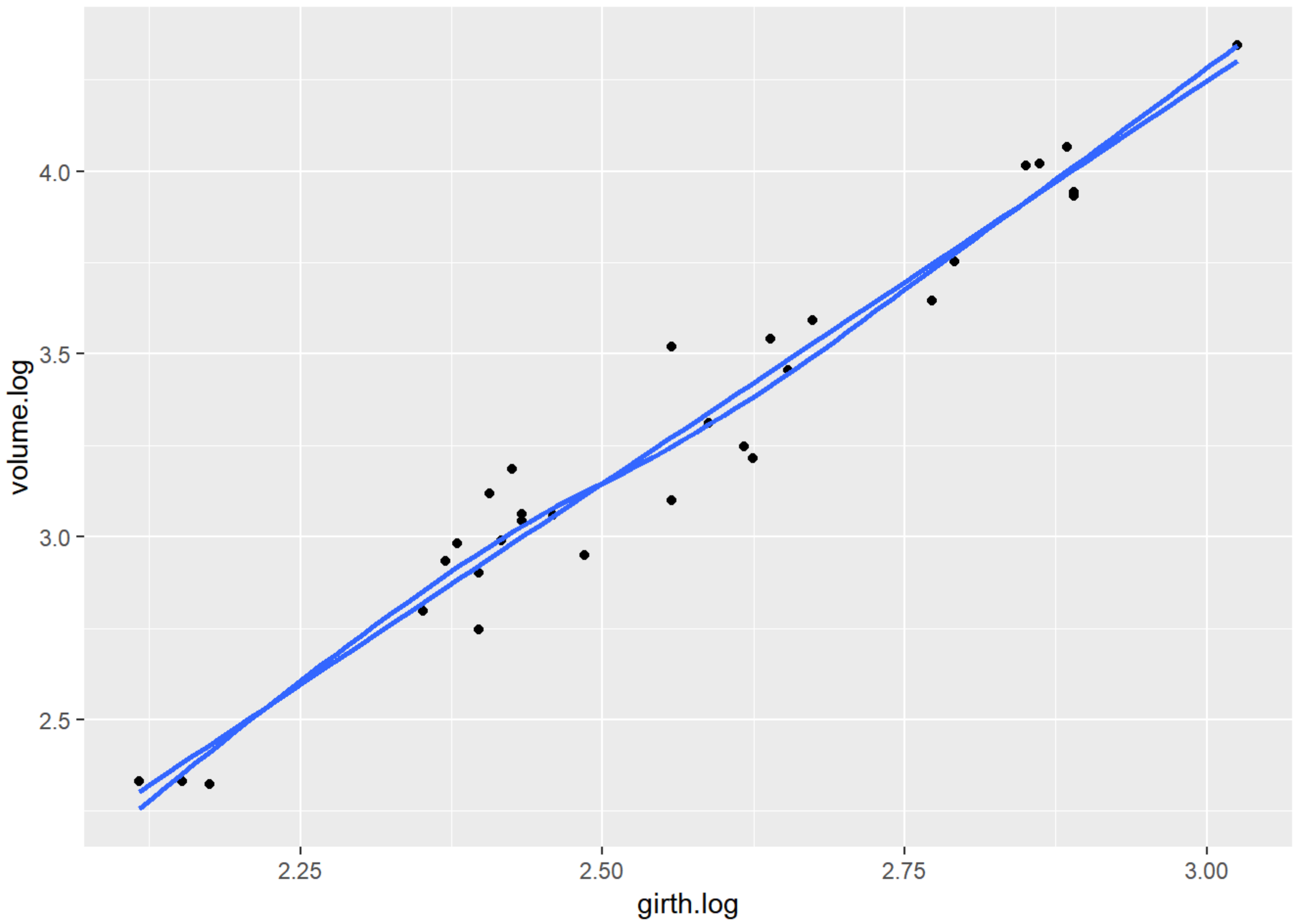
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



iii. taking a log transformation of x and y.

```
ggplot(data=trees,aes(x=girth.log,y=volume.log)) +
  geom_point() +
  geom_smooth(method="lm",se=F) +
  geom_smooth(se=F)
```

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



In each case superimpose a straight line fit and a smooth.

Which plot looks the most linear?

Plot iii) above appears to be the most linear one.

e. Fit the linear model corresponding to your favourite plot of part (d). Construct the residuals versus fitted plot and the Normal plot of residuals.

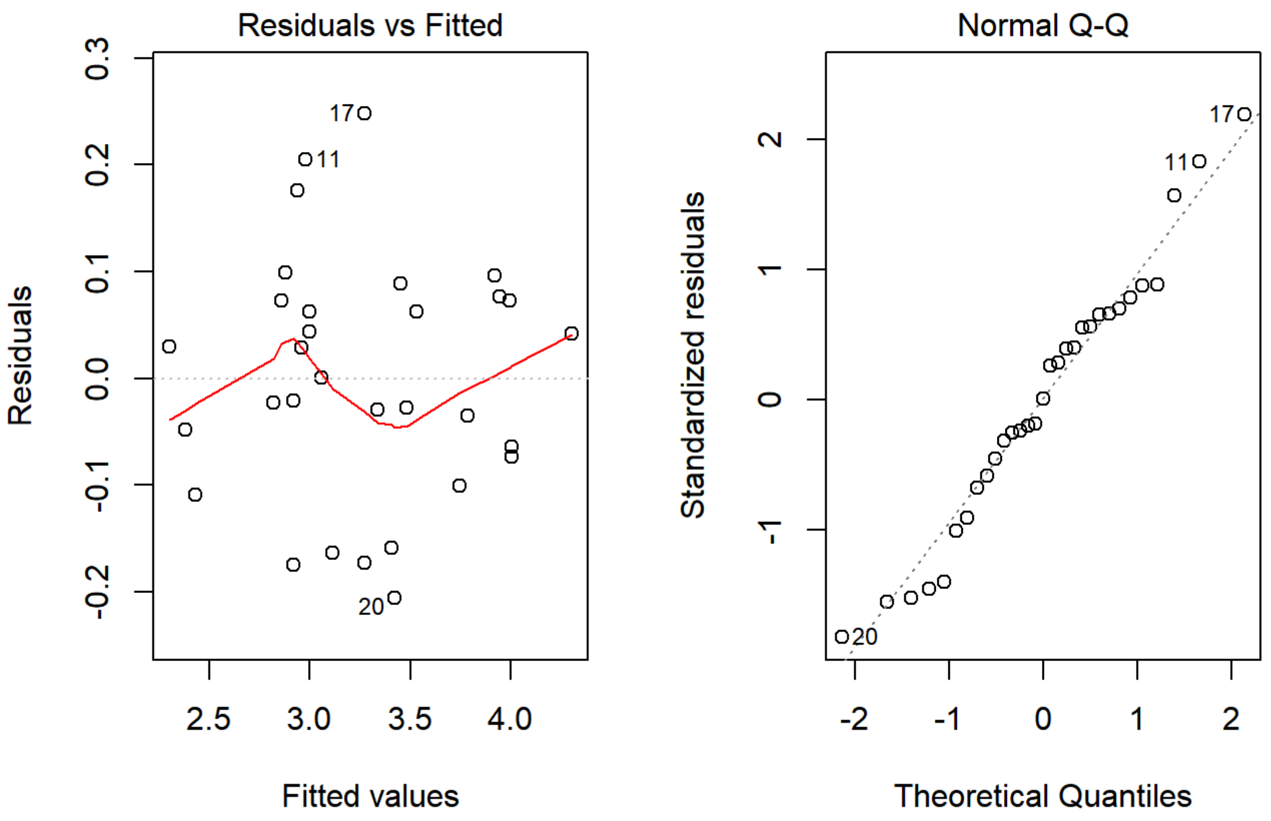
```
f<-lm(volume.log~girth.log);f
```

```
##
## Call:
## lm(formula = volume.log ~ girth.log)
##
## Coefficients:
## (Intercept)    girth.log
##      -2.353        2.200
```

```
summary(f)
```

```
##
## Call:
## lm(formula = volume.log ~ girth.log)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.205999 -0.068702  0.001011  0.072585  0.247963
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.35332    0.23066  -10.20 4.18e-11 ***
## girth.log    2.19997    0.08983   24.49  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.115 on 29 degrees of freedom
## Multiple R-squared:  0.9539, Adjusted R-squared:  0.9523
## F-statistic: 599.7 on 1 and 29 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(1,2))
plot(f,which=1:2)
```



In light of these plots, assess the model assumptions.

f. Use the fit of part (e) to predict and obtain a prediction interval for the Volume of a tree whose Girth is 12.

(Note if your fit uses log(y) you will need to calculate exp(p), if p is your prediction).

```
log12<-log(12)

p<-predict(f,data.frame(girth.log=log12),interval="prediction");p
```

```
##        fit      lwr      upr
## 1 3.113395 2.874147 3.352643
```

```
p.exp<-exp(p);p.exp
```

```
##        fit      lwr      upr
## 1 22.49729 17.71032 28.57815
```

Interpret your result.

We are 95% confident that a tree with a girth measurement of 12 would result in a volume measurement in the 22.49-28.58 range.