Regression Week 1 Question Sheet Solutions

Question 1

The `cars` data are built into R. They describe the speed in miles per hour and stopping distances in feet of 50 cars from the 1920s. To load the data run the following:

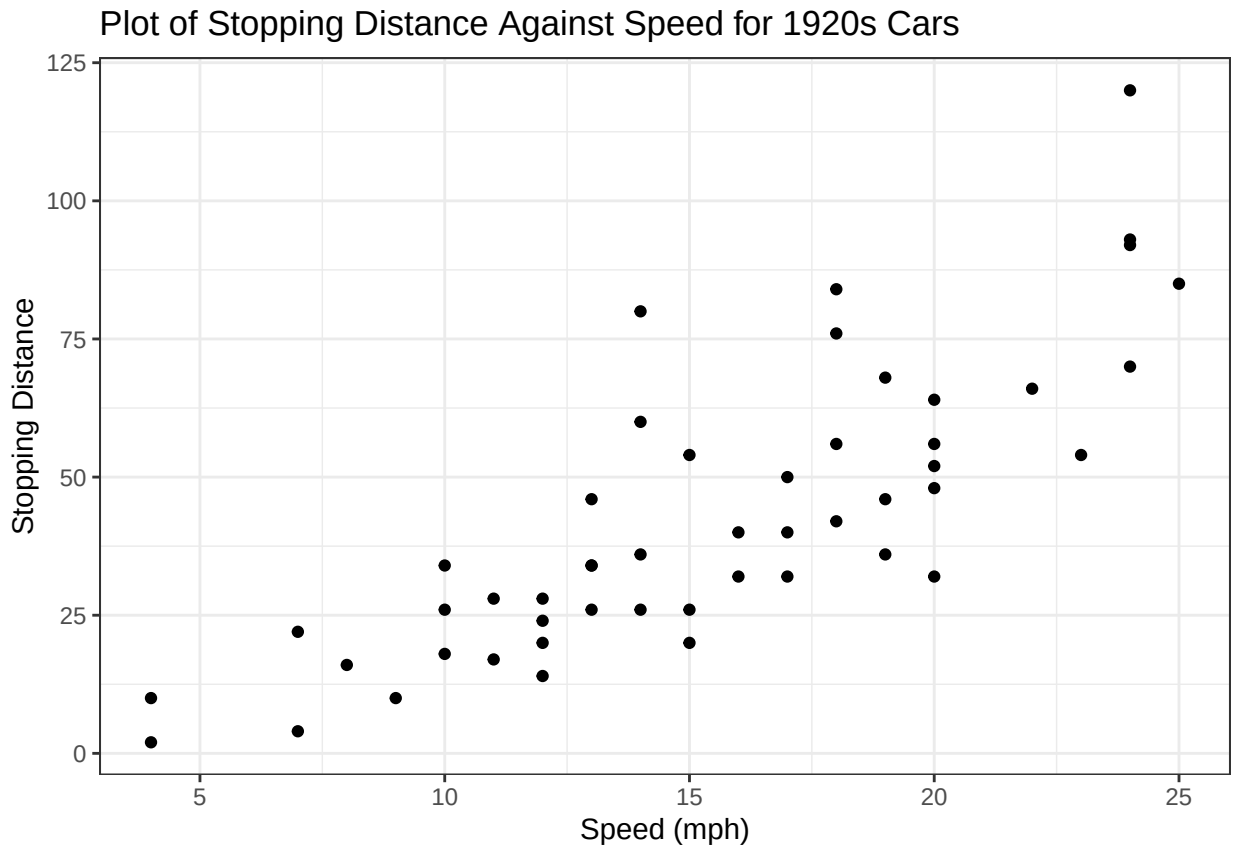
```
data(cars)
```

Consider the following questions:

1. Does a simple linear regression model look appropriate for these data?

```
Want to visualise the relationship between the speed and stopping distance so need to
use a scatterplot. Stopping distance will be our response - this comes from context since
the speed is likely to affect stopping distance (and speed will not be affected by the
stopping distance).
```

```
ggplot(cars) +
  geom_point(aes(speed, dist)) +
  labs(y="Stopping Distance",
       x="Speed (mph)",
       title="Plot of Stopping Distance Against Speed for 1920s Cars") +
  theme_bw()
```



The points look as though they follow a straight line - which is a good indication that linear regression will work

2. What stopping distance might you expect for a similar car travelling at 21.5mph?

A linear regression can be used for prediction - i.e. finding the expected value of the response for some observation of the covariate

```
# fit the linear regression - formula notation
distvspeed <- lm(dist~speed, data=cars)

# Print the estimated model fit
summary(distvspeed)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
```

```
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

```
# Find the prediction

# Option 1: Plug value into the estimated line
-17.5791 + 3.9324*21.5
```

```
## [1] 66.9675
```

```
# Option 2: use predict()
# Better - especially for large models
predict(distvspeed, newdata=data.frame(speed=21.5))
```

```
##        1
## 66.96769
```

When using the predict function the first argument is the model the you want to predict from and the second argument (newdata) is a data frame containing the covariate values that I want to predict at. Note that speed could be a vector if I needed it to be, or the data frame could have multiple columns. We'll revisit this function in week 7

3. Estimate the error variance σ^2 for a simple regression between the two variables.

The error variance is RSS/(n-p-1). This can be calculated in R as follows:

```
sum(resid(distvspeed)^2)/(nrow(cars) - length(coef(distvspeed)))
```

```
## [1] 236.5317
```

Breaking that down:

-resid(distvspeed) extracts the residuals we need to square these and take the sum

- nrow(cars) find the number of observations n
- coef(distvspeed) extracts the model coefficients. length() finds the number of coefficients (p+1) - we can subtract this from nrow(cars) to get n-p-1 by noting that n-p-1 = n-(p+1)

The residual variance can be taken directly from the output:

```
summary(distvspeed)
```

```
##
## Call:
## lm(formula = dist ~ speed, data = cars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -29.069  -9.525  -2.272   9.215  43.201
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.5791     6.7584  -2.601   0.0123 *
## speed         3.9324     0.4155   9.464 1.49e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.38 on 48 degrees of freedom
## Multiple R-squared:  0.6511, Adjusted R-squared:  0.6438
## F-statistic: 89.57 on 1 and 48 DF,  p-value: 1.49e-12
```

In the third line from the bottom it says Residual standard error: 15.38 on 48 degrees of freedom. Squaring 15.38 gives us sigma hat squared

Question 2

Scottish physicist, James D. Forbes collected data at various locations in the Alps and in the Scottish Highlands in 1857, as part of an experiment in measuring altitude using the boiling point of water. It was supposed that a linear relationship existed between Boiling Point (in degrees Fahrenheit) and the log Pressure (in inches of mercury), and that as a result the Atmospheric Pressure, and hence the Altitude, could be estimated simply using the Boiling Point of water. These data can be obtained by running the following in R:

```r
# load the vector of boiling point temperatures (Farenheit)
BP <- c(194.5, 197.9, 199.4, 200.9, 201.4, 203.6, 209.5, 210.7, 212.2,
194.3, 198.4, 199.9, 201.1, 201.3, 204.6, 208.6, 211.9)

# load the vector of pressure values (inches of mercury)
Pressure <- c(20.79, 22.40, 23.15, 23.89, 24.02, 25.14, 28.49, 29.04,
30.06, 20.79, 22.67, 23.35, 23.99, 24.01, 26.57, 27.76, 29.88)

# combine the data into a single data frame, and log-transform the pressure
boiling <- data.frame(logPressure = log(Pressure), BP = BP)

# have a look at the first few rows
head(boiling)
```

```
##   logPressure    BP
## 1    3.034472 194.5
## 2    3.109061 197.9
## 3    3.141995 199.4
## 4    3.173460 200.9
## 5    3.178887 201.4
## 6    3.224460 203.6
```

Explore the relationship between log pressure and boiling point. In particular:

1. What is the estimated least squares regression line?

Again we need to look at the context to find that log pressure is our response we want to estimate log pressure using boiling point and then use log pressure to estimate the altitude.

```
boil_mod <- lm(logPressure~BP, data=boiling)

summary(boil_mod)
```

```
##
## Call:
## lm(formula = logPressure ~ BP, data = boiling)
##
## Residuals:
##        Min         1Q     Median         3Q        Max
## -0.0073622 -0.0033863 -0.0015865  0.0004322  0.0313139
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.9708662  0.0769377  -12.62 2.17e-09 ***
## BP           0.0206224  0.0003789   54.42  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.00873 on 15 degrees of freedom
## Multiple R-squared:  0.995,  Adjusted R-squared:  0.9946
## F-statistic:  2962 on 1 and 15 DF,  p-value: < 2.2e-16
```

The estimated regression line is -0.970908662 + 0.0206224BP (Do calculations to as many decimal places as you have, but feel free to round when presenting coefficients)

2. What is the expected pressure when the boiling point is 195F?

Plug 195 into the estimated regression line: -0.970908662 + 0.0206224*195 = 3.050459 This gives the log pressure. The pressure is then exp(3.050459) = 21.125 (to 3 D.P.)

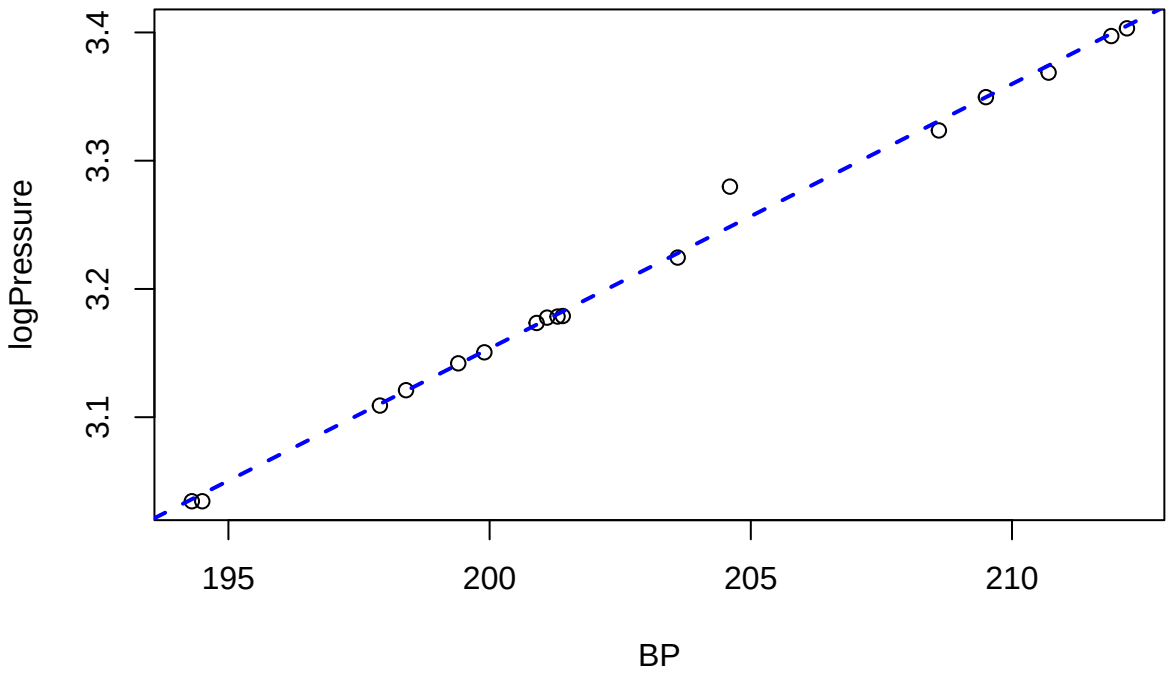
3. Provide and estimate of the error variance σ^2

You could use R as above to extract this, or you could take this directly from the output 0.00873^2 = 7.621e-5

4. Plot `BP` against `Pressure` and add the estimated regression line. Do you think this provides a reasonable fit?

Feel free to do this in base R:

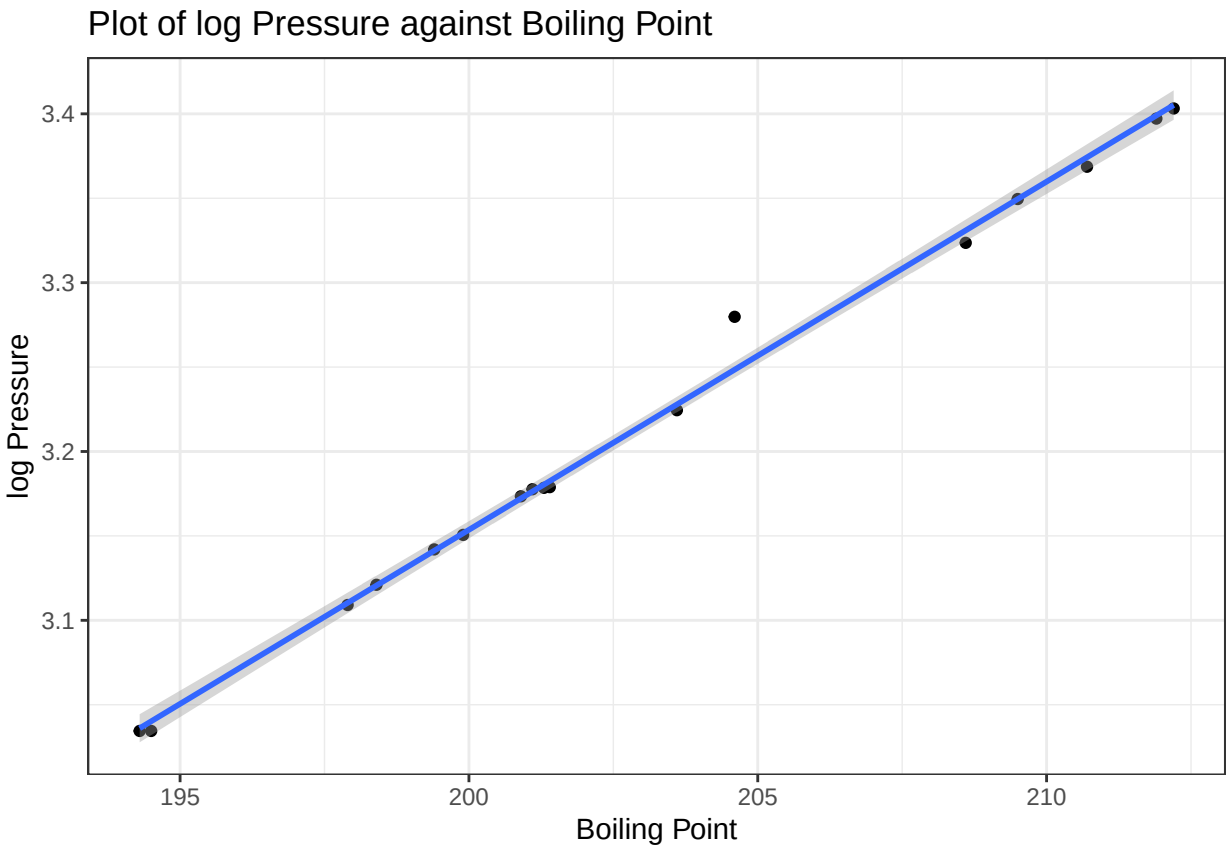
```
plot(logPressure~BP, data=boiling)
abline(boil_mod, col="blue", lty=2, lwd=2)
```



or in ggplot

```
ggplot(boiling, aes(BP, logPressure)) +
  geom_point() +
  # setting method="lm" will add the fitted linear regression
  geom_smooth(method="lm") +
  labs(x="Boiling Point", y="log Pressure",
       title="Plot of log Pressure against Boiling Point") +
  theme_bw()
```

```
## `geom_smooth()` using formula 'y ~ x'
```



No matter which plot type you choose the data are fit by a linear model very well. Notice that we were able to transform one of the variables (Pressure) to get a linear regression. We'll see more of this in Week 8.

Question 3

The `trees` data set provides measurements of the diameter, height and volume of timber in 31 felled black cherry trees. These data can be used in R by running the code

```
data(trees)
```

Then:

1. Find the correlation between `Volume`, `Height` and `Girth`.

```
# Use the cor function
cor(trees)
```

```
##            Girth    Height    Volume
## Girth  1.0000000 0.5192801 0.9671194
## Height 0.5192801 1.0000000 0.5982497
## Volume 0.9671194 0.5982497 1.0000000
```

```
Girth and Volume have much stronger correlation than Height and Volume
```

2. Perform separate simple regressions for `Volume` against each of `Height` and `Girth`. Note the estimated variance `\hat{\sigma}^2` in each case.

Volume~Height

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

summary(height_mod)
```

```
##
## Call:
## lm(formula = Volume ~ Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.274  -9.894  -2.894  12.068  29.852
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -87.1236    29.2731  -2.976 0.005835 **
## Height        1.5433     0.3839   4.021 0.000378 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 13.4 on 29 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3358
## F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784
```

```
# Can also extract sigma-squared
summary(height_mod)$sigma^2
```

```
## [1] 179.4791
```

Volume~Girth

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

summary(girth_mod)
```

```
##
## 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
```

```
# Can also extract sigma-squared
summary(girth_mod)$sigma^2
```

```
## [1] 18.0794
```

Notice that the residual variance is much smaller for the model with girth compared to that for height. This is expected because of the strength of the correlation between girth and volume.

3. Produce scatter plots with the fitted regression lines. Do these relationships look linear?

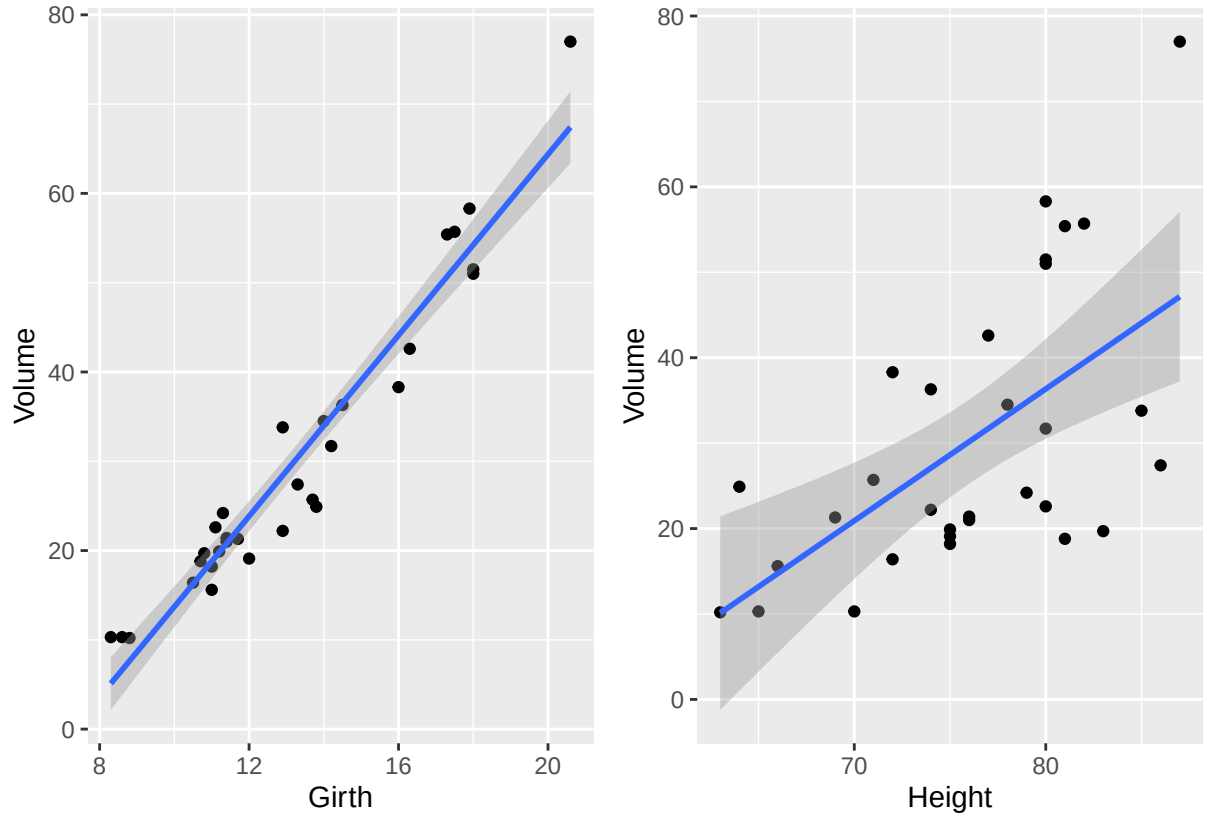
Again feel free to use base R - I've used ggplot:

```
ggplot(trees, aes(Girth, Volume)) +
  geom_point() +
  geom_smooth(method="lm") ->p1

ggplot(trees, aes(Height, Volume)) +
  geom_point() +
  geom_smooth(method="lm") ->p2

require(patchwork)
```

```
## Loading required package: patchwork
```

```
p1+p2
```

```
## `geom_smooth()` using formula 'y ~ x'

## `geom_smooth()` using formula 'y ~ x'
```



Side note: Notice how much closer to the line the points are in the plot of volume against girth compared to that for height - that's why the residual variance is so much smaller for girth compared to height. That's what the residual variance is measuring - how spread out are the points around the fitted line.

There is no obvious non-linear trend between height and volume. There may be a slight curvature in the plot between girth and volume but it isn't obvious (and may be driven by the fact that I am thinking about the relationship between diameter, height and volume in a cylinder!) - based on this plot I would be happy with a straight line fit.

4. Fit a multiple linear regression for `Volume` that contains both `Girth` and `Height`. Calculate the sums of squares and construct an ANOVA table.

```
# separate covariates with + in the formula
mult_mod <- lm(Volume~Girth+Height, data=trees)

summary(mult_mod)
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

```
dfR <- nrow(trees) - length(coef(mult_mod))
dfM <- length(coef(mult_mod))-1
dfT <- nrow(trees) - 1

RSS <- sum(resid(mult_mod)^2)
TSS <- sum((trees$Volume - mean(trees$Volume))^2)
MSS <- TSS - RSS

MSR <- RSS/dfR
MSM <- MSS/dfM

F <- MSM/MSR
```

The table will then be:

	DF	SS	MS	F
Regression	2	7684.16	3842.08	254.97
Residual	28	421.92	15.07	
Total	30	8106.08		

5. Use your ANOVA table to carry out an F-test. You will need to use either `qf()` or `pf()` to find the critical value or the p-value for the test.

If using qf() you reject H0 if the calculated F is greater than the critical value. If using pf() you reject H0 if the output values is small (<0.05 usually)

```
# qf()
# first argument (p) is 1-significance level
# next two arguments are the degrees of freedom
qf(0.95, 2, 28)
```

```
## [1] 3.340386
```

```
# pf()
# first argument is F
# then degrees of freedom
# then lower.tail=FALSE says you want P(X>x)
pf(F, 2, 28, lower.tail=FALSE)
```

```
## [1] 1.071238e-18
```

Our observed F is large in comparison to the critical value obtained from qf() and our p-value is very small so we reject H0, which indicates that there is a linear association between Volume and at least one of Girth and Height.

6. Write down the estimated regression equation and hence interpret the effect of `Girth` of `Volume`.

```
summary(mult_mod)
```

```
##
## Call:
## lm(formula = Volume ~ Girth + Height, data = trees)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.4065 -2.6493 -0.2876  2.2003  8.4847
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
## Girth         4.7082     0.2643  17.816  < 2e-16 ***
## Height        0.3393     0.1302   2.607   0.0145 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.882 on 28 degrees of freedom
## Multiple R-squared:  0.948,  Adjusted R-squared:  0.9442
## F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

The estimated regression line is Volume = -57.9877 + 0.47082Girth + 0.3393Height. This means that for a constant height, tree volume would increase by 4.7082 cubic feet on average for every increase of 1 inch in the tree diameter.

Question 4

The following data describe the marks (out of 100) of ten students in mathematics and physics. We wish to use the mathematics mark of a student to predict their physics mark.

Student	1	2	3	4	5	6	7	8	9	10
Mathematics (X)	17	56	34	83	90	65	42	47	51	59
Physics (Y)	25	62	47	80	88	72	44	47	59	63

An incomplete ANOVA table for the line of regression of Y against X is:

	DF	SS	MS	F
Regression	?	3019.79	?	?
Residual	?	?	?	
Total	?	3144.1		

1. Complete the ANOVA table.

	D.F.	S.S.	M.S.	F
REGRESSION	**1**	3019.79	**3019.79**	**194.3393**
RESIDUAL	**8**	**124.31**	**15.53875**	
TOTAL	**9**	3144.1		

2. State the null and alternative hypotheses associated with the F-test.

$H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$
Where β_1 is the true (population) slope associated with the linear influence of mathematics score on the physics score.

3. What would you conclude based on the following R command and output:

```
qf(p = 0.95, df1 = 1, df2 = 8)
```

```
## [1] 5.317655
```

- Observed F-statistic is 194.34
- Critical value $F_{1,n-2} = 5.31$
- Since 194.32 >> 5.31, reject H_0 and conclude strong evidence that there is significant linear association between maths and physics marks. (Note this is the interpretation because this is simple linear regression - for multiple linear regression we would say there is significant linear association with at least one of the independent variables.)