5.1. Weighted Least Squares

In this chapter we consider Weighted Least Squares regression as a

- technique in its own right
- as a method of assessing lack of fit of regression models

Weighted Least Squares

In the simple linear model ($y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \text{NID}(0, \sigma^2)$) we assume that the values x_i are known **constants.**

i.e. we set $X = x_i$ and observe the response y_i.
Hence $\text{var}(y_i) = \text{var}(\beta_0 + \beta_1 x_i + e_i) = \text{var}(\beta_0) + \text{var}(\beta_1 x_i) + \text{var}(e_i) = 0 + 0 + \text{var}(e_i) = \sigma^2$.

i.e. the response variables y_i have **common variance**. However, there are situations when the response variables y_i do **not** have the same variance.

Example 1:
If the i[th] response y_i is an **mean** of n_i equally variable observations, then

$$\textbf{var}(y_i) = \sigma^2/n_i$$

Recall: If $X_1, X_2, \ldots X_n$ are independent with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$, then

$$\text{var}(\bar{X}) = \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\text{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\left(\sum_{i=1}^{n}\text{var}\, X_i\right) = \frac{1}{n^2}\left(n\sigma^2\right) = \frac{\sigma^2}{n}$$

Example 2:
If the i[th] response y_i is the **total** of n_i equally variable observations, then

$$\textbf{var}(y_i) = n_i\,\sigma^2$$

Recall: If $X_1, X_2, \ldots X_n$ are independent with $E(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$, then

$$\text{var}\left(\sum_{i=1}^{n} X_i\right) = \left(\sum_{i=1}^{n}\text{var}\, X_i\right) = n\sigma^2$$

Example 3:
If the variance of the i[th] response y_i is proportional to some predictor x_i, then

$$\textbf{var}(y_i) = x_i\sigma^2$$

All three examples are particular cases of the **weighted** least squares model:

$y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \text{NID}(0, \sigma^2/ w_i)$, where $\text{var}(y_i) = \text{var}(e_i) = \sigma^2/ w_i$ with **weights** $w_i > 0$.

(Model 1)

If the ith response y_i is an **average** of n_i equally variable observations, then

$\text{var}(y_i) = \sigma^2/n_i = \sigma^2/ w_i$, then $\boldsymbol{w_i = n_i}$

If the ith response y_i is the **total** of n_i equally variable observations, then

$\text{var}(y_i) = n_i \sigma^2 = \sigma^2/ w_i$, then $\boldsymbol{w_i = 1/n_i}$

If the variance of the ith response y_i is **proportional** to some predictor x_i, then

$\text{var}(y_i) = x_i\sigma^2 = \sigma^2/ w_i$, then $\boldsymbol{w_i = 1/x_i}$

Video 5.1

The **weighted** least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are defined to be the estimates of β_0 and β_1 which minimise the **weighted** sum of squared residuals, $\sum w_i\hat{e}_i^2$, where $\hat{e}_i = y_i - \hat{y}_i$.

In ordinary least squares, the least squares estimates are chosen to minimise the sum of squared residuals $\sum \hat{e}_i^2$,

i.e. each residual $\hat{e}_i = y_i - \hat{y}_i$ is given the same weight, since the response variables (y_i) have the same variance.

Here, $\text{var}(y_i) = \sigma^2/ w_i$, so if w_i is **large**, then y_i has **small** variance and **greater** weight should be given to the deviation $\hat{e}_i = y_i - \hat{y}_i$.

Similarly, if w_i is **small**, then y_i has **large** variance and **less** weight should be given to the deviation $\hat{e}_i = y_i - \hat{y}_i$.

This is achieved by choosing least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ which minimise the **weighed** sum of squared residuals $\sum w_i\hat{e}_i^2$.

The **weighted** regression line will, in general, be **close** to points with **large weights** and **not so close** to points **with small weights**.

We solve the weighted regression model ($y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim$ NID(0, $\sigma^2/ \boldsymbol{w_i}$)) by reducing it to a simple linear model in which the response variables do have constant variance.

Multiply $y_i = \beta_0 + \beta_1 x_i + e_i$ across by $\sqrt{w_i}$ to get:

$$\sqrt{w_i}\, y_i = \beta_0 \sqrt{w_i} + \beta_1 \sqrt{w_i}\, x_i + d_i \qquad \text{where } d_i = \sqrt{w_i}\, e_i \sim \text{NID}(0, \sigma^2). \qquad \text{(Model 2)}$$

In this model, the response variables $z_i = \sqrt{w_i}\, y_i$ are regressed on $\sqrt{w_i}$ and $\sqrt{w_i}\, x_i$ (with no intercept term) and the error variables d_i in this model **do have constant variance**.

The least squares estimates of β_0 and β_1 in this model are obtained using ordinary least squares and are the same as the **weighted** least squares estimates of β_0 and β_1 in Model 1.

Summary

In the weighted linear regression model $y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \text{NID}(0, \sigma^2/ \boldsymbol{w_i})$,

the weighted least squares estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to minimise the (Weighted) Residual Sum of Squares, $RSS = \sum w_i \hat{e}_i^2$.

Physics Data Set

Physicists ran an experiment in which a large number of particles were aimed at a target. The particles were beamed at the target at very high speed. The outcome measure (Y) was the degree to which the particles scattered on impact. The particles were beamed at various speeds (X). The estimated variances of the y-values could be obtained from principles of theoretical physics. The estimated standard deviations are recorded in the data set.

	X	Y	Est.SD
1	0.345	367	17
2	0.287	311	9
3	0.251	295	9
4	0.225	268	7
5	0.207	253	7
6	0.186	239	6
7	0.161	220	6
8	0.132	213	6
9	0.084	193	5
10	0.060	192	5

The estimated variances of the y_i and thus the e_i are not constant.
Here $\text{var}(y_i) = \text{var}(e_i) = (\text{Estimated sd}(y_i))^2$ $(= \sigma^2/ w_i)$.
Thus $w_i = \sigma^2/(\text{ Estimated sd}(y_i))^2$.

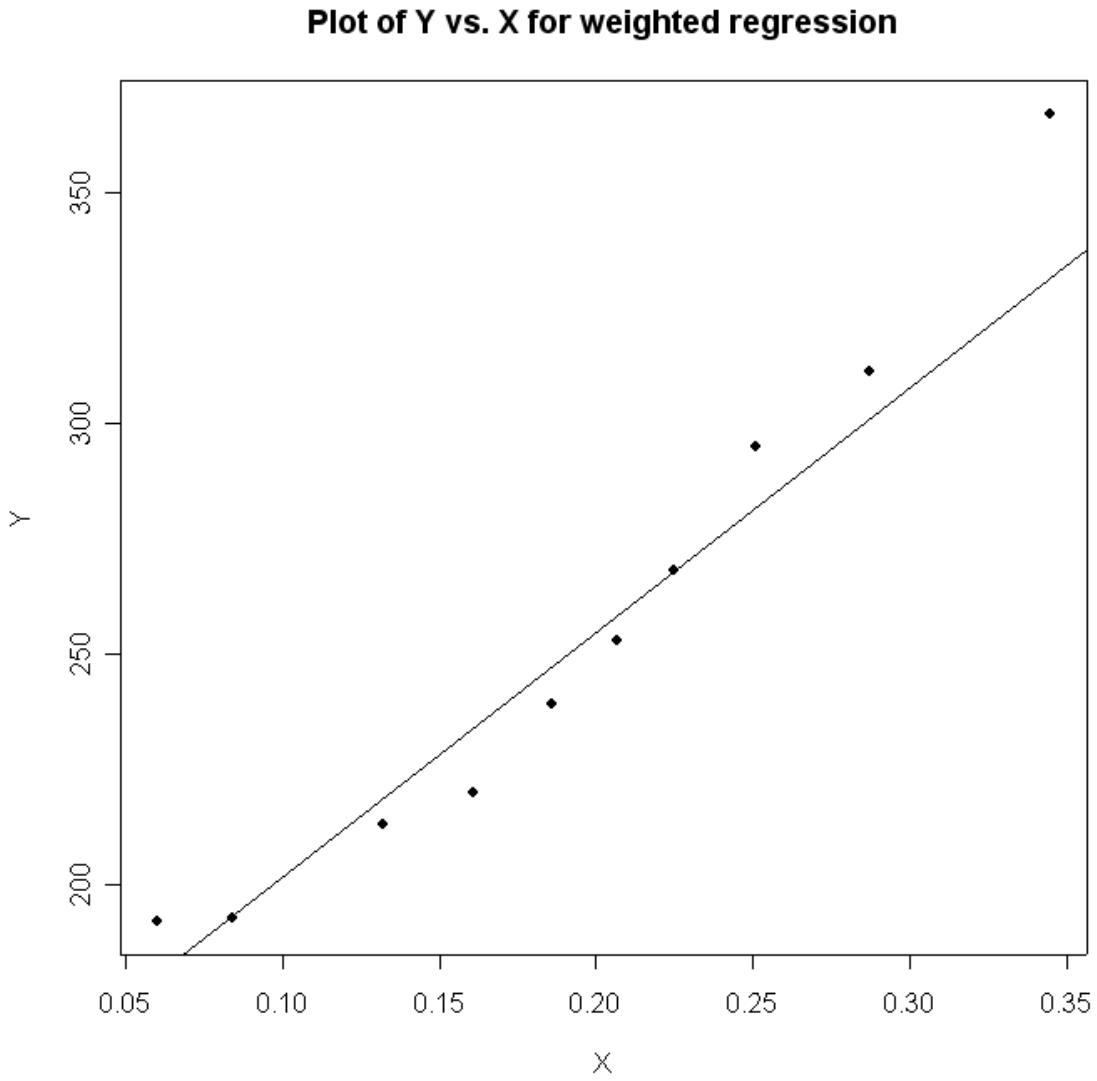
Given the variance is known to be 1.0, $\sigma^2 = 1$, we can fit a weighted regression model

$y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim \text{NID}(0, \sigma^2/ w_i)$, with weights $w_i = 1/(\text{ Estimated sd}(y_i))^2$.

Excerpts from R:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  148.473      8.079   18.38 7.91e-08 ***
X            530.835     47.550   11.16 3.71e-06 ***

Residual standard error: 1.657 on 8 degrees of freedom
Multiple R-Squared: 0.9397,     Adjusted R-squared: 0.9321
F-statistic: 124.6 on 1 and 8 DF,  p-value: 3.710e-06

Analysis of Variance Table
Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 341.99  341.99  124.63 3.710e-06 ***
Residuals  8  21.95    2.74
```

The fitted weighted regression line is:



Note that the weighted regression line is **closer** to the point (0.060, 192) than it is to the point (0.345, 367).

This is because for y = 192, the estimated standard deviation of y is 5, which is **small**, while for y = 367, the estimated standard deviation of y is 17, which is **large**.

Since $w_i = 1/(\text{ Estimated sd}(y_i))^2$;
the weight for y = 192 will be large and the weight for y = 367 will be small.

So the weighted regression line will be pulled **closer** to the point (0.060, 192) than it is to the point (0.345, 367).

Video 5.2

5.2. Testing for Lack of Fit

5.2.1. Variance Known

A model specifies the shape of the relationship between the response variable and the predictor variable(s).

When the chosen shape is correct, the residual mean square gives an unbiased estimate of σ^2, the error variance. That is, $\hat{\sigma}^2 = \frac{\text{RSS}}{\text{n - (p + 1)}}$.

If the chosen shape is incorrect $\hat{\sigma}^2 = \frac{\text{RSS}}{\text{n - (p + 1)}}$ estimates a quantity larger than σ^2, since its size depends on the errors and also the systematic biases due to fitting the incorrect shape.

If σ^2 is known or if a model-free estimate is available, a comparison of $\hat{\sigma}^2$ and σ^2 provides a test for lack of fit.

If $\hat{\sigma}^2$ is too large, we have evidence of a lack of fit. The fitted model is not adequate.

In the case of a simple regression model, we saw that if the model is correct, $\frac{RSS}{\sigma^2} = \frac{(n-2)\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$, (from chapter 1).
Generalising to multiple regression, if the model is correct, $\frac{RSS}{\sigma^2} = \frac{(n-(p+1))\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-(p+1))$.

Thus if σ^2 is known, we can calculate the test statistic, $\frac{RSS}{\sigma^2}$, and test for lack of fit.

The sampling distribution of the test statistic is a chi-square with (n-(p+1)) degrees of freedom.

If the observed value of the test statistic is too big (small p-value), we conclude lack of fit.

(Note: RSS can be requested in R and is known as the Deviance)

Exercise: Testing for lack of fit in the physics data,

RSS = σ^2 = $\frac{RSS}{\sigma^2}$ =

The sampling distribution is with degrees of freedom equal to

The critical value at the 5% level of significance is

Conclusion:

Video 5.3

When the model is inadequate, it is usual to fit alternative models, either by transforming the predictor(s) and/or the response or by adding polynomial terms in the predictors.

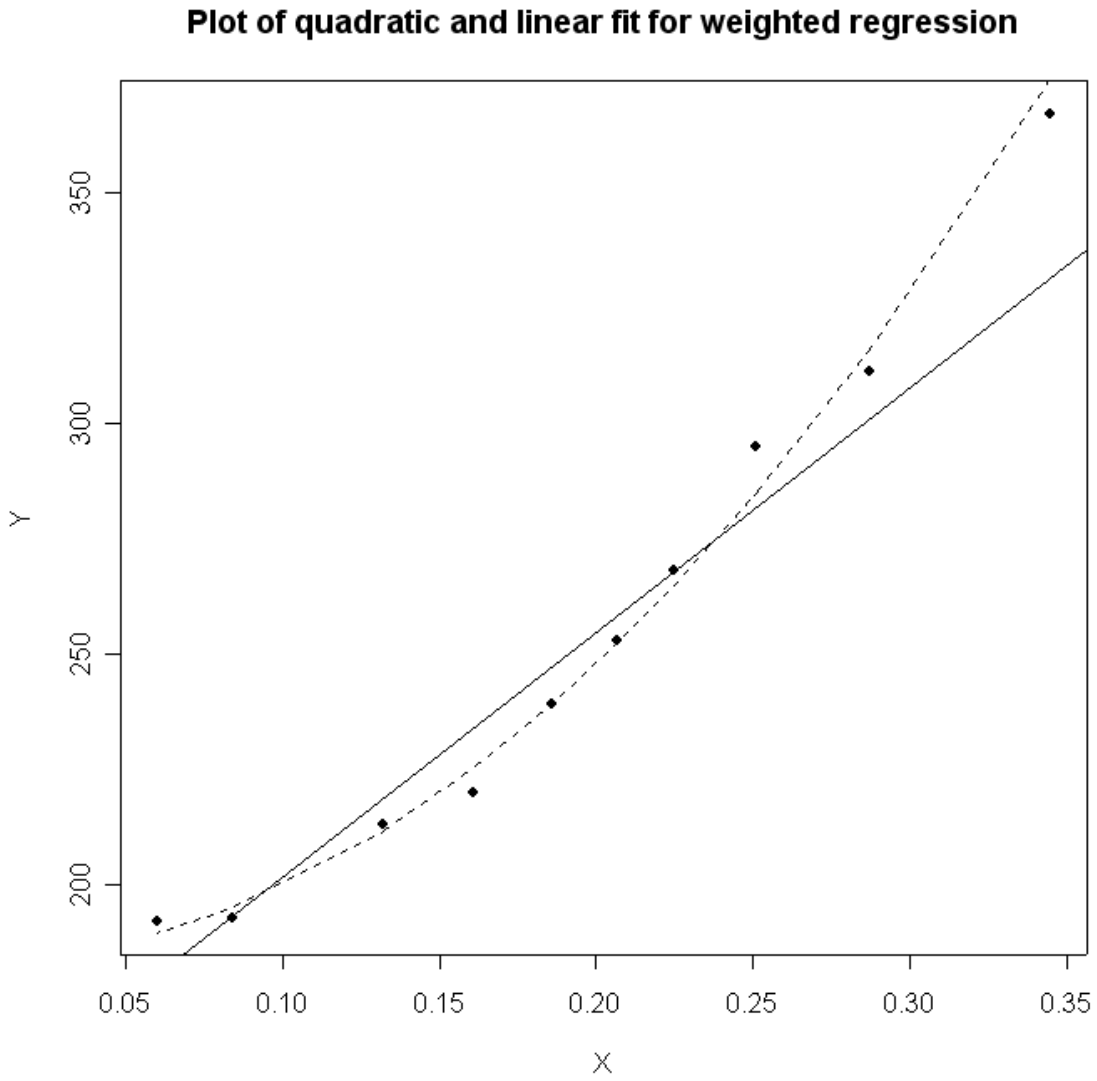
The points in the scatter-plot above suggest that a **quadratic** fit may be more appropriate than a linear fit. The following weighted regression model is fitted

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 \mathbf{x_i^2} + e_i\,,\ e_i \sim \text{NID}(0, \sigma^2/\, w_i),$$ with the same weights w_i as before.

Excerpts from R:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  183.8305     6.4591  28.461  1.7e-08 ***
X              0.9709    85.3688   0.011 0.991243
I(X^2)      1597.5047   250.5869   6.375 0.000376 ***

Residual standard error: 0.6788 on 7 degrees of freedom
Multiple R-Squared: 0.9911,     Adjusted R-squared: 0.9886
F-statistic: 391.4 on 2 and 7 DF,  p-value: 6.554e-08

Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 341.99  341.99 742.185 2.303e-08 ***
I(X^2)     1  18.73   18.73  40.641 0.0003761 ***
Residuals  7   3.23    0.46
```

The fitted weighted regression curve matches the points very closely:



Exercise: Test this model for lack of fit.

RSS = σ^2 = $\frac{RSS}{\sigma^2}$ =

The sampling distribution is with degrees of freedom equal to

The critical value at the 5% level of significance is

Conclusion:

Video 5.4

5.2.2. Variance Unknown

To test for lack of fit when the variance σ^2 is **unknown**, we need an estimate of σ^2 which doesn't depend on the linear model being used, i.e. which is **model-free**. To test our model for fit, we compare the estimate of σ^2 derived from the model with the model-free estimate.

The most common model-free estimate uses the variation between cases with the same values of the predictor variable(s).

Consider the following data:

X	Y	$\bar{y}$	$\sum(y_i-\bar{y})^2$	S.D.	d.f.
1	2.55				
1	2.75				
1	2.57	2.6233	0.0243	0.1102	2
2	2.40	2.4000	0.0000	0.0000	0
3	4.19				
3	4.70	4.4450	0.1301	0.3606	1
4	3.81				
4	4.87				
4	2.93				
4	4.52	4.0325	2.2041	0.8571	3

Recall if $X_1, X_2, \ldots X_n$ are independent with $E(X_i) = \mu$ and $var(X_i) = \sigma^2$, and

$$s^2 = \frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}, \text{ then } E(s^2) = \sigma^2.$$

The sample standard deviation, $SD = s = \sqrt{\frac{\sum_{i=1}^{n}(X_i-\bar{X})^2}{n-1}}$, is an estimate of σ.

Using the y-values for the group of cases with X = 1 to estimate σ^2, gives

$$\bar{y} = \frac{2.55+2.75+2.57}{3} = 2.6233, \sum(y_i-\bar{y})^2 = 0.0243,$$

$$SD = \sqrt{\frac{\sum(y_i-\bar{y})^2}{n-1}} = \sqrt{\frac{0.0243}{2}} = 0.1102, \; SD^2 = (0.1102)^2 = \hat{\sigma}^2, \quad df = 2.$$

This is repeated for each group of cases, giving four estimates of σ.

A **pooled** estimate of the common variance of the y-values is obtained by pooling the SD^2 into a single estimate.

The **Sum of Squares for Pure Error** $SS(pe) = \sum(n-1)SD^2 = \sum\sum(y_i-\bar{y})^2 = 2.3585$

The associated number of degrees of freedom, **Degrees of Freedom for Pure Error**, is

$$\text{df(pe)} = \sum(n-1) = 2+0+1+3 = 6$$

The **pooled** or **pure error estimate of variance** is $\hat{\sigma}^2 = \dfrac{\text{SS(pe)}}{\text{d.f.(pe)}} = \dfrac{2.3585}{6} = 0.3931$

This estimate of σ^2 is **model-free**, since it does not depend on any linear model, but only on the assumptions that the y-values have a **common variance** and are **independent**.

(Note: The Pure Error Sum of Squares occurs in **One-Way Analysis of Variance**, where it is termed the Within-Group Sum of Squares.)

From R, the results a fitting a linear regression model to the data in the table above are:

```
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X          1 4.5693  4.5693   8.669 0.01859 *
Residuals  8 4.2166  0.5271
```

From this we get the model-dependent estimate of σ^2: $\hat{\sigma}^2$ = 0.5271.
This is bigger than the model-free estimate ($\hat{\sigma}^2$ = 0.3931).

If the estimate from the model is **much larger** than the model-free estimate, then the model is inadequate.

We can divide the Residual Sum of Squares (*RSS*) into the Sum of Squares for Pure Error and the remainder called the **Sum of Squares for Lack of Fit**:

RSS = SS(pe) + SS(lof)

SS(lof) = RSS – SS(pe) = 4.2166 –2.3585 = 1.8581.

The corresponding number of degrees of freedom is

df(lof) = df(RSS) – df(pe) = 8 – 6 = 2.

Then, if the null hypothesis of no lack of fit is true, it may be shown that

$$F = \frac{\text{SS(lof)/df(lof)}}{\text{SS(pe)/df(pe)}} \sim F(\text{df(lof)},\text{df(pe)})$$

If we observe a large value of the test statistic (small p-value), we reject the null hypothesis and conclude a lack of fit.

Exercise: Test for lack of fit in this hypothetical data,

RSS = SS(pe) = SS(lof) =

df(RSS) = df(pe) = df(lof) =

$$F = \frac{\text{SS(lof)/df(lof)}}{\text{SS(pe)/df(pe)}} =$$

The critical value at the 5% level of significance is

Conclusion:
Reject/accept hypothesis of no lack of fit.
There is a lack of fit/ There is no lack of fit.

Video 5.5

Note: the data used above was generated from the following model:

$y_i = 2.0 + 0.5\, x_i + e_i$, $e_i \sim$ NID(0,1)

In this model, $\beta_0 = 2.0$, $\beta_1 = 0.5$ and var(y_i) = var(e_i) = $\sigma^2 = 1$.

For X = 1, 2, 3 and 4, values of Y were generated by simulating values from a standard normal distribution.

For X = 1, three such simulated values were $e_1 = 0.05$, $e_2 = 0.25$ and $e_3 = 0.07$.

Then $y_1 = 2.0 + 0.5(1) + 0.05 = 2.55$, $y_2 = 2.75$ and $y_3 = 2.57$.
This was repeated to generate one value of Y for X = 2, two values of Y for X = 3 and four values of Y for X = 4.
As the data was generated from a linear model, if should, unless we are unlucky, have shown no lack of fit.

5.3. Using Weighted Regression to Test for Lack of Fit (in the corresponding un-weighted model)

We can use weighted regression to test for lack of fit if some cases have the same values of the predictor variable(s).

Apple shoots datasets

An experiment was carried out investigating the numbers of stems growing on apple shoots of McIntosh apple trees. The shoots were classified as either long or short. The data provided here are for long apple shoots. The following variables were recorded:

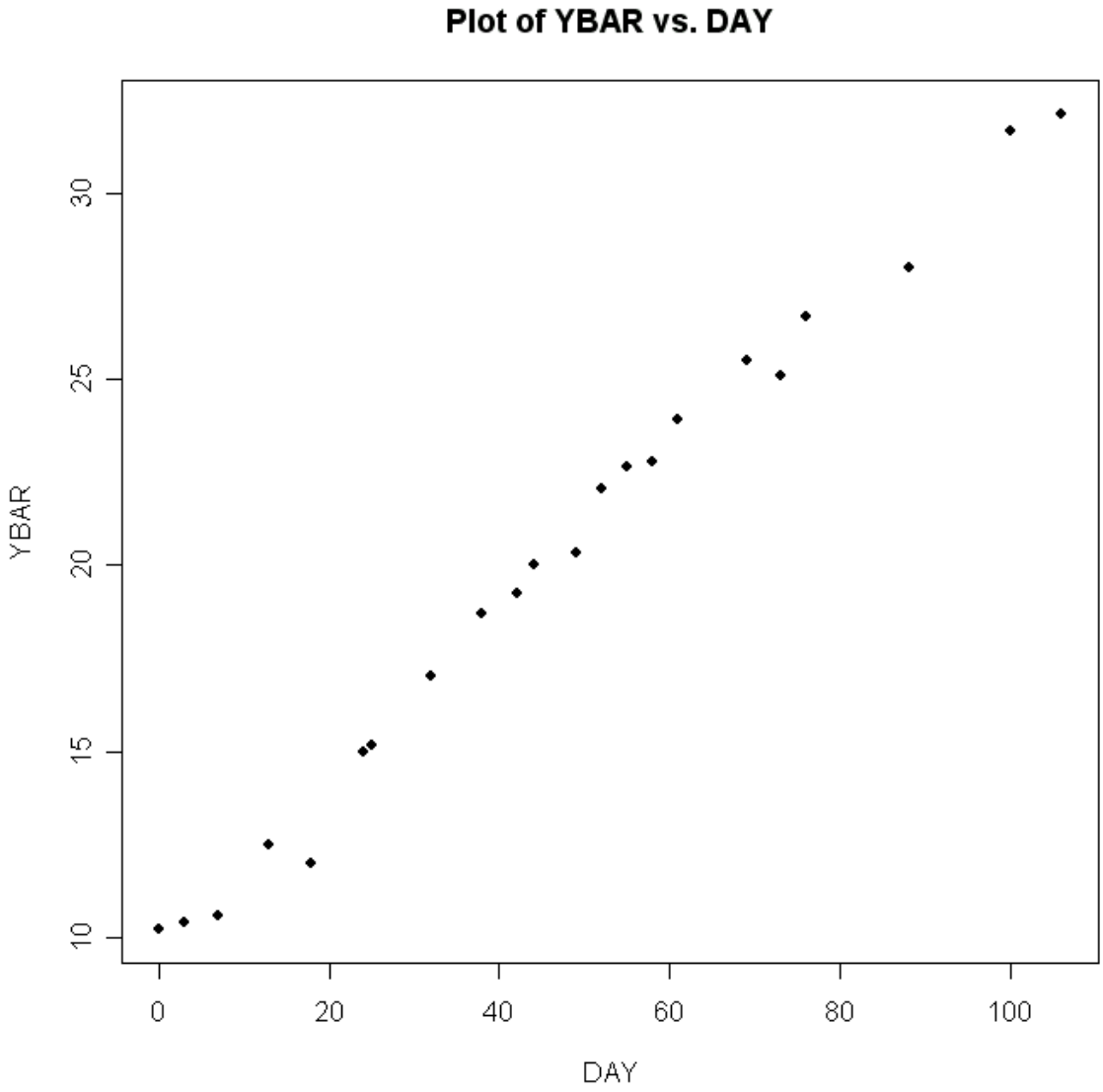
DAY = days from the start of the growing season
N = number of shoots sampled per day
YBAR = average number of stem units per shoot
SD = standard deviation of the number of stem units per shoot

	DAY	N	YBAR	SD
1	0	5	10.20	0.83
2	3	5	10.40	0.54
3	7	5	10.60	0.54
4	13	6	12.50	0.83
5	18	5	12.00	1.41
6	24	4	15.00	0.82
7	25	6	15.17	0.76
8	32	5	17.00	0.72
9	38	7	18.71	0.74
10	42	9	19.22	0.84
11	44	10	20.00	1.26
12	49	19	20.32	1.00
13	52	14	22.07	1.20
14	55	11	22.64	1.76
15	58	9	22.78	0.84
16	61	14	23.93	1.16
17	69	10	25.50	0.98
18	73	12	25.08	1.94
19	76	9	26.67	1.23
20	88	7	28.00	1.01
21	100	10	31.67	1.42
22	106	7	32.14	2.28

In this example**,** the ith response (y_i) is an **average** of n_i equally variable observations, so that var(y_i) = σ^2/n_i.

We use a weighted regression model $y_i = \beta_0 + \beta_1 x_i + e_i$, $e_i \sim$ NID$(0,\sigma^2/ w_i)$, with $\boldsymbol{w_i = n_i}$.

A scatter-plot of $\bar{y}$ vs. DAY is given below. It suggests a strong linear relationship.



Since $\text{var}(\bar{y}) = \sigma^2/n$, we fit a **weighted** regression model of $\bar{y}$ on DAY with weights n:

$$\bar{y} = \beta_0 + \beta_1 \text{DAY} + e,\ e \sim \text{NID}(0, \sigma^2/n)$$

Video 5.6

The results of this fit from R are:

```
Call:
lm(formula = YBAR ~ DAY, data = lshoots.df, weights = N)

Residuals:
     Min       1Q   Median       3Q      Max
-4.21655 -1.17349  0.02198  1.09871  2.97488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.973754   0.314272   31.74   <2e-16 ***
DAY         0.217330   0.005339   40.71   <2e-16 ***
---

Residual standard error: 1.929 on 20 degrees of freedom
Multiple R-Squared: 0.9881,     Adjusted R-squared: 0.9875
F-statistic:  1657 on 1 and 20 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: YBAR
          Df Sum Sq Mean Sq F value    Pr(>F)
DAY        1 6164.30  6164.3  1657.2 < 2.2e-16 ***
Residuals 20   74.39     3.7
```

In particular, note the following: $\hat{\beta}_0$ = 9.9738 and $\hat{\beta}_1$ = 0.2173.

Consider the original $\sum n$ = 189 data points: we can fit an **un-weighted** regression of y on DAY using the model $y_i = \beta_0 + \beta_1 \text{DAY} + e_i$, $e_i \sim \text{NID}(0,\sigma^2)$, where y = number of stem units per shoot.

The results of fitting this un-weighted model would provide the following:
Parameter Estimates:

	Estimate	SE	t-value
Intercept	**9.9738**	0.21630	46.11
DAY	**0.2173**	0.00367	59.12

Note that the estimates of β_0 and β_1 are the **same** in both models.

It can be shown that:

- SS(lof) in the **un-weighted** regression model equals the Residual Sum of Squares in the **weighted** regression model.
- df(lof) in the **un-weighted** regression model equals the Residual Degrees of Freedom in the **weighted** regression model.

Thus we can derive SS(lof) and SS(pe) for a regression model **from the weighted regression model** to the **means of the response variable** of **cases with the same value(s) of the predictor variable(s)**.

Analysis of Variance Table:

	df	Sum Sq	Mean Sq	F-value
DAY	1	**6164.28**	6164.28	3498.35
Residuals	187	329.50	1.76	

Note that RSS **differs** in the two models. All values that use RSS in their calculations, for example, $\hat{\sigma}^2$ and standard errors, differ.
RSS differs, because the in un-weighted model, RSS is the sum of the pure error and the lack of fit error.

We can test the un-weighted regression model for lack of fit by using the test for lack of fit developed in the previous section, using:

- RSS = SS(pe) + SS(lof)
- df(RSS) = df(pe) + df(lof)

SS(lof) = *RSS* – SS(pe) = 329.50 – 255.11 = 74.39.
df(lof) = (n – (p+1)) – df(pe) = 187 – 167 = 20.

Analysis of Variance Table:

	df	Sum Sq	Mean Sq	F-value
DAY	1	**6164.28**	6164.28	3498.35
Residuals	187	329.50	**1.76**	
Lack of Fit	20	74.39	3.72	2.43
Pure Error	167	255.12	1.53	

The test statistic for testing lack of fit is $F = \dfrac{\text{SS(lof)/df(lof)}}{\text{SS(pe)/df(pe)}} = \dfrac{74.39/20}{255.12/167} = 2.43$

Since $F(0.01;20,167) = 1.99$, the *p*-value of the observed test is less than 0.01, indicating that the un-weighted regression model does not appear to be adequate.

However, an *F*-test with this many degrees of freedom is very powerful and will detect very small deviations from the null hypothesis. Thus, while the result here is statistically significant, it may not be scientifically important, and for the purposes of describing the growth of apple shoots, the un-weighted model may be adequate.

Video 5.7

Summer 2006, Question 5

The maintenance cost of tractors seems to increase with the age of the tractor. The following data were recorded for a total of 17 tractors:

Age	0.5	1.0	4.0	4.5	5.0	5.5	6.0
N	2	3	3	3	3	1	2
AvCost	172.5	664.3	633.0	900.3	1202.0	987.0	1068.5

The variables studied were as follows:

Age = Age of tractor (in years)
N = Number of tractors of the same age
AvCost = Average maintenance cost (in pounds) of tractors of the same age over a six month period

(a) A weighted regression model of the following form is fitted to these data:

$$Y_i = \beta_0 + \beta_1 X_i + e_i , \quad e_i \sim \text{IN}(0, \sigma^2 / w_i) ,$$

where Y = AvCost, X = Age and w = N. Explain why weighted regression is required here and why this is an appropriate choice of weights. Excepts from the R output for this model are shown on the following page.

(b) The data were also recorded for the 17 individual tractors. The variables studied were as follows:

Age = Age of tractor (in years)
Cost = Average maintenance cost (in pounds) of tractor over a six month period

The following unweighted regression model is fitted to these data:

$$Y_i = \beta_0 + \beta_1 X_i + e_i, \quad e_i \sim \text{IN}(0, \sigma^2),$$

where Y = Cost and X = Age.

Excerpts from the R output for this model are shown on the following page. Write down the estimate of the coefficient β_1 in this model and interpret it. Justify your identification.

(c) Explain the relationship between the Residual Sum of Squares in the weighted regression model and the Lack of Fit Sum of Squares in the unweighted regression model. Hence, or otherwise, perform a Lack of Fit test for the unweighted regression model.

R output for Question 5

```
> w.tractors.lm <- lm(AvCost ~ Age,
+ data = tractorw.df,weights= N)
> summary(w.tractors.lm)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   323.62     167.27   1.935   0.1108
Age           131.72      40.53   3.250   0.0227 *

> anova(w.tractors.lm)
Analysis of Variance Table

Response: AvCost
          Df  Sum Sq Mean Sq F value  Pr(>F)
Age        1 1099635 1099635   10.56 0.02271 *
Residuals  5  520655  104131

> tractors.lm <- lm(Cost ~ Age, data = tractors.df)

> anova(tractors.lm)
Analysis of Variance Table

Response: Cost
          Df  Sum Sq Mean Sq F value   Pr(>F)
Age        1 1099635 1099635  13.684 0.002143 **
Residuals 15 1205407   80360

> tractors.aov
Call:
   aov(formula = Cost ~ factor(Age), data = tractors.df)

Terms:
                factor(Age) Residuals
Sum of Squares    1620289.7  684752.3
Deg. of Freedom           6        10


> anova(tractors.lm, tractors.aov)
Analysis of Variance Table

Model 1: Cost ~ Age
Model 2: Cost ~ factor(Age)
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1     15 1205407
2     10  684752  5    520655 1.5207 0.2674
```

Practical (Assignment) 5

Instructions for this practical

- Open the template "Surname Forename Chpt x" from Canvas (in "Practicals").
- Complete the grid on the first page.
- Save this file (as a Word document) using your own surname, forename and the appropriate chapter number.

- Practice Question:
- Type the commands one by one into R.
- Compare the results in the R text output and graphics with the corresponding results and figures in your notes.
- Use appropriate R output to answer the questions, adapting the R code if necessary.

- Exam Question:
- Adapt the relevant R code you used for the practice question to answer the questions.
- Copy and paste the relevant R text output and graphics into your Word document to support your answers. Change the text font to "Courier New" to align columns.

- Restrict your Word document to a **maximum of 2 pages** (re-sizing graphics and deleting irrelevant R output will help).
- Submit this Word document **via Canvas** by **5.00pm 7th December 2020** (**STRICT** deadline)
- Note that submitting the practical is a declaration that the practical is your own work. Plagiarism/copying will not be tolerated.

Practice Question (not to be submitted)

Data on the growth of **short** apple shoots was recorded in the dataset `sshoots.txt`. The same variables were recorded as in the `lshoots.txt` dataset.

DAY = days from the start of the growing season
N = number of shoots sampled per day
YBAR = average number of stem units per shoot
SD = standard deviation of the number of stem units per shoot

(a) Fit a weighted regression model of the following form:

$$Y_i = \beta_0 + \beta_1 X_i + e_i , \quad e_i \sim IN(0, \sigma^2 / w_i)$$

where Y = YBAR, X = DAY and w = N. Explain why weighted regression is required here and why this is an appropriate choice of weights.

(b) Draw a scatter-plot of YBAR vs. DAY with the fitted weighted regression line.

(c) Write down the estimates of β_0 and β_1 in the following un-weighted regression model.

$$Y_i = \beta_0 + \beta_1 X_i + e_i , \quad e_i \sim IN(0, \sigma^2),$$

where Y = number of stem units per shoot and X = DAY.
Interpret the estimate of the coefficient β_1.

Exam Question (Winter 2019-20, Question 5) (to be submitted)

An environmentalist is studying the relationship between traffic noise level on a national road and distance from the road. A random sample of locations adjacent to the road was selected. At each location the noise level was recorded at a pre-determined distance from the road.

The variables studied were as follows:

Distance = Perpendicular distance from the road (m)
N = Number of locations at the same distance from the road
MeanNoise = Mean noise level (dB) at locations at the same distance from the road
(The data are stored in **Q5 Noise Means.txt (on Canvas)**)

(a) Fit a weighted regression model of the following form to these data:

$$Y_i = \beta_0 + \beta_1 X_i + e_i , \quad e_i \sim \text{NID}(0, \sigma^2 / w_i) ,$$

where Y = MeanNoise, X = Distance and w = N.

Explain why weighted regression is required here. (5 marks)

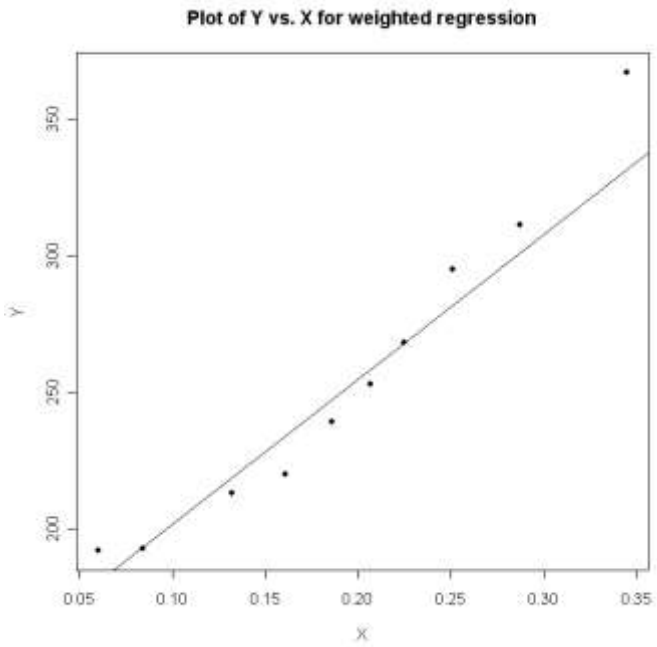
(b) The data were also recorded for the individual locations. The data are stored in **Q5 Noise Individual.txt (on Canvas)**.
Fit an un-weighted regression model to the individual data.
Explain the relationship between the Residual Sum of Squares in the weighted regression model and the Lack of Fit Sum of Squares in the un-weighted regression model.
Use this relationship to perform a Lack of Fit test for the un-weighted regression model. (20 marks)

(c) Perform the Lack of Fit test for the un-weighted regression model in a second way using Analysis of Variance (ANOVA) in R.
Give the R code you used, quote the value of the test statistic and the associated p-value. (15 marks)

(d) Explain why the test statistics in parts (b) and (c) differ slightly. (10 marks)

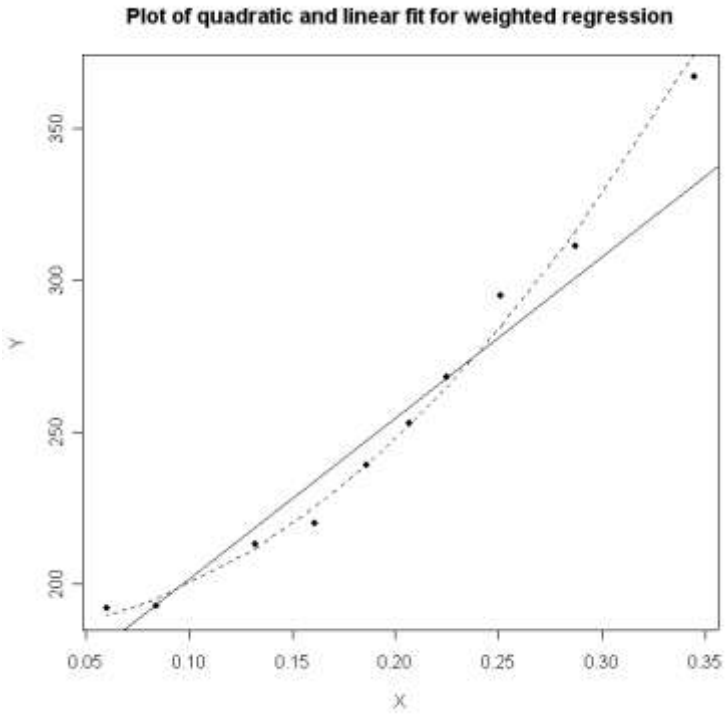
```
> # R code and output for Chapter 5

> physics.df <-
+read.table("P:\\ST2053\\physics.txt",header=T)
> physics.df
       X   Y Est.SD
1  0.345 367     17
2  0.287 311      9
3  0.251 295      9
4  0.225 268      7
5  0.207 253      7
6  0.186 239      6
7  0.161 220      6
8  0.132 213      6
9  0.084 193      5
10 0.060 192      5
>
> # attach physics.df
> attach(physics.df)

> # regression model of Y on X, weights=1/(Est.SD)^2

> physics1.lm <- lm( Y ~ X, data = physics.df,
+ weights= 1/(Est.SD)^2)
> summary(physics1.lm)
Call:
lm(formula = Y ~ X, data = physics.df, weights =
1/(Est.SD)^2)

Residuals:
       Min         1Q     Median         3Q        Max
-2.323e+00 -8.842e-01  1.266e-06  1.390e+00  2.335e+00

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  148.473      8.079   18.38 7.91e-08 ***
X            530.835     47.550   11.16 3.71e-06 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1

Residual standard error: 1.657 on 8 degrees of freedom
Multiple R-Squared: 0.9397,     Adjusted R-squared: 0.9321
F-statistic: 124.6 on 1 and 8 DF,  p-value: 3.710e-06

> anova(physics1.lm)
Analysis of Variance Table

Response: Y
```

```
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 341.99  341.99  124.63 3.710e-06 ***
Residuals  8  21.95    2.74
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1
>
> # The residual sum of squares in weighted regression
> # is the sum of squares of weighted residuals
> sum((resid(physics1.lm)^2/(Est.SD)^2))
[1] 21.95265
> # residual sum of squares = 21.95265
>
> # deviance = weighted residual sum of squares
> deviance(physics1.lm)
[1] 21.95265
> # residual sum of squares = 21.95265

> # chisquare test of goodness of fit
> # observed chisquare =  21.9526
> qchisq(0.99,8)
[1] 20.09024
> # qchisq(0.99,8) = 20.09024, so poor fit


> # plot of Y vs.X for weighted regression
> plot(X,Y,
+ main="Plot of Y vs. X for weighted regression",
+ pch=16)
> abline(physics1.lm)
```



```
> # regression model of Y on X + X^2, weights=1/(Est.SD)^2
> physics2.lm <- lm( Y ~ X + I(X^2), data = physics.df,
+ weights=1/(Est.SD)^2)
> summary(physics2.lm)

Call:
lm(formula = Y ~ X + I(X^2), data = physics.df, weights =
1/(Est.SD)^2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.89928 -0.43508  0.01374  0.37999  1.14238

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  183.8305     6.4591  28.461  1.7e-08 ***
X              0.9709    85.3688   0.011 0.991243
I(X^2)      1597.5047   250.5869   6.375 0.000376 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1

Residual standard error: 0.6788 on 7 degrees of freedom
Multiple R-Squared: 0.9911,     Adjusted R-squared: 0.9886
F-statistic: 391.4 on 2 and 7 DF,  p-value: 6.554e-08

> anova(physics2.lm)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X          1 341.99  341.99 742.185 2.303e-08 ***
I(X^2)     1  18.73   18.73  40.641 0.0003761 ***
Residuals  7   3.23    0.46
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1
>
> # chisquare test of goodness of fit
> # observed chisquare =  3.23
> qchisq(0.99,7)
[1] 18.47531
> # qchisq(0.99,7) = 18.47531, so no evidence of lack of fit
```

```
> # plot of quadratic and linear fit for weighted regression
> plot(X,Y,
+ main="Plot of quadratic and linear fit for weighted
regression",
+ pch=16)
> lines(X,fitted(physics2.lm),lty=2)
> abline(physics1.lm)
```



```
> # Hypothetical example
> hypothet.df <-
+read.table("P:\\ST2053\\hypothet.txt",header=T)
> hypothet.df
   X    Y     SD df
1  1 2.55 0.1102  2
2  1 2.75 0.0000  0
3  1 2.57 0.0000  0
4  2 2.40 0.0000  0
5  3 4.19 0.3606  1
6  3 4.70 0.0000  0
7  4 3.81 0.8571  3
8  4 4.87 0.0000  0
9  4 2.93 0.0000  0
10 4 4.52 0.0000  0
>
> # attach hypothet.df
> attach(hypothet.df)
>
> # regression line of Y on X
> hypothet1.lm <- lm(Y~X,data= hypothet.df)
> anova(hypothet1.lm )
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X          1 4.5693  4.5693   8.669 0.01859 *
Residuals  8 4.2166  0.5271
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1
> deviance(hypothet1.lm)
[1] 4.21664
> hypothet1.lm$df.residual
[1] 8
>
> # calculation of Pure Error Sum of Squares and df
> PE.SS <- sum(df*(SD^2))
> PE.SS
[1] 2.358182
> # Pure Error sum of squares = 2.358182
> PE.df <- sum(df)
> PE.df
[1] 6
> # Pure Error df = 6
```

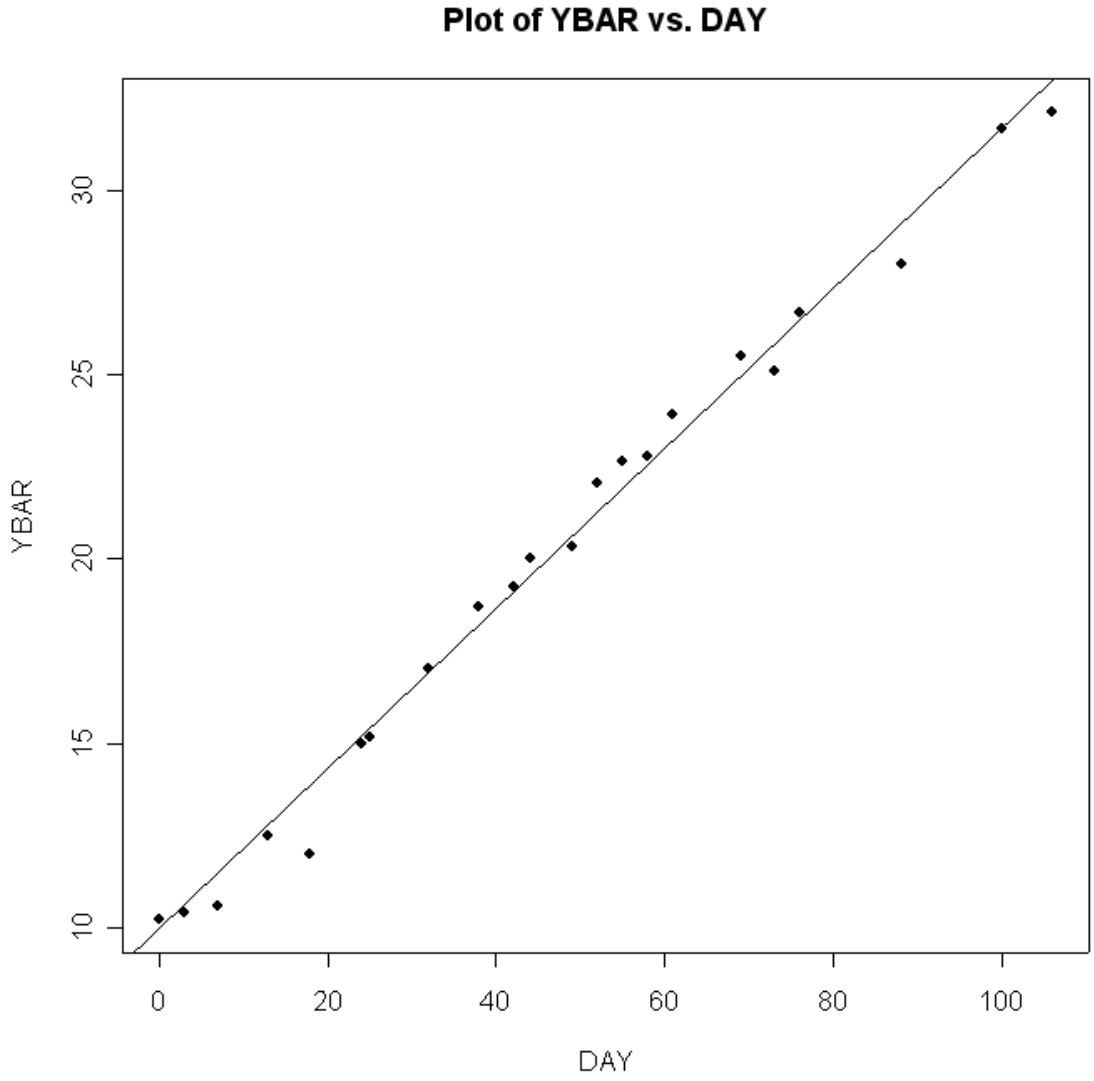
```
> # calculation of Lack of Fit Sum of Squares and df
> LF.SS <- deviance(hypothet1.lm) - PE.SS
> LF.SS
[1] 1.858458
> # Lack of Fit sum of squares = 1.858458
> LF.df <- hypothet1.lm$df.residual - PE.df
> LF.df
[1] 2
> # Lack of fit df = 2
>
> # F-test for lack of fit
> my.F <- (LF.SS/LF.df)/(PE.SS/PE.df)
> my.F
[1] 2.364268
> # F-statistic = 2.364268
> qf(0.95,2,6)
[1] 5.143253
> # qf(0.95,2,6) = 5.143253, so no lack of fit
```

```
> # lack of fit test using one-way anova
>
> # regression line of Y on X
> hypothet1.lm <- lm(Y~X,data= hypothet.df)
> anova(hypothet1.lm )
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value  Pr(>F)
X          1 4.5693  4.5693   8.669 0.01859 *
Residuals  8 4.2166  0.5271
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1
>
> # one-way anova on Y data
> hypothet1.aov <- aov(Y ~ factor(X), data = hypothet.df)
> hypothet1.aov
Call:
   aov(formula = Y ~ factor(X), data = hypothet.df)

Terms:
                factor(X) Residuals
Sum of Squares   6.427498  2.358392
Deg. of Freedom         3         6

Residual standard error: 0.6269492
Estimated effects may be unbalanced
> # residual SS in hypothet1.aov
> # = within-group SS = 2.358392
> # Hence Pure Error SS in hypothet1.lm = 2.358392
>
> # lack of fit test in hypothet1.lm
> anova(hypothet1.lm, hypothet1.aov)
Analysis of Variance Table

Model 1: Y ~ X
Model 2: Y ~ factor(X)
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1      8 4.2166
2      6 2.3584  2    1.8582 2.3638 0.1750
> # Lack of Fit SS = 1.8582 with 2 df;
> # F = 2.3638, as before.
```

```
> # Long Apple shoots
>
> lshoots.df <-
+read.table("P:\\ST2053\\lshoots.txt",header=T)
> lshoots.df
   DAY  N  YBAR   SD
1    0  5 10.20 0.83
2    3  5 10.40 0.54
3    7  5 10.60 0.54
4   13  6 12.50 0.83
5   18  5 12.00 1.41
6   24  4 15.00 0.82
7   25  6 15.17 0.76
8   32  5 17.00 0.72
9   38  7 18.71 0.74
10  42  9 19.22 0.84
11  44 10 20.00 1.26
12  49 19 20.32 1.00
13  52 14 22.07 1.20
14  55 11 22.64 1.76
15  58  9 22.78 0.84
16  61 14 23.93 1.16
17  69 10 25.50 0.98
18  73 12 25.08 1.94
19  76  9 26.67 1.23
20  88  7 28.00 1.01
21 100 10 31.67 1.42
22 106  7 32.14 2.28
>
> attach(lshoots.df)
```

```
> # weighted regression of YBAR on DAY, weights=N
> lshoots1.lm <- lm(YBAR ~ DAY, data = lshoots.df,weights= N)
>
> # regression coefficients for weighted regression,
> summary(lshoots1.lm)

Call:
lm(formula = YBAR ~ DAY, data = lshoots.df, weights = N)

Residuals:
     Min       1Q   Median       3Q      Max
-4.21655 -1.17349  0.02198  1.09871  2.97488

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 9.973754   0.314272   31.74   <2e-16 ***
DAY         0.217330   0.005339   40.71   <2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1

Residual standard error: 1.929 on 20 degrees of freedom
Multiple R-Squared: 0.9881,     Adjusted R-squared: 0.9875
F-statistic:  1657 on 1 and 20 DF,  p-value: < 2.2e-16

> # Note: these regression coefficients are the same
> # as the regression coefficients for the unweighted
regression
> # for the original data in the lecture notes
```

```
> # Plot of YBAR vs. DAY with weighted regression line
> plot(DAY,YBAR,main = "Plot of YBAR vs. DAY",pch=16)
> abline(lshoots1.lm)
```




```
> anova(lshoots1.lm)
Analysis of Variance Table

Response: YBAR
          Df Sum Sq Mean Sq F value    Pr(>F)
DAY        1 6164.3  6164.3  1657.2 < 2.2e-16 ***
Residuals 20   74.4     3.7
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` '
1
> # Note Regression SS in above = 6164.276
> # Note Residual df in above = 20
> # = Lack of Fit df in un-weighted model in notes
> # Note Residual SS in above = 74.392
```

```
> # = Lack of Fit SS in un-weighted model in notes

> # calculation of Pure Error Sum of Squares and df
> # for the un-weighted model in the notes
> PE.SS <- sum((N-1)*(SD^2))
> PE.SS
[1] 255.1215
> # Pure Error Sum of Squares = 255.1215
> PE.df <- sum(N-1)
> PE.df
[1] 167
> # Pure Error degrees of freedom = 167
>
> # calculation of Lack of Fit sum of squares and df
> LF.SS <- 329.50 - PE.SS
> LF.SS
[1] 74.3785
> # Lack of Fit sum of squares = 74.3785
> LF.df <- 187 - PE.df
> LF.df
[1] 20
> # Lack of fit df = 20
> # F-test for lack of fit
> my.F <- (LF.SS/LF.df)/(PE.SS/PE.df)
> my.F
[1] 2.434371
> # F-statistic = 2.434371
> qf(0.99,20,167)
[1] 1.989946
> # qf(0.99,20,167) = 1.989946, so fit is not adequate



# quit R
> q("yes")
```