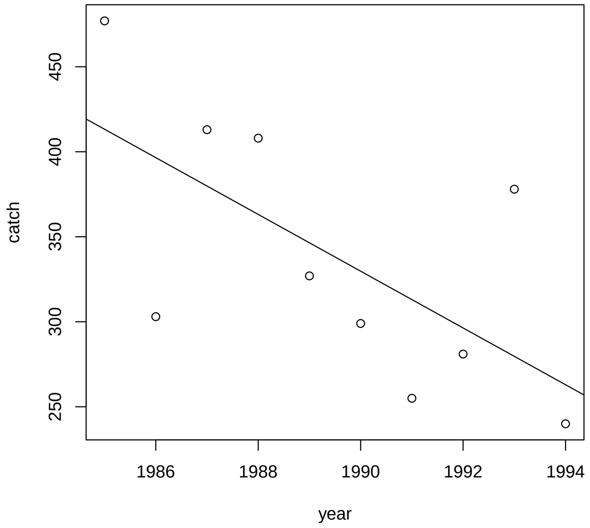
Solutions to Exercises 6.6

1. **Survival of green tree frogs**

```
> frogs <- data.frame(year=c(1985:1994),
+                     catch=c(477, 303, 413, 408, 327, 299, 255, 281, 378, 240))
```

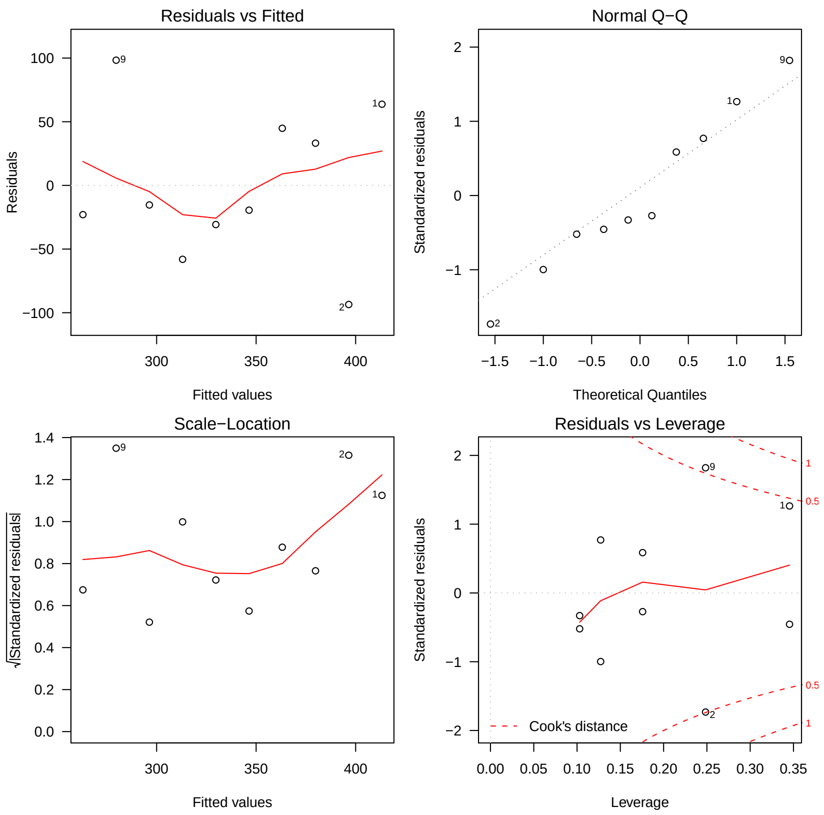
(a)

```
> plot(frogs)
> frogs.lm <- lm(catch ~ year, data = frogs)
> abline(frogs.lm)
```



```
> summary(frogs.lm)

Call:
lm(formula = catch ~ year, data = frogs)

Residuals:
   Min     1Q Median     3Q    Max
-93.54 -28.80 -17.40  41.93  98.34

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 33556.72   13646.99   2.459   0.0394 *
year          -16.70       6.86  -2.434   0.0409 *
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 62.3 on 8 degrees of freedom
Multiple R-squared: 0.4255,      Adjusted R-squared: 0.3537
F-statistic: 5.925 on 1 and 8 DF,  p-value: 0.04094
```

```
> par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 2, 1))
> plot(frogs.lm)
```



These diagnostic plots are basically sound.

(b) The estimated coefficient for year is −16.70 and this is significant at the 5% level. Therefore, we can conclude that the frog population is in decline over the years. This is supported by the first plot in (a).

(c) Substituting 0 for `catch` into the fitted the fitted equation gives year=2009. So if the linear model is accepted, they would be extinct already. This is actually not the case: what does this suggest about our model?

2. Chicks on feeds

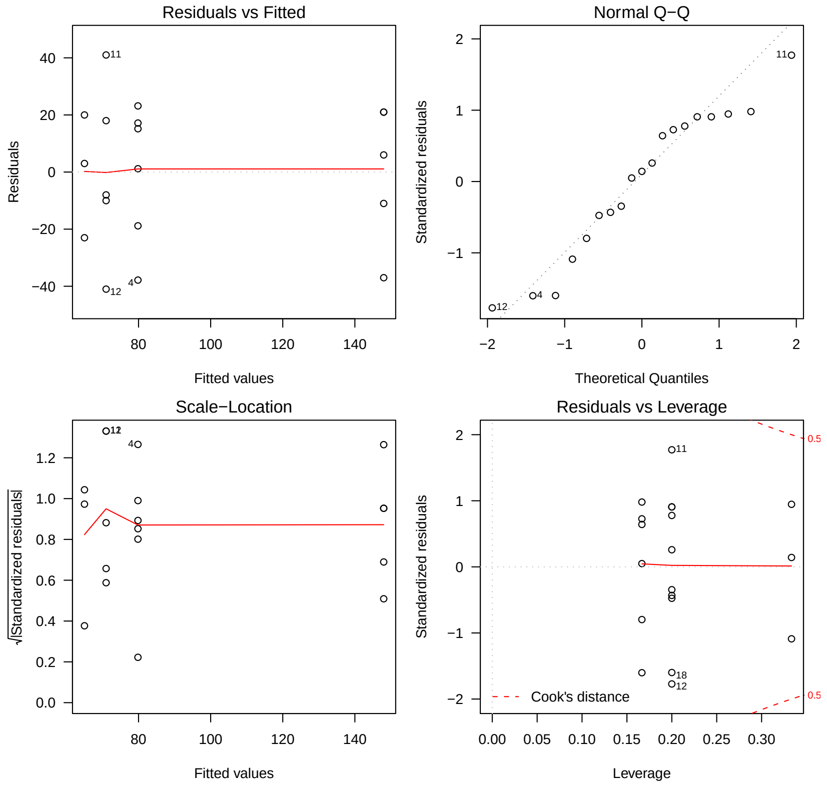
We use the model:

$$y_{ij} = \mu_i + e_{ij}; \qquad e_{ij} \sim \mathrm{N}(0, \sigma), \quad i = \text{A, B, C, D (or 1,2,3,4). j is different for each } i.$$

```
> chicks <- data.frame(feed = rep(c("A","B","C","D"),
+                           times=c(3,6,5,5)),
+                         weight_gain = c(
+                           42, 68, 85,
+                           42, 97, 81, 95, 61, 103,
+                           61, 112, 30, 89, 63,
+                           169, 137, 169, 111, 154))
> chicks.lm <- lm(weight_gain˜feed, data=chicks)
```

To obtain linear model diagnostics, we use:

```
> par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 2, 1))
> plot(chicks.lm)
```



the `las` argument of the `par` function determines the style of axis labels; 1="always horizontal". The `mar` argument determines the margins, with the four numbers corresponding to (bottom, left, top, right). Type `?par` for more details.

The diagnostic plots look acceptable. The variance in the four groups is similar, and the residuals appear to be consistent with a normal distribution.

```
> summary(chicks.lm)

Call:
lm(formula = weight_gain ~ feed, data = chicks)

Residuals:
   Min     1Q Median     3Q    Max 
-41.00 -14.92   3.00  19.00  41.00 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    65.00      14.94   4.351 0.000571 ***
feedB          14.83      18.30   0.811 0.430247    
feedC           6.00      18.90   0.317 0.755251    
feedD          83.00      18.90   4.392 0.000525 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 25.88 on 15 degrees of freedom
Multiple R-squared: 0.6758,	Adjusted R-squared: 0.6109 
F-statistic: 10.42 on 3 and 15 DF,  p-value: 0.0005872
```

The small P-value of the F-test indicates substantial evidence against the null hypothesis of equal means: $\mu_A = \mu_B = \mu_C = \mu_D$.

3. **Newspaper prices**

   (a) To compare these two models, we need to compute the sum of squares for residuals. Let RSS_1 and RSS_2 be the Residual Sum of Squares for the 1st and 2nd model respectively. It's easy to compute that $RSS_1 = 10,500,000$ and $RSS_2 = 6,530,000$. Therefore the second model is better.

   (b) i. If the value of a car does not depend on the actual age but only on whether it is used or not the model should be

$$y_{ij} = \mu_i + e_{ij}, \qquad e_{ij} \sim \mathrm{N}(0, \sigma),$$

where $i = 1, 2$ indicates the status "new" and "used". Here, μ_1 and μ_2 are means of the values of new cars and used cars and σ is the standard deviation of the error distribution. To fit this model, we need

```
> car.value <- data.frame(status = rep(c("new", "used"), times = c(2,
+     2)), value = c(24500, 20000, 20000, 17000))
> car.value.lm <- lm(value ~ status, data = car.value)
> summary(car.value.lm)
Call:
lm(formula = value ~ status, data = car.value)

Residuals:
    1     2     3     4
 2250 -2250  1500 -1500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    22250       1912  11.636   0.0073 **
statusused     -3750       2704  -1.387   0.2999
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 2704 on 2 degrees of freedom
Multiple R-squared: 0.4902,       Adjusted R-squared: 0.2353
F-statistic: 1.923 on 1 and 2 DF,  p-value: 0.2999
```

   ii. From the summary of the model, we can easily observe the parameter estimates:

$$\begin{aligned} \hat{\mu}_1 &= 22250 \\ \hat{\mu}_2 &= 22250 - 3750 = 18500 \\ \hat{\sigma} &= 2704 \end{aligned}$$

   iii. Residual sum of squares $= 2250^2 + (-2250)^2 + 1500^2 + (-1500)^2 = 14,625,000$. So this model is worse.

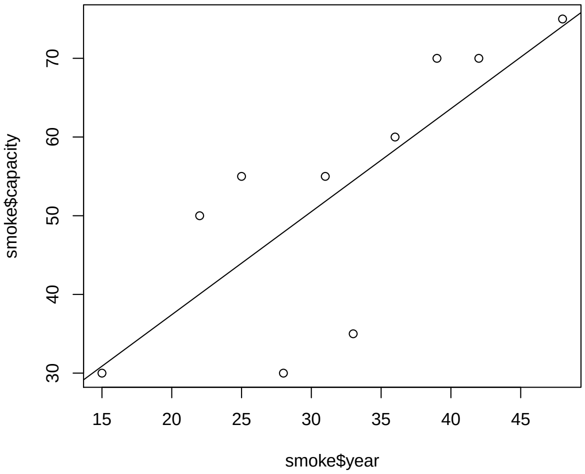
4. **Smoking**

(a) Enter the data as

```
> smoke <- data.frame(patient=c(1:10),
+                     year=c(25, 36, 22, 15, 48, 39, 42, 31, 28, 33),
+                     capacity=c(55, 60, 50, 30, 75, 70, 70, 55, 30, 35))
```

and produce the plots with

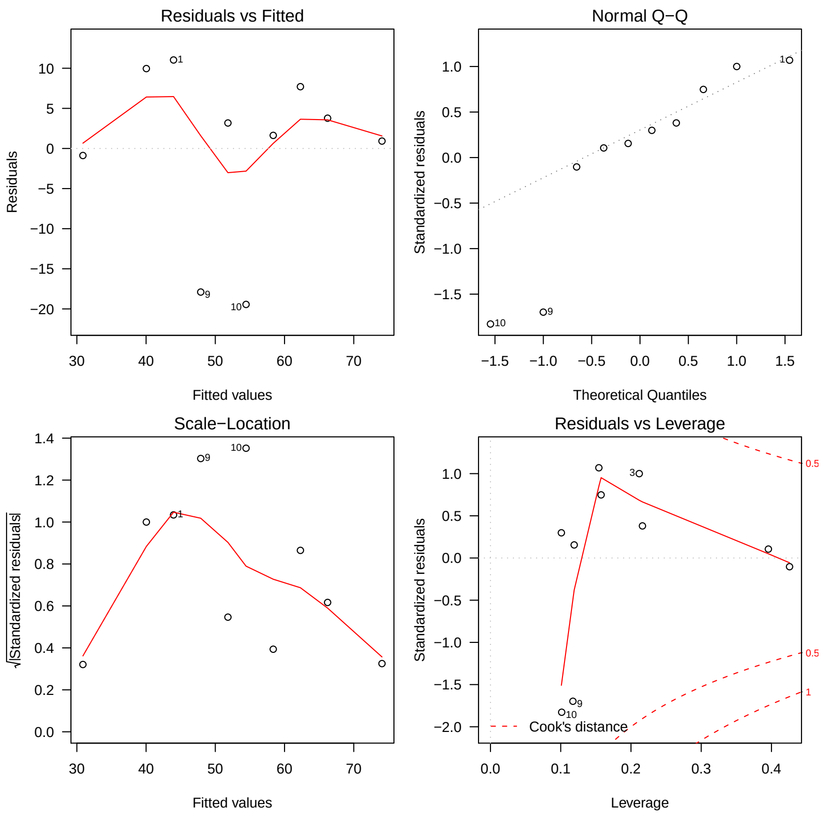
```
> plot(smoke$capacity ~ smoke$year)
> abline(lm(smoke$capacity ~ smoke$year))
```



(b) From the plot, we can see a clear increasing trend (with some noise). This indicates that in general, the more you smoke, the more diminution of lung capacity you will have. However, this demonstrates *association* rather than *causation*.

(c)
```
> smoke.lm <- lm(capacity ~ year, data = smoke)
```

```
> par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 2, 1))
> plot(smoke.lm)
```



These diagnostic plots are largely good. There is a suggestion of a problem in the variance plot (bottom left) but it is driven by two points (obvious in the top left plot, and the bottom right plot). The top right plot shows those two points also skew the distribution a fair bit. However, they do not really affect the estimates all that much as demonstrated by the leverage plot.

```
> summary(smoke.lm)

Call:
lm(formula = capacity ~ year, data = smoke)

Residuals:
     Min       1Q   Median       3Q      Max
-19.4401  -0.4258   2.4053   6.7231  11.0332

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  11.2379    12.5966   0.892  0.39836
year          1.3092     0.3789   3.455  0.00863 **
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 11.22 on 8 degrees of freedom
Multiple R-squared: 0.5988,      Adjusted R-squared: 0.5486
F-statistic: 11.94 on 1 and 8 DF,  p-value: 0.008628
```

The linear relationship between y and x is significant ($P < 0.01$), so it confirms the conclusion about the positive association.

$R^2 = 0.60$ so 60% of the variation in y is explained by variation in x.

(d) The sample correlation coefficient is $r = \sqrt{R^2}$, which is $\sqrt{0.599} = 0.774$.

(e) The slope of the regression line has the following meaning: for patients who have smoked for between about 15 and 50 years, the diminution in lung capacity increases by 1.31 units for each additional year of smoking.

(f) The predictions and the prediction intervals can be found by

```
> predict(smoke.lm, newdata = data.frame(year = c(5, 30)), interval = "prediction")
```

```
       fit       lwr      upr
1 17.78367 -18.11560 53.68294
2 50.51260  23.32592 77.69928
```

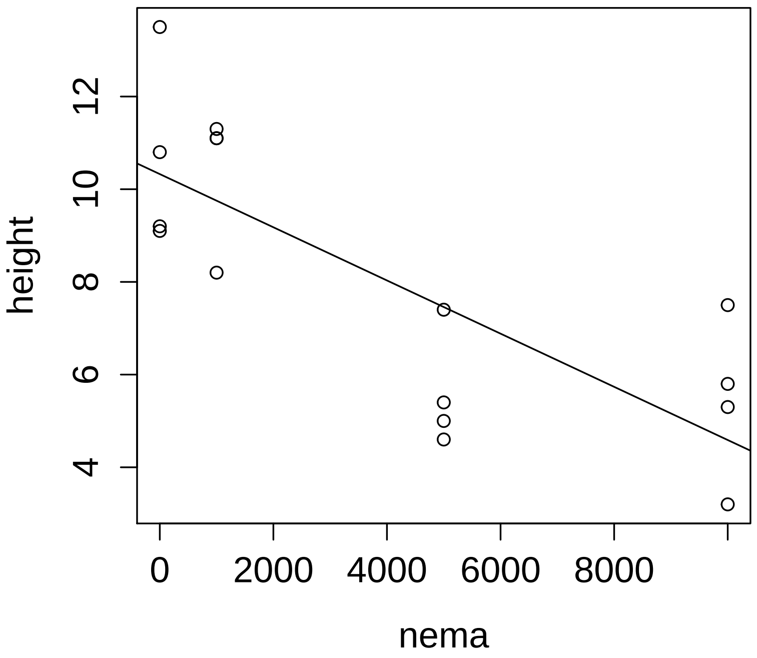
After 5 years, the prediction interval includes 0, which implies that diminution in lung capacity cannot be established after this amount of time — it needs longer.

5. **Nematodes and tomato plant growth**

(a) We can fit the linear model in R using:

```
> worms <- data.frame(nema=rep(c(0, 1000, 5000, 10000), times=c(4,4,4,4)),
+                      height=c(10.8, 9.1, 13.5, 9.2,
+                        11.1, 11.1, 8.2, 11.3,
+                        5.4, 4.6, 7.4, 5.0,
+                        5.8, 5.3, 3.2, 7.5))
```

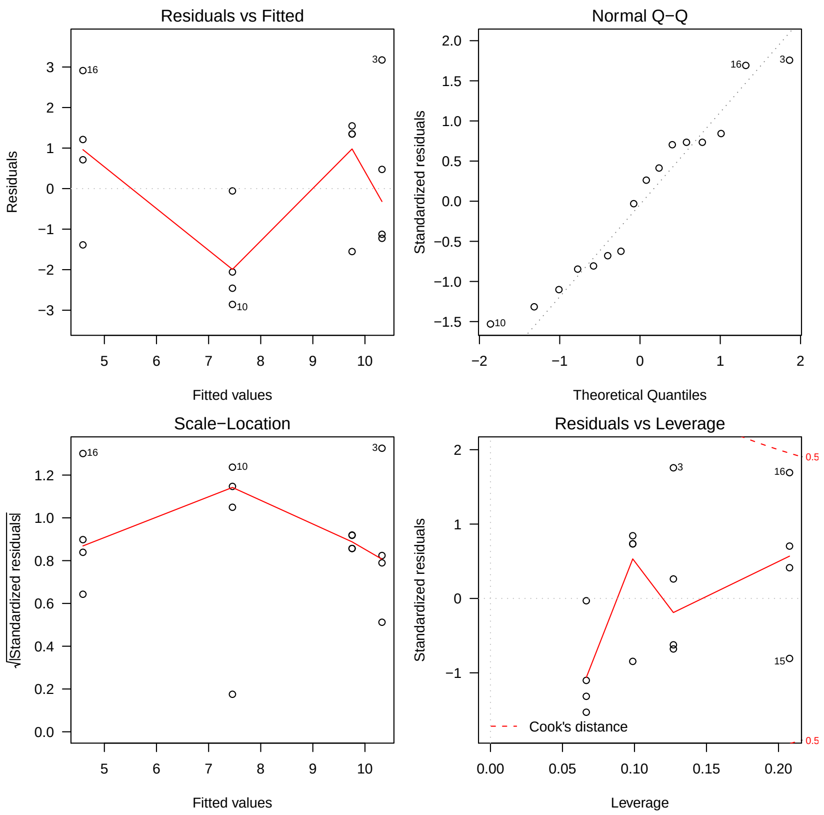
```
> plot(height ~ nema, data = worms)
```



A straight line looks okay, though there may be some curvature.

```
> worms.lm <- lm(height ~ nema, data=worms)
```

```
> par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 2, 1))
> plot(worms.lm)
```



These diagnostics are reasonable.

```
> summary(worms.lm)
Call:
lm(formula = height ~ nema, data = worms)

Residuals:
    Min      1Q  Median      3Q     Max
-2.8575 -1.4295  0.2081  1.3474  3.1736

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 10.3264113  0.6889948  14.988 5.15e-10 ***
nema        -0.0005738  0.0001228  -4.674 0.000358 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 1.933 on 14 degrees of freedom
Multiple R-squared: 0.6094,       Adjusted R-squared: 0.5816
F-statistic: 21.85 on 1 and 14 DF,  p-value: 0.0003584
```

From the summary of the model, the coefficient for `nema` is highly significant. R^2 is about 0.61 which shows that 61% of the variation is explained by the model. In summary, there is a substantial negative linear relationship between number of nematodes and seedling growth.

(b) Now we take the number of nematodes as a factor. First, we change the variable `nema` from numeric to factor.

```
> worms.lm1 <- lm(height ~ factor(nema), data = worms)
> summary(worms.lm1)

Call:
lm(formula = height ~ factor(nema), data = worms)

Residuals:
       Min         1Q     Median         3Q        Max 
-2.250e+00 -1.113e+00  2.276e-15  7.250e-01  2.850e+00 

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)    
(Intercept)        10.6500     0.8333  12.781 2.39e-08 ***
factor(nema)1000   -0.2250     1.1784  -0.191 0.851768    
factor(nema)5000   -5.0500     1.1784  -4.285 0.001059 ** 
factor(nema)10000  -5.2000     1.1784  -4.413 0.000846 ***
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1 

Residual standard error: 1.667 on 12 degrees of freedom
Multiple R-squared: 0.7512,	Adjusted R-squared: 0.689 
F-statistic: 12.08 on 3 and 12 DF,  p-value: 0.0006163 
```

From the summary, we can see a significant effect of the treatments. However, there is not much evidence to show a difference between 0 and 1000 nematodes.

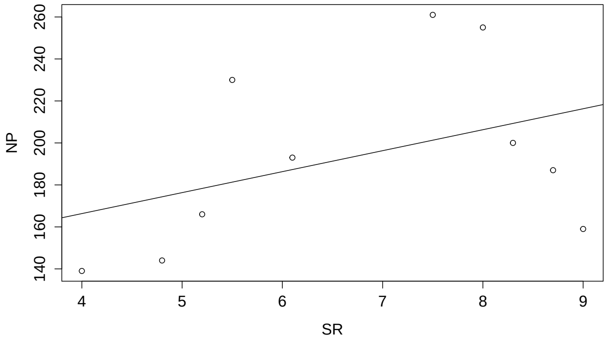
(c) The second model has a larger R^2 and a smaller residual standard error than the straight line model, so it fits better. But it is a more complex model, and so it less adaptable—for example, it cannot be used to predict seedling growth for 3000 nematodes, whereas the first model can.

Other models which may be better are a quadratic model in number of nematodes, or a model which divides the treatments into two groups—low numbers (0 or 1000) vs high numbers (5000 or 10000).

6. **Profitability versus stocking rate for sheep farms**

We draw a scatter-plot and a smoother to examine the data:

```
> farm.record <-
+   data.frame(SR=c(4.0, 4.8, 5.2, 5.5, 6.1, 7.5, 8.0, 8.3, 8.7, 9.0),
+              NP=c(139, 144, 166, 230, 193, 261, 255, 200, 187, 159))

> library(MASS)
> par(las = 1)
> scatter.smooth(farm.record$NP ~ farm.record$SR, xlab = "Stocking Rate (ewes/ha)",
+     ylab = "Nett Profit ($/ha)")
```



It's obvious that the two variables don't have a linear relationship. It seems that the net profit goes up initially when stocking rate goes up, then when the stocking rate reaches about 7.5, it drops. One possibility is a quadratic relationship of the form $NP = c_1 SR^2 + c_2 SR + c_3$ where c_1, c_2 and c_3 are parameters to estimate (c_1 negative).

However, we will fit a linear regression model and assess it.

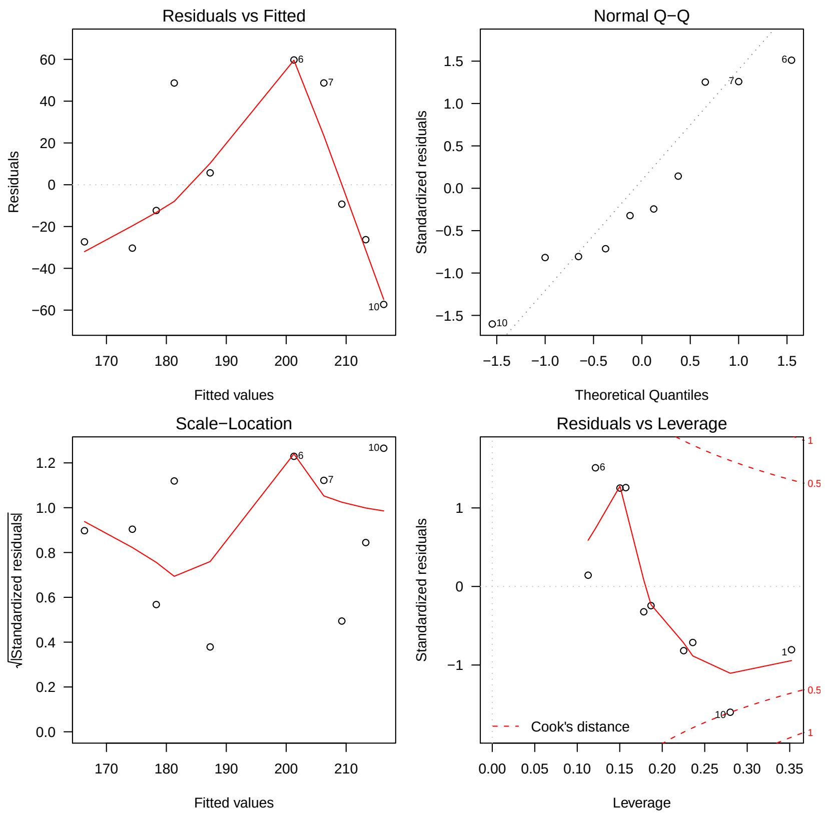
```
> farm.1 <- lm(NP ~ SR, data = farm.record)
```

Now obtain diagnostic plots (see code below and next page).

The model clearly isn't great according to the diagnostics. It looks like we need to add a quadratic term.

Regardless, press on and summarise the model.

```
> par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 2, 1))
> plot(farm.1)
```



```
> summary(farm.1)

Call:
lm(formula = NP ~ SR, data = farm.record)

Residuals:
   Min     1Q Median     3Q    Max
-57.27 -27.07 -10.80  37.94  59.71

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  126.376     54.078   2.337   0.0476 *
SR             9.989      7.811   1.279   0.2368
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 42.15 on 8 degrees of freedom
Multiple R-squared: 0.1697,	Adjusted R-squared: 0.06595
F-statistic: 1.636 on 1 and 8 DF,  p-value: 0.2368
```

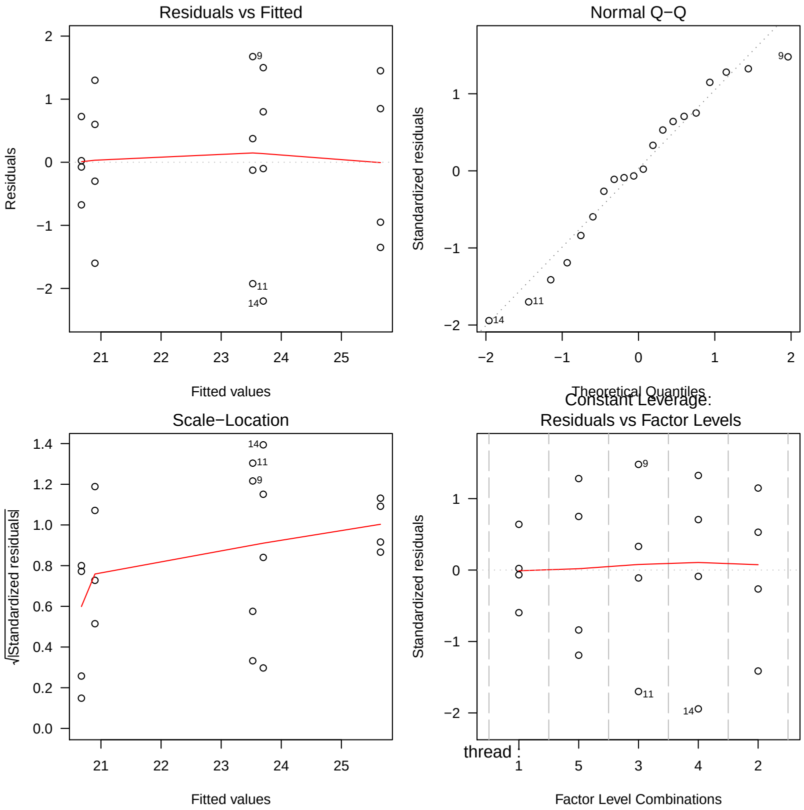
According to this output, there is no significant linear relationship between stocking rate and net profit.

This model is not very useful, as there is a clear relationship, which a simple linear model cannot pick up.

7. Thread strength

We start with the point estimation of the 5 different means, then fit the model.

```
> linen <- data.frame(thread=factor(rep(c(1,2,3,4,5),
+                        times=c(4,4,4,4,4))),
+                     strengths=c(20.6, 20.7, 20.0, 21.4,
+                        24.7, 26.5, 27.1, 24.3,
+                        25.2, 23.4, 21.6, 23.9,
+                        24.5, 21.5, 23.6, 25.2,
+                        19.3, 21.5, 22.2, 20.6))
> tapply(linen$strengths, linen$thread, mean)

     1      2      3      4      5
20.675 25.650 23.525 23.700 20.900

> linen.lm <- lm(strengths ~ thread, data=linen)

> par(mfrow = c(2, 2), las = 1, mar = c(4, 4, 2, 1))
> plot(linen.lm)
```



These diagnostics look really good.

```
> summary(linen.lm)

Call:
lm(formula = strengths ~ thread, data = linen)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.2000 -0.7437 -0.0250  0.8125  1.6750 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  20.6750     0.6538  31.624 3.81e-15 ***
thread2       4.9750     0.9246   5.381 7.64e-05 ***
thread3       2.8500     0.9246   3.083  0.00758 ** 
thread4       3.0250     0.9246   3.272  0.00515 ** 
thread5       0.2250     0.9246   0.243  0.81103    
---
Signif. codes:  0 ,***, 0.001 ,**, 0.01 ,*, 0.05 ,., 0.1 , , 1

Residual standard error: 1.308 on 15 degrees of freedom
Multiple R-squared: 0.7324,	Adjusted R-squared: 0.661 
F-statistic: 10.26 on 4 and 15 DF,  p-value: 0.0003304
```

F-test: At the 0.05 significance level we would reject the null hypothesis that the mean strength of all the threads is the same.