620-371: Linear Models

Practice Class 5

31st March, 2009

In this practice class, we will replicate and extend the results of the practice class from last week, using R. From last week: we model an individual's income at age 30 against the number of years of formal education (with a linear model). The following data is collected:

Years of formal education (x)	Income (\$k) (y)
8	8
12	15
14	16
16	20
16	25
20	40

1. Construct X and $\mathbf{y}$ objects in R. Also set the number of samples, n, and the number of parameters, p.

   Solution:

   ```
   > y <- as.vector(c(8, 15, 16, 20, 25, 40))
   > X <- matrix(c(1, 1, 1, 1, 1, 1, 8, 12, 14, 16, 16, 20), 6, 2)
   > n <- 6
   > p <- 2
   ```

2. Estimate the parameters $\boldsymbol{\beta}$ using the formula for the least squares estimator.

   Solution:

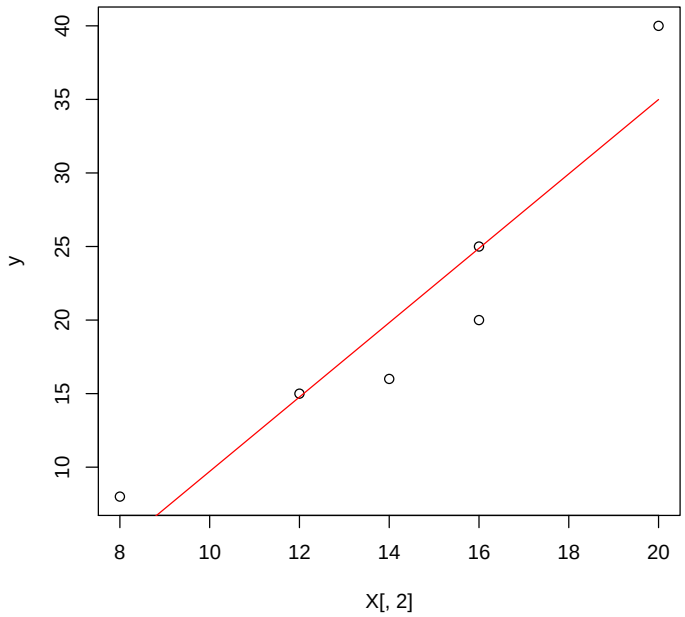
   ```
   > library(car)
   > b <- inv(t(X) %*% X) %*% t(X) %*% y
   > b

           [,1]
   [1,] -15.568
   [2,]   2.528
   ```

3. Create a scatterplot of the data and add the regression line to it.

   Solution:

   ```
   > plot(X[, 2], y)
   > curve(b[1] + b[2] * x, add = TRUE, col = "red")
   ```



4. Create a data frame with the data in two components, `income` and `education`.

   Solution:

   ```
   > income <- data.frame(income = y, education = X[, 2])
   ```

5. Use the `lm` function to estimate the parameters again. Check that you have the same result as in question 2.

   Solution:

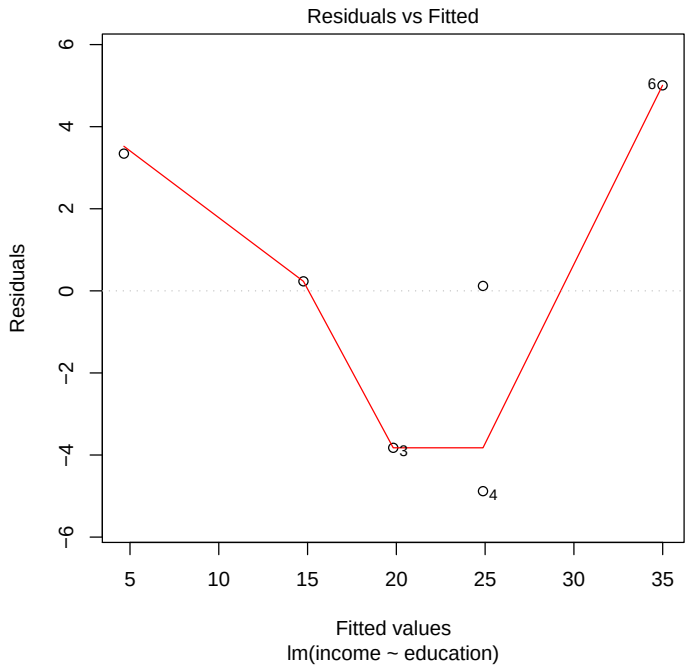
   ```
   > model <- lm(income ~ education, data = income)
   > model$coefficients

   (Intercept)   education
       -15.568       2.528
   ```

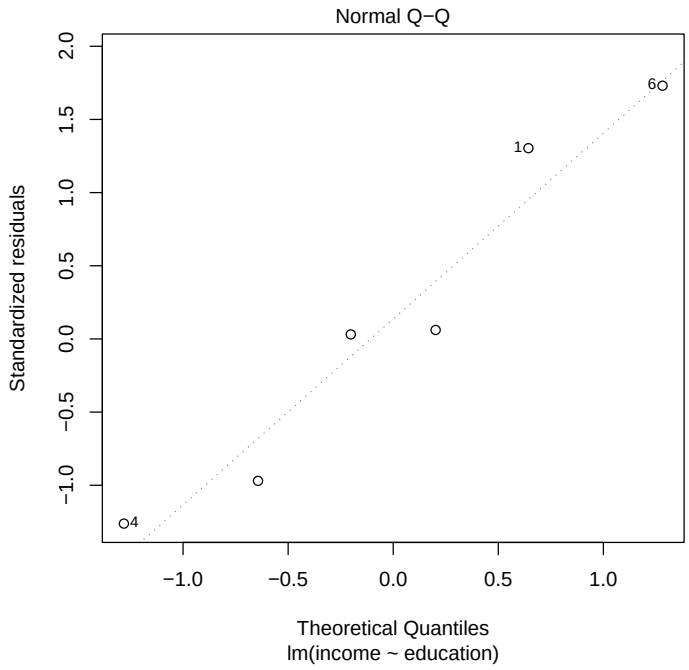
6. Produce diagnostic plots for the model.

   Solution:

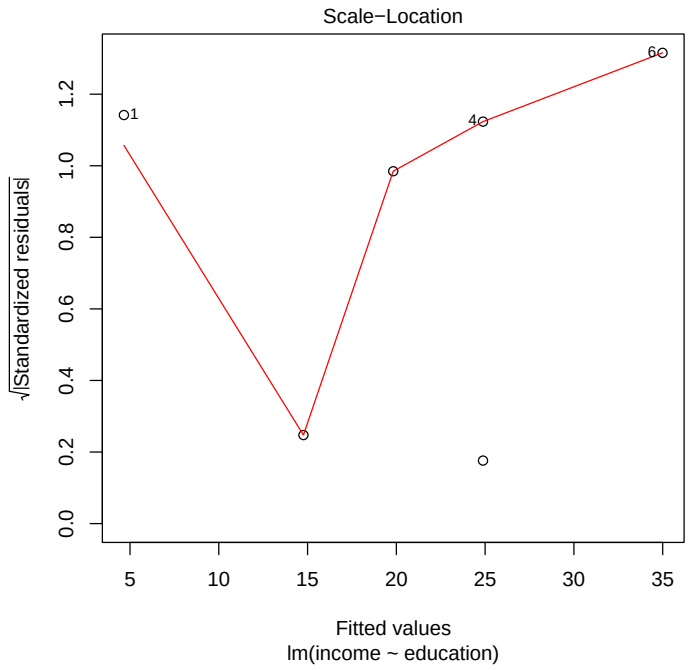
   ```
   > plot(model, which = 1)
   ```



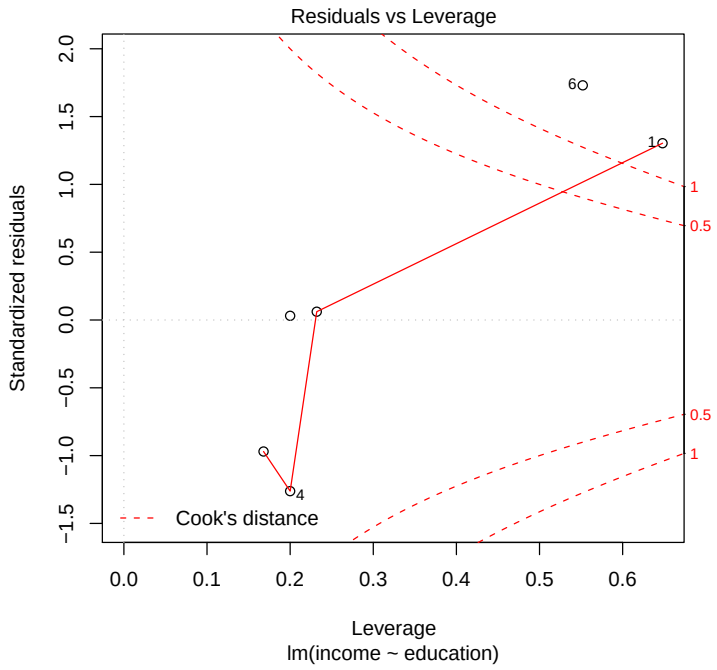

```
> plot(model, which = 2)
```




```
> plot(model, which = 3)
```




```
> plot(model, which = 5)
```



Do the remaining questions in two ways: using matrix manipulation as described in the lecture slides, and using your `lm` model. Check that you get the same results!

7. Estimate the income of a person who has had 18 years of formal education.

   Solution:

```
> tt <- as.vector(c(1, 18))
> tt %*% b

       [,1]
[1,] 29.936

> person <- data.frame(education = 18)
> predict(model, person)

     1
29.936
```

8. Calculate the residuals of the model.

   Solution:

```
> e <- y - X %*% b
> e
```

```
       [,1]
[1,]  3.344
[2,]  0.232
[3,] -3.824
[4,] -4.880
[5,]  0.120
[6,]  5.008
```

```
> model$residuals
```

```
     1      2      3      4      5      6
 3.344  0.232 -3.824 -4.880  0.120  5.008
```

9. Estimate the variance of the error, σ^2.

   Solution:

```
> SSRes <- sum(e^2)
> s2 <- SSRes/(n - p)
> s2
```

```
[1] 18.692
```

```
> deviance(model)/model$df.residual
```

```
[1] 18.692
```

10. Find 95% confidence intervals for both parameters.

    Solution:

```
> C <- inv(t(X) %*% X)
> hw <- qt(0.975, df = n - p) * sqrt(s2 * C[1, 1])
> c(b[1] - hw, b[1] + hw)
```

```
[1] -35.042213   3.906213
```

```
> hw <- qt(0.975, df = n - p) * sqrt(s2 * C[2, 2])
> c(b[2] - hw, b[2] + hw)
```

```
[1] 1.213055 3.842945
```

```
> confint(model, c(1, 2), level = 0.95)
```

```
                2.5 %   97.5 %
(Intercept) -35.042213 3.906213
education     1.213055 3.842945
```

11. Find a 95% confidence interval for the average income of a person who has had 18 years of formal education.

    Solution:

```
> xst <- as.vector(c(1, 18))
> hw <- qt(0.975, df = n - p) * sqrt(s2 * t(xst) %*% C %*% xst)
> c(xst %*% b - hw, xst %*% b + hw)

[1] 23.06130 36.81070

> predict(model, person, interval = "confidence", level = 0.95)

     fit      lwr     upr
1 29.936 23.06130 36.8107
```