Basic Course on R: Linear Regression Practical Answers

Elizabeth Ribble*

20-24 May 2019

Contents

*emcclel3@msudenver.edu

1 Baby Data

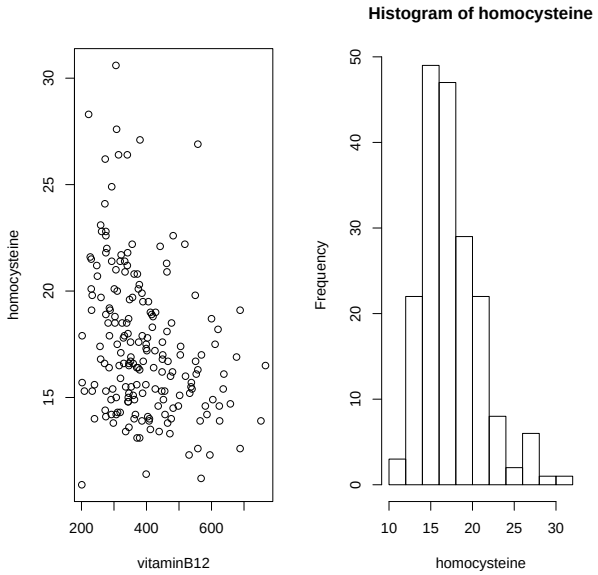
1. Read in the data "R_data_January2015.csv" with a header and row names from the first column. Assign it to the object `babydata` and allow strings be converted to factors. Attach the data to the environment.

```
babydata <- read.csv("R_data_January2015.csv",header=T,row.names=1)
attach(babydata)
```

2. We previously saw that there was an association between `vitaminB12` and `homocysteine`. We will now quantify the magnitude of this relationship and see if we can explain the variabiliy of `homocysteine` with values of `vitaminB12`. Answer the following questions:

   (a) First plot the data: a scatterplot to visualize the association (should be linear) and a histogram of the dependent variable to check for normality (normal errors are required - checking the distribution of the dependent variable is good enough for now). Do these particular assumptions appear to hold?

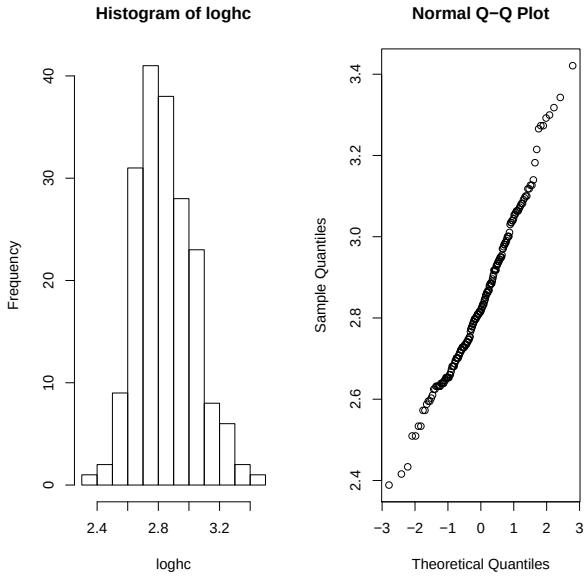
```
par(mfrow=c(1,2))
plot(vitaminB12, homocysteine)
hist(homocysteine)
```



There is a slight linear association but looks like the variable isn't normal...

(b) You should have noticed that the dependent variable `homocysteine` is right-skewed. We need to see if a log-transformation of the data is more normally distributed. Assign the log of `homocysteine` to `loghc` and make a histogram of `loghc`. Check that it is normal by plotting it against a normal distribution (use `qqnorm()`). How does it look? More normal?

```
loghc <- log(homocysteine)
par(mfrow=c(1,2))
hist(loghc)
set.seed(1234)
qqnorm(loghc)
```

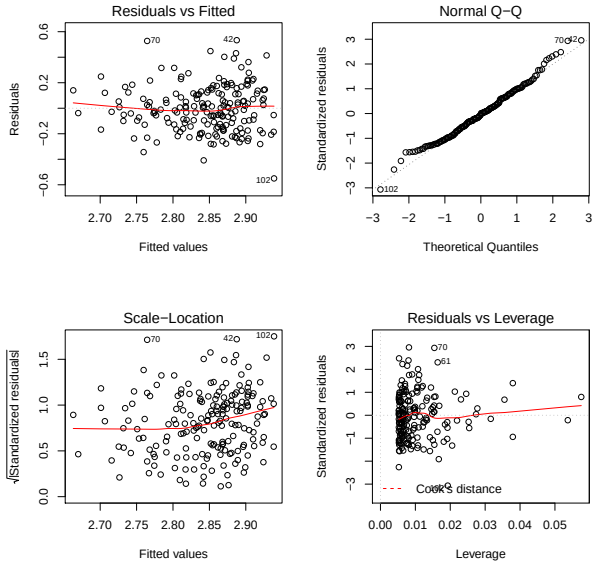


Doesn't look too bad!

(c) We'll just assume everything looks good. Set up (i.e. write down) the linear regression model for modeling `loghc` against `vitaminB12` and then run it in R. Check the assumptions by plotting the residuals (should be no patterns) versus the fitted values and looking at a QQ plot of the residuals (should be straight line). Do the assumptions hold?
The model is

$$\texttt{loghc} \sim b_0 + b_1 * \texttt{vitaminB12} + \epsilon.$$

```
lmbh <- lm(loghc ~ vitaminB12)
par(mfrow=c(2,2))
plot(lmbh)
```



There is a slight fan pattern in the residuals, but it's not too strong. The tails of the distribution deviate from the normal quantile line, but again it's not that bad.

(d) Assuming the assumptions hold (even if they don't), we'll make inference on the slope. Is `vitaminB12` statistically significant in the model? What percent variability does it explain in `loghc`? Write down the model with the estimates.

```
summary(lmbh)

##
## Call:
## lm(formula = loghc ~ vitaminB12)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.5501 -0.1288 -0.0041  0.1176  0.5334
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.0369329  0.0457556  66.373  < 2e-16 ***
## vitaminB12  -0.0004881  0.0001113  -4.386 1.92e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1812 on 188 degrees of freedom
```

```
## Multiple R-squared:  0.09281,Adjusted R-squared:  0.08799
## F-statistic: 19.23 on 1 and 188 DF,  p-value: 1.923e-05
```

```
lmbh
```

```
##
## Call:
## lm(formula = loghc ~ vitaminB12)
##
## Coefficients:
## (Intercept)   vitaminB12
##   3.0369329   -0.0004881
```

`vitaminB12` is statistically significant (p-value less than 0.05). It explains about 9.3 percent of the variability in `loghc`. The model with the estimates is

$$\texttt{loghc} \sim 3.04 - 0.0005 * \texttt{vitaminB12}.$$

(e) What is the predicted `loghc` level for a person with `vitaminB12` equal to 650? What is this value when unlogged (exponentiated)? Does it fall within the original range of values of `homocysteine` (can you guess what the name of the function is to find the range of a vector?)?

```
p650 <- predict(lmbh, newdata=data.frame(vitaminB12=650))
## or
3.04 - 0.0005*650
```

```
## [1] 2.715
```

```
p650
```

```
##        1
## 2.719635
```

```
exp(p650)
```

```
##        1
## 15.17479
```

```
range(homocysteine)
```

```
## [1] 10.9 30.6
```

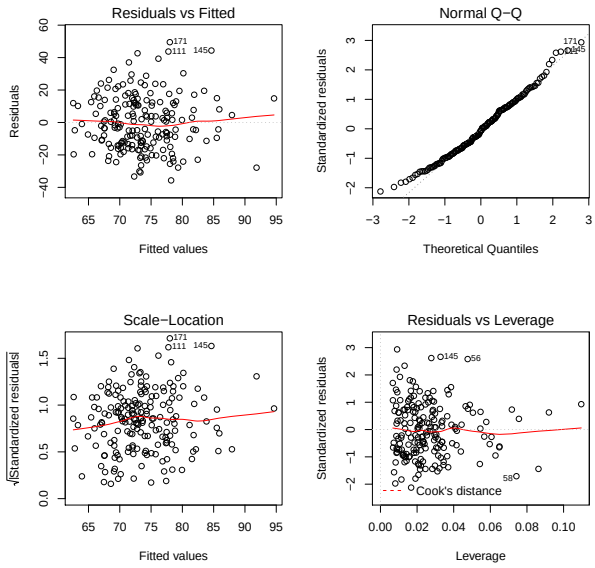
Yes, 15.1 is in the range of the unlogged `homocysteine` values: (10.9, 30.6).

3. Now let's consider a framework where we want to use more than one predictor. We want to build a regression model for `SAM` using `vitaminB12`, `cholesterol`, `homocysteine` and `folicacid_erys` (folic acid red blood cells). Answer the following questions:

   (a) Set up (i.e. write down) the linear regression model and then run it in R. Check the assumptions by plotting the residuals versus the fitted values and looking at a QQ plot of the residuals. Do the assumptions hold?
   The model is

   $$\texttt{SAM} \sim b_0 + b_1 * \texttt{vitaminB12} + b_2 * \texttt{cholesterol} + b_3 * \texttt{homocysteine} + b_4 * \texttt{folicacid\_erys} + \epsilon.$$

```
mlr <- lm(SAM ~ vitaminB12 + cholesterol + homocysteine +
                folicacid_erys)
par(mfrow=c(2,2))
plot(mlr)
```



   The residuals have constant variance and are normal, so the assumptions do hold.

   (b) Assuming the assumptions hold (even if they don't), we'll make inference on the slopes. Are any of the variables statistically significant in the model? What percent variability do the variables explain in `SAM`? Write down the model with the estimates.

```
summary(mlr)
```

```
##
## Call:
## lm(formula = SAM ~ vitaminB12 + cholesterol + homocysteine +
##     folicacid_erys)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -35.704 -12.367  -1.047  11.972  49.505
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     1.800739  27.877405   0.065 0.948566
## vitaminB12     -0.002684   0.011286  -0.238 0.812288
## cholesterol     4.583482   1.369695   3.346 0.000992 ***
## homocysteine   -0.645535   0.413510  -1.561 0.120206
## folicacid_erys  0.005033   0.006812   0.739 0.460939
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.94 on 185 degrees of freedom
## Multiple R-squared:  0.09232,Adjusted R-squared:  0.0727
## F-statistic: 4.704 on 4 and 185 DF,  p-value: 0.001225
```

```
mlr
```

```
##
## Call:
## lm(formula = SAM ~ vitaminB12 + cholesterol + homocysteine +
##     folicacid_erys)
##
## Coefficients:
##    (Intercept)      vitaminB12     cholesterol    homocysteine
##       1.800739       -0.002684        4.583482       -0.645535
## folicacid_erys
##       0.005033
```

Only `cholesterol` is statistically significant (p-value less than 0.05). The 4 variables explain 7.27 percent of the variability in `SAM`. The model with the estimates is

$$\texttt{SAM} \sim 1.8 - 0.003 * \texttt{vitaminB12} + 4.58 * \texttt{cholesterol} - 0.65 * \texttt{homocysteine} + 0.005 * \texttt{folicacid\_erys}.$$

(c) What is the predicted `SAM` level for a person with the following:

$$\begin{aligned} \texttt{vitaminB12} &= 650 \\ \texttt{cholesterol} &= 17 \\ \texttt{homocysteine} &= 16 \\ \texttt{folicacid\_erys} &= 1340 \end{aligned}$$

```
predict(mlr, newdata=data.frame(vitaminB12=650,
                                cholesterol=17,
                                homocysteine=16,
                                folicacid_erys=1340))
```

```
##        1
## 74.39131
```