Applied Statistical Methods – Solution 2

Peter von Rohr

2019-03-04

Problem 1: Regression Model

During the lecture the regression model was explained using the dataset given in Table 1.

Table 1: Dataset for Regression of Body Weight on Breast Circumference for ten Animals

Animal	Breast Circumference	Body Weight
1	176	471
2	177	463
3	178	481
4	179	470
5	179	496
6	180	491
7	181	518
8	182	511
9	183	510
10	184	541

The same dataset is also available from the website at https://charlotte-ngs.github.io/GELASMSS2019/ex/w03/bw_bc_reg.csv.

Your Task

- Setup the linear regression model with an intercept for the data given in Table 1
- Compute the solution for the unknown parameter b
- Verify the result with the output from the function `lm()` in R

Solution

The linear regression model is given by the following equation

$$y = X * b + \epsilon$$

where y is a vector of body weights, X is a matrix with two columns. The first column of X is all ones and the second column contains the breast circumference values, b is the vector with the intercept and the unknown regression coefficient and ϵ is the vector of unknown random residuals. The least squares estimate $\hat{b}$ can be computed as

$$\hat{b} = (X^T X)^{-1} X^T y$$

The matrix X and the vector y are extracted from the dataframe and have the following form

```
n_nr_ani <- nrow(tbl_reg)
mat_x <- matrix(c(rep(1,n_nr_ani), tbl_reg$`Breast Circumference`), ncol = 2)
vec_y <- tbl_reg$`Body Weight`
```

$$X = \begin{bmatrix} 1 & 176 \\ 1 & 177 \\ 1 & 178 \\ 1 & 179 \\ 1 & 179 \\ 1 & 180 \\ 1 & 181 \\ 1 & 182 \\ 1 & 183 \\ 1 & 184 \end{bmatrix}, \quad y = \begin{bmatrix} 471 \\ 463 \\ 481 \\ 470 \\ 496 \\ 491 \\ 518 \\ 511 \\ 510 \\ 541 \end{bmatrix}$$

The result for $\hat{b}$ is then

```
xtx <- crossprod(mat_x)
n_hat_b  <- solve(xtx,crossprod(mat_x, vec_y))
n_hat_b
```

```
##              [,1]
## [1,] -1065.114943
## [2,]     8.673235
```

We can verify this result using the `lm()` function of R

```
lm_bwbc <- lm(`Body Weight` ~ `Breast Circumference`, data = tbl_reg)
summary(lm_bwbc)
```

```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference`, data = tbl_reg)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.3941  -6.5525  -0.0673   9.3707  13.2594
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)            -1065.115    255.483  -4.169 0.003126 **
## `Breast Circumference`     8.673      1.420   6.108 0.000287 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.08 on 8 degrees of freedom
## Multiple R-squared:  0.8234, Adjusted R-squared:  0.8014
## F-statistic: 37.31 on 1 and 8 DF,  p-value: 0.000287
```

Problem 2: Prediction

Given the measurement of the trait `Breast Circumference` for two additional animals. The measurements are shown in the following table

Table 2: Breast Circumference Measurements For Two Animals Used To Predict Body Weight

Animal	Breast Circumference
Animal 11	181.2
Calf 12	99.5

We want to use the results of Problem 1 to computed the predicted values for `Body Weight` for the two animals. The observed value for `Breast Circumference` of "Calf 12" is outside of the range of the values used in Problem 1. Predicting values of response variables based on predictors that are outside of the range of values used for the parameter estimation is called **extrapolation**. Based on the result of the predicted value of the trait `Body Weight` for "Calf 12" what can be said about the process of extrapolation?

Your Tasks

- Compute the predicted value of `Body Weight` for "Animal 11" using the results from Problem 1
- Compute the predicted value of `Body Weight` for "Calf 12" using the results from Problem 1
- Make a statement about the validity of the extrapolated value of `Body Weight` for "Calf 12"

Solution

The equation to predict `Body Weight` from `Breast Circumference` is based on the regression equation that was derived in Problem 1.

$$\widehat{y_k} = \widehat{b_0} + \widehat{b_1} * x_k$$

where $\widehat{b_0}$ and $\widehat{b_1}$ are the estimates of the intercept and the regression coefficient from Problem 1. The variable x_k is the `Breast Circumference` for the newly measured animal k.

Assume that the results from Problem 1 are stored in a variable `n_hat_b`, and that the measured values for `Breast Circumference` are stored in a dataframe with the name `tbl_new_ani` in a column called 'Breast Circumference. The value $\widehat{y_k}$ for the two newly measured animals "Animal 11" and "Calf 12" can be computed as

```
vec_x_k <- tbl_new_ani$`Breast Circumference`
vec_y_k <- n_hat_b[1] + n_hat_b[2] * vec_x_k;vec_y_k
```

```
## [1]  506.4752 -202.1281
```

Collecting these results in a table leads to

Table 3: Prediction Results

Animal	Breast Circumference	Predicted Body Weight
Animal 11	181.2	506
Calf 12	99.5	-202

The result of the predicted `Body Weight` for "Animal 11" corresponds to 506 which is a plausible result. The predicted `Body Weight` for "Calf 12" is -202 which does not make any sense. As a consequence, we can say that extrapolation of response values based on predictors that are so far away from the range of predictors used to estimate the regression equation is not allowed.