Applied Statistical Methods - Solution 5

Peter von Rohr

2023-04-24

Problem 1: Interactions

Use the following dataset on `Breed`, `Breast Circumference` and `Body Weight` and fit a fixed linear effects model with `Body Weight` as response and `Breed` and `Breast Circumference` as predictors and include an interaction term between the two predictors. Compute the expected difference in `Body Weight` for two animals which differ in `Breast Circumference by $1cm$ for every`Breed‘.

The dataset is available under

```
## [1] "https://charlotte-ngs.github.io/asmss2023/data/asm_bw_flem.csv"
```

Solution

- Read the data and select the column that are required for fitting the linear model

```
s_tbl_ex05_p01_path <- "https://charlotte-ngs.github.io/asmss2023/data/asm_bw_flem.csv"
tbl_bw_bc_br <- readr::read_delim(s_tbl_ex05_p01_path, delim = ",")
```

```
## Rows: 10 Columns: 6
## -- Column specification --------------------------------------------------------
## Delimiter: ","
## chr (1): Breed
## dbl (5): Animal, Breast Circumference, Body Weight, BCS, HEI
##
## i Use ‘spec()‘ to retrieve the full column specification for this data.
## i Specify the column types or set ‘show_col_types = FALSE‘ to quiet this message.
```

```
tbl_bw_bc_br <- dplyr::select(tbl_bw_bc_br, Animal, `Body Weight`, `Breast Circumference`, Breed)
tbl_bw_bc_br
```

```
## # A tibble: 10 x 4
##    Animal ‘Body Weight‘ ‘Breast Circumference‘ Breed
##     <dbl>         <dbl>                  <dbl> <chr>
##  1      1           471                    176 Angus
##  2      2           463                    177 Angus
##  3      3           481                    178 Simmental
##  4      4           470                    179 Angus
##  5      5           496                    179 Simmental
##  6      6           491                    180 Simmental
##  7      7           518                    181 Limousin
```

```
##  8      8           511                    182 Limousin
##  9      9           510                    183 Limousin
## 10     10           541                    184 Limousin
```

- Fitting the linear model

```
lm_bw_bc_br_int <- lm(`Body Weight` ~ `Breast Circumference` * Breed, data = tbl_bw_bc_br)
smry_lm_bw_bc_br_int <- summary(lm_bw_bc_br_int)
smry_lm_bw_bc_br_int
```

```
##
## Call:
## lm(formula = `Body Weight` ~ `Breast Circumference` * Breed,
##     data = tbl_bw_bc_br)
##
## Residuals:
##       1       2       3       4       5       6       7       8       9      10
##   3.286  -4.929  -3.333   1.643   6.667  -3.333   8.200  -5.600 -13.400  10.800
##
## Coefficients:
##                                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)                           430.0000   917.1235   0.469    0.664
## `Breast Circumference`                  0.2143     5.1716   0.041    0.969
## BreedLimousin                       -1151.0000  1293.2741  -0.890    0.424
## BreedSimmental                       -835.6667  1685.4451  -0.496    0.646
## `Breast Circumference`:BreedLimousin    6.5857     7.1908   0.916    0.412
## `Breast Circumference`:BreedSimmental   4.7857     9.4420   0.507    0.639
##
## Residual standard error: 11.17 on 4 degrees of freedom
## Multiple R-squared:  0.9103, Adjusted R-squared:  0.7981
## F-statistic: 8.115 on 5 and 4 DF,  p-value: 0.03212
```

- Expected difference in body weight for the three breeds:

Angus: The expected difference in body weight (in kg) of one centimetre increase in breast circumference corresponds to the regression coefficient of `Breast Circumference` and is

```
smry_lm_bw_bc_br_int$coefficients["`Breast Circumference`", "Estimate"]
```

```
## [1] 0.2142857
```

Limousin: Because, for the breed limousin, there is an interaction effect. We have to add the regression coefficient of `Breast Circumference` to the interaction effect `Breast Circumference`:BreedLimousin. From this we get

```
delta_bw_li <- smry_lm_bw_bc_br_int$coefficients["`Breast Circumference`", "Estimate"] +
  smry_lm_bw_bc_br_int$coefficients["`Breast Circumference`:BreedLimousin", "Estimate"]
delta_bw_li
```

```
## [1] 6.8
```

Simmental: The same as for limousin, we have for simmental

```
delta_bw_si <- smry_lm_bw_bc_br_int$coefficients["`Breast Circumference`", "Estimate"] +
  smry_lm_bw_bc_br_int$coefficients["`Breast Circumference`:BreedSimmental", "Estimate"]
delta_bw_si
```

```
## [1] 5
```

Problem 2: Simulation

Use the following values for intercept and regression slope for `Body Weight` on `Breast Circumference` to simulate a dataset of size N. What is the number for N that has to be chosen such that the regression analysis of the simulated data gives the same result as the true regression slope.

The true values are:

- Intercept: -1070
- Regression slope: 8.7
- Residual standard error: 12

Hints

- Start with $N = 10$, simulate a dataset and analyse the data with `lm()`
- If the result (rounded to 1 digits after decimal point) is not the same then double the size of the dataset, hence use, $N = 20$
- Continue until you get close to the true value.
- Assume that the random resiudals follow a normal distribution with mean zero and standard devation equal to 12
- Take breast circumference to be normally distributed with a mean of 180 and a standard deviation of 2.6
- Use a linear regression model with an intercept to model expected body weight based on breast circumference.

Solution

We start with $N = 10$ and first generate the matrix X which consists of a column of all ones and a column of breast circumference values in centimetre taken from the given normal distribution. Whenever, we generate some random numbers it is important to first set the seed with the function `set.seed()` to which an integer number is passed. This makes sure that when repeating the simulation the same results are generated.

```
set.seed(1234)
vec_bc <- rnorm(n_nr_obs, mean = n_mean_bc, sd = n_sd_bc)
mat_X <- matrix(c(rep(1,n_nr_obs), vec_bc), ncol = 2)
mat_X
```

```
##       [,1]     [,2]
##  [1,]    1 176.8616
##  [2,]    1 180.7213
##  [3,]    1 182.8195
##  [4,]    1 173.9012
##  [5,]    1 181.1157
##  [6,]    1 181.3157
```

```
##  [7,]    1 178.5057
##  [8,]    1 178.5788
##  [9,]    1 178.5324
## [10,]    1 177.6859
```

Together with the given true values of intercept and slope, and randomly generated residuals, observations are simulates.

```
vec_b <- c(n_b_intercept, n_b_slope)
vec_y <- crossprod(t(mat_X), vec_b) + rnorm(n_nr_obs, mean=0, sd=n_res_std_error)
vec_y
```

```
##           [,1]
##  [1,] 462.9699
##  [2,] 490.2948
##  [3,] 511.2150
##  [4,] 443.7138
##  [5,] 517.2207
##  [6,] 506.1236
##  [7,] 476.8673
##  [8,] 472.7008
##  [9,] 473.1860
## [10,] 504.8574
```

The simulated data is analysed with a linear regression model

```
tbl_bw_bc_sim <- tibble::tibble(BodyWeight = vec_y, BreastCircumference=vec_bc)
lm_bw_bc_sim <- lm(BodyWeight ~ BreastCircumference, data = tbl_bw_bc_sim)
lm_bw_bc_sim
```

```
##
## Call:
## lm(formula = BodyWeight ~ BreastCircumference, data = tbl_bw_bc_sim)
##
## Coefficients:
##         (Intercept)  BreastCircumference
##            -916.199                7.833
```

The absolute deviation between the true value of the slope and the estimated slope from the simulated data is

```
abs(lm_bw_bc_sim$coefficients[["BreastCircumference"]] - n_b_slope)
```

```
## [1] 0.8671283
```

In the following iteration, the size of the dataset is doubled in each iteration round until, the absolute deviation of the estimated slope from the true value becomes smaller than 0.1.

```r
n_max_iter <- 10
n_iter_round <- 0
while(abs(lm_bw_bc_sim$coefficients[["BreastCircumference"]] - n_b_slope) > n_slope_tol &&
      n_iter_round < n_max_iter){
  # count number of iterations and determine number of observations
  n_iter_round <- n_iter_round + 1
  n_nr_obs <- 2 * n_nr_obs
  # simulate breast circumference
  vec_bc <- rnorm(n_nr_obs, mean = n_mean_bc, sd = n_sd_bc)
  mat_X <- matrix(c(rep(1,n_nr_obs), vec_bc), ncol = 2)
  # simulate body weight
  vec_y <- crossprod(t(mat_X), vec_b) + rnorm(n_nr_obs, mean=0, sd=n_res_std_error)
  # analyse simulated data
  tbl_bw_bc_sim <- tibble::tibble(BodyWeight = vec_y, BreastCircumference=vec_bc)
  lm_bw_bc_sim <- lm(BodyWeight ~ BreastCircumference, data = tbl_bw_bc_sim)
  # results
  cat(" * Iteration: ", n_iter_round, "\n")
  cat(" * Number of observations: ", n_nr_obs, "\n")
  cat(" * Regression slope: ", lm_bw_bc_sim$coefficients[["BreastCircumference"]], "\n")

}
```

```
##  * Iteration:  1
##  * Number of observations:  20
##  * Regression slope:  7.051858
##  * Iteration:  2
##  * Number of observations:  40
##  * Regression slope:  8.936078
##  * Iteration:  3
##  * Number of observations:  80
##  * Regression slope:  8.184633
##  * Iteration:  4
##  * Number of observations:  160
##  * Regression slope:  8.189888
##  * Iteration:  5
##  * Number of observations:  320
##  * Regression slope:  8.361692
##  * Iteration:  6
##  * Number of observations:  640
##  * Regression slope:  8.232463
##  * Iteration:  7
##  * Number of observations:  1280
##  * Regression slope:  8.638768
```