Theoretical Exercise Analysis of Continuous Data

Thomas Sertijn, Bart Smets, Ilja Van Bever, Lieselot Van de Putte

2025-11-26

Continuous data Homework

Thomas Sertijn, Bart Smets, Ilja Van Bever, Lieselot Van De Putte

This theoretical exercise is based on the article on A Material Paradox: Socioeconomic Status, Young People's Disposable Income and Consumer Culture by West et al. (2006). Everything is to be derived from its table 1, that is also presented in Figure 1. It provides information on income of teenagers in West Scotland. Based on the provided summary data, we will estimate regression coefficients for a multivariable regression model.

We expect you to write out and justify calculations, whether in matrix- or simple form, but the calculations themselves can be done in R - restricted to (matrix-) multiplication, addition,…

0. One piece of information missing, is the number of correspondents (by age and gender). Table 5 (not provided here) does provide some info on total sample size. We will assume we work with 2142 correspondents, equally divided over the six groups (combination of age and gender) • From table 1, does assume an equal number of girls and boys within each age-group, seem reasonable?

Total almost always looks like the average of girl and boy so yes?

1. We could consider evaluating age as a continuous variable. • Looking at Figure 1, would such a model provide a good fit? Explain

The effect looks non-linear, as the increase in total income between age 13 and 15 is a lot bigger than between 11 and 13. So it is probably better to keep it categorical. Difficult to say with only 3 datapoints though.

2. Depending on your answer in 1. fit either a model with age and gender as categorical, but without interaction OR the same model without interaction but with age as a continuous predictor • Write out the model

$$Y_i = \beta_0 + \beta_1 \cdot \text{Age13}_i + \beta_2 \cdot \text{Age15}_i + \beta_3 \cdot \text{Female}_i + \varepsilon_i$$

- Estimate the coefficients

$$\hat{\beta} = (X'X)^{-1}X'y$$

with

$$X'X = \sum_{j=1}^{6} n_j \cdot x_j x_j'$$

and

$$X'y = \sum_{j=1}^{6} n_j x_j \bar{y}_j$$

```
# Data input
ntot <- 2142
n <- 357 #(2142/6)

age <- c(11, 11, 13, 13, 15, 15)
gender <- c("Male", "Female", "Male", "Female", "Male", "Female")
mean <- c(4.56, 4.38, 8.78, 8.30, 15.94, 15.13)
sds <- c(4.10, 4.16, 6.97, 6.57, 16.21, 12.90)

#est coeff
XtransX <- matrix(0, 4, 4)

# 11 yo, male
x1 <- c(1, 0, 0, 0)
XtransX <- XtransX + n * outer(x1, x1)

# 11yo, female
x2 <- c(1, 0, 0, 1)
XtransX <- XtransX + n * outer(x2, x2)

# 13 yo, male
x3 <- c(1, 1, 0, 0)
XtransX <- XtransX + n * outer(x3, x3)

# 13yo, female
x4 <- c(1, 1, 0, 1)
XtransX <- XtransX + n * outer(x4, x4)

# 15yo, male
x5 <- c(1, 0, 1, 0)
XtransX <- XtransX + n * outer(x5, x5)

# 15yo, female
x6 <- c(1, 0, 1, 1)
XtransX <- XtransX + n * outer(x6, x6)

print(XtransX)
```

```
##      [,1] [,2] [,3] [,4]
## [1,] 2142  714  714 1071
## [2,]  714  714    0  357
## [3,]  714    0  714  357
## [4,] 1071  357  357 1071
```

```
# Get Xtransposed*y

Xtransy <- n * (x1 * mean[1] +
            x2 * mean[2] +
            x3 * mean[3] +
            x4 * mean[4] +
            x5 * mean[5] +
            x6 * mean[6])

print(Xtransy)
```

```
## [1] 20381.13  6097.56 11091.99  9928.17
```

```
#Solve eq
beta_hat <- solve(XtransX) %*% Xtransy

print(beta_hat[1]) #intercept
```

```
## [1] 4.715
```

```
print(beta_hat[2])
```

```
## [1] 4.07
```

```
print(beta_hat[3])
```

```
## [1] 11.065
```

```
print(beta_hat[4])
```

```
## [1] -0.49
```

- Estimate the residual variance

$$\hat{\sigma}^2 = \frac{SSR}{N-p},$$

where

$$SSR = \sum_{i=1}^{N} (y_i - \hat{y}_i)^2$$

is the sum of squared residuals

For any individual i in group j

$$\hat{y}_i = \bar{y}_j.$$

–> residual is given by

$$y_i - \hat{y}_i = y_i - \bar{y}_j.$$

Therefore,

$$SSR = \sum_{j=1}^{6} \sum_{i \in \text{group } j} (y_i - \bar{y}_j)^2.$$

sample variance in group j is gvien by:

$$s_j^2 = \frac{1}{n_j - 1} \sum_{i \in j} (y_i - \bar{y}_j)^2.$$

$$(n_j - 1) s_j^2 = \sum_{i \in j} (y_i - \bar{y}_j)^2.$$

Thus, with the formula discussed earlier

$$SSR = \sum_{j=1}^{6} (n_j - 1) s_j^2.$$

Plug into formula

$$\hat{\sigma}^2 = \frac{SSR}{N - p} = \frac{\sum_{j=1}^{6} (n_j - 1) s_j^2}{N - p}.$$

```
ssd_in <- sum((n - 1) * sds^2)
dof<-ntot-4
resvar<- ssd_in/dof
print(resvar)
```

```
## [1] 92.41938
```

- Interpret the gender-effect parameter(s) Females have a lower income on average than males, across all age groups together. The difference is small though (0.49 pounds).
- What would change above if sample consisted of brother-sister pairs?

The assumption of independence would be violated as observations are correlated (brothers and sisters come from same family).

- (Continue as if all outcomes are independent)

3. Fit a standard linear model with both age and gender as categorical variables. Include their interaction. • Write out the model • Estimate the coefficients • Estimate the residual variance • Interpret the gender-effect parameter(s)
4. Compare both models in terms of: • Underlying assumptions • Formally test if the model with interaction (3.) fits the data better
5. From the model with interaction (3.), derive a 95% prediction interval for the income of a 15 year old girl. Is this prediction interval likely to have the nominal 95% coverage? Why? If not, will the coverage tend to be higher or lower?