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

By treating age as a continuous variable, assumptions about functional form are made. Here, 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. Therefore it seems more reasonable 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 • Estimate the coefficients • Estimate the residual variance

We will model the income of an individual i of group j by the following formula.

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{Age13}_{ij} + \beta_2 \cdot \text{Age15}_{ij} + \beta_3 \cdot \text{Female}_{ij} + \varepsilon_{ij}$$

Design matrix X:

So we assume six groups with n_j (357) members.

- $x1$ is the group with 11 year old males;
- $x2$ is the group with 11 year old females;

- $x3$ is the group with 13 year old males;
- $x4$ is the group with 13 year old females;
- $x5$ is the group with 15 year old males;
- $x6$ is the group with 15 year old females.

β can be estimated using the following equation:

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

After computation, the following vector is obtained for $\hat{\beta}$:

```
##           beta estimated
## Intercept          4.715
## Age13              4.070
## Age15             11.065
## Female            -0.490
```

These are the estimated slope coefficients.

To estimate the residual variance, the SSE (sum of squared errors of the residuals) will be calculated. The SSE is not restricted to the SSE due to the modelling of the means of the groups (since the model without interaction terms will not give exact predictions of the group means). Also inside the groups there is a natural deviation from the mean. The SSE can be calculated as the sum of

- $n_j \cdot SSE_{model}$: the sum of squared residuals that is due to the fact that a linear model with 4 parameters is fitted to 6 data points. SS_{model} is multiplicated with n_j (the number of observations in each group), because SSE_{model} is the sum for only six datapoints;
- SSE_{groups}: the sum of squared residuals that is due to the fact that there is a variance in the groups.

$$SSE_{total} = n_j \cdot SSE_{model} + SSE_{groups} = n_j \cdot \sum_{j=1}^{6}(\hat{y_j} - y_j)^2 + \sum_{j=1}^{6}\sum_{i=1}^{n_j}(y_{ij} - y_j)^2$$

with $\hat{y_j}$ the estimated average income of individuals of group j, y_j the observed average income of individuals of group j and y_{ij} the observed income of an individual of group j.

To estimate SSE_{model} the following equations are used.

$$H = X(X'X)^{-1}X'$$

$$SSE_{model} = Y'(I - H)Y$$

After calculation for SSE_{model} a value of 0.0993 can be obtained.

To calculate SSE_{groups} for each group the formula $j \sum_{i=1}^{n_j}(y_{ij} - \bar{y_j})^2$ is calculated based on the given standard deviations. We consider that these standard deviations are calculated from the data itself, so consider $n_j - 1$ degrees of freedom.

$$s_j^2 = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(y_{ij} - y_j)^2 \quad \text{so} \quad \sum_{i=1}^{n_j}(y_{ij} - y_j)^2 = (n_j - 1) \cdot s_j^2$$

So SSE_{groups} can be calculated as follows.

$$SSE_{groups} = \sum_{j=1}^{6}\sum_{i=1}^{n_j}(y_{ij} - y_j)^2 = \sum_{j=1}^{6}\left[(n_j - 1) \cdot s_j^2\right]$$

And MSE as follows.

$$MSE = \frac{SSE_{total}}{n_{tot} - p} = \frac{n_j \cdot SSE_{model} + SSE_{groups}}{n_{tot} - p}$$

with:

- $SSE_{model} = 0.0993$;
- $SSE_{groups} = 1.9759264 \times 10^5$;
- $n_{tot} = 2142$;
- $p = 4$.

After calculation we derive that MSE is equal to 92.4359626.

- Interpret the gender-effect parameter(s)

The expected value of the <moet hier nog 'mean' staan?> income of a female is 0.49 pounds less than that of a male, assumed that they are in the same age group (11 years, 13 years or 15 years).

- 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 the same family). This means that SE's are too small as you only have 1071 indep families, not 2142. Siblings give overlapping info. You'd thus underestimate uncertainty as confidence intervals are too small.

- (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)

We will model the income of an individual i of group j by the following formula.

$$Y_{ij} = \beta_0 + \beta_1 \cdot \text{Age13}_{ij} + \beta_2 \cdot \text{Age15}_{ij} + \beta_3 \cdot \text{Female}_{ij} + \beta_4 \cdot \text{Age13}_{ij} \cdot \text{Female}_{ij} + \beta_5 \cdot \text{Age15}_{ij} \cdot \text{Female}_{ij} + \varepsilon_{ij}$$

To solve this exercise the following X-matrix can be used:

```
##    [,1] [,2] [,3] [,4] [,5] [,6]
## x1    1    0    0    0    0    0
## x2    1    0    0    1    0    0
## x3    1    1    0    0    0    0
## x4    1    1    0    1    1    0
## x5    1    0    1    0    0    0
## x6    1    0    1    1    0    1
```

So the X-matrix has two extra columns, because of the two extra parameters.

β can again be estimated using the following equation:

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

After computation, the following vector is obtained for $\hat{\beta}$:

```
##              beta estimated
## Intercept              4.56
## Age13                  4.22
## Age15                 11.38
## Female                -0.18
## Age13*Female          -0.30
## Age15*Female          -0.63
```

The MSE can be calculated using the same formulas as mentioned in exercise 2.

$$MSE = \frac{SSE_{total}}{n_{tot} - p} = \frac{n_j \cdot SSE_{model} + SSE_{groups}}{n_{tot} - p}$$

with:

- $SSE_{model} = 1.7449469 \times 10^{-13}$;
- $SSE_{groups} = 1.9759264 \times 10^{5}$;
- $n_{tot} = 2142$;
- $p = 6$.

After calculation we derive that MSE is equal to 92.5059167. So the MSE is larger now than the MSE in exercise 2, because the denominator in the formula is now smaller (more parameters in the model).

It can be observed that SSE_{model} is negligible. Actually SSE_{model} is (ik zou essentially ipv equal zeggen -> voorzichtiger) equal to zero, because a model with six parameters is fitted to 6 data points. So the group the (bedoel je 'of'?) estimates for the averages of the groups will be exactly equal to the observed averages.

$Y_{ij} = \beta_0 + \beta_1 \cdot \text{Age13}_{ij} + \beta_2 \cdot \text{Age15}_{ij} + \beta_3 \cdot \text{Female}_{ij} + \beta_4 \cdot \text{Age13}_{ij} \cdot \text{Female}_{ij} + \beta_5 \cdot \text{Age15}_{ij} \cdot \text{Female}_{ij} + \varepsilon_{ij}$

(waarom staat die hier nog eens?)

The gender-effect parameters can be interpreted as follows.

- β_3 = -0.18: a female of 11 years old is expected to have an income of 0.18 (-β_3) pound less than a male of 11 years old;
- β_4 = -0.3: -0.18 (β_3) + -0.3 (β_4) = -0.48, so a female of 13 years old is expected to have an income of 0.48 pounds less than a male of 13 years old;
- β_5 = -0.63: -0.18 (β_3) + -0.63 (β_5) = -0.81, so a female of 15 years old is expected to have an income of 0.81 pound less than a male of 15 years old;

4. Compare both models in terms of: • Underlying assumptions

In the model without interaction, gender is assumed to have the same effect regardless of ages (gender-effect independent of age_effect). In theory this means that gender would only influence the intercept and not the slope. Since in this model only qualitative variables are included, we take this to mean that the vertical distance between predicted means for males and females are is constant across age categories. The regression functions for all combinations are parallel to each other.

In the second model, where interaction is included, the effect of gender is no longer independent and changes with age. This is a reinforcement-type effect: with increasing age, the gender gap increases as well (the 'slope' becomes steeper).

- Formally test if the model with interaction (3.) fits the data better

In theorie obs MSE niet? Maar volgens de simulaties wel, omdat een deel van de variantie binnen de groepen ook gemodelleerd wordt door het model? SSE zal wel beter zijn. Rkwadraat? Rkwadraat adjusted?...

Gewoon F-test met berekende SSE's van Q2 en Q3 vglen?

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?

In principe gaat een lineair model wel uit van homoscedasticiteit, dus dan gaat je coverage niet uitkomen (aangezien de data duidelijk heteroscedastisch zijn) Je zou voor de bepalen van je predictie interval wel kunnen rekening houden met de variantie waargenomen voor meisjes van 15 jaar.

Voor 15-jarige meisjes gaat je coverage te klein zijn nromaal, aangezien je lineaire model je mse poolt terwijl je variantie wel groter is voor 15yo meisjes. Weet ook niet of nodig is om te vermelden dat inkomensdistributie afgecut is op 0.