Theoretical Exercise Analysis of Continuous Data

Thomas Sertijn, Ilja Van Bever, Lieselot Van de Putte

2025-12-04

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?

The average of the total group doesn't deviate a lot from the average of the average for girls and the average for boys. So the number of boys and the number of girls in the study will not be very different. Although there are some small differences between the average value of the total group and the average of the average of the subgroups. This can be due to small differences in sample size or number of missing values.

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. Ideally, we would need more than three datapoints to investigate this relationship.

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 i 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 mean 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 equal to zero, because a model with six parameters is fitted to 6 data points. So the estimates of the averages of the groups will be exactly equal to the observed averages.

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 is constant across age categories. The regression functions for all combinations are parallel to each other. This is also valid for the effect of age on gender.

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) and thus the vertical distance between predicted mean for male and female increases with age.

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

We will use the F-statistic to determine if the model with interaction fits the data better: $H0:$ all interaction coefficients = 0

```
##            [,1]
## [1,] 0.1916099
```

```
## [1] 2.999938
```

$F^* = 0.1916 < 2.9999$, therefore we conclude that the model with interaction does not reduce SSE much and we cannot reject the null hypothesis.

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?

For a 15-year old girl in the interaction model, the estimated income would be calculated as

$$\hat{Y}_{15,girl} = \beta_0 + \beta_2 + \beta_3 + \beta_5.$$

And would thus result in $\hat{Y}_{15,\text{girl}} = 15.13$

The variance of the predicton can be calculated using:

$$s^2(\hat{Y}_{15,girl}) = MSE(1 + X_h'(X'X)^{-1}X_h)$$

Which would give us a 95% prediction interval:

$$PI_{95\%}(15, girl) = [\hat{Y}_{15,girl} \pm t_{0.975,df} * s(\hat{Y}_{15,girl})]$$

Resulting in:

```
## lower_bound  prediction upper_bound
##   -11.54436    15.13000    41.80436
```

A linear model assumes linear form is right, errors are independent, homoscedasticty, and normally distributed. For this data, these assumptions are clearly violated, e.g. the data is clearly right-skewed, which is indicated by the large average standard deviation on the money predictor. Since we work with data of children and teens, we expect that the lowest value possible for the money is zero, no debt. Because of this it could be seen that within one or two SDs all total money would go below the zero cutt-off, showing that the data is likely right-skewed.

Next to this, the data is most likely heteroscedastic, as the variances differ massively across ages, showing a dependence of the income variance with the age predictors, showing that the data is not homoscedastic.

These violations will result in the prediction interval not having a nominal 95% coverage. Here, we would expect the prediction interval to be to wide for younger children as they have a lower variance. While the prediction interval will be to narrow for older children. So for the case of a 15 years-old girl, the coverage of the prediction interval will be lower.

These violations will result in the prediction interval not having a nominal 95% coverage. Here, we would expect the prediction interval to be to wide for younger children as they have a lower variance. While the prediction interval will be to narrow for older children. So for the case of a 15 years-old girl, the coverage of the prediction interval will be lower.

appendix

```
knitr::opts_chunk$set(
  echo = FALSE,        # code verbergen
  message = FALSE,     # geen messages
  warning = FALSE      # geen warnings
)
options(width = 100)
# Solution with matrices
```

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

age <- c(11, 11, 13, 13, 15, 15)
#age <- rep(c(11,13,15), each = 2)
gender <- c("Male", "Female", "Male", "Female", "Male", "Female")
#gender <- rep(c("Male", "Female"), times = 3)
mean_vector <- c(4.56, 4.38, 8.78, 8.30, 15.94, 15.13)
sds_vector <- c(4.10, 4.16, 6.97, 6.57, 16.21, 12.90)

x1 <- c(1, 0, 0, 0) # 11yo, male
x2 <- c(1, 0, 0, 1) # 11yo, female
x3 <- c(1, 1, 0, 0) # 13yo, male
x4 <- c(1, 1, 0, 1) # 13yo, female
x5 <- c(1, 0, 1, 0) # 15yo, male
x6 <- c(1, 0, 1, 1) # 15yo, female

X1 <- rbind(x1, x2, x3, x4, x5, x6)
beta1 <- solve(t(X1) %*% X1) %*% t(X1) %*% mean_vector
HAT1 <- X1 %*% solve(t(X1) %*% X1) %*% t(X1)
SSE_mod1 <- t(mean_vector) %*% (diag(6) - HAT1) %*% mean_vector
SSE1 <- sum((n-1)*(sds_vector^2)) + SSE_mod1*n
MSE1 <- SSE1/(ntot-4)
rownames(beta1) <- c("Intercept", "Age13", "Age15", "Female")
colnames(beta1) <- "beta estimated"
print(beta1)
# Solution with matrices

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

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

x1 <- c(1, 0, 0, 0, 0, 0) # 11yo, male
x2 <- c(1, 0, 0, 1, 0, 0) # 11yo, female
x3 <- c(1, 1, 0, 0, 0, 0) # 13yo, male
x4 <- c(1, 1, 0, 1, 1, 0) # 13yo, female
x5 <- c(1, 0, 1, 0, 0, 0) # 15yo, male
x6 <- c(1, 0, 1, 1, 0, 1) # 15yo, female


X2 <- rbind(x1, x2, x3, x4, x5, x6)
beta2 <- solve(t(X2) %*% X2) %*% t(X2) %*% mean_vector

HAT2 <- X2 %*% solve(t(X2) %*% X2) %*% t(X2)
SSE_mod2 <- t(mean_vector) %*% (diag(6) - HAT2) %*% mean_vector
SSE2 <- sum((n-1)*(sds_vector^2)) + SSE_mod2*n
```

```r
MSE2 <- SSE2/(ntot-6)
X2
rownames(beta2) <- c("Intercept", "Age13", "Age15", "Female", "Age13*Female", "Age15*Female")
colnames(beta2) <- "beta estimated"
print(beta2)
f <- ((SSE1 - SSE2) / 2) / (SSE2 / (ntot - 6))
critical_value <- qf(0.95, df1 = 2, df2 = 2136)
f
critical_value
predicted_income_15 <- beta2[1] + beta2[3] + beta2[4] + beta2[6]
XtX_inv <- solve(t(X2) %*% X2)
predicted_var <- MSE2 * (1 + t(x6) %*% XtX_inv %*% x6)
prediction_se <- sqrt(predicted_var)
df <- ntot - nr_of_betas
t95 <- qt(0.975, df)
lower_border <- predicted_income_15 - t95*prediction_se
higher_border <- predicted_income_15 + t95*prediction_se

c(lower_bound = lower_border, prediction = predicted_income_15,  upper_bound = higher_border)
```