Case Study 2: Econometrics I / Ökonometrie I

WS 2025/2026

Deadline: November 10, 2025@ 23:59

One of the most important health measures of newborn babies is their birth weight. Low birth weight, usually defined as weighing less than 2,500 grams, is associated with health and developmental issues of the baby. In this case study, we will analyze factors that influence birth weight using a subset of the data of Almond, Chay and Lee (2005).

1 Data Acquisition (1 point)

Download the zip-folder "Birthweight and Smoking" from the webpage that we already used in Case Study 1, and load the file "birthweight_smoking.xlsx" into R.

Save the variables `birthweight`, `age`, `educ`, `drinks`, `smoker`, and `tripre0` in a separate data frame, and explain what these variables are measuring.

Consider the relationship between (i) `birthweight` and `smoker`; (ii) `birthweight` and `tripre0`; and (iii) `birthweight` and `age`. For each of those 3 pairs of variables answer the following two questions:

- Which relation can one expect between the variables? Provide some scientific evidence (e.g., a journal publication) justifying your answer.
- Does the empirical correlation between the variables support your expectation?

2 Plots (2 points)

Using the **ggplot2** package and the `geom_density` function, create a nice[1] overlayed density plot of:

1. `birthweight` for the two categories of the variable `smoker`.
2. `birthweight` for the two categories of the variable `tripre0`.

Interpret your graphs. Do they align with the answers you provided in question 1?

Hint: In the `aes` option of the `ggplot()` function, use the option `color = factor(smoker)` and `color = factor(tripre0)`.

3 Multiple Linear Regression

3.1 First Part (1 point)

Denote by Y the variable `birthweight`, and denote by X_1, X_2, X_3, X_4 and X_5 the variables `age`, `educ`, `drinks`, `smoker`, and `tripre0`, respectively.

Consider the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u, \quad \mathbb{E}[u \mid X_1, X_2, X_3, X_4, X_5] = 0. \tag{1}$$

Theoretically, using the coefficients from (1), answer:

1. What would be the effect on the weight at birth of an increase of 2 years in the age of the mother?
2. What is the effect on the weight at birth of smoking during pregnancy?

[1] For example, add the title of the graph, the title of the axes and the title of the legend.

3.2 Second Part (2 points)

Estimate the model in (1), and the model in (1) but without the variable `smoker`.

Show a summary of the results and answer:

1. Which of the two models explains more variability in the outcomes? Are you surprised by this result? We shall call this model `model+` from now on.
2. In `model+`, do your estimates align with the answers you provided in question 1?
3. In `model+`, what's the effect on `birthweight` (in grams) of going to at least one prenatal visit?

3.3 Third Part (1 point)

Based on the results of the estimation of `model+`, recall the formula for an unbiased estimator of σ^2 and use this formula and the residuals of that regression to compute an unbiased estimator of the error variance. Show your result.

Recall the formula for the variance-covariance matrix of the OLS estimator. Using this formula and the estimator of the error variance computed above, compute and show an estimator of the variance-covariance matrix of the OLS estimator. What is the value of the covariance of the estimators of `educ` and `tripre0`? Interpret the sign of this covariance?

4 Hypotheses Testing and Prediction (3 points)

To answer the questions in this section you will use the results of `model+`. When answering, state the null and alternative hypotheses, work with $\alpha = 5\%$, and compute (or show using the results of the regression) the respective test statistic. Argue how the value of the test statistic together with a suitable critical value allow one to reject or not reject the respective null hypothesis. Explain how to achieve the same result using the respective p-value. Interpret the results.

1. Does the weekly number of drinks during pregnancy have an impact on the weight at birth?
2. Do prenatal medical visits have an impact on the weight at birth?

Using your estimate for the coefficient and the error variance from question 3, compute the following test statistic for the parameter of the variable `tripre0`

$$t_2 = \frac{\hat{\beta}_2 - 1}{\text{sd}(\hat{\beta}_2)}; \tag{2}$$

Which hypothesis can be tested with (2)? To test this hypothesis using the estimator for the error variance, would you use a standard normal or a student-t distribution? Compute the p-value using the appropriate distribution and interpret your result (use $\alpha = 5\%$).

How you would predict `birthweight` making use of `model+` for a given set of values in the predictors of your choice.

References

Douglas Almond, Kenneth Y. Chay, David S. Lee. *The Costs of Low Birth Weight.* The Quarterly Journal of Economics, 120(3), 2005, pp. 1031–1083. https://doi.org/10.1093/qje/120.3.1031