Econometrics I - Case Study 2

Group 3, Instructor: Univ.Prof. David Preinerstorfer, Ph.D

Julian Hofmaninger, Tsz Lam Hung and Daniel Diederichs

November 10, 2025

Problem 1: Data aquisition

Loading data

```
library(readxl)
data <- read_xlsx("Birthweight_Smoking/birthweight_smoking.xlsx")
```

Describing variables

```
data<- data[, c('birthweight', 'age', 'educ', 'drinks', 'smoker', 'tripre0')]
```

- `birthweight`: Measures the weight (in grams) of babies at the time of their birth.
- `age`: Describes the age of the mother at the time her child was born.
- `educ`: Describes how many years of education the mother of the newborn has. Education of more than 16 years is simply represented as 17.
- `drinks`: Describes the number of drinks a mother used to have during her pregnancy.
- `smoker`: A numeric value that is equal to one if the mother of the newborn smoked during pregnancy and equal to zero if she did not smoke during the pregnancy.
- `tripre0`: A numeric value that is equal to one if the mother had no prenatal visit during her pregnancy and zero if she had at least one.

Predicted correlation

- `birthweight` and `smoker`: We would suggest that newborns whose mother smoked during pregnancy have a lower birthweight. Smoking generally has negative effects on health and as the article [2] suggests, the effects of a mother smoking during pregnancy also has negative effects on the health of the unborn child often resulting in lower birthweight.
- `birthweight` and `tripre0`: We suggest that prenatal care has a positive effect on the health of the baby and therefore leading to higher birthweight. As the studies for Africa mentioned in the article [1] show mother's who have atleast one prenatal visit during their pregnancy tend to rather give birth to a baby of "normal" weight. Concluding we suggest a negative correlation of `birthweight` and `tripre0`.
- `birthweight` and `age`: Pregnancy tends to become unsafer with an increasing age of the mother which would suggest a negative correlation between `birthweight` and `age`. However, the article [3] stresses that younger mothers "compete" for nutrients with their unborn babies. Therefore, birthweight tends to be lower on the "extreme" sides of mother's age and the correlation mainly depends on the which "extreme-side" is weighted heavier in the sample of observation.

Empirical correlation

```
cor(data$birthweight, data$smoker)
```

```
## [1] -0.1691266
```

The computed empirical correlation between `birthweight` and `smoker` is a negative number. This suggests that the birthweight of babies is higher when the mother does not smoke and therfore supports our argument.

```
cor(data$birthweight, data$tripre0)
```

```
## [1] -0.1234999
```

The computed empirical correlation between `birthweight` and `tripre0` is a negative number. This suggests that the birthweight of babies is higher when the mother has at least one prenatal visit during her pregnancy and therefore supports our argument.

```
cor(data$birthweight, data$age)
```

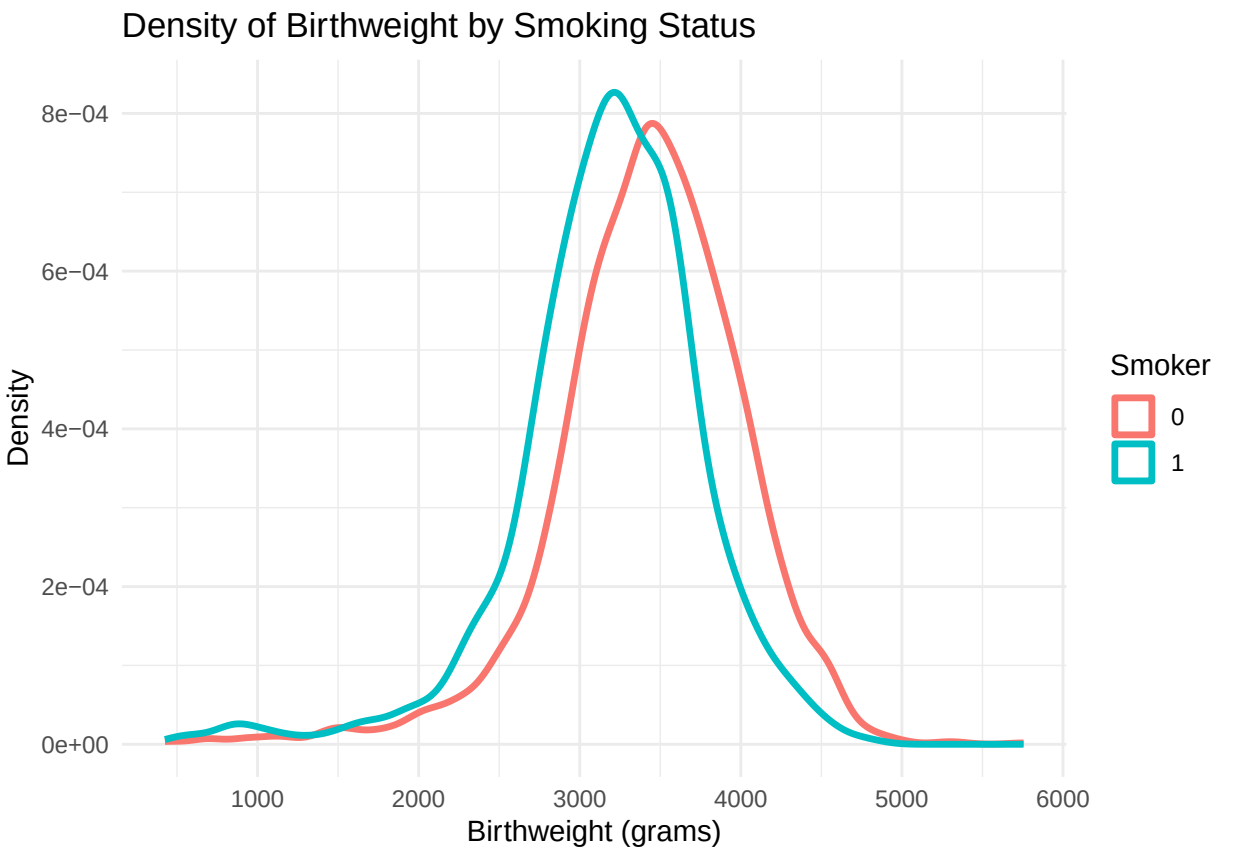
```
## [1] 0.08007321
```

The computed empirical correlation between `birthweight` and `age` is a positive number. This suggests that the birthweight of babies is higher with an increasing age of the mother. However, the correlation is very weak therefore supporting the argument the argument that if the sample is evenly split in terms of age the effects tend to offset each other.

Problem 2: Plots

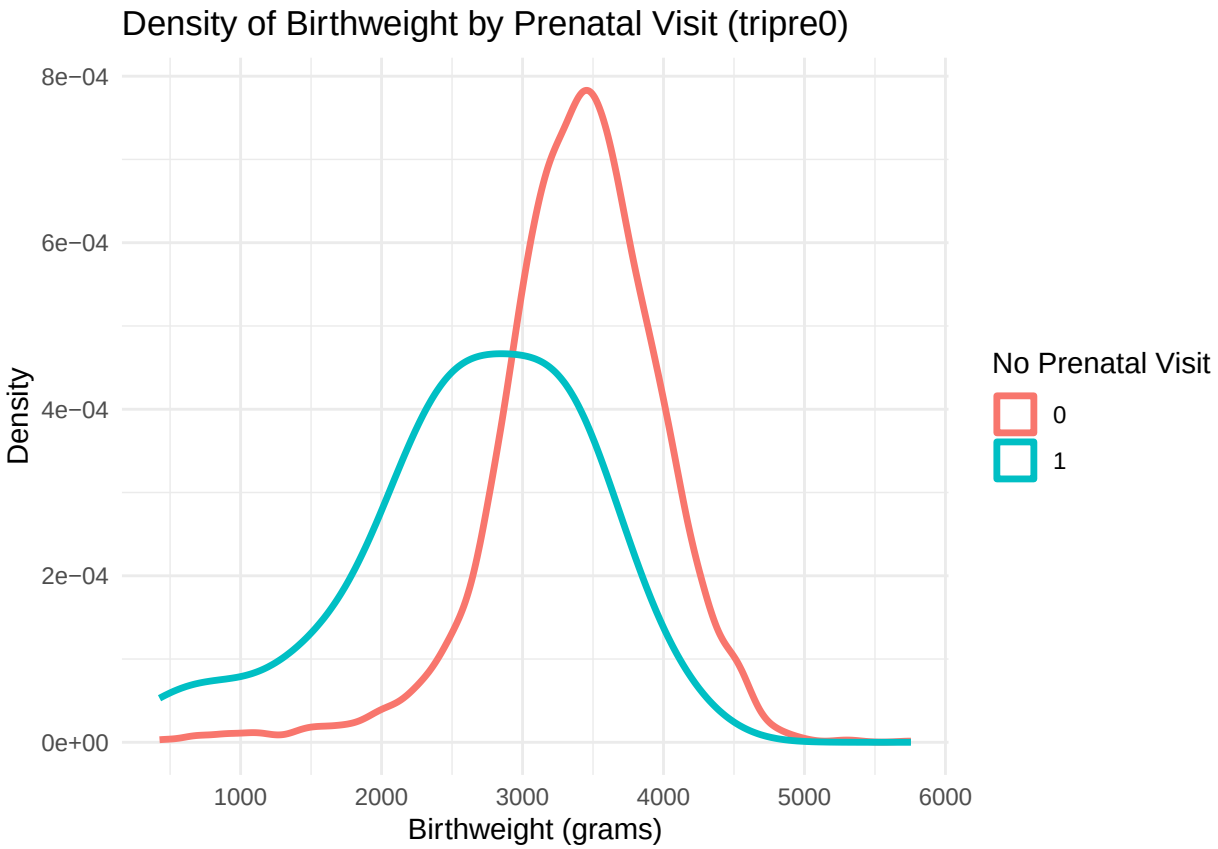
Density plots

```
library(ggplot2)
ggplot(data, aes(x = birthweight, color = factor(smoker))) +
       geom_density(linewidth = 1.2) +
      labs(
             title = "Density of Birthweight by Smoking Status",
             x = "Birthweight (grams)",
             y = "Density",
             color = "Smoker") +     theme_minimal()
```



The graph shows that the distribution of the `birthweight` for the babies of mothers who did not smoke during their pregnancy have a higher mean value for `birthweight`. For the `birthweight` of babies with mothers who did smoke on the other hand the mean value of `birthweight` is lower and therefore supporting the arguments we made prior. Furthermore, the density plot of the smokers appears a little bit narrower which means there is less deviation from the mean and the left (lower birthweight) tail appears a little bit fatter. This aligns with a higher risk of low birthweight for babies of smokers.

```
ggplot(data, aes(x = birthweight, color = factor(tripre0))) +
  geom_density(linewidth = 1.2) +
  labs(
    title = "Density of Birthweight by Prenatal Visit (tripre0)",
    x = "Birthweight (grams)",
    y = "Density",
    color = "No Prenatal Visit"
  ) +
  theme_minimal()
```



The graph shows that the distribution of the `birthweight` for the babies of mothers who had at least one prenatal visit during their pregnancy has a higher mean for `birthweight`. For the `birthweight` of babies with mothers who had no prenatal visit during their pregnancy at all on the other hand has a lower mean value for `birthweight`. Additionally it is clearly visible that the density curve for mother's who had no prenatal visit appears broader therefore indicating a wider spread of the birthweight values. Furthermore, the left tail (lower birthweight) is fatter for the birthweight of babies from mothers with no prenatal visit, thus aligning with the assumption that the risk of lower birthweight increases with no prenatal visits.

Problem 3: Multiple Linear Regression

3.1 First Part: Model definition

```
linear_model <- lm(birthweight ~ age + educ + drinks + smoker + tripre0, data=data)
beta_age <- coef(linear_model)["age"]
2*beta_age
```

```
##      age
## 7.267218
```

If the age of the mother increases by two years (ceteris paribus), the birthweight of the baby increases by 7.27 grams.

```r
beta_smoker <- coef(linear_model)["smoker"]
beta_smoker
```

```
##    smoker
## -216.4648
```

If you take a smoking mother (ceteris paribus), the birthweight of the baby will decrease by 216.46 grams.

3.2 Second Part: Model estimation

```r
summary(linear_model)
```

```
##
## Call:
## lm(formula = birthweight ~ age + educ + drinks + smoker + tripre0,
##     data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2998.01  -304.18    21.97   364.46  2360.34
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3156.187     73.066  43.196  < 2e-16 ***
## age            3.634      2.206   1.647   0.0996 .
## educ          13.817      5.553   2.488   0.0129 *
## drinks       -12.822     15.452  -0.830   0.4067
## smoker      -216.465     27.652  -7.828 6.81e-15 ***
## tripre0     -654.600    106.565  -6.143 9.18e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 578.7 on 2994 degrees of freedom
## Multiple R-squared:  0.04647,    Adjusted R-squared:  0.04488
## F-statistic: 29.18 on 5 and 2994 DF,  p-value: < 2.2e-16
```

```r
linear_model_excluding_smoker <- lm(birthweight ~ age + educ + drinks + tripre0, data=data)
summary(linear_model_excluding_smoker)
```

```
##
## Call:
## lm(formula = birthweight ~ age + educ + drinks + tripre0, data = data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2955.36  -318.80    26.55   369.22  2414.49
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2987.740     70.525  42.364  < 2e-16 ***
```

```
## age            4.477      2.225   2.012   0.0443 *
## educ          21.937      5.510   3.981 7.02e-05 ***
## drinks       -23.895     15.542  -1.537   0.1243
## tripre0     -692.205    107.523  -6.438 1.41e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 584.5 on 2995 degrees of freedom
## Multiple R-squared:  0.02696,    Adjusted R-squared:  0.02566
## F-statistic: 20.74 on 4 and 2995 DF,  p-value: < 2.2e-16
```

The first model has a higher value for `R^2` indicating that it better suited for predicting the variability in the outcomes for the birthweight. This outcome is not suprising as this model has one more predictor variable which allows for better fitting to the outcomes.

```
model_plus <- linear_model
```

- `age`: The estimated coefficient for `age` is a positive value therfore the model suggest increasing birthweight with an increasing age of the mother. That means it contradicts with our suggestion that "extremes" for age will equal out.
- `smoker`: The estimated coefficient for `smoker` is a negative value therfore the model suggests that smoking during the pregnancy leads to a lower birthweight of babies. That aligns with our prior suggestion.
- `tripre0`: The estimated coefficient for `tripre0` is a negative value therfore the model suggests that at least one prenatal visit during pregnancy has a positive effect on the birthweight of a baby. That aligns with our prior suggestion as well.

```
coef(model_plus)["tripre0"]
```

```
##   tripre0
## -654.6002
```

The value of the estimated coefficient for `tripre0` is -654.60 that suggests that if the parameter is 0 (ceteris paribus), meaning that the mother has at least one prenatal visit, the birthweight of an baby increases by 654.60 grams.

3.3 Third Part: Estimator evaluation

```
SSR <- sum(resid(model_plus)^2)
degrees_of_freedom <- model_plus$df.residual
sigma2hat <- SSR / degrees_of_freedom
sigma2hat
```

```
## [1] 334919.1
```

```
X <- cbind(1, data$age, data$educ, data$drinks, data$smoker, data$tripre0)
Xt_X <- t(X) %*% X
vcov_matrix <- sigma2hat * solve(Xt_X)
vcov_matrix[3,6]
```

```
## [1] 19.59575
```

The value for `Cov(educ,tripre0)` can be found at the indices (3,6) or (6,3). There we find a value of 19.596 which is a positive number and therefore indicating that there is a positive dependence on these two variables. Therfore, it can be said that when one of the variables increases the other one tends to increase as well.

Problem 4: Hypotheses Testing and Prediction

Interpretation of the test statistic, p-value and corresponding critical value

The test statistic is the ratio of the difference in the estimate of an parameter and the value it is assumed to have to the standard deviation/error of this parameter. Therefore, it effectively shows how many standard deviations/errors the estimate is away from the assumed value. The critical value is derived from the significance level and the underlying distribution. Say we use the significance level $alpha = 5$ and a standard normal distribution. The corresponding critical value is therefore given by around 1.96. If the t-statistic is higher than this value we reject the null hypothesis because it means that the estimate is less likely due to chance. If it is smaller than the critical value we retain the null hypothesis.

The p-value is associated with the t-statistic and gives the probability in the respective distribution of observing an estimate that is as at least as extreme as the t-value given the condition that the null hypothesis is true. A p-value smaller than the significance level therefore leads us to rejecting the null hypothesis simply because it is very unlikely to observe such "extreme" values given the null hypothesis is true.

Impact of drinks

$$H0 = \beta_3 = 0$$
$$H1 = \beta_3 \neq 0$$

```
alpha <- 0.05
c_alpha <- 1-alpha/ 2
c_alpha
```

```
## [1] 0.975
```

```
critical_value <- qnorm(c_alpha)
t_drinks <- coef(summary(model_plus))[, "t value"]["drinks"]
if(abs(t_drinks) <= critical_value){
  print("Do not reject H0 as the t-statistic is smaller than or equal to the critical value")
} else{
  print("Reject H0 as the t-statistic is greater than the critical value")
}
```

```
## [1] "Do not reject H0 as the t-statistic is smaller than or equal to the critical value"
```

The test statistic is smaller than the `c_alpha` value. Therfore, we do not reject the null hypothesis. This means that under a 5% significance level we can not find evidence that the weekly number of drinks has an influence on the birthweight.

```r
p_drinks <- coef(summary(model_plus))[, "Pr(>|t|)"]["drinks"]
if(p_drinks < alpha){
  print("Reject H0 as the p-value is smaller than the significance level")
} else{
  print("Do not reject H0 as the p-value is greater than or equal to the significance level")
}
```

```
## [1] "Do not reject H0 as the p-value is greater than or equal to the significance level"
```

We can use the `p-value` to show the same thing. In this case the `p-value` is greater than the significance level. Therefore, values that are at least as "extreme" as the value of the t-statistic are plausible to happen under the null hypothesis therefore we refuse to reject the null hypothesis.

Impact of prenatal medical visits

$$H0 = \beta_5 = 0$$
$$H1 = \beta_5 \neq 0$$

```r
t_tripre0 <- coef(summary(model_plus))[, "t value"]["tripre0"]
if(abs(t_tripre0) <= critical_value){
  print("Do not reject H0 as the t-statistic is smaller than or equal to the critical value")
} else{
  print("Reject H0 as the t-statistic is greater than the critical value")
}
```

```
## [1] "Reject H0 as the t-statistic is greater than the critical value"
```

The test statistic is greater than the `c_alpha` value. Therfore, we reject the null hypothesis. This means that unless

```r
p_tripre0 <- coef(summary(model_plus))[, "Pr(>|t|)"]["tripre0"]
if(p_tripre0 < alpha){
  print("Reject H0 as the p-value is smaller than the significance level")
} else{
  print("Do not reject H0 as the p-value is greater than or equal to the significance level")
}
```

```
## [1] "Reject H0 as the p-value is smaller than the significance level"
```

Hypothesis testing for tripre0

With formula (2) we can test the hypothesis that the effect of the parameter for `tripre0` on birthweight is exactly 1 gram. We are dealing with an unknow standard deviation we should use the `student-t distribution` although at the high number of degrees of freedom the crtical value will be quite close to the critical value we would retrieve from the standard normal distribution.

$$H0 = \beta_5 = 1$$
$$H1 = \beta_5 \neq 1$$

```r
critical_value <- qt(c_alpha, df = degrees_of_freedom)
se_tripre0 <- vcov_matrix[6,6]
t_tripre0 <- (coef(model_plus)["tripre0"]-1) / sqrt(se_tripre0)
if(abs(t_tripre0) <= critical_value){
  print("Do not reject H0 as the t-statistic is smaller than or equal to the critical value")
} else{
  print("Reject H0 as the t-statistic is greater than the critical value")
}
```

```
## [1] "Reject H0 as the t-statistic is greater than the critical value"
```

```r
p_tripre0 <- 2 * (1 - pt(abs(t_tripre0), degrees_of_freedom))
print(p_tripre0)
```

```
##      tripre0
## 8.662404e-10
```

```r
if(p_tripre0 < alpha){
  print("Reject H0 as the p-value is smaller than the significance level")
} else{
  print("Do not reject H0 as the p-value is greater than or equal to the significance level")
}
```

```
## [1] "Reject H0 as the p-value is smaller than the significance level"
```

Since the p-value is so small the probability of observing a value as extreme or more extreme than the t-statistic is very unlikely under the null hypothesis. Therefore we reject the null hypothesis and hence the hypothesis that the effect of the `tripre0` parameter is 1 gram.

Predict the birthweight with model+

For a prediction of the birthweight with `model+` we can use the inbuilt `predict` function for a set of given values:

- `age`: 28 years
- `educ`: 12 years
- `drinks`: 2 per week
- `smoker`: 1 (Yes)
- `tripre0`: 1 (No prenatal visit)

```r
newdata <- data.frame(
  age = 28,
  educ = 12,
  drinks = 2,
  smoker = 1,
  tripre0 = 1
)
predicted_bw <- predict(model_plus, newdata = newdata)
predicted_bw
```

```
##        1
## 2527.018
```

References

[1] Garedew Tadege Engdaw et al. “Effect of antenatal care on low birth weight: a systematic review and meta-analysis in Africa, 2022”. In: *Frontiers in Public Health* 11 (2023), p. 1158809.

[2] Mariana Caricati Kataoka et al. “Smoking during pregnancy and harm reduction in birth weight: a cross-sectional study”. In: *BMC pregnancy and childbirth* 18.1 (2018), p. 67.

[3] María Clara Restrepo-Méndez et al. “The association of maternal age with birthweight and gestational age: a cross-cohort comparison”. In: *Paediatric and perinatal epidemiology* 29.1 (2015), pp. 31–40.