Econometrics 1: Case Study 2

Determinants of Birth Weight

Bischoy Bert

November 2025

1 Introduction

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) which contains a census of infant births and deaths. The data in `bw_smoking.data` are for births in Pennsylvania in 1989. To conduct the analysis we will use `R` and aside from its base functions, the packages `tidyverse`, `readxl` and `patchwork` (Almond et al., 2005).

2 Data Acquisition

The raw dataset `birthweight_smoking.xlsx` contains information on birth weight, maternal age, education, weekly alcohol consumption, smoking behavior, prenatal medical visits and more. These variables allow us to examine possible influences on the birth weight of infants.

The dataset used during this analysis shall be titled `bw_smoking` and has been edited to contain only selected variables:

```
birthweight_smoking <- data.frame(readxl::read_xlsx("birthweight_smoking.xlsx"))

bw_smoking <- birthweight_smoking |>
  select(birthweight, age, educ, drinks, smoker, tripre0)
```

The following variables were selected for the analysis:

- **birthweight**: Infant birth weight in grams
- **age**: Age of the mother in years

- **educ**: Years of maternal education (values above 16 coded as 17)
- **drinks**: Number of alcoholic drinks per week during pregnancy
- **smoker**: Indicator variable (1 if mother smoked during pregnancy, 0 otherwise)
- **tripre0**: Indicator variable (1 if no prenatal visits, 0 otherwise)

Now we shall consider the relationship between (i) `birthweight` and `smoker`, (ii) `birthweight` and `tripre0`, (iii) `birthweight` and `age`. For each of those 3 pairs of variables shall answer the following two questions:

1. Which relation can one expect between the variables?
2. Does the empirical correlation between the variables support our expectation?

1.

(i) We expect a negative relationship between the variables `smoker` and `birthweight`. Assuming that higher birth weight is indicative of better infant health, it naturally follows that a greater value for `smoker` (i.e. "1") would imply a lower birth weight and vice versa. According to a study published by the *National Library of Medicine*, there is sufficient evidence to infer a causal link between active or passive maternal smoking and low birth weight or preterm delivery (Delcroix et al., 2023).

(ii) We expect a negative correlation between the variables `tripre0` and `birthweight`. Again, assuming that higher birth weight is indicative of better infant health, it naturally follows that a greater value for `tripre0` (i.e. "1"), in other words no prenatal visits, would imply a lower birth weight and vice versa. This could be due to undetected complications during pregnancy that could have been avoided had there been at least one prenatal visit. According to another study published by the *National Library of Medicine*, analyses indicate that, around the world, the number of prenatal visits is significantly related to birth weight of the infant (Donaldson and Billy, 1984).

(iii) Assuming that lower age is generally associated with better overall health and lower risk for diseases and complications, we would expect a negative correlation between `age` and `birthweight`. A different study published by Wang and colleagues in the *National Library of Medicine*, suggest that the relationship is nonlinear but also that the specific relationship between each additional year of maternal age and birth weight remains unclear (Wang et al., 2020). This result appears counterintuitive to what we expected.

2.

To answer this question, we will compute the empirical correlation between each pair of variables and compare the results to our expectations.

```r
bw_smoking |>
  summarise(cor_smoker = cor(birthweight, smoker),
            cor_tripre0 = cor(birthweight, tripre0),
            cor_age = cor(birthweight, age))
```

```
##   cor_smoker cor_tripre0    cor_age
## 1 -0.1691266  -0.1234999 0.08007321
```

The results match our expectations for the correlation between `birthweight`, and `smoker` and `tripre0`, respectively. The correlation between `birthweight` and `age` is slightly positive.

3 Plots

In this section we shall create two density plots to visualize the distribution of birth weights. One by maternal smoking status and the other by prenatal visit attendance.

```r
bw_sm_plot <- bw_smoking |>
  ggplot() +
  geom_density(aes(x = birthweight, fill = factor(smoker)), alpha = 0.25) +
  scale_fill_manual(
    name = "Smoker",
    values = c("0" = "#619CFF", "1" = "#F8766D"),
    labels = c("0 = Non-Smoker", "1 = Smoker")
  ) +
  labs(
    x = "Birthweight",
    y = "Density",
    title = "Birthweight Density for Smokers vs Non-Smokers"
  )

bw_sm_plot
```

```r
bw_tripre0_plot <- bw_smoking |>
  ggplot() +
  geom_density(aes(x = birthweight, fill = factor(tripre0)), alpha = 0.25) +
  scale_fill_manual(
    name = "Prenatal Visits",
    values = c("0" = "#619CFF", "1" = "#F8766D"),
    labels = c("0 = At least 1 Prenatal Visit", "1 = No Prenatal Visits")
  ) +
```

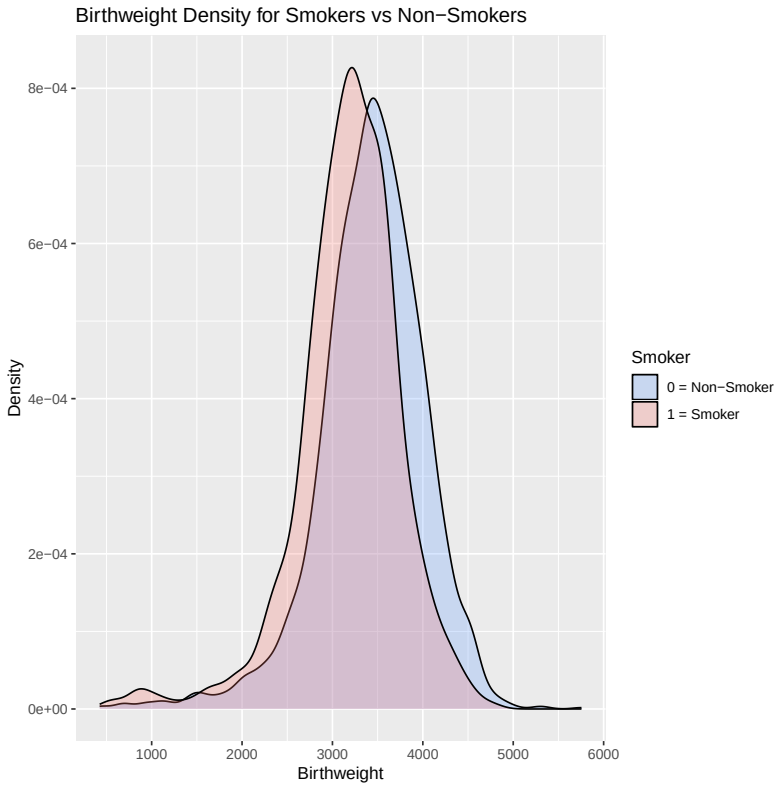

Figure 1: bw density for smokers vs non-smokers

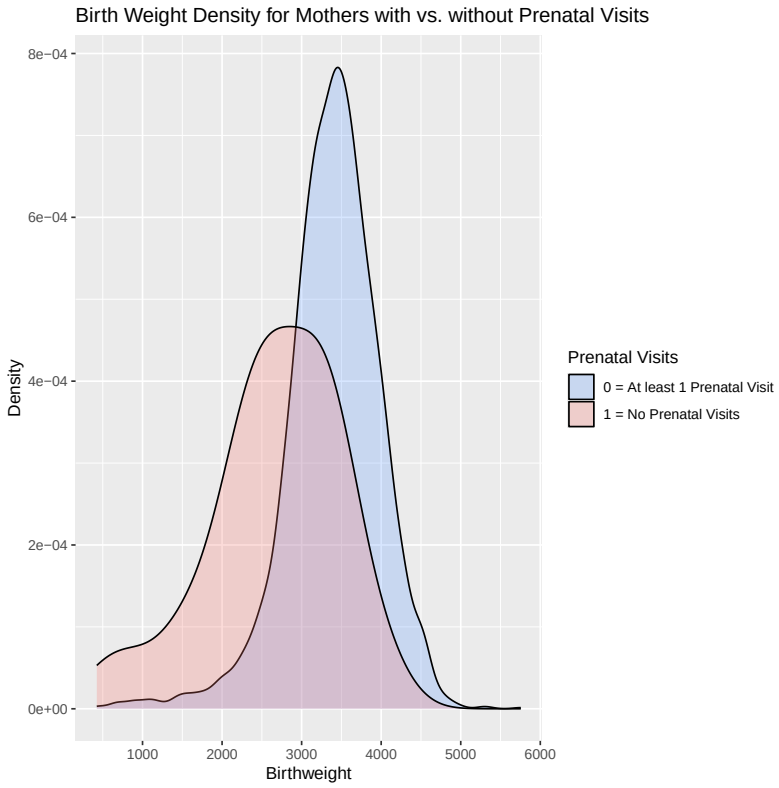

Figure 2: bw density for mothers with vs. without prenatal visits

```
  labs(
    x = "Birthweight",
    y = "Density",
    title = "Birth Weight Density for Mothers with vs. without Prenatal Visits"
  )

bw_tripre0_plot
```

Both figures align with our expectations: The curves for not smoking and attending at least one prenatal visit show a higher mean birth weight than those for smoking mothers or mothers with no prenatal visits, with a dramatically stronger effect for the latter.

4 Multiple Linear Regression

4.1 First Part

By Y we shall denote the variable `birthweight`, and by X_1, X_2, X_3, X_4 and X_5 the variables `age`, `educ`, `drinks`, `smoker`, and `tripre0`, respectively. Now we shall consider the following multiple linear regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + u, \quad E[u \mid X] = 0$$

To create the multiple linear regression model, we use the base function *lm()*. Let `model_full` denote the linear model.

```
model_full <- lm(birthweight ~ age + educ + drinks + smoker + tripre0,
                 data = bw_smoking)

summary(model_full)

##
## Call:
## lm(formula = birthweight ~ age + educ + drinks + smoker + tripre0,
##     data = bw_smoking)
##
## 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
```

Given these coefficients, we shall try and answer the following questions:

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. Consider the estimated coefficient for `age`, which is $\hat{\beta}_1 = 3.634$. This represents the expected change in birthweight for a one-year increase in maternal age, holding all other variables constant. Therefore, a two-year increase in maternal age would increase the expected birthweight by

$$2 \times 3.634 = 7.268 \text{ grams.}$$

2. The estimated coefficient for `smoker` is $\hat{\beta}_4 = -216.465$. This indicates that, ceteris paribus, infants of mothers who smoked during pregnancy are expected to have a birthweight 216.465 grams lower than infants of non-smoking mothers.

4.2 Second Part

In this part we shall estimate the linear model without the variable `smoker` and compare it to our original model including `smoker`. We shall try and answer the following questions:

1. Which of the two models explains more variability in the outcomes? We shall call this model `model+` from now on.

2. In `model+`, do our estimates align with the answers provided in question 1?

3. In `model+`, what is the effect on `birthweight` (in grams) of going to at least one prenatal visit?

Let `model_exsmoker` denote the linear model excluding `smoker`.

```
model_exsmoker <- lm(birthweight ~ age + educ + drinks + tripre0,
                     data = bw_smoking)
summary(model_full)

##
## Call:
## lm(formula = birthweight ~ age + educ + drinks + smoker + tripre0,
##     data = bw_smoking)
##
## 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

summary(model_exsmoker)

##
## Call:
## lm(formula = birthweight ~ age + educ + drinks + tripre0, data = bw_smoking)
##
## 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
```

1. The model including `smoker` explains more variability as it contains a significant influence on `birthweight` (that of smoking vs not-smoking) which is not captured by the other variables. This is confirmed by our obtained R^2 value: `0.0467` for the model

including `smoker` and `0.02696` for the model excluding `smoker`. This result should not be surprising, as `smoker` is a binary variable that directly captures a major behavioral difference affecting pregnancy health.

2. Our estimates do align with our previous answer. See the reported R^2 and the fact that the estimates for the other variables are less significant in `model+` (`model_full`).

3. In `model+`, the effect on `birthweight` of attending at least one prenatal visit is `+654.6` grams.

4.3 Third Part

Based on the results of the estimation of `model+`, we shall 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.

An unbiased estimator of the error variance σ^2 in a homoskedastic multiple regression model is given by:

$$\hat{\sigma}^2 = \frac{SSR}{df}$$

where

$$SSR = \sum_{i=1}^{N} \hat{u}_i^2, \quad df = N - K - 1$$

and N is the number of observations, while K is the number of predictors $X_1, \ldots, X_K$.

```
residuals_full <- residuals(model_full)

SSR <- sum(residuals_full^2)

N <- length(residuals_full)
K <- length(coefficients(model_full))

df <- N - K

err_var_estimate <- SSR / df
err_var_estimate

## [1] 334919.1
```

Or, we can compute the estimated variance error directly.

```r
err_var_estimate_direct <- summary(model_full)$sigma^2
err_var_estimate_direct

## [1] 334919.1
```

Note that R includes the intercept in the coefficient count, hence we do not need to subtract 1 when calculating the degrees of freedom.

Now we shall **(1)** recall the formula for the variance-covariance matrix of the OLS estimator. Using this formula and the estimator of the error variance computed above, we shall **(2)** compute and show an estimator of the variance-covariance matrix of the OLS estimator. Then, we shall (3) determine the value of the covariance of the estimators of `educ` and `tripre0` and interpret the sign of this covariance.

(1) The covariance of two OLS coefficient estimators $\hat{\beta}_j$ and $\hat{\beta}_k$ measures how strongly deviations between the estimator and the true value are correlated and is given by:

$$Cov(\hat{\beta}_j, \hat{\beta}_k) = E[(\hat{\beta}_j - \beta_j)(\hat{\beta}_k - \beta_k)]$$

This information is summarized for all possible pairs of coefficients in the covariance matrix of the OLS estimator:

$$Cov(\hat{\boldsymbol{\beta}}) = E[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T]$$

(2) We can compute the variance-covariance matrix of the OLS estimator manually:

```r
X <- model.matrix(model_full)
err_var_estimate <- summary(model_full)$sigma^2
v_cov_matrix_manual <- err_var_estimate * solve(t(X) %*% X)
v_cov_matrix_manual

##             (Intercept)        age         educ      drinks      smoker
## (Intercept)  5338.67837 -63.754236 -262.9850038  22.4545915 -595.008850
## age           -63.75424   4.865436   -5.2377332  -1.3793463    2.977432
## educ         -262.98500  -5.237733   30.8376588   0.6882034   28.683743
## drinks         22.45459  -1.379346    0.6882034 238.7791968  -39.112391
## smoker       -595.00885   2.977432   28.6837434 -39.1123909  764.626166
## tripre0      -428.17133   3.380294   19.5957463 -58.8971667 -132.831046
##                 tripre0
## (Intercept)  -428.171327
## age             3.380294
## educ           19.595746
```

```
## drinks        -58.897167
## smoker       -132.831046
## tripre0     11356.119084
```

Or directly:

```
v_cov_matrix_direct <- vcov(model_full)
v_cov_matrix_direct

##             (Intercept)        age         educ      drinks      smoker
## (Intercept)  5338.67837 -63.754236 -262.9850038  22.4545915 -595.008850
## age           -63.75424   4.865436   -5.2377332  -1.3793463    2.977432
## educ         -262.98500  -5.237733   30.8376588   0.6882034   28.683743
## drinks         22.45459  -1.379346    0.6882034 238.7791968  -39.112391
## smoker       -595.00885   2.977432   28.6837434 -39.1123909  764.626166
## tripre0      -428.17133   3.380294   19.5957463 -58.8971667 -132.831046
##                 tripre0
## (Intercept)  -428.171327
## age             3.380294
## educ           19.595746
## drinks        -58.897167
## smoker       -132.831046
## tripre0     11356.119084
```

(3) We can determine the covariance of the estimators of `educ` and `tripre0` by looking at the covariance matrix. However, determining the covariance of a particular set of estimators from the covariance matrix can be cumbersome, especially for large matrices. Therefore, we use `R` to determine the covariance directly:

```
v_cov_matrix_direct["educ", "tripre0"]

## [1] 19.59575
```

The covariance between `educ` and `tripre0` is `19.59575`. The positive sign tells us that they are positively correlated. If one variable increases, the other tends to follow, and if one variable decreases, the other tends to decrease as well.

5 Hypothesis Testing and Prediction

To answer the questions in this section, we shall use the results of `model+` (`model_full`).

1. Does the weekly number of drinks during pregnancy have a significant impact on the weight at birth? We shall define our null and alternative hypotheses:

$H_0 : \beta_3 = 0$	The weekly number of drinks has no effect on birth weight
$H_1 : \beta_3 \neq 0$	The weekly number of drinks does affect birth weight

Let the significance level $\alpha = 0.05$

Now we have two equivalent ways of testing the significance of `drinks`. (i) Using the test-statistic and an appropriate critical value. (ii) Using the p-value.

(i) To determine the t-statistic for our estimate for `drinks`, we pull the information needed from our model coefficients. Let `t_drinks` denote the t-statistic for the estimator `drinks`:

```
t_drinks <- summary(model_full)$coefficients["drinks", "t value"]
        abs(t_drinks)

## [1] 0.8297682
```

To find the suitable critical value for a two-sided t-test at significance level $\alpha = 5\%$, we use the t-distribution with the appropriate degrees of freedom and compare it to the absolute value of our t-statistic. Let `t_critical` denote the critical value for our two-sided t-test:

$$t_{\text{critical}} = t_{1-\alpha/2,\,\text{df}}$$

where `df` denotes the degrees of freedom of the regression residuals.

```
alpha <- 0.05
t_critical <- qt(1 - alpha/2, df)
t_critical

## [1] 1.960757
```

Note that the critical t-value is slightly larger than 1.96 because we are using a t-distribution with finite degrees of freedom. Compared to the standard normal distribution, the t-distribution has slightly heavier tails. For the same 5% significance level, the critical value lies a bit further away from zero. As the degrees of freedom increase, the t-distribution approaches the standard normal distribution and the critical value converges to 1.96.

Since |`t_drinks`| is smaller than `t_critical`, we conclude that `drinks` does not have a statistically significant effect on `birthweight` at the 5% significance level.

(ii) By looking at the model coefficients, we do not necessarily need to compare the t-statistic to a critical value. We can determine the significance by looking at the p-value, denoted as `Pr>|t|` in our model Output:, the probability of observing our calculated t-statistic, assuming that the null hypothesis is true (i.e., the combined area under the curve, left and right of our observed t-value). Let `p_drinks` denote the p-value for `drinks`:

```
p_drinks <- summary(model_full)$coefficients["drinks", "Pr(>|t|)"]
p_drinks

## [1] 0.406736
```

The observed p-value of 0.40736 is greater than our significance level of 5%, so we fail to reject the null hypothesis and conclude that the weekly number of drinks during pregnancy does not have a statistically significant impact on birth weight.

2. Do prenatal medical visits have an impact on the weight at birth? We define our null and alternative hypotheses as follows:

$H_0 : \beta_5 = 0$	Attending at least one prenatal visit has no significant effect on birth weight
$H_1 : \beta_5 \neq 0$	Attending at least one prenatal visit does affect the birth weight

Let the significance level $\alpha = 0.05$

Again, we have two equivalent ways of testing the significance of `tripre0`. (i) Using the test-statistic and an appropriate critical value. (ii) Using the p-value.

(i) Let `t_tripre0` denote the t-statistic for the estimator `tripre0`:

```
t_tripre0 <- summary(model_full)$coefficients["tripre0", "t value"]
        abs(t_tripre0)

## [1] 6.142727
```

Since |`t_tripre0`| is larger than our previously calculated `t_critical`, we conclude that `tripre0` does have a statistically significant effect on `birthweight` at the 5% significance level.

(ii) Let `p_tripre0` denote the p-value for `tripre0`:

```
p_tripre0 <- summary(model_full)$coefficients["tripre0", "Pr(>|t|)"]
```

The observed p-value of 9.18×10^{-10} is far smaller than our significance level of 5%, so we reject the null hypothesis and conclude that attending at least one prenatal visit has a statistically significant impact on birth weight.

Now we shall use the estimates for the coefficients and the error variance from section 4 to manually compute the following test statistic for the parameter of the variable `tripre0`.

$$t_5 = \frac{\hat{\beta}_5 - 1}{\text{sd}(\hat{\beta}_5)}$$

where $\hat{\beta}_5$ is the estimated coefficient for `tripre0` and $\text{sd}(\hat{\beta}_5)$ is its standard error.

Let t_5 denote the manually computed test-statistic for `tripre0` using the estimates for the coefficients the error variance:

```
t_5 <- (summary(model_full)$coefficients["tripre0", "Estimate"] - 1) /
  (summary(model_full)$coefficients["tripre0", "Std. Error"])
t_5

## [1] -6.152111
```

The computed test-statistic `t_5` differs from our previously calculated `t_tripre0` insofar as we are not testing whether the true parameter for `tripre0` is equal to 0, but whether or not it is equal to 1. In other words, our hypotheses would look like:

$$H_0 : \beta_5 = 1,$$
$$H_1 : \beta_5 \neq 1.$$

We therefore test whether the true coefficient for `tripre0` differs significantly from 1 rather than from 0.

Since the error variance is estimated from the sample, the test statistic follows a student-t-distribution with residual degrees of freedom $(n - k)$, where n is the sample size and k the number of estimated coefficients (including the intercept).

Now, we shall compute the p-value for a significance level of $\alpha = 0.05$ using said distribution and interpret the results.

Let `p_t_5` denote the p-value for `t_5`, which represents the probability of observing a test statistic at least as extreme as `t_5` under the null hypothesis $H_0 : \beta_5 = 1$.

```
p_t_5 <- 2 * pt(t_5, model_full$df.residual)
```

The observed p-value of approximately 8.66×10^{-10} is far smaller than our significance level of $\alpha = 0.05$, so we reject the null hypothesis $H_0 : \beta_5 = 1$. This indicates that the true coefficient for `tripre0` is significantly different from 1.

If we were testing against the usual null $H_0 : \beta_5 = 0$, the corresponding p-value (from `t_tripre0`) would also be extremely small, so we would similarly conclude that `tripre0` has a statistically significant effect on birth weight. The difference is that here we specifically test whether the effect equals 1 rather than 0.

We can use our estimated model, `model_full`, to predict the expected birth weight for a given set of values of the predictor variables. This is done by creating a new data frame with the chosen values and applying the `predict()` function in `R`.

Example 1

```
newdata_1 <- data.frame(
  age = 40,
  educ = 18,
  drinks = 0,
  smoker = 0,
  tripre0 = 0
)
example_model_1 <- predict(model_full, newdata = newdata_1)
example_model_1
```

```
##       1
## 3550.23
```

Example 2

```
newdata_2 <- data.frame(
  age = 25,
  educ = 0,
  drinks = 6,
  smoker = 1,
  tripre0 = 1
)
example_model_2 <- predict(model_full, newdata = newdata_2)
example_model_2
```

```
##       1
## 2299.03
```

These predictions indicate that, according to our model, the first mother is expected to have a birth weight of approximately 3550 grams, while the second mother is expected to have a lower birth weight of approximately 2299 grams. This shows the combined effects

of age, education, alcohol consumption, smoking, and prenatal visits on birth weight as estimated by our model, `model_full`.

References

Almond, D., Chay, K. Y., and Lee, D. S. (2005). The costs of low birth weight. *The Quarterly Journal of Economics*, 120(3):1031–1083.

Delcroix, M. H., Delcroix-Gomez, C., Marquet, P., Gauthier, T., Thomas, D., and Aubard, Y. (2023). Active or passive maternal smoking increases the risk of low birth weight or preterm delivery: Benefits of cessation and tobacco control policies. *Tobacco Induced Diseases*, 21:72.

Donaldson, P. J. and Billy, J. O. (1984). The impact of prenatal care on birth weight. evidence from an international data set. *Medical Care*, 22(2):177–188.

Wang, S., Yang, L., Shang, L., Yang, W., Qi, C., Huang, L., Xie, G., Wang, R., and Chung, M. C. (2020). Changing trends of birth weight with maternal age: a cross-sectional study in xi'an city of northwestern china. *BMC Pregnancy and Childbirth*, 20(1):744.