Programming Assignment 4

Andrew Shaw

2025-10-17

Problem 1: Using the formula for single variate models (covariances and variances) estimate a model y that's based only on x1.

```
y_bar <- mean(y)
x1_bar <- mean(x1)
cov_1 <- cov(x1, y)
var_1 <- var(x1)
beta1_est <- cov_1 / var_1
alpha <- y_bar - beta1_est*x1_bar
```

$$\hat{\beta}_1 = \frac{cov(x_1, y)}{x_1} = 0.0407551$$

$$\hat{y} = 0.2992 + 0.0408x_1$$

Problem 2: Now estimate the covariance between x1 and x2 and the variance of x1.

```
cov_2 <- cov(x1, x2)
bias_coef <- cov_2 / var_1
```

$$\hat{\beta}_2 = \frac{cov(x_1, x_2)}{x_1} = 0.2524396$$

Problem 3: Based on what you know from the way the data was created what is your best guess for omitted variable bias?

```
OVB <- 0.08* bias_coef
```

$$OVB = 0.08 \times \frac{cov(x_1, x_2)}{x_1} = 0.0201952$$

Problem 4: imagine x1 is a treatment with some drug. Describe why randomizing x1 will give you the correct estimate of the effect of x1 on y?

Because randomizing x_1 will break the correlation that was built into the given model between x_1 and x_2.

Problem 5: run 10 regressions of x1 on y (using the lm(y~x1) function) within groups of x2. I.E. run a set of regressions of x1 on y where x2 equals 0 to 0.1, 0.1 to 0.2, 0.2 to 0.3, etc (don't include the same observations in any regressions. Take the weighted average of the estimates from each of these regressions where you weight by the number of observations in each of these regressions multiplied by the coefficients. What percent of the bias does this estimate have in comparison to your estimate from problem 1?

```
df <- data.frame(y, x1, x2)
betas <- c()
obs <- c()

for (i in 0:9) {
  model <- lm(y ~ x1, data = df, subset = (x2 >= i & x2 < (i + 1)))
  betas <- c(betas, model$coefficients[2])
  obs <- c(obs, length(model$residuals))
}

wtd_avg <- sum(obs * betas) / sum(obs)

bias_binned <- wtd_avg - 0.02
bias_naive <- beta1_est - 0.02

remaining <- abs(bias_binned / bias_naive) * 100
remaining
```

```
## [1] 5.669123
```

Running smaller regressions of y and x_1 lowers the covariance of x_1 and x_2 reducing OVB reducing to 5.669.

Natality Data

```
rm(list = ls())
load("/Users/andrewshaw/Documents/GitHub/6140_econometrics/data/samplebirth2017.RData")
df <- samp


vars <- c("dbwt", "mager", "priorlive", "cig_1", "bmi", "previs", "oegest_comb")
df_small <- df[vars]
df_clean <- df_small |>
  filter(
    priorlive <= 14, cig_1 <= 98, dbwt < 9999,
    bmi != 99.9, previs != 99, oegest_comb != 99
  )
```

Problem 6 Create a multivariate model that best predicts the effects of various characteristics on birth weight which is variable dbwt. Be sure to get rid of the missing variables (see the codebook on canvas). You must use at least five variables (a set of dummy variables counts as one). Show your results in a table in RMarkdown. Describe the results.

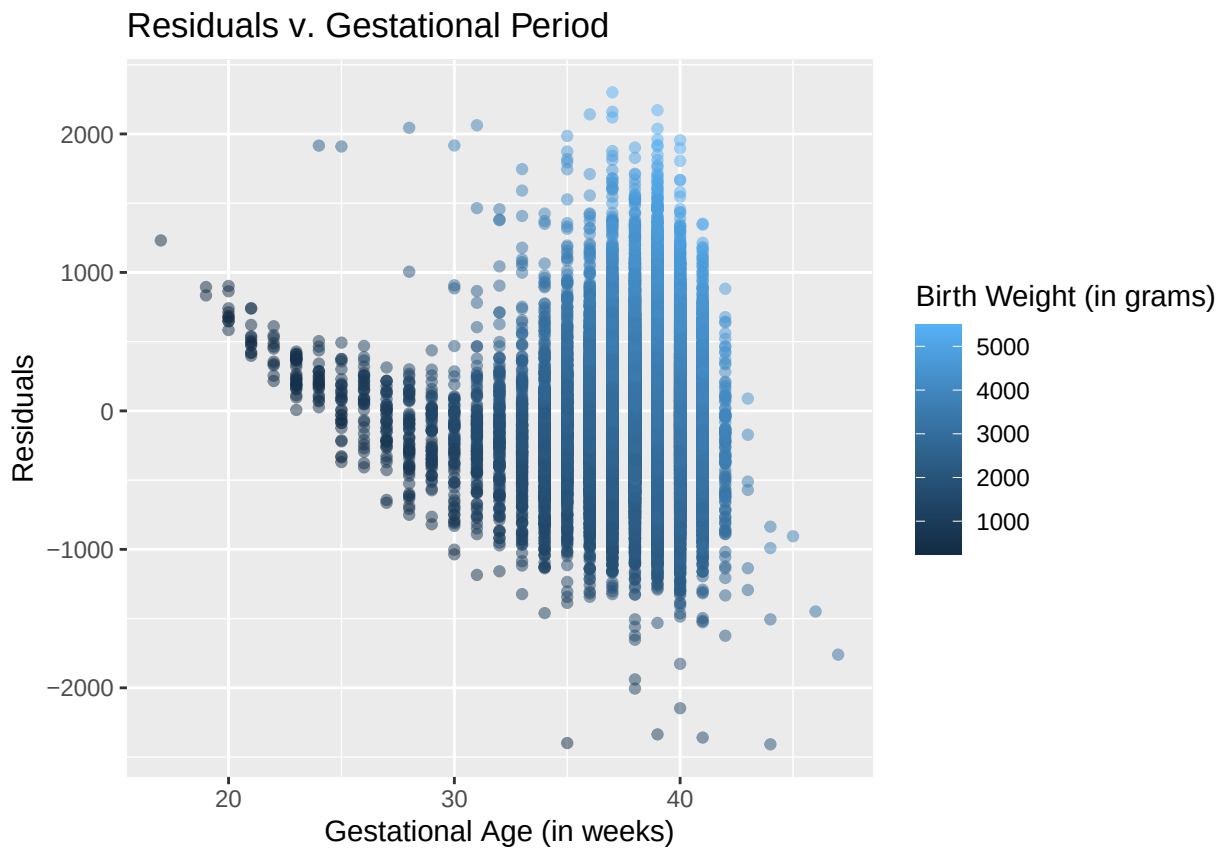
```
model_natality <- lm(dbwt ~ mager + priorlive + cig_1 + bmi + previs + oegest_comb, data = df_clean)
table <- tidy(model_natality)
table$p.value <- formatC(table$p.value, format = "e", digits = 2)
table$term <- c("Intercept", "Mother's Age", "Prior Children (living)", "Cigarettes in 1st Trimester",
tbl_print <- kable(table)
tbl_print
```

term	estimate	std.error	statistic	p.value
Intercept	-4599.649797	45.9655907	-100.067240	0.00e+00
Mother's Age	4.001403	0.4130344	9.687821	3.61e-22
Prior Children (living)	22.291684	1.8667722	11.941298	8.29e-33
Cigarettes in 1st Trimester	-7.517486	0.6559566	-11.460341	2.35e-30
Mother's BMI	8.908787	0.3360109	26.513389	1.91e-153
Prenatal Visits	4.688682	0.5530130	8.478430	2.37e-17
Gestational Period (in weeks)	193.405306	1.1253791	171.857923	0.00e+00

The biggest contributor in this model is gestational period with each additional week adding on average 193.41 grams to the baby's weight. Prenatal visits, mothers more advanced in age, mother's BMI, and prior children are all positively correlated with baby weight but less extreme relative to gestational period. Cigarettes smoked in the first trimester is negatively correlated with the baby weight.

Problem 7: Create a scatter-plot of your residuals on the y-axis against at least one continuous independent variable on the x-axis. Include this in your RMarkdown file along with a description of the results.

```
df_clean$residuals <- residuals(model_natality)
scatter <- ggplot(data = df_clean, aes(x = oegest_comb, y = residuals, color = dbwt)) +
  geom_point(alpha=0.5) +
  labs(
    title = "Residuals v. Gestational Period",
    x = "Gestational Age (in weeks)",
    y = "Residuals",
    color = "Birth Weight (in grams)"
  )
scatter
```



The fanning out of baby weight indicates heteroskedasticity around 40 weeks. The variance of the residuals depends on the gestational age which means that under OLS assumptions homoskedasticity is violated. Using robust standard error measures would account for this. The regression coefficient remains unbiased and consistent.