ECO 5720
Problem Set 2

1. In a study relating college grade point average to time spent in various activities, you distribute a survey to several students. The students are asked how many hours they spend each week in four activities: studying, sleeping, working, and leisure. Any activity is put into one of the four categories, so that for each student, the sum of hours in the four activities must be 168. Can you estimate the following model

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + \beta_4 leisure + u$$

with regressors for all four activities - studying, sleeping, working, and leisure?

Answer: No, since *study*, *sleep*, *work*, and *leisure* are perfectly collinear. Here, *study* + *sleep* + *work* + *leisure* = 168.

2. Which of the following can cause the OLS estimator, i.e., $\hat{\beta}$, to be biased?
(i) Heteroskedasticity.
(ii) Omitting an important variable.
(iii) A sample correlation coefficient of .95 between two independent variables both included in the model.

Answer: An omitted variable that affects the dependent variable and is correlated with the included explanatory variables can cause bias. The homoskedasticity assumption, plays no role in unbiasedness of the OLS estimators. Further, high (but not perfect) collinearity between the explanatory variables in the sample does not affect the assumptions needed for unbiasedness. However, it does inflate the corresponding standard errors. Thus, our answer is only (ii).

3. Use the data in HPRICE1 to estimate the model

$$price = \beta_0 + \beta_1 sqrft + \beta_2 bdrms + u,$$

where *price* is the price of a house in thousands of dollars; *sqrft* represents the size of a house in square feet; *bdrms* denotes the number of bedrooms.

(i) Write out the results in an equation form. While it is sufficient to report the $\hat{\beta}$ estimates, you may also paste the results.

Answer: Estimated equation:

$$\widehat{price} = -19.32 + 0.128 sqrft + 15.198 bdrms.$$

```
===============================================
                          Dependent variable:
                      ---------------------------
                                 price
-----------------------------------------------
sqrft                          0.128***
                               (0.014)

bdrms                          15.198
                               (9.484)

Constant                      -19.315
                              (31.047)

-----------------------------------------------
Observations                     88
```

```
R2                              0.632
Adjusted R2                     0.623
Residual Std. Error       63.045 (df = 85)
F Statistic            72.964*** (df = 2; 85)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

(ii) What is the estimated increase in price for a house with one more bedroom, holding square footage constant?

Answer: Holding square footage constant, price increases by 15.198, i.e., $15,198.

(iii) What percentage of the variation in price is explained by square footage and number of bedrooms?

Answer: About 63.2%.

(iv) The first house in the sample has *sqrft* = 2,438 and *bdrms* = 4. Find the predicted selling price for this house from the OLS regression line.

Answer: The predicted price is –19.315 + 0.128(2,438) + 15.198(4) = 353.541, or $353,541.

(v) The actual selling price of the first house in the sample was $300,000 (so *price* = 300). Find the residual for this house.

Answer: From part (iv), the estimated value of the home based only on square footage and number of bedrooms is $353,541. The actual selling price was $300,000 and the residual is -53541.

4. Use the data in NBASAL to estimate the model

$$wage = \beta_0 + \beta_1 points + \beta_2 rebounds + \beta_3 assists + u.$$

Here, *wage* denotes annual salary in thousands of dollars; *points*, *rebounds*, and *assists* represent points, rebounds, and assists per game, respectively.

(i) What is the estimated value of β_3?

Answer: $\hat{\beta}_3 = 24.35$.

```
===============================================
                        Dependent variable:
                    ---------------------------
                               wage
-----------------------------------------------
points                       81.194***
                             (11.569)

rebounds                     92.236***
                             (19.911)

assists                       24.347
                             (26.987)

Constant                      130.215
                             (96.502)
```

```
-----------------------------------------------
Observations                    269
R2                             0.475
Adjusted R2                    0.470
Residual Std. Error     728.172 (df = 265)
F Statistic           80.070*** (df = 3; 265)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

(ii) Next, estimate the model

$$assists = \delta_0 + \delta_1 points + \delta_2 rebounds + v,$$

and save the residuals, $\hat{v}$.

Answer:

```
===============================================
                    Dependent variable:
                ---------------------------
                          assists
-----------------------------------------------
points                  0.264***
                         (0.021)

rebounds                -0.262***
                         (0.042)

Constant                0.871***
                         (0.213)

-----------------------------------------------
Observations                    269
R2                             0.380
Adjusted R2                    0.375
Residual Std. Error      1.654 (df = 266)
F Statistic           81.475*** (df = 2; 266)
===============================================
Note:               *p<0.1; **p<0.05; ***p<0.01
```

Finally, estimate the model

$$wage = \alpha_0 + \alpha_1 \hat{v} + \varepsilon.$$

What is the estimated value of α_1? How does it compare to the value of β_3 estimated in (i)?

Answer: Here, $\hat{\alpha}_1 = \hat{\beta}_3 = 24.35$.

```
===============================================
                    Dependent variable:
                ---------------------------
                           wage
-----------------------------------------------
uhat                      24.347
                        (37.093)

Constant               1,423.828***
                        (61.022)
```

```
-----------------------------------------------
Observations                    269
R2                             0.002
Adjusted R2                   -0.002
Residual Std. Error    1,000.837 (df = 267)
F Statistic             0.431 (df = 1; 267)
===============================================
Note:                *p<0.1; **p<0.05; ***p<0.01
```

5. Simulate a data set from the following model

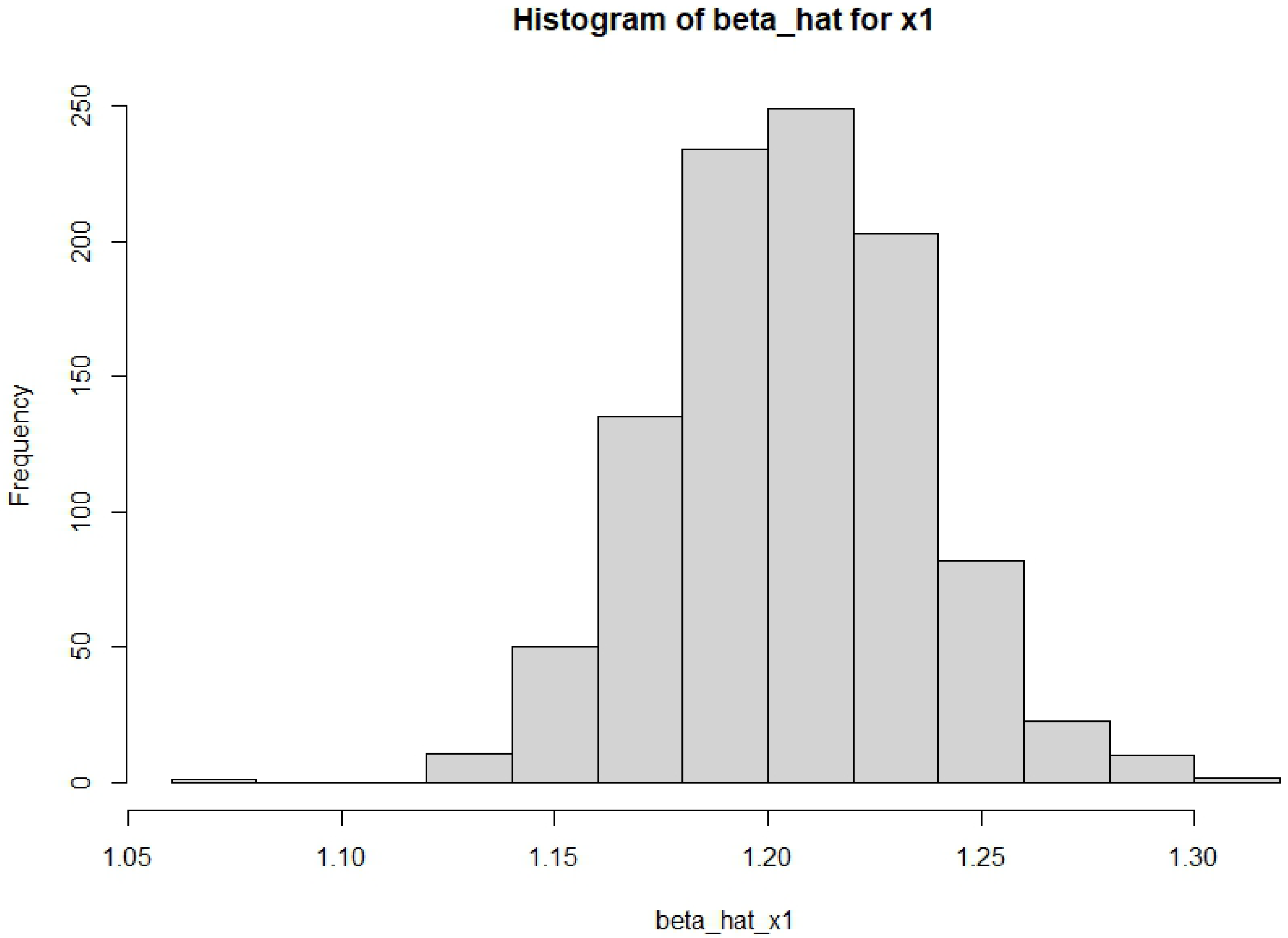
$$y = 1 + 0.7x_1 + 0.5x_2 + u$$
$$x_1 \sim N(0,1)$$
$$x_2 \sim N(0,1)$$
$$u \sim N(0,1)$$
$$\text{Corr}(x_1, u) = 0.4$$
$$\text{Corr}(x_2, u) = 0.2$$
$$\text{Corr}(x_1, x_2) = -0.3.$$

Estimate the model by OLS.

(i) For 1,000 repetitions and 900 observations in each repetition, graph the empirical distribution of $\widehat{\beta_1}$. Please attach your R script and graph.

Answer:

```
###############################################################################
### Simulation - 1000 reps; n=900; corr(x1,u) = 0.4; corr(x2,u) = 0.2; corr (x1,x2) = -0.3 ###
###############################################################################

# Set seed for reproducibility
set.seed(123)

# Number of simulations
n_sim <- 1000

# Store beta hat estimates from each simulation
data_bx1 <- numeric(n_sim)
data_bx2 <- numeric(n_sim)

# Correlation/Covariance matrix: x1, x2, u
C <- matrix(c(1, -0.3, 0.4,
              -0.3, 1, 0.2,
              0.4, 0.2, 1), nrow=3, ncol=3, byrow = TRUE)

# Simulation loop
for (i in 1:n_sim) {

  # Draw normal variables
  draws <- MASS::mvrnorm(n = 900, mu = c(0, 0, 0), Sigma = C)

  x1 <- draws[, 1]
  x2 <- draws[, 2]
  u  <- draws[, 3]

  # Generate outcome variable
  y <- 1 + 0.7*x1 + 0.5*x2 + u

  # Regression
  model <- lm(y ~ x1 + x2)

  # Store coefficient estimates
  data_bx1[i] <- coef(model)["x1"]
  data_bx2[i] <- coef(model)["x2"]
}

# Histograms
hist(data_bx1, main = "Histogram of beta_hat for x1", xlab = "beta_hat_x1")
```



(ii) Is the OLS estimator of $\widehat{\beta_1}$ unbiased? Explain briefly.

Answer: No, since the explanatory variables and u are correlated. The distribution of $\widehat{\beta_1}$ is centered around 1.2.

6. Please answer the following questions from the article entitled "The Credibility Revolution in Empirical

Economics: How Better Research Design is Taking the Con out of Econometrics" by Joshua D. Angrist and Jörn-Steffen Pischke.

(i) On page 4, what is the natural experiment used to study the effect of immigration?

Answer: The natural experiment discussed is the Mariel boatlift from Cuba to Florida. Several people from Cuba emigrated to Miami and increased the latter's labor force by about 7 percent in three months.

(ii) On page 7, what does extreme bounds analysis mean?

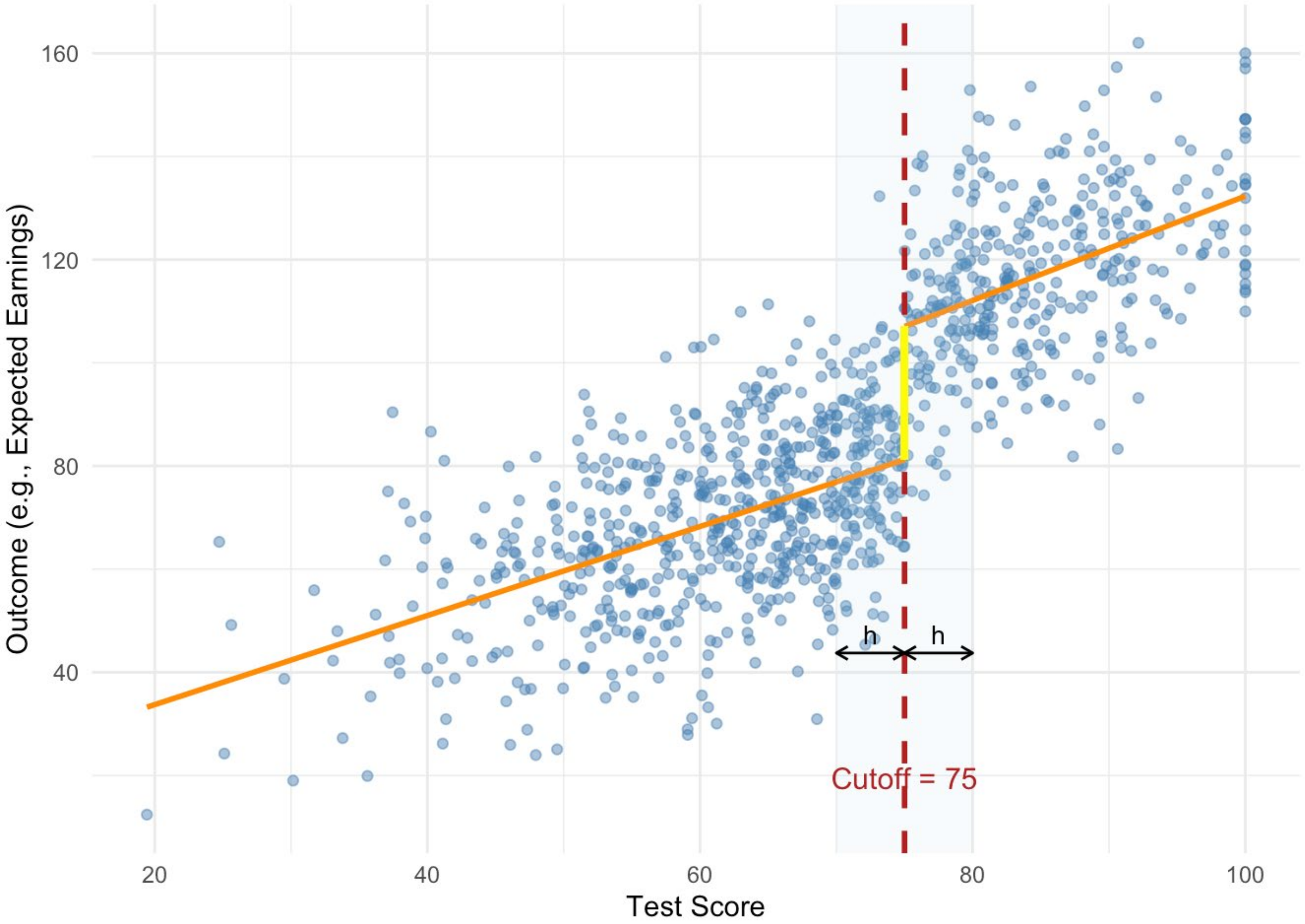
Answer: Extreme bounds analysis is described as an *ad hoc* but intuitive approach. It involves estimating regressions with various combinations of explanatory variables. This yields a range of estimates for the parameter of interest.

(iii) Towards the beginning of page 9, what do the regressions related to education production functions suffer from?

Answer: They suffer from omitted variables bias or reverse causality (i.e., simultaneity).

(iv) On page 14, what is the key assumption for regression discontinuity?

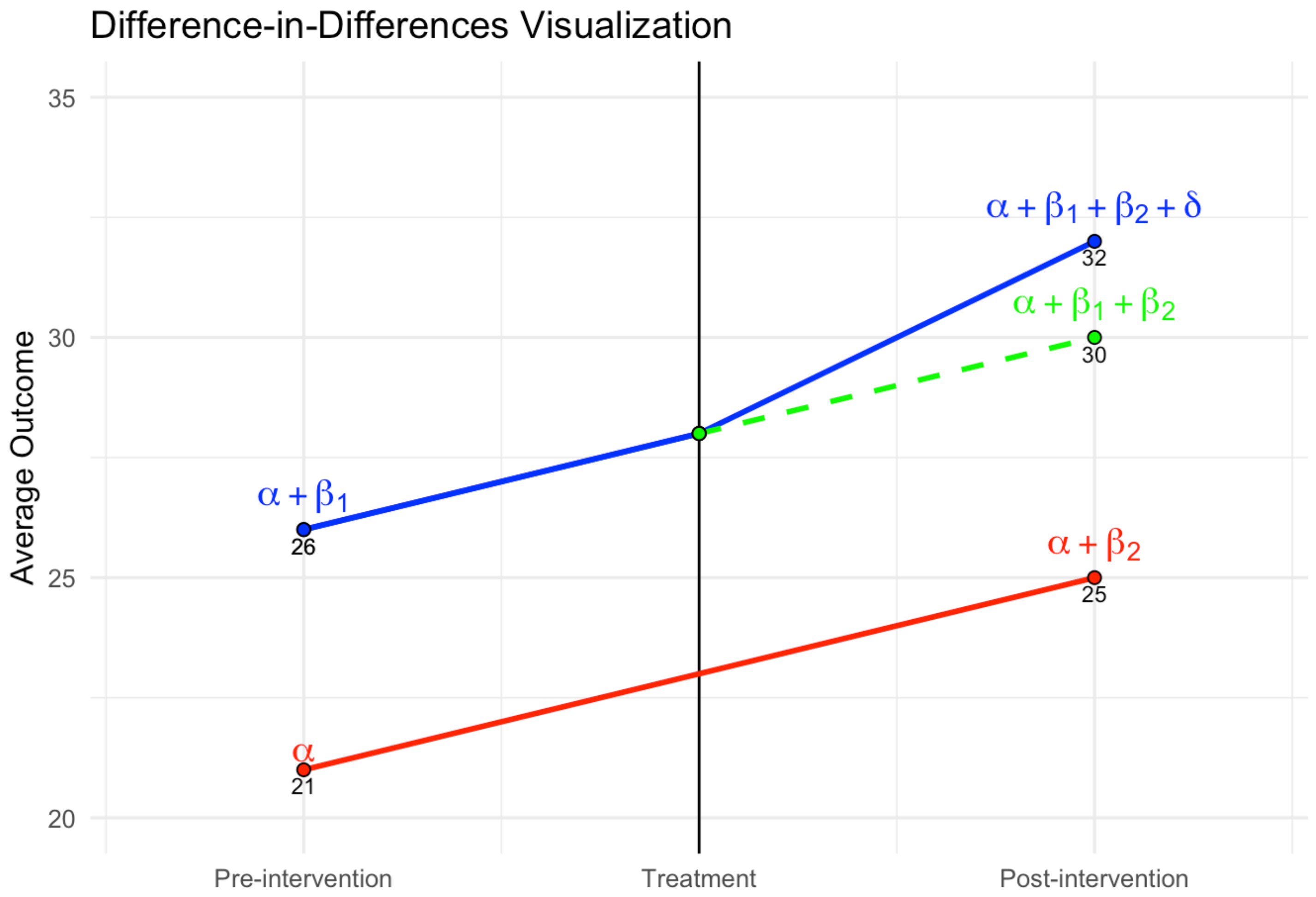
Answer: Regression discontinuity (RD) relies on a running variable with a cutoff value that determines whether observations belong to the treatment or control group. Observations around the cutoff are then analyzed to determine the treatment's effect.



Source: “Causal Inference and Machine Learning: In Economics, Social, and Health Sciences” by Mutlu Yuksel and Yigit Aydede.

(v) On page 14, what is the key assumption for difference-in-differences?

Answer: The difference-in-differences (DID) method compares the evolution of outcomes in groups affected more or less by a policy change. Typically, a treated and a control group’s outcomes are assumed to trend similarly in the absence of treatment. Any deviation in the treated group’s outcome is attributed to the treatment.



Source: “Causal Inference and Machine Learning: In Economics, Social, and Health Sciences” by Mutlu Yuksel and Yigit Aydede.

(vi) On page 17, how many growth regressions does Sala-i-Martin run?

Answer: Sala-i-Martin reported two million regressions.

(vii) On page 24, what leads to weird papers?

Answer: People thinking more about methodologies that research questions can lead to questionable papers.