# ECMA 31360, PSet 1: Solutions

YOUR NAMES, University of Chicago

XX-XX-2025

## ([ ] out of 40p) PART I: Review of OLS for Prediction and Description

### ([ ] out of 22p) Q1: Properties of the OLS Estimator when the CEF is linear-in-parameters

We assume the CEF is linear:

$$E[Y_i \mid X_i] = \beta_0 + \beta_1 X_i.$$

Define the error

$$\varepsilon_i := Y_i - (\beta_0 + \beta_1 X_i), \qquad \Rightarrow \qquad E[\varepsilon_i \mid X_i] = 0.$$

#### (i) Closed-form OLS solution

OLS minimizes $S(b_0, b_1) = \sum_{i=1}^{n}(Y_i - b_0 - b_1 X_i)^2$. FOCs:

$$\frac{\partial S}{\partial b_0} = -2\sum(Y_i - b_0 - b_1 X_i) = 0, \qquad \frac{\partial S}{\partial b_1} = -2\sum X_i(Y_i - b_0 - b_1 X_i) = 0.$$

From the first FOC, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. Substituting into the second yields

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

#### (ii) Unbiasedness

Using $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, we can rewrite

$$\hat{\beta}_1 - \beta_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})\varepsilon_i}{\sum_{i=1}^{n}(X_i - \bar{X})^2}.$$

Condition on $X_1, \ldots, X_n$. The denominator is a function of $X$'s only. For the numerator,

$$E\Big[\sum(X_i - \bar{X})\varepsilon_i \mid X_1, \ldots, X_n\Big] = \sum(X_i - \bar{X})E[\varepsilon_i \mid X_1, \ldots, X_n] = \sum(X_i - \bar{X})E[\varepsilon_i \mid X_i] = 0.$$

Hence $E[\hat{\beta}_1 \mid X_1, \ldots, X_n] = \beta_1$, implying $E[\hat{\beta}_1] = \beta_1$.

For the intercept, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$. Taking expectations:

$$E[\hat{\beta}_0] = E[\bar{Y}] - E[\hat{\beta}_1]E[\bar{X}] = E[E[Y \mid X]] - \beta_1 E[X] = E[\beta_0 + \beta_1 X] - \beta_1 E[X] = \beta_0.$$

Therefore both $\hat{\beta}_1$ and $\hat{\beta}_0$ are unbiased.

**(iii) Consistency**

Write

$$\hat{\beta}_1 - \beta_1 = \frac{\frac{1}{n}\sum(X_i - \bar{X})\varepsilon_i}{\frac{1}{n}\sum(X_i - \bar{X})^2}.$$

Assume i.i.d. sampling with finite second moments and $Var(X) > 0$. By LLN,

$$\frac{1}{n}\sum(X_i - \bar{X})^2 \xrightarrow{p} Var(X) > 0.$$

Also, since $E[\varepsilon \mid X] = 0$ implies $E[(X - E[X])\varepsilon] = 0$, LLN gives

$$\frac{1}{n}\sum(X_i - \bar{X})\varepsilon_i \xrightarrow{p} 0.$$

By Slutsky / Continuous Mapping Theorem for ratios, $\hat{\beta}_1 \xrightarrow{p} \beta_1$.

Finally, $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X}$, and by LLN $\bar{Y} \xrightarrow{p} E[Y]$ and $\bar{X} \xrightarrow{p} E[X]$. Combining with $\hat{\beta}_1 \xrightarrow{p} \beta_1$ and Slutsky yields $\hat{\beta}_0 \xrightarrow{p} \beta_0$.

---

## ([ ] out of 2p) Q2: CEF if linear-in-parameters when eVar is binary 0/1

Suppose $X \in \{0, 1\}$. Let

$$\beta_0 := E[Y \mid X = 0], \qquad \beta_1 := E[Y \mid X = 1] - E[Y \mid X = 0].$$

Then for $X = 0$,

$$\beta_0 + \beta_1 X = \beta_0 = E[Y \mid X = 0],$$

and for $X = 1$,

$$\beta_0 + \beta_1 X = \beta_0 + \beta_1 = E[Y \mid X = 0] + \big(E[Y \mid X = 1] - E[Y \mid X = 0]\big) = E[Y \mid X = 1].$$

Therefore, for all $X \in \{0, 1\}$,

$$E[Y \mid X] = \beta_0 + \beta_1 X,$$

so the CEF is linear-in-parameters when the explanatory variable is binary.

---

## ([ ] out of 4p) Q3: First Response to Manager

Hi Alyson,

Thanks for sharing the regression results — they are definitely useful for summarizing how Prime and non-Prime customers differ on average.

That said, it is important to note that this regression is describing the difference in average spending between customers who already have Prime and those who do not. In other words, the coefficient captures a descriptive difference in conditional means, rather than the causal effect of enrolling in Prime for a given customer.

If our goal is to understand the causal impact of Prime enrollment — i.e., how a customer's spending would change if they were to sign up — we would need additional assumptions or a different research design beyond this simple regression.

Best,
Ty

---

## ([ ] out of 4p) Q4: Second Response to Manager

The key issue is that customers who choose to enroll in Prime may differ systematically from those who do not. For example, Prime members may already be more active shoppers or have higher baseline demand, even in the absence of Prime benefits.

As a result, the regression coefficient reflects both the effect of Prime enrollment and these pre-existing differences between the two groups. Without accounting for this selection, the estimated difference in average spending cannot be interpreted as the causal effect of Prime.

---

## ([ ] out of 2p) Q5: Properties of a linear-in-parameter CEF

**If $E[Y \mid X] = \beta_0 + \beta_1 X$, then $Y = \beta_0 + \beta_1 X + \varepsilon$ with $E[\varepsilon \mid X = x] = E[\varepsilon] = 0$**

Assume the CEF is linear-in-parameters:

$$E[Y \mid X] = \beta_0 + \beta_1 X.$$

Define the residual (error term)

$$\varepsilon := Y - (\beta_0 + \beta_1 X).$$

Then, for any $x$ in the support $\mathcal{X}$,

$$E[\varepsilon \mid X = x] = E[Y - (\beta_0 + \beta_1 X) \mid X = x] = E[Y \mid X = x] - (\beta_0 + \beta_1 x) = (\beta_0 + \beta_1 x) - (\beta_0 + \beta_1 x) = 0.$$

Moreover, by the law of iterated expectations,

$$E[\varepsilon] = E\big(E[\varepsilon \mid X]\big) = E[0] = 0.$$

Hence $Y = \beta_0 + \beta_1 X + \varepsilon$ with $E[\varepsilon \mid X = x] = E[\varepsilon] = 0$ for all $x \in \mathcal{X}$.

**If $Y = \beta_0 + \beta_1 X + \varepsilon$ with $E[\varepsilon \mid X = x] = 0$, then $E[Y \mid X] = \beta_0 + \beta_1 X$**

Assume

$$Y = \beta_0 + \beta_1 X + \varepsilon \quad \text{and} \quad E[\varepsilon \mid X = x] = 0 \;\; \forall x \in \mathcal{X}.$$

Taking conditional expectations given $X$,

$$E[Y \mid X] = E[\beta_0 + \beta_1 X + \varepsilon \mid X] = \beta_0 + \beta_1 X + E[\varepsilon \mid X] = \beta_0 + \beta_1 X.$$

Therefore the CEF is linear-in-parameters.

This establishes the equivalence claimed in Claim 3.

---

## ([ ] out of 6p) Q6: Response to Product Manager

## Q6: Response to Product Manager

### Business-facing numerical example

Consider two types of customers:

- **High-demand customers**: they spend \$100 on average, regardless of whether they have Prime.
- **Low-demand customers**: they spend \$20 on average, regardless of whether they have Prime.

Assume Prime itself has **no causal effect** on spending. However, high-demand customers are much more likely to enroll in Prime.

Suppose the customer base looks like this: - Among Prime users: 80% are high-demand and 20% are low-demand. - Among non-Prime users: 20% are high-demand and 80% are low-demand.

Then average spending is: - Prime users: $0.8 \times 100 + 0.2 \times 20 = 84$ - Non-Prime users: $0.2 \times 100 + 0.8 \times 20 = 36$

A simple regression of spending on a Prime indicator would estimate a difference of $84 - 36 = 48$, suggesting a large positive "Prime effect."
However, in this example Prime has **zero causal impact** on spending for any customer. The entire difference is driven by the fact that customers who choose Prime already have higher baseline demand.

**Technical interpretation**

The regression coefficient identifies the difference in conditional means,

$$E[Y \mid D = 1] - E[Y \mid D = 0],$$

which combines any causal effect of Prime with pre-existing differences between customers who select into Prime and those who do not. Without additional assumptions or an experimental design, this descriptive difference cannot be interpreted as a causal effect.

---

# ([ ] out of 46p) PART II: Review of OLS for Causal Analysis

## ([ ] out of 15p) Q7: Homogeneous Causal Effects

### ([ ] out of 1p) Q7.a: Determinants of Expenditure

---

### ([ ] out of 1p) Q7.b: Interpretation of $\rho$ as Causal Impact

---

### ([ ] out of 1p) Q7.c: Normalization vs Assumption

---

### ([ ] out of 1p) Q7.d: Assumptions Necessary to Run the OLS Algorithm

---

### ([ ] out of 2p) Q7.e: Plan-English Description of $\hat{\rho}$

---

### ([ ] out of 1p) Q7.f: Statistical Properties of the OLS Estimator (first attempt)

---

### ([ ] out of 1p) Q7.g: ZCMA in Plain English

---

**([ ] out of 1p) Q7.h: Identification of $\rho$ under ZCMA**

---

**([ ] out of 1p) Q7.i: Identification of $\rho$ (after weakening ZCMA)**

---

**([ ] out of 1p) Q7.j: Statistical Properties of the OLS Estimator under ZCMA**

---

**([ ] out of 1p) Q7.k: Full Independence of Observed and Unobserved Determinants of the Outcome Variable**

---

**([ ] out of 2p) Q7.l: Do you Expect $\hat{\rho}$ to be unbiased/consistent in the Walmart application?**

---

## ([ ] out of 13p) Q8: Walmart scientists' RCT

**([ ] out of 2p) Q8.a**

---

**([ ] out of 9p) Q8.b**

---

**([ ] out of 2p) Q8.c**

---

## ([ ] out of 14p) Q9: Heterogeneous Treatment Effects

**([ ] out of 2p) Q9.a: Interpretation of $\rho_i$**

---

**([ ] out of 5p) Q9.b: Plain-English description of $\rho_0$, $\rho_1$, and $\rho$. Interpret $\rho_1 = \rho_0$.**

---

**([ ] out of 1p) Q9.c: Model Reformulation**

---

**([ ] out of 4p) Q9.d: Substantive Implications**

---

#### ([ ] out of 2p) Q9.e: Learning about the Average Causal Effect of Treatment on the Entire Customer Population

---

### ([ ] out of 4p) Q10: Take Stock / Learnings

---

# ([ ] out of 10p bonus) PART III: Description vs Causality - the case of Uber One

### ([ ] out of 10p bonus) Q11

### ([ ] out of 5p) Q11.a

A question that produces the answer is "Among Uber One members, what is the average dollar amount of savings per month computed from the pricing rules/discounts applied to their eligible transactions?" This is descriptive because it summarizes observed savings for members (a fact about outcomes among members), not what would happen to the same riders if they did not join.

### ([ ] out of 3p) Q11.b

1. For each member $i$ and month $t$, collect all eligible Uber rides and Uber Eats orders in that month.
2. For each transaction, compute "savings" as non-member price for that transaction less the member price actually paid, accounting for the program's discount and fee rules.
3. Sum savings across transactions for that member-month to get monthly savings ($S_{it}$).
4. Average $S_{it}$ across member-months in the reporting window to get the "$28 on average every month."

### ([ ] out of 2p) Q11.c

A rider typically wants: "If I join Uber One, how much will my monthly out-of-pocket spending change relative to not joining, given my expected usage?" This is causal because it compares outcomes for the same person under their two alternatives.

# ([ ] out of 14p) PART IV: A Look at the Data from the NSW Experiment

### ([ ] out of 14p) Q12: Describe the NSW Data

#### ([ ] out of 1p) Q12.a

The treatment is the offer/assignment to participate in the NSW subsidized employment cum training program (treatment assignment indicator `treat`).

#### ([ ] out of 1p) Q12.b

**Script and Output**

```
# Load
treated <- read.csv("nswre74_treated.csv")
control <- read.csv("nswre74_control.csv")

# Combine (stack rows)
df <- rbind(treated, control)
```

Output:

```
> summary(df)
     treat             age             edu            black             hisp        
 Min.   :0.0000   Min.   :17.00   Min.   : 3.0   Min.   :0.0000   Min.   :0.00000  
 1st Qu.:0.0000   1st Qu.:20.00   1st Qu.: 9.0   1st Qu.:1.0000   1st Qu.:0.00000  
 Median :0.0000   Median :24.00   Median :10.0   Median :1.0000   Median :0.00000  
 Mean   :0.4157   Mean   :25.37   Mean   :10.2   Mean   :0.8337   Mean   :0.08764  
 3rd Qu.:1.0000   3rd Qu.:28.00   3rd Qu.:11.0   3rd Qu.:1.0000   3rd Qu.:0.00000  
 Max.   :1.0000   Max.   :55.00   Max.   :16.0   Max.   :1.0000   Max.   :1.00000  
    married          nodegree          re74               re75            re78      
 Min.   :0.0000   Min.   :0.000   Min.   :    0.0   Min.   :    0   Min.   :    0  
 1st Qu.:0.0000   1st Qu.:1.000   1st Qu.:    0.0   1st Qu.:    0   1st Qu.:    0  
 Median :0.0000   Median :1.000   Median :    0.0   Median :    0   Median : 3702  
 Mean   :0.1685   Mean   :0.782   Mean   : 2102.3   Mean   : 1377   Mean   : 5301  
 3rd Qu.:0.0000   3rd Qu.:1.000   3rd Qu.:  824.4   3rd Qu.: 1221   3rd Qu.: 8125  
 Max.   :1.0000   Max.   :1.000   Max.   :39570.7   Max.   :25142   Max.   :60308  
      u74              u75        
 Min.   :0.0000   Min.   :0.0000  
 1st Qu.:0.0000   1st Qu.:0.0000  
 Median :1.0000   Median :1.0000  
 Mean   :0.7326   Mean   :0.6494  
 3rd Qu.:1.0000   3rd Qu.:1.0000  
 Max.   :1.0000   Max.   :1.0000  
```

**([ ] out of 1p) Q12.c**

**Script and Output**

```
> dplyr::tally(dplyr::group_by(df, treat))
# A tibble: 2 × 2
  treat     n
  <int> <int>
1     0   260
2     1   185
```

**([ ] out of 4p) Q12.d**

**Script and Output**

```
vars <- c("age","edu","nodegree","black","hisp","married","u74","u75","re74","re75","re78","treat")

df %>%
  select(all_of(vars)) %>%
  group_by(treat) %>%
  summarise_all(list(mean))

# A tibble: 2 × 12
  treat   age   edu nodegree black   hisp married   u74   u75  re74  re75  re78
  <int> <dbl> <dbl>    <dbl> <dbl>  <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1     0  25.1  10.1    0.835 0.827 0.108    0.154 0.75  0.685 2107. 1267. 4555.
2     1  25.8  10.3    0.708 0.843 0.0595   0.189 0.708 0.6   2096. 1532. 6349.
```

### ([ ] out of 5p) Q12.e

#### Script and Output

```
results <- lapply(opvs, function(dep){
  formula <- stats::formula(paste(dep, "~ treat"))
  lm_model <- stats::lm(formula = formula, data = df)
  coefs <- summary(lm_model)$coefficients
  data.frame(
    var = dep,
    est = coefs["treat","Estimate"],
    t   = coefs["treat","t value"],
    p   = coefs["treat","Pr(>|t|)"],
    row.names = NULL
  )
})

results_df <- do.call(rbind, results)
results_df$reject_5pct <- results_df$p < 0.05
> results_df
        var          est           t           p reject_5pct
1       age   0.76237006  1.11661493 0.264764269       FALSE
2       edu   0.25748441  1.49582552 0.135411167       FALSE
3  nodegree  -0.12650728 -3.21531660 0.001398352        TRUE
4     black   0.01632017  0.45477598 0.649493182       FALSE
5      hisp  -0.04823285 -1.77566909 0.076473893       FALSE
6   married   0.03534304  0.98043187 0.327408105       FALSE
7       u74  -0.04189189 -0.98286779 0.326208987       FALSE
8       u75  -0.08461538 -1.84662032 0.065468962       FALSE
9      re74 -11.45295788 -0.02217511 0.982318253       FALSE
10     re75 265.14629853  0.87462144 0.382253831       FALSE
```

#### Commentary

At the 5% level, only `nodegree` shows evidence of imbalance between treated and control groups. The remaining 9 predetermined covariates do not show statistically significant differences in means at 5%, though `u75` and `hisp` are closer to conventional thresholds (p-values around 0.06–0.08). Overall, the RA appears broadly consistent with balance in observed covariates, with a notable exception for `nodegree`.

### ([ ] out of 2p) Q12.f

This dataset comes from a job-training experiment in the mid-1970s where eligible men were randomly assigned either to be offered a subsidized employment and training program (185 people) or not offered it (260 people). We observe their background characteristics measured before assignment (such as age, education, race/ethnicity, marital status, prior unemployment, and prior earnings) and their earnings in 1978, about a year after the program.