

ECMA 31360, PSet 1: Solutions

Winston Su, Cagan Gurdal, and Bailey Meche, University of Chicago

01-16-2025

([] 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:

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

Define the error

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

(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:

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

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

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

(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, \dots, X_n . The denominator is a function of X 's only. For the numerator,

$$E\left[\sum (X_i - \bar{X})\varepsilon_i | X_1, \dots, X_n\right] = \sum (X_i - \bar{X})E[\varepsilon_i | X_1, \dots, X_n] = \sum (X_i - \bar{X})E[\varepsilon_i | X_i] = 0.$$

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

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

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

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 $E[\varepsilon | X] = 0$ 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

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

Then for $X = 0$,

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

and for $X = 1$,

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

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

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

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, we would need additional assumptions or a different research design beyond this simple regression.

Best,
Ty

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

Response to Alyson:

Nashant's \$52 is the difference in observed average spend between Prime and non-Prime customers:

$$\bar{Y}_{\text{Prime}} - \bar{Y}_{\text{Non-Prime}}.$$

That statistic is correct as a descriptive fact about group averages, but it does not establish that membership causes higher spend. Customers who choose Prime may differ systematically from non-Prime customers in income, shopping needs, and preferences; those differences can generate a \$52 gap even if the true causal effect of Prime were zero. If we communicate externally, we should state it as "Prime members spend \$52 more on average" (descriptive), not "Prime increases spend by \$52" (causal).

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

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

Assume the CEF is linear-in-parameters:

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

Define the residual (error term)

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

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

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

Moreover, by the law of iterated expectations,

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

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

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

Assume

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

Taking conditional expectations given X ,

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

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,

$$E[Y | D = 1] - E[Y | D = 0],$$

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

Household mean age and having a car can be determinants of spend that may be part of u_i

([] out of 1p) Q7.b: Interpretation of ρ as Causal Impact

If customer i is not a member ($D_i = 0$) :

$$y_i(0) = \alpha + \rho \cdot 0 + u_i = \alpha + u_i$$

If customer i is a member ($D_i = 1$) :

$$y_i(1) = \alpha + \rho \cdot 1 + u_i = \alpha + \rho + u_i$$

Holding u_i fixed,

$$y_i(1) - y_i(0) = (\alpha + \rho + u_i) - (\alpha + u_i) = \rho$$

Because ρ is the same for all customers and $y_i(1) - y_i(0) = \rho$. ρ is the causal impact of Sam's Club Plus membership on a customer's spend. —

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

It is a normalization. Assume,

$$E[u_i] = \mu \neq 0$$

Then

$$E[u_i^*] = E[u_i - \mu] = E[u_i] - E[\mu] = E[u_i] - \mu = \mu - \mu = 0$$

Rewriting the model

$$y_i = \alpha + \rho \cdot D_i + u_i = \alpha + \rho \cdot D_i + (u_i^* + \mu) = (\alpha + \mu) + \rho \cdot D_i + u_i^* = \alpha^* + \rho \cdot D_i + u_i^*$$

. Thus, our new model has $E[u_i^*] = 0$

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

No. To compute $(\hat{\alpha}, \hat{\rho})$ from a given sample, you do not need any assumption on u_i .

You only need variation in D_i , i.e. at least one $D_i = 1$ and one $D_i = 0$

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

$\hat{\alpha}$ is the sample average spending (y_i) among non-members.

$\hat{\rho}$ is the sample difference in average spending between members and non-members.

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

$\hat{\rho}$ is generally biased and inconsistent for ρ when membership D_i is correlated with unobserved determinants u_i (e.g., income, household size).

$\hat{\alpha}$ is also biased and inconsistent if $E[u_i|D_i = 0] \neq 0$

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

ZMCA says members and non-members are same on average in all other unobserved determinants of spend captured by u_i . Equivalently, membership is not related to the unobserved determinants of spend, at least in the mean.

([] out of 1p) Q7.h: Identification of ρ under ZCMA

Assume ZMCA, i.e.

$$E[u_i|D_1] = E[u_i|D_0]$$

Then take the conditional expectation of model.

$$E[y_i|D_i = 1] = \alpha + \rho + E[u_i|D_1]$$

$$E[y_i|D_i = 0] = \alpha + E[u_i|D_0]$$

$$E[y_i|D_i = 1] - E[y_i|D_i = 0] = (\alpha + \rho + E[u_i|D_1]) - (\alpha + E[u_i|D_0]) = \rho$$

Thus under ZMCA we can identify ρ taking conditional expectation of y_i —

([] out of 1p) Q7.i: Identification of ρ (after weakening ZCMA)

Yes, for binary D_i , assuming $\text{cov}(u_i, D_i) = 0$ is equivalent to ZMCA. Thus it is enough for identification.

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

Assume RS and ZCMA: $\mathbb{E}[u_i | D_i] = 0$.

$$\hat{\rho} = \bar{y}_1 - \bar{y}_0.$$

$$\mathbb{E}[y_i | D_i = d] = \alpha + \rho d + \mathbb{E}[u_i | D_i = d] = \alpha + \rho d, \quad d \in \{0, 1\}.$$

$$\begin{aligned} \mathbb{E}[\hat{\rho}] &= \mathbb{E}[\bar{y}_1 - \bar{y}_0] = \mathbb{E}[y | D = 1] - \mathbb{E}[y | D = 0] = (\alpha + \rho) - \alpha = \rho, \\ &\Rightarrow \hat{\rho} \text{ is unbiased.} \end{aligned}$$

By Law of Large Numbers,

$$\begin{aligned}\bar{y}_1 &\xrightarrow{P} \mathbb{E}[y | D = 1], & \bar{y}_0 &\xrightarrow{P} \mathbb{E}[y | D = 0], \\ \hat{\rho} &= \bar{y}_1 - \bar{y}_0 \xrightarrow{P} \mathbb{E}[y | D = 1] - \mathbb{E}[y | D = 0] = \rho, \\ \Rightarrow \hat{\rho} &\text{ is consistent.}\end{aligned}$$

$\hat{\rho} \xrightarrow{P} \rho$ by the Continuous Mapping Theorem (and Slutsky's theorem where relevant).

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

Independence implies ZMCA, thus no change to conclusions.

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

No. Randomly sampling customers gives you a representative mix of members and non-members, but it does not make membership random. Customer will self select into membership based on a number of unobserved factors. Thus, We will probably encounter omitted variable bias. Because in real life there will be unobserved variables like household income and size that correlates both with D_i and u_i . Thus independence assumption or ZMCA assumptions will fail leading to biased and inconsistent estimators for ρ .

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

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

1. Faster and cheaper shipping and other benefits that decrease frictions will enable customer to have a more pleasant shopping experience, which in turn can make customer choose shopping at Walmart rather than other retailers.
 2. Cash pack program reduce the effective price for shopping at Walmart, increasing the quantity purchased.
-

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

Let "treatment" be the offer/assignment of Sam's Club Plus.

- Substitution bias : Control customers (not offered membership) respond by substituting to other actions/programs that affect Walmart.com spend, changing the control outcome relative to what it would have been absent the experiment. Examples: controls switch spending to in-store purchases to avoid shipping costs; controls enroll in a different paid program (or use a partner's account) that changes their Walmart.com spend; controls change shopping behavior because they are denied the offer.
 - Contamination bias: The control group is partially "treated" (or treated group not cleanly treated), so outcomes for control reflect exposure to treatment benefits. Examples: account-sharing of membership benefits with control customers; operationally extending free shipping rules to nonmembers; mis-implementation that accidentally grants membership to some controls; treated customers failing to activate/use the membership benefits (diluting the intended treatment contrast).
-

([] out of 2p) Q8.c

Because membership D_i randomly assigned in the RCT, D_i is independent of u_i . Thus ZMCA holds. The slope estimates the causal effect of assignment (the intention-to-treat effect) on spending in the post period. If compliance is perfect (offered actually treated and used), ITT equals the effect of receiving the membership.

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

([] out of 2p) Q9.a: Interpretation of ρ_i

ρ_i is customer i's individual-specific causal effect of membership on spending, i.e. $\rho_i = Y_i(1) - Y_i(0)$

([] out of 5p) Q9.b: Plain-English description of ρ_0 , ρ_1 , and ρ . Interpret $\rho_1 = \rho_0$.

- $\rho = \mathbb{E}[\rho_i]$: average causal effect across all customers.
- $\rho_1 = \mathbb{E}[\rho_i | D_i = 1]$: average causal effect **among members** (the treated subpopulation).
- $\rho_0 = \mathbb{E}[\rho_i | D_i = 0]$: average causal effect **among non-members** (the untreated subpopulation).

Assumption $\rho = \rho_1 = \rho_0$ means membership status is unrelated to the size of the gains from membership; membership selection is not based on who benefits more. This is why it is called “no selection on gains.”

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

Starting from model $y_i = \alpha + \rho_i D_i + v_i$. Add and subtract $p_1 D_i$

$$y_i = \alpha + p_1 D_i + (v_i + (p_i - p_1) D_i)$$

Since $p_1 = E[p_i | D_i = 1]$

This matches

$$y_i = \alpha + \rho_1 D_i + u_i \quad \text{with} \quad u_i = v_i + (\rho_i - \mathbb{E}[\rho_i | D_i = 1]) D_i.$$

([] out of 4p) Q9.d

- Substantive benefit: Under ZCMA for u_i , the simple treated-control difference $\hat{\rho}_1 = \bar{y}^1 - \bar{y}^0$ identifies ρ_1 , the average causal effect for the treated (members). With heterogeneous effects, this is the primary interpretable causal object you can recover from observational treated/control comparisons under mean-independence conditions.
- What ZCMA implies here: $\mathbb{E}[u_i | D_i = 1] = \mathbb{E}[u_i | D_i = 0]$ requires that the composite unobserved component u_i is mean-independent of D_i . Since u_i contains v_i and the treated-only deviation $(\rho_i - \rho_1) D_i$, ZCMA rules out systematic differences in those unobservables across membership groups (in the mean).

([] out of 2p) Q9.e

You need, in addition, **no selection on gains** so that $\rho_1 = \rho$ (equivalently $\mathbb{E}[\rho_i | D_i = 1] = \mathbb{E}[\rho_i]$). Then identifying ρ_1 also identifies ρ .

([] out of 4p) Q10

In a causal context, the OLS slope from regressing y_i on a constant and D_i equals a difference in group means, and it is causal only under additional assumptions linking D_i to unobserved determinants. In the homogeneous-effects model, the slope equals the causal effect ρ only if the unobserved component has the same mean across treated and control (ZCMA), which is plausibly violated in observational membership data. With heterogeneous effects, the same estimator can still identify an average causal effect for the treated (ρ_1) under an appropriate ZCMA for the redefined error, but it equals the population average effect only under no selection on gains. Random assignment provides a design-based route to these assumptions and supports causal interpretation.

([] 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.   :0.00000  Min.   :0.00000
1st Qu.:0.0000  1st Qu.:20.00  1st Qu.: 9.0  1st Qu.:1.00000  1st Qu.:0.00000
Median :0.0000  Median :24.00  Median :10.0  Median :1.00000  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.:1.00000  3rd Qu.:0.00000
Max.   :1.0000  Max.   :55.00  Max.   :16.0  Max.   :1.00000  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.