PS 200D: Causal Inference
Problem Set 1

Chao-yo Cheng

April 28, 2015

Question 1

Part (a)

Given that Z is discrete, $\mathbb{E}[X|Z]$ represents the expected value of X at different discrete realizations of Z. As a result, the number of values this quantity represents is the same as the number of realizations of the discrete random variable Z.

Part (b)

Given that Z is discrete, $p(X|Z)$ refers to the probability of X at different discrete realizations of Z. As a result, the number of values this quantity represents is the product of the number of realizations of X and that of Z. Given that X is continuous, the number of $p(X|Z)$ should be infinite.

Part (c)

Following from Part (a) and Part (b),

$$\mathbb{E}[X|Z] = \int xp(X = x|Z = z)dX \tag{1}$$

where Z takes discrete values.

Part (d)

See below. Since we have estimates of $p(X|Z)$ and the marginal probabilities $p(X)$, $p(Y)$, and $p(Z)$,

$$p(Z|X) = \frac{p(X|Z)p(Z)}{p(X)} \tag{2}$$

Part (e)

The expression $X \perp Y|Z$ means that X is independent of Y when conditional on Z.

Part (f)

Given that $X \perp Y|Z$, it follows that

$$p(X, Y|Z) = p(X|Z)p(Y|Z) \tag{3}$$

As a result,

$$p(X|Y, Z) = \frac{p(X, Y, Z)}{p(Y, Z)} = \frac{p(X, Y|Z)p(Z)}{p(Y|Z)p(Z)} = \frac{p(X|Z)p(Y|Z)p(Z)}{p(Y|Z)p(Z)} = p(X|Z) \tag{4}$$

Question 2

Part (a)

Given a binary treatment $D_i \in \{0, 1\}$ and observed outcome Y_i, Y_{1i} and Y_{0i} refer to the potential outcomes for observation i when $D_i = 1$ and $D_i = 0$ respectively. In other words,

$$Y_i = \begin{cases} Y_{1i} \text{ if } D_i = 1 \\ Y_{0i} \text{ if } D_i = 0 \end{cases} \tag{5}$$

For instance, if D_i is hospitalization and Y_i is i's health condition, then Y_{1i} and Y_{0i} denote i's potential health condition when i is hospitalized and the potential condition when i is not hospitalized respectively.

Part (b)

Y_{1i} refers to the *potential* outcome of i when i receives $D_i = 1$ rather than the *actually observed* outcome of i when i received the treatment. For unit i, Y_{1i} can be *unobservable* when $D_i = 0$. In the case of hospitalization case, Y_{1i} refers to i's health condition *should* i is hospitalized; this is different from the actually observed health condition when i is hospitalized.

Part (c)

Formally speaking, the average treatment effect (ATE) is defined as

$$\begin{aligned} ATE &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}] \end{aligned} \tag{6}$$

for all i. In other words, ATE the average difference between the two potential outcomes Y_{1i} and Y_{0i} across all units. In reality, it is impossible to compute ATE because for each i we are only able to observe either Y_{1i} or Y_{0i} depending the actual value of D_i.

In contrast, the average treatment effect among the treated (ATT) is defined as

$$\begin{aligned} ATT &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] \end{aligned} \tag{7}$$

for all i. That is, ATE refers to the average difference between the two potential outcomes Y_{1i} and Y_{0i} across all units with $D_i = 1$. In reality, it is also impossible to compute ATT because for each i we are only able to observe either Y_{1i} or Y_{0i} depending the actual value of D_i.

Part (d)

Based on the results in Part (c), it is clear that ATT and ATE by definition are not the same. However, ATT will be equal to ATE when D is independent of Y_0 and Y_1, which yields $\mathbb{E}[Y_{ij}|D_i = 1] = \mathbb{E}[Y_{ij}|D_i = 0]$

for all i and $j \in \{0, 1\}$.

$$\begin{aligned} ATT &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\ &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}] \\ &= ATE \end{aligned} \tag{8}$$

for all i.

Part (e)

Random assignment makes D independent of Y_0 and Y_1.

$$\begin{aligned} \mathbb{E}(Y_{1i}|X = x, D_i = 1) &= \mathbb{E}(Y_{1i}|X = x, D_i = 0) \\ \mathbb{E}(Y_{0i}|X = x, D_i = 1) &= \mathbb{E}(Y_{0i}|X = x, D_i = 0) \end{aligned} \tag{9}$$

The average treatment effect (ATT) conditional on X is then defined as follows.

$$\begin{aligned} \alpha_{ATE(x)} &= \mathbb{E}[Y_1 - Y_0|X = x] \\ &= \mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x] \\ &= \mathbb{E}[Y_{1i}|X = x, D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{1i}|X = x, D_i = 0]P(D_i = 0) \\ &- \mathbb{E}[Y_{0i}|X = x, D_i = 1]P(D_i = 1) - \mathbb{E}[Y_{0i}|X = x, D_i = 0]P(D_i = 0) \\ &= \mathbb{E}[Y_{1i}|X = x, D_i = 0] - \mathbb{E}[Y_{0i}|X = x, D_i = 0] \\ &= \mathbb{E}[Y_{1i}|X = x, D_i = 1] - \mathbb{E}[Y_{0i}|X = x, D_i = 0] \end{aligned} \tag{10}$$

Part (f)

If, conditional on X, D is independent of Y_0 and Y_1, it follows tat

$$\begin{aligned} \mathbb{E}(Y_{1i}|X = x, D_i = 1) &= \mathbb{E}(Y_{1i}|X = x, D_i = 0) \\ \mathbb{E}(Y_{0i}|X = x, D_i = 1) &= \mathbb{E}(Y_{0i}|X = x, D_i = 0) \end{aligned} \tag{11}$$

As a result,

$$\begin{aligned} \alpha_{ATE(x)} &= \mathbb{E}[Y_1 - Y_0|X = x] \\ &= \mathbb{E}[Y_1|X = x] - \mathbb{E}[Y_0|X = x] \\ &= \mathbb{E}[Y_{1i}|X = x, D_i = 1]P(D_i = 1|X = x) + \mathbb{E}[Y_{1i}|X = x, D_i = 0]P(D_i = 0|X = x) \\ &- \mathbb{E}[Y_{0i}|X = x, D_i = 1]P(D_i = 1|X = x) - \mathbb{E}[Y_{0i}|X = x, D_i = 0]P(D_i = 0|X = x) \\ &= \mathbb{E}[Y_{1i}|X = x, D_i = 0] - \mathbb{E}[Y_{0i}|X = x, D_i = 0] \\ &= \mathbb{E}[Y_{1i}|X = x, D_i = 1] - \mathbb{E}[Y_{0i}|X = x, D_i = 0] \end{aligned} \tag{12}$$

Question 3

Part (a)

See below. The difference in means estimand can be decomposed into the ATT and a bias term.

$$\begin{aligned} \mathbb{E}[Y_{1i} - Y_{0i}] &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\ &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\ &= ATT + \mathbb{E}[Y_{0i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \end{aligned} \tag{13}$$

where the bias term refers to the average difference of the potential outcome Y_{0i} when i is treated and not treated.

Part (b)

See below. The difference in means estimand can also be decomposed into the ATC and a bias term.

$$\begin{aligned} \mathbb{E}[Y_{1i} - Y_{0i}] &= \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 0] \\ &= \mathbb{E}[Y_{1i}|D_i = 0] - \mathbb{E}[Y_{0i}|D_i = 0] + \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{1i}|D_i = 0] \\ &= \mathbb{E}[Y_{1i} - Y_{0i}|D_i = 0] + \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{1i}|D_i = 0] \\ &= ATC + \mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{1i}|D_i = 0] \end{aligned} \tag{14}$$

where the bias term refers to the average difference of the potential outcome Y_{1i} when i is treated and not treated.

Part (c)

See below.

$$\begin{aligned} ATE &= \mathbb{E}[Y_{1i}] - \mathbb{E}[Y_{0i}] \\ &= \mathbb{E}[Y_{1i}|D_i = 1]P(D_i = 1) + \mathbb{E}[Y_{1i}|D_i = 0]P(D_i = 0) \\ &\quad - \mathbb{E}[Y_{0i}|D_i = 1]P(D_i = 1) - \mathbb{E}[Y_{0i}|D_i = 0]P(D_i = 0) \\ &= P(D_i = 1)(\mathbb{E}[Y_{1i}|D_i = 1] - \mathbb{E}[Y_{0i}|D_i = 1]) + P(D_i = 0)(\mathbb{E}[Y_{1i}|D_i = 0] - \mathbb{E}[Y_{0i}|D_i = 0]) \\ &= P(D_i = 1)\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 1] + P(D_i = 0)\mathbb{E}[Y_{1i} - Y_{0i}|D_i = 0] \\ &= P(D_i = 1)ATT + P(D_i = 0)ATC \\ &= P(D_i = 1)\left(\mathbb{E}[Y_{1i} - Y_{0i}] - E[Y_{0i}|D_i = 1] + E[Y_{0i}|D_i = 0]\right) \\ &\quad + P(D_i = 0)\left(\mathbb{E}[Y_{1i} - Y_{0i}] - E[Y_{1i}|D_i = 1] + E[Y_{1i}|D_i = 0]\right) \\ &= \mathbb{E}[Y_{1i} - Y_{0i}] + P(D_i = 1)\left(E[Y_{0i}|D_i = 0] - E[Y_{0i}|D_i = 1]\right) \\ &\quad + P(D_i = 0)\left(E[Y_{1i}|D_i = 0] - E[Y_{1i}|D_i = 1]\right) \end{aligned} \tag{15}$$

where the bias term equals to a weighted sum of the biases derived the decomposition of ATE by ATT and ATC. The weights are the probabilities of being treated and not being treated espectively.

Question 4

Part (a)

The difference in means estimator can be written as follows. Given randomization, D_i (the treatment) is independent of Y_{1i} and Y_{0i}.

$$\begin{aligned}\mathbb{E}[Y_i|D_i=1]-\mathbb{E}[Y_i|D_i=0] &= \mathbb{E}[Y_{1i}|D_i=1]-\mathbb{E}[Y_{0i}|D_i=0] \\ &= \mathbb{E}[Y_{1i}]-\mathbb{E}[Y_{0i}] \\ &= \overline{Y_1}-\overline{Y_0}\end{aligned} \tag{16}$$

The following `R` code estimate the average treatment effect by computing the difference-in-means between the treatment group and the control group with respect to the outcome variable `pct_missing`.

```
> y1 = dta[dta$treat_invite == 1,]
> y0 = dta[dta$treat_invite == 0,]
> mean(y1$pct_missing, na.rm=T)-mean(y0$pct_missing, na.rm=T)
[1] -0.02314737
```

Part (b)

Under the specified condition, we can assume $cov(\overline{Y_1}, \overline{Y_0}) = 0$. As a result, the variance of the above difference-in-means can be estimated by

$$\begin{aligned}\mathbb{V}(\overline{Y_1}-\overline{Y_0}) &= \mathbb{V}(\overline{Y_1})+\mathbb{V}(\overline{Y_0}) \\ &= \frac{\sigma^2_{Y_1}}{N_1}+\frac{\sigma^2_{Y_0}}{N_0}\end{aligned} \tag{17}$$

where N_1 and N_0 refer to the number of observations in the treatment group and the control group respectively. The standard error of the above difference-in-mean is accordingly the square of $\mathbb{V}(\overline{Y_1}-\overline{Y_0})$.

Part (c)

Based on the results derived in Part (b), see the following `R` code.

```
> var.y1 = var(y1$pct_missing, na.rm=T)
> var.y0 = var(y0$pct_missing, na.rm=T)
> n1 = nrow(y1)-length(is.na(y1$pct_missing)[is.na(y1$pct_missing)==T])
> n0 = nrow(y0)-length(is.na(y0$pct_missing)[is.na(y0$pct_missing)==T])
> var.estimate = (var.y1/n1) + (var.y0/n0)
> se.estimate = sqrt(var.estimate); se.estimate
[1] 0.03285978
```

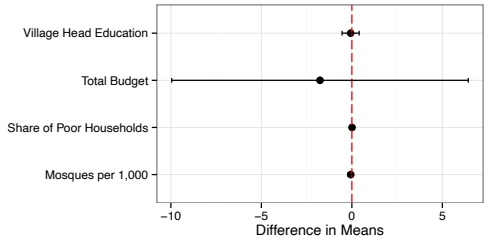
The standard error of the difference-in-means estimate is around 0.032860.

Part (d)

We can calculate the standard error of the difference-in-means and derive the 95% confidence interval to examine how different these two groups differ in terms of the covariates. See the following `R` code. The function also returns the p-value based on the derived t statistic.

```
b.test = function(var.y1,var.y0){
  diff = mean(var.y1, na.rm=T)-mean(var.y0, na.rm=T)
  v.y1 = var(var.y1, na.rm=T)
  v.y0 = var(var.y0, na.rm=T)
  n1 = nrow(y1)-length(is.na(var.y1)[is.na(var.y1)==T])
  n0 = nrow(y0)-length(is.na(var.y0)[is.na(var.y0)==T])
  var = (v.y1/n1) + (v.y0/n0)
  se = sqrt(var)
  ub = diff + 1.96*se
  lb = diff - 1.96*se
  t = diff/se
  p = 2*pt(-abs(t), df=n1+n0-1)
  return(list(diff=diff, se=se, ci=c(lb, ub), t=t, p=p))
}
```

Ideally, the confidence interval should cover 0. The figure below confirms that this is the case, although the standard error of the difference-in-means for `total_budget` is quite large.



The following table summarizes the results from the balance test. Again, in terms of the covariates, the treatment group does not significantly differ from the control group.

	Diff in Mean	Std Err	t value	p value	Upper Bound of 95% CI	Lower Bound of 95% CI
Village Head Education	−0.069	0.241	−0.285	0.776	0.404	−0.541
Mosques per 1,000	−0.062	0.074	−0.833	0.405	0.083	−0.207
Share of Poor Households	0.009	0.019	0.468	0.640	0.045	−0.028
Total Budget	−1.760	4.180	−0.421	0.674	6.433	−9.954

Part (e)

The proper choice of standard error will be the robust standard error because we would like to allow the variance differs between $D = 1$ and $D = 0$, which implies heteroscedasticity.

```
> library(sandwich); library(lmtest)
> mod <- lm(pct_missing ~ treat_invite, data=dta)
> modr <- coeftest(mod, vcov = vcovHC(mod, type = "HC2"))
> modr
t test of coefficients:

             Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.252106   0.026393  9.5521   <2e-16 ***
treat_invite -0.023147   0.032860 -0.7044   0.4815
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The SATE estimate is the same as the difference-in-means estimate in Part (a).

Part (f)

Based on the standard error derived in Part (b), we can examine whether the difference-in-means between the treatment group and the control group is significant different. Specifically, with $\widehat{\tau}_{ATE} = -0.02314737$ (see Part (a)) and $\sigma_{ATE} = 0.032860$, we can derive that the 95% confidence interval of the estimated effect lies between $(-0.08755254, 0.04125780)$. As a result, $\widehat{\tau}_{ATE}$ is not statistically significant at the .5 level of significance.

Part (h)

The robust standard error of the OLS estimate is the same as the standard error of the difference-in-means estimate because the OLS estimate applies the "HC2" heteroscedasticity-robust variance such that

$$\widehat{\sigma}^2_{HC2} = \frac{\mathbb{V}(Y_1)}{N_1} + \frac{\mathbb{V}(Y_0)}{N_0} \tag{18}$$

Part (i)

I use the following regression models to reestimate SATE. All estimation applies robust standard errors. The covariates are denoted as follows:

- $X1$: Village head education.
- $X2$: Mosques per 1,000.
- $X3$: Share of poor households.
- $X4$: Total budget.

In brief, compared with the results in Part (d), all SATE estimates remain largely unchanged while the standard error decreases when other covariates are included. In other words, controlling for the covariates makes the SATE estimate more efficient. Meanwhile, it is also noticeable that our SATE estimates are robust to alternative specifications. To see why this is the case, recall that the Frisch-Waugh-Lovell (FWL) theorem says that

$$\beta_k = \frac{cov(Y, \widetilde{Xk_i})}{var(\widetilde{Xk_i})} \tag{19}$$

where $\widetilde{Xk_i}$ is the residual derived by regressing Xk on all other X's. In the case of randomized experiment,

$$\widehat{\tau}_{ATE} = \frac{\widehat{cov}(Y, \widetilde{D}_i)}{\widehat{var}(\widetilde{D}_i)}. \tag{20}$$

Given that the treatment variable D_i ideally is independent of all other covariates, $\tilde{D}_i$ should almost the same as D_i because we are not able to use other covariates to predict D_i. Therefore, adding covariates should not bias our experiment finding. Finally, the final specification with demeaned covariates and interaction allows to estimate the average treatment effect for any $x \in X$, hence providing more nuances on our experiment findings.

Model w/ All Pre-treatment Covariates

The model specification is as follows:

$$y_i = \beta_0 + \tau_{ATE} D_i + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 X4_i + \epsilon_i \tag{21}$$

where i refers to individual observations.

```
> mod1 <- lm(pct_missing ~ ., data=dta)
> mod1r <- coeftest(mod1, vcov = vcovHC(mod1, type = "HC2"))
> mod1r

t test of coefficients:

                Estimate  Std. Error t value  Pr(>|t|)
(Intercept)   0.39044550  0.09373578  4.1654 3.705e-05 ***
treat_invite -0.02641825  0.03263124 -0.8096  0.418583
head_edu     -0.00550816  0.00614516 -0.8963  0.370533
mosques      -0.04819141  0.01844077 -2.6133  0.009257 **
pct_poor     -0.11771251  0.07336727 -1.6044  0.109297
total_budget  0.00053070  0.00032211  1.6475  0.100120
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model w/ Arbitrary Functions of the Covariates

In this model, I reestimate the previous model with all covariates but control for `total_budget`^{2} because government budget might impose a non-linear impact on the level of corruption (measured by `pct_missing`). Intuitively, the level of corruption might increase as the budget increases as the size of available rents grows large. However, beyond a certain level of government budget, government officials responsible for inspect might instead acquire sufficient resources to support their work, which leads to the reduction in corruption. The results, despite being insignificant, confirm this conjecture. To sum up, the model specification is as follows:

$$y_i = \beta_0 + \tau_{ATE} D_i + \beta_1 X1_i + \beta_2 X2_i + \beta_3 X3_i + \beta_4 X4_i + \beta_5 X4_i^2 + \epsilon_i \tag{22}$$

where i refers to individual observations.

```
> mod2 <- lm(pct_missing ~ . + I(total_budget^2), data=dta)
> mod2r <- coeftest(mod2, vcov = vcovHC(mod2, type = "HC2"))
> mod2r

t test of coefficients:

                     Estimate  Std. Error t value  Pr(>|t|)
(Intercept)        3.4639e-01  9.8119e-02  3.5303 0.0004564 ***
treat_invite      -2.4244e-02  3.2667e-02 -0.7421 0.4583730
head_edu          -5.5033e-03  6.1643e-03 -0.8928 0.3724485
mosques           -4.5665e-02  1.8291e-02 -2.4965 0.0128872 *
pct_poor          -1.1970e-01  7.3485e-02 -1.6289 0.1040172
total_budget       1.1363e-03  4.9340e-04  2.3030 0.0217193 *
I(total_budget^2) -1.0718e-06  6.5423e-07 -1.6383 0.1020341
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model w/ a Demeaned Covariate and Interaction

Finally, the model specification is as follows:

$$y_i = \beta_0 + \tau_{ATE} D_i + \beta_1 (X1_i - \overline{X1}) + \beta_2 D_i (X1_i - \overline{X1}) + \epsilon_i \tag{23}$$

where i refers to individual observations.

```
> treat <- dta[,1]
> Y <- dta[,2]
> head_edu.d <- dta[,3]-mean(dta[,3], na.rm=T)
> mod3 <- lm(Y ~ treat + head_edu.d + treat*head_edu.d)
> mod3r <- coeftest(mod3, vcov = vcovHC(mod3, type = "HC2"))
> mod3r

t test of coefficients:

                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       0.2529566  0.0265599  9.5240   <2e-16 ***
treat            -0.0250565  0.0330662 -0.7578   0.4490
head_edu.d       -0.0076937  0.0110056 -0.6991   0.4849
treat:head_edu.d  0.0049438  0.0133114  0.3714   0.7105
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Part (j)

Given that $y_i = \alpha + \beta_1 D_i + \epsilon_1$ (Model 1), comparing it to $y_i = \alpha + \beta_1 D_i + \beta_2 X_i + \epsilon_2$ (Model 2) suggests that $\epsilon_1 = \beta_2 X_i + \epsilon_2 > \epsilon_2$. As a result, the variance without controlling for X_i is generally larger than the variance derived from the model that controls X_i as

$$\mathbb{V}(\widehat{\beta_1}) = \frac{\sigma^2}{\sum x_i^2} \tag{24}$$

where $\sigma^2 = \mathbb{E}(\epsilon^2|X)$. As a result, $\mathbb{V}(\widehat{\beta_1})$ increases as ϵ grows large. We can also show this in matrix form. Recall that the covariance-variance matrix is

$$\begin{aligned}
\mathbb{E}[(\beta_{OLS} - \beta)(\beta_{OLS} - \beta)^\top |X] &= \mathbb{E}[((X^\top X)^{-1} X^\top \epsilon)((X^\top X)^{-1} X^\top \epsilon)^\top |X] \\
&= \mathbb{E}[(X^\top X)^{-1} X^\top \epsilon\epsilon^\top X (X^\top X)^{-1} |X] \\
&= (X^\top X)^{-1} X^\top \mathbb{E}[\epsilon\epsilon^\top |X] X (X^\top X)^{-1} \\
&= (X^\top X)^{-1} X^\top (\sigma^2 I_N) X (X^\top X)^{-1} \\
&= (X^\top X)^{-1} X^\top (\sigma^2) X (X^\top X)^{-1} \\
&= X^{-1} (X^\top)^{-1} X^\top (\sigma^2) X (X)^{-1} (X^\top)^{-1} \\
&= X^{-1} (\sigma^2) (X^\top)^{-1} \\
&= (\sigma^2) X^{-1} (X^\top)^{-1} \\
&= (\sigma^2) (X^\top X)^{-1}
\end{aligned} \tag{25}$$

where $\mathbb{E}[\epsilon\epsilon^\top |X] = \sigma^2 I_N$. As a result, σ^2 increases as ϵ increases. After we fit the model, we can still see that the residuals from Model 1 is greater than Model 2, leading to greater estimate variance for Model 1.

Question 5

Part (a)

Sharp null of no effect from the Fisher's Exact Test posits that $H_0 : Y_{1i} = Y_{0i}$ for all i. That is, the treatment imposes no effect for all units since the potential outcomes for unit i are the same when $D_i = 1$ and $D_i = 0$. In contrast, the null hypothesis we tested in the previous question says $H_0 : \mathbb{E}(Y_{1i}) = \mathbb{E}(Y_{0i})$, which means that the average non-treated potential outcome (i.e. Y_0) is the same as the average treated potential outcome (i.e. Y_1).

Part (b)

Sharp null is convenient because it allows us to "observe" all potential outcomes for all units, from which we can derive the distribution of the estimated average treated effect (ATE) by changing different randomization schemes.

Part (c)

See the function below. The function takes three arguments. First, `P` specified the number of permuted randomization schemes for alternative treated and non-treated potential outcomes to be observed. I set the default $S = 10,000$ so we can have a large number of permutations although the number of all possible permutations is 2^n (n is the total number of units in the sample). Second, Y is the outcome variable of interest. Finally, `D` is the treatment vector of D, as provided by the data. Based on `D`, I can create P different randomization schemes by using `sample(D)`.

```
sharp <- function(P=100000,Y,D){
  mean.diff <- rep(NA,P)
  for(i in 1:P){
    t <- sample(D)
    Exp = data.frame(Y,t)
    mean.diff[i] = mean(Exp$Y[t==1], na.rm=T)-mean(Exp$Y[t==0], na.rm=T)
  }
  plot(density(mean.diff), main="")
  t.diff = mean(Y[D==1], na.rm=T)-mean(Y[D==0], na.rm=T)
  abline(v=t.diff, col="red", lty=2)
  t = (mean(mean.diff)-t.diff)/sd(mean.diff)
  p = 2*pt(-abs(t), df=P-1)
  return(list(t.diff=t.diff,p=p))
}
```

The above function creates a plot showing the distribution of the difference in means statistic based on different randomization schemes with a red vertical dash line that represents the observed difference in means. The function also returns `p`, the p-value for the difference in means statistic against the sharp null. The object `t.diff` refers to the observed difference in the data.

Part (d)

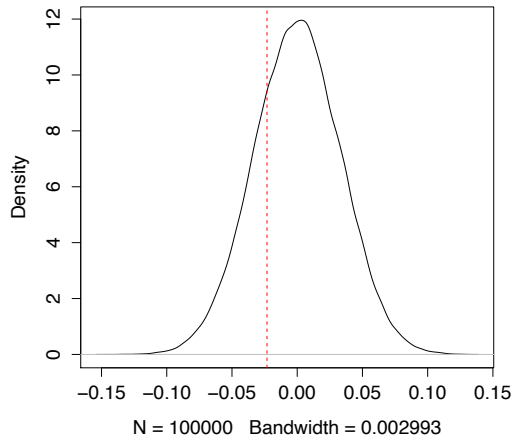
See below. The derived p-value suggests that we cannot reject the sharp null.

```
> sharp(P=100000, Y=dta$pct_missing, D=dta$treat_invite)
$t.diff
[1] -0.02314737

$p
```

```
[1] 0.48686
```

The distribution of the difference-in-mean statistic under shape null is illustrated below.



Appendix: R Code for Problem Set 1

```
#### 200D Problem Set 1
#### Chao-yo Cheng

##### Question 4 #####
setwd("C:/Users/CYCheng/Desktop/Dropbox/2015 Spring/Coursework/200D/Problem Set/Pset 1")
setwd("/Users/chaoyocheng/Dropbox/2015 Spring/Coursework/200D/Problem Set/Pset 1")
dta = read.csv("OlkenData.csv")
head(dta)

#### Part (a)
y1 = dta[dta$treat_invite == 1,]
y0 = dta[dta$treat_invite == 0,]
mean(y1$pct_missing, na.rm=T)-mean(y0$pct_missing, na.rm=T)

#### Part (b) and (c)
var.y1 = var(y1$pct_missing, na.rm=T)
var.y0 = var(y0$pct_missing, na.rm=T)
n1 = nrow(y1)-length(is.na(y1$pct_missing)[is.na(y1$pct_missing)==T])
n0 = nrow(y0)-length(is.na(y0$pct_missing)[is.na(y0$pct_missing)==T])

var.estimate = (var.y1/n1) + (var.y0/n0)
se.estimate = sqrt(var.estimate); se.estimate

#### Part (d)
b.test = function(var.y1,var.y0){
  diff = mean(var.y1, na.rm=T)-mean(var.y0, na.rm=T)
  v.y1 = var(var.y1, na.rm=T)
  v.y0 = var(var.y0, na.rm=T)
  n1 = nrow(y1)-length(is.na(var.y1)[is.na(var.y1)==T])
  n0 = nrow(y0)-length(is.na(var.y0)[is.na(var.y0)==T])
  var = (v.y1/n1) + (v.y0/n0)
  se = sqrt(var)
  ub = diff + 1.96*se
  lb = diff - 1.96*se
  t = diff/se
  p = 2*pt(-abs(t), df=n1+n0-1)
  return(list(diff=diff, se=se, ci=c(lb, ub), t=t, p=p))
}
edu <- b.test(y1$head_edu, y0$head_edu)
mos <- b.test(y1$mosques, y0$mosques)
poor <- b.test(y1$pct_poor, y0$pct_poor)
bud <- b.test(y1$total_budget, y0$total_budget)

res.diff <- c(edu$diff, mos$diff, poor$diff, bud$diff)
res.se <- c(edu$se, mos$se, poor$se, bud$se)
res.lb <- c(edu$ci[1], mos$ci[1], poor$ci[1], bud$ci[1])
res.ub <- c(edu$ci[2], mos$ci[2], poor$ci[2], bud$ci[2])
res.t <- c(edu$t, mos$t, poor$t, bud$t)
res.p <- c(edu$p, mos$p, poor$p, bud$p)

res <- cbind(res.diff, res.se, res.t, res.p, res.ub, res.lb)
rownames(res) <- c("Village Head Education", "Mosques per 1,000", "Share of Poor
    Households", "Total Budget")
colnames(res) <- c("Diff in Mean", "Std Err", "t value", "p value", "Upper Bound of 95%
    CI", "Lower Bound of 95% CI")
library(xtable); xtable(res, digits=3)
```

```
coef <- res[,1] # coefficients/center point
ub <- res[,5] # upper bound
lb <- res[,6] # lower bound
var <- rownames(res) # label vars etc.
all <- data.frame(var, coef, lb, ub)

library(ggplot2)
pd <- position_dodge(.1)

ggplot(all, aes(coef, var)) +
  geom_point(size=3, position=pd) +
  geom_vline(xintercept = 0, colour="red", linetype = "longdash") +
  xlab("Difference in Means") + ylab("") +
  geom_errorbarh(aes(xmin=lb, xmax=ub), height = 0.1) + theme_bw()

#### Part (e)
library(sandwich); library(lmtest)
mod <- lm(pct_missing ~ treat_invite, data=dta); summary(mod)
modr <- coeftest(mod, vcov = vcovHC(mod, type = "HC2"))
modr

#### Part (f)
b.test(y1$pct_missing,y0$pct_missing)

#### Part (i)
library(sandwich); library(lmtest)

mod1 <- lm(pct_missing ~ ., data=dta)
mod1r <- coeftest(mod1, vcov = vcovHC(mod1, type = "HC2"))
mod1r

mod2 <- lm(pct_missing ~ . + I(total_budget^2), data=dta)
mod2r <- coeftest(mod2, vcov = vcovHC(mod2, type = "HC2"))
mod2r

treat <- dta[,1]
Y <- dta[,2]
head_edu.d <- dta[,3]-mean(dta[,3], na.rm=T)
mosques <- dta[,4]
pct_poor <- dta[,5]
total_budget <- dta[,6]
mod3 <- lm(Y ~ treat + head_edu.d + mosques + pct_poor + total_budget + treat*head_edu.d
    )
mod3r <- coeftest(mod3, vcov = vcovHC(mod3, type = "HC2"))
mod3r

res.coef <- c(coef(mod1)[2], coef(mod2)[2], coef(mod3)[2])
res.se <- c(mod1r[2,2], mod2r[2,2], mod3r[2,2])
res.ub <- res.coef + 1.96*res.se
res.lb <- res.coef - 1.96*res.se

res <- cbind(res.coef, res.ub, res.lb)
rownames(res) <- c("Model 1", "Model 2 ", "Model 3")

coef <- res[,1] # coefficients/center point
ub <- res[,2] # upper bound
lb <- res[,3] # lower bound
```



```
var <- rownames(res) # label vars etc.
all <- data.frame(var, coef, lb, ub)

library(ggplot2)
pd <- position_dodge(.1)

ggplot(all, aes(coef, var)) +
  geom_point(size=3, position=pd) +
  geom_vline(xintercept = 0, colour="red", linetype = "longdash") +
  xlab("Estimated Marginal Effect") + ylab("") +
  geom_errorbarh(aes(xmin=lb, xmax=ub), height = 0.1) + theme_bw()

##### Question 5 #####
sharp <- function(P=100000,Y,D){
  mean.diff <- rep(NA,P)
  for(i in 1:P){
    t <- sample(D)
    Exp = data.frame(Y,t)
    mean.diff[i] = mean(Exp$Y[t==1], na.rm=T)-mean(Exp$Y[t==0], na.rm=T)
  }
  plot(density(mean.diff), main="")
  t.diff = mean(Y[D==1], na.rm=T)-mean(Y[D==0], na.rm=T)
  abline(v=t.diff, col="red", lty=2)
  t = (mean(mean.diff)-t.diff)/sd(mean.diff)
  p = 2*pt(-abs(t), df=P-1)
  return(list(t.diff=t.diff,p=p))
}
sharp(1000, dta$pct_missing, dta$treat_invite)
```