Assignment 3

Yongchao Zhao

9/30/2018

Part A: Linear Parametric form

Question 1: Simulate the data

(a) Start with the marginal distribution of X. Simulate as X~N(0,1) with sample size of 1000. Set the seed to be 1234.

```
# DGP
set.seed(1234)
X <- rnorm(1000, mean=0, sd=1)

summary(X)

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -3.39606 -0.67325 -0.03979 -0.02660  0.61582  3.19590
```

(b) Look at the DGP. What role does X play?

X serves as the confounding covariate that might influence the treatment participants attend as well as the potential outcomes.

(c) The distribution of binary Z depends on the value of X. Therefore, the next step is to simulate Z from p(Z|X) = Binomial(p), where the vector of probabilities can vary across observations. Come up with a strategy for generating the vector Z conditional on X that forces you to create be explicit about how these probabilities are conditional on X (an inverse logit function would be one strategy but there are others). Make sure that X is significantly associated with Z and that the vector of probabilities used to draw Z doesn't vary below .05 or above .95.

```
# Probability of treatment assignment by using inverse logit function
library(gtools)
p <- inv.logit(X, min = .05, max = .95)
summary(p)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.07918 0.35399 0.49105 0.49473 0.63434 0.91461

# treatment assignment
Z <- as.factor(rbinom(1000, 1, p))
table(Z)

## Z
##   0   1
## 491 509
```

(d) The last step is to simulate Y from p(Y0,Y1|Z,X). Come up with a strategy for simulating each potential outcome with appropriate conditioning on Z and X with the following stipulations.

(i) Make sure that E[Y(1)|X] - E[Y(0)|X] = 5.

(ii) Make sure that X has a linear and statistically significant relationship with the outcome.

(iii) Finally, set your error term to have a standard deviation of 1 and allow the residual standard error to be different for the same person across potential outcomes.

(iv) Create a data frame containing X,Y,Y0,Y1 and Z and save it for use later.

```
# simulate potential outcomes assuming a linear relationship (set beta0 = 10,
beta1 = 1, and tao = 5)
Y0 <- 10 + X + rnorm(1000, 0, 1)
Y1 <- 10 + X + 5 + rnorm(1000, 0, 1)

# generate the simulated dataset
data.lin <- data.frame(X, Z, Y0, Y1)
data.lin$Y <- ifelse(data.lin$Z == 1, data.lin$Y1, data.lin$Y0)


head(data.lin)
##            X Z        Y0       Y1         Y
## 1 -1.2070657 0 10.587009 12.42670 10.587009
## 2  0.2774292 1  8.912880 15.81664 15.816639
## 3  1.0844412 1 10.377001 14.76251 14.762509
## 4 -2.3456977 0  7.098018 12.37301  7.098018
## 5  0.4291247 1 10.119044 13.32418 13.324178
## 6  0.5060559 0 10.129877 13.88845 10.129877
```

(e) Think about the difference between the DGP used in this homework and the first DGP from previous homework (completely randomized experiment). How is the difference in the study design encoded?

Compared with the first DGP from previous homework where the treatment has been randomly assigned, the DGP used in this homework does not randomly assign treatment to participants. According to the encoded study design, the probability of treatment assignment in this homework is significantly associated with the covariate (or the pretest variable X).

(f) Calculate the SATE from 1.d.iv

The SATE from the simulated dataset is 5.003.

```
(SATE <- mean(data.lin$Y1 - data.lin$Y0))

## [1] 5.002736
```

Question 2: Playing the role of the researcher

(a) Estimate the treatment effect using a difference in mean outcomes across treatment groups (save it for use later).

The estimated treatment effect using mean outcome differences across treatment groups is 5.803.

```
(avgs <- tapply(data.lin$Y, data.lin$Z, mean))

##         0         1
##  9.548737 15.351830

(t_mean_diff <- avgs[[2]] - avgs[[1]])

## [1] 5.803093
```

(b) Estimate the treatment effect using a regression of the outcome on the treatment indicator and covariate (save it for use later).

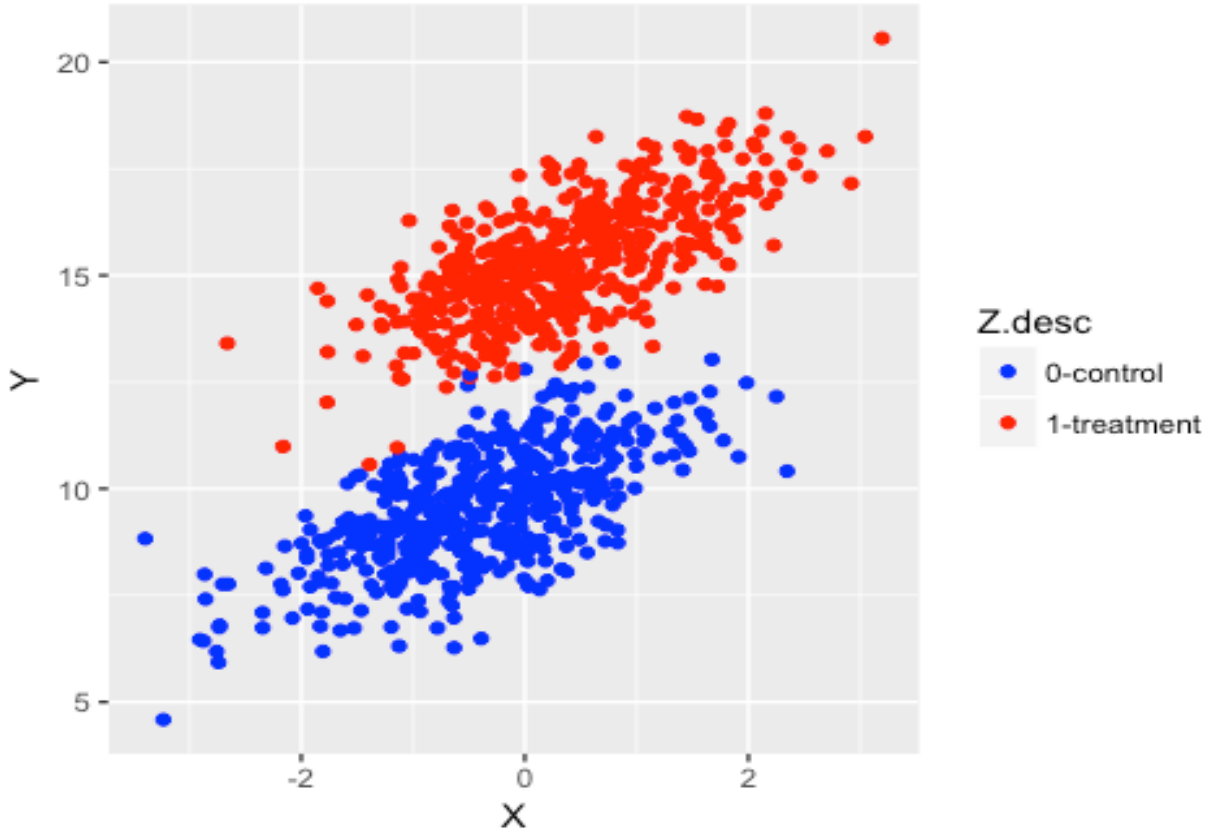
The estimated treatment effect using regression is 5.029.

```
summary(lm(Y ~ Z + X, data = data.lin))
## Call:
## lm(formula = Y ~ Z + X, data = data.lin)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0898 -0.7083  0.0355  0.6861  3.2000
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.96948    0.04927  202.36   <2e-16 ***
## Z1           5.02890    0.07133   70.50   <2e-16 ***
## X            1.00315    0.03577   28.04   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.04 on 997 degrees of freedom
## Multiple R-squared:  0.8958, Adjusted R-squared:  0.8956
## F-statistic:  4285 on 2 and 997 DF,  p-value: < 2.2e-16

(t_regreesion <- coef(lm (Y ~ Z + X, data = data.lin))[[2]])

## [1] 5.028901
```

(c) Create a scatter plot of X versus the observed outcome with different colors for treatment and control observations (suggested: red for treated and blue for control). If you were the researcher would be comfortable using linear regression in this setting?

According to the scatterplot, we do see overlap in the pretest variable and linearity for both treatment and control groups. Thus, from a researcher's perspective, linear regression seems a comfortable option in this setting.

```
library(ggplot2)
data.lin$Z.desc <- ifelse(data.lin$Z == 1, "1-treatment", "0-control")
ggplot(data.lin, aes(X, Y, color=Z.desc)) + geom_point() +
scale_color_manual(values = c("blue","red"))
```



Question 3: Exploring the properties of estimators

a) Create a scatter plot of X versus each potential outcome with different colors for treatment and control observations (suggested: red for Y(1) and blue for Y(0)). Is linear regression a reasonable model to estimate causal effects for the observed data set? Why or why not?

Similarly, according to the scatterplot, we see great overlap in the pretest variable and the linearity also holds for both potential outcomes. Furthermore, if X is the only confounding covariate, the linear regression seems a reasonable model to estimate the causal effects for this observed data set.

```
# reshape dataset from wide to long for plotting
library(tidyr)
data.lin_new <- data.lin
data.lin_new$ID <- 1:1000
data.lin_new_long <- gather(data.lin_new, po_outcomes, po_values, Y0:Y1,
factor_key = T)
```

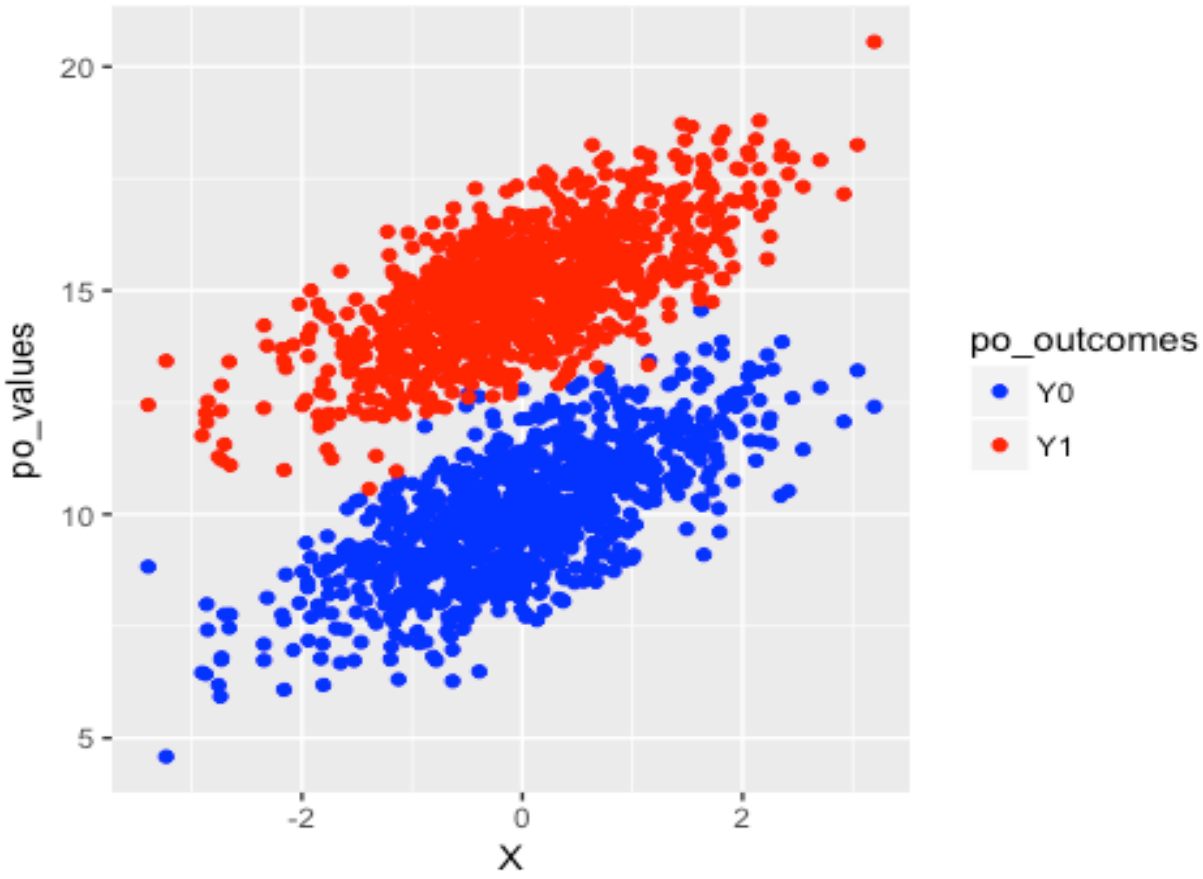
```r
# check the reshape
head(data.lin_new, 3)
```

```
##            X Z       Y0       Y1        Y       Z.desc ID
## 1 -1.2070657 0 10.58701 12.42670 10.58701    0-control  1
## 2  0.2774292 1  8.91288 15.81664 15.81664 1-treatement  2
## 3  1.0844412 1 10.37700 14.76251 14.76251 1-treatement  3
```

```r
subset(data.lin_new_long, ID %in% c(1:3))
```

```
##               X Z        Y       Z.desc ID po_outcomes po_values
## 1    -1.2070657 0 10.58701    0-control  1          Y0  10.58701
## 2     0.2774292 1 15.81664 1-treatement  2          Y0   8.91288
## 3     1.0844412 1 14.76251 1-treatement  3          Y0  10.37700
## 1001 -1.2070657 0 10.58701    0-control  1          Y1  12.42670
## 1002  0.2774292 1 15.81664 1-treatement  2          Y1  15.81664
## 1003  1.0844412 1 14.76251 1-treatement  3          Y1  14.76251
```

```r
# scatterplot
ggplot(data.lin_new_long, aes(X, po_values, color = po_outcomes)) +
geom_point() + scale_color_manual(values=c("blue", "red"))
```



b) Find the bias of each of the estimates calculated by the researcher in Question 2 relative to SATE.

Evidently, the regression controlling the covariate yields a better estimate of the treatment effect.

```
# bias of mean difference approach
(bias_mean_diff <- t_mean_diff - SATE)
## [1] 0.8003567
# bias of regression approach
(bias_regression <- t_regreesion - SATE)
## [1] 0.02616469
```

c) Think harder about the practical significance of the bias by dividing this estimate by the standard deviation of the observed outcome Y.

```
# practical significance of bias of mean difference approach
bias_mean_diff/sd(data.lin$Y)
## [1] 0.2487226
# practical significance of bias of regression approach
bias_regression/sd(data.lin$Y)
## [1] 0.008131062
```

d) Find the bias of each of the estimators by creating randomization distributions for each. [Hint: When creating randomization distributions remember to be careful to keep the original sample the same and only varying treatment assignment and the observed outcome.]

By creating randomization distributions of the treatment, the new bias of estimated treatment effect from differences in means is <u>0.152</u>, while the new bias from regression approach is <u>0.078</u>.

```
data.lin.rand <- data.lin[, c("X", "Z", "Y0", "Y1", "Y")]
data.lin.rand$ID <- 1:1000

# randomly assign the treatment
set.seed(100)
sampling <- sample(data.lin.rand$ID, 500, replace = F)
data.lin.rand$Z_rand<-as.factor(ifelse(data.lin.rand$ID %in% sampling, 1, 0))

# new observed outcome
data.lin.rand$Y_rand <- ifelse(data.lin.rand$Z_rand == 1, data.lin.rand$Y1,
data.lin.rand$Y0)
summary(data.lin.rand)
##        X            Z             Y0               Y1
##  Min.   :-3.39606   0:491   Min.   : 4.587   Min.   :10.57
##  1st Qu.:-0.67325   1:509   1st Qu.: 8.968   1st Qu.:14.01
##  Median :-0.03979           Median :10.030   Median :15.02
```

```
##  Mean   :-0.02660           Mean   : 9.998   Mean   :15.00
##  3rd Qu.: 0.61582           3rd Qu.:10.992   3rd Qu.:15.91
##  Max.   : 3.19590           Max.   :14.563   Max.   :20.56
##        Y                ID         Z_rand      Y_rand
##  Min.   : 4.587   Min.   :   1.0   0:500   Min.   : 5.929
##  1st Qu.: 9.589   1st Qu.: 250.8   1:500   1st Qu.: 9.913
##  Median :12.807   Median : 500.5           Median :12.467
##  Mean   :12.503   Mean   : 500.5           Mean   :12.464
##  3rd Qu.:15.354   3rd Qu.: 750.2           3rd Qu.:15.128
##  Max.   :20.556   Max.   :1000.0           Max.   :20.556
```

```
# new bias of mean differences
(avgs_new <- tapply(data.lin.rand$Y_rand, data.lin.rand$Z_rand, mean))
```

```
##         0         1
##  9.886097 15.040906
```

```
(t_mean_diff_new <- avgs_new[[2]] - avgs_new[[1]])
```

```
## [1] 5.15481
```

```
(bias_mean_diff_new <- t_mean_diff_new - SATE)
```

```
## [1] 0.1520738
```

```
# new bias of regression
summary(lm(Y_rand ~ Z_rand + X, data = data.lin.rand))
## Call:
## lm(formula = Y_rand ~ Z_rand + X, data = data.lin.rand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0862 -0.6806  0.0084  0.6917  3.0577
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.94879    0.04548  218.74   <2e-16 ***
## Z_rand1      5.08047    0.06430   79.01   <2e-16 ***
## X            0.95942    0.03225   29.75   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.016 on 997 degrees of freedom
## Multiple R-squared:  0.8801, Adjusted R-squared:  0.8799
## F-statistic:  3660 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
(t_regreesion_new <- coef(lm(Y_rand ~ Z_rand + X, data =
data.lin.rand))[[2]])
```

```
## [1] 5.080466
```

```
(bias_regression_new <- t_regreesion_new - SATE)
```

```
## [1] 0.0777299
```

Part B: Non-Linear Parametric form

Question 1: Simulate the data

(a) Create function sim.nlin with the following DGP.

(i) X should be drawn from a uniform distribution between 0 and 2.

(ii) Treatment assignment should be drawn from a Binomial distribution with the following properities (make sure you save the p vector for use later).
$E[Z \mid X] = p = \text{logit}^{-1}(-2 + X2)$ $Z \sim \text{Binom}(N, p)$

(iii) The response surface (model for Y(0) and Y(1)) should be drawn from the following distributions:
$Y(0) = 2X + \varepsilon_0$
$Y(1) = 2X+3X^2 +\varepsilon_1$
where both error terms are normally distributed with mean 0 and standard deviation of 1.

(iv) Make sure the returned dataset has a column for the probability of treatment assignment as well.

```
# Create function sim.nin for DGP
library(gtools)
sim.nlin<- function(N) {
  set.seed(1234)
  X = runif(N, min = 0, max = 2) # uniform distribution between 0 and 2
  p = inv.logit(-2+X^2) # probability of treatment assignment
  Z = as.factor(rbinom(N,1,p)) # treatment assignment

  # potential outcomes
  Y0 = 2*X + rnorm(N, 0, 1)
  Y1 = 2*X + 3*X^2 + rnorm(N, 0, 1)

 return(data.frame(X, Z, p, Y0, Y1))
}
```

(b) Simulate a data set called data.nlin with sample size 1000.

```
data.nlin <- sim.nlin(1000)
# create observed outcome variable
data.nlin$Y <- ifelse(data.nlin$Z == 1, data.nlin$Y1, data.nlin$Y0)

summary(data.nlin)
```

```
##        X             Z             p                Y0
##  Min.   :0.0006836   0:649   Min.   :0.1192   Min.   :-2.2664
##  1st Qu.:0.5163177   1:351   1st Qu.:0.1502   1st Qu.: 0.9445
##  Median :1.0203848           Median :0.2771   Median : 2.0579
##  Mean   :1.0145469           Mean   :0.3719   Mean   : 2.0436
##  3rd Qu.:1.5168771           3rd Qu.:0.5747   3rd Qu.: 3.1643
##  Max.   :1.9986061           Max.   :0.8802   Max.   : 5.7451
```

```
##        Y1                Y          
##  Min.   :-2.145   Min.   :-2.266  
##  1st Qu.: 2.031   1st Qu.: 1.076  
##  Median : 5.174   Median : 2.468  
##  Mean   : 6.163   Mean   : 4.421  
##  3rd Qu.: 9.796   3rd Qu.: 6.832  
##  Max.   :18.815   Max.   :18.815
```

(c) Make the following plots.

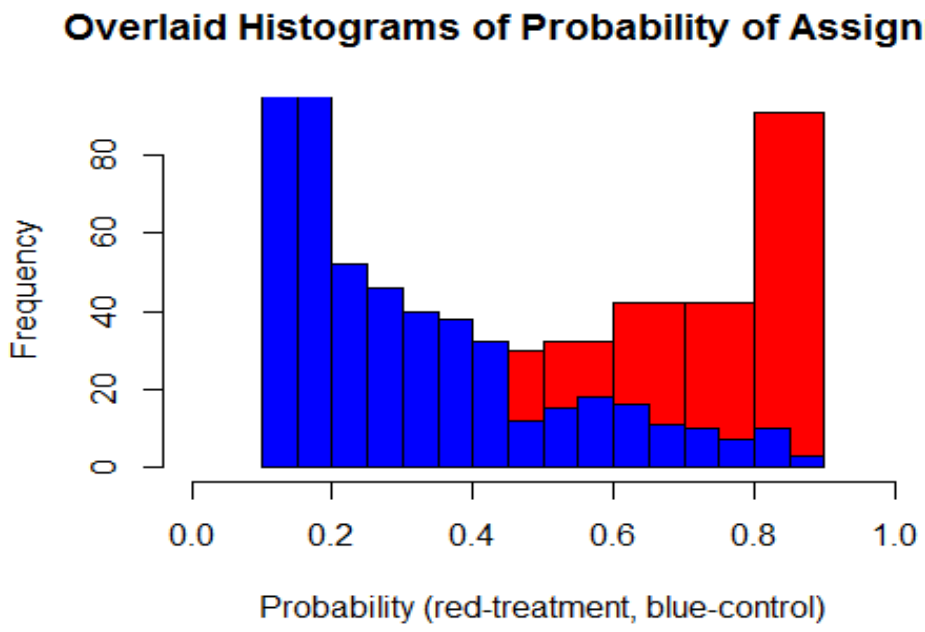
(i) Create overlaid histograms of the probability of assignment.

(ii) Make a scatter plot of X versus the observed outcomes versus X with different colors for each treatment group.

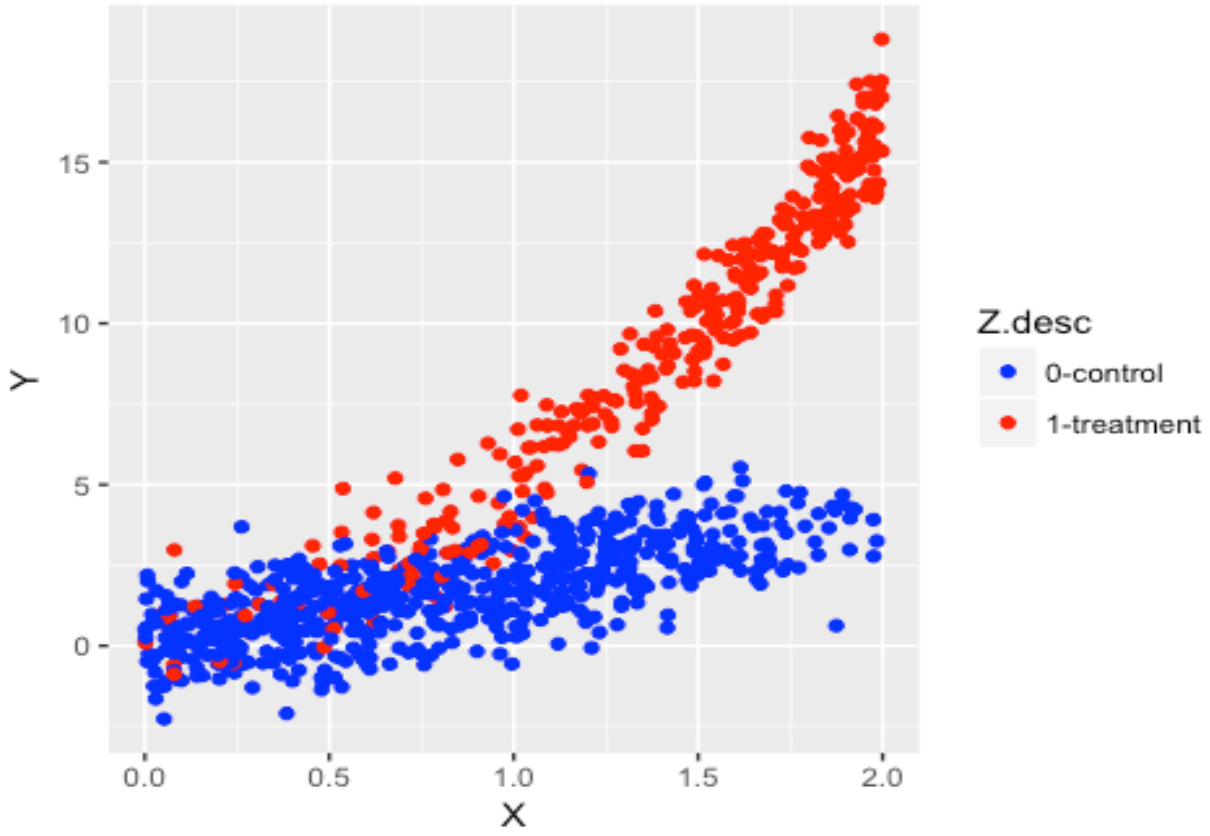
(iii) Create a scatter plot of X versus each potential outcome with different colors for treatment and control observations (suggested: red for Y(1) and blue for Y(0)). Does linear regression of Y and X seem like a good model for this response surface?

According to the scatterplots, clearly, linearity does not hold in this case, therefore, the linear regression of Y and X is not a good model for this response surface.

```
# Overlaid histograms of probability of assignment
hist(data.nlin[data.nlin$Z == 1,]$p, col="red", main = "Overlaid Histograms
of Probability of Assignment", xlab = "Probability (red-treatment, blue-
control)", xlim = c(0,1))

hist(data.nlin[data.nlin$Z == 0,]$p, col="blue", add =T)
```



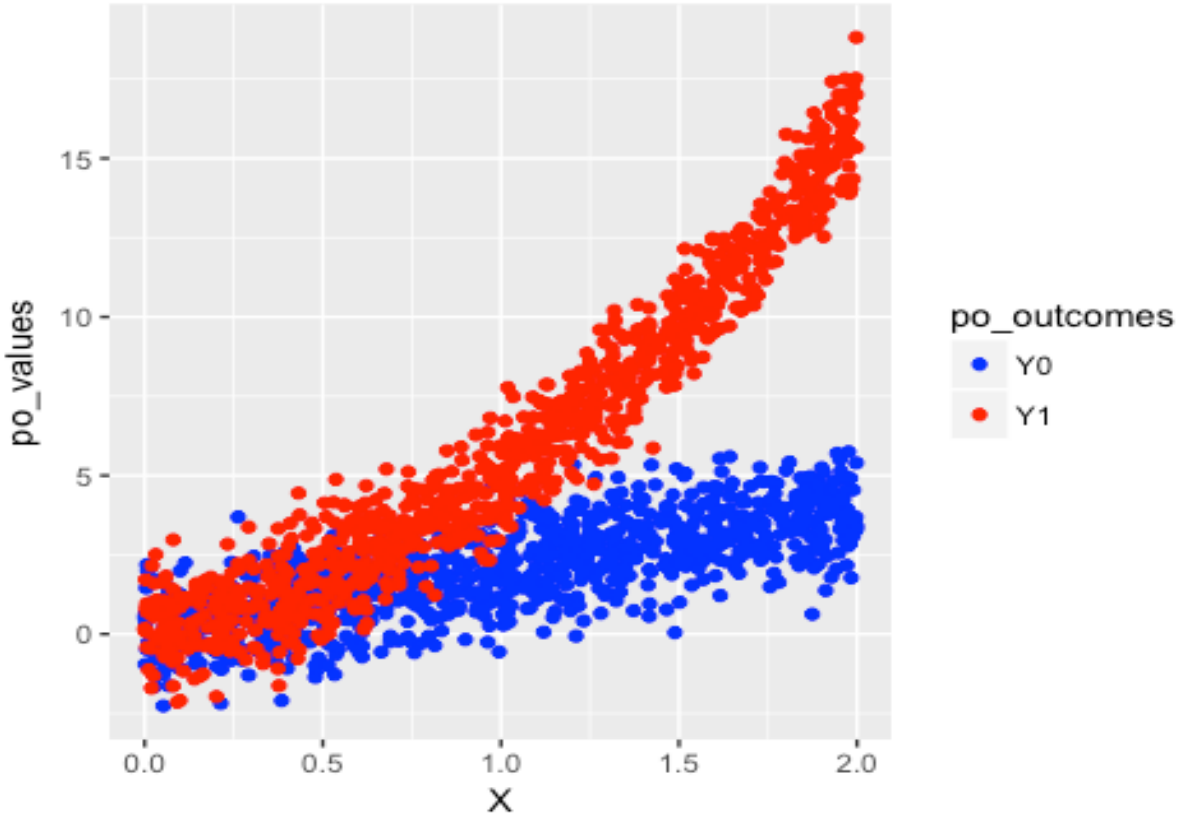
```
# scatterplot of X and observed outcomes
data.nlin$Z.desc <- ifelse(data.nlin$Z == 1, "1-treatment", "0-control")

library(ggplot2)

ggplot(data.nlin, aes(X, Y, col = Z.desc)) + geom_point() +
scale_color_manual(values = c("blue","red"))
```




```
# Scatterplot of X and potential outcomes
 # reshape data from wide to ling for plotting
library(tidyr)


data.nlin_new<- data.nlin
data.nlin_new$ID <- 1:1000
data.nlin_new_long <- gather(data.nlin_new, po_outcomes, po_values, Y0:Y1,
factor_key = T)

 # check the reshape
head(data.nlin_new, 3)
```

```
##           X Z         p         Y0         Y1          Y    Z.desc ID
## 1 0.2274068 0 0.1247404 -0.7505198 -0.3638634 -0.7505198 0-control  1
## 2 1.2445988 0 0.3891293  2.7906644  7.0366450  2.7906644 0-control  2
## 3 1.2185495 0 0.3739900  0.8979537  6.7809523  0.8979537 0-control  3
```

```
subset(data.nlin_new_long, ID %in% c(1:3))

##              X Z         p          Y    Z.desc ID po_outcomes  po_values
## 1    0.2274068 0 0.1247404 -0.7505198 0-control  1          Y0 -0.7505198
## 2    1.2445988 0 0.3891293  2.7906644 0-control  2          Y0  2.7906644
## 3    1.2185495 0 0.3739900  0.8979537 0-control  3          Y0  0.8979537
## 1001 0.2274068 0 0.1247404 -0.7505198 0-control  1          Y1 -0.3638634
## 1002 1.2445988 0 0.3891293  2.7906644 0-control  2          Y1  7.0366450
## 1003 1.2185495 0 0.3739900  0.8979537 0-control  3          Y1  6.7809523

 # Scatterplot
ggplot(data.nlin_new_long, aes(X, po_values, colour = po_outcomes)) +
geom_point() + scale_color_manual(values = c("blue", "red"))
```



(d) Create randomization distributions to investigate the properties of each of 3 estimators with respect to SATE: (1) difference in means, (2) linear regression of the outcome on the treatment indicator and X, (3) linear regression of the outcome on the treatment indicator, X, and X^2.

In terms of the output shown below,

- **SATE from the simulated data set is <u>4.119</u>:**
- **Estimated SATE from mean differences is <u>4.035</u>;**
- **Estimated SATE from regression of outcome on treatment and X is <u>4.084</u>;**
- **Estimated SATE from regression of outcome on treatment, X and X^2 is <u>4.075</u>.**

```r
# generate randomization distributions of the treatment
data.nlin.rand <- data.nlin
data.nlin.rand$ID <- 1:1000
set.seed(123)
sampling <- sample(data.nlin.rand$ID, 500, replace = F)
data.nlin.rand$Z_rand <- as.factor(ifelse(data.nlin.rand$ID %in% sampling, 1,
0))
data.nlin.rand$Y_rand <- ifelse(data.nlin.rand$Z_rand == 1,data.nlin.rand$Y1,
data.nlin.rand$Y0)

head(data.nlin.rand)
```

```
##           X Z         p         Y0         Y1          Y    Z.desc ID
## 1 0.2274068 0 0.1247404 -0.7505198 -0.3638634 -0.7505198 0-control  1
## 2 1.2445988 0 0.3891293  2.7906644  7.0366450  2.7906644 0-control  2
## 3 1.2185495 0 0.3739900  0.8979537  6.7809523  0.8979537 0-control  3
## 4 1.2467589 0 0.3904092  3.1288885  8.3489355  3.1288885 0-control  4
## 5 1.7218308 0 0.7240621  4.1466133 10.6818792  4.1466133 0-control  5
## 6 1.2806212 0 0.4109573  0.6553596  6.4355712  0.6553596 0-control  6
##   Z_rand     Y_rand
## 1      1 -0.3638634
## 2      0  2.7906644
## 3      0  0.8979537
## 4      0  3.1288885
## 5      1 10.6818792
## 6      1  6.4355712
```

```r
# SATE of difference in means
(avgs <- tapply(data.nlin.rand$Y_rand, data.nlin.rand$Z_rand, mean))
```

```
##        0        1
## 2.082298 6.117443
```

```r
(t_diff <- avgs[[2]] -avgs[[1]])
```

```
## [1] 4.035145
```

```r
# SATE of regression 1
mod1 <- lm(Y_rand ~ Z_rand + X, data=data.nlin.rand)
summary(mod1)
```

```
## Call:
## lm(formula = Y_rand ~ Z_rand + X, data = data.nlin.rand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.1779 -1.5691 -0.2889  1.5336  6.4189
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -3.0702     0.1501  -20.45   <2e-16 ***
## Z_rand1       4.0842     0.1334   30.62   <2e-16 ***
## X             5.0544     0.1145   44.12   <2e-16 ***
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Residual standard error: 2.109 on 997 degrees of freedom
## Multiple R-squared:  0.7417, Adjusted R-squared:  0.7412
## F-statistic:  1431 on 2 and 997 DF,  p-value: < 2.2e-16
```

```
(t_regression01 <- coef(mod1)[[2]])
```

```
## [1] 4.084222
```

```
# SATE of regression 2
mod2 <- lm(Y_rand ~ Z_rand + X + I(X^2), data=data.nlin.rand)
summary(mod2)
## Call:
## lm(formula = Y_rand ~ Z_rand + X + I(X^2), data = data.nlin.rand)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8991 -1.5169 -0.0564  1.5317  5.6758
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.2452     0.2098 -10.701  < 2e-16 ***
## Z_rand1       4.0753     0.1314  31.008  < 2e-16 ***
## X             2.6384     0.4502   5.860 6.29e-09 ***
## I(X^2)        1.1918     0.2150   5.543 3.81e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.078 on 996 degrees of freedom
## Multiple R-squared:  0.7494, Adjusted R-squared:  0.7487
## F-statistic: 992.9 on 3 and 996 DF,  p-value: < 2.2e-16
```

```
(t_regression02 <- coef(mod2)[[2]])
```

```
## [1] 4.075257
```

(e) Calculate the standardized bias (bias divided by the standard deviation of Y) of these estimators relative to SATE.

```
# SATE
(SATE <- mean(data.nlin.rand$Y1) - mean(data.nlin.rand$Y0))
```

```
## [1] 4.118968
```

```
# standardized bias of mean differences
(t_diff-SATE)/sd(data.nlin.rand$Y_rand)
```

```
## [1] -0.02022516
```

```
# standardized bias of regression 01
(t_regression01-SATE)/sd(data.nlin.rand$Y_rand)
```

```
## [1] -0.008383712
```

```
# standardized bias of regression 02
(t_regression02-SATE)/sd(data.nlin.rand$Y_rand)
```

```
## [1] -0.01054692
```