Homework 05

Causal Inference

Name

October 15, 2019

Design of an experiment

1. Suppose you are interested in the effect of the presence of vending machines in schools on childhood obesity. What randomized experiment would you want to do (in a perfect world) to evaluate this question?

- In perfect experiment design, we can perfectly randomly assign the experiment units into two group, which means each units have the same propability being assigned to any of the two groups, one with vending machine and another without vending machine, so the estimated treatment effect is the difference between the obesity level of these two groups.

2. Suppose you are interested in the effect of smoking on lung cancer. What randomized experiment could you plausibly perform (in the real world) to evaluate this effect?

- In real world we can only observe the units whether they smoke or not, which means each units don‘t have the same probability of smoking or not, due to some other facts one may be more likely to smoke than another one, so they are perfectly ramdom, so we use propensity matching to control other factors, which is to ensure that in each matched group they have same probability smoking or not, than the treatment effect is the difference between the mean ratio or possibility of having lung cancer.

3. Suppose you are a consultant for a researcher who is interested in investigating the effects of teacher quality on student test scores. Use the strategy of mapping this question to a randomized experiment to help define the question more clearly. Write a memo to the researcher asking for needed clarifications to this study proposal.

- First we have to make clear that how to define the teaching quality, how to measure the teaching quality, and as for the experiment design, we should ask whether we can assign student to different teaching quality level, if so we can randomly assign student to different teaching quality level and the effect of teaching quality is the defference between the mean test score.

Causal effect

The table below describes a hypothetical experiment on 2400 persons. Each row of the table specifies a category of person, as defined by his or her pre-treatment predictor x, treatment indicator T , and potential outcomes y_0, y_1. (For simplicity, we assume unrealistically that all the people in this experiment fit into these eight categories.)

Category	# persons in category	x	T	y_0	y_1
1	300	0	0	4	6
2	300	1	0	4	6
3	500	0	1	4	6
4	500	1	1	4	6
5	200	0	0	10	12
6	200	1	0	10	12
7	200	0	1	10	12
8	200	1	1	10	12

In making the table we are assuming omniscience, so that we know both y_0 and y_1 for all observations. But the (nonomniscient) investigator would only observe x, T, and y^T for each unit. (For example, a person in category 1 would have $x = 0$,$T = 0$,$y = 4$, and a person in category 3 would have $x = 0$,$T = 1$,$y = 6$.)

(a) What is the average treatment effect in this population of 2400 persons?

- As we can see that the pre-treatment predictor x does not affect the potential outcome. The mean of y_0 is 7, the mean of y_1 is 9, so the difference is -2

(b) Is it plausible to believe that these data came from a randomized experiment? Defend your answer.

- Calculate the $\mathbb{E}(y_1|T = 1) - \mathbb{E}(y_0|T = 0) = 2$ which is equals to $\mathbb{E}(y_1) - \mathbb{E}(y_0) = 2$, so I conclude that this experiment is randomized.

(c) Another population quantity is the mean of y for those who received the treatment minus the mean of y for those who did not. What is the relation between this quantity and the average treatment effect?

- This is a randomized experiment, so the mean of y for those who received the treatment minus the mean of y for those who did not equals to the average treatment effect.

(d) For these data, is it plausible to believe that treatment assignment is ignorable given sex? Defend your answer.

- If x represents sex, we can see that the value of x does not affect the potential outcome, so it is plausible to say that the treatment assignment is ignorable given sex.

(e) Figure out the estimate and the standard error of the coefficient of T in a regression of y on T and x.

- Just do linear regression

```
y <- c(rep(4,300),rep(4,300),rep(6,500),rep(6,500),rep(10,400),rep(12,400))
x <- c(rep(0,300),rep(1,300),rep(0,500),rep(1,500),rep(0,200),rep(1,200),rep(0,200),rep(1,200))
t <- c(rep(0,600),rep(1,1000),rep(0,400),rep(1,400))
fit1 <- lm(y~t+x)
summary(fit1)
```

```
##
## Call:
## lm(formula = y ~ t + x)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -2.400 -1.886 -1.714  3.600  4.286
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 6.400e+00  1.058e-01   60.51   <2e-16 ***
## t           1.314e+00  1.163e-01   11.30   <2e-16 ***
## x           2.901e-15  1.147e-01    0.00        1
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.81 on 2397 degrees of freedom
## Multiple R-squared:  0.05055,    Adjusted R-squared:  0.04976
## F-statistic: 63.81 on 2 and 2397 DF,  p-value: < 2.2e-16
```

```
sqrt(vcov(fit1)[2,2])
```

```
## [1] 0.1163413
```

Consulting

You are consulting for a researcher who has performed a randomized trial where the treatment was a series of 26 weekly therapy sessions, the control was no therapy, and the outcome was self-report of emotional state one year later. However, most people in the treatment group did not attend every therapy session. In fact there was a good deal of variation in the number of therapy sessions actually attended. The researcher is concerned that her results represent "watered down" estimates because of this variation and suggests adding in another predictor to the model: number of therapy sessions attended. What would you advise her?

- The variable "number of therepy session attended" is correlated to the outcome and the treantment, so ignoring this variable will generate the endogenity problem, the best way is to add the missing variable.

Gain-score models:

In the discussion of gain-score models in [GH] Section 9.3, we noted that if we include the pre-treatment measure of the outcome in a gain score model, the coefficient on the treatment indicator will be the same as if we had just run a standard regression of the outcome on the treatment indicator and the pre-treatment measure. Show why this is true.

- Let y_1 represents the post-treatment score, and y_0 represents the pre-treatment score, so we can define the Gain-Score as $g = y_1 - y_0$, following define two regression model:

$$(1) : y_1 = \beta_0 + \beta_1 y_0 + \beta_2 T + \varepsilon$$
$$(2) : g = \alpha_0 + \alpha_1 T + \varepsilon$$
$$(3) : g = y_1 - y_0$$

Plug (3) into (1), we get

$$(4) : g = \beta_0 + (\beta_1 - 1) y_0 + \beta_2 T + \varepsilon$$

In (2),

$$\alpha_0 = \mathbb{E}[g|T=0] = \mathbb{E}[y_1 - y_0|T=0] = \mathbb{E}[\beta_0 + \beta_1 y_0 + \beta_2 T + \varepsilon - y_0|T=0] =$$
$$\mathbb{E}[\beta_0 + (\beta_1 - 1) y_0 + \beta_2 T + \varepsilon|T=0] = \beta_0 + (\beta_1 - 1) y_0$$

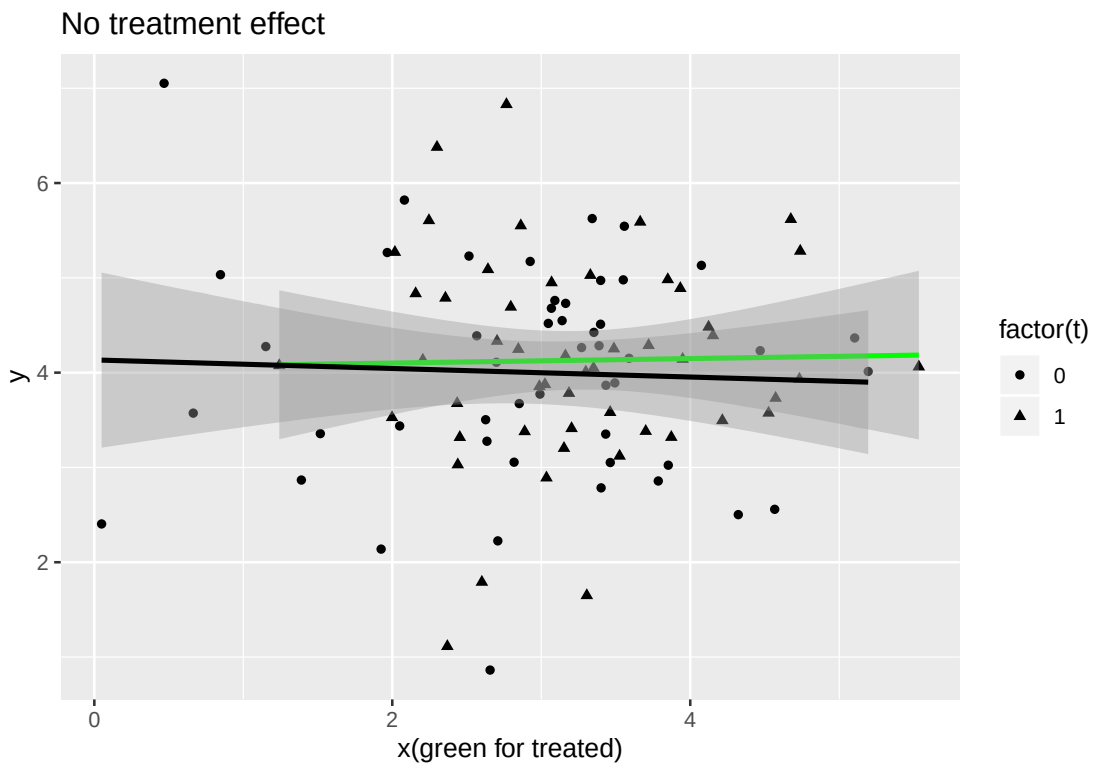
so $\alpha_0 = \beta_0 + (\beta_1 - 1) y_0$, plug this result into (2), we get $g = \beta_0 + (\beta_1 - 1) y_0 + \alpha_1 T + \varepsilon$, compare this with (4), we can deduce that $\alpha_1 = \beta_2$, so the coefficient of treatment in standard regression and Gain-Score regression are the same.

linear regression

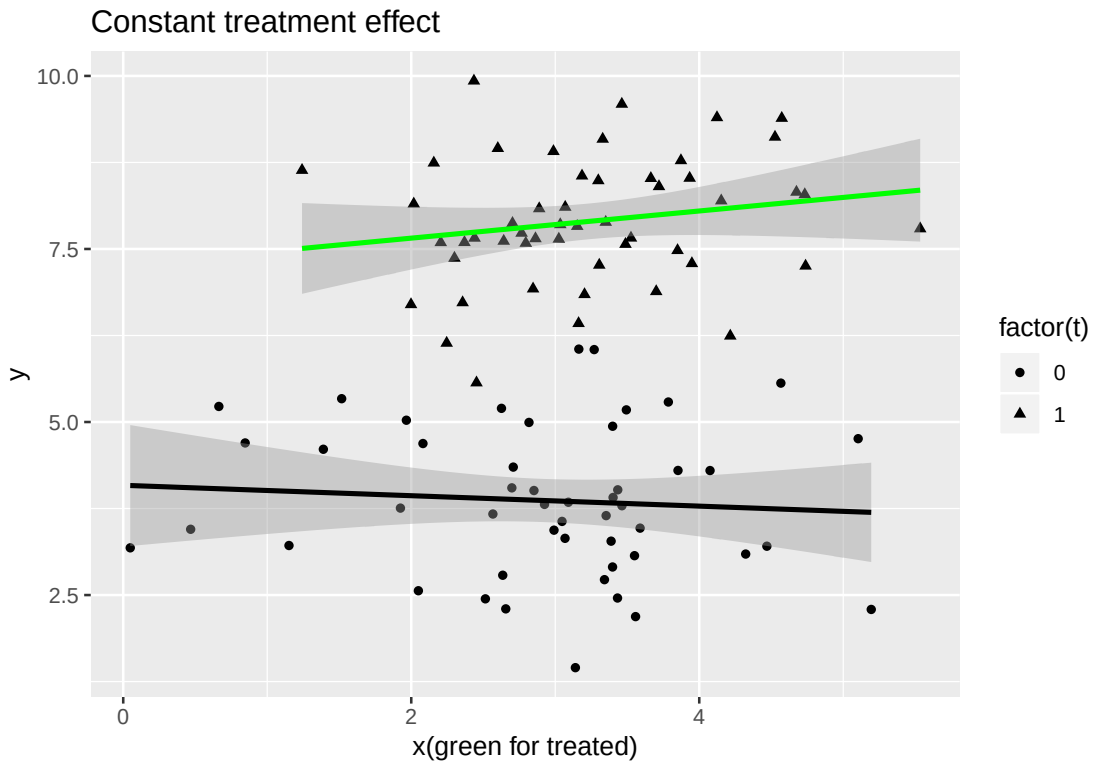
Assume that linear regression is appropriate for the regression of an outcome, y, on treatment indicator, T, and a single confounding covariate, x. Sketch hypothetical data (plotting y versus x, with treated and control units indicated by circles and dots, respectively) and regression lines (for treatment and control group) that represent each of the following situations: a. No treatment effect, b. Constant treatment effect, c. Treatment effect increasing with x.

```
y1 <- rnorm(100,4,1)
t <- Rlab::rbern(100,0.5)
x <- rnorm(100,3,1)
treated <- t==1
control <- t==0
ggplot()+
  geom_point(aes(y=y1, x=x, shape=factor(t)))+
  geom_smooth(aes(y=y1[treated], x=x[treated]), colour="green", method = lm)+
  geom_smooth(aes(y=y1[control], x=x[control]),colour="black", method = lm)+
```

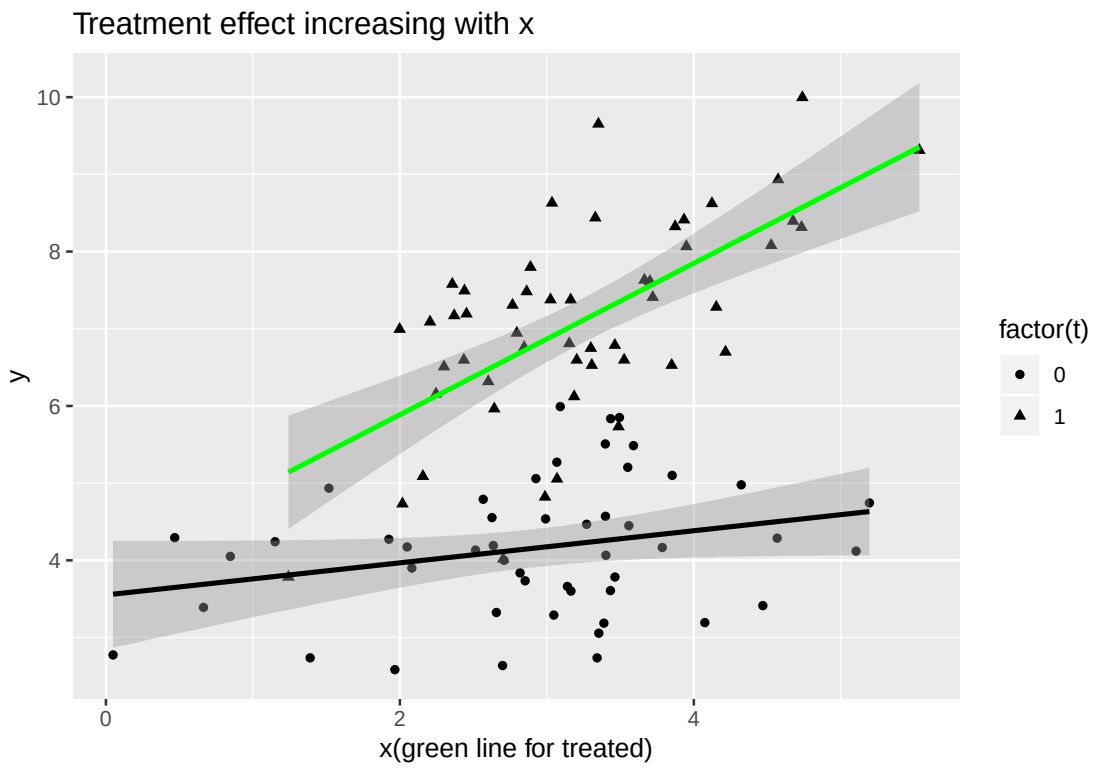
```
  labs(title="No treatment effect",
       x ="x(green for treated)", y = "y")
```




```
y2 <- rep(0,100)
for (i in 1:100){
  if (t[i]==1){
    y2[i] <- rnorm(1,8,1)
  }else{
    y2[i] <-rnorm(1,4,1)
  }
}
ggplot()+
  geom_point(aes(y=y2, x=x, shape=factor(t)))+
  geom_smooth(aes(y=y2[treated], x=x[treated]), colour="green", method = lm)+
  geom_smooth(aes(y=y2[control], x=x[control]),colour="black", method = lm)+
  labs(title="Constant treatment effect",
       x ="x(green for treated)", y = "y")
```




```
y3 <- rep(0,100)
for (i in 1:100){
  if (t[i]==1){
    y3[i] <- rnorm(1,4+x[i],1)
  }else{
    y3[i] <-rnorm(1,4,1)
  }
}
ggplot()+
  geom_point(aes(y=y3, x=x, shape=factor(t)))+
  geom_smooth(aes(y=y3[treated], x=x[treated]), colour="green", method = lm)+
  geom_smooth(aes(y=y3[control], x=x[control]),colour="black", method = lm)+
  labs(title="Treatment effect increasing with x",
       x ="x(green line for treated)", y = "y")
```



Hypothetical Study

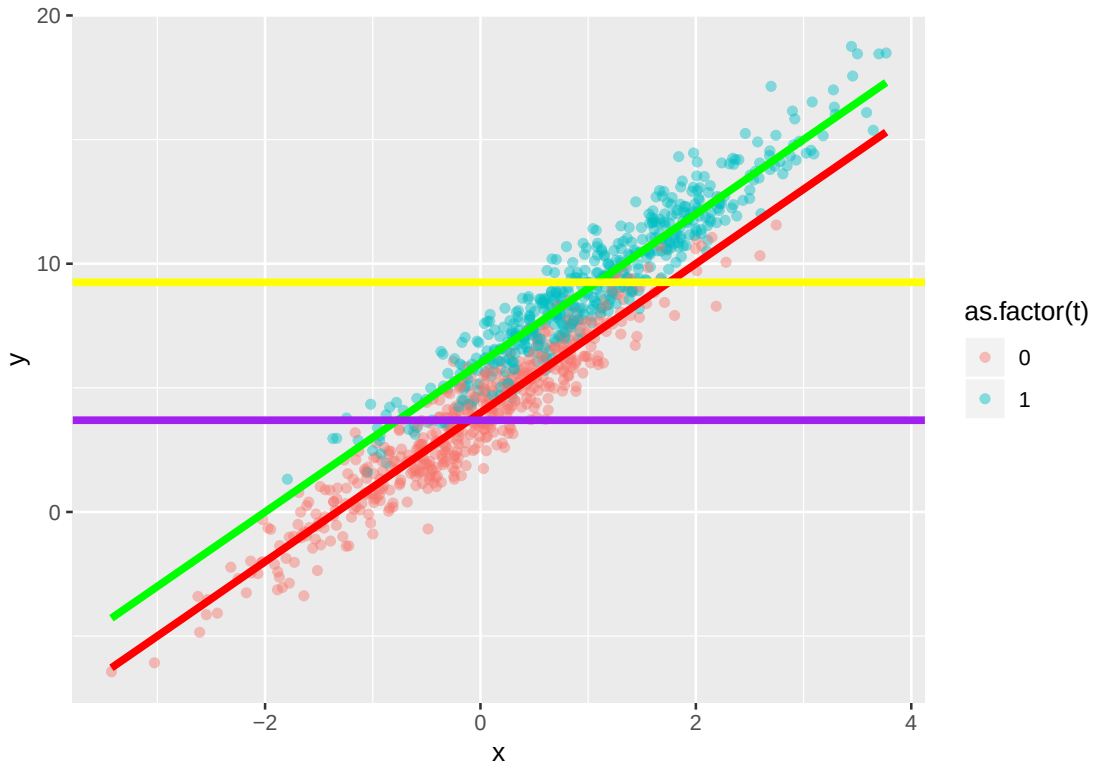
Consider a study with an outcome, y, a treatment indicator, T , and a single confounding covariate, x. Draw a scatterplot of treatment and control observations that demonstrates each of the following: (a) A scenario where the difference in means estimate would not capture the true treatment effect but a regression of y on x and T would yield the correct estimate.

```
n = 1000

t = sample(c(0,1),size = n,replace = T, prob = c(0.5,0.5))
x = rnorm(n,mean = t,sd = 1)
y = 4 + 3*x + 2*t + rnorm(n)

da = data.frame(cbind(y,x,t))

library(ggplot2)
f1 = function(x){return(4+3*x)}
f2 = function(x){return(4+3*x+2)}
ggplot(da) + geom_point(aes(y = y,x = x, color = as.factor(t)),alpha = 0.45) +
  stat_function(fun = f1,color = "red",size = 1.5)+
  stat_function(fun = f2,color = "green",size = 1.5) +
  geom_hline(yintercept = mean(da$y[which(da$t==1)]),color = "yellow",size = 1.5) +
  geom_hline(yintercept = mean(da$y[which(da$t==0)]),color = "purple",size = 1.5)
```



As we can see that we sample T from 0 and 1 with same possibility, and set x comply with normal distribution with constant variance but mean equals to T, so T is correlated with x, let $y = 4 + 3X + 2T + \varepsilon$, above plot shows the scater plot of X and Y, Green and Red line represents the fitted line of Y with different T value, horizental line represents the mean of treatment group and control group. the true treatment effect should be the difference bewteen the slope of Green and Red line, so the estimation through means are clearly bigger than true treatment effect.

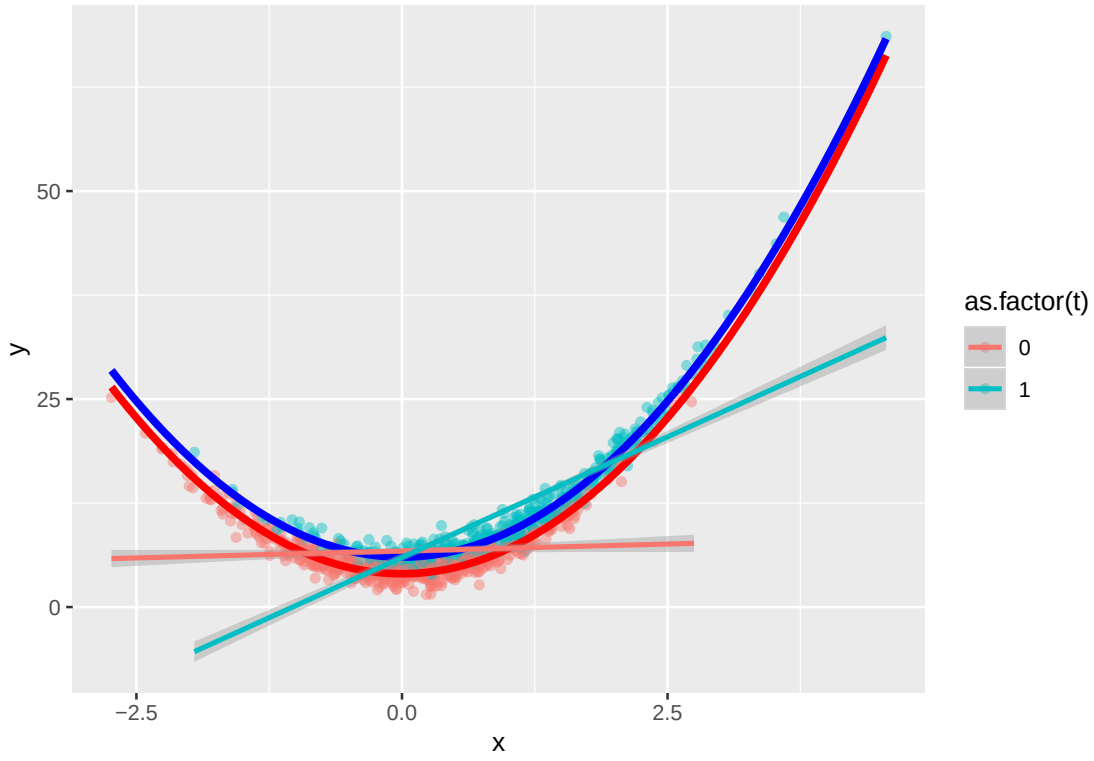
(b) A scenario where a linear regression would yield the wrong estimate but a nonlinear regression would yield the correct estimate.

```
n = 1000

t = sample(c(0,1),size = n,replace = T, prob = c(0.5,0.5))
x = rnorm(n,mean = t,sd = 1)
y = 4 + 3*x^2 + 2*t + rnorm(n)

da = data.frame(cbind(y,x,t))

library(ggplot2)
f1 = function(x){return(4+3*x^2)}
f2 = function(x){return(4+3*x^2+2)}
ggplot(da) + geom_point(aes(y = y,x = x, color = as.factor(t)),alpha = 0.45) +
  stat_function(fun = f1,color = "red",size = 1.5)+
  stat_function(fun = f2,color = "blue",size = 1.5) +
  geom_smooth(aes(y = y,x = x, color = as.factor(t)),method = "lm")
```



Above plot shows this situation, two convex line is the real relation ship between Y and X, so the streight line failed to show the true relationship.

Messy randomization

The folder `cows` contains data from an agricultural experiment that was conducted on 50 cows to estimate the effect of a feed additive on six outcomes related to the amount of milk fat produced by each cow.

Four diets (treatments) were considered, corresponding to different levels of the additive, and three variables were recorded before treatment assignment: lactation number (seasons of lactation), age, and initial weight of cow.

Cows were initially assigned to treatments completely at random, and then the distributions of the three covariates were checked for balance across the treat- ment groups; several randomizations were tried, and the one that produced the "best"balance with respect to the three covariates was chosen. The treatment assignment is ignorable (because it depends only on fully observed covariates and not on unrecorded variables such as the physical appearances of the cows or the times at which the cows entered the study) but unknown (because the decisions whether to rerandomize are not explained). We shall consider different estimates of the effect of additive on the mean daily milk fat produced. a. Consider the simple regression of mean daily milk fat on the level of additive. Compute the estimated treatment effect and standard error, and explain why this is not a completely appropriate analysis given the randomization used. b. Add more predictors to the model. Explain your choice of which variables to include. Compare your estimated treatment effect to the result from (a). c. Repeat (b), this time considering additive level as a categorical predictor with four letters. Make a plot showing the estimate (and standard error) of the treatment effect at each level, and also showing the inference the model fit in part (b).

sesame

The folder `sesame` contains data from an experiment in which a randomly selected group of children was encouraged to watch the television program Sesame Street and the randomly selected control group was

not. a. The goal of the experiment was to estimate the effect on child cognitive development of watching more Sesame Street. In the experiment, encouragement but not actual watching was randomized. Briefly explain why you think this was done. (Hint: think of practical as well as statistical reasons.) b. Suppose that the investigators instead had decided to test the effectiveness of the program simply by examining how test scores changed from before the intervention to after. What assumption would be required for this to be an appropriate causal inference? Use data on just the control group from this study to examine how realistic this assumption would have been. c. Did encouragement (the variable viewenc in the dataset) lead to an increase in post-test scores for letters (postlet) and numbers (postnumb)? Fit an appropriate model to answer this question. d. We are actually more interested in the effect of watching Sesame Street regu- larly (regular) than in the effect of being encouraged to watch Sesame Street. Fit an appropriate model to answer this question. e. Comment on which of the two previous estimates can plausibly be interpreted causally.

```
data = read.dta("http://www.stat.columbia.edu/~gelman/arm/examples/sesame/sesame.dta")
```