Andrew ElHabr
ECON 7022
10/21/17

Problem Set 2

Note: I worked with Cyrus Rich on this assignment.

Reading (25pts):

Question 1: Reading Angrist & Pischke [10pts]

1. - Completed.
   - Completed.
2. - The Conditional Independence Assumption (CIA) asserts that the treatment outcomes, Y_{0i} and Y_{1i}, conditioned on some known observed covariates X_i, are independent of the treatment assignment, Z_i, i.e.
   $$\{Y_{0i}, Y_{1i}\} \perp\!\!\!\perp Z_i | X_i.$$
   Note that this definition can be extended to treatment assignments that take on more than two values. This assumption reduces our selection bias, $\mathbb{E}[Y_{01}|Z_i = 1] - \mathbb{E}[Y_{01}|Z_i = 0]$ to 0, and thus, gives us that
   $$\mathbb{E}[Y_i|X_i, Z_i = 1] - \mathbb{E}[Y_i|X_i, Z_i = 0] = \mathbb{E}[Y_{1i} - Y_{0i}|X_i],$$
   i.e. the difference in our average outcome across different treatments is equal to the average treatment effect given the observed covariates. If this assumption is not satisfied, we cannot assume that selection bias is equal to 0, and then we have to find some way to estimate this bias, which may not be easy to do. One example in which the CIA would not hold is the following:
   Let X_i represent the age of a person. Now, if we let $Z_i = 0$ if this person is not approved for tax relief by signing up for a wellness and benefit program and $Z_i = 1$ if this person is approved for tax relief by signing up for a wellness and benefit program, and let Y_i be the amount of money out-of-pocket that this person pays in taxes. Clearly, the CIA will not hold because no matter what the observed age of a person is, the amount that they pay in taxes out-of-pocket is highly dependent on whether or not they are granted tax relief. In order for the CIA to hold, we need to find observed variables that are much more strongly correlated to our outcomes than our treatment assignments are to our outcomes.
   - Regression can be useful in finding causal relationships given the experimental ideal. However, regression has nothing to do with the design or the selection process of the experimental ideal. It is simply a tool that one can use with the experimental ideal to quantify causal relationships in given data.
   - Matching is the process of calculating covariate-specific treatment-control comparisons that are weighted together to produce a single overall average treatment effect. Matching and regression are both control strategies in which the causal interpretation of their outputs is based on the validity of the CIA. Regression can be motivated as a specific type of matching, and thus the differences between the two are unlikely to be of major empirical importance.

Question 2: Hansen [10pts]

1. We covered most of Chapter 8 (Hypothesis Testing) and only some of Chapter 9 (Regression Extensions), specifically the section about generalized least squares and testing for heteroskedasticity, in class

already. Obviously, the Hansen lecture notes go into much more detail than the topics that we briefly discussed in class.

2. ***Exercise 9.1*** *Suppose that $y_i = g(\boldsymbol{x}_i, \boldsymbol{\theta}) + e_i$ with $\mathbb{E}(e_i|\boldsymbol{x}_i) = 0, \hat{\boldsymbol{\theta}}$ is the NLLS estimator, and $\hat{\boldsymbol{V}}$ is the estimator of var($\hat{\boldsymbol{\theta}}$). You are interested in the conditional mean function $\mathbb{E}(y_i|\boldsymbol{x}_i = \boldsymbol{x}) = g(\boldsymbol{x})$ at some $\boldsymbol{x}$. Find an asymptotic 95% confidence interval for $g(\boldsymbol{x})$.*

   Let our t-statistic be $t_n = |t_n(\theta_0)|$ where $t_n(\theta) = (\hat{\theta} - \theta)/se(\hat{\theta})$, and let us set $T_n = t_n$. Then, we have that $t_n \xrightarrow{d} |Z|$ as $n \to \infty$ where $Z \sim N(0,1)$. This means that our asymptotic null distribution T is the symmetric standard normal distribution. Therefore, our 95% confidence interval for $g(\boldsymbol{x})$ is $[g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) - 1.96se(\hat{\boldsymbol{\theta}}), g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) + 1.96se(\hat{\boldsymbol{\theta}})] = [g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) - 1.96\sqrt{\hat{\boldsymbol{V}}}, g(\boldsymbol{x}, \hat{\boldsymbol{\theta}}) + 1.96\sqrt{\hat{\boldsymbol{V}}}]$.

Question 3: Reading summary - Identification [5pts]

- When there is an identification issue/endogeneity problem, the OLS assumption that is not met is strict exogeneity (A.2), i.e. $\mathbb{E}[u_i|\boldsymbol{X}] = 0$.
- Three threats to identification that we have mentioned in class are omitted variable bias, measurement errors in regressors, and simultaneous equations.
- When we omit variables, the error term in the regression equation will contain the effects of these omitted variables, which we cannot necessarily assume are uncorrelated to the covariates included in the regression equation. This will also lead to inflated/deflated coefficient estimates for the variables included in the linear model. When there are measurement errors in the regressors, once again, the error term in the regression equation will contain the implicit effect of the true measures of the regressors, which is clearly correlated to the incorrectly measured regressors. This often leads to a downward bias in the coefficient estimates of the variables included in the linear model. When there are simultaneous equations, we cannot infer the exact cause of an observed change in the response variable because the explanatory variable is dependent on multiple intertwined relationships.

Theory (10pts):

Question 1: Sample Bias [4pts]

1. They are corrupted by selection bias since it is likely that the beer/soft drink consumption of person i, Y_i, would be relatively high for someone who went to the game since these types of people tend to have a higher affinity for beer than people who do not care to go to games, i.e. $\mathbb{E}[Y_{0i}|A_i = 1] \neq \mathbb{E}[Y_{0i}|A_i = 0]$ and $\mathbb{E}[Y_{0i}|A_i = 1] - \mathbb{E}[Y_{0i}|A_i = 0] > 0$, where $A_i = 1$ if the person saw the advertisement for beer in the game and $A_i = 0$ if the person did not see the advertisement for beer in Publix.
2. I think it is likely that the overall beer consumption of person i would be greater if he/she is at the game, so $\mathbb{E}[Y_{0i}|A_i = 1] - \mathbb{E}[Y_{0i}|A_i = 0] > 0$. Therefore, I think the selection bias is positive, i.e. the bias works upwards. This would indicate that the effectiveness of the advertisement is overrated.
3. The observed treatment effect of the advertisement is

   $$\mathbb{E}[Y_i|A_i = 1] - \mathbb{E}[Y_i|A_i = 0],$$

   where A_i is the indicator if the advertisement was seen at the football game and Y_i is beer/soft drink consumption. Note that

   $$Y_i = \begin{cases} Y_{1i} & \text{if } A_i = 1, \\ Y_{0i} & \text{if } A_i = 0, \end{cases} \tag{1}$$

   so $Y_i = Y_{0i} + (Y_{1i} - Y_{0i})A_i$.

4. The observed treatment effect of the advertisement can be decomposed in the following way:

$$\mathbb{E}[Y_i|A_i=1]-\mathbb{E}[Y_i|A_i=0]=\mathbb{E}[Y_{1i}-Y_{0i}|A_i=1]+\mathbb{E}[Y_{0i}|A_i=1]-\mathbb{E}[Y_{0i}|A_i=0],$$

where $\mathbb{E}[Y_{1i}-Y_{0i}|A_i=1]$ is the unobserved treatment effect and $\mathbb{E}[Y_{0i}|A_i=1]-\mathbb{E}[Y_{0i}|A_i=0]$ is the unobserved selection bias. Since we believe that the selection bias is positive and we are given the observed treatment effect, we can see that the unobserved treatment effect is smaller than the observed treatment effect that there was a tripled amount in the likelihood that a customer that sees the advertisement will buy beer.

5. If we do this, we are making an assumption about the selection bias. So then yes, you can derive the treatment effect. For example, if I assume that the difference in the beer/soft drink consumption of a stadium visitor and a Publix visitor is B, then true treatment effect (unobserved treatment effect would be

$$\begin{aligned}\mathbb{E}[Y_{1i}-Y_{0i}|A_i=1]=&\mathbb{E}[Y_i|A_i=1]-\mathbb{E}[Y_i|A_i=0]-B\\ =&60\%-20\%-B\\ =&40\%-B\end{aligned}$$

6. The treatment effect will simply be our given observable treatment effect minus this assumed selection bias. This estimate definitely depends on our assumption from the previous part. If our assumption about beer/soft drink consumption between a stadium visitor and a Publix visitor decreases (e.g. by a factor of 2), then our estimate of the treatment effect will increase (e.g. $40\%-B/2>40\%-B$). If this assumption about the difference consumption is 0 ($B=0$), then the estimate of the treatment effect would be equal ($40\%-0=40\%$) to the observable statistics that the advertisement tripled the likelihood of a customer choosing beer. Clearly, given an observed effect, the true treatment effect is directly related to the negative of the selection bias, and given a selection bias, the true treatment effect is directly related to the observed effect.

7. In general, we could improveme the initial estimation by quantifying the selection bias in some way and incorporating this into our estimation. We could also improve the initial estimate by trying to get people who do not normally go to football games to go to a football game and see the advertisement. This way, selection bias would be reduced since the groups are more randomly assigned than before.

Question 2: Delta Method and Asymptotic Distribution: [6pts]

- ***Exercise 5.4*** *Find the moment estimator $\widehat{\mu}_3$ of $\mu_3=\mathbb{E}y_i^3$, and show that $\sqrt{n}(\widehat{\mu}_3-\mu_3)\xrightarrow{d}N(0,v^2)$ for some v^2. Write v^2 as a function of the moments of y_i.*

  First, we assume that y is not multi-dimensional since it is not in bold. The moment estimator $\widehat{\mu}_3$ of $\mu_3=\mathbb{E}(y_i^3)$ is

$$\frac{1}{n}\sum_{i=1}^{n}y_i^3.$$

  Now, if we assume that $\mathbb{E}||y^3||<\infty$, then Theorem 5.8.2 gives us that $\sqrt{n}(\widehat{\mu}_3-\mu_3)\xrightarrow{d}N(0,v^2)$ where

$$\begin{aligned}v^2=&\mathbb{E}((y^3-\mu_3)(y^3-\mu_3))\\ =&\mathbb{E}((y^3-\mathbb{E}(y^3))(y^3-\mathbb{E}(y^3)))\\ =&\mathbb{E}(y^6-2y^3\mathbb{E}(y^3)+(\mathbb{E}(y^3))^2)\\ =&\mathbb{E}(y^6)-2(\mathbb{E}(y^3))^2+(\mathbb{E}(y^3))^2\\ =&\mathbb{E}(y^6)-(\mathbb{E}(y^3))^2.\end{aligned}$$

- ***Exercise 5.6*** *Suppose $\sqrt{n}(\hat{\mu}-\mu) \xrightarrow{d} N(0,v^2)$ and set $\beta=\mu^2$ and $\hat{\beta}=\hat{\mu}^2$.*

  1. *Use the Delta Method to obtain an asymptotic distribution for $\sqrt{n}(\hat{\beta}-\beta)$.*
  2. *Now suppose $\mu=0$. Describe what happens to the asymptotic distribution from the previous part.*
  3. *Improve on the previous answer. Under the assumption $\mu=0$, find the asymptotic distribution for $n\hat{\beta}=n\hat{\mu}^2$.*
  4. *Comment on the differences between the answers in parts 1 and 3.*

  1. Since $g(u)=u^2$ is continuously differentiable in a neighborhood μ, then by Theorem 5.10.3 (Delta Method), we have that $\sqrt{n}(\hat{\beta}-\beta) \xrightarrow{d} N(0,4\mu^2v^2)=N(0,4\beta v^2)$ since $G(u)=\partial/\partial u(u^2)=2u$ and $(2u)^2=4u^2$.
  2. If $\mu=0$, then the asymptotic distribution for $\sqrt{n}(\hat{\beta}-\beta)$ becomes $N(0,0)$, which is a degenerate distribution, i.e. the underlying random variable X equals 0 with probability 1.
  3. If $\mu=0$, then $\sqrt{n}\hat{\mu} \xrightarrow{d} N(0,v^2)=v^2N(0,1)=v^2Z$ since $Z\sim N(0,1)$. By squaring both sides, we get that $n\hat{\mu}^2=n\hat{\beta} \xrightarrow{d} v^2\chi_1^2$.
  4. The answers in parts 1 and 3 are similar but not the same. In the case when $\mu=0$, the answer for part 3 seems to be more useful since we no longer have to deal with a degenerate distribution. Therefore, the answer to part 3 is more robust than the answer to part 1.

Empirical Problems (25pts):

Question E1: Omitted Variables [15pts]

1. The sample variances are $\sigma_y^2=1.0262, \sigma_{x_1}^2=1.0251, \sigma_{x_2}^2=0.9801, \sigma_{Z_1}^2=1.0338$, and $\sigma_{Z_2}^2=1.0065$, and the sample covariance matrix is
$$\begin{bmatrix} 1.0262 & 0.2015 & 0.1012 & 0.3705 & 0.0141 \\ 0.2015 & 1.0251 & -0.0017 & 0.4203 & 0.0295 \\ 0.1012 & -0.0017 & 0.9801 & 0.0079 & 0.3876 \\ 0.3705 & 0.4203 & 0.0079 & 1.0338 & 0.6299 \\ 0.0141 & 0.0295 & 0.3876 & 0.6299 & 1.0065 \end{bmatrix}.$$

2. a) The regression coefficient when regressing Y on only x_1 calculated by using only covariances and variances is
$$\frac{\mathrm{Cov}(x_1,Y)}{\mathrm{Var}(x_1)}=\frac{0.2015}{1.0251}=0.1966.$$

   b) The regression coefficient when regressing Y on only x_1 using standard regression commands is 0.1966, which is essentially the same number as calculated above.

   c) The coefficient is biased by the effects of the omitted variables. When all of the variables are included in a linear regression model, the regression coefficient for x_1 becomes -0.0900. The coefficients are especially biased by the significant covariance of x_1 and Z_1.

   d) Let the model estimate for the previous work be $Y=\tilde{\beta}_0+\tilde{\beta}_1x_1$ and the model estimate for a more robust model be $Y=\hat{\beta}_0+\hat{\beta}_1x_1+\hat{\beta}_2Z_1$. (We only add Z_1 in the more robust model since it is the variable that has the most significant effect on both x_1 and Y, indicating that it is the variable that is most likely inducing the bias in the coefficient estimate of x_1.) Then, the theoretical bias is $\hat{\beta}_2\tilde{\delta}$ where $\tilde{\delta}$ is the regression coefficient from regression Z_1 on x_1 since $\tilde{\beta}_1=\hat{\beta}_1+\hat{\beta}_2\tilde{\delta}$. Therefore, the theoretical bias of this $\tilde{\beta}_1$ is 0.1370, which is equivalent to the product of the regression coefficient of Z_1 when it is regressed on x_1 (0.4100) and the regression coefficient of Z_1 when Y is regressed on x_1 and Z_1 (0.3341).

3. a) Completed in ECON_7022_Problem_Set_2.R file.

   b) Note that in the design matrix X, the vector of ones is the last column.

$$X'X = \begin{bmatrix} 5480.7500 & -8.5135 & 2246.4983 & 158.0292 & -43.1976 \\ -8.5135 & 5241.2882 & 43.1301 & 2072.3314 & -97.1913 \\ 2246.4983 & 43.1301 & 5527.8084 & 3367.0424 & 72.4113 \\ 158.0292 & 2072.3314 & 3367.0424 & 5381.0453 & -16.8654 \\ -43.1976 & -97.1913 & 72.4113 & -16.8654 & 5347 \end{bmatrix}.$$

   c)

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} -0.0900 \\ 0.3229 \\ 0.7403 \\ -0.5710 \\ 0.0148 \end{bmatrix}.$$

   The following code is used to compute this:

```
Y <- m %>% pull(Y)
beta_hat <- solve(Xs) %*% t(X) %*% Y
beta_hat.
```

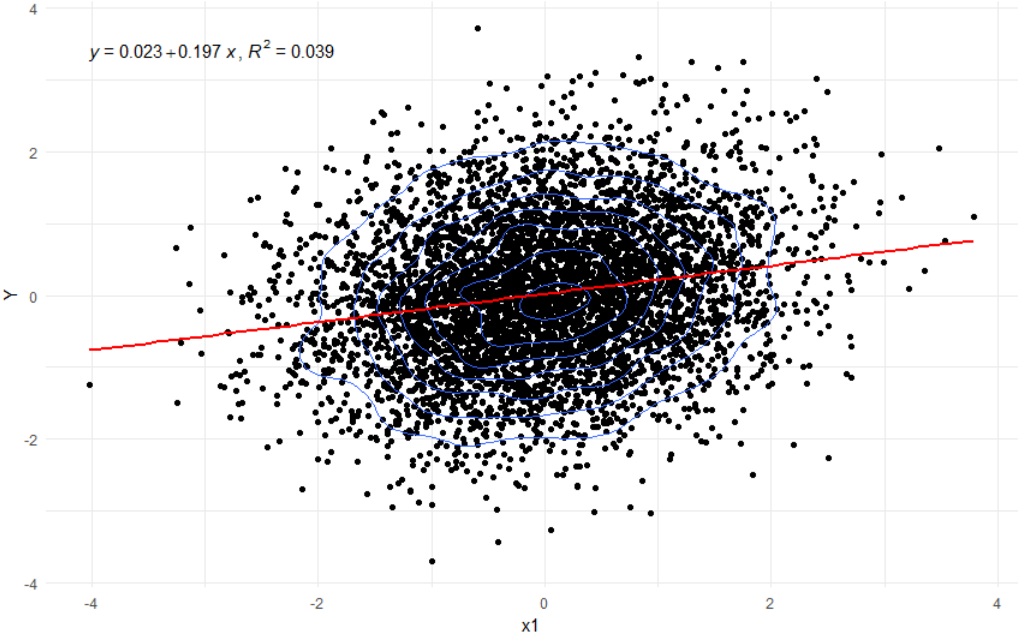
   d) Completed in ECON_7022_Problem_Set_2.R file.

   e) Completed in ECON_7022_Problem_Set_2.R file, but the regression coefficients are clearly the same.

   f) Completed in ECON_7022_Problem_Set_2.R file, but the regression coefficients are clearly different, especially since there are more coefficients for the multivariate model.

4. a) The CEF for $Y|x_1$ is

$$\mathbb{E}[Y|x_1] = 0.0230 + 0.1966x_1.$$



   b) When comparing points 2, 3, and 4, I observe that the residual values are all positive. For the fitted model, the omission bias is positive because the correlation between Y and x_1 is positive and the coefficient for x_1 is positive. The positive residuals on these points indicate that the positive bias in the model has caused the model to miss predict these values in all in the same direction.

Question E2: IV for starters (E1 cont'd) [10pts]

- Completed in ECON_7022_Problem_Set_2.R file. The coefficient estimates are different for the truncated and full models because different sets of variables are considered in each model. However, the estimates for β_0 (0.0230 and 0.0249, respectively) and β_1 (0.1966 and 0.1967, respectively) are very similar for both models since x_1 and x_2 are barely correlated. In addition, the variables that are omitted from both models have approximately the same level of significance for both models. (Note that $\hat{\beta}_2 = 0.1036$ for the full model.)
- I would expect the exclusion of Z_2 to have almost no effect on the coefficient estimate of x_1 since x_1 and Z_2 are barely correlated. Now, since x_2 and Z_2 are significantly positively correlated, I would expect that the exclusion of Z_2 to inflate the coefficient estimate of x_2 if Z_2 were positively correlated with Y. However, since Z_2 is barely correlated with Y, I would expect that the exclusion of Z_2 would have little to no effect on the coefficient estimate of x_2.
- If one of the x variables were endogenous with Y, we could only Z_2 as an instrumental variables for x_2. First of all, we can use Z_2 as an instrumental variable because it is not correlated with the dependent variable Y. We could only use Z_2 as an instrumental variable for x_2 because there is positive correlation between x_2 and Z_2, but no correlation between x_1 and Z_2. If we were to use Z_2 as an instrumental variable for x_2, we would have to assume Z_2 is correlated with x_2 given the other covariates and that Z_2 is exogenous with Y, i.e. that Z_2 is not correlated with the error term of the original regression model. This is the exclusion restriction.

Project 1 (20pts)

Note: I will be changing my following project idea completely and resubmitting it tomorrow since there are too many issues with this current project.

- The research question that I would like to pursue is "How do the 3 pointers made per game, 2 pointers made per game, and free throws made per game of two NBA teams predict the total points scored in a game between these two teams?" This question can be of interest to sport books who set the over/under lines on total points for each NBA game and/or to the gambler who is seeking to make money off betting these over/under lines. The estimation equation for this question is

  total points scored = β_0 + β_1*(3 pointers made per game of home team) + β_2*(3 pointers made per game of away team) + β_3*(2 pointers made per game of home team) + β_4*(2 pointers made per game of away team) + β_5*(free throws made per game of home team) + β_6*(free throws made per game of away team) + ϵ.

- The unit of observation will be one set of average made attempts in the three categories for each team going into a game and the total points scored between the teams in said game. I have scraped NBA data from the past 11 years, and there are 1,230 NBA games per year, so I have approximately 13,530 observations. Yes, all x-variables vary at the same level of my units of observation.
- There are many unobserved variables that may affect an NBA game's win total. Some of these include injuries, total distance traveled by each team in the past week, average number of turnovers by each team, the skill level of all of the individual players, the playing style of each team, home court advantage, etc. With lots of time and effort, some of these variables, like total number of injuries by each team going into a game and total distance traveled by each team in the past week, could become observed. However, most of these variables are very difficult to quantify, and thus must be considered unobserved. However, most of these unobserved variables should not cause much of a problem since they are likely highly correlated to the current covariates. I will try to verify this claim in the paperette. I think a very relevant, but currently unobserved variable, is the total distance traveled by each team in the past

week. I think the amount that a team has traveled in the days before a game can cause that team to play worse than normal, and thus, score fewer points. However, this effect may by counteracted by the amount of extra points that its opposing team may score do the team's sloppier play. As mentioned before, with some extra effort, I may be able to scrape travel data so that this variable is no longer unobserved. In general, unobserved variables that cannot be omitted from the model will bias the coefficient estimates for the observed variables in the model, which is a problem if we are trying to measure true relationships. However, I do not think that this will be much of a problem for me since I believe I already have the most significant/predictive variables included in my model.

- There are no other sources of endogeneity like simultaneity, selection bias, or measurement errors that I can see since the the relationship between the dependent and independent variables is not overly complex, we are using all teams over the past 11 years, and observations are scraped from an official website.

- There is no real experiment to conduct here since all observations have already been completed. As discussed before, the only source of endogeneity that I am worried about is omitted variable bias, but I think I will be able to overcome this. Ideally, I would have observations of various types so that the OLS estimator would be as robust as possible. By this, I mean that I would like to have sufficient observations of some extreme scenarios like a team that makes a lot of 3 pointers and free throws per game against a team that does not make many free throws per game but makes a lot of 2 pointers per game, etc. Basically, this would be an attempt to reduce selection bias as much as possible, but I feel comfortable that with the amount of data that I have, selection bias is not very troublesome here.

- I think there are a few limitations to my project, the biggest of which is omitted variable bias. As mentioned before, I think the biggest source of this bias can potentially be observed with a good amount of time put into scraping data from the Internet, so in the end, it may not be a big problem. I also think that in general I am limited by time since there are so many different combinations of covariates and interactions among covariates that I could use to potentially predict an NBA game's point total. I will obviously not be able to attempt every possible combination, but with some intuition, I can reduce the number of combinations that I try out to only the ones that are likely to be significant.

- As already mentioned, I can improve my modeling strategy by spending more time on scraping information about distance traveled by each team. I can also investigate the covariance among all potential omitted covariates and my current model's covariates to justify why I am only using the covariates that I am currently using.

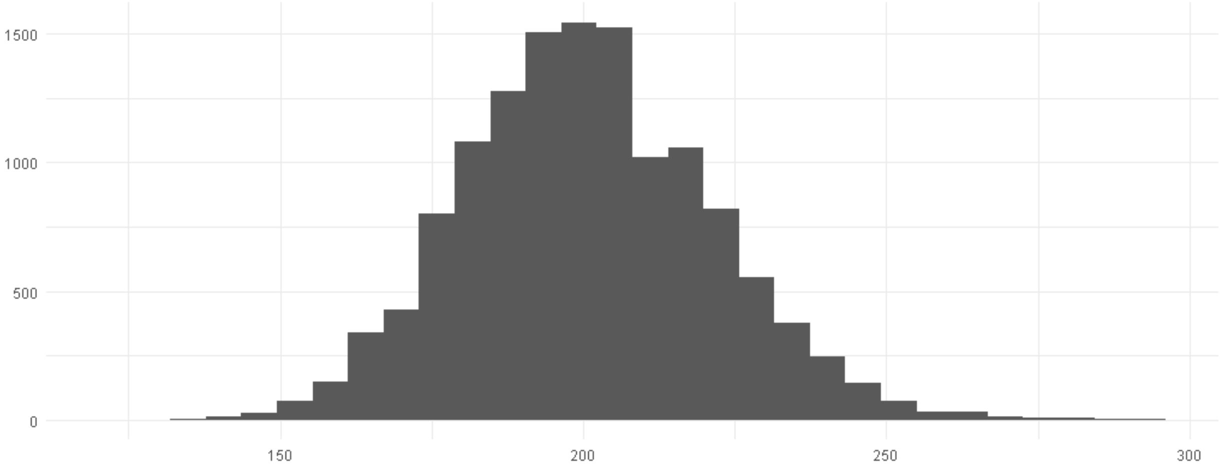
Project 2: Data Cleaning (15pts)

1. No proxies or simulated data were needed here since I scraped the data from the Internet already.

2. The following images are some of the visualizations I created to start looking at my data. I plotted the distributions of the dependent variable and each of the individual covariates. In addition, I plotted the conditional expectation function of each individual covariate with the dependent variable. Other methods for data cleaning were performed locally.

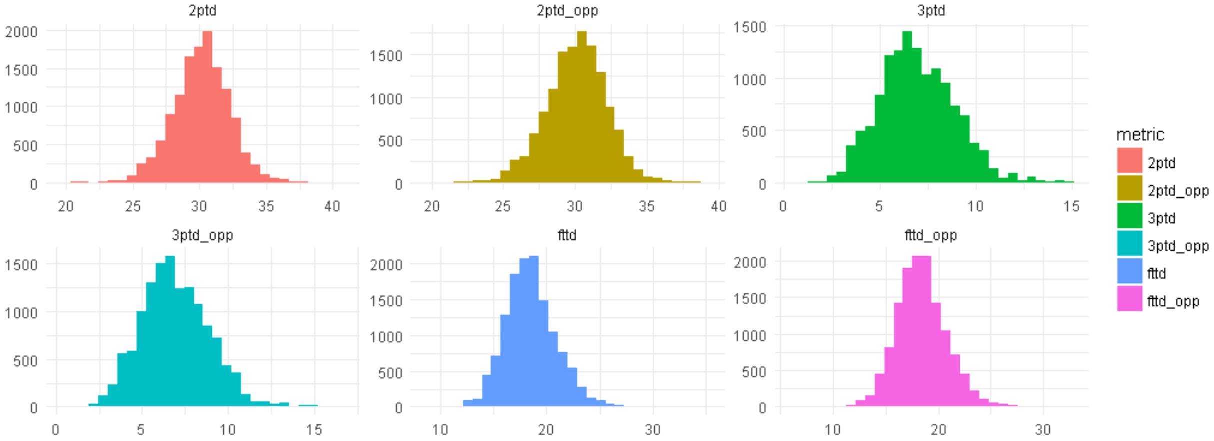
Distribution of Response Variable

pts_total



Distribution of Covariates

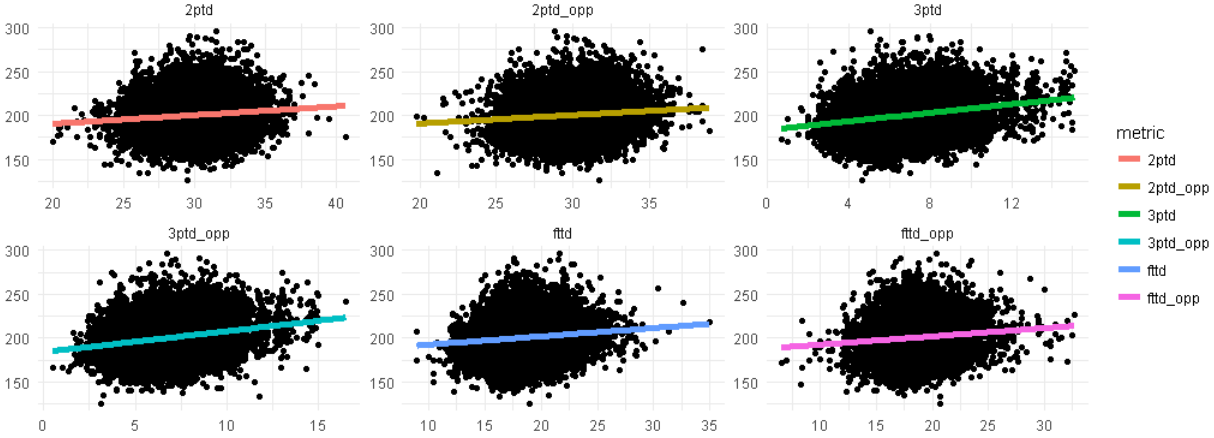
pts_total ~ 2ptd + 3ptd + fttd + 2ptd_opp + 3ptd_opp + fttd_opp



*'td' stands for 'to date'.

Covariates VS. Response Variable

pts_total ~ 2ptd + 3ptd + fttd + 2ptd_opp + 3ptd_opp + fttd_opp



*'td' stands for 'to date'.

3. It appears as if the distribution for the dependent variable is approximately normal. This is great! And it appears as if there are no outliers, so outlier removal will not be necessary.

4. It appears as if each of the distributions is close to normal with only some minor skewing. Once again, this is great! This means that I will likely not have to assert any untrue assumptions in my model. It appears as if there are almost no outliers in any of the distributions, so outlier removal will not be necessary.

5. I learned that all of my data is approximately normally distributed as hoped and that outlier removal is not necessary. Most of the work for the data is put into scraping it, and from there, the manipulation is fairly easy.

6. No, there are no corrections needed since all of the data has been scraped from an official website and follows distributions that were expected.