Applied Linear Regression: Analysis of Experimental Training Program Data and Prediction

Jay Bendre, Grant Gambetta, and Collin Kennedy

November 2021

1 Abstract

Using both observational and experimental data from economist Robert LaLonde's paper, "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", we explore the application of linear regression in quantifying the effect of participation in a jobs training program. Whereas LaLonde's focuses on comparing estimates obtained using *both* experimental and observational data, we delve deeper into verifying the statistical assumptions required to appropriately apply regression to this problem using the experimental data. We then go a completely different route, and attempt to (with minimal success) develop a model for predicting an individual's earnings using the observational data.

2 Introduction

The conduction of causal inference in econometrics has led to some of the most major and eye-opening discoveries in the field of economics. From medical marijuana laws being found to reduce opioid-related addictions and deaths [8], the conclusion that the implementation of a national mask mandate could have reduced deaths related to COVID-19 by up to 47% prior to May 2020 [2], to John Donohue and Steven Levitt's finding that legalized abortion in the United States accounts for up to 50% of the reduction in crime witnessed during the 1990s [5], it is evident that econometrics has found a place in the world of revealing causal relationships with observational data.

It essentially goes without saying that econometricians intend for the results of their analyses of typically observational data to reproduce the results of randomized controlled experiments. Otherwise, why bother?

In Robert J. Lalonde's "Evaluating the Econometric Evaluations of Training Programs with Experimental Data", which was published in the *American Economic Review* in 1986, he explores this issue. More specifically, he sought

to evaluate the efficacy (or potentially lack thereof) of using econometric methods to determine the effect of participating in an employment program on one's future earnings. The data used in his study, and now our project, is unique because it contains both experimental and observational data [4]. Greater detail about the nuances and specifics of the data is provided later in the report.

Whereas Lalonde's paper focused on quantifying the specification error that can occur when applying econometric methods on non-experimental (i.e., observational) data, we focus more on exploring the potential effect that assignment to the jobs program has on earnings. We then take our exploration of the data in a completely different direction and consider how well can we *predict* an individuals earnings given whether or not they participated in the jobs program, along with attributes related to age, years of education, and ethnicity.

Our report is organized as follows: First, we conduct some exploratory data analysis in order to verify that our assumptions of linear regression are met and that it is appropriate to model the relationship between earnings and participation in the NSW employment program with a linear model. Next, we perform model fitting to estimate the average treatment effect of participating in the training program on earnings. We then shift gears, and consider the comparison of predictive models trained on both the experimental data and observational data using RMSE and R^2 metrics.

2.1 Data

We obtained the data from "Evaluating the Econometric Evaluations" which was made publicly available and can be found in the *qte* package in R `https://rdrr.io/cran/qte/man/lalonde.html`. To briefly summarize how this data was created and collected, during the 1970s the National Supported Work Demonstration was a federally sponsored employment training program designed to temporarily employ disadvantaged workers. What was unique however, was that workers were *randomly* assigned to training positions. The Manpower Demonstration Research Corporation (MDRC) operated the program and collected data on earnings and demographics of the workers in both the training and control groups. Lalonde also then obtained data from the *Panel Study of Income Dynamics* and the *Current Population Survey-Social Security Administration File*. This was/is the sort of data that an econometrician would analyze if randomly-assigned experimental data was not available, and the overall objective of his study was to compare the estimates from analyzing this data with the estimates obtained from analyzing the experimental data. All the data LaLonde used in his research can be found in the *qte* package, and we make ample use of both the experimental and observational data for the purposes of analyzing the treatment effect and developing a predictive model, respectively.

The data consists of control and testing datasets for people undergoing national supported work demonstration. Table 1 explains the types of the variables used in this paper.

3 Methods & Results

3.1 Exploratory Data Analysis

For our exploratory data analysis, we began by creating a series of plots to understand the relationships that were present in our earnings data. First, we created a boxplot (Figure 1) to determine whether the employment training program had an effect on peoples' earnings in 1978. From looking at the plot, we found that in 1978, the earnings of the treatment group were higher than the earnings of the control group. The treatment group had median earnings of roughly $2,200 and the control group had median earnings of roughly $1,300. This implies that people who completed the employment training program generally had higher earnings in 1978. Also, we created two histograms (Figures 2 and 3) to analyze the distribution of earnings of the control and treatment groups in only 1978. We observed that the majority of people in the control group had earnings anywhere between $1,000 to $8,000, with a few people having earnings in the $10,000 to $20,000 range. For the treatment group, we noticed that the majority of people had earnings between $5000 and $20,000 with some earnings as high as $30,000. Therefore, this also shows that people who participated in the employment training program had higher overall earnings in 1978 than those who did not participate.

Luckily, this dataset was in a highly user-friendly form in the *qte* package, however, we still needed to perform some data preprocessing. First, we checked to ensure there were no missing values, and converted all factor variables that had been interpreted by R as integers into factors. Lastly, we dropped the *id* column because it served no purpose in our analysis and could potentially cause problems when fitting our predictive model.

3.2 Assumptions and Model Fitting

We began our model fitting procedure by fitting the model specified by LaLonde in "Evaluating the Econometric Evaluations":

$$Y_{re78} = intercept + treat + age + age^2 + education + black + hispanic + nodegree$$

To ensure that LaLonde had not overlooked any worthwhile transformations of the response variable to dampen potential heteroskedasticity of the residuals, we also considered square root, inverse, and log transformations of the response variable (*re78, real earnings in 1978*). We then inspected each of the four model's respective diagnostic plots, paying special attention to the *Residuals vs Fitted Values* and *Normal QQ* plots (see Figure 4 and Figure 5).

While no transformation seemed perfect, the square root transformation appeared to best dampen the variance in the residuals (see Figure 4c). Meanwhile, the inverse and log transformations seemed to do much more harm than good. Because the square root transformation appeared to produce similar diagnostic

plots as the specified model with no transformations (see Figure 4c and Figure 5), we selected this as our model:

$$Y_{re78} = intercept + treat + age + age^2 + education + black + hispanic + nodegree$$

before continuing with verifying model assumptions. We then supplement our visual inspection diagnostics with some statistical tests.

First, we wanted to determine if the constant variance assumption holds. This assumption is perhaps the most important, as regression coefficients estimated with *heteroskedastic* data will have larger sampling variability, which will in turn result in larger standard errors and negatively impact inference of the coefficients.

While the residuals for the most part appear to be evenly and randomly dispersed about 0 over the entire domain of the *Residuals vs Fitted Values* plot seen in 5, there also appears to be a slight pattern in a subsection of the plot. In order to verify our intuition with formal test, we perform a White test, which can be used to determine whether there is any heteroscedasticity present [1]. The null hypothesis H_0 says that the variances of the errors are equal, while the alternative H_a suggests that heteroskedasticity is present. We *fail to reject* the null hypothesis at $\alpha = .05$, and conclude that that variance is homoskedastic (p-value = .6575).

Turning our attention to the exogeneity assumption, such that

$$\mathbb{E}(\epsilon|X_i) = 0,$$

we initially noted that this assumption is likely met due to the linearity that is present in the *Residuals vs. Fitted Values* plot in Figure 5. To be thorough, we employed the Wu-Hausman test (also known as that Durbin-Wu-Hausman test). The null hypothesis H_0 states that the regressors are exogenous ie. the residuals obtained and the X variables with which the model was fitted are uncorrelated with each other. The alternative, H_a states that there is *endogeneity* [3].

A violation of the exogeneity assumption (i.e., there is endogeneity) is indicative of omitted variable bias, and means that the estimated regression coefficients are no longer the minimum variance *unbiased* estimator. We *fail to reject* the null hypothesis, and conclude the regressors are exogenous (p-value = .4765).

The results of both tests (White test for constant variance and Wu-Hausman test for Exogeneity), the hypothesis and the p-values are presented in the Table 2.

The regression output of our final model can be seen in Table 3. Implications and interpretation of the fitted coefficients will be discussed later in the report.

3.3 Prediction

For the prediction part of this project, we had the goal of building a model that could predict an individuals earnings based on whether they participated in the

jobs program, as well as other factors such as ethnicity, years of education, and age. To accomplish this, we began by preprocessing the observational data in the same way as the experimental data. Then, we fit an initial model on the observational data with 1978 earnings (re78) as the response variable and the rest of the variables as predictors. This model had an R_a^2 of 0.585.

After fitting the initial model on the observational data, we split the data into a training and testing set (70% train, 30% test) and ran a forward selection procedure with the null model as the starting model and the two way interaction model as the full model. The model obtained from the forward selection is shown below:

$$Y_{re78} = intercept + re75 + education + re74 + age + u74 + married + re75 : re74$$

$$+re74 : age + age : married + education : u74 + age : u74$$

We then fit that model on training dataset and experimented with three different resampling methods to evaluate the model fit: k-fold cross validation, leave one out cross validation, and bootstrap. The resulting in-sample R^2 and RMSE for each of these evaluation methods can be found in Table 4. The out-of-sample R^2 and RMSE for this model were 0.54 and 11,174, respectively.

4 Discussion

There are perhaps three important things to note about the results of our fitted model.

First, holding all else constant, the estimated average treatment effect of participating in the jobs program is \$1672.43 (p-value $< .0001$). In other words, individuals that were assigned to the jobs training program earned about \$1600 more on average compared to individuals that were *not* assigned to the training program. Adjusting for inflation, this amounts to about \$4905.36 in 2021 dollars [7].

Another noteworthy finding is the effect of education on real earnings in 1978. According to our findings, each year of education is associated with approximately a \$379.50 increase in earnings (p-value $< .05$), holding all else constant (\$1,087.73 in 2021 dollars).

Lastly, our findings also highlight racial economic disparities that were highly prolific during the 1970s. Our model indicates that black individuals earned about \$2,208 less than their non-black counterparts (p-value $< .05$). Adjusted for inflation, this amounts to \$6,328.59 in 2021 dollars [7].

The specification of our experimental model is not without its potential drawbacks. For example, our estimates do not account for differences in assigned profession. For example, some individuals that participated in the employment training program worked in construction (primarily males) while others worked at gas stations and printing shops [4]. The potential differences in earnings between these professions is masked by our single estimate of the treatment effect.

The assumptions of a normal error model may also be violated (see Figure 3), since our residuals do not appear to follow a normal distribution. This does not completely invalidate our analysis, but it is worth pointing out. Our model may be mispecified, and the power of our statistical tests may be negatively impacted as a result. With that said, as pointed out by Amand Schmit and Chris Finan in their work "Linear Regression and the Normality Assumption," violations of the normality assumption do not typically impact results [9]. Violations of homoskedasticity and are often far more problematic. Thankfully, our model did not demonstrate any statistically significant departures from constant variance.

While we were not able to produce a model with high predictive accuracy, there are several tasks that could be done in the future to improve the prediction capabilities of our model. First, obtaining more data could have a major effect on the prediction accuracy of our model because our observational dataset consisted of only *2,675 samples*. This meant we trained the model on only *1,872 samples*, which is probably not enough given the noise that is present in the data. Also, since the dataset contains quite a bit of noise, more feature engineering could help improve the quality of the data and therefore increase the prediction accuracy of the model. Lastly, using more advanced machine learning techniques that are more suited for prediction such as random forest, gradient boosted trees, bagging, or neural networks could lead to increased prediction accuracy. All of these new approaches could be of great use and would be worth exploring in the future.

4.1 Conclusion

The results of our analysis are consistent with that of LaLonde's, and participation in the employment training program had a significant effect on a given individual's earnings 3 years following participation in the program, compared to individuals in the control group. And, for the most part, our model is robust to concerns about misspecification. Unsurprisingly, we found that continued education is associated with significantly higher earnings, which is consistent with much of the economic literature on the relationship between schooling and human capital accumulation. We also found that black individuals made significantly less money than their non-black counterparts, even if they had participated in the employment training program, and this racial disparity in earnings has actually worsened in recent years[6]. This has major implications for the US economy that we could not explore here, and potential remedies to this pressing issue should continue to be proposed and analyzed.

As for as prediction is concerned, we failed to produce a model that could predict an individual's earnings with a high degree of accuracy, based on out-of-sample R^2 and $RMSE$. With that said, being able to accurately predict earnings could be of great use to policy makers and employment program administrators to better facilitate and assist individuals looking to upskill in our dynamic economy, so a thorough investigation into the application of more advanced prediction methods could be of great use down the road.

5 Appendices

5.1 Appendix 1: Figures and Tables

5.1.1 Figures

This section shows all the figures and plots that were being referred to in any earlier sections of the paper.

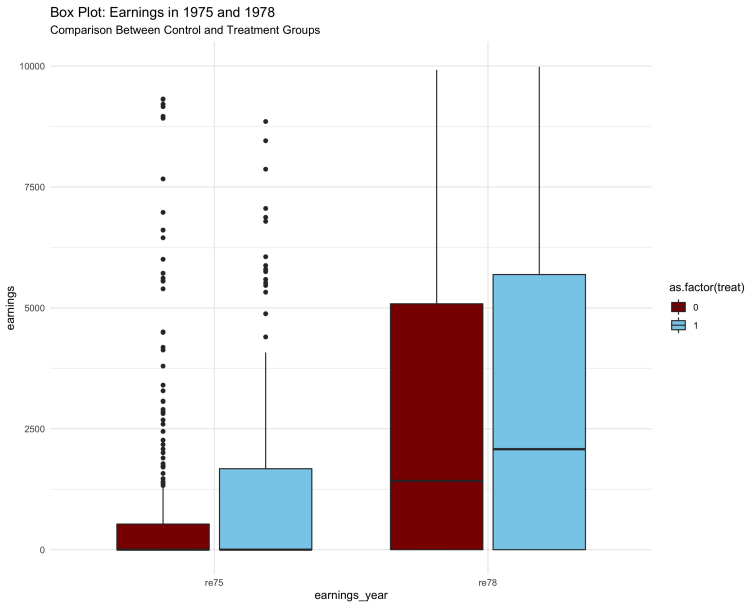


Figure 1: Earnings in 1975 and 1978 based on whether people participated in the training program.

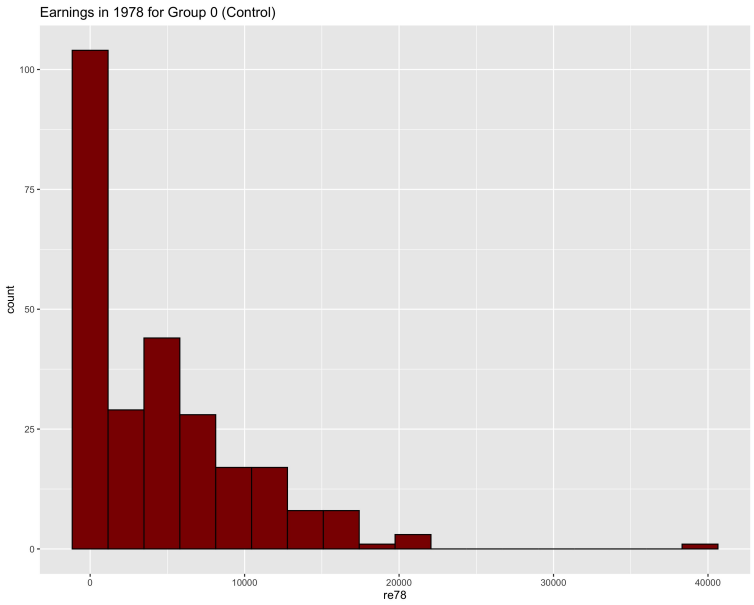


Figure 2: Earnings in 1978 for people who did not participate in the training program.

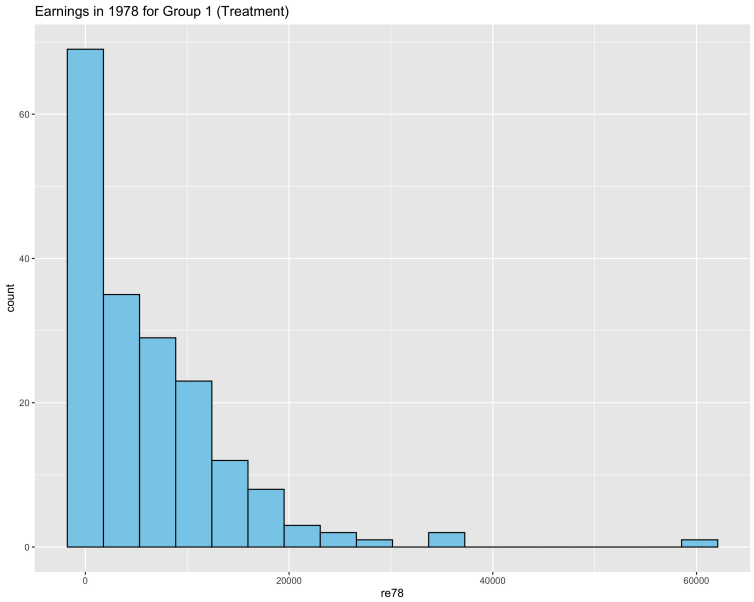
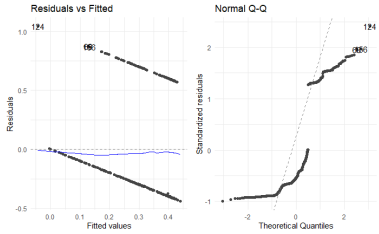
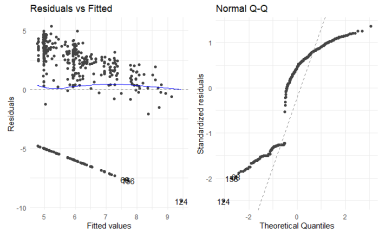


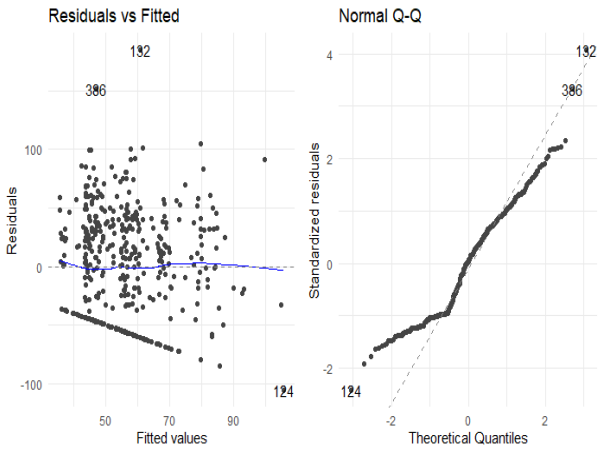
Figure 3: Earnings in 1978 for people who participated in the training program.



(a) Diagnostic Plots for $1/Y$



(b) Diagnostic Plots for $log(1+Y)$



(c) Diagnostic Plots for $\sqrt{Y}$

Figure 4: Shows all the transformations considered and their respective diagnostic plots.

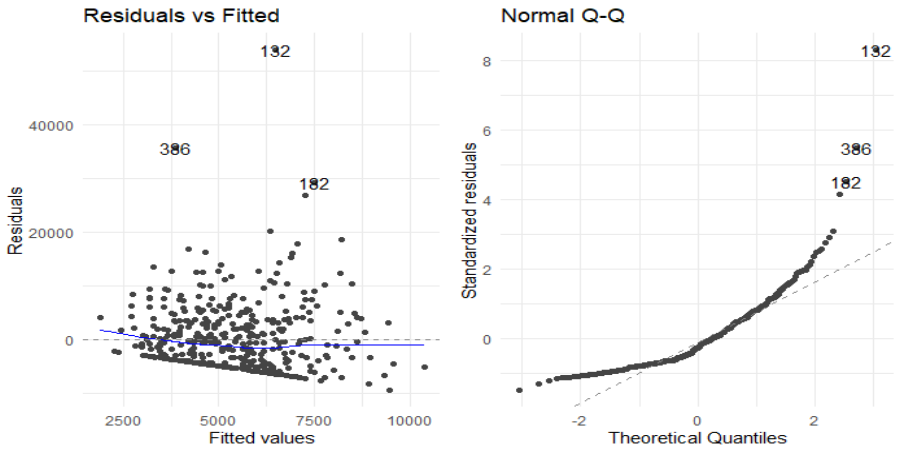


Figure 5: Shows the Fitted vs Residuals and the QQ plots of the model.

5.1.2 Tables

This section shows all the tables that were being referred to in any earlier sections of the paper.

Variable Name	Variable Type	Description
Age	Quantitative	Age in years
Educ	Quantitative	Years of schooling
Black	Qualitative	Indicator variable for blacks
Hispanic	Qualitative	Indicator variable for Hispanic
Married	Qualitative	Indicator variable for marital status
Nondegr	Qualitative	Indicator variable for high school diploma
u74,u75	Qualitative	Indicator for unemployed in 1974 and 1975
Re74,Re75,Re78	Quantitative	Real earnings in 1974,1975 and 1978
Treat	Qualitative	Indicator for treatment status

Table 1: The Variable Description of the 'lalonde' dataset

Tests	H_0 & H_a	P- Values
White Test	H_0 : Error Variances are equal H_a : Non Constant Error Variances	**0.6575**
Wu-Hausman Test	H_0 : Regressors are Exogenous ie. uncorrelated to each other H_a: Regressors are correlated to each other	**0.4765**

Table 2: The Hypothesis Test Results

Dependent Variable: re78	Experimental Model
Treat	$1{,}672.427^{***}$
	(637.310)
Age	184.031
	(266.802)
Age^2	-2.164
	(4.399)
Education	379.502^{*}
	(229.088)
Nodegree	-101.666
	(994.792)
Black	$-2{,}208.032^{*}$
	(1,167.230)
Hispanic	128.446
	(1,548.397)
Intercept	-521.888
	(4,677.166)
R^2	0.048
Adjusted R^2	0.033
Residual Std. Error	6,520.820 (df = 437)
F Statistic	3.171^{***} (df = 7; 437)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

Table 3: Summary of the Observational Model

Testing Methods	R^2	RMSE
Leave One Out Cross Validation	**0.61**	10,933
Bootstrap	**0.608**	9,733
5 Fold Cross Validation	**0.623**	9,474

Table 4: Prediction Re-sampling Test Results

5.2 Appendix 2: R Codes

```
# Dataset Description:

# http://sekhon.berkeley.edu/matching/lalonde.html description of the dataset
#AER article where the dataset is from https://business.baylor.edu//scott_cunningham
#/teaching/lalonde-1986.pdf

library(qte)
library(AER)
library(dplyr)
library(ggfortify)
library(ggplot2)
library(tidyr)
library(PerformanceAnalytics)
library(gridExtra)
library(MASS)
library(stargazer)
library(caret)
work_training_df = lalonde.exp
head(work_training_df)


# Performing Exploratory Data Analysis

# Summarising the dataframe based on the treatment variable
# 0 - Didnt go through the program 1 - went through the program

work_training_df %>%
  group_by(treat) %>%
  summarise(count = n())


#Look at earnings for both groups in 75 and 78
plot_data = work_training_df %>%
  pivot_longer(cols = c(re75,re78), names_to = "earnings_year",values_to = "earnings")

plot_data %>%
```

```
  group_by(treat,earnings_year) %>%
  summarise(average_earnings = mean(earnings))

plot_data

ggplot(data = plot_data,
mapping = aes(x = earnings_year, y = earnings, fill = as.factor(treat)))+
  geom_boxplot()+
  theme_minimal()+
  ylim(c(0, 20000)) +
  ggtitle("Box Plot: Earnings in 1975 and 1978",
  subtitle = "comparison between control and treatment groups")


# Checking for missing values in the data
colSums(is.na(work_training_df))

# Looking for all the classes that need to be converted into factors
#depending upon them being indicators
sapply(work_training_df, class)


# convert variables to factors
work_training_df <- work_training_df %>% mutate_if(is.integer, as.factor)
sapply(work_training_df, class)


# Instead of storing all the levels on years of education ranging
#from 3-16 , in this study we maintain 3 primary level ie. 0 -
#Elementary 1 - High School and 2 - College

encode <- function(col) {
  if (col %in% c(3:8)) {
    return(0)
  } else if (col %in% c(9:12)) {
    return(1)
  } else {
    return(2)
  }
}

work_training_df$education =  as.factor(sapply(work_training_df$education,encode))

# Seeing the effect of having a high level education on the earnings of the people
ggplot(data = work_training_df, mapping = aes(x = treat, y = re78, fill = education))+
  geom_boxplot()+
```

```
  theme_minimal() +
  ylim(c(0,20000))

# Seeing the distribution of the earnings based on the treatment variable
filtered_df_0 <- work_training_df %>% filter(treat == 0)
filtered_df_1 <- work_training_df %>% filter(treat == 1)

p1 <- ggplot(data = filtered_df_0, aes(re78)) +
  geom_histogram()

p2 <- ggplot(data = filtered_df_1, aes(re78)) +
  geom_histogram()

grid.arrange(p1, p2, nrow=1)

# Working on getting the appropriate transformation
# Transformations considerd were as follows:
# 1. Inverse of Y + 1
# 2. Square Root of Y
# 3. Logarithm of Y + 1

work_training_df_exp <- lalonde.exp

sapply(work_training_df_exp %>% dplyr::select_if(is.integer),as.factor)

work_training_df_exp$education =  as.factor(sapply(work_training_df$education,encode))

# Deciding on the sqrt transformation as the graph is most normal like.
work_training_df_exp <- work_training_df %>% dplyr::select(-c(id))
work_training_df_exp$re78.transformed <- sqrt(work_training_df_exp$re78)
hist(work_training_df_exp$re78.transformed, breaks = 30)

# Fitting a model to see whether the performance improves with the transformation

m_test = lm(re78.transformed ~ ., data = work_training_df_exp %>%
dplyr::select(-c("re78")))
m_test.summary = summary(m_test)
m_test.summary
plot(m_test)

# Implementing StepAIC to get a better model than the full model
work_training_df_exp <- lalonde.exp
work_training_df_exp <- work_training_df_exp %>% dplyr::select(-c(id))


sapply(work_training_df_exp %>% dplyr::select_if(is.integer),as.factor)
```

```r
work_training_df_exp$education =  as.factor(sapply(work_training_df_exp$education,encode))

work_training_df_exp$re78.transformed <- sqrt(work_training_df_exp$re78)
hist(work_training_df_exp$re78.transformed, breaks = 30)

# Dataframe being used atm = work_training_df_exp
colnames(work_training_df_exp)

# Selecting the required columns
work_training_df_exp <- work_training_df_exp %>% dplyr::select(-c("re78"))

# Scaling all the variables
work_training_df_exp <- work_training_df_exp %>% dplyr::mutate_if(is.numeric, scale)

# Model 0
m0 <- lm(re78.transformed ~ 1, data = work_training_df_exp)

# Full Model
m.full <- lm(re78.transformed ~ .^2, data = work_training_df_exp)

# Implementing forward selection
forward.aic <- stepAIC(m0, scope = list(lower = ~1, upper = m.full), direction = "both",
k = 2, trace = 0)

forward.aic$anova
forward.aic.summary <- summary(forward.aic)

# Checking for constant variance using the White Test

white_lm(forward.aic)

# Checking for endogenity between the X variables and the residuals using Wu-Hausman Test

ivreg1 <- ivreg(re78.transformed ~ black + treat + education + re75 +
    black:education | u75, data = work_training_df_exp)
wu_hausman_test <- summary(ivreg1 ,diagnostics = TRUE)$diagnostics[8,]

wu_hausman_test

# Model proposed.
m_1 <- lm(re78 ~ treat + age + I(age^2) + education + nodegree + black + hispanic ,
data = work_training_df_exp)

# Summary of the model
summary(m_1)
```

```r
# Perfoming Whites test
white_lm(m_1)

# Performing Wu-Hausman Test
ivreg1 <- ivre(re78 ~  treat + age + I(age^2) + education + nodegree + black + hispanic
| u74, data = work_training_df_exp)

# Extracting the Wu-Hausman Values from the summary
wu_hausman_test <- summary(ivreg1 ,vcov = sandwich, diagnostics = TRUE)$diagnostics[8,]
wu_hausman_test

# Plotting the diagnostic plots using ggplot2
autoplot(m_1, which = 1:2) + theme_minimal()

# Fitting the predictive model on the observational dataset

# Loading the dataset
obs_df <- lalonde.psid

# Feature Engineering
obs_df$treat <- as.factor(obs_df$treat)
obs_df$black <- as.factor(obs_df$black)
obs_df$hispanic <- as.factor(obs_df$hispanic)
obs_df$married <- as.factor(obs_df$married)
obs_df$nodegree <- as.factor(obs_df$nodegree)
obs_df$u74 <- as.factor(obs_df$u74)
obs_df$u75 <- as.factor(obs_df$u75)
obs_df$education <- as.factor(sapply(obs_df$education, encode))
obs_df <- obs_df %>% dplyr::select(-c(id))
head(obs_df)

# Split data into training and testing datasets
smp_size <- floor(0.7 * nrow(obs_df))

## set the seed to make your partition reproducible
set.seed(123)
train_ind <- sample(seq_len(nrow(obs_df)), size = smp_size)

# Getting the training and testing datasets
train_obs_df <- obs_df[train_ind, ]
test_obs_df <- obs_df[-train_ind, ]

# Fitting the model using 1. Leave One Out Cross Validation control
ctrl <- trainControl(method = "LOOCV")
```

```r
model <- train(re78 ~ re75 + education + re74 + age + u74 + married + re75:re74 +
    re75:u74 + re74:age + age:married + education:u74 +
    age:u74, data = train_obs_df, method = "lm", trControl = ctrl)

model

# Extracting the predictions
pred <- predict.train(model, newdata = test_obs_df, level = 0.95)

# Getting the out of sample R^2
postResample(pred = predictions_model1, obs = test_obs_df$re78)

# Fitting the model using 5 Fold Cross Validation
ctrl2 <- trainControl(method = "CV")

model2 <- train(re78 ~ re75 + education + re74 + age + u74 + married + re75:re74 +
    re75:u74 + re74:age + age:married + education:u74
    + age:u74, data = train_obs_df, method = "lm", trControl = ctrl2)
model2

# Extracting the predictions
pred2 <- predict.train(model2, newdata = test_obs_df, level = 0.95)
predictions_model2 = predict(model2,newdata = test_obs_df, level = .95)

# Getting the out of sample R^2
postResample(pred = predictions_model2, obs = test_obs_df$re78)

# Fitting the model using 3. Bootstrapping
ctrl3 <- trainControl(method = "boot")


model3 <- train(re78 ~ re75 + education + re74 + age + u74 + married + re75:re74 +
    re75:u74 + re74:age + age:married + education:u74 +
    age:u74, data = train_obs_df, method = "lm", trControl = ctrl3)
model3

# Extracting the predictions
pred3 <- predict.train(model3, newdata = test_obs_df, level = 0.95)
predictions_model3 = predict(model3,newdata = test_obs_df, level = .95)

# Getting the out of sample R^2
postResample(pred = predictions_model3, obs = test_obs_df$re78)
```

References

[1] John Black, Nigar Hashimzade, and Gareth D. Myles. *A dictionary of economics*. Oxford University Press, 2009.

[2] Victor Chernozhukov, Hiroyuki Kasahara, and Paul Schrimpf. Causal impact of masks, policies, behavior on early covid-19 pandemic in the u.s. *Journal of Econometrics*, 220(1):23–62, 2021.

[3] J. A. Hausman. Specification tests in econometrics. *Econometrica*, 46(6):1251–1271, 1978.

[4] Robert LaLonde. Evaluating the econometric evaluations of training programs with experimental data. *The Journal of Health Economics*, 76(4):604–620, 1986.

[5] Steven Levitt and John Donohue. Impact of legalized abortion on crime. *The Quarterly Journal of Economics*, 116(2):379–420, 2001.

[6] Nick Noel, Duwain Pinder, Shelley Stewart, and Jason Write. "the economic impact of closing the racial wealth gap. "https://www.mckinsey.com/industries/public-and-social-sector/our-insights/the-economic-impact-of-closing-the-racial-wealth-gap", Year = 2021, Note = "[Online; accessed 05-Dec-2021]".

[7] US Bureau of Labor Statistics. Cpi inflation calculator provided by the us bureau of labor statistics. "https://www.mckinsey.com/industries/public-and-social-sector/our-insights/the-economic-impact-of-closing-the-racial-wealth-gap", 2021. [Online; accessed 05-Dec-2021].

[8] David Powell, Rosalie Pacula, and Mireille Jacobson. Do medical marijuana laws reduce addictions and deaths related to pain killers? *The Journal of Health Economics*, 58(2):29–42, 2018.

[9] Amand Schmidt and Chris Finan. Linear regression and the normality assumption. *Journal of Clinical Epidemiology*, 98(4):146–151, 2017.