Data Analysis Stories

Casey Canfield

24 January, 2021

In this file, I will demonstrate how to perform the data analysis stories. This project uses data from:

Ludwig, J., Duncan, G. J., Gennetian, L. A., Katz, L. F., Kessler, R. C., Kling, J. R., and Sanbonmatsu, L. (2012). Neighborhood effects on the long-term well-being of low-income adults. *Science*, 337(6101), 1505–1510.

The data is available at the National Bureau of Economic Research website. You should download the Cell-Level PUF (Public Use Files) for the *Science* paper that was last updated on 9/21/2012.

First, I call any needed libraries.

```
# LIBRARIES
# install.packages() if needed

library(tidyverse) # always
library(haven) # for read_dta
library(car) # for qqPlot
library(Hmisc) # for rcorr
```

Then I import data.

```
# using results ='hide' makes it so that this doesn't have an output
# when you knit to pdf

# IMPORT DATA
mto_data <- read_dta("Data/mto_sci_puf_cells_20130206.dta")
mto_data

# CLEAN
str(mto_data$ra_group)
mto_data$ra_group <- as_factor(mto_data$ra_group)
# ra_group needs to be a factor so that R
# understands it's a categorical variable

# SUMMARY STATS
#summary(mto_data)
#names(mto_data) # names of all the variables in order
#objects(mto_data) # names in alphabetical order
```

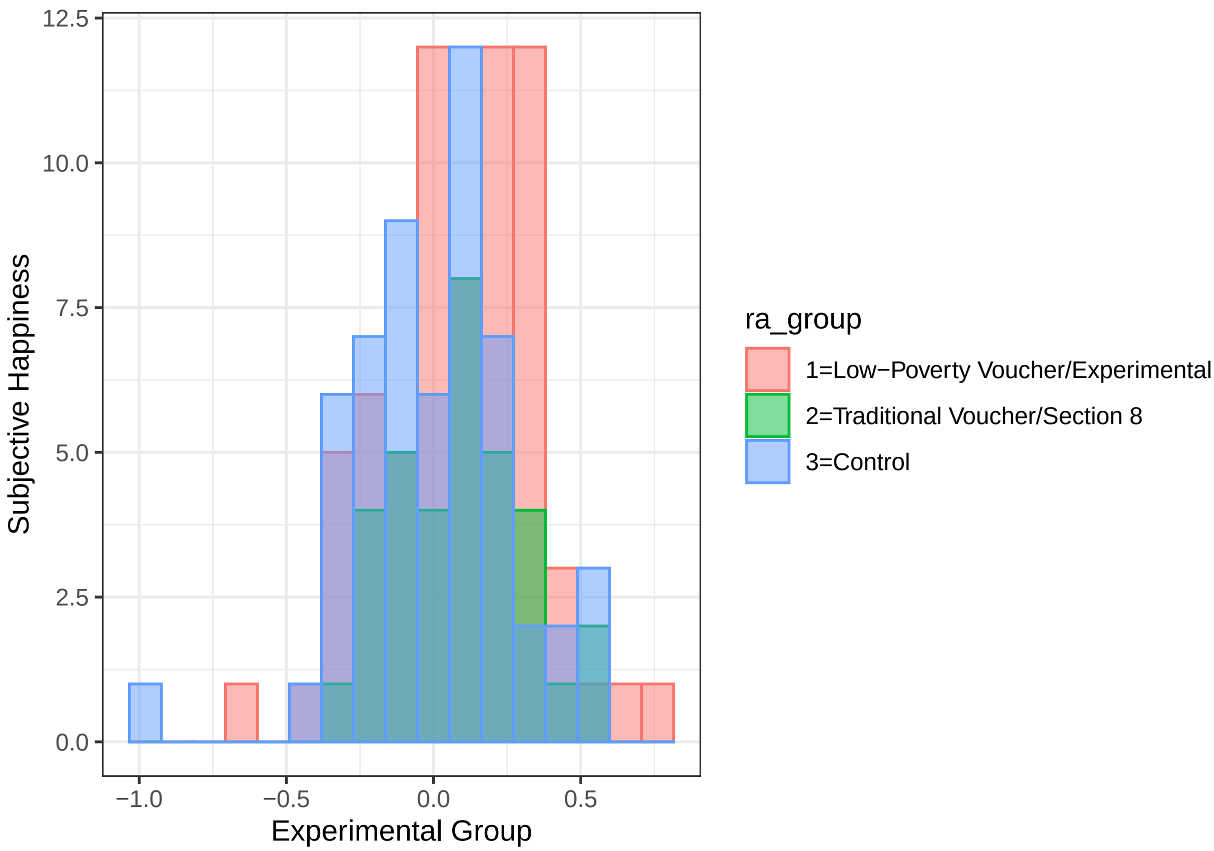
Now I can perform the data analysis stories!

Data Summary Story

```
# histogram showing how subjective happiness changes
# depending on the experimental group

# identify the appropriate bin size
bw <- density(mto_data$mn_happy_scale123_z_ad,
              kernel = "gaussian",
              bw = "ucv")$bw
```

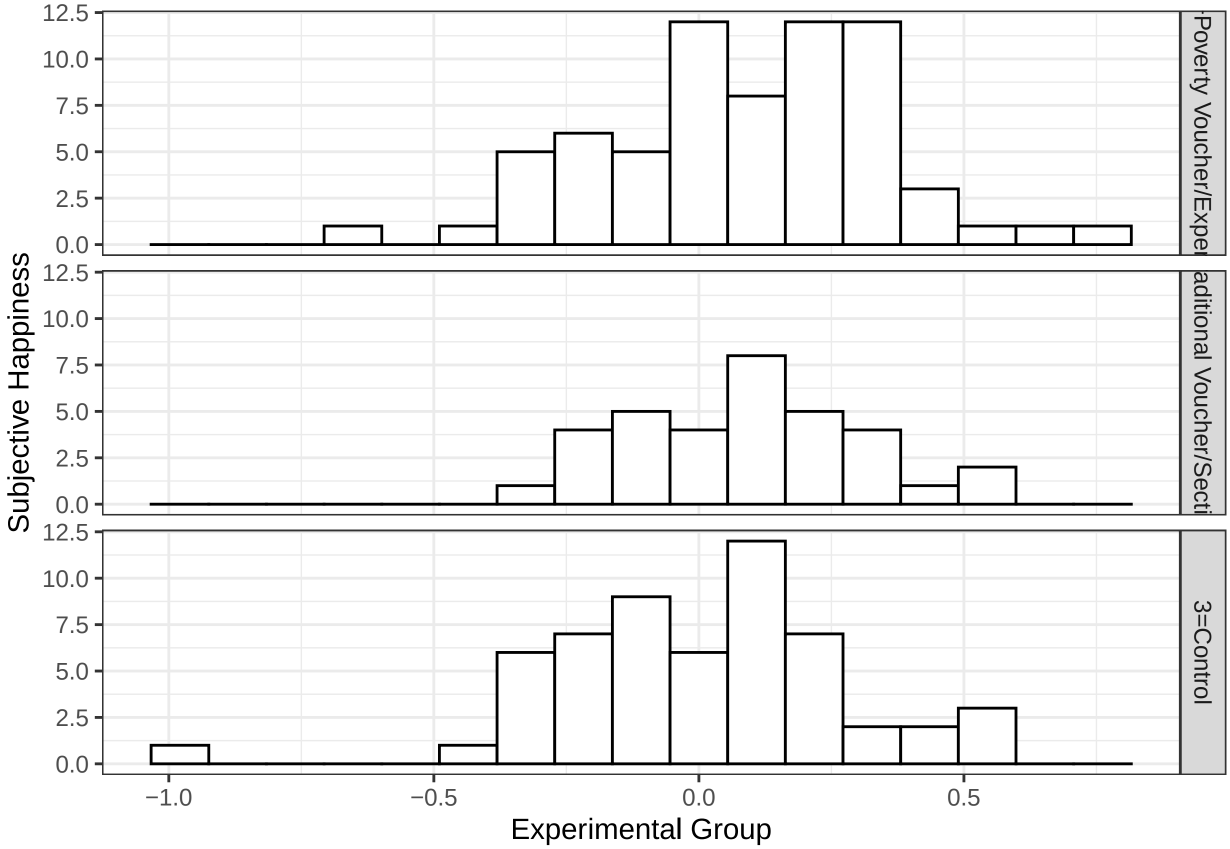
```
## Warning in bw.ucv(x): minimum occurred at one end of the range
```

```
# plot histograms on top of each other
ggplot(mto_data, aes(x = mn_happy_scale123_z_ad,
                     color = ra_group,
                     fill=ra_group)) +
  geom_histogram(position="identity",
                 binwidth = bw,
                 alpha=0.5) +
  theme_bw() +
  xlab("Experimental Group") + ylab("Subjective Happiness")
```



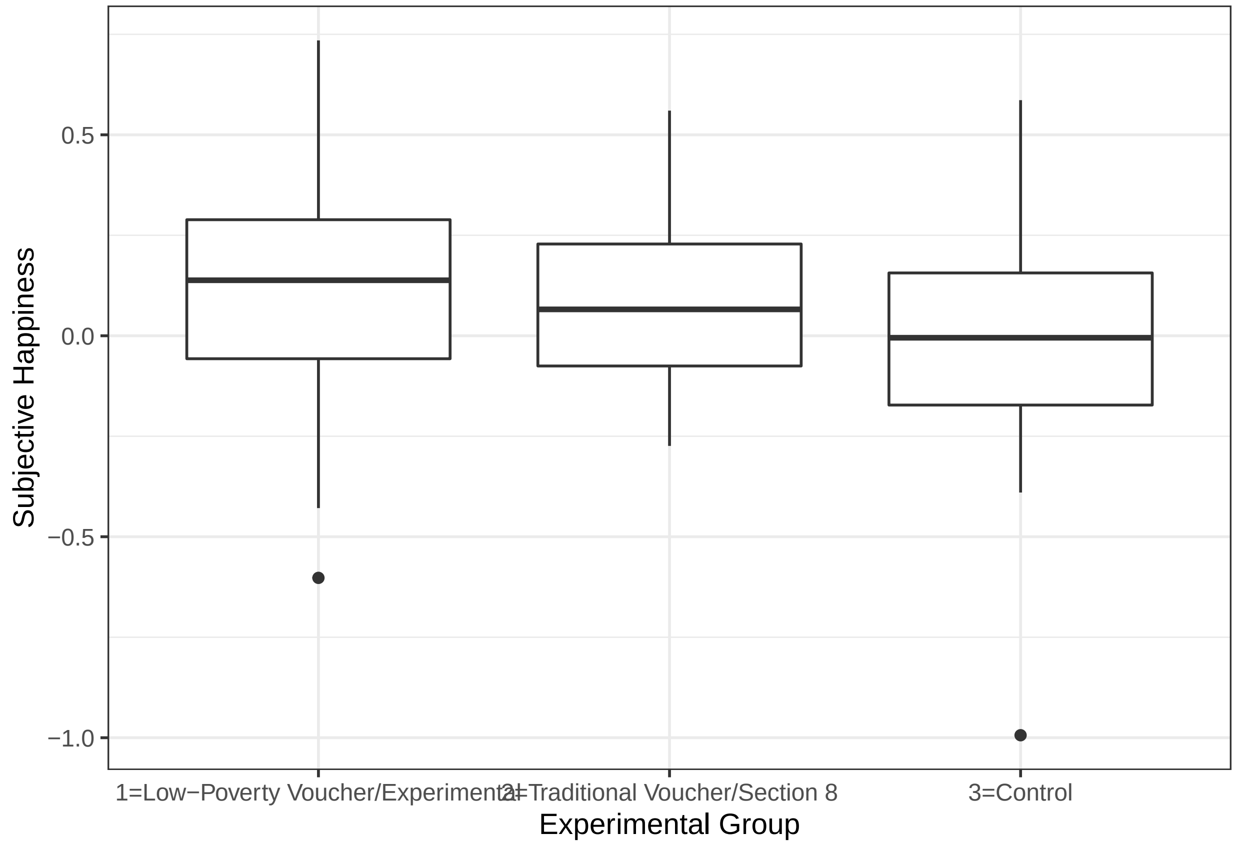
Sometimes histograms are easier to see when they are separated. We want to understand the distributions of the data.

```
# use facets to plot the histograms separately
ggplot(mto_data, aes(x=mn_happy_scale123_z_ad)) +
  geom_histogram(color="black",
                 fill="white",
                 binwidth = bw) +
  facet_grid(ra_group ~ .) +
  theme_bw() +
  xlab("Experimental Group") + ylab("Subjective Happiness")
```



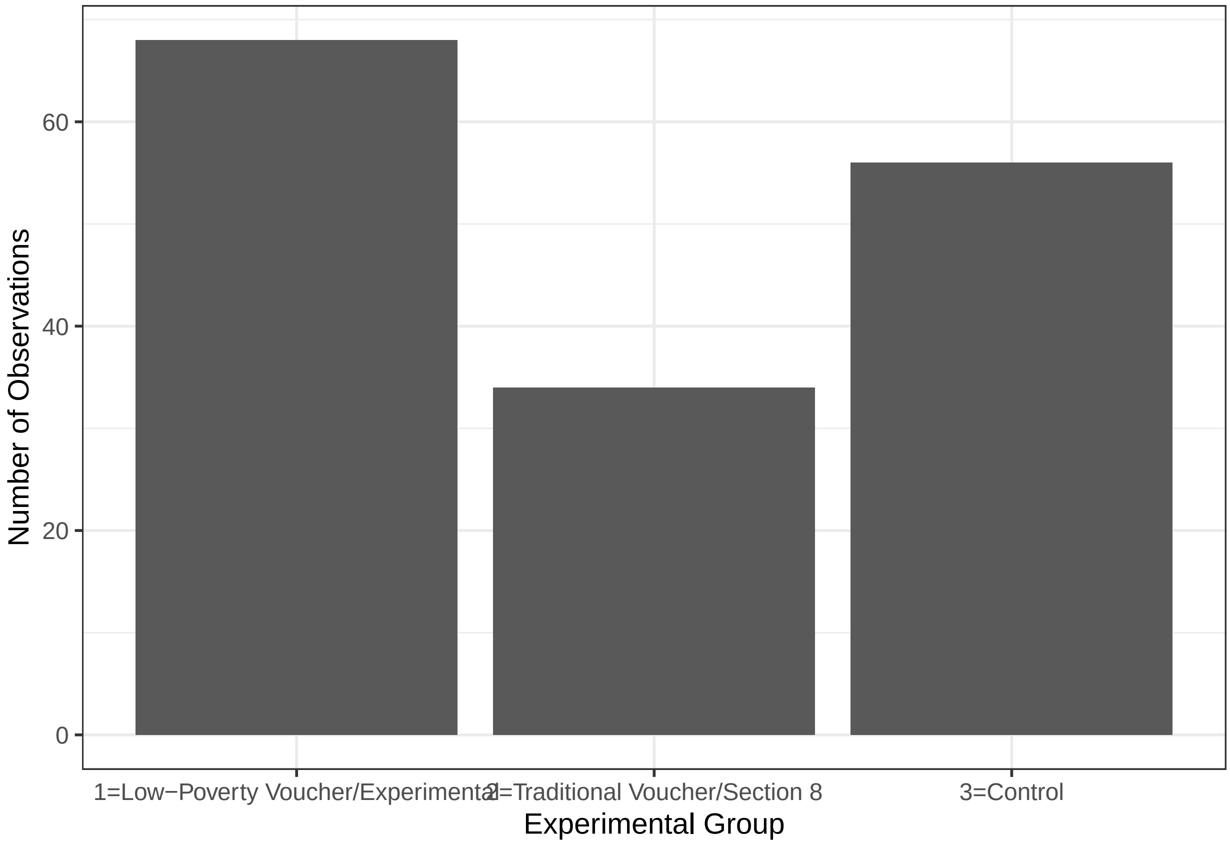
Using box plots, we can better understand medians and outliers.

```
# boxplot
ggplot(mto_data, aes(x = ra_group, y = mn_happy_scale123_z_ad)) +
  geom_boxplot() +
  theme_bw() +
  xlab("Experimental Group") + ylab("Subjective Happiness")
```



Bar plots are useful for understanding categorical variables.

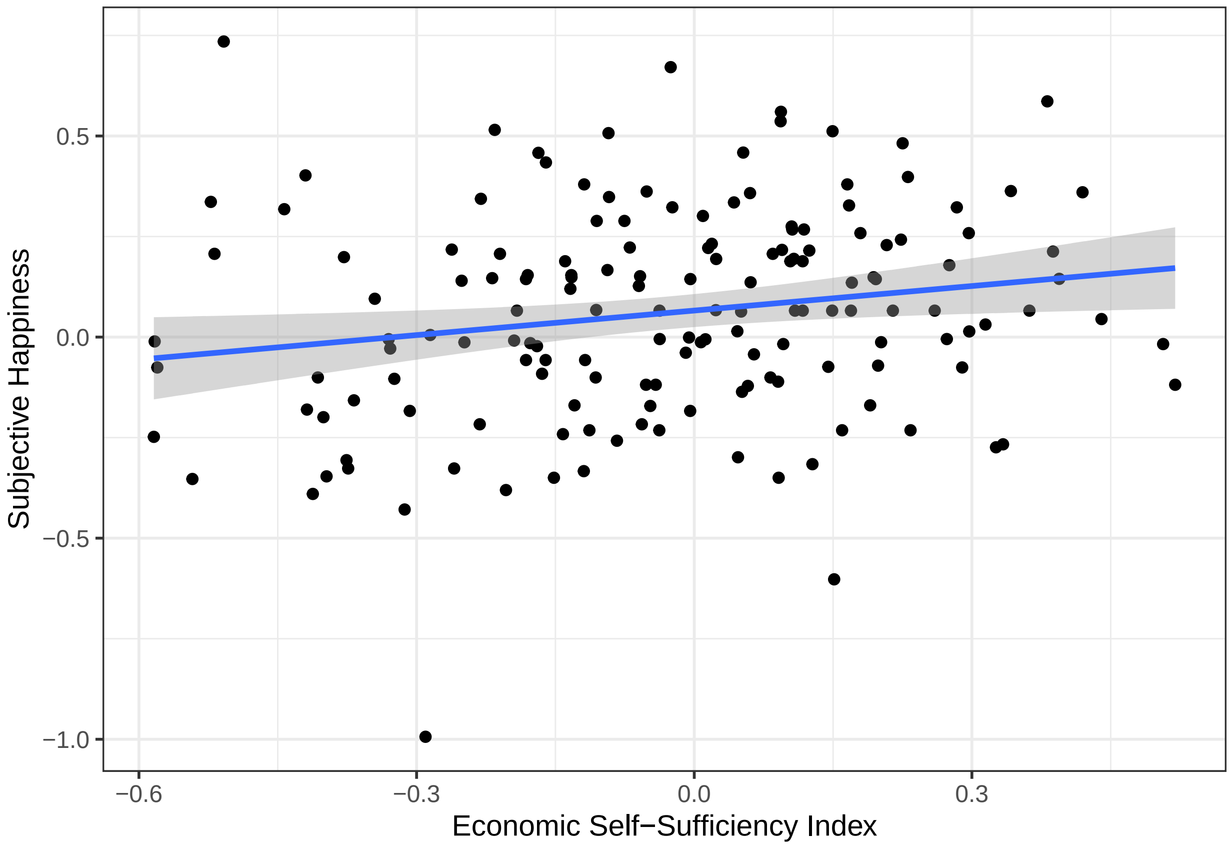
```
ggplot(mto_data, aes(x = ra_group)) +
  geom_bar() +
  theme_bw() +
  xlab("Experimental Group") + ylab("Number of Observations")
```



Scatter plots help us understand the relationship between two variables

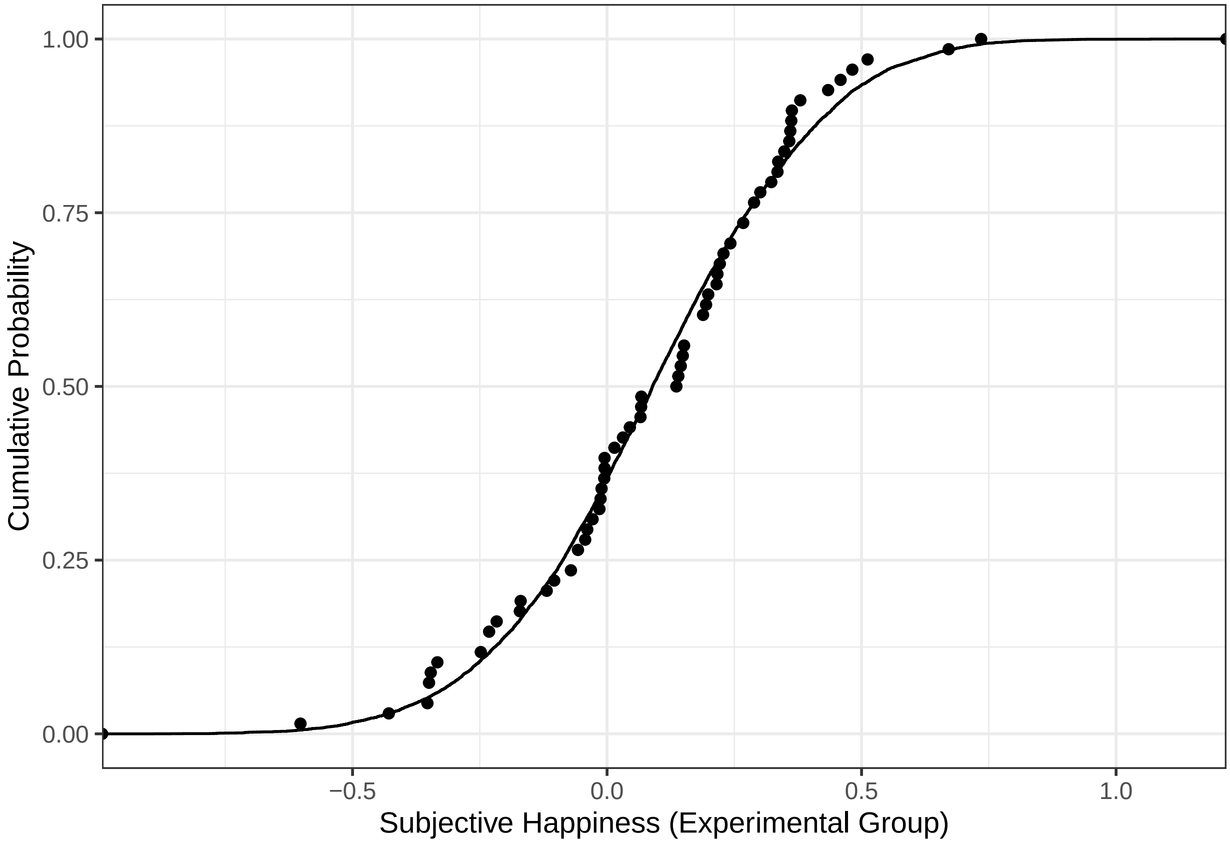
```
# scatter plot with regression line
ggplot(mto_data, aes(x = mn_f_ec_idx_z_ad, y = mn_happy_scale123_z_ad)) +
  geom_point() +
  geom_smooth(method='lm') +
  theme_bw() +
  xlab("Economic Self-Sufficiency Index") + ylab("Subjective Happiness")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Conditional Distribution Story

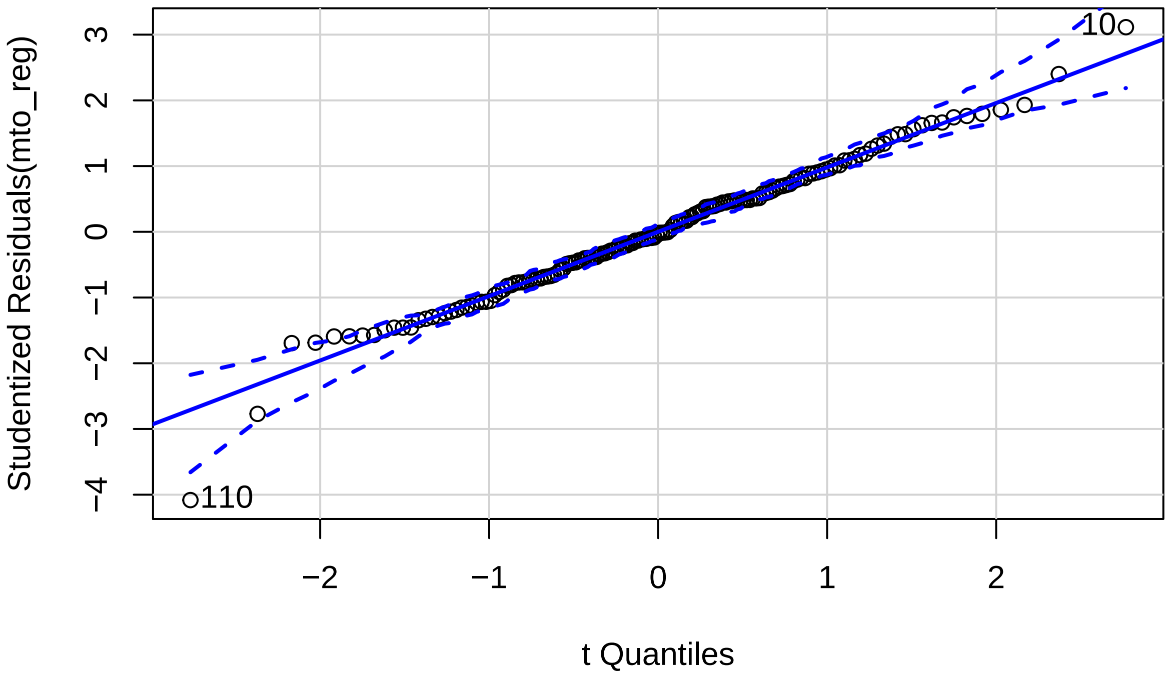
```
# plot cumulative distribution
# compare data to Normal distribution

distribution_info <- mto_data %>%
  group_by(ra_group) %>%
  summarise(mean = mean(mn_happy_scale123_z_ad),
            sd = sd(mn_happy_scale123_z_ad))

set.seed(1) # Set seed for the random number generator, to reproduce results
n.experimental <- as_tibble(rnorm(10000,
                                  distribution_info$mean[1],
                                  distribution_info$sd[1]))
n.section8 <- rnorm(10000,
                    distribution_info$mean[2],
                    distribution_info$sd[2])
n.control <- rnorm(10000,
                   distribution_info$mean[3],
                   distribution_info$sd[3])

# just plot experimental group data for now
mto_data %>%
  filter(ra_group == "1=Low-Poverty Voucher/Experimental") %>%
ggplot(aes(x = mn_happy_scale123_z_ad)) +
  stat_ecdf(geom = "point") +
  theme_bw() +
  xlab("Subjective Happiness (Experimental Group)") + ylab("Cumulative Probability") +
  stat_ecdf(aes(value), n.experimental)
```




```r
mto_reg <- lm(mn_happy_scale123_z_ad ~ mn_f_ec_idx_z_ad, data = mto_data)
summary(mto_reg)
```

```
##
## Call:
## lm(formula = mn_happy_scale123_z_ad ~ mn_f_ec_idx_z_ad, data = mto_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.0007 -0.1770 -0.0084  0.1605  0.7723
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)        0.06596    0.02076   3.178  0.00179 **
## mn_f_ec_idx_z_ad  0.20334    0.08558   2.376  0.01871 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.259 on 156 degrees of freedom
## Multiple R-squared:  0.03492,    Adjusted R-squared:  0.02874
## F-statistic: 5.645 on 1 and 156 DF,  p-value: 0.01871
```

```r
# q-q plot
qqPlot(mto_reg)
```



Studentized Residuals(mto_reg)
t Quantiles
−4
−3
−2
−1
0
1
2
3
−2
−1
0
1
2
10
110


```
## [1]  10 110
```

Forecasting Story

```
# cross validation

complex <- c() # Create an empty vector
simple <- c()
set.seed(11)

for(i in 1:10){  # Loop i from 1 to 100
  train <- sample(mto_data$cell_id,
                  2*length(mto_data$cell_id)/3,
                  replace = FALSE)
  test <- mto_data$cell_id[ - train]
  train1 <- lm(mn_happy_scale123_z_ad ~ mn_f_ec_idx_z_ad,
               data = mto_data[mto_data$cell_id %in% train, ])
  train2 <- lm(mn_happy_scale123_z_ad ~ 1,
               data = mto_data[mto_data$cell_id %in% train, ])
  test1 <- (mto_data$mn_happy_scale123_z_ad[mto_data$cell_id %in% test] -
              predict(train1, mto_data[mto_data$cell_id %in% test, ]))^2
  test2 <- (mto_data$mn_happy_scale123_z_ad[mto_data$cell_id %in% test] -
              predict(train2,mto_data[mto_data$cell_id %in% test, ]))^2
  rMSEtest1 <- sqrt(sum(test1)/length(test1))
  rMSEtest2 <- sqrt(sum(test2)/length(test2))
  # Append the rMSE from this iteration to vectors
  complex <- append(complex, rMSEtest1)
  simple <- append(simple, rMSEtest2)
}

summary(complex)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2244  0.2482  0.2571  0.2607  0.2702  0.3103
```

```
summary(simple)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.2248  0.2376  0.2635  0.2628  0.2747  0.3104
```

Statistical Inference Story

```r
# correlation
rcorr(as.matrix(mto_data[,c("mn_happy_scale123_z_ad", "mn_f_ec_idx_z_ad")]))
```

```
##                        mn_happy_scale123_z_ad mn_f_ec_idx_z_ad
## mn_happy_scale123_z_ad                   1.00             0.19
## mn_f_ec_idx_z_ad                         0.19             1.00
##
## n= 158
##
##
## P
##                        mn_happy_scale123_z_ad mn_f_ec_idx_z_ad
## mn_happy_scale123_z_ad                        0.0187
## mn_f_ec_idx_z_ad       0.0187
```

```r
# hypothesis testing

# t.test
experimental <- mto_data %>%
  filter(ra_group == "1=Low-Poverty Voucher/Experimental")
control <- mto_data %>%
  filter(ra_group == "3=Control")
t.test(experimental$mn_happy_scale123_z_ad,
       control$mn_happy_scale123_z_ad)
```

```
##
##  Welch Two Sample t-test
##
## data:  experimental$mn_happy_scale123_z_ad and control$mn_happy_scale123_z_ad
## t = 1.8605, df = 116.23, p-value = 0.06535
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.005904634  0.188807160
## sample estimates:
##   mean of x   mean of y
## 0.093573791 0.002122528
```

```r
# regression
mto_group_lm <- lm(mn_happy_scale123_z_ad ~ ra_group,
                   data = mto_data)
summary(mto_group_lm)
```

```
##
## Call:
## lm(formula = mn_happy_scale123_z_ad ~ ra_group, data = mto_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.99593 -0.16295 -0.00698  0.17397  0.64133
##
## Coefficients:
##                                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                                0.093574   0.031637   2.958  0.00358
```

```
## ra_group2=Traditional Voucher/Section 8 -0.005756   0.054797  -0.105  0.91648
## ra_group3=Control                        -0.091451   0.047077  -1.943  0.05388
##
## (Intercept)                              **
## ra_group2=Traditional Voucher/Section 8
## ra_group3=Control                        .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2609 on 155 degrees of freedom
## Multiple R-squared:  0.0268, Adjusted R-squared:  0.01425
## F-statistic: 2.134 on 2 and 155 DF,  p-value: 0.1218
```

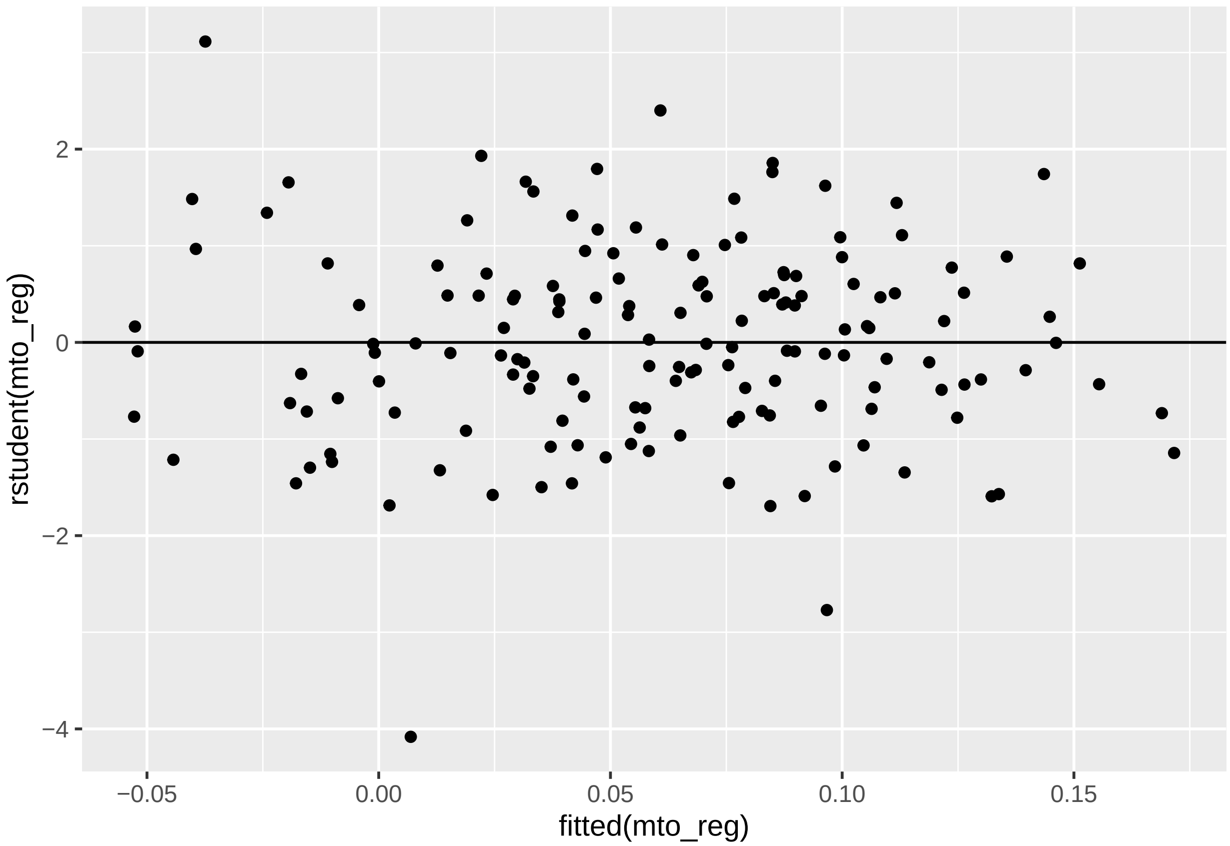
```
# ANOVA
mto_group_aov <- aov(mn_happy_scale123_z_ad ~ ra_group,
                     data = mto_data)
summary(mto_group_aov)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## ra_group      2  0.291 0.14527   2.134  0.122
## Residuals   155 10.549 0.06806
```

```
# post hoc test
TukeyHSD(mto_group_aov, which = 'ra_group')
```

```
##   Tukey multiple comparisons of means
##     95% family-wise confidence level
##
## Fit: aov(formula = mn_happy_scale123_z_ad ~ ra_group, data = mto_data)
##
## $ra_group
##                                                                     diff
## 2=Traditional Voucher/Section 8-1=Low-Poverty Voucher/Experimental -0.005756045
## 3=Control-1=Low-Poverty Voucher/Experimental                       -0.091451263
## 3=Control-2=Traditional Voucher/Section 8                          -0.085695218
##                                                                     lwr
## 2=Traditional Voucher/Section 8-1=Low-Poverty Voucher/Experimental -0.1354297
## 3=Control-1=Low-Poverty Voucher/Experimental                       -0.2028571
## 3=Control-2=Traditional Voucher/Section 8                          -0.2199202
##                                                                     upr
## 2=Traditional Voucher/Section 8-1=Low-Poverty Voucher/Experimental 0.12391758
## 3=Control-1=Low-Poverty Voucher/Experimental                       0.01995455
## 3=Control-2=Traditional Voucher/Section 8                          0.04852974
##                                                                   p adj
## 2=Traditional Voucher/Section 8-1=Low-Poverty Voucher/Experimental 0.9939354
## 3=Control-1=Low-Poverty Voucher/Experimental                       0.1303044
## 3=Control-2=Traditional Voucher/Section 8                          0.2886327
```

```
# resampling with bootstrap

# jacknife residuals plot
ggplot(mto_data, aes(x = fitted(mto_reg), y = rstudent(mto_reg))) +
  geom_point() +
  geom_hline(yintercept = 0)
```



rstudent(mto_reg)
2
0
−2
−4
−0.05
0.00
0.05
0.10
0.15
fitted(mto_reg)

Causal Inference Story

There's no R code for this story.