Data Visualizations

Brian Anderson

Load Libraries

```
library(tidyverse)
library(broom)
```

Simulated Data

Lets set some values, and we'll simulate a 500 observation dataset.

$$y = 2.5 + .65x + 1.5m + .85xm + .5c + \mathcal{N}(0, 3)$$

```
# Set our (quasi) random number generator seed
set.seed(1)

# Set our number of observations
obs <- 500

# Define our model parameters
a <- 2.5
b_x <- .65
b_m <- 1.5
b_xm <- .85
b_c <- .5

df <- tibble(x = rnorm(obs, 0, 2),  # Gaussian (normal) distribution
             m = rbinom(obs, 1, .4),  # Bernulli (binomial) distribution
             c = rpois(obs, 15),  # Poisson distribution
             y = a + (b_x * x) + (b_m * m) + (b_xm * (x * m)) + (b_c * c) +
               rnorm(obs, 0, 3))

df
```

```
## # A tibble: 500 x 4
##         x     m     c     y
##     <dbl> <int> <int> <dbl>
##  1 -1.25      0     9  5.00
##  2  0.367     1    15 12.8
##  3 -1.67      0    13  7.07
##  4  3.19      1    16 19.4
##  5  0.659     0    16 13.1
##  6 -1.64      0    16 12.6
##  7  0.975     0    18 12.9
##  8  1.48      0    13 13.3
##  9  1.15      1    14 12.5
## 10 -0.611     1    14 12.9
## # ... with 490 more rows
```

Interaction Model

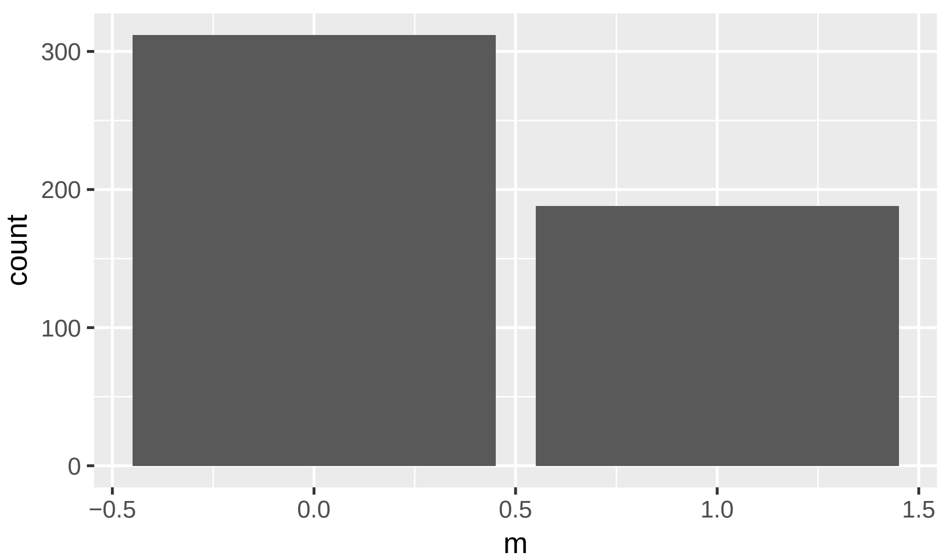
```r
y.model <- lm(y ~ x * m + c, data = df)
summary(y.model)
```

```
##
## Call:
## lm(formula = y ~ x * m + c, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.9670 -1.9139  0.0009  2.2957  9.7059
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.55565    0.53812   4.749 2.68e-06 ***
## x            0.55031    0.08772   6.273 7.72e-10 ***
## m            2.00123    0.29215   6.850 2.20e-11 ***
## c            0.48668    0.03547  13.720  < 2e-16 ***
## x:m          0.77254    0.14408   5.362 1.27e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.141 on 495 degrees of freedom
## Multiple R-squared:  0.4824, Adjusted R-squared:  0.4782
## F-statistic: 115.3 on 4 and 495 DF,  p-value: < 2.2e-16
```

Basic Visualizations

Bar Chart

```r
ggplot(data = df, aes(x = m)) +
  geom_bar()
```



Data summary

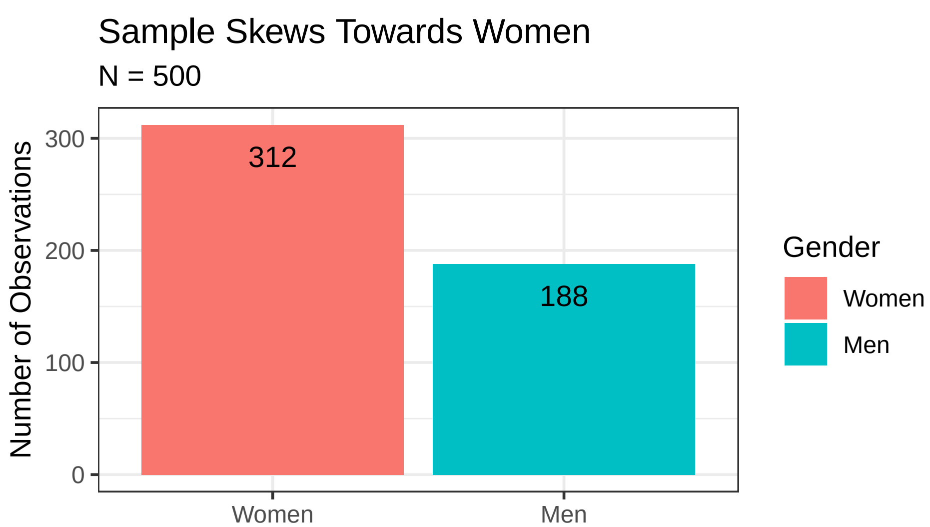
```r
df %>%
  count(m)
```

```
## # A tibble: 2 x 2
##       m     n
##   <int> <int>
## 1     0   312
## 2     1   188
```

Tidy Bar Chart

```r
tidy.box <- ggplot(data = df %>%
                     mutate(m = as_factor(m)),
                   aes(x = m, fill = m)) +
  geom_bar() +
  geom_text(stat = 'count', aes(label = ..count..), vjust = 2) +
  scale_x_discrete(breaks = c("0", "1"),
                   labels = c("Women", "Men")) +
  scale_fill_discrete(name = "Gender",
                      breaks = c("0", "1"),
                      labels = c("Women", "Men")) +
  labs(title = "Sample Skews Towards Women",
       subtitle = "N = 500",
       y = "Number of Observations",
       x = "") +
  theme_bw()

tidy.box
```

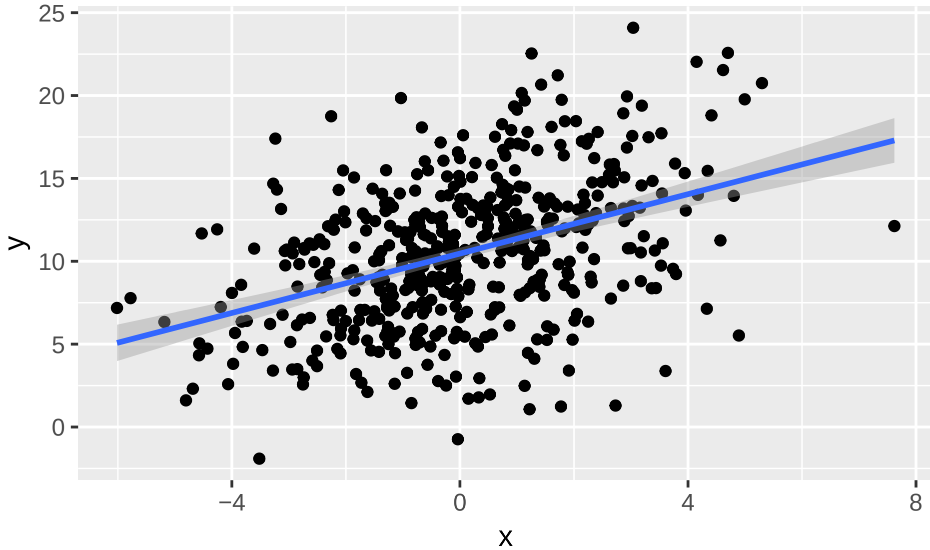


Save the Plot

```r
ggsave("TidyBox.png", tidy.box, width = 6, height = 4)
```

Scatterplot

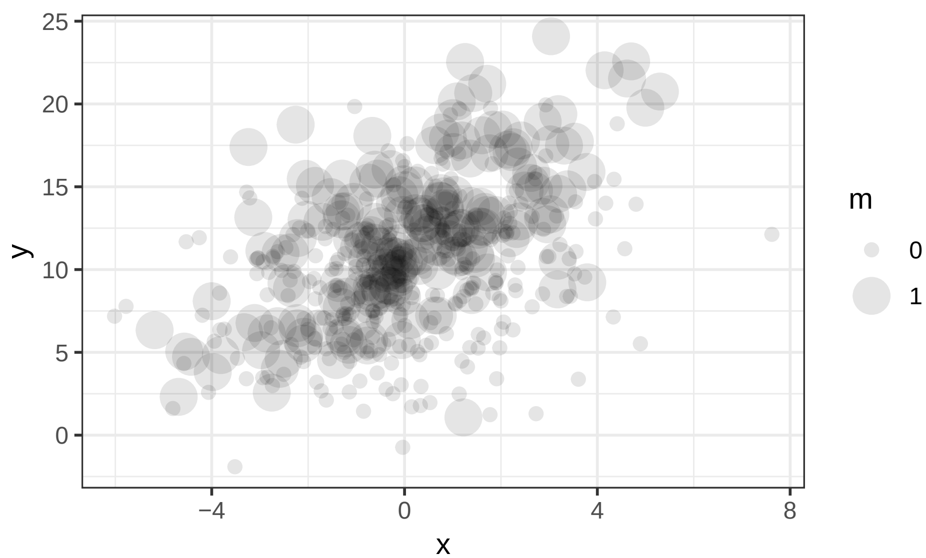
```r
ggplot(data = df, aes(y = y, x = x)) +
  geom_point() +
  geom_smooth(method = "lm")
```



Bubble Chart

This incorporates the dichotomous mediator…

```r
ggplot(data = df %>%
         mutate(m = as_factor(m)),
       aes(y = y, x = x, size = m)) +
  geom_point(alpha = .1) +
  theme_bw()
```



Plotting Simnple Slopes

Create New Dataframe

```r
# M = 0 condition
m.0 <- tibble(x = seq(min(df$x), max(df$x), .1),
              m = 0)

# M = 1 condition
m.1 <- tibble(x = seq(min(df$x), max(df$x), .1),
              m = 1)

# Bind the dataframes together
m.df <- bind_rows(m.0, m.1) %>%
  mutate(c = mean(df$c))

m.df
```

```
## # A tibble: 274 x 3
##        x     m     c
##    <dbl> <dbl> <dbl>
##  1 -6.02     0  14.6
##  2 -5.92     0  14.6
##  3 -5.82     0  14.6
##  4 -5.72     0  14.6
##  5 -5.62     0  14.6
##  6 -5.52     0  14.6
##  7 -5.42     0  14.6
##  8 -5.32     0  14.6
##  9 -5.22     0  14.6
## 10 -5.12     0  14.6
## # ... with 264 more rows
```

Create predicted values

```r
y.pred <- augment(y.model, newdata = m.df)
y.pred
```

```
## # A tibble: 274 x 5
##        x     m     c .fitted .se.fit
##    <dbl> <dbl> <dbl>   <dbl>   <dbl>
##  1 -6.02     0  14.6    6.37   0.554
##  2 -5.92     0  14.6    6.42   0.546
##  3 -5.82     0  14.6    6.48   0.538
##  4 -5.72     0  14.6    6.53   0.530
##  5 -5.62     0  14.6    6.59   0.521
##  6 -5.52     0  14.6    6.64   0.513
##  7 -5.42     0  14.6    6.70   0.505
##  8 -5.32     0  14.6    6.75   0.497
##  9 -5.22     0  14.6    6.81   0.489
## 10 -5.12     0  14.6    6.86   0.480
## # ... with 264 more rows
```

Quantify Uncertainty

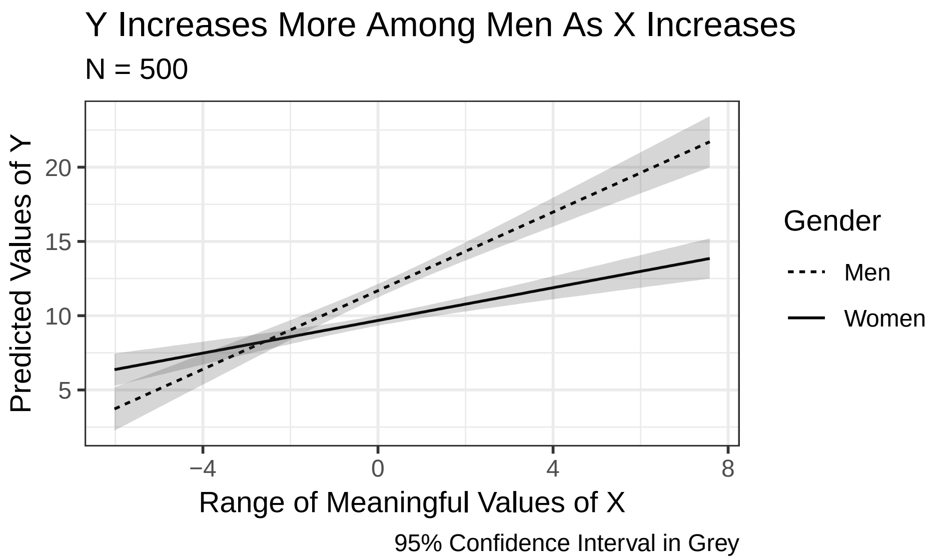
```r
y.pred <- y.pred %>%
  mutate(lower.ci = .fitted - (1.96 * .se.fit),
         upper.ci = .fitted + (1.96 * .se.fit)) %>%
  mutate_if(is.numeric, funs(round(., 2)))

y.pred
```

```
## # A tibble: 274 x 7
##        x     m     c .fitted .se.fit lower.ci upper.ci
##    <dbl> <dbl> <dbl>   <dbl>   <dbl>    <dbl>    <dbl>
##  1 -6.02     0  14.6    6.37    0.55     5.28     7.45
##  2 -5.92     0  14.6    6.42    0.55     5.35     7.49
##  3 -5.82     0  14.6    6.48    0.54     5.42     7.53
##  4 -5.72     0  14.6    6.53    0.53     5.49     7.57
##  5 -5.62     0  14.6    6.59    0.52     5.57     7.61
##  6 -5.52     0  14.6    6.64    0.51     5.64     7.65
##  7 -5.42     0  14.6    6.7     0.5      5.71     7.69
##  8 -5.32     0  14.6    6.75    0.5      5.78     7.73
##  9 -5.22     0  14.6    6.81    0.49     5.85     7.76
## 10 -5.12     0  14.6    6.86    0.48     5.92     7.8
## # ... with 264 more rows
```

Build the Plot

```r
ggplot(data = y.pred %>%
         mutate(m = as_factor(m)),
       aes (y = .fitted, x = x, group = m)) +
  geom_line(aes(linetype = m)) +
  geom_ribbon(alpha = .2, aes(ymin = lower.ci, ymax = upper.ci)) +
  scale_linetype_discrete(name = "Gender",
                          breaks = c("1", "0"),
                          labels = c("Men", "Women")) +
  labs(title = "Y Increases More Among Men As X Increases",
       subtitle = "N = 500",
       y = "Predicted Values of Y",
       x = "Range of Meaningful Values of X",
       caption = "95% Confidence Interval in Grey") +
  theme_bw()
```



Multilevel Data

```
rnd.df <- read_csv("https://www.drbanderson.com/data/FirmRND.csv")
rnd.df
```

```
## # A tibble: 330 x 9
##    FirmID  Year TickerSymbol CompanyName SICCode    RND Revenue     at
##     <dbl> <dbl> <chr>        <chr>         <dbl>  <dbl>   <dbl>  <dbl>
##  1   1820  2013 ALOT         ASTRONOVA ~    3577 5.07e0    68.6 7.80e1
##  2   1820  2014 ALOT         ASTRONOVA ~    3577 5.80e0    88.3 7.43e1
##  3   1820  2015 ALOT         ASTRONOVA ~    3577 6.94e0    94.7 7.80e1
##  4   1820  2016 ALOT         ASTRONOVA ~    3577 6.31e0    98.4 8.37e1
##  5   1820  2017 ALOT         ASTRONOVA ~    3577 7.45e0   113.  1.22e2
##  6   2721  2013 CAJ          CANON INC      3577 2.91e3 35453.  4.03e4
##  7   2721  2014 CAJ          CANON INC      3577 2.58e3 31099.  3.72e4
##  8   2721  2015 CAJ          CANON INC      3577 2.73e3 31598.  3.68e4
##  9   2721  2016 CAJ          CANON INC      3577 2.59e3 29127.  4.40e4
## 10   2721  2017 CAJ          CANON INC      3577 2.93e3 36206.  4.61e4
## # ... with 320 more rows, and 1 more variable: NetIncome <dbl>
```

Wrangling…

```
# Industries...
# 3570 - Computer & Office Eqpmt
# 3571 - Electronic Computers
# 3572 - Computer Storage Devices
# 3576 - Computer Communications Eqpmt
# 3577 - Computer Peripheral Eqpmt
# 3578 - Calculating & Accounting Machines
# 3579 - Office Machines

rnd.df <- rnd.df %>%
  mutate(SICCode = as_factor(SICCode))
```

What are we looking at?

```r
rnd.df %>%
  summarise(NumberFirms = n_distinct(FirmID))
```

```
## # A tibble: 1 x 1
##   NumberFirms
##         <int>
## 1          79
```

```r
rnd.df %>%
  distinct(Year)
```

```
## # A tibble: 5 x 1
##    Year
##   <dbl>
## 1  2013
## 2  2014
## 3  2015
## 4  2016
## 5  2017
```

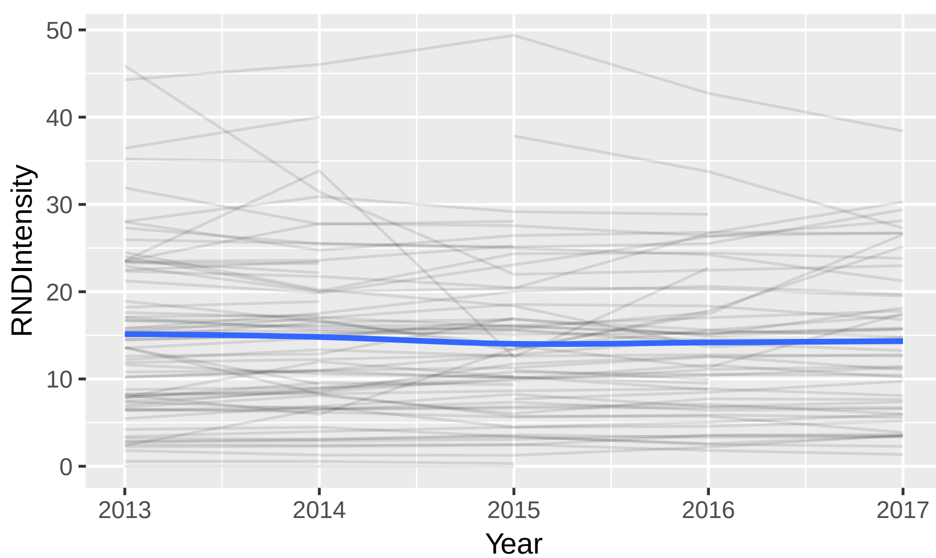
Lets create a couple of variables...

```r
rnd.df <- rnd.df %>%
  mutate(ROS = 100 * (NetIncome / Revenue),
         RNDIntensity = 100 * (RND / Revenue)) %>%
  filter(ROS > -50,
         RNDIntensity < 100)  # Eliminate outliers
```

Visualizations

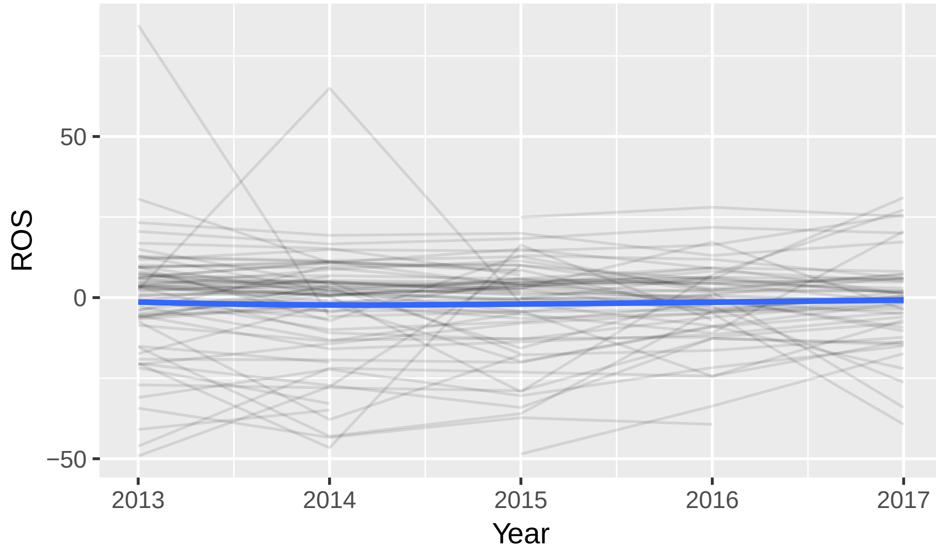
R&D Intensity Over Time By Firm

```r
ggplot(rnd.df, aes(y = RNDIntensity, x = Year)) +
  geom_line(aes(group = FirmID), alpha = 1/10) +
  geom_smooth(method = "loess", se = FALSE)
```



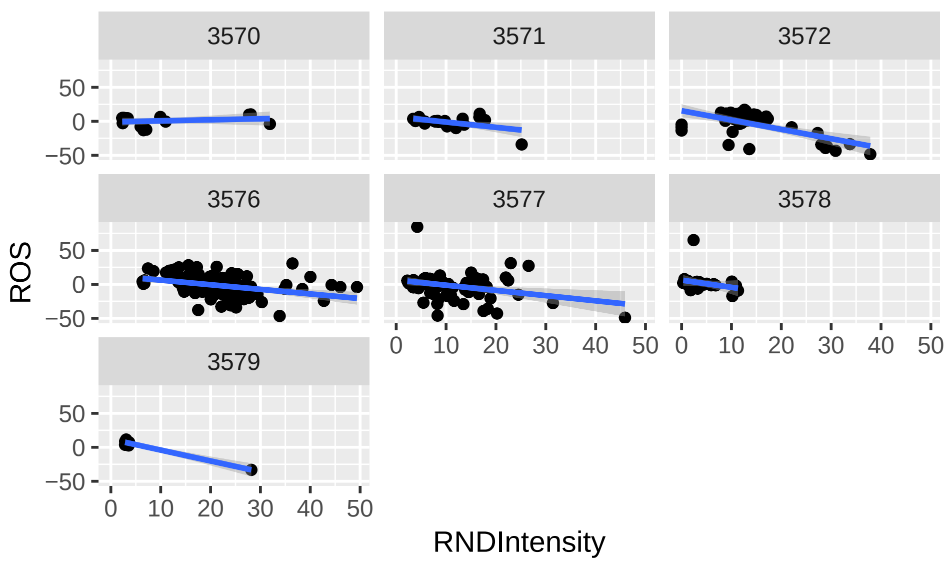
ROS Over Time By Firm

```
ggplot(rnd.df, aes(y = ROS, x = Year)) +
  geom_line(aes(group = FirmID), alpha = 1/10) +
  geom_smooth(method = "loess", se = FALSE)
```



ROS & R&D By Industry

```
ggplot(rnd.df, aes(y = ROS, x = RNDIntensity)) +
  geom_point() +
  geom_smooth(method = "lm") +
  facet_wrap(~ SICCode)
```



Data Summaries

```
ggplot(data = rnd.df %>%
         group_by(SICCode, Year) %>%
           summarise(meanRND = median(RNDIntensity)),
       aes(y = meanRND, x = Year, group = SICCode)) +
  geom_line(aes(linetype = SICCode)) +
  scale_linetype_discrete(name = "SIC Code") +
  labs(title = "Median R&D Intensity of 35x Industries",
       subtitle = "2013 - 2017",
       y = "Median R&D Intensity",
       x = "") +
  theme_bw()
```