W271-2 – Spring 2016 – **Lab 2**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

March 7, 2016

Contents

Question 1: Broken Rulers **2**

Question 2: Investing **8**

Question 3: Turtles **10**

Question 4: CLM 1 **12**

Background . . . 12
The Data . . . 12
Question 4.1 . . . 12
Question 4.2 . . . 18
Question 4.3 . . . 18
Question 4.4 . . . 18
Question 4.5 . . . 18
Question 4.6 . . . 19

Question 5: CLM 2 **20**

Question 6: CLM 3 **27**

Question 1: Broken Rulers

You have a ruler of length 1 and you choose a place to break it using a uniform probability distribution. Let random variable X represent the length of the left piece of the ruler. X is distributed uniformly in $[0, 1]$. You take the left piece of the ruler and once again choose a place to break it using a uniform probability distribution. Let random variable Y be the length of the left piece from the second break.

1. **Find the conditional expectation of Y given X, $\mathbf{E}(\mathbf{Y}|\mathbf{X})$.**

$f_X = U(0, 1)$ and $f_{Y|X} = U(0, X)$ (because the maximum length of the second left piece cannot be greater than the length of the first left piece). As we know, the probability density function for a variable Z that follows a uniform distribution $U(a, b)$ is:

$$f_Z(z) = \begin{cases} \dfrac{1}{b-a} & a \leq z \leq b \\ 0 & \text{otherwise} \end{cases}$$

So:

$$f_{Y|X}(y|x) = \begin{cases} \dfrac{1}{x} & 0 \leq z \leq x \\ 0 & \text{otherwise} \end{cases}$$

And:

$$\mathbf{E(Y|X)} = \int_{\mathbb{Y}} y \cdot f_{Y|X}(y|x) \cdot dy = \int_{y=0}^{x} y \cdot \frac{1}{x} \cdot dy = \frac{1}{x}\left[\frac{y^2}{2}\right]_{y=0}^{x} = \frac{x^2}{2x} = \mathbf{\frac{x}{2}}$$

We'll make use of some simulations through this Question to confirm the results.

```
simulations <- 1e4 # number of simulations
set.seed(123)
x <- runif(simulations, min=0, max=1) # X ~ U(0,1)
y <- runif(simulations, min=0, max=x) # Y|X ~ U(0,X)
par(mfrow = c(1, 2))
hist(x, main = "pdf(x)", freq = FALSE)
lines(density(x), col = 'red')
hist(y, main = "pdf(y)", freq = FALSE)
lines(density(y), col = 'red')
```

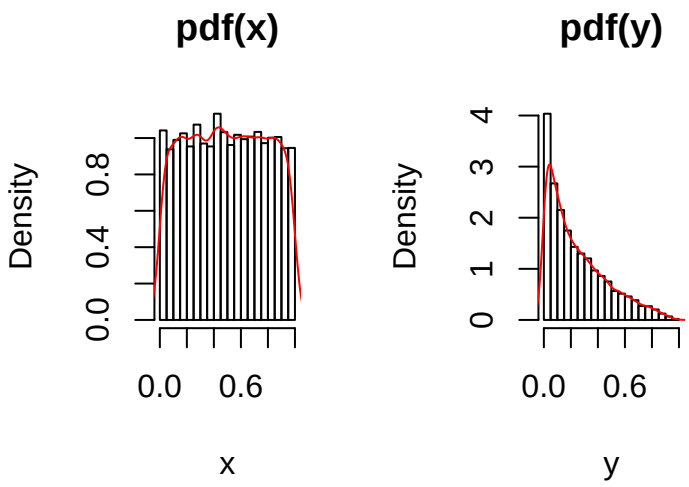


Figure 1: Histogram and approximate pdf of X and Y

```
# y1 <- runif(simulations, min = 0, max = 0.2) # Fix X to 0.2
y1 <- y[x > 0.2 - 1e-2 & x < 0.2 + 1e-2] # Using previous simulation
hist(y1, main = 'pdf(y|x=0.2)', freq = FALSE)
lines(density(y1), xlim = c(0, 1), main = 'pdf(y|x=0.2)', col = 'red')
abline(v = mean(y1), col = 'green', lty = 2, lwd = 4)
```

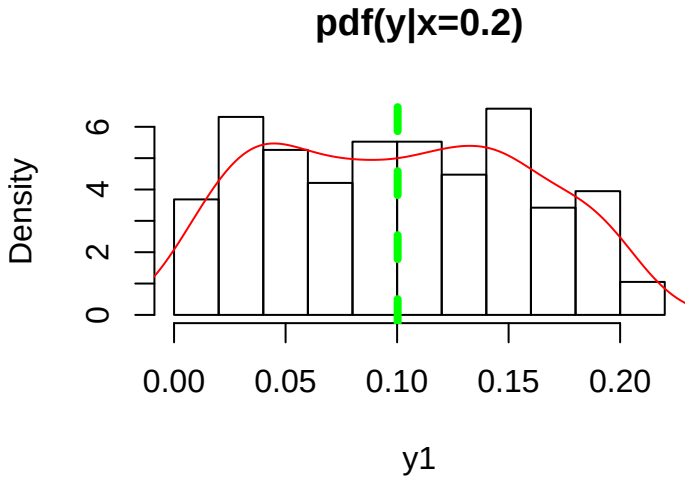


Figure 2: Histogram and approximate pdf of Y conditional on X for a given value of X (0.2)

```
# legend("topright", "E(Y|X=0.2)", lty = 1, bty="n", col = 'red')
```

2. **Find the unconditional expectation of Y. One way to do this is to apply the law of iterated expectations, which states that $\mathbf{E(Y) = E(E(Y|X))}$. The inner expectation is the conditional expectation computed above, which is a function of X. The outer expectation finds the expected value of this function.**

$$\mathbf{E(Y)} = E\left[E(Y|X)\right] = \int_{\mathbb{X}} E(Y|X) \cdot f_X(x) \cdot dx = \int_{x=0}^{1} \frac{x}{2} \cdot 1 \cdot dx = \left[\frac{x^2}{4}\right]_{x=0}^{1} = \mathbf{\frac{1}{4} = 0.25}$$

3. **Write down an expression for the joint probability density function of X and Y, $\mathbf{f_{X,Y}(x,y)}$.**

$$\mathbf{f_{X,Y}(x,y)} = f_{Y|X}(y|x) \cdot f_X(x) = \begin{cases} \mathbf{\frac{1}{x}} & \mathbf{x \in (0,1), y \in (0,x)} \\ \mathbf{0} & \mathbf{\text{otherwise}} \end{cases}$$

Let's check that this is a valid joint *pdf*:

$$\int_{\mathbb{X}} \int_{\mathbb{Y}} f_{X,Y}(x,y) \cdot dx \cdot dy = \int_{x=0}^{1} \int_{y=0}^{x} \frac{1}{x} \cdot dy \cdot dx = \int_{x=0}^{1} \left[\frac{y}{x}\right]_{y=0}^{x} dx = \int_{x=0}^{1} dx = [x]_{x=0}^{1} = 1$$

Simulations:

```
pdf_x <- function(x) ifelse(x<1 & x>0, 1, 0) # f(x)
integrate(pdf_x, -Inf, Inf) # integral
```

```
## 1 with absolute error < 4.2e-11
```

```
pdf_y_given_x <- function(x,y) ifelse(y<x & y>0 & x<1 & x>0, 1/x, 0) # f(y|x)
pdf_xy <- function(x,y) pdf_x(x)*pdf_y_given_x(x,y) # f(x,y)
# integral
integrate(function(y) sapply(y, function(y) integrate(function(x)
  pdf_xy(x,y), 0, 1)$value), -Inf, Inf)
```

```
## 1 with absolute error < 9.3e-05
```

```
# Plot f(x,y)
x0 <- y0 <- seq(0, 1, by = 0.01)
grid <- mesh(x0, y0)
z0 <- with(grid, pdf_x(x)*pdf_y_given_x(x,y))
# contour(x0, y0, z0, asp=1)
# par(mfrow = c(1, 2))
persp3D(z = z0, x = x0, y = y0)
```

```
# Confirm that f(x,y) = 1/x
# z2 <- with(grid, ifelse(x<=y | x==0 | y == 0, 0, 1/x))
# persp3D(z = z2, x = x0, y = y0)
```

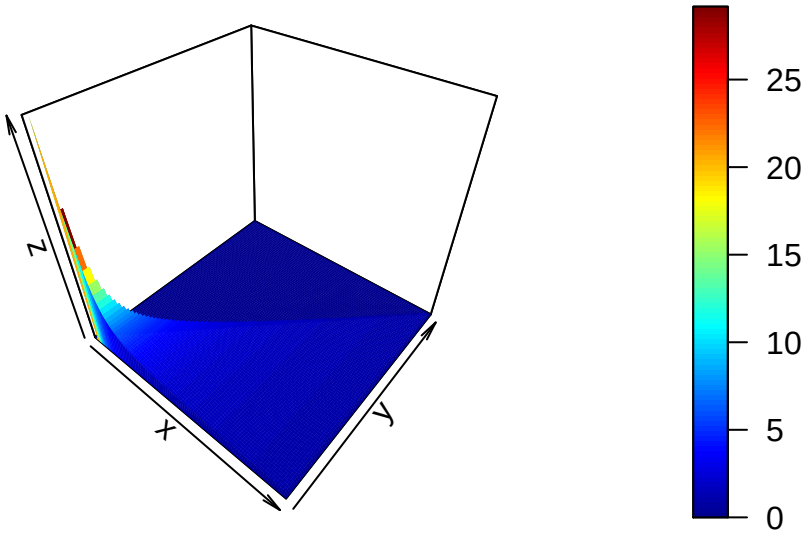


Figure 3: Approximate joint pdf of X and Y

4. **Find the conditional probability density function of X given Y, $\mathbf{f_{X|Y}}$.**

In order to find $f_{X|Y}$ we need the marginal pdf of Y.

$$f_Y(y) = \int_{\mathbb{X}} f_{X,Y}(x,y) \cdot dx = \int_{y=0}^{x} \frac{1}{x} \cdot dx = \int_{x=y}^{1} \frac{dx}{x} = [\log(x)]_{x=y}^{1} dx = -\log(y) = \log\left(\frac{1}{y}\right)$$

This result confirms what the shape of $f_Y(y)$ in Figure 1 suggested.

```
# f(y)
pdf_y <- function(y)
  sapply(y, function(y) integrate(function(x)
    pdf_y_given_x(x,y)*pdf_x(x), 0, 1)$value)
integrate(pdf_y, -Inf, Inf) # integral
```

```
## 1 with absolute error < 9.3e-05
```

```
plot(sort(y), pdf_y(sort(y)), type = 'l', main = 'pdf(y)', xlab = 'y')
# Confirm that f(y) = log(1/y)
lines(sort(y), log(1/sort(y)), type = 'l', main = 'pdf(y)')
```

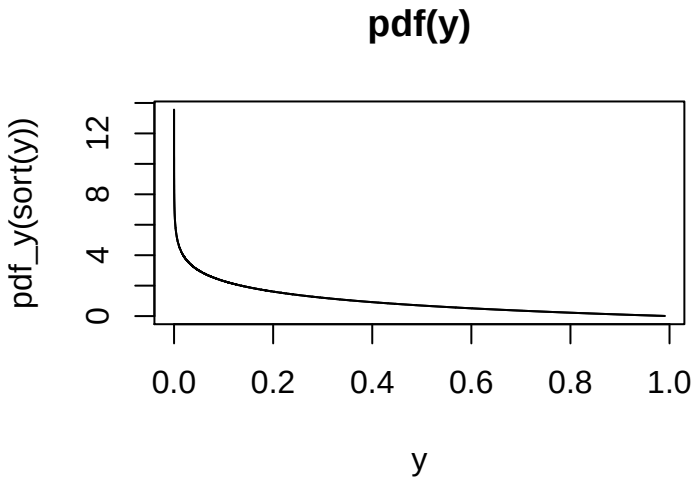


Figure 4: Approximate pdf of Y conditional on X for two values of X

$$\mathbf{f_{X|Y}(x|y)} = \frac{f_{X,Y}(x,y)}{f_Y(y)} = \frac{\mathbf{-1}}{\mathbf{x} \cdot \log(\mathbf{y})}$$

5. **Find the expectation of X, given that Y is 1/2, E(X|Y = 1/2).**

$$\mathbf{E(X|Y=1/2)} = \int_{\mathbb{X}} x \cdot f_{X|Y}(x|y=1/2) \cdot dx = \int_{x=1/2}^{1} x \cdot \left(\frac{-1}{x \cdot \log(1/2)} \right) \cdot dx$$
$$= \frac{1}{\log(2)} \int_{x=1/2}^{1} dx = \frac{1}{\log(2)} [x]_{x=1/2}^{1} = \mathbf{\frac{1}{2 \cdot \log(2)} = 0.721}$$

```
# Confirm E(X|Y=0.5) (use values of Y around 0.5 in the previous simulation)
mean(x[y > 0.5 - 1e-2 & y < 0.5 + 1e-2])
```

```
## [1] 0.72847
```

```
1/(2*log(2))
```

```
## [1] 0.7213475
```

Question 2: Investing

Suppose that you are planning an investment in three different companies. The payoff per unit you invest in each company is represented by a random variable. A represents the payoff per unit invested in the first company, B in the second, and C in the third. A, B, and C are independent of each other. Furthermore, Var(A) = 2Var(B) = 3Var(C).

You plan to invest a total of one unit in all three companies. You will invest amount a in the first company, b in the second, and c in the third, where $a, b, c \in [0, 1]$ and $a + b + c = 1$. Find, the values of a, b, and c that minimize the variance of your total payoff.

Let's call P the total payoff:

$$Var(P) = Var(aA + bB + cC) = a^2 Var(A) + b^2 Var(B) + c^2 Var(C)$$

because A, B, and C are independent of each other. And since $Var(A) = 2Var(B) = 3Var(C)$, we can derive that:

$$Var(P) = Var(A)\left(a^2 + \frac{b^2}{2} + \frac{c^2}{3}\right)$$

We want to:

$$\begin{array}{ll} \text{minimize} & P(a, b, c) \\ \text{subject to} & g(a, b, c) = 0 \end{array}$$

where $g(a, b, c) = a + b + c - 1$, so $g(a, b, c) = 0$ is our constraint.

Using the Lagrange multiplier method, we can define:

$$\mathcal{L}(a, b, c, \lambda) = P(a, b, c) - \lambda \cdot g(a, b, c)$$

So to find our local minima we need to solve:

$$\nabla_{a,b,c,\lambda} \mathcal{L} = 0$$

$$\left(\frac{\partial \mathcal{L}}{\partial a}, \frac{\partial \mathcal{L}}{\partial b}, \frac{\partial \mathcal{L}}{\partial c}, \frac{\partial \mathcal{L}}{\partial \lambda}\right) = \left(2a - \lambda, b - \lambda, \frac{2}{3}c - \lambda, -(a + b + c - 1)\right) = \mathbf{0}$$

$$\Rightarrow \begin{cases} 2a - \lambda = 0 \\ b - \lambda = 0 \\ 2c/3 - \lambda = 0 \\ a + b + c - 1 = 0 \end{cases} \Rightarrow \begin{cases} a = \lambda/2 \\ b = \lambda \\ c = 3\lambda/2 \\ \frac{\lambda}{2} + \lambda + \frac{3\lambda}{2} = 3\lambda = 1 \end{cases} \Rightarrow \begin{cases} \mathbf{a = \frac{1}{6} = 0.1667} \\ \mathbf{b = \frac{1}{3} = 0.3333} \\ \mathbf{c = \frac{1}{2} = 0.5} \end{cases}$$

Let's prove the result in R:

```
payoff <- function(x) {
  a <- x[1]
  b <- x[2]
  c <- x[3]
  a^2 + b^2/2 + c^2/3
}
gradient_payoff <- function(x) {
  a <- x[1]
  b <- x[2]
  c <- x[3]
  c(2*a, b, 2*c/3)
}
sol <- constrOptim(theta = c(.3, .3, .4), f = payoff, grad = gradient_payoff,
                   ui = rbind(c(1, 0, 0), c(0, 1, 0), c(0, 0, 1),
                              c(-1, 0, 0), c(0, -1, 0), c(0, 0, -1),
                              c(1, 1, 1), c(-1, -1, -1)),
                   ci = c(0, 0, 0, -1, -1, -1, 1-1e-6, -1-1e-6))
sol$par
```

```
## [1] 0.1666989 0.3333673 0.4999327
```

Question 3: Turtles

Next, suppose that the lifespan of a species of turtle follows a uniform distribution over $[\mathbf{0}, \theta]$. Here, parameter θ represents the unknown maximum lifespan. You have a random sample of n individuals, and measure the lifespan of each individual i to be $\mathbf{y_i}$.

1. **Write down the likelihood function, $\mathbf{l}(\theta)$ in terms of $\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n}$.**

$$l(\theta; y_1, \ldots, y_n) = f(y_1, \ldots, y_n|\theta) = \prod_{i=1}^{n} f(y_i|\theta) = \begin{cases} \prod_{i=1}^{n} \frac{1}{\theta} = \theta^{-n} & 0 \leq y_i \leq \theta \\ 0 & \text{otherwise} \end{cases}$$

2. **Based on the previous result, what is the maximum-likelihood estimator for θ?**

The MLE of θ must be a value of θ for which $\theta \geq y_i$ for $i = 1, \ldots, n$ and which maximizes $1/\theta^n$ among all such values. I.e., the maximum value of y_i within the sample.

$$\hat{\theta}_{ml} = \arg\max_{\theta \in \Theta} \; \hat{l}(\theta; y_1, \ldots, y_n) = \max\{y_1, \ldots, y_n\}$$

3. **Let $\hat{\theta}_{\mathbf{ml}}$ be the maximum likelihood estimator above. For the simple case that $\mathbf{n = 1}$, what is the expectation of $\hat{\theta}_{\mathbf{ml}}$, given θ?**

$$E\left(\hat{\theta}_{ml} \mid \theta\right) = E(y_1) = E(y) = \int_{y=0}^{\theta} \frac{y}{\theta} \cdot dy = \left[\frac{y^2}{2\theta}\right]_{y=0}^{\theta} = \frac{\theta}{2}$$

4. **Is the maximum likelihood estimator biased?**

Yes, it is:

$$E\left(\hat{\theta}_{ml}\right) - \theta = \frac{\theta}{2} \neq 0$$

5. **For the more general case that $\mathbf{n \geq 1}$, what is the expectation of $\hat{\theta}_{\mathbf{ml}}$?**

Without loss of generality, let's suppose the individual n is the one with the maximum lifespan among the sample, i.e., $y_n \geq y_i$ for $i = 1, \ldots, n-1$. Call z that maximum value of y_i.

$$E[max\{y_1, \ldots, y_n\}] = E(y_n) = E(z) = \int_{z=0}^{\theta} z \cdot f(z) dz$$

But what is the distribution of z?

$$F(z) = \Pr(y_n \leq z) = \Pr(y_1 \leq z \cap \cdots \cap y_n \leq z) \overset{\text{i.i.d}}{=} \prod_{i=1}^{n} \Pr(y_i \leq z) = \left(\frac{z}{\theta}\right)^n \Rightarrow f(z) = \frac{n z^{n-1}}{\theta^n}$$

$$E[max\{y_1, \ldots, y_n\}] = \frac{n}{\theta^n} \int_{z=0}^{\theta} z^n dz = \frac{n}{\theta^n} \left[\frac{z^{n+1}}{n+1}\right]_{z=0}^{\theta} = \frac{n}{n+1} \cdot \theta$$

(Which confirms the previous result, for $n = 1$.)

6. **Is the maximum likelihood estimator consistent?**

It is:

$$\Pr\left(|\hat{\theta}_{ml} - \theta| > \varepsilon\right) = \Pr\left(\frac{\theta}{n+1} > \varepsilon\right) \to 0 \text{ as } n \to \infty$$

```
simulations <- 1e3 # number of simulations
theta <- 100 # an arbitrary value of theta
y <- runif(n = simulations, min = 0, max = theta) # Y ~ U(0,theta)
# any(y == theta); all(y < theta) # FALSE and TRUE, respectively
# No matter how large is the sample, Yi is always lower than 1
set.seed(1)
num_simulations <- sort(c(1, sample(c(2:simulations), 49)))
theta_mle <- unlist(lapply(num_simulations, function(n)
  mean(max(runif(n = n, min = 0, max = theta)))))
plot(num_simulations, theta_mle, ylim = c(floor(min(theta_mle)), theta),
     xlab = "Number of simulations", ylab = "MLE of theta", pch = '*')
lines(num_simulations, theta_mle, lwd = 0.5, col = 'blue')
```

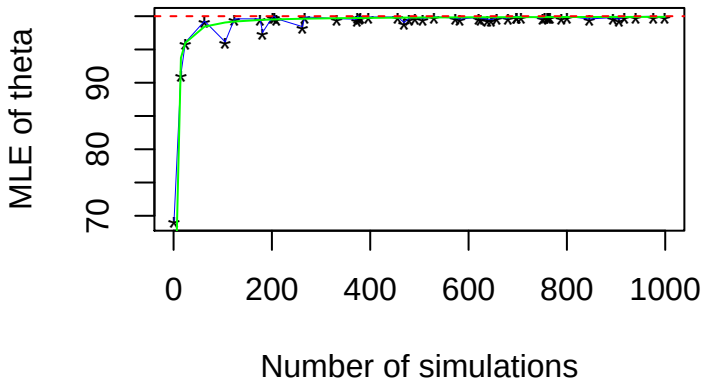


Figure 5: Trend of the MLE of θ depending on the sample size

Question 4: CLM 1

Background

The file `WageData2.csv` contains a dataset that has been used to quantify the impact of education on wage. One of the reasons we are proving another wage-equation exercise is that this area by far has the most (and most well-known) applications of instrumental variable techniques, the endogenity problem is obvious in this context, and the datasets are easy to obtain.

The Data

You are given a sample of 1000 individuals with their wage, education level, age, working experience, race (as an indicator), father's and mother's education level, whether the person lived in a rural area, whether the person lived in a city, IQ score, and two potential instruments, called z1 and z2.

The dependent variable of interest is `wage` (or its transformation), and we are interested in measuring "return" to education, where return is measured in the increase (hopefully) in wage with an additional year of education.

Question 4.1

Conduct an univariate analysis (using tables, graphs, and descriptive statistics found in the last 7 lectures) of all of the variables in the dataset.

```
# QUESTION 4 ------------------------------------------------------------
# setwd('Lab2/data')
data <- read.csv('WageData2.csv')
summary(data)
```

```
##        X              wage          education       experience
##  Min.   :   5.0   Min.   : 127.0   Min.   : 2.00   Min.   : 0.000
##  1st Qu.: 715.5   1st Qu.: 400.0   1st Qu.:12.00   1st Qu.: 6.000
##  Median :1431.5   Median : 543.0   Median :12.00   Median : 8.000
##  Mean   :1466.7   Mean   : 578.8   Mean   :13.22   Mean   : 8.788
##  3rd Qu.:2212.0   3rd Qu.: 702.5   3rd Qu.:16.00   3rd Qu.:11.000
##  Max.   :3009.0   Max.   :2404.0   Max.   :18.00   Max.   :23.000
##
##       age          raceColor     dad_education   mom_education
##  Min.   :24.00   Min.   :0.000   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:25.00   1st Qu.:0.000   1st Qu.: 8.00   1st Qu.: 8.00
##  Median :27.00   Median :0.000   Median :11.00   Median :12.00
##  Mean   :28.01   Mean   :0.238   Mean   :10.18   Mean   :10.45
##  3rd Qu.:30.00   3rd Qu.:0.000   3rd Qu.:12.00   3rd Qu.:12.00
##  Max.   :34.00   Max.   :1.000   Max.   :18.00   Max.   :18.00
##                                  NA's   :239     NA's   :128
##      rural            city             z1              z2
##  Min.   :0.000   Min.   :0.000   Min.   :0.00   Min.   :0.000
##  1st Qu.:0.000   1st Qu.:0.000   1st Qu.:0.00   1st Qu.:0.000
##  Median :0.000   Median :1.000   Median :0.00   Median :1.000
##  Mean   :0.391   Mean   :0.712   Mean   :0.44   Mean   :0.686
##  3rd Qu.:1.000   3rd Qu.:1.000   3rd Qu.:1.00   3rd Qu.:1.000
```

```
##  Max.   :1.000   Max.   :1.000   Max.   :1.00   Max.   :1.000
##
##     IQscore         logWage
##  Min.   : 50.0   Min.   :4.844
##  1st Qu.: 93.0   1st Qu.:5.991
##  Median :103.0   Median :6.297
##  Mean   :102.3   Mean   :6.263
##  3rd Qu.:113.0   3rd Qu.:6.555
##  Max.   :144.0   Max.   :7.785
##  NA's   :316
```

```
round(stat.desc(data, desc = TRUE, basic = TRUE), 2)
```

```
##                      X      wage education experience      age raceColor
## nbr.val         1000.00   1000.00   1000.00    1000.00  1000.00   1000.00
## nbr.null           0.00      0.00      0.00       3.00     0.00    762.00
## nbr.na             0.00      0.00      0.00       0.00     0.00      0.00
## min                5.00    127.00      2.00       0.00    24.00      0.00
## max             3009.00   2404.00     18.00      23.00    34.00      1.00
## range           3004.00   2277.00     16.00      23.00    10.00      1.00
## sum          1466678.00 578783.00  13219.00    8788.00 28007.00    238.00
## median          1431.50    543.00     12.00       8.00    27.00      0.00
## mean            1466.68    578.78     13.22       8.79    28.01      0.24
## SE.mean           27.40      8.43      0.09       0.13     0.10      0.01
## CI.mean.0.95      53.77     16.54      0.17       0.26     0.19      0.03
## var           750882.89  71058.93      7.45      17.82     9.72      0.18
## std.dev          866.53    266.57      2.73       4.22     3.12      0.43
## coef.var           0.59      0.46      0.21       0.48     0.11      1.79
##              dad_education mom_education   rural    city      z1      z2
## nbr.val             761.00        872.00 1000.00 1000.00 1000.00 1000.00
## nbr.null              7.00          4.00  609.00  288.00  560.00  314.00
## nbr.na              239.00        128.00    0.00    0.00    0.00    0.00
## min                   0.00          0.00    0.00    0.00    0.00    0.00
## max                  18.00         18.00    1.00    1.00    1.00    1.00
## range                18.00         18.00    1.00    1.00    1.00    1.00
## sum                7748.00       9113.00  391.00  712.00  440.00  686.00
## median               11.00         12.00    0.00    1.00    0.00    1.00
## mean                 10.18         10.45    0.39    0.71    0.44    0.69
## SE.mean               0.14          0.11    0.02    0.01    0.02    0.01
## CI.mean.0.95          0.27          0.21    0.03    0.03    0.03    0.03
## var                  14.05          9.77    0.24    0.21    0.25    0.22
## std.dev               3.75          3.13    0.49    0.45    0.50    0.46
## coef.var              0.37          0.30    1.25    0.64    1.13    0.68
##               IQscore logWage
## nbr.val        684.00 1000.00
## nbr.null         0.00    0.00
## nbr.na         316.00    0.00
## min             50.00    4.84
## max            144.00    7.78
## range           94.00    2.94
## sum          69955.00 6262.77
## median         103.00    6.30
## mean           102.27    6.26
## SE.mean          0.61    0.01
```

```
## CI.mean.0.95     1.19    0.03
## var            250.99    0.20
## std.dev         15.84    0.45
## coef.var         0.15    0.07
```

```
# Lots of NAs, and some binary variables (raceColor, rural, city, z1, z2)
ggplot(data, aes(X)) +
  geom_histogram(aes(y = ..count..), color = "black", fill = "white",
                 bins = 20) +
  labs(x = "Value", y = "Number of observations",
       title = "Histogram of unnamed variable")
```

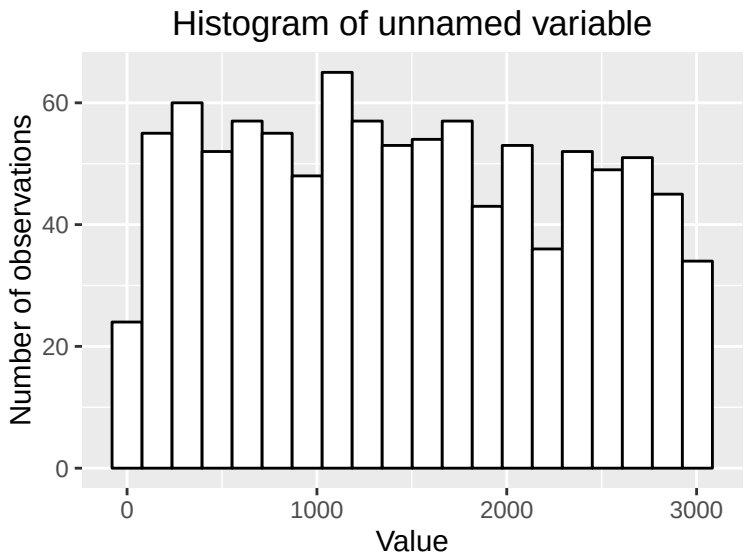


Figure 6: Histogram of unnamed variable X

```
data <- data %>% select(-X)
data_melt <- data %>% gather(variable, value)
ggplot(data_melt, aes(value)) +
  geom_histogram(aes(y = ..count..), color = "black", fill = "white",
                 bins = 20) +
  facet_wrap(~ variable, scales = "free") +
  labs(x = "Variable Value", y = "Number of observations",
       title = "Histogram of all variables in the dataset")
```

```
data_reduced <- data %>%
  select(which(names(data) %in% names(data)[sapply(data, function(x)
    length(levels(as.factor(x)))) > 2]))
pairs(data_reduced)
```

```
ggpairs(data_reduced %>% na.omit()) +
  theme(axis.ticks = element_blank(), axis.text =  element_blank())
```

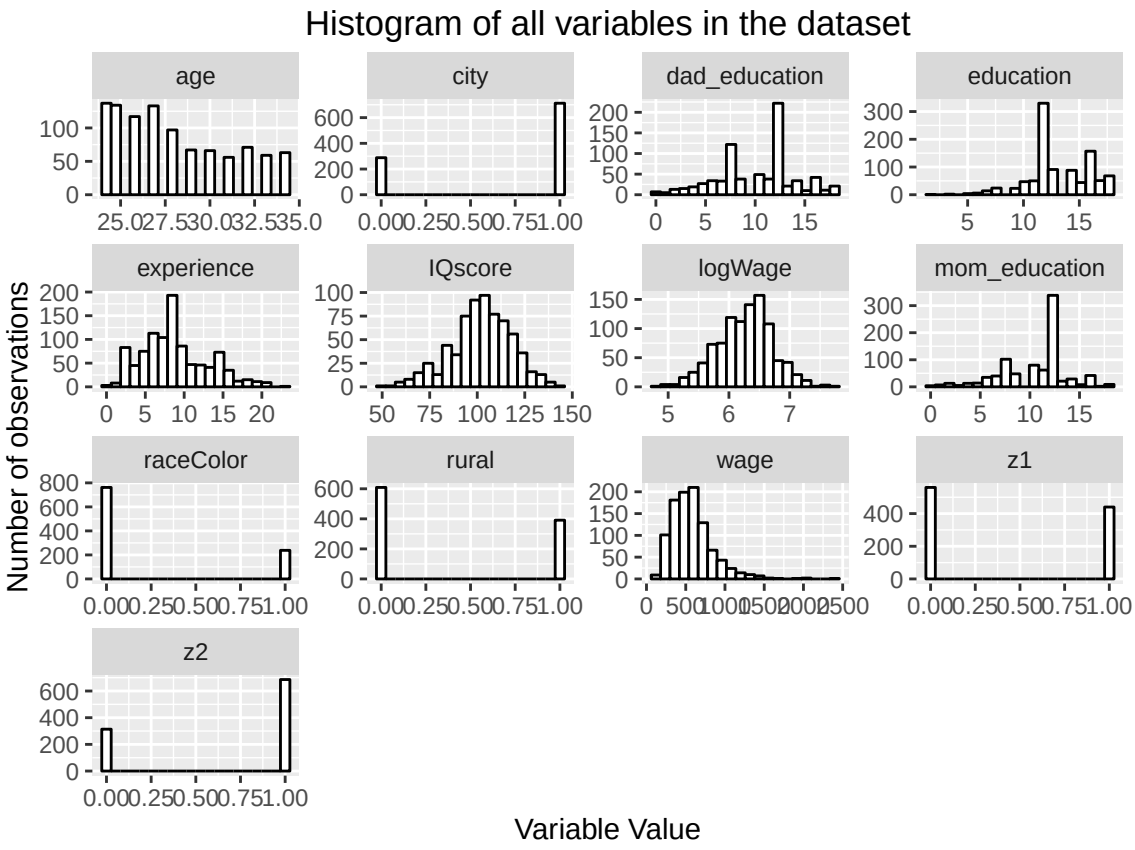


Figure 7: Histogram of all variables in the dataset

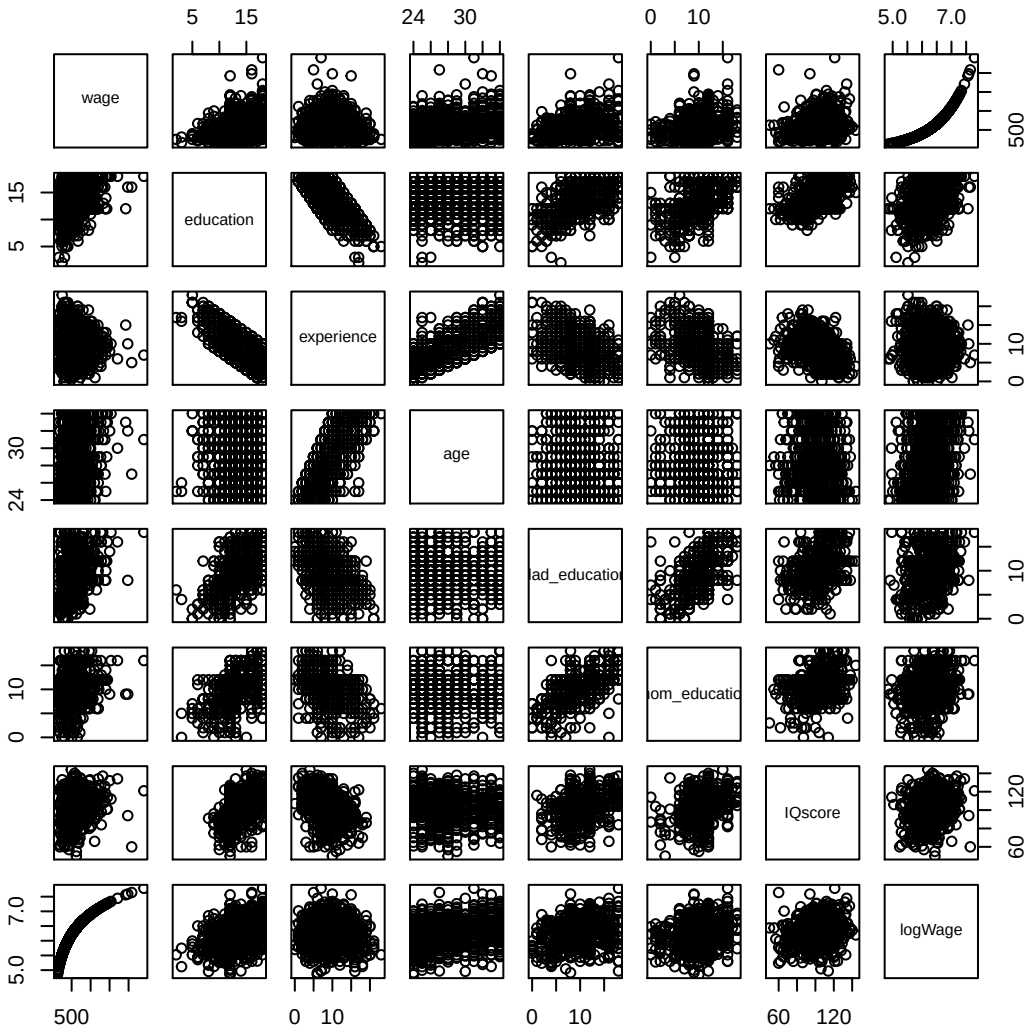


Figure 8: Scatterplot matrix

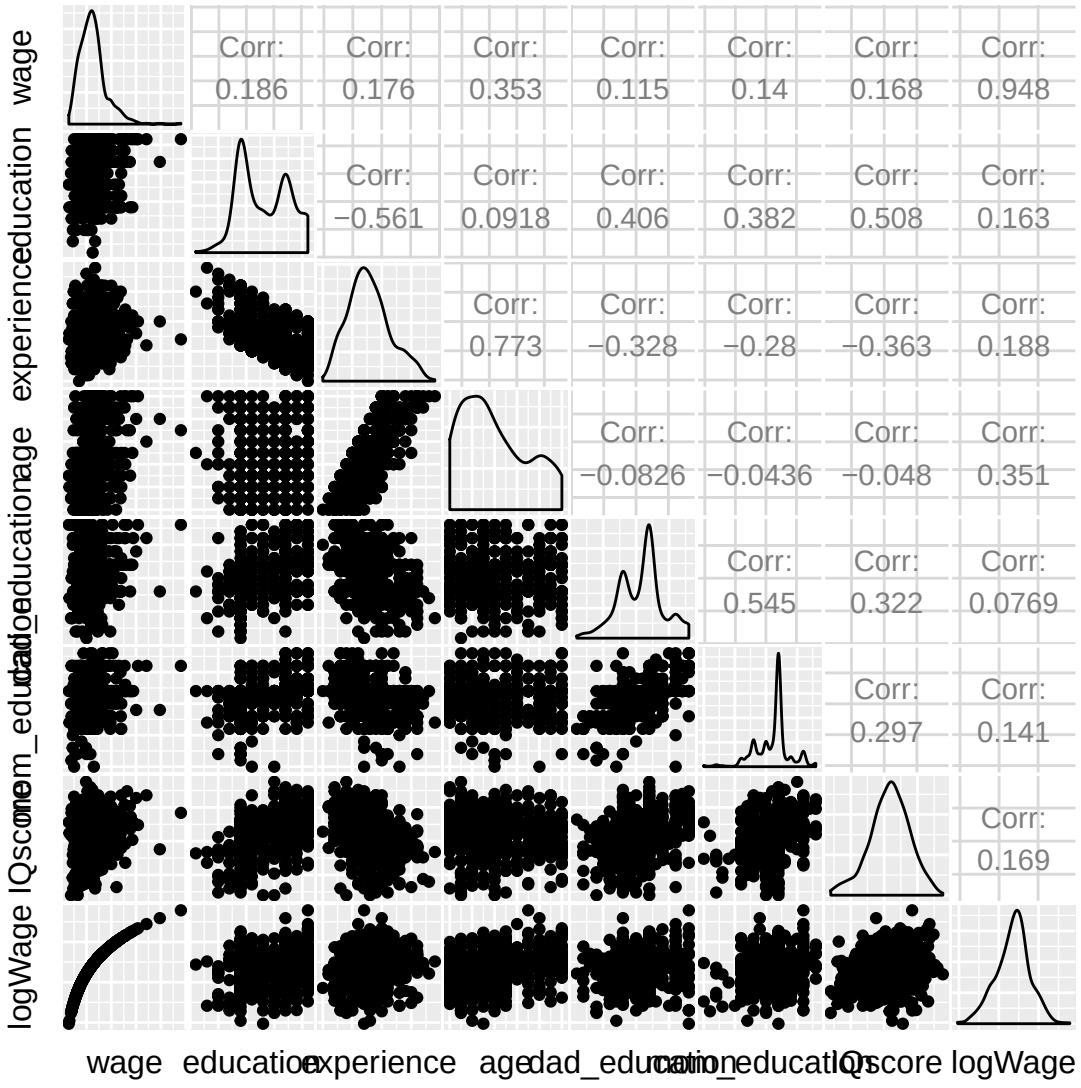


Figure 9: Scatterplot matrix omitting missing values (with correlations)

Also, create two variables: (1) natural log of wage (name it `logWage`) (2) square of experience (name it `experienceSquare`)

```
# Create two variables:
# (1) natural log of wage (name it `logWage`)
# (2) square of experience (name it `experienceSquare`)**
data <- data %>%
  mutate(logWage = log(wage), experienceSquare = experience^2)
```

Question 4.2

Conduct a bivariate analysis (using tables, graphs, descriptive statistics found in the last 7 lectures) of `wage` and `logWage` and all the other variables in the datasets.

Question 4.3

Regress *log(wage)* on education, experience, age, and raceColor.

1. **Report all the estimated coefficients, their standard errors, t-statistics, F-statistic of the regression, R^2, R^2_{adj} , and degrees of freedom.**
2. **Explain why the degrees of freedom takes on the specific value you observe in the regression output.**
3. **Describe any unexpected results from your regression and how you would resolve them (if the intent is to estimate return to education, condition on race and experience).**
4. **Interpret the coefficient estimate associated with education.**
5. **Interpret the coefficient estimate associated with experience.**

Question 4.4

Regress *log(wage)* on education, experience, experienceSquare, and race-Color.

1. **Plot a graph of the estimated effect of experience on wage.**
2. **What is the estimated effect of experience on wage when experience is 10 years?**

Question 4.5

Regress `logWage` on education, experience, `experienceSquare`, `raceColor`, `dad_education`, `mom_education`, rural, city.

1. **What are the number of observations used in this regression? Are missing values a problem? Analyze the missing values, if any, and see if there is any discernible pattern with wage, education, experience, and raceColor.**
2. **Do you just want to "throw away" these observations?**
3. **How about blindly replace all of the missing values with the average of the observed values of the corresponding variable? Rerun the original regression using all of the observations?**

4. **How about regress the variable(s) with missing values on education, experience, and raceColor, and use this regression(s) to predict (i.e., "impute") the missing values and then rerun the original regression using all of the observations?**

5. **Compare the results of all of these regressions. Which one, if at all, would you prefer?**

Question 4.6

1. **Consider using z_1 as the instrumental variable (IV) for education. What assumptions are needed on z_1 and the error term (call it, u)?**

2. **Suppose z_1 is an indicator representing whether or not an individual lives in an area in which there was a recent policy change to promote the importance of education. Could z_1 be correlated with other unobservables captured in the error term?**

3. **Using the same specification as that in Question 4.5, estimate the equation by 2SLS, using both z_1 and z_2 as instrument variables. Interpret the results. How does the coefficient estimate on education change?**

Question 5: CLM 2

The dataset, `wealthy candidates.csv`, contains candidate level electoral data from a developing country. Politically, each region (which is a subset of the country) is divided in to smaller electoral districts where the candidate with the most votes wins the seat. This dataset has data on the financial wealth and electoral performance (voteshare) of electoral candidates. We are interested in understanding whether or not wealth is an electoral advantage. In other words, do wealthy candidates fare better in elections than their less wealthy peers?

1. **Begin with a parsimonious, yet appropriate, specification. Why did you choose this model? Are your results statistically significant? Based on these results, how would you answer the research question? Is there a linear relationship between wealth and electoral performance?**

```
# QUESTION 5 ------------------------------------------------------------------
# setwd('Lab2/data')
data <- read.csv('wealthy_candidates.csv')
ggplot(data, aes(X)) +
  geom_histogram(aes(y = ..count..), color = "black", fill = "white",
                 bins = 20) +
  labs(x = "Value", y = "Number of observations",
       title = "Histogram of unnamed variable")
```

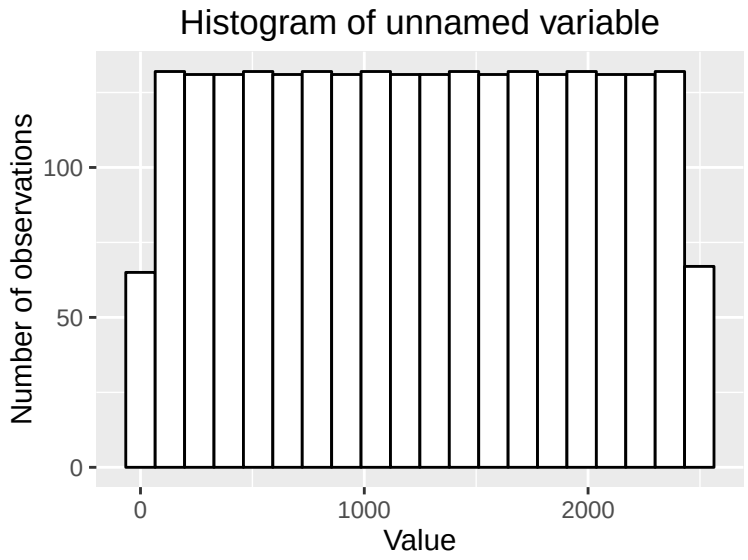


Figure 10: Histogram of unnamed variable X

```
data <- data %>% select(-X)
summary(data)
```

```
##       region          urb               lit           voteshare
##  Region 1:1183   Min.   :0.02835   Min.   :0.2418   Min.   :0.006037
##  Region 2: 690   1st Qu.:0.08387   1st Qu.:0.3846   1st Qu.:0.199620
##  Region 3: 625   Median :0.14657   Median :0.4602   Median :0.293398
```

```
##                  Mean   :0.18729   Mean   :0.4512   Mean   :0.287860
##                  3rd Qu.:0.24319   3rd Qu.:0.5105   3rd Qu.:0.367978
##                  Max.   :0.80234   Max.   :0.6524   Max.   :0.693324
##
##  absolute_wealth
##  Min.   :2.000e+00
##  1st Qu.:1.875e+05
##  Median :1.337e+06
##  Mean   :5.034e+06
##  3rd Qu.:4.092e+06
##  Max.   :1.216e+09
##  NA's   :1
```

```
# Only 1 NA in absolute_wealth
data[is.na(data$absolute_wealth), ]
```

```
##       region       urb       lit voteshare absolute_wealth
## 177 Region 2 0.4172694 0.5199646 0.1484079              NA
```

```
# The values of the other variables for that observations are not outliers
# We can omit that observation from our sample
data <- data %>% filter(!is.na(absolute_wealth))
# Region is categorical (3 possible values)
round(stat.desc(data[, names(sapply(data,
                                    is.factor))[!sapply(data, is.factor)]],
                desc = TRUE, basic = TRUE), 2)
```

```
##                    urb     lit voteshare absolute_wealth
## nbr.val        2497.00 2497.00   2497.00    2.497000e+03
## nbr.null          0.00    0.00      0.00    0.000000e+00
## nbr.na            0.00    0.00      0.00    0.000000e+00
## min               0.03    0.24      0.01    2.000000e+00
## max               0.80    0.65      0.69    1.216399e+09
## range             0.77    0.41      0.69    1.216399e+09
## sum             467.43 1126.64    718.93    1.257016e+10
## median            0.15    0.46      0.29    1.336629e+06
## mean              0.19    0.45      0.29    5.034105e+06
## SE.mean           0.00    0.00      0.00    6.223434e+05
## CI.mean.0.95      0.01    0.00      0.00    1.220362e+06
## var               0.02    0.01      0.02    9.671163e+14
## std.dev           0.15    0.09      0.12    3.109849e+07
## coef.var          0.79    0.20      0.43    6.180000e+00
```

```
data_melt <- data %>% gather(variable, value, - region)
ggplot(data_melt, aes(x=value)) +
  geom_histogram(aes(y = ..count..), alpha=0.6,
                 bins = 20, position = "dodge", fill = "white", color = "black") +
  facet_wrap(~ variable, scales = "free")
```

```
ggplot(data_melt, aes(x=value, fill=region, color = region)) +
  geom_histogram(aes(y = ..count..), alpha=0.6,
```

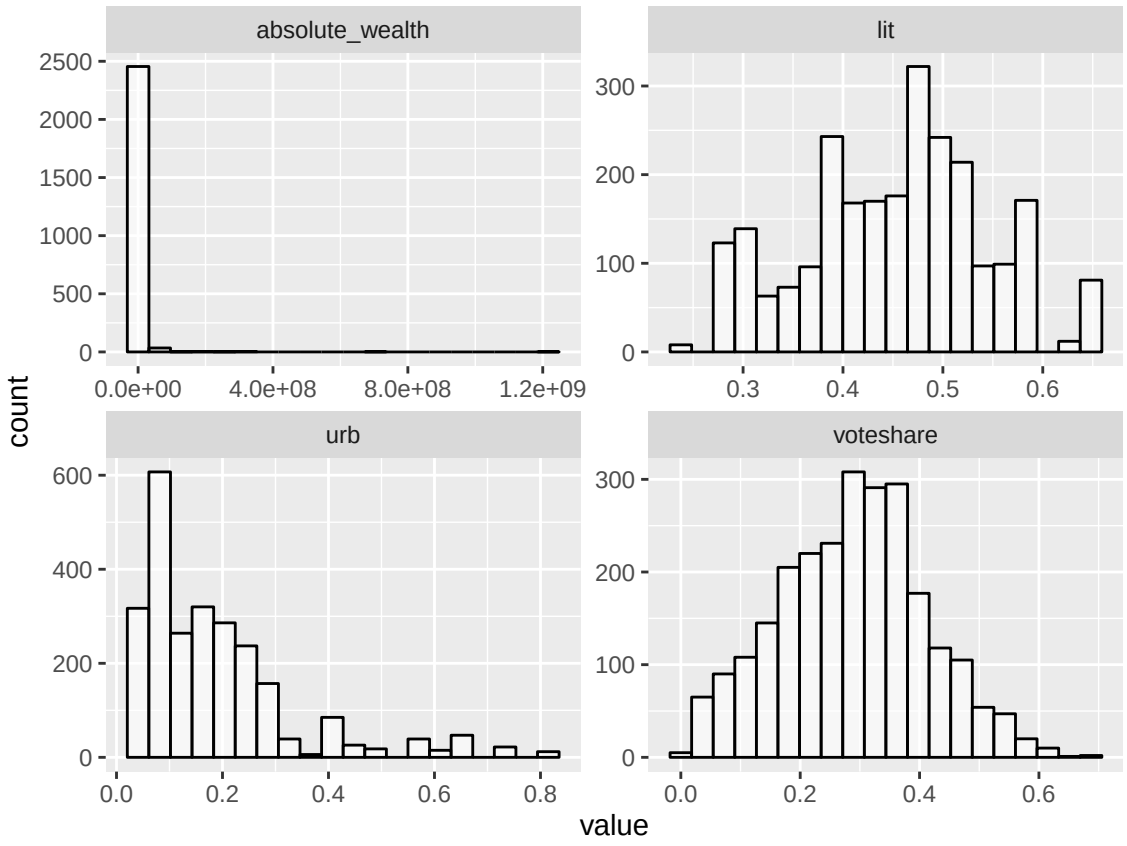


Figure 11: Histogram of all variables in the dataset

```
                bins = 20, position = "dodge") +
  facet_wrap(~ variable, scales = "free") +
  labs(x = "Variable Value", y = "Number of observations",
       title = "Histogram of all variables in the dataset per Region")
```

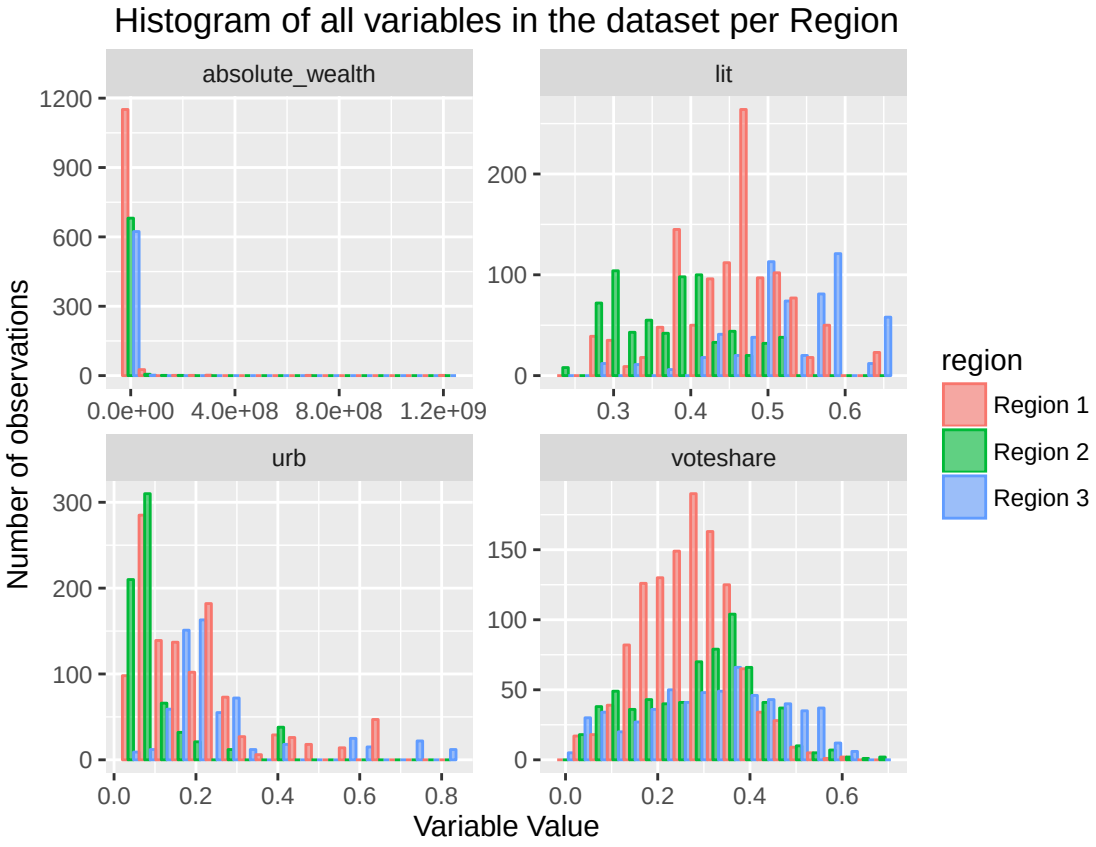


Figure 12: Histogram of all variables in the dataset per Region

```
data_reduced <- data %>%
  select(-region)
pairs(data_reduced)
```

```
ggpairs(data_reduced %>% sample_n(500)) +
  theme(axis.ticks = element_blank(), axis.text =  element_blank())
```

2. **A team-member suggests adding a quadratic term to your regression. Based on your prior model, is such an addition warranted? Add this term and interpret the results. Do wealthier candidates fare better in elections?**

3. **Another team member suggests that it is important to take into account the fact that different regions have different electoral contexts. In particular, the relationship between candidate wealth and electoral performance might be different across states. Modify your model and report your results. Test the hypothesis that this addition is not needed.**

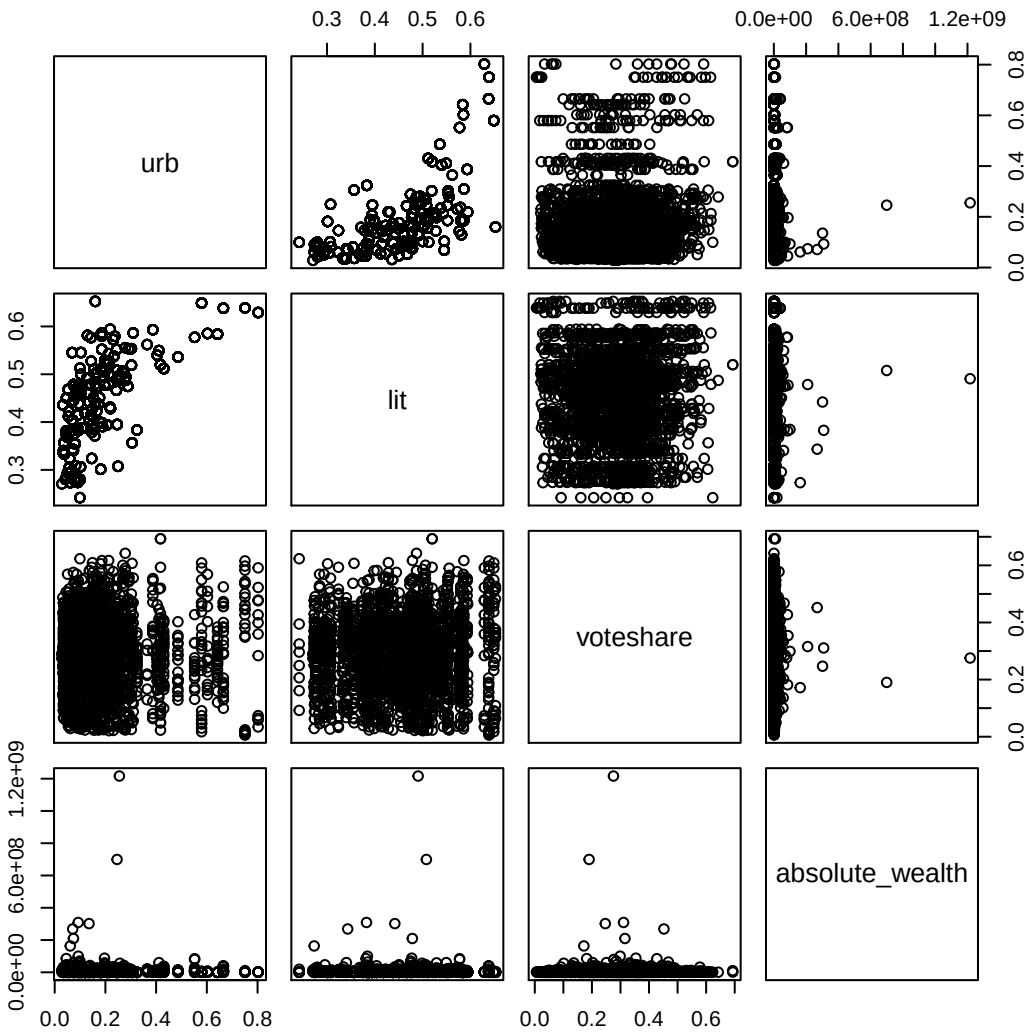


Figure 13: Scatterplot matrix

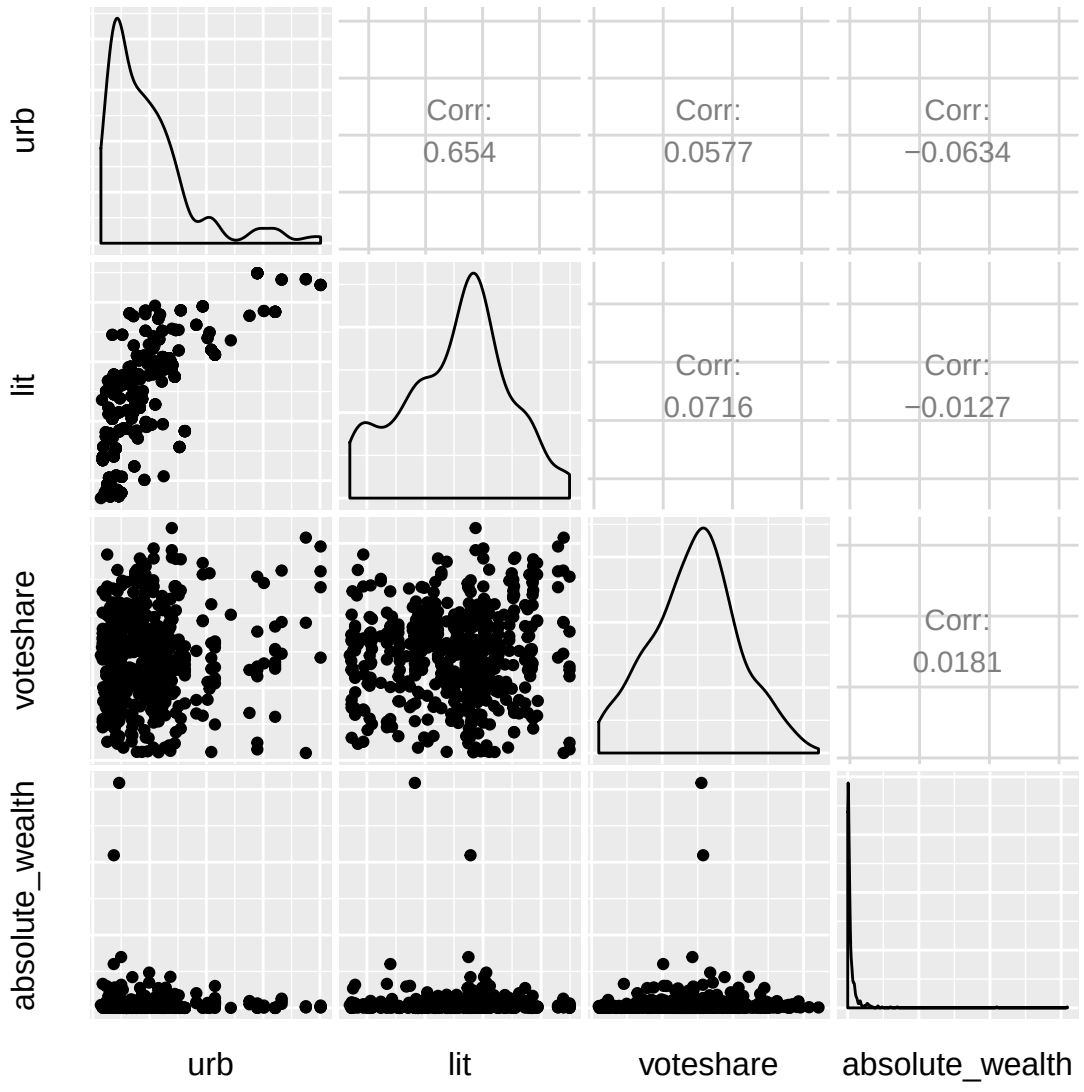


Figure 14: Scatterplot matrix of a sample of the dataset (with correlations)

4. **Return to your parsimonious model. Do you think you have found a causal and unbiased estimate? Please state the conditions under which you would have an unbiased and causal estimates. Do these conditions hold?**

5. **Someone proposes a difference in difference design. Please write the equation for such a model. Under what circumstances would this design yield a causal effect?**

Question 6: CLM 3

Your analytics team has been tasked with analyzing aggregate revenue, cost and sales data, which have been provided to you in the R workspace/data frame `retailSales.Rdata`.

Your task is two fold. First, your team is to develop a model for predicting (forecasting) revenues. Part of the model development documentation is a backtesting exercise where you train your model using data from the first two years and evaluate the model's forecasts using the last two years of data.

Second, management is equally interested in understanding variables that might affect revenues in support of management adjustments to operations and revenue forecasts. You are also to identify factors that affect revenues, and discuss how useful management's planned revenue is for forecasting revenues.

Your analysis should address the following:

- **Exploratory Data Analysis: focus on bivariate and multivariate relationships.**
- **Be sure to assess conditions and identify unusual observations.**

First we explore the whole dataset.

```
load("retailSales.Rdata")
data <- retailSales; rm(retailSales)
summary(data)
```

```
##       Year                          Product.line
##  Min.   :2004   Camping Equipment       :24108
##  1st Qu.:2005   Golf Equipment          : 8820
##  Median :2006   Mountaineering Equipment:12348
##  Mean   :2006   Outdoor Protection      : 8820
##  3rd Qu.:2006   Personal Accessories    :30576
##  Max.   :2007
##
##                Product.type                  Product
##  Eyewear              : 9408   Aloe Relief       :  588
##  Watches              : 7644   Astro Pilot       :  588
##  Lanterns             : 7056   Auto Pilot        :  588
##  Cooking Gear         : 5880   Bear Edge         :  588
##  Navigation           : 5880   Bear Survival Edge:  588
##  Climbing Accessories: 4116   Bella             :  588
##  (Other)              :44688   (Other)           :81144
##    Order.method.type  Retailer.country    Revenue
##  E-mail     :12096   Australia: 4032   Min.   :       0
##  Fax        :12096   Austria  : 4032   1st Qu.:   18579
##  Mail       :12096   Belgium  : 4032   Median :   59867
##  Sales visit:12096   Brazil   : 4032   Mean   :  189418
##  Special    :12096   Canada   : 4032   3rd Qu.:  190193
##  Telephone  :12096   China    : 4032   Max.   :10054289
##  Web        :12096   (Other)  :60480   NA's   :59929
##  Planned.revenue     Product.cost        Quantity        Unit.cost
##  Min.   :      16   Min.   :      6   Min.   :     1   Min.   :  0.85
##  1st Qu.:   19557   1st Qu.:   9432   1st Qu.:   328   1st Qu.: 11.43
```

```
##  Median :   63907   Median :  32784   Median :  1043   Median : 36.83
##  Mean   :  198818   Mean   : 111625   Mean   :  3607   Mean   : 84.89
##  3rd Qu.:  203996   3rd Qu.: 111371   3rd Qu.:  3288   3rd Qu.: 80.00
##  Max.   :10054289   Max.   :6756853   Max.   :313628   Max.   :690.00
##  NA's   :59929      NA's   :59929     NA's   :59929    NA's   :59929
##    Unit.price       Gross.profit     Unit.sale.price
##  Min.   :   2.06   Min.   : -18160   Min.   :   0.00
##  1st Qu.:  23.00   1st Qu.:   8333   1st Qu.:  20.15
##  Median :  66.77   Median :  25794   Median :  62.65
##  Mean   : 155.99   Mean   :  77793   Mean   : 147.23
##  3rd Qu.: 148.30   3rd Qu.:  78254   3rd Qu.: 140.96
##  Max.   :1359.72   Max.   :3521098   Max.   :1307.80
##  NA's   :59929     NA's   :59929     NA's   :59929
```

The dataset contains 84,672 observations of 14 variables. 5 of them are categorical (`Product.line`, `Product.type`, `Product`, `Order.method.type`, `Retailer.country`), and `Year` should also be considered as categorical, since there are data from only 4 years (from 2004 to 2007).

```
data <- data %>%
  mutate(Year = as.factor(Year))
```

We also notice (from the output of `summary`) that some of the variables (all of them numerical) has a high number of `NA`s, the same in all cases (59929, i.e., 70.78% of the total number of observations). Do the `NA`s appear in the same observations for all those variables? Yes, they do.

```
# data_isNA <- as.data.frame(sapply(data, is.na))
data_isNA <- data %>% mutate_each(funs(is.na(.)))
head(data_isNA)
```

```
##    Year Product.line Product.type Product Order.method.type
## 1 FALSE        FALSE        FALSE   FALSE             FALSE
## 2 FALSE        FALSE        FALSE   FALSE             FALSE
## 3 FALSE        FALSE        FALSE   FALSE             FALSE
## 4 FALSE        FALSE        FALSE   FALSE             FALSE
## 5 FALSE        FALSE        FALSE   FALSE             FALSE
## 6 FALSE        FALSE        FALSE   FALSE             FALSE
##   Retailer.country Revenue Planned.revenue Product.cost Quantity Unit.cost
## 1            FALSE   FALSE           FALSE        FALSE    FALSE     FALSE
## 2            FALSE   FALSE           FALSE        FALSE    FALSE     FALSE
## 3            FALSE    TRUE            TRUE         TRUE     TRUE      TRUE
## 4            FALSE    TRUE            TRUE         TRUE     TRUE      TRUE
## 5            FALSE   FALSE           FALSE        FALSE    FALSE     FALSE
## 6            FALSE    TRUE            TRUE         TRUE     TRUE      TRUE
##   Unit.price Gross.profit Unit.sale.price
## 1      FALSE        FALSE           FALSE
## 2      FALSE        FALSE           FALSE
## 3       TRUE         TRUE            TRUE
## 4       TRUE         TRUE            TRUE
## 5      FALSE        FALSE           FALSE
## 6       TRUE         TRUE            TRUE
```

```
# vars_with_NAs <- apply(data_isNA, 2, sum)
vars_with_NAs <- data_isNA %>% summarise_each(funs(sum))
(vars_with_NAs <- names(vars_with_NAs)[vars_with_NAs>0])
```

```
## [1] "Revenue"         "Planned.revenue" "Product.cost"    "Quantity"
## [5] "Unit.cost"       "Unit.price"      "Gross.profit"    "Unit.sale.price"
```

```
sapply(data_isNA[, vars_with_NAs[-1]], identical,
       as.vector(data_isNA[, vars_with_NAs[1]]))
```

```
## Planned.revenue    Product.cost        Quantity       Unit.cost
##            TRUE            TRUE            TRUE            TRUE
##      Unit.price    Gross.profit Unit.sale.price
##            TRUE            TRUE            TRUE
```

And the amount of NAs per category is roughly the same for all categorical values (or at least there are non-missing data for all categories; below we just show the percentage per category for three of the numerical variables).

```
data_categorical <- data %>%
  select(which(names(data) %in% names(data)[sapply(data, is.factor)])) %>%
  mutate_each(funs(as.character(.))) %>% mutate(Revenue = data$Revenue)
data_categorical %>%
  select(Revenue, Year) %>%
  group_by(Year) %>%
  summarise_each(funs(100*mean(is.na(.)))) %>%
  rename("% of NAs in numerical variables" = Revenue)
```

```
## Source: local data frame [4 x 2]
##
##    Year % of NAs in numerical variables
##   (chr)                           (dbl)
## 1  2004                        67.95163
## 2  2005                        65.49981
## 3  2006                        71.70257
## 4  2007                        77.95729
```

```
data_categorical %>%
  select(Revenue, Product.line) %>%
  group_by(Product.line) %>%
  summarise_each(funs(100*mean(is.na(.)))) %>%
  rename("% of NAs in numerical variables" = Revenue)
```

```
## Source: local data frame [5 x 2]
##
##               Product.line % of NAs in numerical variables
##                      (chr)                           (dbl)
## 1        Camping Equipment                        65.26049
## 2           Golf Equipment                        68.67347
## 3 Mountaineering Equipment                        76.13379
## 4       Outdoor Protection                        66.62132
## 5     Personal Accessories                        74.77106
```

```
## Source: local data frame [21 x 2]
##
##    Retailer.country % of NAs in numerical variables
##                (chr)                           (dbl)
## 1          Australia                        77.15774
## 2            Austria                        72.44544
## 3            Belgium                        75.99206
## 4             Brazil                        81.49802
## 5             Canada                        57.66369
## 6              China                        77.33135
## 7            Denmark                        80.28274
## 8            Finland                        79.46429
## 9             France                        60.49107
## 10           Germany                        59.37500
## 11             Italy                        69.07242
## 12             Japan                        58.60615
## 13             Korea                        74.47917
## 14            Mexico                        73.36310
## 15       Netherlands                        70.03968
## 16         Singapore                        70.70933
## 17             Spain                        71.55258
## 18            Sweden                        74.25595
## 19       Switzerland                        80.03472
## 20    United Kingdom                        70.23810
## 21     United States                        52.28175
```

So we can ommit all those missing observations (reducing our sample size to 24743), and continue with a further analysis of the numerical variables:

```
data <- data %>% na.omit()
data_categorical <- data %>%
  select(which(names(data) %in% names(data)[sapply(data, is.factor)]))
data_non_categorical <- data %>%
  select(which(names(data) %in% names(data)[!sapply(data, is.factor)]))
round(stat.desc(data_non_categorical, desc = TRUE, basic = TRUE), 2)
```

```
##                   Revenue Planned.revenue Product.cost    Quantity
## nbr.val      2.474300e+04    2.474300e+04 2.474300e+04    24743.00
## nbr.null     7.600000e+01    0.000000e+00 0.000000e+00        0.00
## nbr.na       0.000000e+00    0.000000e+00 0.000000e+00        0.00
## min          0.000000e+00    1.569000e+01 5.760000e+00        1.00
## max          1.005429e+07    1.005429e+07 6.756853e+06   313628.00
## range        1.005429e+07    1.005427e+07 6.756847e+06   313627.00
## sum          4.686776e+09    4.919342e+09 2.761941e+09 89237091.00
## median       5.986727e+04    6.390684e+04 3.278372e+04     1043.00
## mean         1.894182e+05    1.988175e+05 1.116251e+05     3606.56
## SE.mean      2.484130e+03    2.559050e+03 1.515680e+03       55.80
## CI.mean.0.95 4.869040e+03    5.015880e+03 2.970830e+03      109.38
## var          1.526863e+11    1.620349e+11 5.684198e+10 77048387.56
## std.dev      3.907509e+05    4.025355e+05 2.384156e+05     8777.72
## coef.var     2.060000e+00    2.020000e+00 2.140000e+00        2.43
##               Unit.cost Unit.price  Gross.profit Unit.sale.price
## nbr.val        24743.00   24743.00  2.474300e+04        24743.00
```

```
## nbr.null           0.00       0.00  0.000000e+00          76.00
## nbr.na             0.00       0.00  0.000000e+00           0.00
## min                0.85       2.06 -1.815960e+04           0.00
## max              690.00    1359.72  3.521098e+06        1307.80
## range            689.15    1357.66  3.539257e+06        1307.80
## sum          2100344.99 3859701.42  1.924835e+09     3642909.71
## median            36.83      66.77  2.579376e+04          62.65
## mean              84.89     155.99  7.779311e+04         147.23
## SE.mean            0.83       1.57  1.005230e+03           1.48
## CI.mean.0.95       1.63       3.08  1.970320e+03           2.89
## var            17190.71   60912.60  2.500267e+10       53846.65
## std.dev          131.11     246.80  1.581223e+05         232.05
## coef.var           1.54       1.58  2.030000e+00           1.58
```

All numerical variables are right-skewed, with long right tails (i.e., with several observations more than 2 standard deviations far from the mean), especially the ones corresponding to aggregate—non-unit—results.

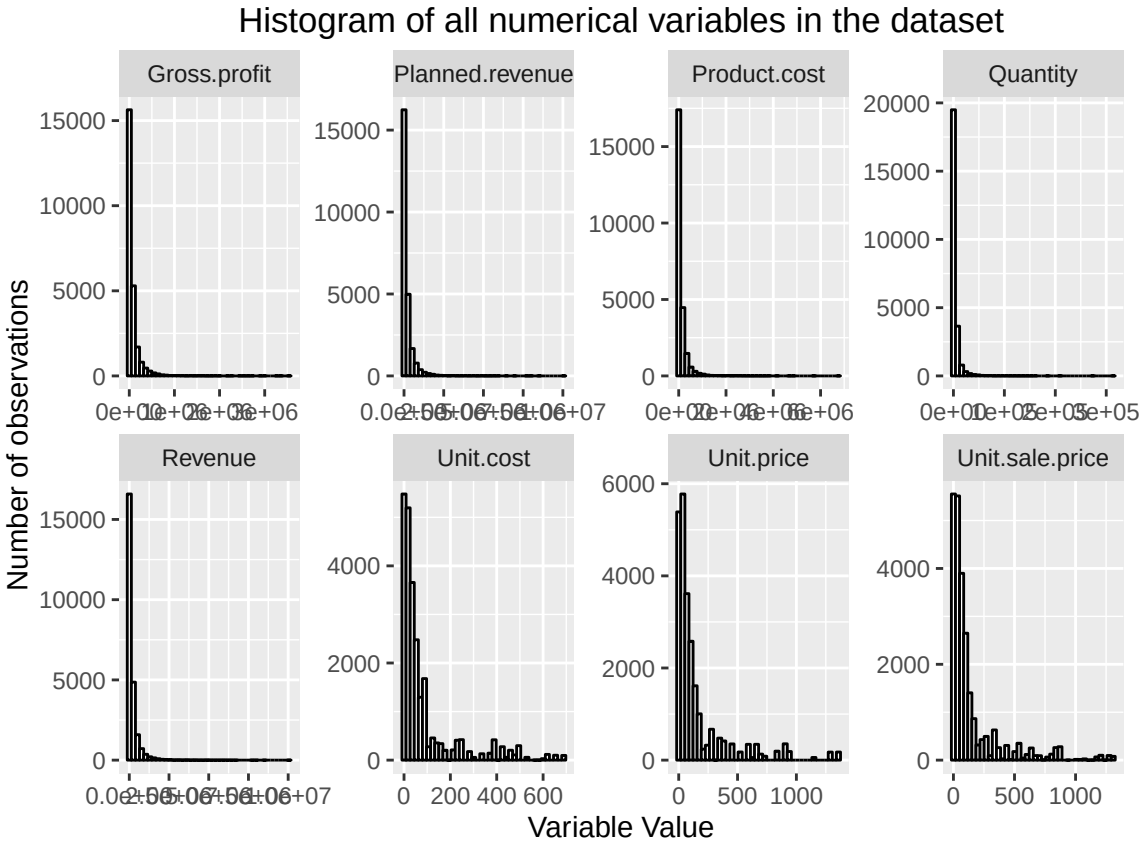


Figure 15: Histogram of all non-categorical variables in the dataset

Below we show the correlation matrix of the numerical variables, as well as two different representations of the scatterplot matrix (where we've used a sample of the data of size 500 because the plotting functions consume a lot of resources; that's why the correlations shown in the second Figure, only approximate, differ from the ones shown right below). As we might have expected, the correlations between `Revenue`,

`Planned.revenue`, `Product.cost`, and `Gross.profit` (i.e., the aggregate values), as well as those between `Unit.cost`, `Unit.price`, and `Unit.sale.price` (i.e., the values per unit), are positive and very high. `Quantity` is negatively correlated with the unitary variables (but that correlation is negligible in absolute value), and is moderately correlated ($\rho \simeq 0.5$) with the aggregate values.

```
cor(data_non_categorical)
```

```
##                   Revenue Planned.revenue Product.cost   Quantity
## Revenue         1.0000000       0.9990586    0.9903575  0.5055979
## Planned.revenue 0.9990586       1.0000000    0.9895792  0.4994770
## Product.cost    0.9903575       0.9895792    1.0000000  0.5061298
## Quantity        0.5055979       0.4994770    0.5061298  1.0000000
## Unit.cost       0.2463441       0.2550054    0.2415089 -0.1687497
## Unit.price      0.2332806       0.2421026    0.2194407 -0.1677662
## Gross.profit    0.9779407       0.9767878    0.9395732  0.4862920
## Unit.sale.price 0.2360448       0.2444078    0.2220105 -0.1674531
##                  Unit.cost Unit.price Gross.profit Unit.sale.price
## Revenue          0.2463441  0.2332806    0.9779407       0.2360448
## Planned.revenue  0.2550054  0.2421026    0.9767878       0.2444078
## Product.cost     0.2415089  0.2194407    0.9395732       0.2220105
## Quantity        -0.1687497 -0.1677662    0.4862920      -0.1674531
## Unit.cost        1.0000000  0.9886870    0.2446187       0.9889263
## Unit.price       0.9886870  1.0000000    0.2456107       0.9992750
## Gross.profit     0.2446187  0.2456107    1.0000000       0.2485667
## Unit.sale.price  0.9889263  0.9992750    0.2485667       1.0000000
```

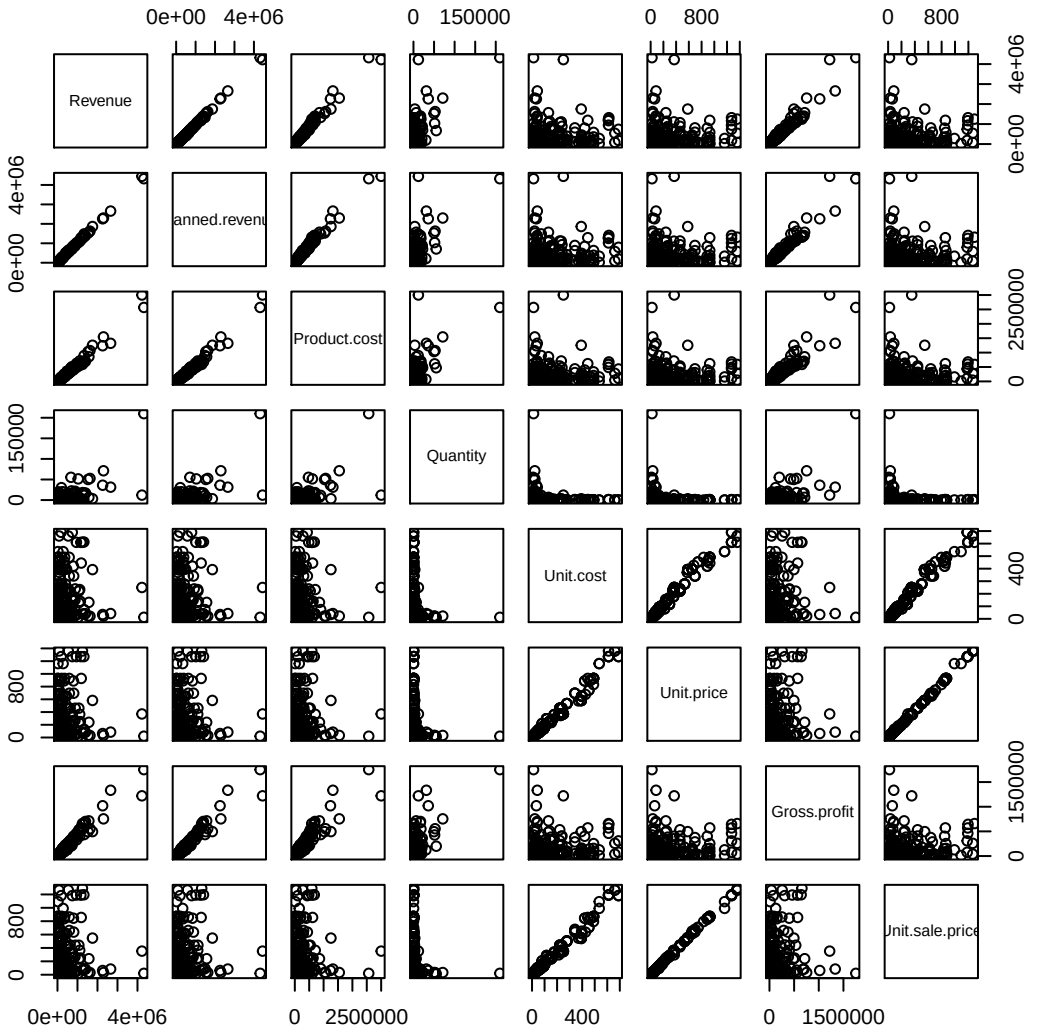


Figure 16: Scatterplot matrix of a sample of the dataset

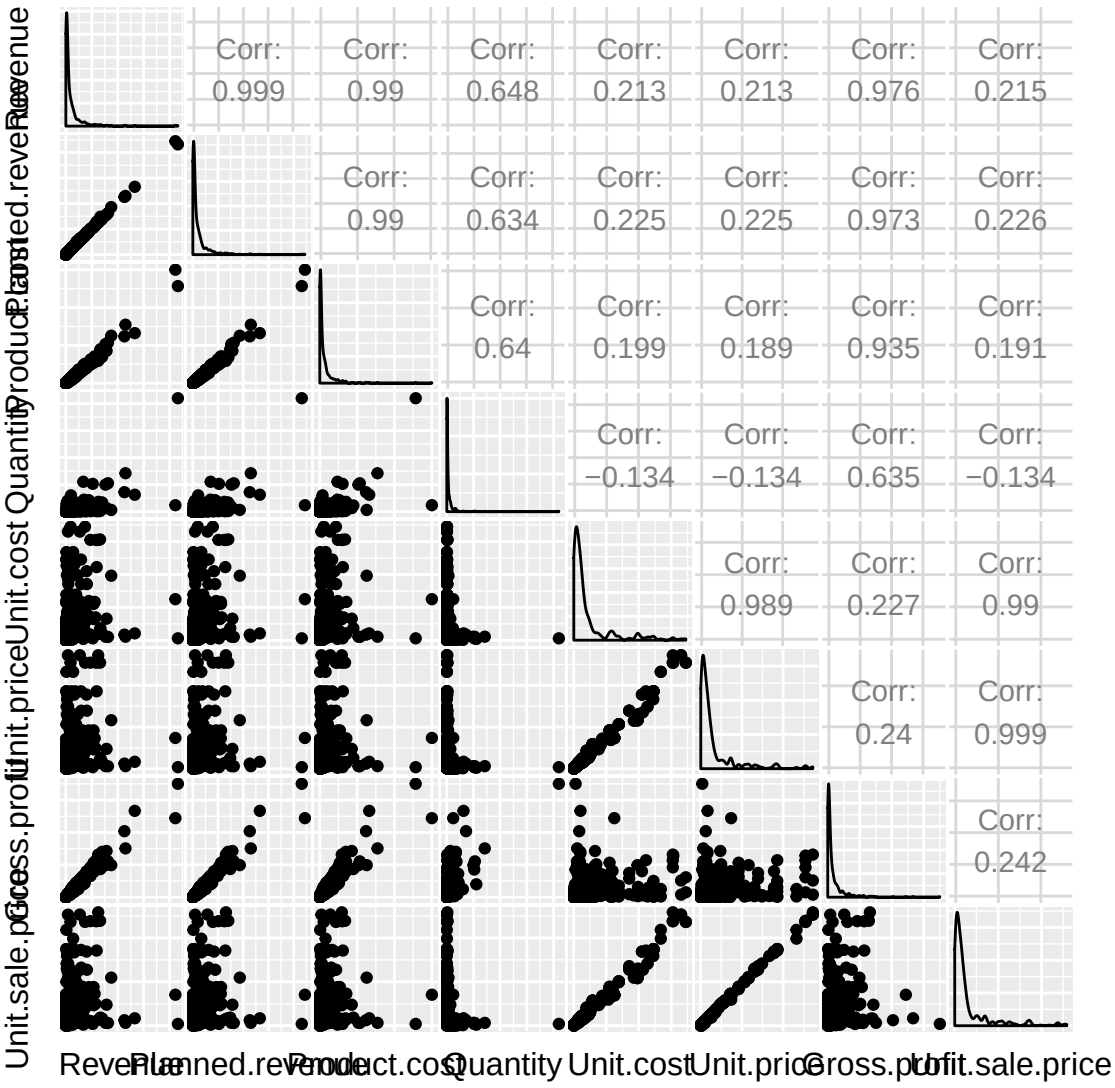


Figure 17: Scatterplot matrix of a sample of the dataset (with correlations)

After our EDA, we can divide the dataset into two separate ones, to train and evaluate the model. Now we can convert `Year` back to a numerical variable (subtracting 2004 so the baseline is 0; that will make the intercept more intuitive when including `Year` in the regression model).

```
# Year back to integer (factor only useful for vizs)
data <- data %>% mutate(Year = as.numeric(levels(Year))[Year] - 2004)
# One dataset per couple of years
# data_200405 <- data %>% filter(Year <= 2005)
# data_200607 <- data %>% filter(Year > 2005)
data_200405 <- data %>% filter(Year <= 1)
data_200607 <- data %>% filter(Year > 1)
```

Not all products appear in both periods so some re-factoring is needed:

```
# Re-factor Product (since the levels differ by period)
products_200405 <- data.frame(Product = levels(droplevels(data_200405$Product)))
products_200607 <- data.frame(Product = levels(droplevels(data_200607$Product)))
continuing_products <- intersect(products_200405, products_200607)
(new_or_discontinuted_products <- union(products_200405, products_200607) %>%
  setdiff(continuing_products))
```

```
##        Product
## 1 Trail Master
## 2   Trail Star
## 3   Auto Pilot
```

```
# Products present in one period and not the other are labelled as "Other"
data_200405 <- data_200405 %>%
  mutate(Product = ifelse(Product %in% new_or_discontinuted_products$Product,
                          "Other", as.character(Product))) %>%
  mutate(Product = factor(Product))
data_200607 <- data_200607 %>%
  mutate(Product = ifelse(Product %in% new_or_discontinuted_products$Product,
                          "Other", as.character(Product))) %>%
  mutate(Product = factor(Product))
```

There are some variables that are calculated from `Revenue` (or vice versa) so including them in any regression model would lead to a perfect fit. In particular, `Gross.profit` = `Revenue` - `Product.cost`. And `Revenue` should be equal to `Unit.sale.price` times `Quantity`, though this is not always the case, and there are differences in many cases (53.4% of the total number of observations).

```
head(data %>% select(Revenue, Product.cost, Gross.profit) %>%
       mutate(Revenue2 = Product.cost + Gross.profit))
```

```
##     Revenue Product.cost Gross.profit  Revenue2
## 1 315044.33    158371.76    156672.57 315044.33
## 2  13444.68      6298.80      7145.88  13444.68
## 3 181120.24     89413.06     91707.18 181120.24
## 4  69608.15     35326.25     34281.90  69608.15
## 5  30940.35     16370.97     14569.38  30940.35
## 6  74321.18     36531.63     37789.55  74321.18
```

```
all(round(data$Revenue - data$Product.cost, 2) == round(data$Gross.profit, 2))
```

```
## [1] TRUE
```

```
head(data %>% select(Revenue, Unit.sale.price, Quantity) %>%
       mutate(Revenue2 = Unit.sale.price * Quantity))
```

```
##     Revenue Unit.sale.price Quantity  Revenue2
## 1 315044.33        5.195714    66385 344917.49
## 2  13444.68        6.190000     2172  13444.68
## 3 181120.24        5.488000    35696 195899.65
## 4  69608.15        5.040000    15205  76633.20
## 5  30940.35        3.950000     7833  30940.35
## 6  74321.18        5.585000    14328  80021.88
```

So `Revenue` and `Product.cost` should definitely not be included in the regression model, but `Unit.sale.price` and `Quantity` might.

Let's start with the simplest model:

```
# Simplest model
params = c("Planned.revenue")
model1 <- lm(as.formula(paste("Revenue", paste(params, sep = "",
                                               collapse = " + "),
                              sep = " ~ ")), data_200405)
coeftest(model1, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                    Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     -3.2766e+03  3.5760e+02  -9.163 < 2.2e-16 ***
## Planned.revenue  9.6938e-01  2.7204e-03 356.333 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
new_data <- data.frame(data_200607[, params])
```

We'll use the RMSE to compare different models:

```
(RMSE <- sqrt(sum((model1_predictions[, 1] - data_200607$Revenue)^2) /
                dim(data_200607)[1]))
```

```
## [1] 19593.67
```

- **Is the change in the average revenue different from 95 cents when the planned revenue increases by $1?**

As shown below, the change in the average revenue is significantly different from \$0.95 when the revenue increases by \$1 (while the F statistic of the exact value, which is quite close to \$0.95, has a p value equal to 1):

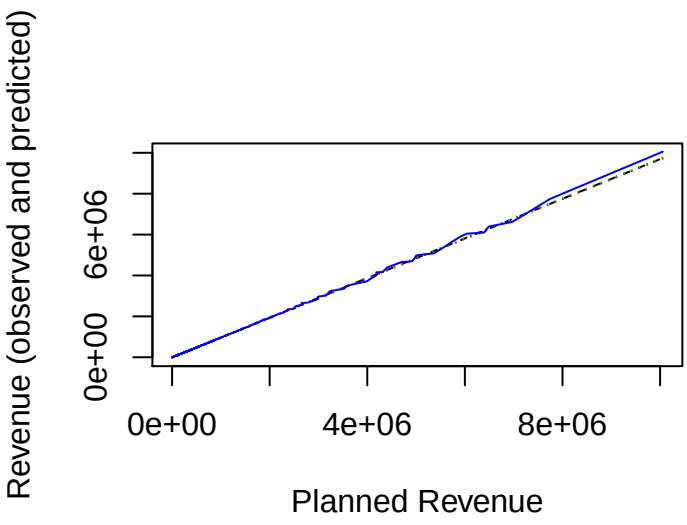


Figure 18: Planned Revenue vs. Revenue (observed and predicted) in 2006 and 2007

```
model1_full <- lm(as.formula(paste("Revenue", paste(params, sep = "",
                                                    collapse = " + "),
                                   sep = " ~ ")), data)
coeftest(model1_full, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                    Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)     -3.3970e+03  3.0377e+02 -11.183 < 2.2e-16 ***
## Planned.revenue  9.6981e-01  1.8151e-03 534.297 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
linearHypothesis(model1_full, "Planned.revenue = 0.95", vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Planned.revenue = 0.95
##
## Model 1: restricted model
## Model 2: Revenue ~ Planned.revenue
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F    Pr(>F)
## 1  24742
## 2  24741  1 119.12 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

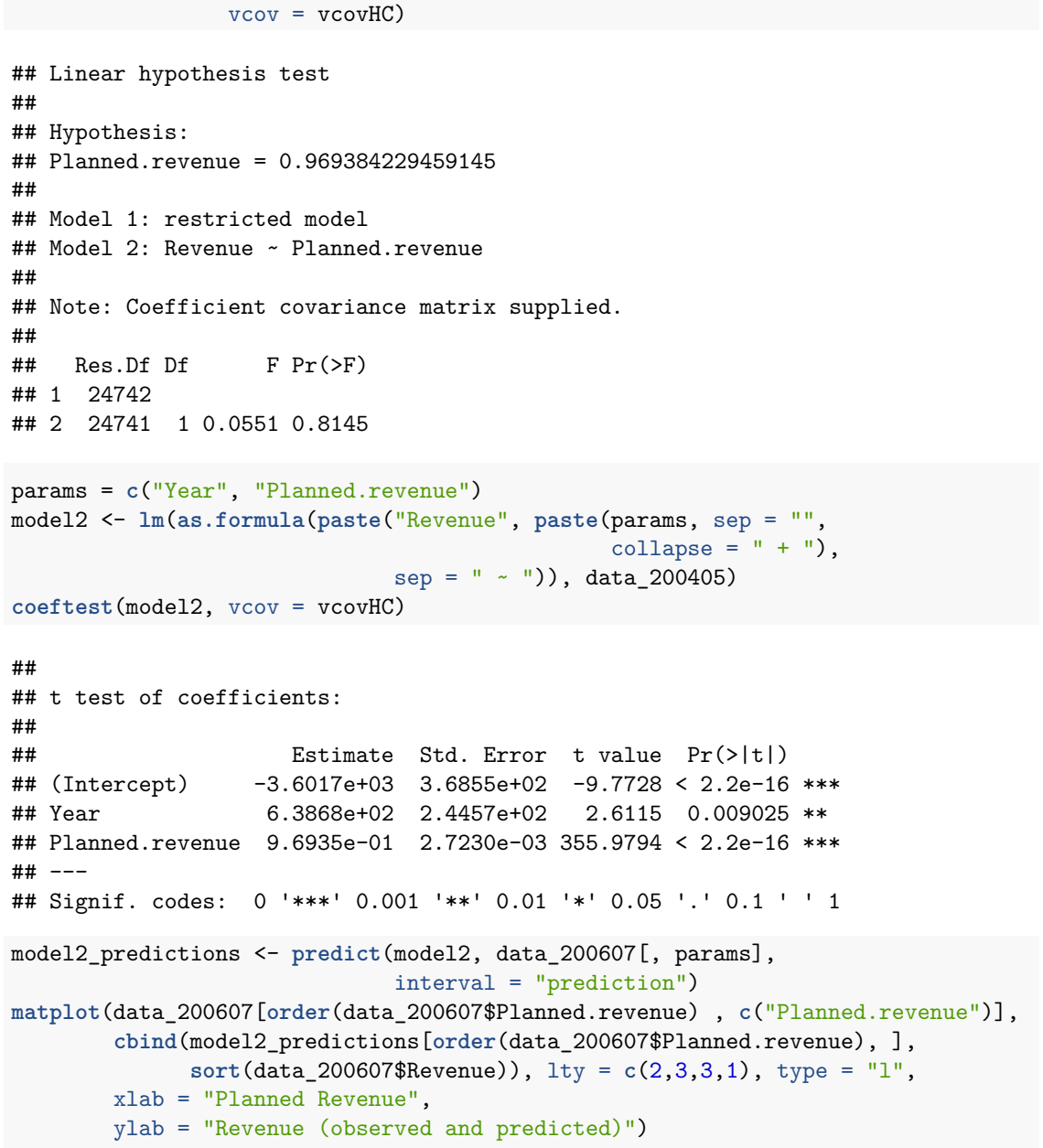
```
linearHypothesis(model1_full, paste("Planned.revenue =",
                                    coeftest(model1, vcov = vcovHC)[2, 1]),
                 vcov = vcovHC)
```

```
## Linear hypothesis test
##
## Hypothesis:
## Planned.revenue = 0.969384229459145
##
## Model 1: restricted model
## Model 2: Revenue ~ Planned.revenue
##
## Note: Coefficient covariance matrix supplied.
##
##   Res.Df Df      F Pr(>F)
## 1  24742
## 2  24741  1 0.0551 0.8145
```

```
params = c("Year", "Planned.revenue")
model2 <- lm(as.formula(paste("Revenue", paste(params, sep = "",
                                               collapse = " + "),
                              sep = " ~ ")), data_200405)
coeftest(model2, vcov = vcovHC)
```

```
##
## t test of coefficients:
##
##                     Estimate  Std. Error  t value  Pr(>|t|)
## (Intercept)     -3.6017e+03  3.6855e+02  -9.7728 < 2.2e-16 ***
## Year             6.3868e+02  2.4457e+02   2.6115  0.009025 **
## Planned.revenue  9.6935e-01  2.7230e-03 355.9794 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model2_predictions <- predict(model2, data_200607[, params],
                              interval = "prediction")
matplot(data_200607[order(data_200607$Planned.revenue) , c("Planned.revenue")],
        cbind(model2_predictions[order(data_200607$Planned.revenue), ],
              sort(data_200607$Revenue)), lty = c(2,3,3,1), type = "l",
        xlab = "Planned Revenue",
        ylab = "Revenue (observed and predicted)")
```



```
## [1] 19649.12
```

- **Explain what interaction terms in your model mean in context supported by data visualizations.**
- **Give two reasons why the OLS model coefficients may be biased and/or not consistent, be specific.**
- **Propose (but do not actually implement) a plan for an IV approach to improve your forecasting model.**

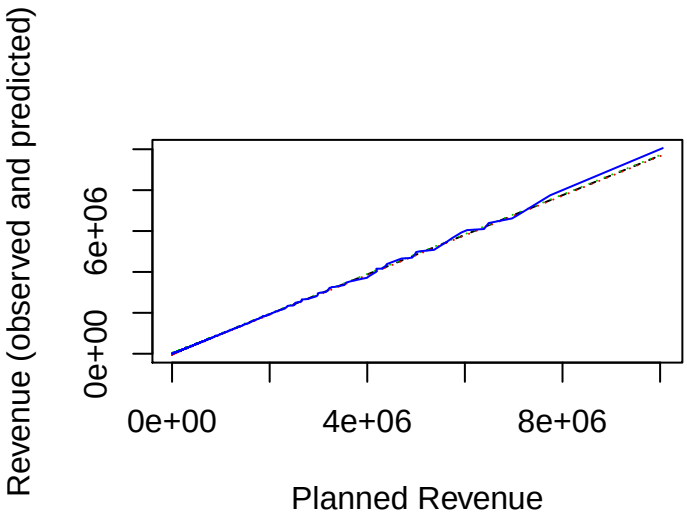


Figure 19: Planned Revenue vs. Revenue (observed and predicted) in 2006 and 2007