W271-2 – Spring 2016 – **Lab 2**

Juanjo Carin, Kevin Davis, Ashley Levato, Minghu Song

March 7, 2016

Contents

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')
```

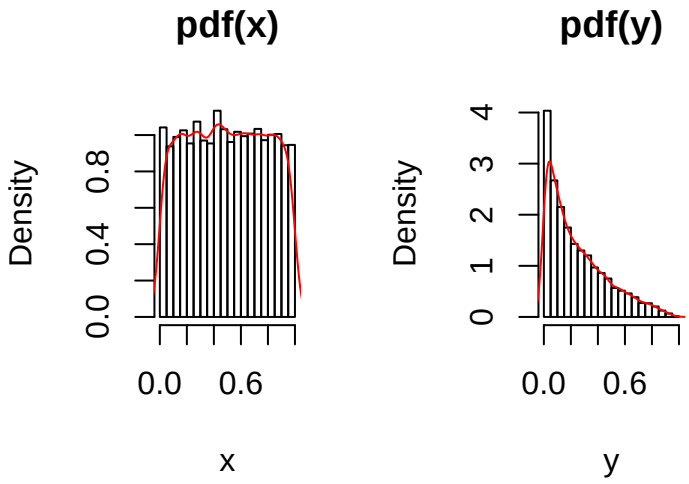

Figure 1: Histogram and approximate pdf of X and Y

```
# 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)
```

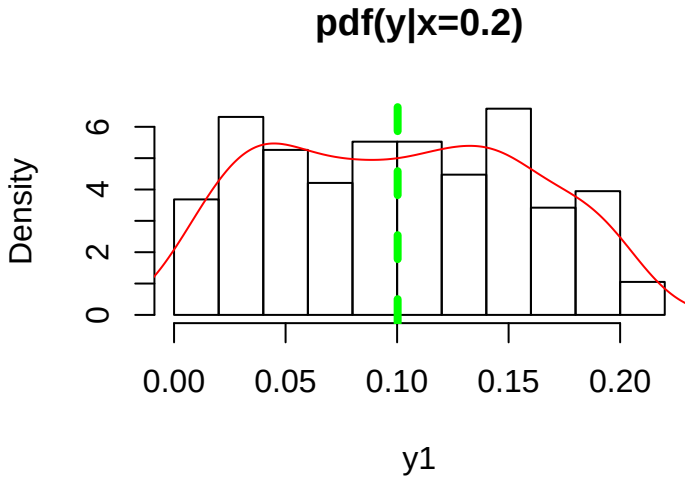

Figure 2: Histogram and approximate pdf of Y conditional on X for a given value of X (0.2)

```
# 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)
```

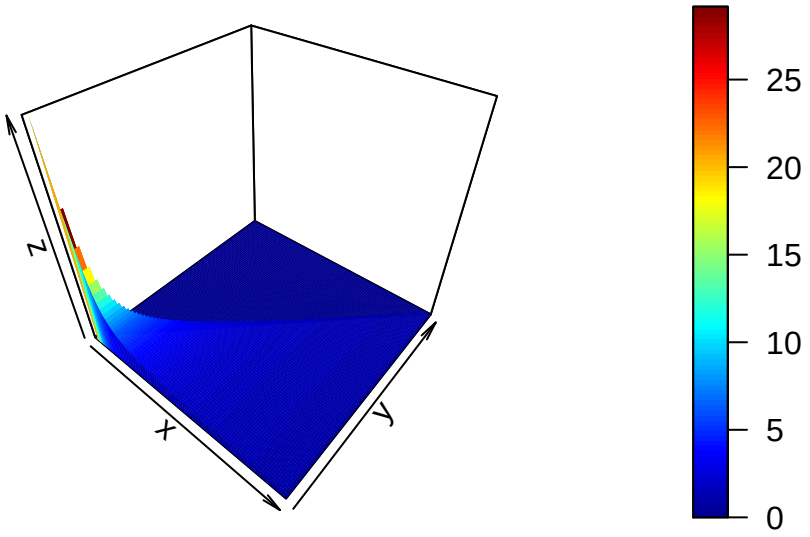

Figure 3: Approximate joint pdf of X and Y

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)')
```

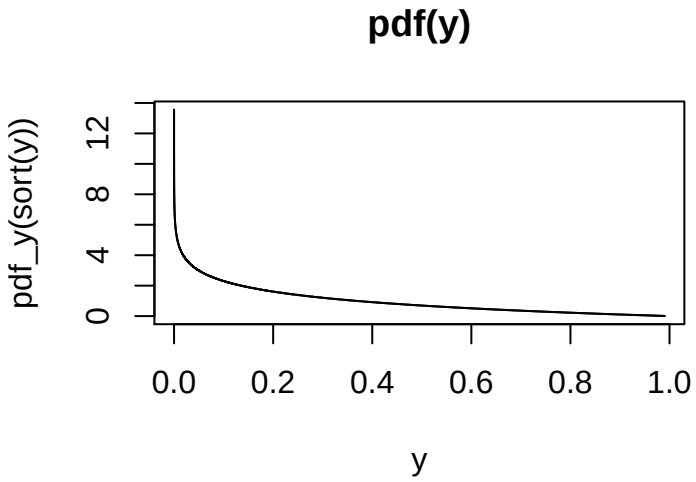

Figure 4: Approximate pdf of Y conditional on X for two values of X

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

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 $\mathbf{a, b, c \in [0, 1]}$ and $\mathbf{a + b + c = 1}$. 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:

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

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

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

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')
```

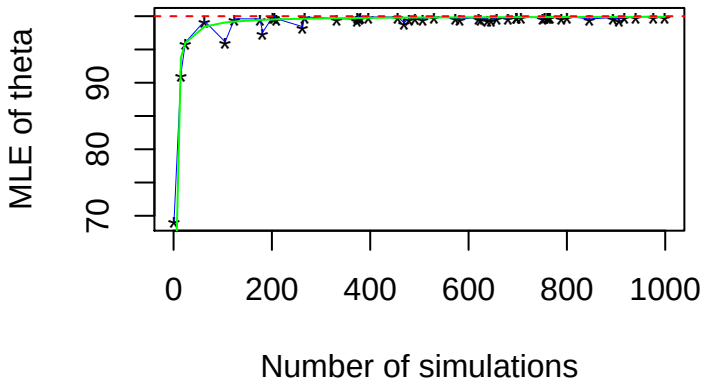

Figure 5: Trend of the MLE of θ depending on the sample size

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.

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

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