# Lab 1

Barum Park

Department of Sociology
New York University

Jan. 25, 2018

# Announcements

1. **Office Hours:**

| | | | |
|---|---|---|---|
| Christina Nelson | Mondays | 13:00-15:00 | Puck Building |
| Barum Park | Thursdays | 13:00-15:00 | Puck Building |

2. **Lost purple water bottle?**

# Before we start ...

- Any questions regarding the last class?
- Please interrupt me with questions AS OFTEN AS YOU CAN
- The lab will be hold in `STATA`

# Before we start ...

- Any questions regarding the last class?
- Please interrupt me with questions AS OFTEN AS YOU CAN
- The lab will be hold in `STATA`

# Before we start ...

- Any questions regarding the last class?
- Please interrupt me with questions AS OFTEN AS YOU CAN
- The lab will be hold in `STATA`

# Before we start ...

- Any questions regarding the last class?
- Please interrupt me with questions AS OFTEN AS YOU CAN
- The lab will be hold in `STATA`

!!! WARNING !!!

!!! PLEASE CONSULT YOUR TEXTBOOKS RATHER THAN USING THESE SLIDES TO STUDY !!!

THE TEXTBOOKS THAT YOU WERE ASSIGNED WENT THROUGH MANY REVISIONS. SO YOU CAN TRUST THEIR CONTENT

These slides, on the other hand, were created by a poor GRADUATE STUDENT from the top of his head !!

# Regression review (Single Predictor)

- Consider the (population) regression equation from last class

$$NONE = \beta_0 + \beta_1 RONE + \epsilon$$

where

$$NONE = \begin{cases} 1, & \text{no religious preference} \\ 0, & \text{otherwise} \end{cases}$$

$$RONE = \begin{cases} 1 & \text{raised with no religion} \\ 0 & \text{otherwise} \end{cases}$$

# Regression review (Single Predictor)

- Consider the (population) regression equation from last class

$$NONE = \beta_0 + \beta_1 RONE + \epsilon$$

- We assume that $E[\epsilon \mid RONE] = 0$.
- The assumption implies that

$$E[NONE|RONE] = \beta_0 + \beta_1 RONE$$

# Regression review (Single Predictor)

- Consider the (population) regression equation from last class

$$NONE = \beta_0 + \beta_1 RONE + \epsilon$$

- We assume that $E[\epsilon \mid RONE] = 0$.
  Q. What is the meaning of this assumption? What is the difference between $E[\epsilon]$ and $E[\epsilon \mid RONE]$?
- The assumption implies that

$$E[NONE|RONE] = \beta_0 + \beta_1 RONE$$

## Regression review (Single Predictor)

- Consider the (population) regression equation from last class

$$NONE = \beta_0 + \beta_1 RONE + \epsilon$$

- We assume that $E[\epsilon \mid RONE] = 0$.
- The assumption implies that

$$E[NONE|RONE] = \beta_0 + \beta_1 RONE$$

# Regression review (Single Predictor)

- Consider the (population) regression equation from last class

$$NONE = \beta_0 + \beta_1 RONE + \epsilon$$

- We assume that $E[\epsilon \mid RONE] = 0$.
- The assumption implies that

$$E[NONE|RONE] = \beta_0 + \beta_1 RONE$$

Q. As *NONE* is a dummy variable...What is $E[NONE]$?

# Regression review (Single Predictor)

- Consider the (population) regression equation from last class

$$NONE = \beta_0 + \beta_1 RONE + \epsilon$$

- We assume that $E[\epsilon \mid RONE] = 0$.
- The assumption implies that

$$E[NONE|RONE] = \beta_0 + \beta_1 RONE$$

Q. What does $\beta_0$ represent?

# Regression review (Single Predictor)

- Consider the (population) regression equation from last class

$$NONE = \beta_0 + \beta_1 RONE + \epsilon$$

- We assume that $E[\epsilon \mid RONE] = 0$.
- The assumption implies that

$$E[NONE|RONE] = \beta_0 + \beta_1 RONE$$

Q. What does $\beta_1$ represent?

# Regression review (Single Predictor)

- Consider the (population) regression equation from last class

$$NONE = \beta_0 + \beta_1 RONE + \epsilon$$

- Yes!

$$E[NONE|RONE = 0] = \beta_0 + \beta_1 \times 0 = \beta_0$$

and

$$E[NONE|RONE = 1] = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

so

$$E[NONE|RONE = 1] - E[NONE|RONE = 0] = \beta_1$$

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon | RONE, FEMALE] = 0$.

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon | RONE, FEMALE] = 0$.
Q. What does $\beta_0$ represent now?

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon | RONE, FEMALE] = 0$.
Q. According to the equation, what is the proportion of NONEs among women who were raised with no religion?

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon | RONE, FEMALE] = 0$.
Q. According to the equation, what is the proportion of NONEs among women who were raised with no religion?
Indeed,

$$E[NONE | RONE = 1, FEMALE = 1] = \beta_0 + \beta_1 + \beta_2$$

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon|RONE, FEMALE] = 0$.

- Notice that this model assumes that the difference in the proportion of NONEs between women ($FEMALE = 1$) and men ($FEMALE = 0$) does not depend on $RONE$.

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon | RONE, FEMALE] = 0$.

- Notice that this model assumes that the difference in the proportion of NONEs between women ($FEMALE = 1$) and men ($FEMALE = 0$) does not depend on $RONE$.
Q. Why?

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon | RONE, FEMALE] = 0$.

- This is because

$$\begin{aligned}
&\underbrace{E[NONE|RONE, FEMALE = 1]}_{\text{Proportion of nones among women}} - \underbrace{E[NONE|RONE, FEMALE = 0]}_{\text{Proportion of nones among men}} \\
&= (\beta_0 + \beta_1 RONE + \beta_2) - (\beta_0 + \beta_1 RONE) \\
&= (\beta_0 - \beta_0) + (\beta_1 RONE - \beta_1 RONE) + \beta_2 \\
&= \beta_2
\end{aligned}$$

regardless of whether $RONE = 1$ or $RONE = 0$.[1]

[1] The description in the underbraces is actually not correct as we need to "integrate out" *RONE* in order to obtain the proportion of nones among women, i.e., $E[NONE|FEMALE = 1]$. Yet, the conclusion that the difference between men and women does not depend on *RONE* is correct.

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon | RONE, FEMALE] = 0$.

- Is this assumption plausible?
- If not, what can we do about it?

# Regression review (Multiple Predictors)

- Next, consider the regression equation

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \epsilon$$

where we, again, assume that $E[\epsilon | RONE, FEMALE] = 0$.

- Is this assumption plausible?
- If not, what can we do about it?

# Regression Review (Interactions)

- We add an interaction term!

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \beta_{12}(RONE \times FEMALE) + \epsilon$$

# Regression Review (Interactions)

- We add an interaction term!

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \beta_{12}(RONE \times FEMALE) + \epsilon$$

- Now, we have

$$\begin{aligned} & E[NONE|RONE, FEMALE] \\ & \qquad = \beta_0 + \beta_1 RONE + \underbrace{(\beta_2 + \beta_{12} RONE)}_{\text{coefficient of FEMALE}} \times FEMALE \end{aligned}$$

so that the difference in the proportions of Nones between men and women depend on $RONE$.

# Regression Review (Interactions)

- We add an interaction term!

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \beta_{12}(RONE \times FEMALE) + \epsilon$$

Q. What does $\beta_0$ represent?

# Regression Review (Interactions)

- We add an interaction term!

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \beta_{12}(RONE \times FEMALE) + \epsilon$$

Q. What is the proportion of Nones among women raised in a religious family?

# Regression Review (Interactions)

- We add an interaction term!

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \beta_{12}(RONE \times FEMALE) + \epsilon$$

Q. What is the proportion of Nones among women raised in a religious family?

$$E[NONE|RONE = 0, FEMALE = 1] = \beta_0 + \beta_2$$

# Regression Review (Interactions)

- We add an interaction term!

$$NONE = \beta_0 + \beta_1 RONE + \beta_2 FEMALE + \beta_{12}(RONE \times FEMALE) + \epsilon$$

- In fact, with the interaction model, we can express the proportion of Nones within each cell of the following cross-table in terms of the regression coefficients:

| | RONE=0 | RONE=1 |
|---|---|---|
| FEMALE=0 | $\beta_0$ | $\beta_0 + \beta_1$ |
| FEMALE=1 | $\beta_0 + \beta_2$ | $\beta_0 + \beta_1 + \beta_2 + \beta_{12}$ |

# From Population to Samples

- Clearly, we do not observe the population but have to *estimate* the parameters from a sample
- Suppose we have a *simple random sample* of size $n$ for these variables. The data would look like this:

$$\begin{bmatrix} NONE_1 & RONE_1 & FEMALE_1 \\ NONE_2 & RONE_2 & FEMALE_2 \\ \vdots & \vdots & \vdots \\ NONE_n & RONE_n & FEMALE_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \end{bmatrix}$$

- and we would use the model

$$None_i = \hat{\beta}_0 + \hat{\beta}_1 Rone_i + \hat{\beta}_2 Female_i + \hat{\beta}_{12}(Female_i \times Rone_i) + e_i$$

to estimate $\beta_0$, $\beta_1, \beta_2$, and $\beta_{12}$.

# From Population to Samples

- Clearly, we do not observe the population but have to *estimate* the parameters from a sample
- Suppose we have a *simple random sample* of size $n$ for these variables. The data would look like this:

$$\begin{bmatrix} NONE_1 & RONE_1 & FEMALE_1 \\ NONE_2 & RONE_2 & FEMALE_2 \\ \vdots & \vdots & \vdots \\ NONE_n & RONE_n & FEMALE_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \end{bmatrix}$$

- and we would use the model

$$None_i = \hat{\beta}_0 + \hat{\beta}_1 Rone_i + \hat{\beta}_2 Female_i + \hat{\beta}_{12}(Female_i \times Rone_i) + e_i$$

to estimate $\beta_0$, $\beta_1, \beta_2$, and $\beta_{12}$.

# From Population to Samples

- Clearly, we do not observe the population but have to *estimate* the parameters from a sample
- Suppose we have a *simple random sample* of size $n$ for these variables. The data would look like this:

$$\begin{bmatrix} NONE_1 & RONE_1 & FEMALE_1 \\ NONE_2 & RONE_2 & FEMALE_2 \\ \vdots & \vdots & \vdots \\ NONE_n & RONE_n & FEMALE_n \end{bmatrix} = \begin{bmatrix} 0 & 1 & 1 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \end{bmatrix}$$

- and we would use the model

$$None_i = \hat{\beta}_0 + \hat{\beta}_1 Rone_i + \hat{\beta}_2 Female_i + \hat{\beta}_{12}(Female_i \times Rone_i) + e_i$$

to estimate $\beta_0$, $\beta_1, \beta_2$, and $\beta_{12}$.

# From Population to Samples

- Recall that in the population the following relationship holds:

| | RONE=0 | RONE=1 |
|---|---|---|
| FEMALE=0 | $\beta_0$ | $\beta_0 + \beta_1$ |
| FEMALE=1 | $\beta_0 + \beta_2$ | $\beta_0 + \beta_1 + \beta_2 + \beta_{12}$ |

  where each cell of the table is the proportion of Nones expressed in regression coefficients

- It turns out that the OLS estimator satisfies

| | RONE=0 | RONE=1 |
|---|---|---|
| FEMALE=0 | $\hat{\beta}_0$ | $\hat{\beta}_0 + \hat{\beta}_1$ |
| FEMALE=1 | $\hat{\beta}_0 + \hat{\beta}_2$ | $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_{12}$ |

  where, now, the cells are the sample proportions of Nones within each category (we will discuss this further in the `STATA` session).

# From Population to Samples

- Recall that in the population the following relationship holds:

| | RONE=0 | RONE=1 |
|---|---|---|
| FEMALE=0 | $\beta_0$ | $\beta_0 + \beta_1$ |
| FEMALE=1 | $\beta_0 + \beta_2$ | $\beta_0 + \beta_1 + \beta_2 + \beta_{12}$ |

  where each cell of the table is the proportion of Nones expressed in regression coefficients

- It turns out that the OLS estimator satisfies

| | RONE=0 | RONE=1 |
|---|---|---|
| FEMALE=0 | $\hat{\beta}_0$ | $\hat{\beta}_0 + \hat{\beta}_1$ |
| FEMALE=1 | $\hat{\beta}_0 + \hat{\beta}_2$ | $\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_{12}$ |

  where, now, the cells are the sample proportions of Nones within each category (we will discuss this further in the `STATA` session).

# Logit

- What is a "logit"?
- What is an "odds-ratio"?
- The connection between them is

$$\text{logit}(p_1) - \text{logit}(p_2) = \ln\left[OR(p_1, p_2)\right].$$

# Logit

- What is a "logit"?

$$\text{logit}(p) = \text{logged-odds}(p) = \ln\left(\frac{p}{1-p}\right)$$

- What is an "odds-ratio"?

- The connection between them is

$$\text{logit}(p_1) - \text{logit}(p_2) = \ln\left[OR(p_1, p_2)\right].$$

# Logit

- What is a "logit"?

$$\text{logit}(p) = \text{logged-odds}(p) = \ln\left(\frac{p}{1-p}\right)$$

- What is an "odds-ratio"?

- The connection between them is

$$\text{logit}(p_1) - \text{logit}(p_2) = \ln\left[OR(p_1, p_2)\right].$$

# Logit

- What is a "logit"?

$$\mathrm{logit}(p) = \text{logged-odds}(p) = \ln\left(\frac{p}{1-p}\right)$$

- What is an "odds-ratio"? Suppose you have two probabilities $p_1$ and $p_2$, then their odds-ratio is

$$OR(p_1, p_2) = \left(\frac{p_1}{1-p_1}\right) \Big/ \left(\frac{p_2}{1-p_2}\right)$$

- The connection between them is

$$\mathrm{logit}(p_1) - \mathrm{logit}(p_2) = \ln\left[OR(p_1, p_2)\right].$$

# Logit

- What is a "logit"?

$$\text{logit}(p) = \text{logged-odds}(p) = \ln\left(\frac{p}{1-p}\right)$$

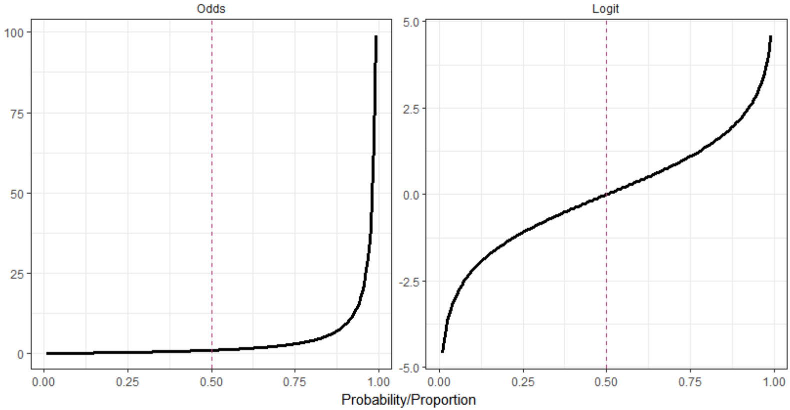
- What is an "odds-ratio"? Suppose you have two probabilities $p_1$ and $p_2$, then their odds-ratio is

$$OR(p_1, p_2) = \left(\frac{p_1}{1-p_1}\right) \Big/ \left(\frac{p_2}{1-p_2}\right)$$

- The connection between them is

$$\text{logit}(p_1) - \text{logit}(p_2) = \ln\left[OR(p_1, p_2)\right].$$

# Modeling Odds?



# Logistic Regression

- Logistic regression (next week) models conditional probabilities/proportions as

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

  where $p = E[y|x_1, x_2, ..., x_k]$ and $y$ is a dummy variable.
- If there is only one predictor, $x_1$, which is dummy-coded, then

$$x_1 = 0 \implies \text{logit}(p_0) = \beta_0$$
$$x_1 = 1 \implies \text{logit}(p_1) = \beta_0 + \beta_1$$

  and

$$\beta_1 = \text{logit}(p_1) - \text{logit}(p_0) = \ln[OR(p_1, p_0)]$$

# Logistic Regression

- Logistic regression (next week) models conditional probabilities/proportions as

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

where $p = E[y|x_1, x_2, ..., x_k]$ and $y$ is a dummy variable.

- If there is only one predictor, $x_1$, which is dummy-coded, then

$$\begin{aligned} x_1 = 0 &\implies \text{logit}(p_0) = \beta_0 \\ x_1 = 1 &\implies \text{logit}(p_1) = \beta_0 + \beta_1 \end{aligned}$$

and

$$\beta_1 = \text{logit}(p_1) - \text{logit}(p_0) = \ln[OR(p_1, p_0)]$$

# Logistic Regression

- Logistic regression (next week) models conditional probabilities/proportions as

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k,$$

where $p = E[y|x_1, x_2, ..., x_k]$ and $y$ is a dummy variable.

- If there is only one predictor, $x_1$, which is dummy-coded, then

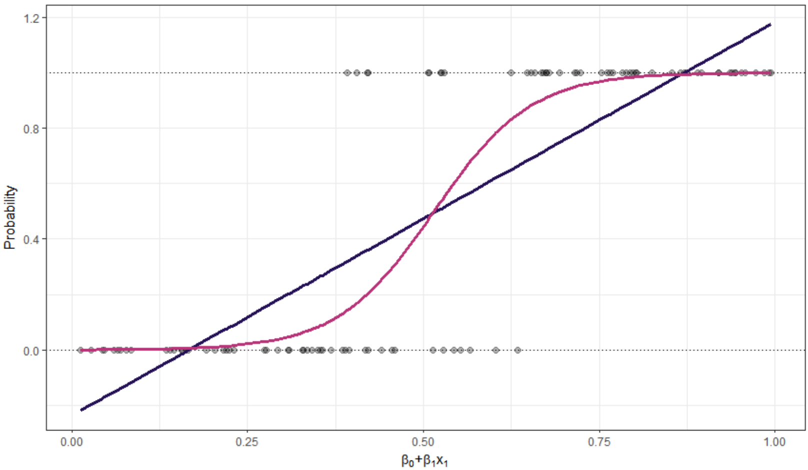
$$x_1 = 0 \implies \text{logit}(p_0) = \beta_0$$
$$x_1 = 1 \implies \text{logit}(p_1) = \beta_0 + \beta_1$$

and

$$\beta_1 = \text{logit}(p_1) - \text{logit}(p_0) = \ln[OR(p_1, p_0)]$$

Q.We saw that we can model proportions/probabilities with linear regression, why using "logits"?

# Linear Regression?



Let's turn to `STATA` ...