Lab 2

Barum Park

Department of Sociology
New York University

Feb. 1, 2018

Before We Start ...

Any questions regarding last class?

!!! 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 !!

Centering in Regressions with Interactions

Centering in Regressions with Interactions

- Consider the regression from last class

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2 + \epsilon$$

where

$$\begin{aligned} y &: \text{attitude toward abortion} \\ x_1 &: \text{female=1, male=0} \\ x_2 &: \text{political views} \in \{1, 2, \ldots, 7\} \end{aligned}$$

- What is β_1 representing?
- What is β_2 representing?

Centering in Regressions with Interactions

- Suppose that the coefficients of the model are as follows:

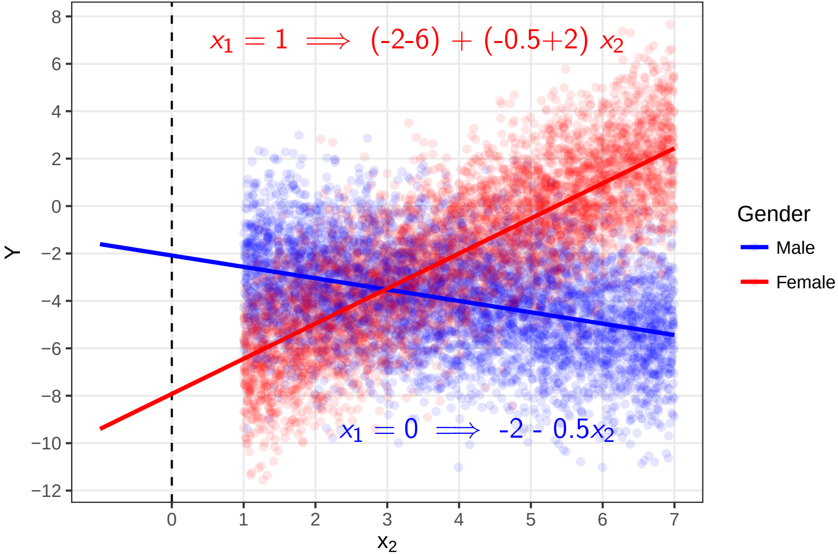
$$y = -2 - 6x_1 - 0.5x_2 + 2(x_1 x_2) + \epsilon$$

- Note that $\beta_0 = -2$ represents the level of support for abortion when $x_2 = 0$ and $x_1 = 0$.
- Similarly, $\beta_1 = -6$ represents the gender gap when $x_2 = 0$.

- The problem is that there is no respondent in our sample for which $x_2 = 0$!
- Let's look at a simulated dataset that has these patterns ...

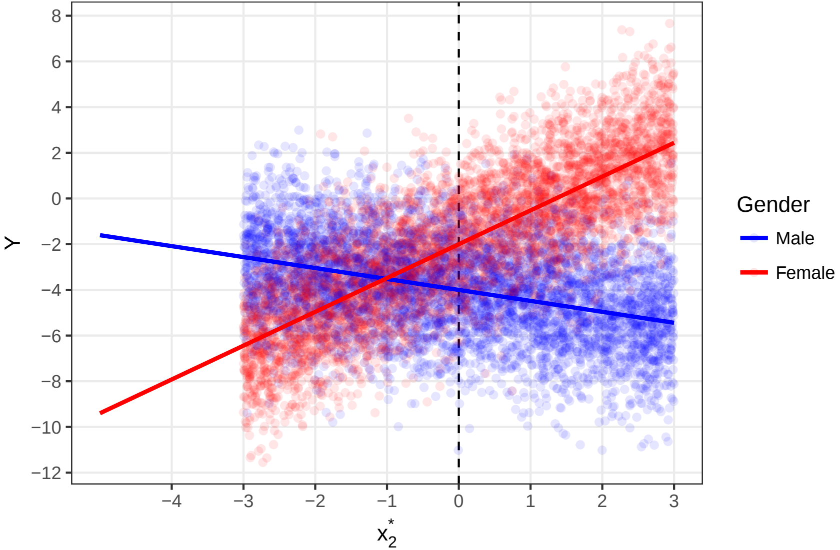
Centering in Regressions with Interactions

$$y = -2 - 6x_1 - 0.5x_2 + 2(x_1 x_2) + \epsilon$$



Centering in Regressions with Interactions

$$y = -4 - 2x_1 - 0.5x_2^* + 2(x_1 x_2^*) + \epsilon$$



Logistic Regression

General Structure of the Logistic Regression Model

- The logistic regression model has the form

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

where $p = p(\mathbf{x}) = E[y \mid x_1, x_2, ..., x_k]$ is the probability that $y = 1$ *given the values of the predictors*.

- Note that

$$\ln(x) = y \iff x = e^y$$

We also write e^y as $\exp(y)$..

- Thus,

$$\frac{p(\mathbf{x})}{1-p(\mathbf{x})} = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k).$$

Logistic Regression and Odds

In what follows, I will show how exponentiated logistic regression coefficients translate into odds-ratios, as the question came up in class.

However ...

I HIGHLY RECOMMEND THAT YOU CONVERT ALL RESULTS FROM YOUR LOGISTIC REGRESSIONS INTO "PROBABILITIES" NOT "ODDS-RATIO"S !!

NOT MANY PEOPLE UNDERSTAND WHAT ODDS-RATIOS ARE!!

(EVEN I DON'T UNDERSTAND THEM !!! MIKE PROBABLY DOES ...)

AND EVEN HE USES PLOTS !!!

Logistic Regression and Odds

- Let us concentrate on a simple model:

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1.$$

- By exponentiating both sides, we have

$$Odds(x_1) = \frac{p(x_1)}{1-p(x_1)} = e^{\beta_0+\beta_1 x_1} = e^{\beta_0} e^{\beta_1 x_1}$$

- Thus,

$$Odds(x_1 = 0) = e^{\beta_0} \text{ and } Odds(x_1 = 1) = e^{\beta_0} e^{\beta_1}$$

- It follows that

$$OR(x_1) = \frac{Odds(x_1 = 1)}{Odds(x_1 = 0)} = \frac{e^{\beta_0} e^{\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

Logistic Regression and Odds

- Thus, when we exponentiate the coefficient of a dummy variable, we get the ratio of the odds for the event that $y = 1$.
- What do we get when we exponentiate the coefficient of a continuous variable?
- Let the variable x_1 from the above example be continuous, then

$$OR(x_1) = \frac{e^{\beta_0} e^{\beta_1 x_1}}{e^{\beta_0}} = e^{\beta_1 x_1} = \left(e^{\beta_1}\right)^{x_1} = \gamma_1^{x_1}.$$

so for $x_1 = 1$ we get γ_1, for $x_1 = 2$ we get γ_1^2, and so on ...

Logistic Regression and Odds

- Note that γ_1 gets multiplied by γ_1 every time x_1 increases by one unit. Thus, we can interpret the coefficient $\gamma_1 = e^{\beta_1}$ as follows:

  `the model predicts that every unit increase in` x_1 `is associated with an increase/decrease in the odds that` $y = 1$ `by a` **factor** `of` γ_1.

- If $\gamma_1 > 1$, this means that the odds are increasing and if $\gamma_1 < 1$ the odds are decreasing.
- For example, if $\gamma_1 = .33 \approx 1/3$, the odds are decreasing by a factor of 3 for every unit increase in x_1 (this means that the odds are cut into one third); if $\gamma_1 = 2$, the odds are increasing by a factor of 2 (this means that the odds are doubled).

Bonus: Interaction Term? (Do not take this seriously...)

- In a model with an interaction term (both variables are binary)

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2,$$

what is the interpretation of β_{12} in terms of odds-ratios?

-

$$Odds(x_1 = 0, x_2 = 0) = e^{\beta_0} = \gamma_0$$
$$Odds(x_1 = 1, x_2 = 0) = e^{\beta_0+\beta_1} = \gamma_0\gamma_1$$
$$Odds(x_1 = 0, x_2 = 1) = e^{\beta_0+\beta_2} = \gamma_0\gamma_2$$
$$Odds(x_1 = 1, x_2 = 1) = e^{\beta_0+\beta_1+\beta_2+\beta_{12}} = \gamma_0\gamma_1\gamma_2\gamma_{12}$$

so

$$\gamma_{12} = \frac{\gamma_0\gamma_1\gamma_2\gamma_{12}\gamma_0}{\gamma_0\gamma_1\gamma_0\gamma_2} = \frac{Odds(x_1 = 1, x_2 = 1)Odds(x_1 = 0, x_2 = 0)}{Odds(x_1 = 0, x_2 = 1)Odds(x_1 = 1, x_2 = 0)}$$

- A RATIO OF ODDS RATIOS!!!

THIS IS WHY YOU SHOULD TRY TO CONVERT LOGISTIC REGRESSION RESULTS INTO PROBABILITIES !!!

Predicted Probabilities

Logistic Regression and Probabilities

- But if not using odds-ratios, what to do?
- We can go one step further and transform predicted logits into predicted probabilities!
- How?

- Consider again the simple logistic regression

$$\ln\left(\frac{p(x_1)}{1-p(x_1)}\right) = \beta_0 + \beta_1 x_1$$

- By exponentiating both sides, we obtain the odds

$$\frac{p(x_1)}{1-p(x_1)} = e^{\beta_0+\beta_1 x_1}$$

- Next, just to make the equations look less complicated, let us define $\mathbf{xb} = \beta_0 + \beta_1 x_1$ (this is simply a number!)

- So far we have that the odds are

$$Odds(x_1) = \frac{p(x_1)}{1 - p(x_1)} = e^{\beta_0 + \beta_1 x_1} = e^{\mathbf{xb}}$$

- Next, let us do some arithmetics

$$\begin{aligned}
\frac{p(x_1)}{1 - p(x_1)} &= e^{\mathbf{xb}} \\
p(x_1) &= e^{\mathbf{xb}}[1 - p(x_1)] \\
p(x_1) &= e^{\mathbf{xb}} - e^{\mathbf{xb}} p(x_1) \\
p(x_1) + e^{\mathbf{xb}} p(x_1) &= e^{\mathbf{xb}} \\
p(x_1)[1 + e^{\mathbf{xb}}] &= e^{\mathbf{xb}} \\
p(x_1) &= \frac{e^{\mathbf{xb}}}{1 + e^{\mathbf{xb}}}
\end{aligned}$$

Interpretation

- Actually, we can go one step further!

$$p(x_1) = \frac{e^{\mathbf{xb}}}{1 + e^{\mathbf{xb}}} = \frac{1}{\left(\frac{1}{e^{\mathbf{xb}}}\right) + 1} = \frac{1}{1 + e^{-\mathbf{xb}}}$$
$$= \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1)}}$$

- Note that this is a complicated non-linear function in x_1. This means that interpretations such as "the model predicts that an unit increase in x_1 is associated with a such and such increase/decrease in p" does not hold anymore!
- These interpretations are only valid on the logit-scale (note that the equation is linear in its coefficients!)

$$logit(p) = \beta_0 + \beta_1 x_1$$

  Here you can say that "a unit increase in x_1 is associated with a β_1-unit increase in the the logit."

Interpretation

- What should we do ??
- Note that, we can always **PLOT!** the predicted probabilities of the model
  1. If $x_1 = 0$ the probability that $y = 1$ is
  $$p(x_1 = 0) = \frac{1}{1 + e^{-\beta_0}}$$
  2. if $x_1 = 1$ the corresponding probability is:
  $$p(x_1 = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1)}}$$
- We can thereafter give the reader a visual representation of the predictions of the model

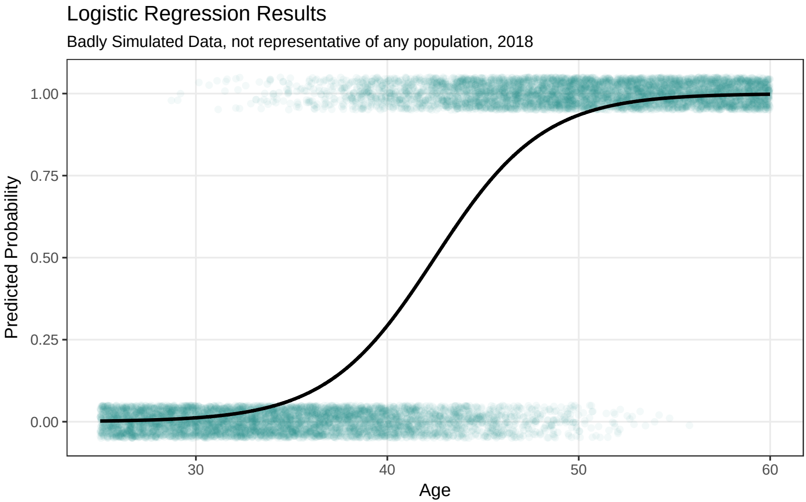
Interpretation

- Consider, for example, the following description :

  `Age was a significant predictor of whether respondents turn out to vote. The model predicts that a unit increase in age corresponds to a 40% increase in the odds of voting.`

- How strong is this association? What is the likelihood of a person of age 45 to vote?

Interpretation



1) Line shows shows predicted probabilities from the logistic regression model

2) jittered points at the top and bottom show the observed data points

3) Fake data, can you see why?

Interactions

- Next, consider the model from last class

$$logit(p) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$$

  where x_1 is gender (dummy, 1=female) and x_2 is political views (continuous).

- We know, by now, that the predicted probabilities of the model are

$$p(x_1, x_2) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2)}}$$

- By plugging in different values of x_1 and x_2, we can therefore plot the predicted probabilities.

Interpretation

- Again, we start with interpretations in terms of odds :

  ```
  All predictors in the model, including the
  interaction term, were statistically significant.
  The model predicts that a unit increase on the
  ideological self-placement scale is associated
  with an increase in the odds of turning out to
  vote by a factor of approximately 6 for women,
  while the same increase in ideology corresponds
  to a decrease in the odds to vote by
  approximately 55 percent for males.
  ```

- How likely are women to vote? How likely are men, who identify as extremely liberal,to vote?

Interpretation

1) Line shows shows predicted probabilities from the logistic regression model

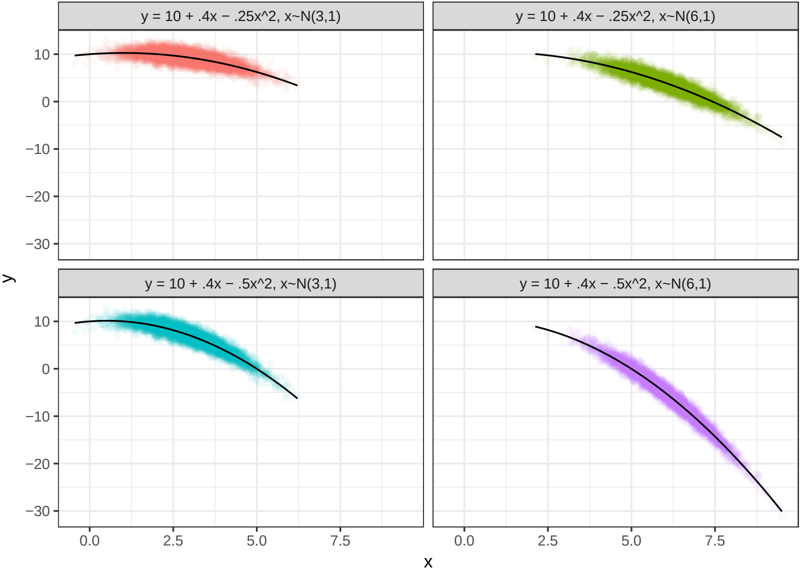
2) jittered points at the top and bottom show the observed data points

Interactions and Polynomials

- Lastly, let us look at polynomial regressions. Here we focus on the linear model, but everything will carry over to logistic regression (as the regression of the logit on the predictors is a linear model)
- Consider the polynomial regression

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \epsilon$$

- Say that $\beta_1 > 0$ and $\beta_2 < 0$. What does this imply?
- When does the regression line hit its highest prediction?
- Both depend on the relative size of the coefficients *and* the distribution of x_1! Just plot it!


y = 10 + .4x − .25x^2, x~N(3,1)
y = 10 + .4x − .25x^2, x~N(6,1)
y = 10 + .4x − .5x^2, x~N(3,1)
y = 10 + .4x − .5x^2, x~N(6,1)
y
x
10
0
−10
−20
−30
0.0
2.5
5.0
7.5

- What if we have a polynomial and an interaction term? For example, consider

$$E[y|x,z] = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + \beta_4 xz + \beta_5 x^2 z$$

where y is income (continuous) x is age (also continuous) and z is a gender (female=1, dummy). How would you interpret this equation?

- First way: gather terms

$$y = \beta_0 + (\beta_1 + \beta_4 z)x + (\beta_2 + \beta_5 z)x^2 + \epsilon$$

Now,

$$E[y|x,z=0] = \beta_0 + \beta_1 x + \beta_2 x^2$$
$$E[y|x,z=1] = (\beta_0 + \beta_3) + (\beta_1 + \beta_4)x + (\beta_2 + \beta_5)x^2$$

- Thus, the regression curve for both males and females follow a quadratic trend, but the lines might differ to the extend that β_4 and β_5 deviate from zero
- But, again, the equation per se gives us not a good sense of how this curve looks like, so we have to PLOT THEM