Week 12: Logistic regression

Marcelo Coca Perraillon

University of Colorado
Anschutz Medical Campus

Health Services Research Methods I
HSMP 7607
2017
Outline

- Logistic regression
- Parameter interpretation
- Log odds, odds ratios, probability scale
- Goodness of fit
- Marginal effects (average predicted probabilities)
Big picture

- Last class we saw that there are many ways to derive a logistic model
- Perhaps the most straightforward is to assume a probability density function for the outcome (Bernoulli or Binomial), write, the likelihood function, and find the MLE solution
- Today, we will focus on interpreting the logistic coefficients
- We will use a dataset on married women’s labor force participation from Wooldridge
Example

- Women’s labor force participation (inlf); main predictor is "extra" money in family

\[ \text{inlf} = 1 \text{ if in labor force, 1975} \]
\[ \text{nwifeinc} = \frac{\text{faminc} - \text{wage*hours}}{1000} \]
\[ \text{educ} \quad \text{years of schooling} \]
\[ \text{exper} \quad \text{actual labor mkt exper} \]
\[ \text{age} \quad \text{woman’s age in yrs} \]
\[ \text{kidslt6} \quad \# \text{kids < 6 years} \]
\[ \text{kidsge6} \quad \# \text{kids 6-18} \]

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>inlf</td>
<td>753</td>
<td>0.568393</td>
<td>0.4956295</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>nwifeinc</td>
<td>753</td>
<td>20.12896</td>
<td>11.6348</td>
<td>-0.0290575</td>
<td>96</td>
</tr>
<tr>
<td>educ</td>
<td>753</td>
<td>12.28685</td>
<td>2.280246</td>
<td>5</td>
<td>17</td>
</tr>
<tr>
<td>exper</td>
<td>753</td>
<td>10.63081</td>
<td>8.06913</td>
<td>0</td>
<td>45</td>
</tr>
<tr>
<td>age</td>
<td>753</td>
<td>42.53785</td>
<td>8.072574</td>
<td>30</td>
<td>60</td>
</tr>
<tr>
<td>kidslt6</td>
<td>753</td>
<td>0.2377158</td>
<td>0.523959</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>kidsge6</td>
<td>753</td>
<td>1.353254</td>
<td>1.319874</td>
<td>0</td>
<td>8</td>
</tr>
</tbody>
</table>

©2017 PERRAILLON ARR
Labor force participation

- The probability of working decreasing as a function of "extra" income

```stata
lowess inlf nwifeinc, gen(lflow) nograph
scatter inlf nwifeinc, jitter(5) msize(small) || line lflow nwifeinc, sort ///
    legend(off) saving(lblow.gph, replace)
graph export lblow.png, replace
```
Writing down the model

- We want to estimate the following model:
  \[ P(\text{inlf}_i = 1|\text{nwifeinc}_i) = \Lambda(\beta_0 + \beta_1 \text{nwifeinc}_i) \]

- **By convention**, when we write capital lambda, \( \Lambda() \), we imply a logistic model (\( \Lambda \) is not a non-linear function). When we write phi, \( \phi() \), we imply a probit model.

- The other common way of writing the logistic model is:
  \[ \log\left(\frac{\text{inlf}_i}{1-\text{inlf}_i}\right) = \beta_0 + \beta_1 \text{nwifeinc}_i \]

- Or
  \[ \logit(\text{inlf}_i) = \beta_0 + \beta_1 \text{nwifeinc}_i \]

- **Note**, no additive error anywhere.

- My preferred way is to write \( \log\left(\frac{\text{inlf}_i}{1-\text{inlf}_i}\right) = \beta_0 + \beta_1 \text{nwifeinc}_i \) because this will match Stata’s (or any other statistical package) output. Remember, we are not directly estimating \( P(\text{inlf}_i = 1|\text{nwifeinc}_i) \)
Estimating the model

- So, we will estimate \( \log\left(\frac{\text{inlf}_i}{1-\text{inlf}_i}\right) = \beta_0 + \beta_1 \text{nwifeinc}_i \)

\[
\text{logit inlf nwifeinc, nolog}
\]

Logistic regression

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of obs</td>
<td>753</td>
</tr>
<tr>
<td>LR chi2(1)</td>
<td>10.44</td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.0012</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-509.65435</td>
</tr>
<tr>
<td>Pseudo R2</td>
<td>0.0101</td>
</tr>
</tbody>
</table>

|          | Coef.   | Std. Err. | z    | P>|z|   | [95% Conf. Interval] |
|----------|---------|-----------|------|-------|---------------------|
| inlf     |         |           |      |       |                     |
| nwifeinc | -.0207569 | .0065907 | -3.15 | 0.002 | -.0336744 ,-.0078394 |
| _cons    | .6946059  | .1521569  | 4.57  | 0.000 | .396384 , .9928279 |

- A one thousand increase in “extra” income decreases the log-odds of participating in the labor force by 0.021. And it’s statistically significant (p-value = 0.002). Same Wald test as before: 
  \(-.0207569/.0065907 = -3.1494227\). The difference is that it’s not t-student distributed but normally distributed
Overall significance

- The $\chi^2$ (chi-square) test of the overall significance should look familiar. It compares the current model to the null model (without covariates); the null hypothesis is that all the coefficients in current model are zero.

- It’s the **likelihood ratio test** that we have seen before; the equivalent of ANOVA:

```plaintext
* LRT
qui logit inlf nwifeinc, nolog
est sto full

qui logit inlf, nolog
est sto redu

lrtest full redu

Likelihood-ratio test
(Assumption: redu nested in full)
LR chi2(1) = 10.44
Prob > chi2 = 0.0012
```
What about that Pseudo $R^2$?

- We can’t partition variance into explained and unexplained as before so we don’t have a nice $R^2$.
- But one way to come up with a measure of fit is to use the likelihood function to compare the current model to the model without any explanatory variable (the null model).
- The formula is: $1 - \frac{ll_{cm}}{ll_{nul}}$, where $ll_{cm}$ is the log-likelihood of the current model and $ll_{nul}$ is the log-likelihood of the null model.
- If the current model is as good as the null model, then $\frac{ll_{cm}}{ll_{nul}}$ is going to close to 1 and the pseudo $- R^2$ is going to be close to zero.
- If not, it will be greater than 0. Recall that log-likelihood is usually negative so the ratio is positive.
Pseudo-$R^2$

- Replicate Pseudo $R^2$

    qui logit inlf nwifeinc, nolog
    scalar ll_cm = e(ll)

    qui logit inlf, nolog
    scalar ll_n = e(ll)

    di 1 - (ll_cm/ll_n)

    .0101362

- Psuedo $R^2$ is not a measure of how good the model is at prediction; just how better it fits compared to null model. I don’t think that calling it pseudo $R^2$ is a good idea

- Big picture: comparing the log-likelihood of models is a way of comparing goodness of fit. If nested, we have the a test (LRT); if not nested, we have BIC or AIC
Let’s try a different predictor

- We will estimate \( \log \left( \frac{\text{inlf}_i}{1 - \text{inlf}_i} \right) = \beta_0 + \beta_1 \text{hsp}_i \), where \( \text{hsp} \) if education > 12

```stata
gen hsp = 0
replace hsp = 1 if educ > 12 & educ ~= .
logit inlf hsp, nolog
```

Logistic regression

<table>
<thead>
<tr>
<th>Number of obs</th>
<th>753</th>
</tr>
</thead>
<tbody>
<tr>
<td>LR chi2(1)</td>
<td>15.08</td>
</tr>
<tr>
<td>Prob &gt; chi2</td>
<td>0.0001</td>
</tr>
<tr>
<td>Log likelihood</td>
<td>-507.33524</td>
</tr>
<tr>
<td>Pseudo R2</td>
<td>0.0146</td>
</tr>
</tbody>
</table>

| inlf | Coef.  | Std. Err. | z    | P>|z| | [95% Conf. Interval] |
|------|--------|-----------|------|------|----------------------|
| hsp  | .6504074 | .1704773 | 3.82 | 0.000 | .3162781 - .9845368 |
| _cons | .0998982 | .086094 | 1.16 | 0.246 | -.068843 - .2686393 |

- The log-odds of entering the labor force is 0.65 higher for those with more than high school education compared to those with high-school completed or less than high-school
Odds ratios

- Let’s do our usual math to make sense of coefficients. We just estimated the model \( \log\left(\frac{\text{inlf}_i}{1-\text{inlf}_i}\right) = \beta_0 + \beta_1 hsp_i \)
- For those with \( hsp = 1 \), the model is \( \log\left(\frac{\text{inlf}_{hsp}}{1-\text{inlf}_{hsp}}\right) = \beta_0 + \beta_1 \)
- For those with \( hsp = 0 \), the model is \( \log\left(\frac{\text{inlf}_{nohsp}}{1-\text{inlf}_{nohsp}}\right) = \beta_0 \)
- The difference of the two is \( \log\left(\frac{\text{inlf}_{hsp}}{1-\text{inlf}_{hsp}}\right) - \log\left(\frac{\text{inlf}_{nohsp}}{1-\text{inlf}_{nohsp}}\right) = \beta_1 \)
- Applying the rules of logs: \( \log\left(\frac{\text{inlf}_{hsp}}{1-\text{inlf}_{hsp}}\right) = \beta_1 \)
- Taking \( e() \): \( \frac{\text{inlf}_{hsp}}{1-\text{inlf}_{hsp}} \cdot \frac{\text{inlf}_{nohsp}}{\text{inlf}_{nohsp}} = e^{\beta_1} \)
Odds ratios

\[ \frac{\text{inlf}_{hsp}}{1-\text{inlf}_{hsp}} \cdot \frac{1-\text{inlf}_{nohsp}}{\text{inlf}_{nohsp}} = e^{\beta_1} \]

- And that’s the *(in)famous odds-ratio*
- In our example, \( e^{.6504074} = 1.92 \). So the odds of entering the labor force is almost twice as high for those with more than high school education compare to those without

- That’s the way most reporters would report this finding. **And it’s correct.** The problem is that we would then interpret this as saying that the **probability** of entering the labor force is twice as high for those with more than high school

- **That interpretation is wrong.** A ratio of odds is more often than not far away from the ratio of probabilities
Odds ratios are NOT relative risks or relative probabilities

- One quick way to see this is by doing some algebra
- Changing the notation to make it easier:
  \[
  \frac{P_A}{1-P_A} = \frac{P_B}{1-P_B} = e^{\beta_1}
  \]
- After some simple algebra:
  \[
  \frac{P_A}{P_B} = \frac{1-P_A}{1-P_B} e^{\beta_1}
  \]
- Only when rare events (both \(P_A\) and \(P_B\) are small) are odds ratios close to relative probabilities (\(\frac{1-P_A}{1-P_B}\) will be close to 1)
- For a more epi explanation, see http://www.mdedge.com/jfponline/article/65515/relative-risks-and-odds-ratios-whats-difference
Relative probabilities

- With only a dummy variable as predictor we can very easily calculate the **probabilities**
- Remember, we are modeling $\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X$. We also know that we can solve for $p$:
  \[
  p = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}
  \]
- So we can calculate the probability for those with more than high school education and the probability for those with less
Probabilities

```
qui logit inlf hsp, nolog

* hsp = 1
  di exp(_b[_cons] + _b[hsp]) / (1 + exp(_b[_cons] + _b[hsp]))
  .67924528

* hsp = 0
  di exp(_b[_cons]) / (1 + exp(_b[_cons]))
  .52495379

* Relative
  di .67924528/ .52495379
  1.2939144

* Difference
  di .67924528 - .52495379
  .15429149
```

Before, we said that the odds were doubled, or 100% higher. Now in, 
the scale that matters, we say that they are only 30% higher. Or 15% 
percent points different
Big picture

- A ratio of odds is hard to interpret at best. At worse, it is misleading.
- We tend to think of them as a ratio of probabilities, but they are NOT.
- Often there is little resemblance between relative probabilities and odds ratios (unless events are rare).
- They tend to be often misreported and confusing; same with ratio of probabilities.
- For example, it sounds bad that event A is 10 times more likely to make you sick than event B, but that could be because $P_A = 0.001$ and $P_B = 0.0001$; their difference is 0.0009.
- My personal opinion: A ratio of probabilities can be confusing, a ratio of odds is EVIL.
Back to the continuous case

- Let’s go back to the model $\log\left( \frac{inlf_i}{1-inlf_i} \right) = \beta_0 + \beta_1 nwifeinc_i$
- We can also take $\exp(\beta_1)$. In this case, $\exp(-.0207569) = .97945704$
- A thousand dollars of extra income decreases the odds of participating in the labor force by a factor of 0.98
- Again, same issue. We can also solve for $p$ or $\text{inlf}$ in this case but not as easy as before because $nwifeinc$ is continuous
- We could take, as with the linear model, the derivative of $p$ with respect to $nwifeinc$, but we know that it’s non-linear so there is not a single effect; it depends on the values of $nwifeinc$
- Solution: We will do it numerically
We will do something that is conceptually very simple to numerically get the derivative

1. Estimate the model
2. For each observation, calculate predictions in the probability scale
3. Increase the nwifeinc by a “small” amount and calculate predictions again
4. Calculate the change in the two predictions as a fraction of the change in nwifeinc. In other words, calculate $\frac{\Delta Y}{\Delta X}$, which is the definition of the derivative
5. Take the average of the change in previous step across observations

That’s it
Numerical derivative

```
preserve
    qui logit inlf nwifeinc, nolog

    predict inlf_0 if e(sample)
    replace nwifeinc = nwifeinc + 0.011
    predict inlf_1 if e(sample)

    gen dydx = (inlf_1 - inlf_0) / 0.011

    sum dydx
restore
```

<table>
<thead>
<tr>
<th>Variable</th>
<th>Obs</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>Min</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>dydx</td>
<td>753</td>
<td>-0.0050217</td>
<td>0.001554</td>
<td>-0.005191</td>
<td>-0.0034977</td>
</tr>
</tbody>
</table>

- A small increase in extra income decreases the probability of entering the labor force by 0.005
That’s what Stata calls **marginal effects**

```stata
qui logit inlf nwifeinc, nolog
margins, dydx(nwifeinc)

Average marginal effects  Number of obs  =  753
Model VCE      : OIM
Expression     : Pr(inlf), predict()
dy/dx w.r.t.   : nwifeinc

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dy/dx</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------</td>
</tr>
<tr>
<td>nwifeinc</td>
<td>-.0050217</td>
</tr>
</tbody>
</table>
```

See, piece of cake!
Comments

- If you do it “by hand,” you can calculate any change, not just small changes
- Small is relative. A small change in age is not the same as a small change in income when income is measured in thousands
- I got the 0.011 by dividing the standard deviation of nwifeinc by 1000; that seems to be close to what Stata does
- Once we have other variables, we have to “hold them constant” at some values
- For marginal effects, it won’t matter at which value you hold them constant
Margins for indicator variables

qui logit inlf i.hsp, nolog
margin, dydx(hsp)

Conditional marginal effects

| Expression : Pr(inlf), predict() |
| dy/dx w.r.t. : 1.hsp |
| Number of obs : 753 |
Model VCE : OIM

| dy/dx | Std. Err. | z | P>|z| | [95% Conf. Interval] |
|-------|-----------|---|-----|-----------------------|
| 1.hsp | 0.1542915 | 0.038583 | 4.00 | 0.000 | 0.0786701 | 0.2299128 |

Note: dy/dx for factor levels is the discrete change from the base level.

- Same as what we found before doing it by hand. If we have covariates, we need to hold them constant at some value
- **Always use factor variable notation** with margins to avoid mistakes
Please be **fearful** of the margin command; it’s healthy

```
margin, dydx(hsp)

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>dy/dx</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td>1.hsp</td>
<td>.1542915</td>
<td>.038583</td>
<td>4.00</td>
<td>0.000</td>
<td>.0786701 - .2299128</td>
</tr>
</tbody>
</table>
```

Note: dy/dx for factor levels is the discrete change from the base level.

```
margin i.hsp
```

```
<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Margin</td>
<td>Std. Err.</td>
<td>z</td>
<td>P&gt;</td>
<td>z</td>
</tr>
<tr>
<td>hsp</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.5249538</td>
<td>.0214699</td>
<td>24.45</td>
<td>0.000</td>
<td>.4828736 - .567034</td>
</tr>
<tr>
<td>1</td>
<td>.6792453</td>
<td>.0320577</td>
<td>21.19</td>
<td>0.000</td>
<td>.6164134 - .7420772</td>
</tr>
</tbody>
</table>
```

```
margin
```

```
<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>_cons</td>
<td>.5683931</td>
<td>.0178717</td>
<td>31.80</td>
<td>0.000</td>
<td>.5333652 - .603421</td>
</tr>
</tbody>
</table>
```

**Small syntax changes make a big difference.** The third version is just the average prediction; same as observed proportion

©2017 PERRAILLON ARR
Note on predictions and Stata and odds ratios

- By default, Stata calculates predictions in the probability scale.
- You can also request predictions in the log-odds or logit scale.
- By default, Stata shows you the coefficients in the estimation scale (that is, log-odds).
- You can also request coefficients in the odds-ration scale.
- But since you know they are evil, don’t do it.
Sata things

qui logit inlf i.hsp nwifeinc, nolog

* Default, probability scale
predict hatp if e(sample)
(option pr assumed; Pr(inlf))

* Logit scale
predict hatp_l, xb

* Request odds ratios
logit inlf i.hsp nwifeinc, or nolog

|     | Odds Ratio | Std. Err. |    z  |   P>|z|  | [95% Conf. Interval] |
|-----|------------|-----------|------|------|---------------------|
| 1.hsp | 2.461153  | .4532018  | 4.89 | 0.000 | 1.715523 3.530861   |
| nwifeinc | .9689898 | .0069954  | -4.36| 0.000 | .9553756 .9827981   |
| _cons | 1.95093   | .303736   | 4.29 | 0.000 | 1.437872 2.647058   |

* That 2.46? 0.20 in probability scale, 39% more in relative probability:
margins hsp

<table>
<thead>
<tr>
<th></th>
<th>Delta-method</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Margin</td>
</tr>
<tr>
<td>-----</td>
<td>--------------</td>
</tr>
<tr>
<td>hsp</td>
<td></td>
</tr>
<tr>
<td>0</td>
<td>.5100289</td>
</tr>
<tr>
<td>1</td>
<td>.7137439</td>
</tr>
</tbody>
</table>

di .7137439/.5100289
1.3994185
di .7137439-.5100289
.203715
Summary

- Main difficulty with logistic models is to interpret parameters.
- We estimate models in log-odds scale, we can easily convert coefficients into odds ratios but we really care about probabilities because a ratio of odds is not that informative (they are EVIL).
- We can use numerical “derivatives” to come up with average predicted differences, what economists and Stata call marginal effects.
- With more covariates, we just add our usual “holding other factors constant” or “taking into account other factors”.
- We will do more of that next class.