RELATIVE RISKS, RELATIVE ODDS, LOGISTIC REGRESSION

RELATIVE RISKS: Suppose we are interested in the association between lung cancer and smoking. Consider the following table for the whole population:

<table>
<thead>
<tr>
<th>Lung cancer</th>
<th>Smoker</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>LC&amp;S</td>
<td>m₁₁</td>
<td>m₁₂</td>
<td>m₁</td>
</tr>
<tr>
<td>No</td>
<td>nLC&amp;S</td>
<td>m₂₁</td>
<td>m₂₂</td>
<td>m₂</td>
</tr>
<tr>
<td>Total</td>
<td>S</td>
<td>m₁₁⁺m₂₁</td>
<td>m₁₂⁺m₂₂</td>
<td>m₁⁺m₂</td>
</tr>
</tbody>
</table>

m₁₁ = number of people who both have lung cancer and smoke
m₁₂ = number of people who both have lung cancer and don’t smoke
m₂₁ = number of people who both do not have lung cancer and smoke
m₂₂ = number of people who both do not have lung cancer and do not smoke

The dot in the subscript shows that the values were summed over all values in that position

e.g. m₁ = m₁₁ + m₁₂

We could ask whether the risk of lung cancer is higher among smokers than non-smokers: is

$$\frac{m_{11}}{m_1} > \frac{m_{12}}{m_2}$$

This question can be asked in another way: is the relative risk of cancer for smokers compared to non-smokers greater than 1? The relative risk is the ratio of the risk for smokers compared to non-smokers, i.e.

$$\text{relative risk} = \frac{\text{risk for smokers}}{\text{risk for nonsmokers}} = \frac{\frac{m_{11}}{m_1}}{\frac{m_{12}}{m_2}}$$

A relative risk greater than 1 means the smokers have higher risk of lung cancer than non-smokers.

A made-up example:

<table>
<thead>
<tr>
<th>Lung cancer</th>
<th>Smoker</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>120</td>
<td>12</td>
<td>132</td>
<td></td>
</tr>
<tr>
<td>No</td>
<td>1,199,880</td>
<td>1,199,988</td>
<td>2,399,868</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>1,200,000</td>
<td>1,200,000</td>
<td>2,400,000</td>
<td></td>
</tr>
</tbody>
</table>

Risk .0001 .00001 Relative Risk: .0001/.00001 = 10

1/10,000 1/100,000

The risk is 10 times higher for smokers than non-smokers - but because lung cancer is rare, we had to look at a lot of
people - here over 2 million - to find the relative risk.
RELATIVE ODDS: Instead of calculating the risk of lung cancer, we can calculate the odds of lung cancer for smokers and for non-smokers. The odds of lung cancer is the number with lung cancer/ number without lung cancer.

For smokers: \( \frac{m_{11}}{m_{21}} = \frac{120}{1,199,880} \)
For nonsmokers: \( \frac{m_{12}}{m_{22}} = \frac{12}{1,199,988} \)

For those of you who know something about betting on horse races, then you know that the chances that a horse will will are given in terms of odds - odds of 3:2 mean the bookies guess that the horse has a 60% chance of winning.

The relative odds is the ratio of the odds. The relative odds for smokers compared to nonsmokers is given by

\[
\frac{\frac{m_{11}}{m_{21}}}{\frac{m_{12}}{m_{22}}} = \frac{m_{11}m_{22}}{m_{12}m_{21}}
\]

In our case, the relative odds and the relative risk are near identical - which happens when risks are small. Relative odds are important here for two reasons. It turns out that they can also be calculated in what are called case-control studies. They are also used in logistic regression.

CASE - CONTROL STUDIES

In a case-control study, we pick cases of the rare condition - here lung cancer - and match to them individuals with similar characteristics but who do not have lung cancer. In this case we are sampling on the dependent variable. What we will demonstrate is that, from this type of study, the odds ratio can be estimated.

<table>
<thead>
<tr>
<th>Lung cancer</th>
<th>Smoker</th>
<th>Yes</th>
<th>No</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>a</td>
<td>b</td>
<td></td>
<td>1000</td>
</tr>
<tr>
<td>No</td>
<td>c</td>
<td>d</td>
<td></td>
<td>1000</td>
</tr>
</tbody>
</table>

In this case, we have selected 1000 people with lung cancer and 1000 without. We are trying to estimate the relative odds of lung cancer for smokers as compared to nonsmokers. The relative odds in this case are \( \frac{a}{c} / \frac{b}{d} \). The question is whether this is the same as the relative odds when we have a true population sample, \( m_{11}m_{22}/m_{12}m_{21} \).

To see that this is indeed the case, we note that

\[
a = 1000 x \left( \frac{m_{11}}{m_{1.}} \right)
b = 1000 x \left( \frac{m_{21}}{m_{1.}} \right)
c = 1000 x \left( \frac{m_{12}}{m_{2.}} \right)
d = 1000 x \left( \frac{m_{22}}{m_{2.}} \right)
\]

Then \( \frac{ad}{bc} = \frac{m_{11}/m_{1.} \times m_{22}/m_{2.}}{m_{12}/m_{1.} \times m_{21}/m_{2.}} = \frac{m_{11}m_{22}}{m_{12}m_{21}} \)
Therefore, case-control studies allow us to estimate relative odds (which is also referred to as the odds ratio).

**LOGISTIC REGRESSION**

Logistic regression is a direct extension of the analysis of odds and odds ratios. Logistic regression is used when the dependent variable is a dichotomy - e.g. cancer, yes or no. What logistic regression does is consider the odds of being in category 1 of the dependent variable. If

\[
P = \text{probability of being in category 1 (e.g. of having lung cancer), and}
\]

\[
1-P = \text{probability of being in category 0 (of not having lung cancer),}
\]

then the odds of having cancer are \(p/(1-p)\). The log odds is called the logit of \(p\), i.e.,

\[
\text{logit } P = \log \frac{P}{1-P}
\]

Therefore, logit regression is concerned with predicting the probability or the odds of being in category 1. One of the benefits of using logs is that we never estimate a probability greater than 1 or less than zero, which could happen if we used OLS. (The logit of 0 is -\(4\) and the logit of 1 is \(4\).) Logit regression also overcomes the problem that, with only two possible values of the dependent variable, 0 or 1, the error term in the standard regression model cannot be normally distributed.

We should note that if we know \(L\), we can convert back to \(P\):

\[
P = \frac{1}{1 + e^{-L}}
\]

The model in logit regression (or logistic regression) is:

\[
\text{logit } P = \log \frac{P}{1-P} = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_3 + \ldots \alpha_k X_k + \ldots
\]

Before discussing how we estimate the \(\alpha\)'s, let's look at an example. It comes from an analysis by Furstenberg and Morgan of the National Survey of Children and looks at 15 and 16 year-olds. The dependent variable of interest is \(Y = \text{Ever had intercourse}\). \(Y=1\) if they have and 0 if not.

In one of their papers, they considered two dummy variable predictors: Black and Female.
The interpretation of the constant, -1.12, is that the odds of having intercourse for a white male (female=0, black=0) are $e^{-1.12}$.

The coefficient of female, -0.65: The relative odds of having ever had intercourse for girls compared to boys of the same race is $e^{-0.65}$.

The coefficient of black, 1.31: The relative odds of having ever had intercourse for blacks compared to whites of the same gender are $e^{1.31}$.

We can either calculate the relative odds ourselves (using the `display` command in STATA) or ask STATA to compute them for us, by using the `logistic` command.

```
. logistic inter female black
```

We also note that when $\hat{\alpha} < 0$ odds decrease as $X$ increases (e.g. as we go from female = 0 to female = 1 $\hat{\alpha} = 0$ X is unrelated to Y. $\hat{\alpha} > 0$ odds are greater as $X$ increases (e.g. as we go from black = 0 to black =1).

In general, $e^{\alpha}$ is the relative odds that $Y=1$ when an individual with $X=x$ is compared to an individual with exactly the same characteristics except $X=x-1$. 

---

```
.logit inter female black
Iteration 0:  Log Likelihood = -264.6267
Iteration 1:  Log Likelihood = -246.37474
Iteration 2:  Log Likelihood = -245.89753
Iteration 3:  Log Likelihood = -245.89738

Logit Estimates
Number of obs = 462
chi2(2) = 37.46
Prob > chi2 = 0.0000
Log Likelihood = -245.89738
Pseudo R2 = 0.0708

|         | Coef.   | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|---------|---------|-----------|-------|-------|---------------------|
| female  | -.6478314 | .2250376 | -2.879 | 0.004 | -1.090063 to -.2055996 |
| black   | 1.31346 | .2377786 | 5.524 | 0.000 | .8461905 to 1.78073 |
| _cons   | -1.12112 | .1624118 | -6.903 | 0.000 | -1.440282 to -.8019567 |
```

---

```
.logistic inter female black
Logit Estimates
Number of obs = 462
chi2(2) = 37.46
Prob > chi2 = 0.0000
Log Likelihood = -245.89738
Pseudo R2 = 0.0708

|         | Odds Ratio | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|---------|------------|-----------|-------|-------|---------------------|
| female  | .5231791   | .117735   | -2.879 | 0.004 | .3361953 to .814159 |
| black   | 3.71902    | .8843034  | 5.524 | 0.000 | 2.330751 to 5.934185 |
```

---

```
.logistic inter female black
Logit Estimates
Number of obs = 462
chi2(2) = 37.46
Prob > chi2 = 0.0000
Log Likelihood = -245.89738
Pseudo R2 = 0.0708

|         | Odds Ratio | Std. Err. | t     | P>|t|   | [95% Conf. Interval] |
|---------|------------|-----------|-------|-------|---------------------|
| female  | .5231791   | .117735   | -2.879 | 0.004 | .3361953 to .814159 |
| black   | 3.71902    | .8843034  | 5.524 | 0.000 | 2.330751 to 5.934185 |
```
ESTIMATING THE PROBABILITY THAT Y=1

STATA will estimate P from these results. For example, for a black female, we estimate

\[ \logit P = -1.12 - .65 \text{ (female=1)} + 1.31 \text{ (black=1)} = L \]

Then the estimated \( P = \frac{1}{1+e^{-L}} \).

After we use the command logit, we can then issue the command:

\[ . \text{predict phat} \]

\[ . \text{tab phat} \]

<table>
<thead>
<tr>
<th>phat</th>
<th>Freq.</th>
<th>Percent</th>
<th>Cum.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>.1456728</td>
<td>175</td>
<td>37.88</td>
<td>37.88</td>
</tr>
<tr>
<td>.2458037</td>
<td>177</td>
<td>38.31</td>
<td>76.19</td>
</tr>
<tr>
<td>.3880561</td>
<td>58</td>
<td>12.55</td>
<td>88.74</td>
</tr>
<tr>
<td>.5479375</td>
<td>52</td>
<td>11.26</td>
<td>100.00</td>
</tr>
<tr>
<td>Total</td>
<td>462</td>
<td>100.00</td>
<td></td>
</tr>
</tbody>
</table>

white females
white males
black females
black males

LINEAR VS MULTIPLICATIVE MODEL

The logit model is linear:

\[ \logit P = \log \frac{P}{1-P} = \hat{\alpha}_0 + \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2 + \hat{\alpha}_3 X_3 + \cdots + \hat{\alpha}_K X_K \]

But the model for the odds is multiplicative (please note that the subscripts go on the same line when they are already in a superscript):

\[ \frac{P}{1-P} = e^{\hat{\alpha}_0 + \hat{\alpha}_1 X_1 + \hat{\alpha}_2 X_2 + \hat{\alpha}_3 X_3 + \cdots + \hat{\alpha}_K X_K} \]

HYPOTHESIS TESTS

The STATA output gives us approximate t-tests for the coefficients and approximate confidence intervals. There is also a test analogous to the F-test that allows us to compare nested models - and this test is actually the more precise one. It is a \( F^2 \) test. The degrees of freedom for the test are equal to the number of X variables included in the model. The statistic compares the model with only the intercept to the one with K-1 predictors. In the case of our example above, we have two predictors, and the chi2(2) tells us that, with two variables, we fit our data significantly better than if we only use an intercept.

The actual test statistic uses the Log Likelihood, which is calculated because the method of estimation is that of maximum likelihood. In this case, the quantity -2 Log Likelihood has a chi-square distribution, with degrees of freedom = n-K (n=sample size, K=number of estimated parameters).
The logit and logistic output in STATA gives the information for testing whether our model fits better than the one with intercept only. The difference between two likelihood statistics also has a chi-square distribution with df= K-H where H is the number of parameters estimated in the smaller model. Therefore, the chi2(K-H) statistic given in the STATA output is actually

\[-2 \text{Log Likelihood}_k - -2 \text{Log Likelihood}_h = -2 (\text{Log Likelihood}_k - \text{Log Likelihood}_h)\]

Logit regression results differ in one important way from OLS. In OLS, the F-test was completely equivalent to the t-test when we were comparing a model with a nested model with one fewer predictors. Here, the chi-square test may give a different result from the individual t-tests - and the chi-square test is more reliable.

Next time we will discuss how we get the estimates of the coefficients - what this likelihood is - and some diagnostics.