9 Review of Discrete Data Analysis

The material in this section was covered last semester. Since Stata differs from Minitab in how the methods are implemented, we will review those methods and see how to use Stata for them. The huge difference from what we have been doing is that the response or outcome variable is now categorical instead of continuous. Our goal is to extend all the t-test, regression, ANOVA, and ANCOVA methods we have studied to the case of categorical outcomes.

Comparing Two Proportions: Independent Samples

The New Mexico state legislature is interested in how the proportion of registered voters that support Indian gaming differs between New Mexico and Colorado. Assuming neither population proportion is known, the state's statistician might recommend that the state conduct a survey of registered voters sampled independently from the two states, followed by a comparison of the sample proportions in favor of Indian gaming.

Statistical methods for comparing two proportions using independent samples can be formulated as follows. Let p_1 and p_2 be the proportion of populations 1 and 2, respectively, with the attribute of interest. Let $\hat{p}_1$ and $\hat{p}_2$ be the corresponding sample proportions, based on independent random or representative samples of size n_1 and n_2 from the two populations.

Large Sample CI and Tests for $p_1 - p_2$

A large sample CI for $p_1 - p_2$ is $(\hat{p}_1 - \hat{p}_2) \pm z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2)$, where z_{crit} is the standard normal critical value for the desired confidence level, and

$$SE_{CI}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

is the CI standard error.

A large sample p-value for a test of the null hypothesis $H_0 : p_1 - p_2 = 0$ against the two-sided alternative $H_A : p_1 - p_2 \neq 0$ is evaluated using tail areas of the standard normal distribution (identical to 1 sample evaluation) in conjunction with the test statistic

$$z_s = \frac{\hat{p}_1 - \hat{p}_2}{SE_{test}(\hat{p}_1 - \hat{p}_2)},$$

where

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{\frac{\bar{p}(1-\bar{p})}{n_1} + \frac{\bar{p}(1-\bar{p})}{n_2}} = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

is the test standard error for $\hat{p}_1 - \hat{p}_2$. The **pooled proportion**

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

is the proportion of successes in the two samples combined. The test standard error has the same functional form as the CI standard error, with $\bar{p}$ replacing the individual sample proportions.

The pooled proportion is the best guess at the common population proportion when $H_0 : p_1 = p_2$ is true. The test standard error estimates the standard deviation of $\hat{p}_1 - \hat{p}_2$ assuming H_0 is true.

Example Two hundred and seventy nine French skiers were studied during two one-week periods in 1961. One group of 140 skiers receiving a placebo each day, and the other 139 receiving 1 gram of ascorbic acid (Vitamin C) per day. The study was double blind - neither the subjects

nor the researchers knew who received what treatment. Let p_1 be the probability that a member of the ascorbic acid group contracts a cold during the study period, and p_2 be the corresponding probability for the placebo group. Linus Pauling and I are interested in testing whether $p_1 = p_2$. The data are summarized below as a two-by-two table of counts (a contingency table)

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

The sample sizes are $n_1 = 139$ and $n_2 = 140$. The sample proportion of skiers developing colds in the placebo and treatment groups are $\hat{p}_2 = 31/140 = .221$ and $\hat{p}_1 = 17/139 = .122$, respectively. The pooled proportion is the number of skiers that developed colds divided by the number of skiers in the study: $\bar{p} = 48/279 = .172$.

The test standard error is:

$$SE_{test}(\hat{p}_1 - \hat{p}_2) = \sqrt{.172 * (1 - .172)\left(\frac{1}{139} + \frac{1}{140}\right)} = .0452.$$

The test statistic is

$$z_s = \frac{.122 - .221}{.0452} = -2.19.$$

The p-value for a two-sided test is twice the area under the standard normal curve to the right of 2.19 (or twice the area to the left of -2.19), which is $2 * (.014) = .028$ At the 5% level, we reject the hypothesis that the probability of contracting a cold is the same whether you are given a placebo or Vitamin C.

A CI for $p_1 - p_2$ provides a measure of the size of the treatment effect. For a 95% CI

$$z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{.221 * (1 - .221)}{140} + \frac{.122 * (1 - .122)}{139}} = 1.96 * (.04472) = .088.$$

The 95% CI for $p_1 - p_2$ is $(.122 - .221) \pm .088$, or $(-.187, -.011)$. We are 95% confident that p_2 exceeds p_1 by at least .011 but not by more than .187.

On the surface, we would conclude that a daily dose of Vitamin C decreases a French skier's chance of developing a cold by between .011 and .187 (with 95% confidence). This conclusion was somewhat controversial. Several reviews of the study felt that the experimenter's evaluations of cold symptoms were unreliable. Many other studies refute the benefit of Vitamin C as a treatment for the common cold.

To implement this test and obtain a CI using **Stata**'s `prtesti` command (immediate from of prtest command – uses data on the command line rather than in memory), we must provide the raw number of skiers receiving ascorbic acid (139) along with the proportion of these skiers that got a cold ($\hat{p}_1 = 0.122$), as well as the raw number of skiers receiving placebo (140) along with the proportion of these skiers that got a cold ($\hat{p}_2 = 0.221$). I actually like using the GUI `(Statistics -> Summaries, tables & tests -> Classical tests of hypotheses -> Two sample proportion calculator)` instead of the command line for this, in which case it all looks just like Minitab. Options and entries are a little more obvious from the GUI.

```
. prtesti 139 0.122 140 0.221
Two-sample test of proportion                     x: Number of obs =      139
                                                  y: Number of obs =      140
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |       .122     .02776                      .0675914    .1764086
           y |       .221   .0350672                      .1522696    .2897304
-------------+----------------------------------------------------------------
        diff |      -.099    .044725                     -.1866594   -.0113406
             |  under Ho:    .045153    -2.19   0.028
------------------------------------------------------------------------------

              Ho: proportion(x) - proportion(y) = diff = 0
      Ha: diff < 0            Ha: diff != 0             Ha: diff > 0
       z = -2.193              z = -2.193               z = -2.193
   P < z =  0.0142        P > |z| =  0.0283          P > z =  0.9858
```

It actually is a little more direct to use counts instead of proportions you calculate, by typing `prtesti 139 17 140 31, count`.

Example A case-control study was designed to examine risk factors for cervical dysplasia (Becker et al. 194). All the women in the study were patients at UNM clinics. The 175 cases were women, aged 18-40, who had cervical dysplasia. The 308 controls were women aged 18-40 who did not have cervical dysplasia. Each women was classified as positive or negative, depending on the presence of HPV (human papilloma virus).

The data are summarized below.

HPV Outcome	Cases	Controls
Positive	164	130
Negative	11	178
Sample size	175	308

Let p_1 be the probability that a case is HPV positive and let p_2 be the probability that a control is HPV positive. The sample sizes are $n_1 = 175$ and $n_2 = 308$. The sample proportions of positive cases and controls are $\hat{p}_1 = 164/175 = .937$ and $\hat{p}_2 = 130/308 = .422$.

For a 95% CI

$$z_{crit} SE_{CI}(\hat{p}_1 - \hat{p}_2) = 1.96\sqrt{\frac{.937 * (1 - .937)}{175} + \frac{.422 * (1 - .422)}{308}} = 1.96 * (.03336) = .0659.$$

A 95% CI for $p_1 - p_2$ is $(.937 - .422) \pm .066$, or $.515 \pm .066$, or $(.449, .581)$. I am 95% confident that p_1 exceeds p_2 by at least .45 but not by more than .58.

Not surprisingly, a two-sided test at the 5% level would reject $H_0 : p_1 = p_2$. In this problem one might wish to do a one-sided test, instead of a two-sided test. Can you find the p-value for the one-sided test in the **Stata** output below?

```
. prtesti 175 0.937 308 0.422
Two-sample test of proportion                     x: Number of obs =      175
                                                  y: Number of obs =      308
------------------------------------------------------------------------------
    Variable |       Mean   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |       .937   .0183663                      .9010028    .9729972
           y |       .422   .0281413                       .366844     .477156
-------------+----------------------------------------------------------------
        diff |       .515   .0336044                      .4491366    .5808634
             |  under Ho:   .0462016    11.15   0.000
------------------------------------------------------------------------------

              Ho: proportion(x) - proportion(y) = diff = 0
      Ha: diff < 0            Ha: diff != 0             Ha: diff > 0
       z = 11.147              z = 11.147               z = 11.147
   P < z =  1.0000        P > |z| =  0.0000          P > z =  0.0000
```

Appropriateness of the Large Sample Test and CI

The standard two sample CI and test used above are appropriate when each sample is large. A rule of thumb suggests a minimum of at least five successes (i.e. observations with the characteristic of interest) and failures (i.e. observations without the characteristic of interest) in each sample before using these methods. This condition is satisfied in our two examples.

Effect Measures in Two-by-Two Tables

Consider a study of a particular disease, where each individual is either exposed or not-exposed to a risk factor. Let p_1 be the proportion diseased among the individuals in the exposed population, and p_2 be the proportion diseased among the non-exposed population. This population information can be summarized as a two-by-two table of population proportions:

Outcome	Exposed population	Non-Exposed population
Diseased	p_1	p_2
Non-Diseased	$1-p_1$	$1-p_2$

A standard measure of the difference between the exposed and non-exposed populations is the **absolute difference:** $p_1 - p_2$. We have discussed statistical methods for assessing this difference.

In many epidemiological and biostatistical settings, other measures of the difference between populations are considered. For example, the relative risk

$$\text{RR} = \frac{p_1}{p_2}$$

is commonly reported when the individual risks p_1 and p_2 are small. The odds ratio

$$\text{OR} = \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$

is another standard measure. Here $p_1/(1-p_1)$ is the odds of being diseased in the exposed group, whereas $p_2/(1-p_2)$ is the odds of being diseased in the non-exposed group.

Note that each of these measures can be easily estimated from data, using the sample proportions as estimates of the unknown population proportions. For example, in the vitamin C study:

Outcome	Ascorbic Acid	Placebo
# with cold	17	31
# with no cold	122	109
Totals	139	140

the proportion with colds in the placebo group is $\hat{p}_2 = 31/140 = .221$. The proportion with colds in the vitamin C group is $\hat{p}_1 = 17/139 = .122$.

The estimated absolute difference in risk is $\hat{p}_1 - \hat{p}_2 = .122 - .221 = -.099$. The estimated risk ratio and odds ratio are

$$\widehat{RR} = \frac{.122}{.221} = .55$$

and

$$\widehat{OR} = \frac{.122/(1-.122)}{.221/(1-.221)} = .49,$$

respectively.

In the literature it probably is most common to see OR (actually $\widehat{OR}$ or adjusted $\widehat{OR}$) reported, usually from a logistic regression analysis — that will be covered in the next section). We will be interested in testing $H_0 : OR = 1$ (or $H_0 : RR = 1$). We will estimate OR with $\widehat{OR}$ and will need the sampling distribution of $\widehat{OR}$ in order to construct tests and confidence intervals.

Testing for Homogeneity of Proportions

Example The following two-way table of counts summarizes the location of death and age at death from a study of 2989 cancer deaths (Public Health Reports, 1983):

(Obs Counts)	Location of death			
Age	Home	Acute Care	Chronic care	Row Total
15-54	94	418	23	535
55-64	116	524	34	674
65-74	156	581	109	846
75+	138	558	238	934
Col Total	504	2081	404	2989

The researchers want to compare the age distributions across locations. A one-way ANOVA would be ideal if the actual ages were given. Because the ages are grouped, the data should be treated as categorical. Given the differences in numbers that died at the three types of facilities, a comparison of proportions or percentages in the age groups is appropriate. A comparison of counts is not.

The table below summarizes the proportion in the four age groups at each location. For example, in the acute care facility 418/2081 = .201 and 558/2081 = .268. The **pooled proportions** are the Row Totals divided by the total sample size of 2989. The pooled summary gives the proportions in the four age categories, ignoring location of death.

The age distributions for home and for the acute care facilities are similar, but are very different from the age distribution at chronic care facilities.

To formally compare the observed proportions, one might view the data as representative sample of ages at death from the three locations. Assuming independent samples from the three locations (populations), a chi-squared statistic is used to test whether the population proportions of ages at death are identical (homogeneous) across locations. The **chi-squared test for homogeneity** of population proportions can be defined in terms of proportions, but is traditionally defined in terms of counts.

(Proportions)	Location of death			
Age	Home	Acute Care	Chronic care	Pooled
15-54	.187	.201	.057	.179
55-64	.230	.252	.084	.226
65-74	.310	.279	.270	.283
75+	.273	.268	.589	.312
Total	1.000	1.000	1.000	1.000

In general, assume that the data are independent samples from c populations (strata, groups, sub-populations), and that each individual is placed into one of r levels of a categorical variable. The raw data will be summarized as a $r \times c$ **contingency table** of counts, where the columns correspond to the samples, and the rows are the levels of the categorical variable. In the age distribution problem, $r = 4$ and $c = 3$. (SW uses k to identify the number of columns.)

To implement the test:

1. Compute the (estimated) **expected** count for each cell in the table as follows:

$$\mathrm{E} = \frac{\text{Row Total} * \text{Column Total}}{\text{Total Sample Size}}.$$

2. Compute the Pearson test statistic

$$\chi^2_S = \sum_{\text{all cells}} \frac{(O-E)^2}{E},$$

where O is the **observed** count.

3. For a size α test, reject the hypothesis of homogeneity if $\chi^2_S \geq \chi^2_{crit}$, where χ^2_{crit} is the upper α critical value from the chi-squared distribution with $df = (r-1)(c-1)$.

The p-value for the chi-squared test of homogeneity is equal to the area under the chi-squared curve to the right of χ^2_S; see Figure 1.

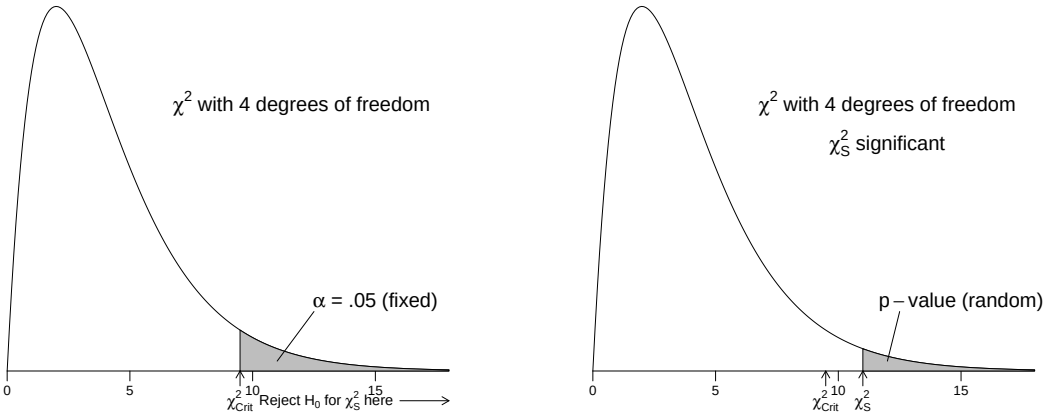

Figure 1: The p-value is the shaded area on the right

For a two-by-two table of counts, the chi-squared test of homogeneity of proportions is identical to the two-sample proportion test we discussed earlier.

Stata Analysis

One way to obtain the test statistic and p-value in **Stata** is to use the `tabi` command. The tables put out from that command are too poorly labelled to be very useful, though, so it's preferable to put the data into the worksheet so that it looks like this:

```
      Age   Location   Count
 1.     1          1      94
 2.     1          2     418
 3.     1          3      23
 4.     2          1     116
 5.     2          2     524
 6.     2          3      34
 7.     3          1     156
 8.     3          2     581
 9.     3          3     109
10.     4          1     138
11.     4          2     558
12.     4          3     238
```

The Hills and De Stavola book explains the following sequence,

```
. label define agemap 1  "15-54"  2  "55-64"  3   "65-74"   4  "75+"
. label define locmap 1 "Home" 2 "Acute Care" 3 "Chronic Care"
. label values Age agemap
. label values Location locmap
. list,clean
        Age       Location   Count
  1.   15-54           Home      94
  2.   15-54     Acute Care     418
  3.   15-54   Chronic Care      23
  4.   55-64           Home     116
  5.   55-64     Acute Care     524
  6.   55-64   Chronic Care      34
  7.   65-74           Home     156
  8.   65-74     Acute Care     581
  9.   65-74   Chronic Care     109
 10.     75+           Home     138
 11.     75+     Acute Care     558
 12.     75+   Chronic Care     238
```

If I typed `list,clean nolabel` I would get the original listing.

Why am I bothering with this? I actually could put those labels in as variable values, and not bother with labels. When I form tables, though, Stata wants to alphabetize according to variable values which will force Home as the last column. By keeping values numeric I can get Stata to order correctly and print the correct labels.

I find it easiest to go through the menu path `Summaries, tables, & tests -> Tables -> Two-way tables with measures of association` to generate the following commands. Note in particular the `[fweight = Count]` (frequency weight given by Count variable) syntax to tell Stata that each line represents many observations. Minitab and SAS have similar options.

```
. tabulate Age Location [fweight = Count], chi2 column expected lrchi2 row
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
| column percentage  |
+--------------------+

           |            Location
       Age |      Home  Acute Car  Chronic C |     Total
-----------+---------------------------------+----------
     15-54 |        94        418         23 |       535
           |      90.2      372.5       72.3 |     535.0
           |     17.57      78.13       4.30 |    100.00
           |     18.65      20.09       5.69 |     17.90
-----------+---------------------------------+----------
     55-64 |       116        524         34 |       674
           |     113.6      469.3       91.1 |     674.0
           |     17.21      77.74       5.04 |    100.00
           |     23.02      25.18       8.42 |     22.55
-----------+---------------------------------+----------
     65-74 |       156        581        109 |       846
           |     142.7      589.0      114.3 |     846.0
           |     18.44      68.68      12.88 |    100.00
           |     30.95      27.92      26.98 |     28.30
-----------+---------------------------------+----------
       75+ |       138        558        238 |       934
           |     157.5      650.3      126.2 |     934.0
           |     14.78      59.74      25.48 |    100.00
           |     27.38      26.81      58.91 |     31.25
-----------+---------------------------------+----------
     Total |       504      2,081        404 |     2,989
           |     504.0    2,081.0      404.0 |   2,989.0
           |     16.86      69.62      13.52 |    100.00
           |    100.00     100.00     100.00 |    100.00

          Pearson chi2(6) = 197.6241   Pr = 0.000
 likelihood-ratio chi2(6) = 200.9722   Pr = 0.000
```

The Pearson statistic is 197.6241 on 6 = (4-1)(3-1) *df*. The p-value is 0 to three places. The data strongly suggest that there are differences in the age distributions among locations.

Testing for Homogeneity in Cross-Sectional and Stratified Studies

Two-way tables of counts are often collected either by **stratified sampling** or by **cross-sectional** sampling.

In a stratified design, distinct groups, strata, or sub-populations are identified. Independent samples are selected from each group, and the sampled individuals are classified into categories. The HPV study is an illustration of a stratified design (and a case-control study). Stratified designs provide estimates for the strata (population) proportion in each of the categories. A test for **homogeneity of proportions** is used to compare the strata.

In a **cross-sectional design**, individuals are randomly selected from a population and classified by the levels of **two** categorical variables. With cross-sectional samples you can test homogeneity of proportions by comparing either the row proportions or by comparing the column proportions.

Example The following data (*The Journal of Advertising,* 1983, p. 34-42) are from a cross-sectional study that involved soliciting opinions on anti-smoking advertisements. Each subject was asked whether they smoked and their reaction (on a five-point ordinal scale) to the ad. The data are summarized as a two-way table of counts, given below:

	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	8	14	35	21	19	97
Non-smoker	31	42	78	61	69	281
Col Total	39	56	113	82	88	378

The row proportions are

(Row Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like	Row Tot
Smoker	.082	.144	.361	.216	.196	1.000
Non-smoker	.110	.149	.278	.217	.245	1.000

For example, the entry for the (Smoker, Str. Dislike) cell is: $8/97 = .082$.

Similarly, the column proportions are

(Col Prop)	Str. Dislike	Dislike	Neutral	Like	Str. Like
Smoker	.205	.250	.310	.256	.216
Non-smoker	.795	.750	.690	.744	.784
Total	1.000	1.000	1.000	1.000	1.000

Although it may be more natural to compare the smoker and non-smoker row proportions, the column proportions can be compared across ad responses. There is no advantage to comparing "rows" instead of "columns" in a formal test of homogeneity of proportions with cross-sectional data. The Pearson chi-squared test treats the rows and columns interchangeably, so you get the same result regardless of how you view the comparison. However, one of the two comparisons may be more natural to interpret.

Note that checking for homogeneity of proportions is meaningful in stratified studies only when the comparison is across strata! Further, if the strata correspond to columns of the table, then the column proportions or percentages are meaningful whereas the row proportions are not.

Question: How do these ideas apply to the age distribution problem?

Testing for Independence in a Two-Way Contingency Table

The row and column classifications for a population where each individual is cross-classified by two categorical variables are said to be independent if each **population** cell proportion in the two-way table is the product of the proportion in a given row and the proportion in a given column. One can show that independence is equivalent to homogeneity of proportions. In particular, the two-way table of population cell proportions satisfies independence if and only if the population column proportions are homogeneous. If the population column proportions are homogeneous then so are the population row proportions.

This suggests that a test for independence or **no association** between two variables based on a cross-sectional study can be implemented using the chi-squared test for homogeneity of proportions. This suggestion is correct. If independence is not plausible, I tend to interpret the dependence as a deviation from homogeneity, using the classification for which the interpretation is most natural.

Example: **Stata** output for testing independence between smoking status and ad reaction is given below. The Pearson chi-squared test is not significant (p-value = 0.559). The observed association between smoking status and the ad reaction is not significant. This suggests, for example, that the smoker's reactions to the ad were not statistically significantly different from the non-smoker's reactions, which is consistent with the smokers and non-smokers attitudes being fairly similar. The data were coded as opinion from 1 to 5 and smoke as 1 or 2, and then label define applied as before.

```
. tabulate Smoke Opinion [fweight=count],chi2 lrchi2 exp col row
+--------------------+
| Key                |
|--------------------|
|     frequency      |
| expected frequency |
|   row percentage   |
| column percentage  |
+--------------------+
           |                          Opinion
     Smoke | Str. Disl    Dislike    Neutral       Like  Str. Like |     Total
-----------+-------------------------------------------------------+----------
    Smoker |         8         14         35         21         19 |        97
           |      10.0       14.4       29.0       21.0       22.6 |      97.0
           |      8.25      14.43      36.08      21.65      19.59 |    100.00
           |     20.51      25.00      30.97      25.61      21.59 |     25.66
-----------+-------------------------------------------------------+----------
Non-smoker |        31         42         78         61         69 |       281
           |      29.0       41.6       84.0       61.0       65.4 |     281.0
           |     11.03      14.95      27.76      21.71      24.56 |    100.00
           |     79.49      75.00      69.03      74.39      78.41 |     74.34
-----------+-------------------------------------------------------+----------
     Total |        39         56        113         82         88 |       378
           |      39.0       56.0      113.0       82.0       88.0 |     378.0
           |     10.32      14.81      29.89      21.69      23.28 |    100.00
           |    100.00     100.00     100.00     100.00     100.00 |    100.00
          Pearson chi2(4) =   2.9907   Pr = 0.559
 likelihood-ratio chi2(4) =   2.9797   Pr = 0.561
```

One-sample procedures

Last semester we spent some time on the situation where we obtained a SRS of n observations from a binomial population (binary outcome variable) with probability p of Success. We learned how to calculate CIs for p and tests of $H_0 : p = p_0$ for some fixed p_0. The large sample form of this is also done with the `prtesti` command or through the GUI, and the (preferable) exact binomial test is done through the `bitesti` command (or through the menus). The extension to 3 or more categories was the chi-squared goodness of fit test, done in Stata using the `csgof` command. That command is not automatically installed but you can locate and install it from the `findit csgof`

command. Since we do these one sample procedures relatively infrequently, I am going to leave it to you to learn them in Stata if you need them.

10 Logistic Regression - Two Introductory Examples

The chi-squared tests in the previous section are used very frequently, along with Fisher's exact test (asked for with the `,fisher` option in tabulate – note that it is often feasible to calculate only for small sample sizes). Those "classical" methods have been around a very long time and are often the best choice for analysis. In order to consider problems with more complicated predictors we need newer technology, so we now turn to logistic regression.

The data below are from a study conducted by Milicer and Szczotka on pre-teen and teenage girls in Warsaw. The subjects were classified into 25 age categories. The number of girls in each group (sample size) and the number that reached menarche (# RM) at the time of the study were recorded. The age for a group corresponds to the midpoint for the age interval.

Sample size	# RM	Age	Sample size	# RM	Age
376	0	9.21	106	67	13.33
200	0	10.21	105	81	13.58
93	0	10.58	117	88	13.83
120	2	10.83	98	79	14.08
90	2	11.08	97	90	14.33
88	5	11.33	120	113	14.58
105	10	11.58	102	95	14.83
111	17	11.83	122	117	15.08
100	16	12.08	111	107	15.33
93	29	12.33	94	92	15.58
100	39	12.58	114	112	15.83
108	51	12.83	1049	1049	17.58
99	47	13.08			

The researchers were interested in whether the proportion of girls that reached menarche (# RM/ sample size) varied with age. One could perform a test of homogeneity by arranging the data as a 2 by 25 contingency table with columns indexed by age and two rows: ROW1 = # RM and ROW2 = # that have not RM = sample size − # RM. A more powerful approach treats these as regression data, using the proportion of girls reaching menarche as the "response" and age as a predictor.

The data were imported into **Stata** using the `infile` command and labelled `menarche`, `total`, and `age`. A plot of the observed proportion of girls that have reached menarche (obtained in **Stata** with the two commands `generate phat = menarche / total` and `twoway (scatter phat age)`) shows that the proportion increases as age increases, but that the relationship is nonlinear.

The observed proportions, which are bounded between zero and one, have a lazy S-shape (a **sigmoidal function**) when plotted against age. The change in the observed proportions for a given change in age is much smaller when the proportion is near 0 or 1 than when the proportion is near 1/2. This phenomenon is common with regression data where the response is a proportion.

The trend is nonlinear so linear regression is inappropriate. A sensible alternative might be to transform the response or the predictor to achieve near linearity. A better approach is to use a non-linear model for the proportions. A common choice is the **logistic regression model**.

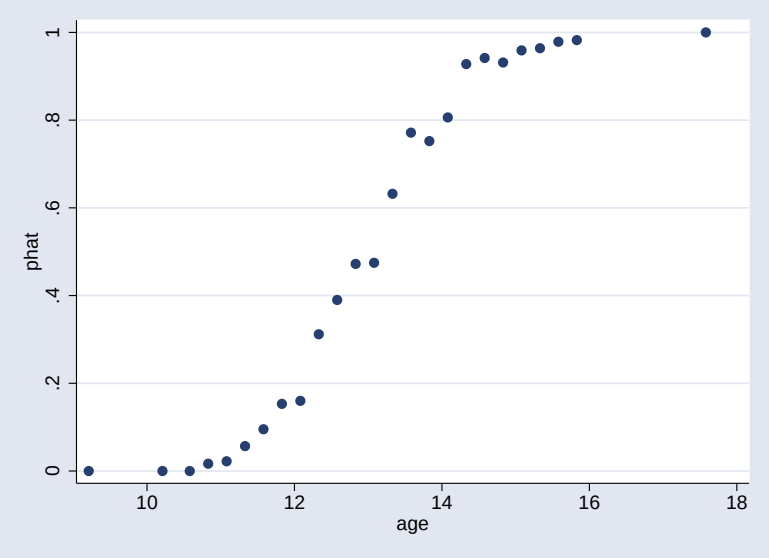

Figure 2: Estimated proportions $\hat{p}_i$ versus AGE_i, for $i = 1, \ldots, 25$.

The Simple Logistic Regression Model

The simple logistic regression model expresses the population proportion p of individuals with a given attribute (called a success) as a function of a single predictor variable X. The model assumes that p is related to X through

$$\text{logit}(p) = \log\left(\frac{p}{1-p}\right) = \alpha + \beta X \tag{1}$$

or, equivalently, as

$$p = \frac{exp(\alpha + \beta X)}{1 + exp(\alpha + \beta X)}.$$

The logistic regression model is a **binary response model**, where the response for each case falls into one of 2 exclusive and exhaustive categories, often called success (cases with the attribute of interest) and failure (cases without the attribute of interest). In many biostatistical applications, the success category is presence of a disease, or death from a disease.

I will often write p as $p(X)$ to emphasize that p is the proportion of all individuals with score X that have the attribute of interest. In the menarche data, $p = p(X)$ is the population proportion of girls at age X that have reached menarche.

The odds of success are $p/(1-p)$. For example, the odds of success are 1 (or 1 to 1) when $p = 1/2$. The odds of success are 2 (or 2 to 1) when $p = 2/3$. The logistic model assumes that the log-odds of success is linearly related to X. Graphs of the logistic model relating p to X are given in Figure 3. The sign of the slope refers to the sign of β.

There are a variety of other binary response models that are used in practice. The **probit** regression model or the **complementary log-log** regression model might be appropriate when the logistic model does not fit the data.

Data for Simple Logistic Regression

For the formulas below, I assume that the data are given in summarized or **aggregate** form:

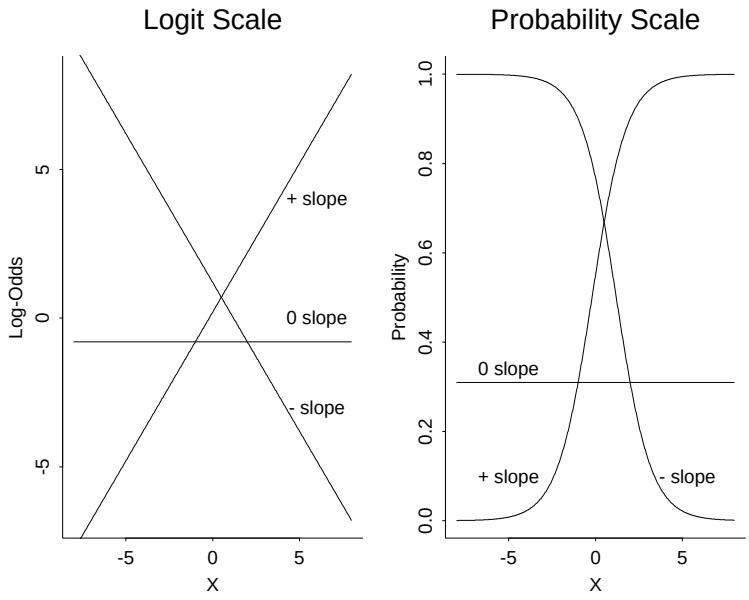

Figure 3: logit(p) and p as a function of X

X	n	D
X_1	n_1	d_1
X_2	n_2	d_2
.	.	.
.	.	.
X_m	n_m	d_m

where d_i is the number of individuals with the attribute of interest (number of diseased) among n_i randomly selected or representative individuals with predictor variable value X_i. The subscripts identify the group of cases in the data set. In many situations, the sample size is 1 in each group, and for this situation d_i is 0 or 1.

For **raw data** on individual cases, the sample size column n is usually omitted and D takes on 1 of two coded levels, depending on whether the case at X_i is a success or not. The values 0 and 1 are typically used to identify "failures" and "successes" respectively.

Estimating Regression Coefficients

The principle of maximum likelihood is commonly used to estimate the two unknown parameters in the logistic model:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta X.$$

The **maximum likelihood estimates** (MLE) of the regression coefficients are estimated iteratively by maximizing the so-called Binomial likelihood function for the responses, or equivalently, by minimizing the **deviance** function (also called the likelihood ratio LR chi-squared statistic)

$$\text{LR} = 2\sum_{i=1}^{m}\left\{d_i\log\left(\frac{d_i}{n_i p_i}\right) + (n_i - d_i)\log\left(\frac{n_i - d_i}{n_i - n_i p_i}\right)\right\}$$

over all possible values of α and β, where the p_is satisfy

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta X_i.$$

The ML method also gives standard errors and significance tests for the regression estimates.

The deviance is an analog of the residual sums of squares in linear regression. The choices for α and β that minimize the deviance are the parameter values that make the observed and fitted proportions as close together as possible in a "likelihood sense".

Suppose that $\hat{\alpha}$ and $\hat{\beta}$ are the MLEs of α and β. The deviance evaluated at the MLEs:

$$\text{LR} = 2\sum_{i=1}^{m}\left\{d_i \log\left(\frac{d_i}{n_i\tilde{p}_i}\right) + (n_i - d_i)\log\left(\frac{n_i - d_i}{n_i - n_i\tilde{p}_i}\right)\right\},$$

where the fitted probabilities $\tilde{p}_i$ satisfy

$$\log\left(\frac{\tilde{p}_i}{1-\tilde{p}_i}\right) = \hat{\alpha} + \hat{\beta}X_i,$$

is used to test the adequacy of the model. The deviance is small when the data fits the model, that is, when the observed and fitted proportions are close together. Large values of LR occur when one or more of the observed and fitted proportions are far apart, which suggests that the model is inappropriate.

If the logistic model holds, then LR has a chi-squared distribution with $m-r$ degrees of freedom, where m is the number of groups and r (here 2) is the number of estimated regression parameters. A p-value for the deviance is given by the area under the chi-squared curve to the right of LR. A small p-value indicates that the data does not fit the model.

Stata does not provide the deviance statistic, but rather the Pearson chi-squared test statistic, which is defined similarly to the deviance statistic and is interpreted in the same manner:

$$X^2 = \sum_{i=1}^{m}\frac{(d_i - n_i\tilde{p}_i)^2}{n_i\tilde{p}_i(1-\tilde{p}_i)}.$$

This statistic can be interpreted as the sum of standardized, squared differences between the *observed* number of successes d_i and *expected* number of successes $n_i\tilde{p}_i$ for each covariate X_i. When what we expect to see under the model agrees with what we see, the Pearson statistic is close to zero, indicating good model fit to the data. When the Pearson statistic is *large*, we have an indication of lack of fit. Often the *Pearson residuals* $r_i = (d_i - n_i\tilde{p}_i)/\sqrt{n_i\tilde{p}_i(1-\tilde{p}_i)}$ are used to determine exactly *where* lack of fit occurs. These residuals are obtained in **Stata** using the `predict` command after the `logistic` command. Examining these residuals is very similar to looking for large values of $\frac{(O-E)^2}{E}$ in a χ^2 analysis of a contingency table as discussed in the last lecture. We will not talk further of logistic regression diagnostics.

Age at Menarche Data: Stata Implementation

A logistic model for these data implies that the probability p of reaching menarche is related to age through

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta \text{ AGE}.$$

If the model holds, then a slope of $\beta = 0$ implies that p does not depend on AGE, i.e. the proportion of girls that have reached menarche is identical across age groups. However, the power of the logistic regression model is that if the model holds, and if the proportions change with age, then you have a way to quantify the effect of age on the proportion reaching menarche. This is more appealing and useful than just testing homogeneity across age groups.

A logistic regression model with a single predictor can be fit using one of the many commands available in **Stata** depending on the data type and desired results: `logistic` (raw data, outputs

odds ratios), `logit` (raw data, outputs model parameter estimates), and `blogit` (grouped data). The `logistic` command has many more options than either `logit` or `blogit`, but requires you to reformat the data into individual records, one for each girl. For an example of how to do this, check out the online **Stata** help at **http://www.stata.com/support/faqs/stat/grouped.html**. The **Stata** command `blogit menarche total age` yields the following output:

```
Logit estimates                                   Number of obs   =       3918
                                                  LR chi2(1)      =    3667.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -819.65237                       Pseudo R2       =     0.6911
------------------------------------------------------------------------------
    _outcome |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.631968   .0589509    27.68   0.000     1.516427     1.74751
       _cons |  -21.22639   .7706558   -27.54   0.000    -22.73685   -19.71594
------------------------------------------------------------------------------
```

The output tables the MLEs of the parameters: $\hat{\alpha} = -21.23$ and $\hat{\beta} = 1.63$. Thus, the fitted or predicted probabilities satisfy:

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = -21.23 + 1.63\text{AGE}$$

or

$$\tilde{p}(AGE) = \frac{\exp(-21.23 + 1.63\text{AGE})}{1 + \exp(-21.23 + 1.63\text{AGE})}.$$

The p-value for testing $H_0 : \beta = 0$ (i.e. the slope for the regression model is zero) based upon the chi-squared test p-value (`P>|z|`) is 0.000, which leads to rejecting H_0 at any of the usual test levels. Thus, the proportion of girls that have reached menarche is not constant across age groups.

The likelihood ratio test statistic of no logistic regression relationship (`LR chi2(1) = 3667.18`) and p-value (`Prob > chi2 = 0.0000`) gives the logistic regression analogue of the overall F-statistic that *no predictors are important* to multiple regression. In general, the chi-squared statistic provided here is used to test the hypothesis that the regression coefficients are zero for each predictor in the model. There is a single predictor here, AGE, so this test and the test for the AGE effect are *both* testing $H_0 : \beta = 0$.

To obtain the Pearson goodness of fit statistic and p-value we must reformat the data and use the `logistic` command as described in the webpage above:

```
generate w0 = total - menarche
rename menarche w1
generate id = _n
reshape long w, i(id) j(y)
logistic y age [fw=w]
lfit
```

We obtain the following output:

```
Logistic regression                               Number of obs   =       3918
                                                  LR chi2(1)      =    3667.18
                                                  Prob > chi2     =     0.0000
Log likelihood = -819.65237                       Pseudo R2       =     0.6911
------------------------------------------------------------------------------
           y | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   5.113931   .3014706    27.68   0.000     4.555917    5.740291
------------------------------------------------------------------------------

Logistic model for y, goodness-of-fit test

       number of observations =      3918
 number of covariate patterns =        25
             Pearson chi2(23) =        21.87
                  Prob > chi2 =         0.5281
```

Using properties of exponential functions, the odds of reaching menarche is $\exp(1.632) = 5.11$ times larger for every year older a girl is. To see this, let $p(\text{Age}+1)$ and $p(\text{Age})$ be probabilities of reaching menarche for ages one year apart. The odds ratio OR satisfies

$$\begin{aligned}
\log(OR) &= \log\left(\frac{p(\text{Age}+1)/(1-\text{p}(\text{Age}+1))}{p(\text{Age})/(1-\text{p}(\text{Age}))}\right) \\
&= \log\left(p(\text{Age}+1)/(1-\text{p}(\text{Age}+1))\right) - \log\left(p(\text{Age})/(1-\text{p}(\text{Age}))\right) \\
&= (\alpha+\beta(\text{Age}+1)) - (\alpha+\beta\ \text{Age}) \\
&= \beta
\end{aligned}$$

so $OR = e^{\beta}$. If we considered ages 5 years apart, the same derivation would give us $OR = e^{5\beta} = (e^{\beta})^5$. You often see a continuous variable with a significant though apparently small OR, but when you examine the OR for a reasonable range of values (by raising to the power of the range in this way), then the OR is substantial.

You should pick out the estimated regression coefficient $\hat{\beta} = 1.632$ and the estimated odds ratio $\exp(\hat{\beta}) = \exp(1.632) = 5.11$ from the output obtained using the `blogit` and `logistic` commands respectively. We would say that, for example, that the odds of 15 year old girls having reached menarche are between 4.5 and 5.7 times larger than for 14 year old girls.

The Pearson chi-square statistic is 21.87 on 23 df, with a p-value of 0.5281. The large p-value suggests no gross deficiencies with the logistic model.

Logistic Regression with Two Effects: Leukemia Data

Feigl and Zelen reported the survival time in weeks and the white cell blood count (WBC) at time of diagnosis for 33 patients who eventually died of acute leukemia. Each person was classified as AG+ or AG- (coded as IAG = 1 and 0, respectively), indicating the presence or absence of a certain morphological characteristic in the white cells. The researchers are interested in modelling the probability p of surviving at least one year as a function of WBC and IAG. They believe that WBC should be transformed to a log scale, given the skewness in the WBC values. Where `Live`=0, 1 indicates whether the patient died or lived respectively, the data are

IAG	WBC	Live	IAG	WBC	Live	IAG	WBC	Live
1	75	1	1	230	1	1	430	1
1	260	1	1	600	0	1	1050	1
1	1000	1	1	1700	0	1	540	0
1	700	1	1	940	1	1	3200	0
1	3500	0	1	5200	0	1	10000	1
1	10000	0	1	10000	0	0	440	1
0	300	1	0	400	0	0	150	0
0	900	0	0	530	0	0	1000	0
0	1900	0	0	2700	0	0	2800	0
0	3100	0	0	2600	0	0	2100	0
0	7900	0	0	10000	0	0	10000	0

As an initial step in the analysis, consider the following model:

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 \text{IAG},$$

where LWBC = log WBC. This is a logistic regression model with 2 effects, fit using the `logistic` command. The parameters α, β_1 and β_2 are estimated by maximum likelihood.

The model is best understood by separating the AG+ and AG- cases. For AG- individuals, IAG=0 so the model reduces to

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 * 0 = \alpha + \beta_1 \text{LWBC}.$$

For AG+ individuals, IAG=1 and the model implies

$$\log\left(\frac{p}{1-p}\right) = \alpha + \beta_1 \text{LWBC} + \beta_2 * 1 = (\alpha + \beta_2) + \beta_1 \text{LWBC}.$$

The model without IAG (i.e. $\beta_2 = 0$) is a simple logistic model where the log-odds of surviving one year is linearly related to LWBC, and is independent of AG. The reduced model with $\beta_2 = 0$ implies that there is no effect of the AG level on the survival probability *once LWBC has been taken into account.*

Including the **binary predictor** IAG in the model implies that there is a linear relationship between the log-odds of surviving one year and LWBC, with a constant slope for the two AG levels. This model includes an effect for the AG morphological factor, but more general models are possible. Thinking of IAG as a **factor**, the proposed model is a logistic regression analog of ANCOVA.

The parameters are easily interpreted: α and $\alpha + \beta_2$ are intercepts for the population logistic regression lines for AG- and AG+, respectively. The lines have a common slope, β_1. The β_2 coefficient for the IAG indicator is the difference between intercepts for the AG+ and AG- regression lines. A picture of the assumed relationship is given below for $\beta_1 < 0$. The population regression lines are parallel on the logit (i.e. log odds) scale only, but the order between IAG groups is preserved on the probability scale.

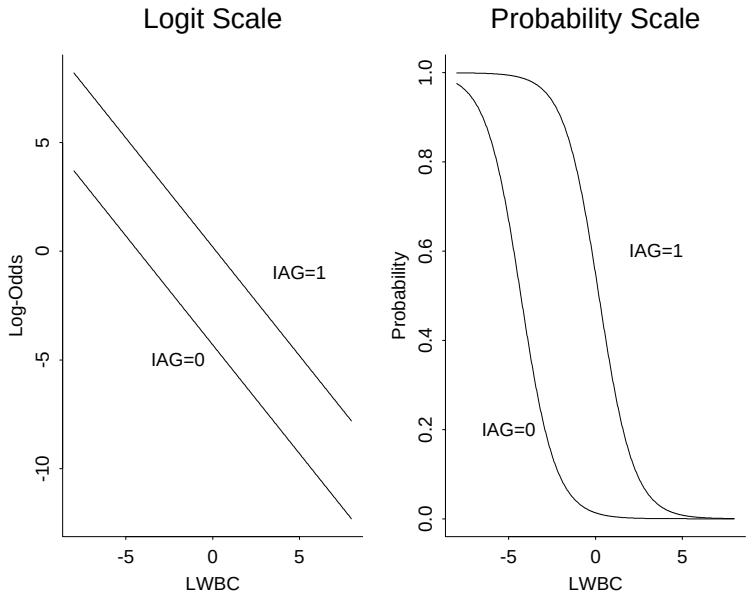

Figure 4: Predicted relationships on the logit and probability scales

The data are in the **raw data** form for individual cases. There are three columns: the binary or **indicator variable** `iag` (with value 1 for AG+, 0 for AG-), `wbc` (continuous), `live` (with value 1 if the patient lived at least 1 year and 0 if not). Note that a frequency column is not needed with

raw data (and hence using the `logistic` command) and that the success category corresponds to surviving at least 1 year.

Before looking at output for the equal slopes model, note that the data set has 30 distinct IAG and WBC combinations, or 30 "groups" or samples that could be constructed from the 33 individual cases. Only two samples have more than 1 observation. The majority of the observed proportions surviving at least one year (number surviving $\geq$ 1 year/ group sample size) are 0 (i.e. 0/1) or 1 (i.e. 1/1). This sparseness of the data makes it difficult to graphically assess the suitability of the logistic model (Why?). Although significance tests on the regression coefficients do not require large group sizes, the chi-squared approximations to the deviance and Pearson goodness-of-fit statistics are suspect in sparse data settings. With small group sizes as we have here, most researchers would not interpret the p-values for the deviance or Pearson tests literally. Instead, they would use the p-values to informally check the fit of the model. Diagnostics would be used to highlight problems with the model.

We obtain the following modified output:

```
. infile iag wbc live using c:/biostat/notes/leuk.txt
. generate lwbc = log(wbc)
. logistic live iag lwbc
. logit
. lfit
Logistic regression                                     Number of obs   =         33
                                                        LR chi2(2)      =      15.18
                                                        Prob > chi2     =     0.0005
Log likelihood = -13.416354                             Pseudo R2       =     0.3613
------------------------------------------------------------------------------
       live | Odds Ratio   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        iag |   12.42316    13.5497     2.31   0.021     1.465017    105.3468
       lwbc |   .3299682   .1520981    -2.41   0.016     .1336942    .8143885
------------------------------------------------------------------------------
Logit estimates                                         Number of obs   =         33
                                                        LR chi2(2)      =      15.18
                                                        Prob > chi2     =     0.0005
Log likelihood = -13.416354                             Pseudo R2       =     0.3613
------------------------------------------------------------------------------
       live |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
------------+-----------------------------------------------------------------
        iag |   2.519562   1.090681     2.31   0.021     .3818672    4.657257
       lwbc |  -1.108759   .4609479    -2.41   0.016      -2.0122   -.2053178
      _cons |   5.543349   3.022416     1.83   0.067     -.380477    11.46718
------------------------------------------------------------------------------
Logistic model for live, goodness-of-fit test
       number of observations =        33
 number of covariate patterns =        30
            Pearson chi2(27) =        19.81
                 Prob > chi2 =         0.8387
```

The large p-value (0.8387) for the lack-of-fit chi-square (i.e. the Pearson statistic) indicates that there are no gross deficiencies with the model. Given that the model fits reasonably well, a test of $H_0 : \beta_2 = 0$ might be a primary interest here. This checks whether the regression lines are identical for the two AG levels, which is a test for whether AG affects the survival probability, after taking LWBC into account. The test that $H_0 : \beta_2 = 0$ is equivalent to testing that the odds ratio $\exp(\beta_2)$ is equal to 1: $H_0 : e^{\beta_2} = 1$. The p-value for this test is 0.021. The test is rejected at any of the usual significance levels, suggesting that the AG level affects the survival probability (assuming a very specific model). In fact we estimate that the odds of surviving past a year in the AG+ population is 12.4 times the odds of surviving past a year in the AG- population, with a 95% CI of (1.4, 105.4); see below for this computation carried out explicitly.

The estimated survival probabilities satisfy

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52\text{IAG}.$$

For AG- individuals with IAG=0, this reduces to

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC},$$

or equivalently,

$$\tilde{p} = \frac{\exp(5.54 - 1.11\text{LWBC})}{1 + \exp(5.54 - 1.11\text{LWBC})}.$$

For AG+ individuals with IAG=1,

$$\log\left(\frac{\tilde{p}}{1-\tilde{p}}\right) = 5.54 - 1.11\text{LWBC} + 2.52 * (1) = 8.06 - 1.11\text{LWBC},$$

or

$$\tilde{p} = \frac{\exp(8.06 - 1.11\text{LWBC})}{1 + \exp(8.06 - 1.11\text{LWBC})}.$$

Using the **logit scale**, the difference between AG+ and AG- individuals in the estimated log-odds of surviving at least one year, at a fixed but arbitrary LWBC, is the estimated IAG regression coefficient:

$$(8.06 - 1.11\text{LWBC}) - (5.54 - 1.11\text{LWBC}) = 2.52.$$

Using properties of exponential functions, the odds that an AG+ patient lives at least one year is $\exp(2.52) = 12.42$ times larger than the odds that an AG- patient lives at least one year, regardless of LWBC.

Although the equal slopes model appears to fit well, a more general model might fit better. A natural generalization here would be to add an **interaction**, or product term, IAG $*$ LWBC to the model. The logistic model with an IAG effect and the IAG $*$ LWBC interaction is equivalent to fitting separate logistic regression lines to the two AG groups. This interaction model provides an easy way to test whether the slopes are equal across AG levels. I will note that the interaction term is not needed here.