Lecture 10: Logistic Regression - Two Introductory Examples

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

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`

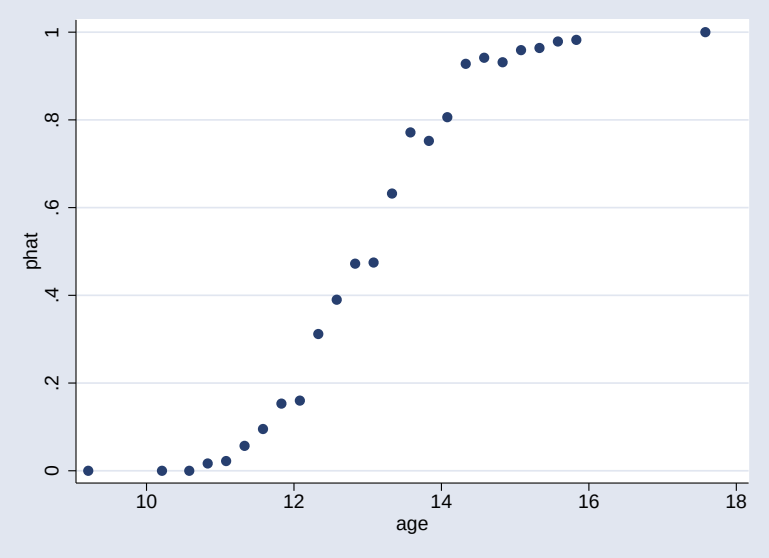


Figure 1: Estimated proportions $\hat{p}_i$ versus AGE_i, for $i = 1, \ldots, 25$.

(`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**.

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

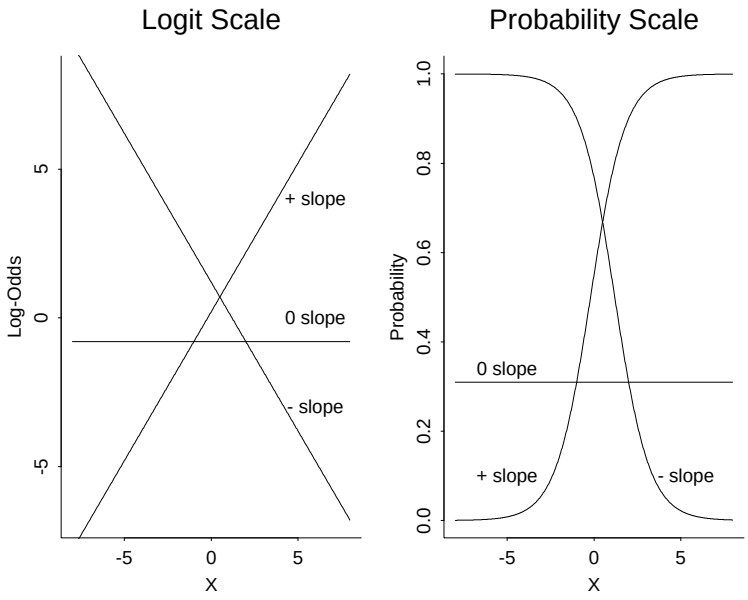


Figure 2: logit(p) and p as a function of X

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 above. 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:

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:

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

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

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

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:

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

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 $n_i\hat{p}_i$ 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* $r_i = (d_i - n_i\hat{p}_i)/\sqrt{n_i\hat{p}_i(1-\hat{p}_i)}$ 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:

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

or

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

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 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

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

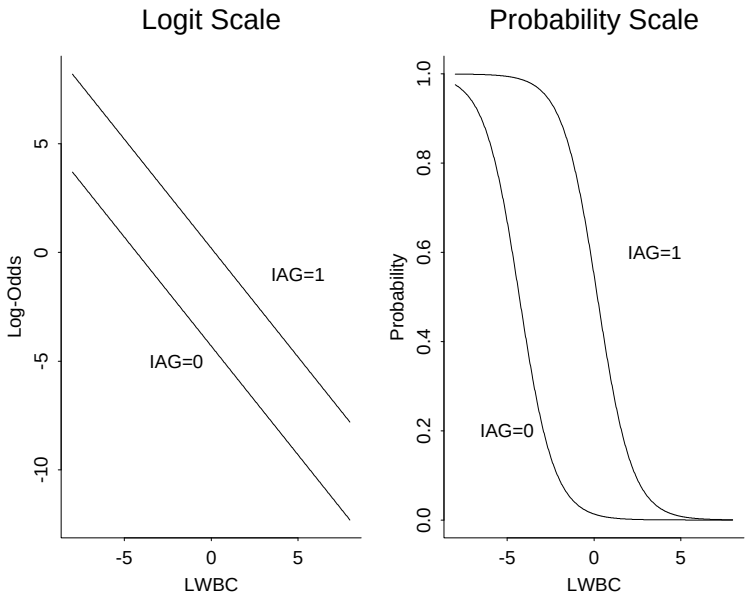
The model without IAG (i.e. $\beta_2 = 0$) is a simple logistic model where the log-odds of surviving one year is linearly related to LWBC, and is independent of AG. The reduced model with $\beta_2 = 0$ implies that there is no effect of the AG level on the survival probability *once LWBC has been taken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

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

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

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

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

or equivalently,

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

For AG+ individuals with IAG=1,

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

or

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

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.