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.

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

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.

A plot of the observed proportion of girls that have reached menarche (labeled proportion on page I of the JMP-IN output) shows that the proportion increases as age increases, but that the relationship is nonlinear. This is reinforced by the spline fit superimposed on the data plot. The plot and spline fit were done in the Fit Y-by-X platform.
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 model assumes that $p$ is related to $X$ through

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

(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 9 (or 9 to 1) when $p = .9$. 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 $\beta$.

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 is given in summarized or aggregate form:

\[
\begin{array}{ccc}
X & n & D \\
X_1 & n_1 & d_1 \\
X_2 & n_2 & d_2 \\
\vdots & \vdots & \vdots \\
X_m & n_m & d_m \\
\end{array}
\]

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. **A different format is used to enter the data in JMP-IN!**

**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)

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

over all possible values of \( \alpha \) and \( \beta \), where the \( p_i \)s 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 \( \alpha \) and \( \beta \) 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 \( \alpha \) and \( \beta \). The deviance evaluated at the MLEs:

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

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.

**Age at Menarche Data: JMP-IN 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 fitted in either the **FIT Y by X** platform or the **FIT MODEL** platform. The **FIT MODEL** platform must be used when the model has 2 or more effects. As with simple linear regression, the **FIT Y by X** platform plots the fitted model. Summaries from the two platforms are given on pages II and III of the JMP-IN output.

Regardless of which platform you choose, the data is entered in a similar form to a two-way contingency table of counts. The spreadsheet for this analysis includes 3 columns: AGE (the predictor - continuous with 25 levels), RM (the response variable, which should be defined as nominal with 2 coded levels to identify those that reached menarche and those that did not), and FREQ (a frequency variable that specifies the number of cases for each combination of AGE and RM). I defined RM as 0 for girls that reached menarche and 1 for those that did not. The actual numerical codes are unimportant, provided you remember that in JMP-IN
the success group corresponds to the LOWER of the two levels of the response variable RM. The JMP-IN output provides a subset of the spreadsheet to illustrate the data format.

The **Fit Model** output is more detailed, so let us start there. The form of the output is consistent with the output for multiple regression (last lecture!). The **Parameter Estimates** table gives 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

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

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 in the **Parameter Estimates** table is less than .0001, 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 **Lack of Fit** table gives the deviance chi-square statistic as as 26.70 on 23 df, with a p-value of .268. The large p-value suggests no gross deficiencies with the logistic model.

The **Whole Model Test** table gives the logistic regression analog of the Whole Model ANOVA table. 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 in the **Effect Test** table are also testing \( H_0 : \beta = 0 \).

The **Fit Y-by-X** platform only provides the **Whole Model Test** table, the **Parameter Estimates** table, and a plot of the fitted probabilities. The dots on the plot are supposed to be the data, but I do not know how to interpret the scatter.

**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.

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, so this model is fitted in the **FIT MODEL** platform. The parameters \( \alpha, \beta_1 \) and \( \beta_2 \) 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

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

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

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

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: \( \alpha \) and \( \alpha + \beta_2 \) are intercepts for the population logistic regression lines for AG- and AG+, respectively. The lines have a common slope, \( \beta_1 \). The \( \beta_2 \) 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 JMP-IN data sheet contains raw data for individual cases. There are four columns: the binary or indicator variable IAG (treated as continuous, with value 1 for AG+, 0 for AG-), WBC (continuous), LIVE (nominal, with value 0 if the patient lived at least 1 year and 1 if not), and Log WBC (natural log of WBC). Note that a frequency column is not needed with raw data 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 LWBC 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 ≥ 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 approximation to the deviance is suspect in sparse data settings. With small group sizes as we have here, most researchers would not interpret the p-values for the deviance 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.

The large p-value (.684) for the lack-of-fit chi-square (i.e. the deviance) 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 p-value for this test is .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).

The estimated survival probabilities satisfy

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

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

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

or equivalently,

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

For AG+ individuals with IAG=1,

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

or

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

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:

$$(8.06 - 1.11LWBC) - (5.54 - 1.11LWBC) = 2.52.$$ 

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.