Logistic Regression

• Logistic regression is an example of a large class of regression models called *generalized linear models (GLM)*

An Observational Case Study: The Donner Party (Gayson, D.K., 1990, "Donner Party deaths: A demographic assessment," *Journal of Anthropological Research*, **46**, 223-42, and Ramsey, F.L. and Schafer, D.W., 2002, *The Statistical Sleuth, 2nd Ed*, Duxbury Press, p. 580.

• In 1846, the Donner and Reed families left Illinois for California by covered wagon (87 people, 20 wagons). They attempted a new and untried crossing of the region between Ft. Bridger, Wyoming and the Sacramento Valley. After numerous problems and delays in Utah, they reached the Eastern Serria Nevada in late October. They were stranded near Lake Tahoe by a series of snowstorms that left as much as 8 feet of snow by some accounts. By the time they were rescued in April of the following year, 40 members had died. Some (or perhaps all) of those that survived did so by resorting to cannibalism

• The researchers attempted to address questions such as whether females are better able to withstand harsh conditions than men, and whether the odds of survival varied with age. Grayson was able to reconstruct records on survival, age and gender for 45 individuals.

Summary of Findings The odds of survival of females were estimated to be 4.9 times the odds of survival for men of the same age. An approximate 95% confidence interval for this odds ratio is 1.1 to 21.6

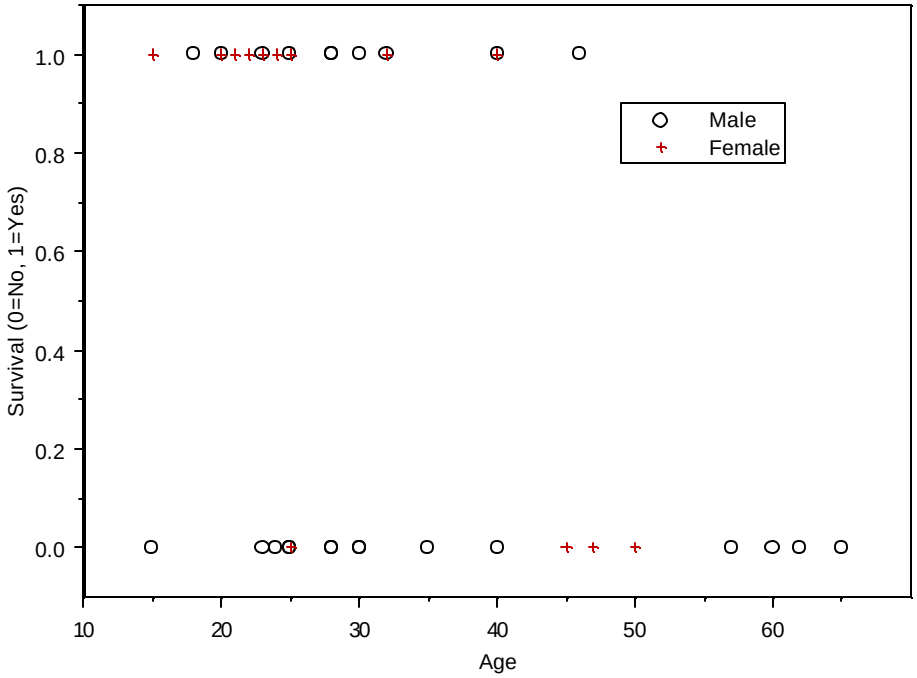
• We call $\pi/(1-\pi)$ the odds of the event of interest. Here, π is the probability of the event of interest (e.g., survival), so $\pi = P(S)$. The complement of S isdenoted by $\overline{S}$. If $\pi = 0.5,$ the odds of S (relative to $\overline{S}$) is 1

• If $\pi = 0.75,$ the odds of S (relative to $\overline{S}$) is $0.75/0.25 = 3$, or 3 to 1

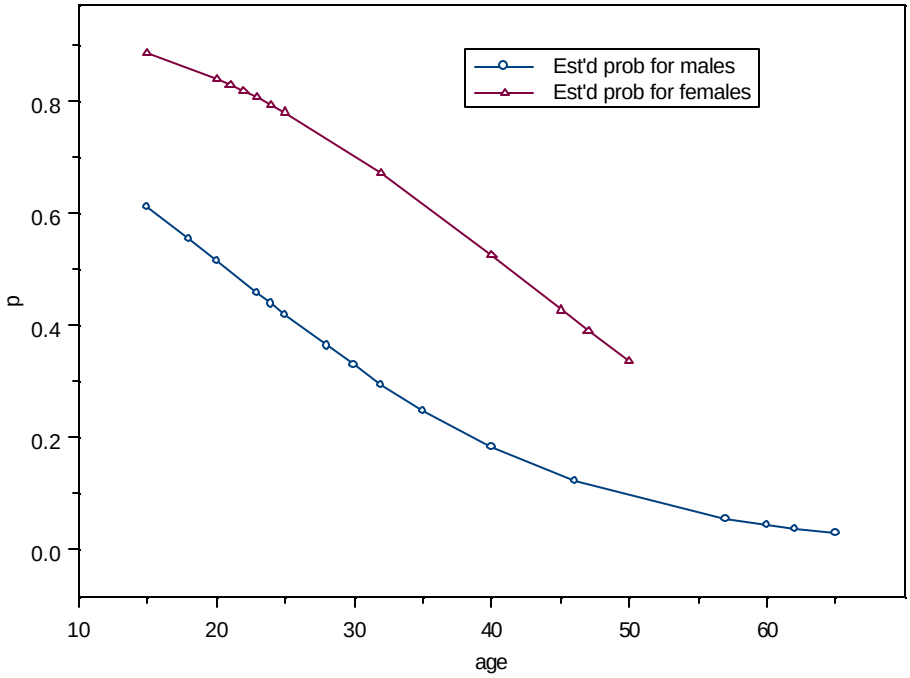
• If $\pi = 0.25,$ the odds of S (relative to $\overline{S}$) is $0.25/0.75 = 0.33$, or 1 in 3

• The logistic regression model is of log(odds of S), though it is easy to recover the estimated odds of S, and the probability of S, from the fitted model

• The figure below is a plot of outcome (survival or not) against age, by gender



• The Figure below shows the fitted logistic regression model. Here the estimated probability of survival is plotted against age for males and females



Scope of Inference Because the data are observational, the result cannot be used to infer that women are more apt to survive than men. There may be unquantified confounding variables that account for the apparent difference (e..g., behavior). Moreover, these 45 individuals are not a randomly sampled from any identifiable population to which inference may be made

A Retrospective Case Study (see Ramsey and Schafer, The Statistical Sleuth, 2nd ed.) Holst, P.A., Kromhout, D. and Brand, R. 1988. "For debate: pet birds as an independent risk factor for lung cancer", *British Medical Journal*, **297**, 13-21. A health survey in The Hague, Netherlands presented evidence of an association between keeping pet birds and increased risk of lung cancer. To investigate this link further, researchers conducted a *case-control* study of patients at four hospitals in The Hague. They identified 49 cases of lung cancer among patients that were younger than 65 and long-term residents of the city. The also selected 98 controls from the population of city residents having the same general age structure as the cancer cases. Data were gathered on the following variables:

FM: Sex ($1 =$ F, $0 =$ M)
AG: Age in years
SS: Socioeconomic status ($1 =$ High, $0 =$ Low), determined by occupation of the household's principal wage earner
YR: Years of smoking prior to diagnosis or examination
CD: Average rate of smoking (cigarettes/day)
BK: Indicator of birdkeeping. Birdkeeping was defined as keeping caged birds in the home for more than 6 consecutive months from 5 to 14 years before diagnosis (cases) or examination (controls)

Summary of Statistical Findings: The odds of lung cancer among bird keepers are estimated to be 3.8 times as large as the odds of lung cancer among nonbirdkeepers, after accounting for the effects of smoking, sex, age, and socioeconomic status. An approximate 95% confidence interval for true ratio is 1.7 to 8.5. The data present convincing evidence of increased risk associated with birdkeeping even after accounting for the effect of smoking (p-value $= 0.0008$).

Scope of Inference: Inference extends to the population of lung cancer patients and unaffected individuals in The Hague in 1985 (the study year). Statistical analysis of these observational data cannot be used as evidence that birdkeeping causes lung cancer. However, there is medical rationale supporting this conclusion: people who keep birds inhale and expectorate

excess allergens and dust particles which increases the likelihood of dysfunction of lung macrophages which in turn may lead to diminished immune system response.

• This is a retrospective study. A retrospective study is one in which two (sub)populations (cases and controls) are sampled at different rates. A comparison of the number of individuals in each sample provides *no* information regarding the probability that a randomly sampled birdkeeper has lung cancer. However, a comparison of the cases and controls provides a relative measure of increased risk (the actual level of risk is not addressed)

• In essence, if the proportion of lung cancer cases among the cases (birdkeepers) is twice that among the controls (not birdkeepers), then the data indicate that the odds of lung cancer is twice as great for birdkeepers compared to nonbirdkeepers

Generalized Linear Models (GLMs)

• A GLM is a probability model in which the mean, or expected value, of a response variable is related to a set of explanatory variables through a regression equation

• The usual multiple regression model is an example:

$$\mu = E(Y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

• To handle distributions besides the normal distribution, we must model a nonlinear function of μ. For example, if the response variable is Poisson in distribution, then the GLM is

$$\log \mu = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

• If the response variable is Binomial in distribution, then the GLM is

$$\log\left(\frac{\mu}{n-\mu}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

where n is the number of trials

• If the response variable is Binary(or Bernoulli) in distribution, then the GLM is

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

because $\mu = n\pi = \pi$ when the distribution is Bernoulli

• In general, there is some link function, say $g(\mu)$ that links the linear model to μ

• Provided that the link function is appropriate, nearly all of the ordinary multiple regression methods (residual analysis, hypothesis testing, confidence intervals) carry over to the GLM with only minor modifications

Logistic regression is used for modeling the mean of a binary, or Bernoulli random variable.

• A Bernoulli random variable is a Binomial random variable for which the number of trials is $n = 1$.

• Only one of two outcomes (S or $\overline{S}$) is possible, and the probability of S is denoted by $P(S) = \pi$. Consequently, $P(\overline{S}) = 1 - P(S) = 1 - \pi$.

• The mean of a Binomial random variable is $\mu = n\pi,$ and the variance is $\sigma^2 = n\pi(1-\pi)$

• Thus, for a binary random variable Y defined as

$$Y = \begin{cases} 1, & \text{if the outcome is } S \\ 0, & \text{if the outcome is } \overline{S}, \end{cases}$$

$$E(Y) = \mu = \pi$$

and

$$\text{Var}(Y) = \pi(1-\pi)$$

The **logit function** is the link between μ and the linear model. Let η denote the linear portion of the model, i.e.,

$$\eta = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1}$$

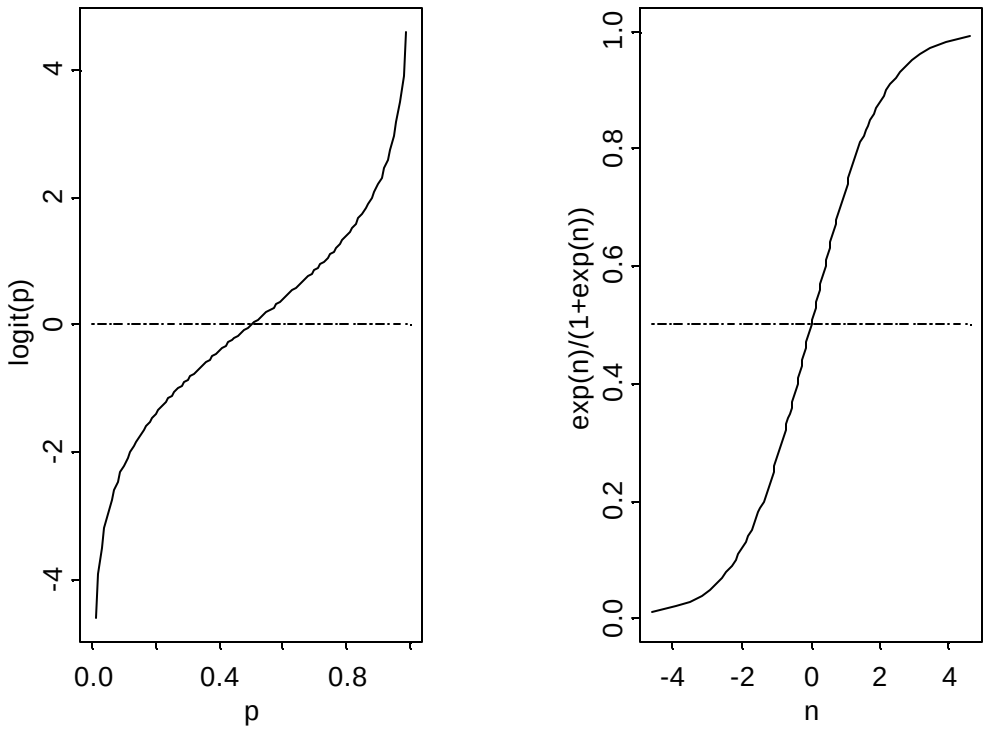
• The logit function of π is

$$\text{logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right)$$

- Thus, for logistic regression $\text{logit}(\pi) = \eta$
- Recall that $\pi/(1-\pi)$ is the the odds of S. If $\pi = 0.5$, the odds of S (relative to $\overline{S}$) is 1
- The logit function has the effect of stretching the possible values of π from 0 to 1, to $-\infty$ to ∞. This is very helpful with respect to the computational aspects of model fitting
- The inverse of the logit is important. We calculate it as follows

$$\begin{aligned}
\eta &= \log\Big(\frac{\pi}{1-\pi}\Big) \\
\Rightarrow e^{\eta} &= \frac{\pi}{1-\pi} \\
\Rightarrow \frac{1}{e^{\eta}} &= \frac{1-\pi}{\pi} = \frac{1}{\pi} - 1 \\
\Rightarrow 1 + \frac{1}{e^{\eta}} &= \frac{1}{\pi} \\
\Rightarrow \frac{1+e^{\eta}}{e^{\eta}} &= \frac{1}{\pi} \\
\Rightarrow \frac{e^{\eta}}{1+e^{\eta}} &= \pi
\end{aligned}$$

• The logit (as a function of π) is shown in the left panel, and the inverse function is shown in the right panel



• A useful interpretation of β_1 is obtained by considering the odds ratio. The odds ratio expresses how much more (or less) likely event A is to occur than event B.

• Suppose that the probability of A is $\pi_A = P(A)$ and that the probability of B is $\pi_B = P(B)$

• The odds that A will occur is $\pi_A/(1 - \pi_A)$ and the odds of B is $\pi_B/(1 - \pi_B)$. The odds ratio is

$$\frac{\pi_A/(1 - \pi_A)}{\pi_B/(1 - \pi_B)} = \frac{\pi_A(1 - \pi_B)}{\pi_B(1 - \pi_A)}$$

• For example, if $P(A) = \pi_A = 0.75$, and $P(B) = \pi_B = 0.25$, then the odds of A relative to B is

$$\frac{\pi_A(1 - \pi_B)}{\pi_B(1 - \pi_A)} = \frac{0.75(1 - .25)}{0.25(1 - .75)} = 3 \times 3 = 9$$

• It is 9 times more likely that A will occur than B. We can also say that it is 9 times less likely that B will occur than A

• For example, if $P(A) = \pi_A = 0.75$, and $P(B) = \pi_B = 0.6$, then the odds of A relative to B is

$$\frac{\pi_A(1-\pi_B)}{\pi_B(1-\pi_A)} = \frac{0.75(1-.60)}{0.60(1-.75)} = \frac{5}{4} \times \frac{8}{5} = 2$$

• Said another way, the odds of A is $3:1$, and the odds of B is $0.6/0.4 = 1.5$, or $3:2$, hence, the odds of A is twice that of B

Interpretation of logistic regression coefficients

• Consider the logit model of π as a function of a single explanatory variable x

$$\log\left(\frac{\pi}{1-\pi}\right) = \eta = \beta_0 + \beta_1 x$$

• Let π_A denote the probability when $x = A$:

$$\frac{\pi_A}{1-\pi_A} = e^{\beta_0+\beta_1 x_A}$$

• Let π_B denote the probability when $x = B$:

$$\frac{\pi_B}{1-\pi_B} = e^{\beta_0+\beta_1 x_B}$$

• The ratio of the odds of success (i.e., π) when $x = A$ relative to $x = B$ is

$$\begin{aligned}\frac{\pi_A/(1-\pi_A)}{\pi_B/(1-\pi_B)} &= \frac{e^{\beta_0+\beta_1 x_A}}{e^{\beta_0+\beta_1 x_B}} \\ &= e^{\beta_0+\beta_1 x_A-(\beta_0+\beta_1 x_B)} \\ &= e^{\beta_1(x_A-x_B)}\end{aligned}$$

• This means that if x_A differs from x_B by one unit, i.e., $x_A - x_B = 1$, then

$$e^{\beta_1(x_A-x_B)} = e^{\beta_1},$$

and the odds of success will change by a multiplicative factor of e^{β_1}

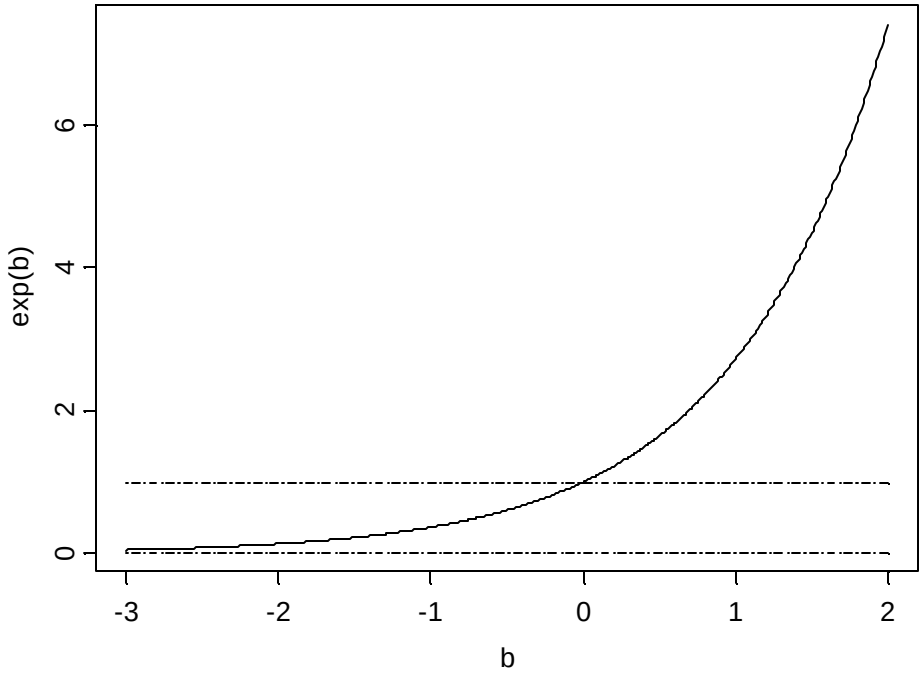
• For example, suppose that $\beta_1 = 0.5$; then $e^{\beta_1} = e^{0.5} = 1.649$. So, a one unit change in x will increase the odds of success by 1.65 times

• For example, suppose that $\beta_1 = -0.5$; then

$$e^{\beta_1} = e^{-0.5} = 1/1.649 = 0.6065.$$

Thus, a one unit change in x will decrease the odds of success by a (multiplicative) factor of 0.6065. The odds of success are $100 - 60.6 = 39.4\%$ less as a result of a 1 unit change in x

• The figure below plots β on the x-axis and e^{β} on the y-axis, and shows how a one-unit change in x will change the odds of success



• Suppose that all variables $x_2, \ldots, x_{p-1}$ are held constant, and we compare the effect of a 1-unit change in x_1 on the odds of S. The change is best measured with respect to the odds ratio.

• Let π_1 denote the probability of S for some value, say x_1, and π_2 denote the probability of S at $x_1 + 1$. The odds ratio is

$$\left(\frac{\pi_1}{1-\pi_1}\right)\Big/\left(\frac{\pi_2}{1-\pi_2}\right) = \frac{e^{\beta_0+\beta_1(x_1+1)+\cdots+\beta_{p-1}x_{p-1}}}{e^{\beta_0+\beta_1 x_1+\cdots+\beta_{p-1}x_{p-1}}}$$
$$= \frac{e^{\beta_1(x_1+1)}}{e^{\beta_1 x_1}} = e^{\beta_1}$$

• We say that if x_1 changes by 1-unit, then the odds of S will change by a multiplicative factor of e^{β_1} provided that all other variables are held constant

• More generally, if x_1 changes from A to B, then the odds of S will change by a multiplicative factor of $e^{\beta_1(A-B)}$ provided that all other variables are held constant because

$$\left(\frac{\pi_A}{1-\pi_A}\right)\Big/\left(\frac{\pi_B}{1-\pi_B}\right) = \frac{e^{\beta_0+\beta_1 A+\cdots+\beta_{p-1}x_{p-1}}}{e^{\beta_0+\beta_1 B+\cdots+\beta_{p-1}x_{p-1}}}$$
$$= \frac{e^{\beta_1 A}}{e^{\beta_1 B}} = e^{\beta_1(A-B)}$$

• Logistic regression produces a fitted model for the Donner party as follows. The model is of log-odds of $\pi = P(\text{Death})$, and it is

$$\text{logit}(\widehat{\pi}) = -3.23+0.078x^{AGE}-1.597x^{GENDER},$$

where

$$x^{GENDER} = \begin{cases} 1, & \text{if Female} \\ 0, & \text{if Male} \end{cases}$$

• The odds of death increase by a multiplicative factor of $e^{0.078} = 1.081$ for every year of life if gender is held constant. In other words, the risk increases by 8% for every year of life. An approximate 95% CI for this factor is 1.005 to 1.163

• The odds of death of females are estimated to be $e^{-1.597} = 0.2025$ times the odds of survival for males of the same age. An approximate 95% confidence interval for this odds ratio is 0.046 to 0.89

• We also can say that the odds of death of males are estimated to be

$$\frac{1}{e^{-1.597}} = e^{1.597} = 4.9$$

times the odds of death for females of the same age. An approximate 95% confidence interval for this odds ratio is $1/0.89 = 1.124$ to $1/0.046 = 21.71$

Estimation of Logistic Regression Coefficients

• In contrast to ordinary multiple regression, the regression parameters are estimated by maximizing the *likelihood* of the data (instead of minimizing the sum of the residuals)

• Logistic regression parameter estimates are *maximum likelihood* estimates

• The likelihood of the data, in the case of discrete data[1], is the probability of obtaining the sample as a function of the *parameters*

• By maximizing the likelihood through the choice of the parameter estimates, we make the observed sample more likely to have been the outcome of sampling any other possible sample of the same size

• An example involving 1 parameter is a random sample of $n = 20$ Bernoulli observations. Suppose the sample is $\boldsymbol{y} = (0,0,1,0,0,0,0,0,1,0,0,0,0,0,1,0,0,0,0,1)$, and that each observation is a single outcome from the Bin$(1,\pi)$ distribution. For example,

$$P(Y_1 = y_1) = \pi^{y_1}(1-\pi)^{1-y_1}, \text{ for } y_1 = 0 \text{ or } 1$$

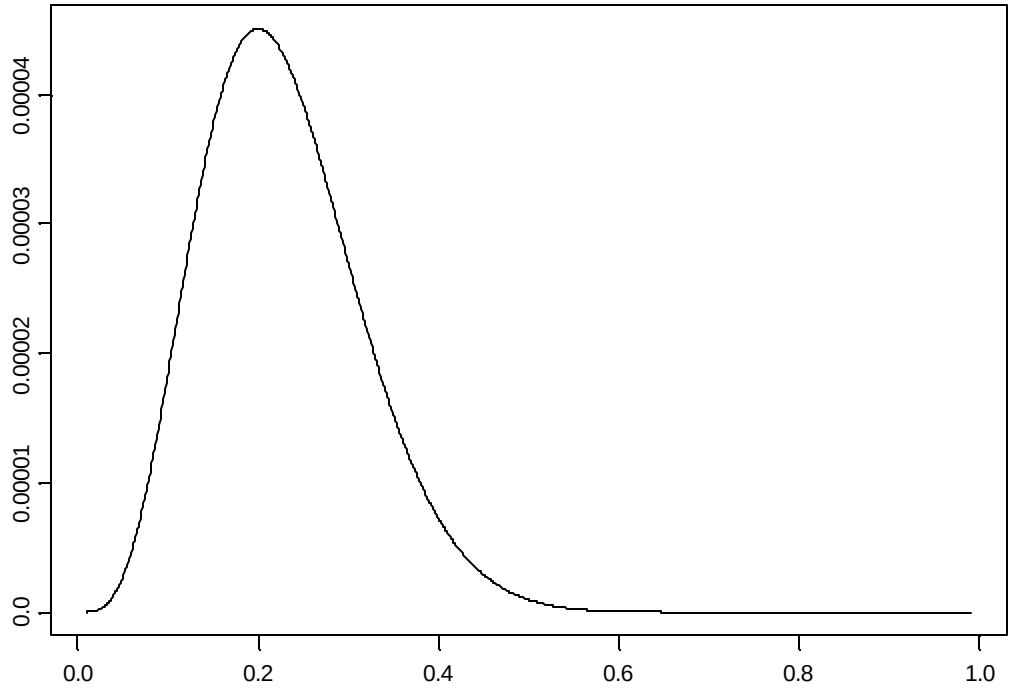
• Because the sample is random, the observations are independent, and the likelihood $L(\pi;\boldsymbol{y})$ of observing this sample is the probability of each outcome multiplied to together, that is

$$\begin{aligned} L(\pi;\boldsymbol{y}) &= P(Y_1 = 0, Y_2 = 0, \ldots, Y_{20} = 1) \\ &= P(Y_1 = 0) \times P(Y_2 = 0) \times \cdots \times P(Y_{20} = 1) \\ &= \pi^0(1-\pi)^1 \times \pi^0(1-\pi)^1 \times \cdots \times \pi^1(1-\pi)^0 \\ &= \pi^4(1-\pi)^{16} \end{aligned}$$

because there are four 1's and sixteen 0's in the sample $\boldsymbol{y}$

• The following graph of $L(\pi;\boldsymbol{y})$ versus π shows how the likelihood of the sample varies with π:

[1] In the case of continuous data, the definition is somewhat more complicted



• There are $2 \times \cdots \times 2 = 2^{20} = 1{,}048{,}576$ possible samples, so the likelihood of any particular sample is very small. In any case, the maximum of the likelihood occurs at $\pi = 0.2$ and so the maximum likelihood estimate (MLE) is $\widehat{\pi} = 0.2$

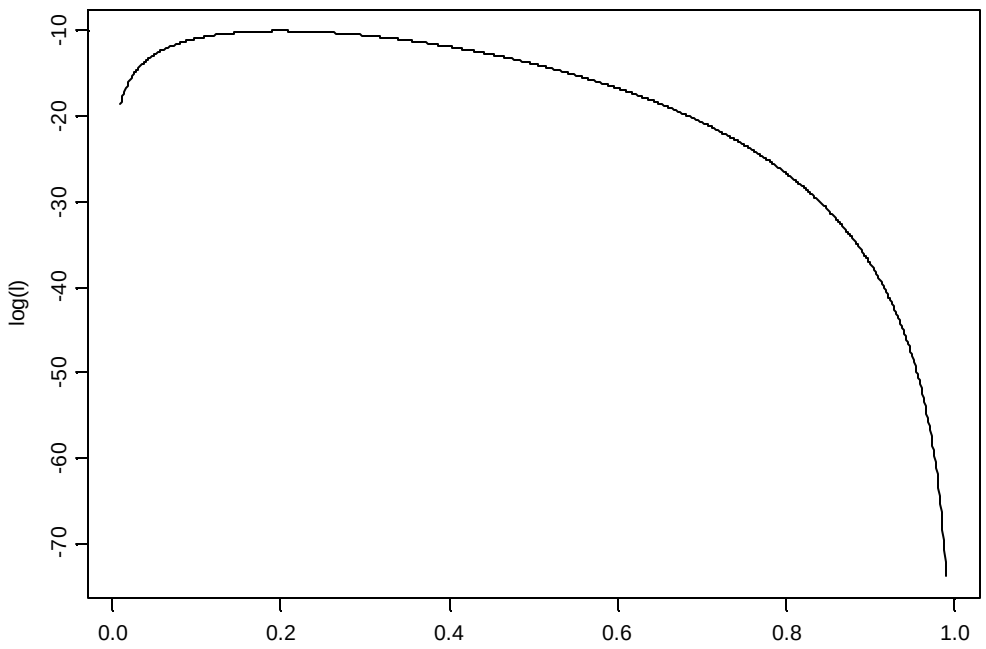
• In this example, the MLE is the same as our previous estimator of π, namely, the sample proportion

$$p = \frac{\text{\# successes}}{\text{\# trials}} = \frac{4}{20}$$

• It is easier to work with the log-likelihood, i.e., $\log L(\pi;\boldsymbol{y})$, instead of the likelihood because the logarithm converts products to sums. For example,

$$2.996 = \log(20) = \log(4 \times 5\,) = \log(4) + \log(5) = 1.386 + 1.610$$

• Also, the value of π that maximizes $\log L(\pi;\boldsymbol{y})$ also maximizes $L(\pi;\boldsymbol{y})$. For example, a plot of $\log L(\pi;\boldsymbol{y})$ versus π reveals the same maximizing value of $\pi = 0.2$:



• For the logistic regression model, the probability of survival for the ith individual is a function of the parameters β_0, β_1 and β_2. The log-likelihood is obtained in two steps. First,

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_i^{AGE}+\beta_2 x_i^{GENDER}}}{1+e^{\beta_0+\beta_1 x_i^{AGE}+\beta_1 x_i^{GENDER}}} = \frac{e^{\eta_i}}{1+e^{\eta_i}}$$

and

$$1-\pi_i = \frac{1}{1+e^{\beta_0+\beta_1 x_i^{AGE}+\beta_2 x_i^{GENDER}}} = \frac{1}{1+e^{\eta_i}}$$

• Secondly,

$$\begin{aligned}
\log[P(Y_i = y_i)] &= \log[\pi_i^{y_i}(1-\pi_i)^{1-y_i}] \\
&= \log \pi_i^{y_i} + \log(1-\pi_i)^{(1-y_i)} \\
&= y_i \log \pi_i + (1-y_i)\log(1-\pi_i) \\
&= y_i \log\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right) + (1-y_i)\log\left(\frac{1}{1+e^{\eta_i}}\right)
\end{aligned}$$

Then,

$$\log[L(\beta_0, \beta_1, \beta_2; \boldsymbol{y})] = \sum_{i=1}^{45} \log[P(Y_i = y_i)]$$
$$= \sum_{i=1}^{45} y_i \log\left(\frac{e^{\eta_i}}{1+e^{\eta_i}}\right) + (1 - y_i)\log\left(\frac{1}{1+e^{\eta_i}}\right)$$

• The final substitution replaces η_i by $\beta_0 + \beta_1 x_i^{AGE} + \beta_2 x_i^{GENDER}$. This is too complicated to be informative; in fact, a formula for the maximum likelihood estimates in logistic regression cannot be expressed in closed-form. Instead, numerical optimization methods are used to compute the estimates

• For the Donner Party data, the following table[2] shows $\log[L(\beta_0, \beta_1, \beta_2; \boldsymbol{y})]$ for a few of the possible values of β_0, β_1, and β_2:

β_0	β_1	β_2	$\log[L(\beta_0, \beta_1, \beta_2; \boldsymbol{y})]$
1.50	−0.50	1.25	−27.7083
		1.80	−28.7931
	−0.80	1.25	−26.1531
		1.80	−25.7272
1.70	−0.50	1.25	−29.0467
		1.80	−30.3692
	−0.80	1.25	−25.7972
		1.80	−25.6904
1.63	−0.078	1.60	−25.6282

• The last line in the table shows the value of $\log[L(\beta_0, \beta_1, \beta_2; \boldsymbol{y})]$ at the MLE's. These MLEs are different than those above because Ramsey and Schafer set $\pi = P(\text{Survival})$ whereas I set $\pi = P(\text{Death})$

Properties of MLEs

If the model is correct and the sample size is large, then

• MLE's are nearly unbiased

• The standard errors of MLEs can be estimated, and the estimates are nearly unbiased

• MLE's are more precise than nearly all other estimators

[2] From Ramsey, F.L. and Schafer, D.W. 2002. The Statistical Sleuth, 2nd Ed. p. 589.

• The distribution of an MLE is approximately normal (a consequence of the Central Limit Theorem)

Tests and Confidence Intervals Involving A Single Parameter

• The last property above implies that if $\widehat{\beta}_i$ is the MLE of β_i, then

$$\widehat{\beta}_i \dot{\sim} N(\beta_i, \sigma_{\widehat{\beta}_i})$$

where $\sigma_{\widehat{\beta}_i}$ is the standard error of the $\widehat{\beta}_i$

• This gives a approximate test of

$$H_0 : \beta_i = 0 \text{ versus } H_1 : \beta_i \neq 0$$

• The test statistic is

$$Z = \frac{\widehat{\beta}_i}{\widehat{\sigma}_{\widehat{\beta}_i}}.$$

• If H_0 is true, then $Z \dot{\sim} N(0,1)$.

• Suppose that the observed value of the test statistic is z; then, the p-value is $2P(Z \geq |z|)$.

• For example, in the Donner Party data analysis, the linear portion of a logistic regression model of $P(\text{Death})$ is

$$\eta_i = \beta_0 + \beta_1 x_i^{AGE} + \beta_2 x_i^{GENDER} + \beta_3 x_i^{AGE \times GENDER},$$

• SPSS reports

$$\widehat{\beta}_3 = 0.162, \widehat{\sigma}_{\widehat{\beta}_3} = 0.092, \text{ and Sig.} = 0.086$$

• The p-value is approximately 0.086 because

$$Z = \frac{\widehat{\beta}_i}{\widehat{\sigma}_{\widehat{\beta}_i}} = \frac{0.162}{0.092} = 1.73 \text{ and } 2P(Z \geq 1.73) = 0.086$$

(ignoring rounding error)

• Also, SPSS reports that $e^{\widehat{\beta}_3} = e^{0.162} = 1.175$. This means that the difference in odds of death resulting from a 1 year increase in age is greater

for females ($GENDER = 1$) than males by a factor of 1.175 given that age is held constant. A approximate 95% CI for this factor is 0.977 to 1.141

• In my opinion, interaction between gender and age is not supported by the data, and I adopt the no-interaction model.

• Earlier, it was reported that the odds of death increase by a multiplicative factor of $e^{0.078} = 1.081$ for every year of life if gender is held constant. So, if I compare a female at age $= 20$ versus a female at age $= 30$, the odds of death are $1.081^{10} = 2.18$ times greater for the older female. An approximate 95% CI for this factor is

$$1.005^{10} = 1.05 \text{ to } 1.163^{10} = 4.56$$

The Drop-in-Deviance Test

• Testing the significance of a covariate or a factor in logistic regression is conducted using the same principle as in ordinary regression.

• Specifically, we compare the fit of two models: the full model which contains the covariate or the factor of interest, and a reduced model that is the same as the full except that the covariate or factor is not in the model

• If we are assessing the importance of a factor with k levels, so $k - 1$ indicator variables are used to account for it, then all $k - 1$ indicators are removed from the full model to get the reduced model

• Suppose that g variables are used to account for the term (if the term is a covariate $g = 1$), and the parameters that multiply the variables are $\beta_1, \beta_2, \ldots, \beta_g$ then the null hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_g = 0$$

and the alternative is

$$H_1 : \text{at least one of } \beta_1, \beta_2, \ldots, \beta_g \text{ is not } 0.$$

• The test statistic is called the likelihood ratio statistic because it is 2 times logarithm of the ratio of the two likelihoods

$$\begin{aligned} LRT &= 2\log\left\{ \frac{L(\text{full model}; \boldsymbol{y})}{L(\text{reduced model}; \boldsymbol{y})} \right\} \\ &= 2\log[L(\text{full model}; \boldsymbol{y})] - 2\log[L(\text{reduced model}; \boldsymbol{y})]. \end{aligned}$$

• If the H_0 is true, then the LRT is approximately chi-square in distribution and the degrees of freedom are g, i.e.,

$$LRT \dot{\sim} \chi^2_g$$

• The entire procedure is often called the drop-in-deviance test (analogous to the extra-sums-of-squares test which uses an F-statistic). The term deviance arises from the *deviance* of a model, say model M

$$D(\text{M}) = -2\log[L(\text{M}; \boldsymbol{y})]$$

• The larger the deviance, the worse the fit of the model. Like the error sums of squares in ordinary regression, we cannot interpret the deviance in isolation (without comparing it to some other model)

• The likelihood ratio statistic comparing two models can be expressed in terms of the deviance:

$$LRT = D(\text{reduced}) - D(\text{full})$$

• Example: Donner Party data. SPSS reports the following from a model containing both age and gender:

Likelihood Ratio Tests

Effect	-2 Log Likehood of of Reduced Model	Chi-Square	df	Sig
Intercept	38.754	2.408	1	.121
AGE	42.377	6.030	1	.014
GENDER	41.381	5.034	1	.025

The chi-square statistic is the difference in -2 log-likelihoods between the final model and a reduced model. The reduced model is formed by omitting an effect from the final model. The null hypothesis is that all parameters of that effect are 0.

• Note that SPSS does not use the term deviance, but instead `-2 Log Likehood`. It is the same quantity

• The test of significance for AGE compares the deviance of the full model ($D = 36.347$) to the deviance to the model without AGE but including GENDER ($D = 42.377$). The test statistic is

$$LRT = 42.377 - 36.347 = 6.030,$$

and $P(\chi^2_1 > 6.030) = 0.014$. Hence, there is strong evidence (p-

value ≈ 0.014) that there differences in the probability of survival at different ages

- **ACase-Control Study - Birdkeeping and Lung Cancer**

- The objective was to determine if there is an increased probability of lung cancer associated with birdkeeping, even after accounting for other factors (e.g., smoking)

- Factors (and covariates) are

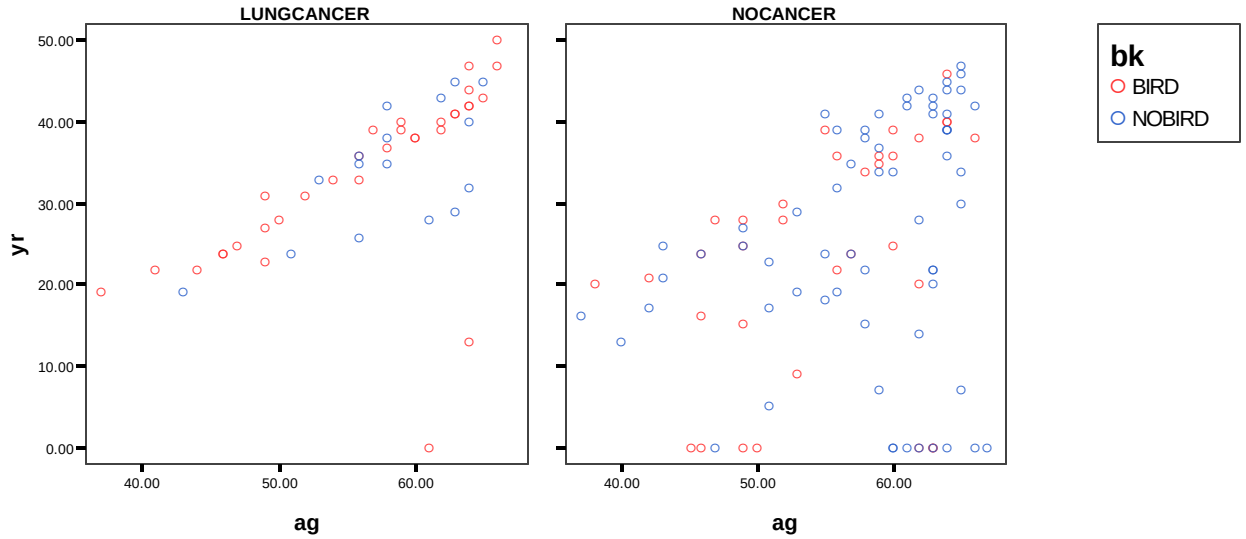
  FM: Sex ($1 =$ F, $0 =$ M)
  AG: Age in years
  SS: Socioeconomic status ($1 =$ High, $0 =$ Low), determined by occupation of the household's principal wage earner
  YR: Years of smoking prior to diagnosis or examination
  CD: Average rate of smoking (cigarettes/day)
  BK: Indicator of birdkeeping.

- The first step is to examine the effect of smoking on lung cancer. The following plot compares the relationship between age and smoking, and whether the individuals kept birds, for the cases and the controls

- The plot shows that smoking is strongly associated with lung cancer because of the cases, all but one has been smoking for more than 10 years

- Comparing cases (LUNGCANCER) to controls (NOCANCER) reveals that relatively more individuals were birdkeepers among the cases than the controls

- Birdkeeping does not appear to be associated with years of smoking or age



• This plot was obtained from the Graphs/Interactive/Scatterplot window. I set the y-axis and x-axis variables to be *YR* and *AG*, respectively. The legend variable was *bk* and the panel variable was *lc*.

• LR test results for main effects (besides birdkeeping) are shown below

Likelihood Ratio Tests

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	160.323	.000	0	.
SEX	162.539	2.216	1	.137
SS	160.351	.028	1	.868
AG	163.344	3.021	1	.082
YR	172.526	12.203	1	.000
CD	160.987	.664	1	.415

• Years of smoking has the largest observed effect on the odds of lung cancer, as measured by the LR test. The estimated effect of one additional year of smoking on the odds of cancer is 1.083 and an approximate 95% CI is 1.028 to 1.140, i.e., between 2.8% and 14.0%

• Removing SS yields

Effect	-2 Log Likelihood of Reduced Model	Chi-Square	df	Sig.
Intercept	160.351	.000	0	.
SEX	162.810	2.459	1	.117
AG	163.473	3.122	1	.077
YR	173.395	13.044	1	.000

```
CD              161.004          .653          1    .419
```

• Removing CD yields

```
          -2 Log Likelihood
Effect    of Reduced Model   Chi-Square     df   Sig.
Intercept      161.004          .000          0    .
SEX            163.283         2.280          1    .131
AG             164.866         3.863          1    .049
YR             181.583         20.580         1    .000
```

• I am inclined to retain SEX in the model. My next step is to attempt to add interaction terms among the 3 variables above. None were significant according to the drop-in-deviance tests

• Adding beekeeping and all possible 2-way interactions with the three variables, followed by removal of non-significant terms yields the following drop-in-deviance tests

```
          -2 Log Likelihood
Effect    of Reduced Model   Chi-Square     df   Sig.
SEX            152.058          .896          1    .344
AG             152.939         1.776          1    .183
YR             168.051         16.889         1    .000
BK             162.390         11.228         1    .001
```

• After eliminating SEX as a factor, the parameter estimates, estimates of the effect on odds of lung cancer, and 95% CI's are

				Approximate 95% confidence interval	
Variable	$\widehat{\beta}$	Std. Err.	$e^{\widehat{\beta}}$	Lower bound	Upper bound
Intercept	-1.034	1.661	$\cdot$	$\cdot$	$\cdot$
BK = BIRD	1.377	0.401	3.961	1.806	8.688
AGE	-0.046	0.034	0.955	0.893	1.021
YEAR	0.0748	0.023	1.078	1.030	1.127

• There is little evidence that AGE is significant ($\chi^2_1 = 1.897$, p-value ≈ 0.16) even after removing SEX

• The estimated effect of one additional year of smoking on the odds of cancer is 1.078 and an approximate 95% CI is 1.030 to 1.127, given that all other factors are held constant

• The estimated effect of birdkeeping on the odds of cancer is 3.961 and an approximate 95% CI is 1.806 to 8.688, given that all other factors are held constant

• **The Cervical Cancer Study.** A case-control study conducted under the auspices of the World Health Organization attempted to determine if the use of oral contraceptives affects the risk of invasive carcinoma of the cervix[3]. Data obtained from one of the hospitals involved in the study (Siriraj in Thailand) consist of 899 records. Of these 141, are cases, and the remaining 758 are controls. The recorded variables are

Caco: Case-control status: ($0 =$ control, $1 =$ case)

Age: Age group ($1 =$ 15-30, $2 =$ 31-35, $3 =$ 36-40, $4 =$ 41-45, $5 =$ 46-60)

Ocuse: Previous oral contraceptive use ($0 =$ no, $1 =$ yes)

Dur: Duration of contraceptive use in months ($0 =$ never, $1 =$ 1-12, $2 =$ 13-24, $3 =$ 25-60, $4 =$ 61 or more)

Sexrel: Number of sexual relations ($0 =$ none, $1 =$ one, $2 =$ more than one)

Agesex: Age at first sexual relationship ($0 =$ never, $1 =$ 8-16, $2 =$ 17-18, $3 =$ 19-20, $4 =$ 21-23, $5 =$ 24 or older)

Npaps: Number of Pap smears taken ($0 =$ none, $1 =$ 1-2, $2 =$ 3-4, $3 =$ 5-6, $4 =$ 7 or more)

Vagdis: Doctor visited for abnormal vaginal discharge ($0 =$ no, $1 =$ yes)

Tpreg:Total number of pregnancies ($0 =$ none, $1 =$ 1-2, $2 =$ 3-4, $3 =$ 5-6, $4 =$ 7 or more)

• These data pose several problems, ranging from analysis strategy to computational

• The first problem is that of adopting a data analysis strategy. To a large extent, the objective (which dictates the strategy) is fairly limited: determine if contraceptive use is associated with an increased risk of cervical cancer

• In principle, a good strategy to use is the one that was pursued in the sex discrimination analysis of wages, namely find a model that explains as much of the variation in wages as possible, ignoring gender, and then test whether

[3]From: Collett, D. 2003. Modeling Binary Data, 2nd Ed. Chapman and Hall/CRC.

gender significantly improves the fit of the model. Then, estimate the differences between gender after accounting for all significant variables

• The explanatory factors in this data set can be roughly classified as physical and behavioral (*Age*, *Tpreg*, *Agesex*) or symptomatic (*Npaps* and *Vagdis*). The symptomatic variables may *account* for variation in the odds of cancer, but they don't *explain* variation. This implies that it may be justified to use only the physical and behavioral variables in the analysis. Moreover, the symptomatic variables may mask the effect of contraceptive use. For example, if we are attempting to determine what factors are associated with HIV infection from among a list of physical and behavioral factors, we would not include white blood cell count because it is a symptomatic variable and may be such a good predictor of HIV status that all other interesting variables are not significant

• Computational problems occur when attempting to introduce interaction variables between some of the factors with relatively large numbers of levels. The problem occurs when, for some combination of levels (say *Agesex* $= 1$ and *Age* $= 1$, all observations on *Caco* are cases (or perhaps all are controls). This means that the estimated odds are either $1/0$ or $0/0$, and both situations lead to a parameter estimate that, by definition, should be ∞ or $-\infty$. (We say that the parameter estimate is undefined because ∞ is not a number). If this is the case, then SPSS (or any other standard statistical package) cannot compute reliable parameter estimates of any model parameters. The message warning of this problem is

Unexpected singularities in the Hessian matrix are encountered. There may be a quasi-complete separation in the data. Some parameter estimates will tend to infinity. The NOMREG procedure continues despite the above warning(s). Subsequent results shown are based on the last iteration. Validity of the model fit is uncertain.

• One solution to this problem is to represent factors that are ordered and have many levels (say, > 3) as two covariates. For example, *Age* has 5 ordered levels. It can be represented by two covariates: *Age* (use the age class labels, or the mid-points of the age classes) and the squared version of this covariate. Using these variables will provide a simple flexible model of the age effect, and avoid the problem of undefined parameter estimates

• For example, an initial rich model contains the following terms:

Main effects: *Agesex* (covariate), $Agesex^2$ (covariate), *Age* (covariate), Age^2 (covariate), *Tpreg* (covariate), $Tpreg^2$ (covariate), *Sexrel* (factor)

Interaction effects: $Agesex \times Age$, $Agesex \times Tpreg$, $Agesex \times Sexrel$, $Age \times Tpreg$, $Age \times Sexrel$, $Tpreg \times Sexrel$

- For simplicity, I have assumed that interaction is negligible between $Agesex^2$, Age^2 and $Tpreg^2$ and all other terms

- After removing nonsignificant terms, test for the significance of *Ocuse* and *Dur*, but not both simultaneously. Investigate interaction with other factors (provided that the model parameters can be computed by SPSS)

- Note: if $Agesex^2$ is significant, then retain both *Agesex* and $Agesex^2$, regardless of the significance of *Agesex* (for the same reason that we keep main the effect of a factor in a model if any interaction term involving the factor is significant)