Standard Errors & Confidence Intervals

\[ \beta - \bar{\beta} \overset{asy}{\sim} N(0, I(\bar{\beta})^{-1}), \]

where

\[ I(\bar{\beta}) = \left[ \frac{\partial^2 l(\beta, \phi; y)}{\partial \beta_i \partial \beta_j} \right]_{\beta = \bar{\beta}} \]

We can obtain asymptotic \(100(1 - \alpha)\)% confidence intervals for \(\beta_j\) using:

\[ \bar{\beta}_j \pm Z_{1-\alpha/2} se(\bar{\beta}_j) \equiv \bar{\beta}_j \pm 1.96 se(\bar{\beta}_j) \text{ for } \alpha = 0.05, \]

where \(Z_p\) denotes the \(p\)th percentile of the N(0,1) density.
Profile Likelihood Confidence Intervals

Express log-likelihood as $l(\alpha, \beta)$, where $\beta$ is the parameter of interest and $\alpha$ are nuisance parameters.

Maximum the log-likelihood with respect to $\alpha$ for a range of values of $\beta$ to give the profile log-likelihood function

$$l(\hat{\alpha}_\beta, \beta),$$

where $\hat{\alpha}_\beta$ is the maximum likelihood estimate of $\alpha$ for a given $\beta$.

A confidence region for $\beta$ is formed from those values of $\beta$ giving $l(\hat{\alpha}_\beta, \beta)$ sufficiently close to the overall maximum.

For a 95% interval, the allowable difference from the maximum is

$$\frac{1}{2} \times \chi^2_{1,0.95} = \frac{1}{2} \times 3.841 = 1.920.$$
Model Selection and Goodness of Fit

In most data analyses, there is uncertainty about the model & you need to do some form of model comparison.

1. Select $q$ out of $p$ predictors to form a parsimonious model
2. Select the link function (e.g., logit or probit)
3. Select the distributional form (normal, t-distributed)

We focus first on problem 1 - Variable Selection
Frequentist Strategies for Variable Selection

It is standard practice to sequentially add or drop variables from a model one at a time and examine the change in model fit.

If model fit does not improve much (or significantly at some arbitrary level) when adding a predictor, then that predictor is left out of the model (forward selection).

Similarly, if model fit does not decrease much when removing a predictor, then that predictor is removed (backwards elimination).

Implementation: *step() function in S-PLUS.*
Some Issues with Stepwise Procedures

Normal Linear & Orthogonal Case

• In normal models with orthogonal $\mathbf{X}$, forward and backwards selection will yield the same model (i.e., the selection process is not order-dependent).

• However, the selection of the significance level for inclusion in the model is arbitrary and can have a large impact on the final model selected. Potentially, one can use some goodness of fit criterion (e.g., AIC, BIC).

• In addition, if interest focus on inference on the $\beta$’s, stepwise procedures can result in biased estimates and invalid hypothesis tests (i.e., if one naively uses the final model selected without correction for the selection process)
Issues with Stepwise Procedures (General Case)

For GLMs other than orthogonal, linear Gaussian models, the order in which parameters are added or dropped from the model can have a large impact on the final model selected.

This is not good, since choices of the order of selection are typically (if not always) arbitrary.

Model selection is a challenging area, but there are alternatives
Goodness of Fit Criteria

There are a number of criteria that have been proposed for comparing models based on a measure of goodness of fit penalized by model complexity

1. Akaike’s Information Criterion (AIC):

\[ AIC_M = D_M + 2p_M, \]

where \( D_M \) is the deviance for model \( M \) and \( p_M \) is the number of predictors

2. Bayesian Information Criterion (BIC):

\[ BIC_M = D_M + p_M \log(n). \]
Note that deviance decreases as variables are added and the likelihood increases.

The AIC and BIC differ in the penalty for model complexity, with the AIC using twice the number of parameters and the BIC using the number of parameters multiplied by the logarithm of the sample size.

The BIC tends to place a larger penalty on the number of predictors and hence more parsimonious models are selected.
Probability of Selecting a Model

Posterior probability of selecting model \( j \), for \( j \in \{1, \ldots, J\} \):

\[
\Pr(M = j \mid y, X) = \frac{\exp(-BIC_j/2)}{\sum_{k=1}^{J} \exp(-BIC_k/2)},
\]
Latent Variable Models for Binary Data

Suppose that for a given vector of explanatory variables $\mathbf{x}$, the latent variable, $U$, has a continuous cumulative distribution function $F(u; \mathbf{x})$ and that the binary response $Y = 1$ is recorded if and only if $U > 0$:

$$\theta = \Pr(Y = 1 \mid \mathbf{x}) = 1 - F(0; \mathbf{x}).$$

Since $U$ is not directly observed there is no loss of generality in taking the critical (i.e., cutoff point) to be 0.

In addition, we can take the standard deviation of $U$ (or some other measure of dispersion) to be 1, without loss of generality.
Probit Models

For example, if \( U \sim N(x'\beta, 1) \) it follows that

\[
\theta_i = \Pr(Y = 1 \mid x_i) = \Phi(x_i'\beta),
\]

where \( \Phi(\cdot) \) is the cumulative normal distribution function

\[
\Phi(t) = (2\pi)^{-1/2} \int_{-\infty}^{t} \exp\left(-\frac{1}{2}z^2\right) dz.
\]

The relation is linearized by the inverse normal transformation

\[
\Phi^{-1}(\theta) = x_i'\beta = \sum_{j=1}^{p} x_{ij}\beta_j.
\]
We have regarded the cutoff value of $U$ as fixed and the mean of $U$ to be changing with $x$.

Alternatively, one could assume that the distribution of $U$ is fixed and allow the critical value to vary with $x$ (e.g., dose).

In toxicology studies where dose is the explanatory variable it makes sense to let $V$ denote the minimum level of dose needed to produce a response (i.e., tolerance).
Under the second formulation, $y_i = 1$ if $x_i \beta > v_i$

It follows that

$$
Pr(Y = 1 \mid x_i) = Pr(V \leq x'_i \beta).
$$

Note that the shape of the dose-response curve is determined by the distribution function of $V$

If $V \sim N(0, 1)$, then

$$
Pr(Y = 1 \mid x_i) = \Phi(x'_i \beta),
$$

and it follows that the $U$ and $V$ formulations are equivalent

The $U$ formulation is more common
Suppose that Fred is choosing between 2 brands of a product (say, Ben & Jerry’s or Haigen Daz)

Fred has a utility for Ben & Jerry’s (denoted by $Z_{i1}$) and a utility for Haigen Daz (denotes by $Z_{i2}$)

Letting the difference in utilities be represented by the normal linear model, we have

$$U_i = Z_{i1} - Z_{i2} = x_i' \beta + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$.

If Fred has a higher utility for Ben & Jerry’s, then $Z_{i1} > Z_{i2}, U_i > 0$, and Fred will choose Ben & Jerry’s ($Y_i = 1$)
This latent utility formulation is again equivalent to a probit model for the binary response.

The generalization to a multinomial response is straightforward by introducing $k$ latent utilities instead of 2, and letting an individual’s response (i.e., choice) correspond to the category with the maximum utility.

Although the probit model is preferred in bioassay and social sciences applications, the logistic model is preferred in the biomedical sciences.

Of course, the choice of distribution function for $U$ (and hence the choice of link in the binary response GLM) should be motivated by model fit.
Logistic Regression

The normal form is only one possibility for the distribution of $U$.

Another is the logistic distribution with location $x_i'\beta$ and unit scale.

The logistic distribution has cumulative distribution function

$$F(u) = \frac{\exp(u - x_i'\beta)}{1 + \exp(u - x_i'\beta)},$$

so that

$$F(0; x_i) = 1/\{1 + \exp(x_i'\beta)\},$$

It follows that

$$\Pr(Y = 1 \mid x_i) = \Pr(U > 0 \mid x_i) = 1 - F(0; x_i) = 1/\{1+\exp(-x_i'\beta)\}.$$

To linearize this relation, we take the logit transformation of both sides,

$$\log\{\theta_i/(1 - \theta_i)\} = x_i'\beta.$$
Homework Exercise: For \(x_i = (1, x)'\) and \(\beta_2 > 0\), reformulate the logistic regression in terms of a threshold model (i.e., the \(V\) formulation of the probit model described above). Derive the probability density function (pdf) obtained by differentiating \(\Pr(Y = 1 \mid x_i)\) with respect to \(x\). Reparameterize in terms of \(\tau = 1/\beta_2\) and \(\mu = -\beta_1/\beta_2\). Plot this pdf for \(\mu = 0\) and \(\pi \tau / \sqrt{3} = 1\) along with the \(N(0,1)\) pdf in S-PLUS. Which density has the fatter tails? Is the pdf for \(x\) in the logistic case in the exponential family?
Some Generalizations of the Logistic Model

The logistic regression model assumes a restricted dose-response shape and it is possible generalize the model to relax the restriction

Aranda-Ordaz (1981) proposed two families of linearizing transformation, which can easily be inverted and which span a range of forms. The first, which is restricted to symmetric cases (i.e., invariant to interchanging success & failure) is

\[
\frac{2 \theta^{\nu} - (1 - \theta)^{\nu}}{\nu \theta^{\nu} + (1 - \theta)^{\nu}}.
\]

In the limit as \( \nu \to 0 \), this is logistic and for \( \nu = 1 \) this is linear

The second family has

\[
\log\{(1 - \theta)^{-\nu} - 1\}/\nu, \]

which reduces to the extreme value model when \( \nu = 0 \) and the logistic when \( \nu = 1 \).
When there is doubt about the transformation, a formal approach is to use one or the other of the above transformations and to fit the resulting model for a range of possible values for $\nu$.

A profile likelihood can be obtained for $\nu$ by plotting the maximized likelihood against $\nu$.

Potentially, one could choose a standard form, such as the logistic, if the corresponding value of $\nu$ falls within the 95% profile likelihood confidence region.
Another possibility for generalizing the logistic regression model is to allow the effects of one or more of the predictors to be non-linear

\[
\text{logitPr}(Y = 1 \mid x_i) = \alpha + s(x_i)\beta,
\]

where \(s(\cdot)\) is an unknown function

This type of model is referred to as a generalized additive model and we will discuss this in more detail next class - motivating the approach using data from Longnecker et al. (2001).