

6.5: Power and sample size*

Recall: $\alpha = P(\text{reject } H_0 | H_0 \text{ true})$ and $\beta = P(\text{accept } H_0 | H_1 \text{ true})$.

Power is $1 - \beta = P(\text{reject } H_0 | H_1 \text{ true})$. Often we want to find an overall sample size n such that, for example, $1 - \beta = 0.9$ while capping off $\alpha = 0.05$.

One sample proportion

Say we want to test $H_0 : \pi = \pi_0$ for $Y \sim \text{bin}(n, \pi)$. The score test statistic is $Z_0 = \frac{\hat{\pi} - \pi_0}{\sigma_0}$ where $\hat{\pi} = Y/n$ and $\sigma_0 = \sqrt{\pi_0(1 - \pi_0)/n}$.

Under $H_0 : \pi = \pi_0$, $Z \overset{\bullet}{\sim} N(0, 1)$; this determines $z_{\alpha/2}$. The power $1 - \beta$ is a function of the hypothesized π_0 , the true π_1 , and the sample size through σ_0 and $\sigma_1 \sqrt{\pi_1(1 - \pi_1)/n}$. We compute:

$$\begin{aligned}
1 - \beta &= P(\text{reject } H_0 | H_1 \text{ true}) \\
&= P(|Z_0| > z_{\alpha/2} | \pi = \pi_1) \\
&= 1 - P(-z_{\alpha/2} \leq Z_0 \leq z_{\alpha/2} | \pi = \pi_1) \\
&= 1 - P(-z_{\alpha/2}\sigma_0 + \pi_0 \leq \hat{\pi} \leq z_{\alpha/2}\sigma_0 + \pi_0 | \pi = \pi_1) \\
&= 1 - P\left(\frac{-z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \leq \frac{\hat{\pi} - \pi_1}{\sigma_1} \leq \frac{z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1}\right) \\
&= 1 - P\left(\frac{-z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1} \leq Z \leq \frac{z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1}\right) \\
&= 1 - \left[\Phi\left(\frac{z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1}\right) - \Phi\left(\frac{-z_{\alpha/2}\sigma_0 + \pi_0 - \pi_1}{\sigma_1}\right) \right].
\end{aligned}$$

For a given β , α , π_0 , and π_1 , we can solve this equation for the sample size n . Check out <http://www.cs.uiowa.edu/~rlenth/Power/>

6.5.1 Testing $H_0 : \pi_1 = \pi_2$ from two samples

Recall the two-sample proportion problem. Assume the same number of observations n will be collected in each group $X = 1$ and $X = 2$.

$$Y_1 \sim \text{bin}(n_1, \pi_1) \perp Y_2 \sim \text{bin}(n_2, \pi_2).$$

Let $\hat{\pi}_1 = Y_1/n$ and $\hat{\pi}_2 = Y_2/n$. The CLT gives us

$$\hat{\pi}_1 \overset{\bullet}{\sim} N\left(\pi_1, \frac{\pi_1(1-\pi_1)}{n_1}\right) \perp \hat{\pi}_2 \overset{\bullet}{\sim} N\left(\pi_2, \frac{\pi_2(1-\pi_2)}{n_2}\right),$$

and so

$$\hat{\pi}_1 - \hat{\pi}_2 \overset{\bullet}{\sim} N\left(\pi_1 - \pi_2, \frac{\pi_1(1-\pi_1)}{n_1} + \frac{\pi_2(1-\pi_2)}{n_2}\right).$$

Under $H_0 : \pi_1 = \pi_2$ and $n_1 = n_2$ the test statistic is

$$Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{2\hat{\pi}(1 - \hat{\pi})/n}},$$

where $\hat{\pi} = (Y_1 + Y_2)/(2n)$ is the pooled estimator (the MLE under H_0).

Similar computations as in the one-sample case leads to

$$n_1 = n_2 = (z_{\alpha/2} + z_{\beta})^2 \frac{\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)}{(\pi_1 - \pi_2)^2}.$$

Note that for $\alpha = 0.05$ and $\beta = 0.1$ we have $z_{0.025} = 1.960$ and $z_{0.1} = 1.282$. $1 - \beta = 0.99$ yields $z_{0.01} = 2.326$.

What happens when $\pi_1 \approx \pi_2$?

6.5.2 Sample size for simple logistic regression*

Let

$$\text{logit } \pi(x) = \alpha + \beta X,$$

where $X \sim N(\mu, \sigma^2)$ and

$$\tau = \log \left\{ \frac{\pi(\mu + \sigma)/[1 - \pi(\mu + \sigma)]}{\pi(\mu)/[1 - \pi(\mu)]} \right\},$$

the log of the ratio of event odds when $x = \mu + \sigma$ and $x = \mu$. Then to test $H_0 : \beta \leq 0$ versus $H_0 : \beta > 0$ (or the other direction) at significance α and power $1 - \beta$ we need sample size

$$n = [z_\alpha + z_\beta e^{-\tau^2/4}]^2 [1 + 2\pi(\mu)\delta] / [\pi(\mu)\tau^2],$$

where

$$\delta = [1 + (1 + \tau^2)e^{5\tau^2/4}] / [1 + e^{-\tau^2/4}].$$

Text example.

- X is cholesterol level, Y indicates “severe heart disease.”
- Know $\pi(\mu) = 0.08$. Want to be able to detect a 50% increase in probability for a standard deviation increase in cholesterol. 50% increase in probability is $1.5 \times 0.08 = 0.12$.
- $\pi(\mu)/[1 - \pi(\mu)] = 0.08/0.92 = 0.087$.
- $\pi(\mu + \sigma)/[1 - \pi(\mu + \sigma)] = 0.12/0.88 = 0.136$. So the odds ratio is $0.136/0.087 = 1.57$, and $\tau = \log(1.57) = 0.45$.
- Then for $\alpha = 0.05$, $1 - \beta = 0.9$, we have $\delta = 1.306$ and $n = 612$.
- Note: didn't need to know μ and σ , but rather $\pi(\mu)$ and $\pi(\mu + \sigma)$.

6.5.3 Sample size for one effect in multiple logistic regression*

Say now that we're interested in X_1 but there's $p - 2$ more more predictors X_2, \dots, X_{p-1} . Let R denote the multiple correlation between X and the remaining predictors:

$$R = \max_{\|\mathbf{a}\|=1} \{\text{corr}(X_1, a_2X_2 + \dots + a_{p-1}X_{p-1})\}.$$

Let $\pi(\boldsymbol{\mu}) = \pi(\mu_1, \mu_2, \dots, \mu_{p-1})$ be the probability at the mean of all $p - 1$ variables.

τ is the now the log odds ratio comparing $\pi(\mu_1 + \sigma_1, \mu_2, \dots, \mu_{p-1})$ to $\pi(\mu_1, \mu_2, \dots, \mu_{p-1})$.

$$n = [z_\alpha + z_\beta e^{-\tau^2/4}]^2 [1 + 2\pi(\boldsymbol{\mu})\delta] / [\pi(\boldsymbol{\mu})\tau^2(1 - R^2)].$$

Text example (continued):

- Say we have another variable X_2 is blood pressure and $R = \text{corr}(X_1, X_2) = 0.4$.
- Then $n = 612 / (1 - 0.4^2) = 729$.
- What happens when $\text{corr}(X_1, X_2) \approx 1$. Is this problematic? Hint: think about the interpretation of β_1 .

6.5.4, 6.5.5, & 6.5.6 Misc. power and sample size considerations

Read over if interested.

6.6 Alternative links in binary regression*

There are three common links considered in binary regression: logistic, probit, and complimentary log-log. All three are written

$$\pi(\mathbf{x}) = F(\mathbf{x}'\boldsymbol{\beta}).$$

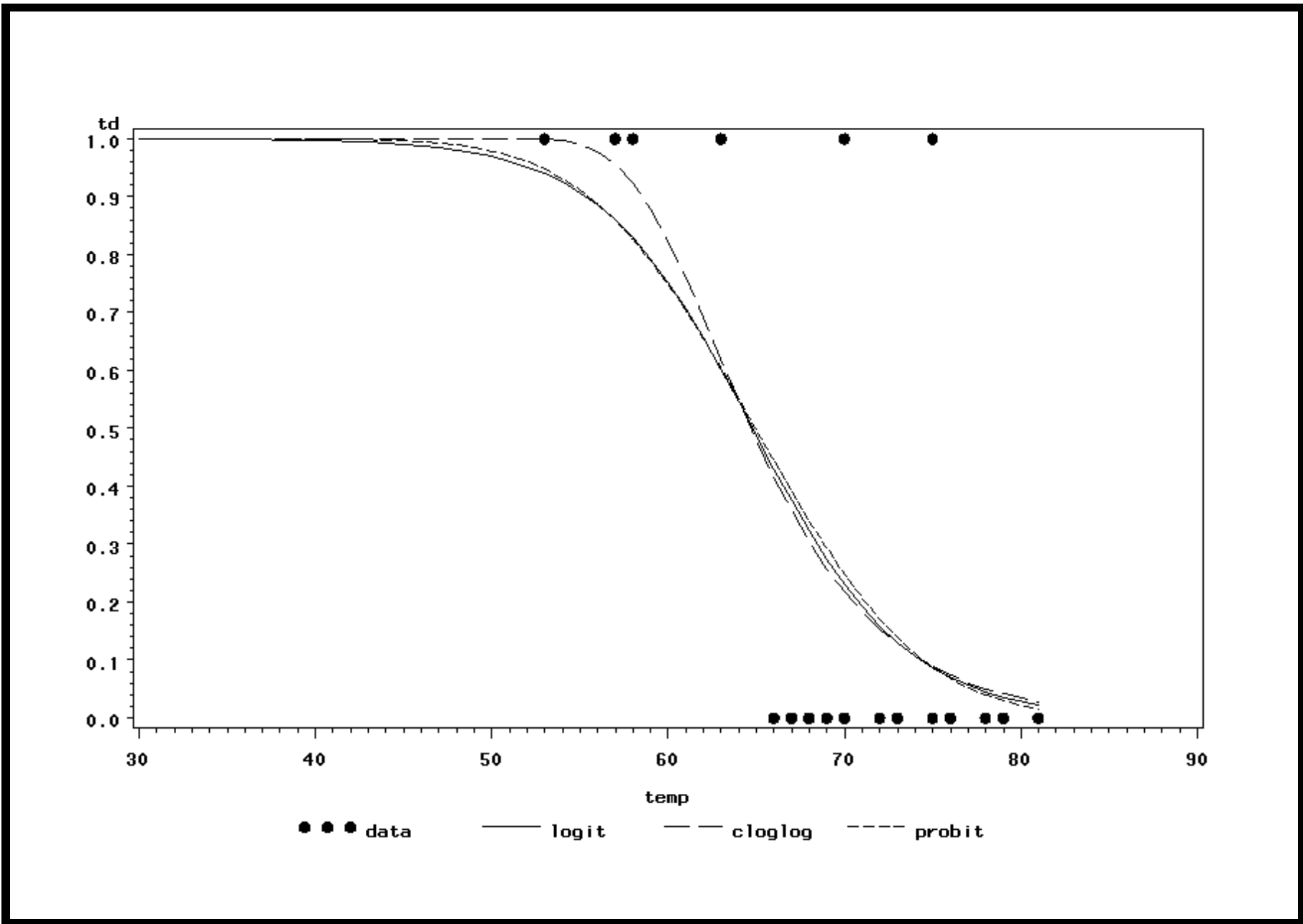
- Logistic regression: $F(x) = \frac{e^x}{1+e^x}$.
- Probit regression: $F(x) = \Phi(x)$ where $\Phi(x) = \int_{-\infty}^x \frac{e^{-0.5z^2}}{\sqrt{2\pi}} dz$.
- Complimentary log-log binary regression:
 $F(x) = 1 - \exp\{-\exp(x)\}$.

They differ primarily in the tails, but the logistic and probit links are symmetric in that rare and very common events are treated similarly in the tails. The CLL link approaches 1 faster than 0, so obtaining “rare event” status requires more extreme values of \mathbf{x} than reaching “likely event” status.

```

data shut1; input temp td @@; datalines;
  66 0 70 1 69 0 68 0 67 0 72 0 73 0 70 0 57 1 63 1
  70 1 78 0 67 0 53 1 67 0 75 0 70 0 81 0 76 0 79 0
  75 1 76 0 58 1
;
data shut2;
  do i=1 to 50; temp=i+29; td=.; output; end;
data shut3; set shut1 shut2;
proc logistic descending data=shut3; model td = temp / link=logit;
  output out=shut4 p=p1;
proc logistic descending data=shut3; model td = temp / link=cloglog;
  output out=shut5 p=p2;
proc logistic descending data=shut3; model td = temp / link=probit;
  output out=shut6 p=p3;
data shut7; set shut4 shut5 shut6;
proc sort data=shut7; by temp;
goptions;
  symbol1 color=black value=dot  interpol=none;
  symbol2 color=black value=none l=1  interpol=join;
  symbol3 color=black value=none l=2  interpol=join;
  symbol4 color=black value=none l=3  interpol=join;
  legend1 label=none value=('data' 'logit' 'cloglog' 'probit');
proc gplot data=shut7;
  plot td*temp p1*temp p2*temp p3*temp / overlay legend=legend1;

```



From the output:

Statistic	logit	cloglog	probit
AIC	24.3	23.5	24.4
$\hat{\alpha}$	15.0	12.3	8.8
$\hat{\beta}$	-0.23	-0.20	-0.14

- Complimentary log-log chosen as “best” out of three according to AIC.
- Fitted cloglog model is

$$\hat{\pi}(\text{temp}) = 1 - \exp\{-\exp(12.3 - 0.2 \text{ temp})\}.$$

- H-L p -values are 0.21, 0.23, 0.22 respectively.

Hierarchical model building:

“When using a polynomial regression model as an approximation to the true regression function, statisticians will often fit a second-order or third-order model and then explore whether a lower-order model is adequate...With the hierarchical approach, if a polynomial term of a given order is retained, then all related terms of lower order are also retained in the model. Thus, one would not drop the quadratic term of a predictor variable but retain the cubic term in the model. Since the quadratic term is of lower order, it is viewed as providing more basic information about the shape of the response function; the cubic term is of higher order and is viewed as providing refinements in the specification of the shape of the response function.” – *Applied Statistical Linear Models* by Neter, Kutner, Nachtsheim, and Wasserman.

“It is not usually sensible to consider a model with interaction but not the main effects that make up the interaction.” – *Categorical Data Analysis* by Agresti.

“Consider the relationship between the terms $\beta_1 x$ and $\beta_2 x^2$. To fit the term $\beta_0 + \beta_2 x^2$ without including $\beta_1 x$ implies that the maximum (or minimum) of the response occurs at $x = 0$...ordinarily there is no reason to suppose that the turning point of the response is at a specified point in the x -scale, so that the fitting of $\beta_2 x^2$ without the linear term is usually unhelpful.

A further example, involving more than one covariate, concerns the relation between a cross-term such as $\beta_{12} x_1 x_2$ and the corresponding linear terms $\beta_1 x_1$ and $\beta_2 x_2$. To include the former in a model formula without the latter two is equivalent to assuming the point $(0, 0)$ is a col or saddle-point of the response surface. Again, there is usually no reason to postulate such a property for the origin, so that the linear terms must be included with the cross-term.” – *Generalized Linear Models* by McCullagh and Nelder.

A polynomial model as an approximation to an unknown surface:

Real model

$$y_i = f(x_{i1}, x_{i2}) + e_i.$$

First order approximation to $f(x_1, x_2)$ about some (\bar{x}_1, \bar{x}_2) :

$$\begin{aligned} f(x_1, x_2) &= f(\bar{x}_1, \bar{x}_2) + \frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_1} (x_1 - \bar{x}_1) + \frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_2} (x_2 - \bar{x}_2) \\ &\quad + \text{HOT.} \\ &= \left[f(\bar{x}_1, \bar{x}_2) - \bar{x}_1 \frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_1} - \bar{x}_2 \frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_2} \right] \\ &\quad + \left[\frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_1} \right] x_1 + \left[\frac{\partial f(\bar{x}_1, \bar{x}_2)}{\partial x_2} \right] x_2 + \text{HOT} \\ &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \text{HOT} \end{aligned}$$

$E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$ is an approximation to unknown, infinite-dimensional $f(x_1, x_2)$ characterized by $(\beta_0, \beta_1, \beta_2)$.

Now let $\mathbf{x} = (x_1, x_2)$ and

$$f(\mathbf{x}) = f(\bar{\mathbf{x}}) + Df(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \frac{1}{2}(\mathbf{x} - \bar{\mathbf{x}})' D^2 f(\bar{\mathbf{x}})(\mathbf{x} - \bar{\mathbf{x}}) + \text{HOT}.$$

This similarly reduces to

$$f(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \text{HOT},$$

where $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$ correspond to various (unknown) partial derivatives of $f(x_1, x_2)$. Depending on the shape of the true (unknown) $f(x_1, x_2)$, some or many of the terms in the approximation $E(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2$ may be unnecessary.

We work backwards via Wald tests *hierarchically* getting rid of HOT first to get at more general trends/shapes, e.g. the first order approximation.

BTW, this directly relates to generalized additive models (GAM) where instead we approximate

$$f(x_1, x_2) = \beta_0 + f_1(x_1) + f_2(x_2),$$

where often

$$f_1(x_1) = \sum_{j=1}^J \theta_{1j} g_{1j}(x_1) \text{ and } f_2(x_1) = \sum_{j=1}^J \theta_{2j} g_{2j}(x_2),$$

functional expansions in terms of basis functions. Here, $(\theta_{11}, \dots, \theta_{1J})$ and $(\theta_{21}, \dots, \theta_{2J})$ are estimated from the data and the functions $\{g_{ij}(\cdot)\}$ are known; e.g. spline basis functions.

A simple additive model is a special case where $J = 1$ and $g_{11}(x) = g_{21}(x) = x$ yielding $f(x_1, x_2) = \beta_0 + \theta_{11}x_1 + \theta_{21}x_2$.
More on this later!