Stat 5102 Lecture Slides
Deck 6

Charles J. Geyer
School of Statistics
University of Minnesota

The Gauss-Markov Theorem

Suppose we do not want to assume the response vector is normal (conditionally given covariates that are random). What then?

One justification for still using least squares estimators (LSE), no longer MLE when normality is not assumed, is the following.

Theorem (Gauss-Markov). Suppose $\mathbf{Y}$ has mean vector $\boldsymbol{\mu}$ and variance matrix $\sigma^2\mathbf{I}$, and suppose $\boldsymbol{\mu} = \mathbf{M}\boldsymbol{\beta}$, where $\mathbf{M}$ has full rank. Then the LSE

$$\hat{\boldsymbol{\beta}} = (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{Y}$$

is the best linear unbiased estimator (BLUE) of $\boldsymbol{\beta}$, where "best" means

$$\text{var}(\mathbf{a}^T\hat{\boldsymbol{\beta}}) \leq \text{var}(\mathbf{a}^T\tilde{\boldsymbol{\beta}}), \qquad \text{for all } \mathbf{a} \in \mathbb{R}^p$$

where $\tilde{\boldsymbol{\beta}}$ is any other linear and unbiased estimator.

The Gauss-Markov Theorem (cont.)

We do not assume normality. We do assume the same first and second moments of $\mathbf{Y}$ as in the linear model. We get the conclusion that the LSE are BLUE, rather than MLE.

They can't be MLE because we don't have a statistical model, having specified only moments, not distributions, so there is no likelihood.

By the definition of "best" all linear functions of $\hat{\boldsymbol{\beta}}$ are also BLUE. This includes $\hat{\boldsymbol{\mu}} = \mathbf{M}\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\mu}}_{\text{new}} = \mathbf{M}_{\text{new}}\hat{\boldsymbol{\beta}}$.

The Gauss-Markov Theorem (cont.)

Proof of Gauss-Markov Theorem. The condition that $\tilde{\boldsymbol{\beta}}$ be linear and unbiased is $\tilde{\boldsymbol{\beta}} = \mathbf{AY}$ for some matrix $\mathbf{A}$ satisfying

$$E(\tilde{\boldsymbol{\beta}}) = \mathbf{A}\boldsymbol{\mu} = \mathbf{AM}\boldsymbol{\beta} = \boldsymbol{\beta}$$

for all $\boldsymbol{\beta}$. Hence, if $\mathbf{AM}$ is full rank, then $\mathbf{AM} = \mathbf{I}$. It simplifies the proof if we define

$$\mathbf{B} = \mathbf{A} - (\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T$$

so

$$\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} + \mathbf{BY}$$

and $\mathbf{BM} = 0$.

The Gauss-Markov Theorem (cont.)

For any vector $\mathbf{a}$

$$\mathrm{var}(\mathbf{a}^T\tilde{\boldsymbol{\beta}}) = \mathrm{var}(\mathbf{a}^T\hat{\boldsymbol{\beta}}) + \mathrm{var}(\mathbf{a}^T\mathbf{B}\mathbf{Y}) + 2\,\mathrm{cov}(\mathbf{a}^T\hat{\boldsymbol{\beta}}, \mathbf{a}^T\mathbf{B}\mathbf{Y})$$

If the covariance here is zero, that proves the theorem. Hence it only remains to prove that.

$$\begin{aligned}\mathrm{cov}(\mathbf{a}^T\hat{\boldsymbol{\beta}}, \mathbf{a}^T\mathbf{B}\mathbf{Y}) &= \mathbf{a}^T(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\,\mathrm{var}(\mathbf{Y})\mathbf{B}^T\mathbf{a} \\ &= \sigma^2\mathbf{a}^T(\mathbf{M}^T\mathbf{M})^{-1}\mathbf{M}^T\mathbf{B}^T\mathbf{a}\end{aligned}$$

is zero because $\mathbf{B}\mathbf{M} = 0$ hence $\mathbf{M}^T\mathbf{B}^T = 0$. And that finishes the proof of the theorem.

The Gauss-Markov Theorem (cont.)

Criticism of the theorem. The conclusion that LSE are BLUE can seem to say more than it actually says. It doesn't say the LSE are the best estimators. It only says they are best among linear and unbiased estimates. Presumably there are better estimators that are either biased or nonlinear. Otherwise a stronger theorem could be proved.

The Gauss-Markov theorem drops the assumption of exact normality, but it keeps the assumption that the mean specification $\boldsymbol{\mu} = \mathbf{M}\boldsymbol{\beta}$ is correct. When this assumption is false, the LSE are not unbiased. More on this later.

Not specifying a model, the assumptions of the Gauss-Markov theorem do not lead to confidence intervals or hypothesis tests.

Bernoulli Response

Suppose the data vector $\mathbf{Y}$ has independent Bernoulli components.

The assumption $\boldsymbol{\mu} = \mathbf{M}\boldsymbol{\beta}$ now seems absurd, because

$$E(Y_i) = \mathsf{Pr}(Y_i = 1)$$

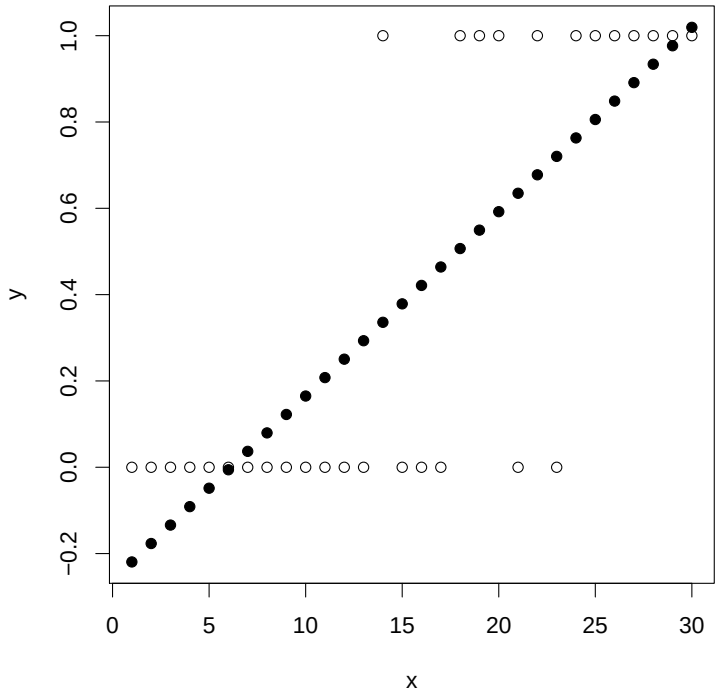
is between zero and one, and linear functions are not constrained this way. Moreover

$$\mathsf{var}(Y_i) = \mathsf{Pr}(Y_i = 1)\,\mathsf{Pr}(Y_i = 0)$$

so we cannot have constant variance $\mathsf{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Bernoulli Response (cont.)

Here is what happens if we try to apply LSE to Bernoulli data with the simple linear regression model $\mu_i = \beta_1 + \beta_2 x_i$. Hollow dots are the data, solid dots the LSE predicted values.



Bernoulli Response (cont.)

The predicted values go outside the range of possible values. Not good.

Also there is no way to do statistics — confidence intervals and hypothesis tests – based on this model. Also not good.

We need a better idea.

Sufficiency

Given a statistical model with parameter vector $\boldsymbol{\theta}$ and data vector $\mathbf{Y}$, a statistic $\mathbf{Z} = g(\mathbf{Y})$, which may also be vector-valued, is called *sufficient* if the conditional distribution of $\mathbf{Y}$ given $\mathbf{Z}$ does not depend on $\boldsymbol{\theta}$.

A sufficient statistic incorporates all of the information in the data $\mathbf{Y}$ about the parameter $\boldsymbol{\theta}$ (assuming the correctness of the statistical model).

Sufficiency (cont.)

The *sufficiency principle* says that all statistical inference should depend on the data only through the sufficient statistic.

The likelihood is

$$L(\boldsymbol{\theta}) = f(\mathbf{Y} \mid \mathbf{Z}) f_{\boldsymbol{\theta}}(\mathbf{Z})$$

and we may drop terms that do not contain the parameter so the likelihood is also

$$L(\boldsymbol{\theta}) = f_{\boldsymbol{\theta}}(\mathbf{Z})$$

Hence likelihood inference and Bayesian inference automatically obey the sufficiency principle. Non-likelihood frequentist inference (such as the method of moments) does not automatically obey the sufficiency principle.

Sufficiency (cont.)

The converse of this is also true. The *Neyman-Fisher factorization criterion* says that if the likelihood is a function of the data $\mathbf{Y}$ only through a statistic $\mathbf{Z}$, then $\mathbf{Z}$ is sufficient.

This is because

$$f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z}) = f_{\boldsymbol{\theta}}(\mathbf{y} \mid \mathbf{z}) f_{\boldsymbol{\theta}}(\mathbf{z}) = h(\mathbf{y}) L(\boldsymbol{\theta})$$

where $L(\boldsymbol{\theta})$ depends on $\mathbf{Y}$ only through $\mathbf{Z}$ and $h(\mathbf{Y})$ does not contain $\boldsymbol{\theta}$. Write $L_{\mathbf{z}}(\boldsymbol{\theta})$ for $L(\boldsymbol{\theta})$ to remind us of the dependence on $\mathbf{Z}$. Then

$$f_{\boldsymbol{\theta}}(\mathbf{z}) = \int_A f_{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z})\, d\mathbf{y} = L_{\mathbf{z}}(\boldsymbol{\theta}) \int_A h(\mathbf{y})\, d\mathbf{y}$$

where $A = \{\, \mathbf{y} : g(\mathbf{y}) = \mathbf{z} \,\}$.

Sufficiency (cont.)

Hence

$$f_{\boldsymbol{\theta}}(\mathbf{Y} \mid \mathbf{Z}) = \frac{f_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})}{f_{\boldsymbol{\theta}}(\mathbf{Z})} = \frac{h(\mathbf{y})}{\int_A h(\mathbf{y})\, d\mathbf{y}}$$

does not depend on $\boldsymbol{\theta}$. That finishes (a sketchy but correct) proof of the Neyman-Fisher factorization criterion. For the discrete case, replace integrals by sums.

Sufficiency (cont.)

The whole data is always sufficient, that is, the criterion is trivially satisfied when $\mathbf{Z} = \mathbf{Y}$.

There need not be any non-trivial sufficient statistic.

Sufficiency and Exponential Families

Recall the theory of exponential families of distributions (deck 3, slides 105–113). A statistical model is called an *exponential family of distributions* if the log likelihood has the form

$$l(\boldsymbol{\theta}) = \sum_{i=1}^{p} t_i(\mathbf{x}) g_i(\boldsymbol{\theta}) - c(\boldsymbol{\theta})$$

By the Neyman-Fisher factorization criterion

$$\mathbf{Y} = \Big(t_1(\mathbf{X}), \ldots, t_p(\mathbf{X})\Big)$$

is a p-dimensional sufficient statistic. It is called the *natural statistic* of the family. Also

$$\boldsymbol{\psi} = \Big(g_1(\boldsymbol{\theta}), \ldots, g_p(\boldsymbol{\theta})\Big)$$

is a p-dimensional parameter vector for the family, called the *natural parameter*.

Sufficiency and Exponential Families (cont.)

We want to use $\boldsymbol{\theta}$ for the natural parameter vector instead of $\boldsymbol{\psi}$ from here on. Then the log likelihood is

$$l(\boldsymbol{\theta}) = \mathbf{y}^T \boldsymbol{\theta} - c(\boldsymbol{\theta})$$

A *natural affine submodel* is specified by a parametrization

$$\boldsymbol{\theta} = \mathbf{a} + \mathbf{M}\boldsymbol{\beta}$$

where $\mathbf{a}$ is a known vector and $\mathbf{M}$ is a known matrix, called the *offset vector* and *model matrix*. Usually $\mathbf{a} = 0$, in which case we have a *natural linear submodel*.

Sufficiency and Exponential Families (cont.)

The log likelihood for the natural affine submodel is

$$l(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{a} + \mathbf{y}^T\mathbf{M}\boldsymbol{\beta} - c(\mathbf{a} + \mathbf{M}\boldsymbol{\beta})$$

and the term that does not contain $\boldsymbol{\beta}$ can be dropped, giving

$$l(\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{M}\boldsymbol{\beta} - c(\mathbf{a} + \mathbf{M}\boldsymbol{\beta}) = (\mathbf{M}^T\mathbf{y})^T\boldsymbol{\beta} - c(\mathbf{a} + \mathbf{M}\boldsymbol{\beta})$$

which we see also has the exponential family form. We have a new exponential family, with natural statistic $\mathbf{M}^T\mathbf{y}$ and natural parameter $\boldsymbol{\beta}$.

Sufficiency and Exponential Families (cont.)

The log likelihood derivatives are

$$\nabla l(\boldsymbol{\beta}) = \mathbf{M}^T \mathbf{y} - \mathbf{M}^T \nabla c(\mathbf{a} + \mathbf{M}\boldsymbol{\beta})$$
$$\nabla^2 l(\boldsymbol{\beta}) = -\mathbf{M}^T \nabla^2 c(\mathbf{a} + \mathbf{M}\boldsymbol{\beta}) \mathbf{M}$$

The log likelihood derivative identities say

$$E_{\boldsymbol{\beta}}\{\nabla l(\boldsymbol{\beta})\} = 0$$
$$\text{var}_{\boldsymbol{\beta}}\{\nabla l(\boldsymbol{\beta})\} = -E_{\boldsymbol{\beta}}\{\nabla^2 l(\boldsymbol{\beta})\}$$

Sufficiency and Exponential Families (cont.)

Combining these we get

$$E_{\boldsymbol{\beta}}\{\mathbf{M}^T\mathbf{Y}\} = \mathbf{M}^T\nabla c(\mathbf{a} + \mathbf{M}\boldsymbol{\beta})$$
$$\text{var}_{\boldsymbol{\beta}}\{\mathbf{M}^T\mathbf{Y}\} = \mathbf{M}^T\nabla^2 c(\mathbf{a} + \mathbf{M}\boldsymbol{\beta})\mathbf{M}$$

Hence the MLE is found by solving

$$\mathbf{M}^T\mathbf{y} = \mathbf{M}^T E_{\boldsymbol{\beta}}(\mathbf{Y})$$

for $\boldsymbol{\beta}$ ("observed equals expected"), and observed and expected Fisher information for $\boldsymbol{\beta}$ are the same

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbf{M}^T\nabla^2 c(\mathbf{a} + \mathbf{M}\boldsymbol{\beta})\mathbf{M}$$

If the distribution of the natural statistic vector $\mathbf{M}^T\mathbf{Y}$ is non-degenerate, then the log likelihood is strictly concave and the MLE is unique if it exists and is the global maximizer of the log likelihood.

Bernoulli Response (cont.)

Let us see how this helps us with Bernoulli response models.

The Bernoulli distribution is an exponential family. The log likelihood is

$$\begin{aligned} l(p) &= y \log(p) + (1 - y) \log(1 - p) \\ &= y \Big[\log(p) - \log(1 - p) \Big] + \log(1 - p) \\ &= y \log \left(\frac{p}{1 - p} \right) + \log(1 - p) \end{aligned}$$

so the natural statistic is y and the natural parameter is

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

This function is called logit and pronounced with a soft "g" (low-jit).

Bernoulli Response (cont.)

The notion of natural affine submodels, suggests we model the natural parameter affinely. If $Y_1, \ldots, Y_n$ are independent Bernoulli random variables with

$$Y_i \sim \text{Ber}(\mu_i)$$

let

$$\theta_i = \text{logit}(\mu_i)$$

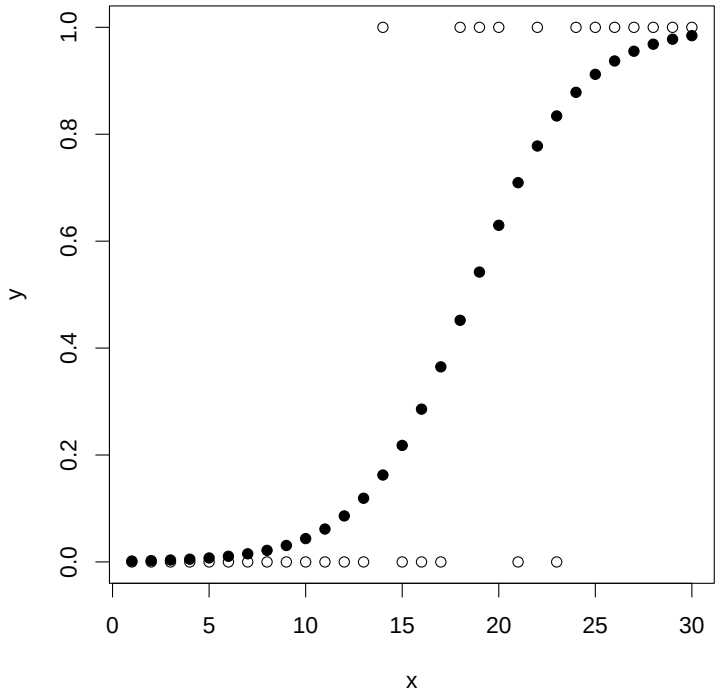
and

$$\boldsymbol{\theta} = \mathbf{a} + \mathbf{M}\boldsymbol{\beta}$$

This idea is called *logistic regression*.

Bernoulli Response (cont.)

Here is what happens if we apply logistic regression to Bernoulli data with the simple linear regression model $\theta_i = \beta_1 + \beta_2 x_i$. Hollow dots are the data, solid dots the MLE mean values $\widehat{\mu}_i$.



Bernoulli Response (cont.)

The R commands to make the picture on the preceding page are

```
Rweb:> lout <- glm(y ~ x, family = binomial)
Rweb:> plot(x, y)
Rweb:> points(x, predict(lout, type = "response"), pch = 19)
```

There are differences between *generalized linear models* (GLM) fit by the R function `glm` and *linear models* (LM) fit by the R function `lm`. We need to specify `family = binomial` because `glm` can fit response distributions other than Bernoulli. We need to specify `type = "response"` in the `predict` function because this function "predicts" (estimates, actually) either natural parameters or mean-value parameters. The formula `y ~ x` is the same for GLM and LM.

Bernoulli Response (cont.)

```
Rweb:>  summary(lout)

Call:
glm(formula = y ~ x, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.8959  -0.3421  -0.0936   0.3460   1.9061

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.7025     2.4554  -2.730  0.00634 **
x             0.3617     0.1295   2.792  0.00524 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Bernoulli Response (cont.)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.7025     2.4554  -2.730  0.00634 **
x             0.3617     0.1295   2.792  0.00524 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

More differences between GLM and LM. The "`Estimate`" column contains MLE (not LSE) of the regression coefficients. The "`Std. Error`" column contains estimated standard deviations of the regression coefficients, which are approximate (not exact) obtained from the inverse Fisher information matrix. The "`z value`" column gives the asymptotic (not exact) test statistic for a two-tailed test of whether the regression coefficient is zero; its reference distribution is standard normal (not Student t). The "`Pr(>|z|)`" column gives the P-value for this test.

Bernoulli Response (cont.)

```
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -6.7025     2.4554  -2.730  0.00634 **
x             0.3617     0.1295   2.792  0.00524 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

An approximate, large sample 95% confidence interval for the second regression coefficient is

```
Rweb:> 0.3617 + c(-1,1) * qnorm(0.975) * 0.1295
[1] 0.1078847 0.6155153
```

Bernoulli Response (cont.)

So far everything is similar for GLM and LM. There are differences inherent in the nature of GLM. There are some extra arguments to functions because GLM are more complicated. Hypothesis tests and confidence intervals are approximate, based on the asymptotics of maximum likelihood.

Bernoulli Response (cont.)

Estimate of the natural parameter $\theta = \text{logit}(p)$ for a new individual with covariate value $x = 25$, and its standard error derived from inverse Fisher information and the delta method.

```
Rweb:> tout <- predict(lout, newdata = data.frame(x = 25),
+     se.fit = TRUE)
Rweb:> print(tout)
$fit
       1
2.339301

$se.fit
[1] 1.051138
```

Bernoulli Response (cont.)

Asymptotic 95% confidence interval

```
Rweb:> tout$fit + c(-1,1) * qnorm(0.975) * tout$se.fit
[1] 0.2791077 4.3994944
```

Bernoulli Response (cont.)

Estimate of the mean-value parameter p for a new individual with covariate value $x = 25$, and its standard error derived from inverse Fisher information and the delta method.

```
Rweb:> pout <- predict(lout, newdata = data.frame(x = 25),
+     se.fit = TRUE, type = "response")
Rweb:> print(pout)
$fit
      1
0.91208

$se.fit
         1
0.08429082
```

Bernoulli Response (cont.)

Asymptotic 95% confidence intervals

```
Rweb:> pout$fit + c(-1,1) * qnorm(0.975) * pout$se.fit
[1] 0.7468731 1.0772870
Rweb:> invlogit <- function(theta) 1 / (1 + exp(- theta))
Rweb:> invlogit(tout$fit + c(-1,1) * qnorm(0.975) * tout$se.fit)
[1] 0.5693274 0.9878655
```

These intervals are asymptotically equivalent, but the sample size is not large enough for them to be close. Clearly, the delta method does not work so well here. Perhaps the second interval is preferred.

Likelihood Ratio Tests

Suppose l_n is the log likelihood for a statistical model that satisfies the "usual regularity conditions" for maximum likelihood. Suppose we have a nested submodel specified by $\boldsymbol{\theta} = \mathbf{M}\boldsymbol{\beta}$, and $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\beta}}_n$ are the MLE for the supermodel and submodel, respectively. Suppose the Fisher information matrix for the supermodel $\mathbf{I}(\boldsymbol{\theta})$ and submodel $\mathbf{M}^T\mathbf{I}(\boldsymbol{\theta})\mathbf{M}$ are both full rank. Then the asymptotic distribution of

$$2\Big[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\mathbf{M}\hat{\boldsymbol{\beta}}_n)\Big]$$

is chi-square with degrees of freedom that is the difference in dimensions of $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$.

Likelihood Ratio Tests (cont.)

We now repeat the argument in Deck 3, Slides 32–50, 86–87, and 90–91 to get simultaneous asymptotics for $\hat{\boldsymbol{\theta}}_n$ and $\hat{\boldsymbol{\beta}}_n$.

Expanding the gradient of the log likelihood in a Taylor series gives

$$\nabla l_n(\boldsymbol{\theta}) \approx \nabla l_n(\boldsymbol{\theta}_0) + \Big[\nabla^2 l_n(\boldsymbol{\theta}_0)\Big](\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

from which we obtain

$$o_p(1) = n^{-1/2}\nabla l_n(\boldsymbol{\theta}_0) + \Big[n^{-1}\nabla^2 l_n(\boldsymbol{\theta}_0)\Big] n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)$$

$$o_p(1) = n^{-1/2}\nabla l_n(\boldsymbol{\theta}_0)\mathbf{M} + \Big[n^{-1}\mathbf{M}^T\nabla^2 l_n(\boldsymbol{\theta}_0)\mathbf{M}\Big] n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)$$

where $\boldsymbol{\theta}_0 = \mathbf{M}\boldsymbol{\beta}_0$ is the true parameter value.

Likelihood Ratio Tests (cont.)

This gives

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \left[-n^{-1}\nabla^2 l_n(\boldsymbol{\theta}_0)\right]^{-1} n^{-1/2}\nabla l_n(\boldsymbol{\theta}_0) + o_p(1)$$

$$n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) = \left[-n^{-1}\mathbf{M}^T\nabla^2 l_n(\boldsymbol{\theta}_0)\mathbf{M}\right]^{-1} n^{-1/2}\nabla l_n(\boldsymbol{\theta}_0)\mathbf{M} + o_p(1)$$

from which

$$n^{-1/2}\nabla l_n(\boldsymbol{\theta}_0) \xrightarrow{\mathcal{D}} \mathcal{N}\Big(0, \mathbf{I}(\boldsymbol{\theta}_0)\Big)$$

$$-n^{-1}\nabla^2 l_n(\boldsymbol{\theta}_0) \xrightarrow{P} \mathbf{I}(\boldsymbol{\theta}_0)$$

and Slutsky's theorem give

$$\begin{pmatrix} n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\ n^{-1/2}\nabla l_n(\boldsymbol{\theta}_0) \end{pmatrix} \xrightarrow{\mathcal{D}} \begin{pmatrix} \mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{Z} \\ \left[\mathbf{M}^T\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{M}\right]^{-1}\mathbf{M}^T\mathbf{Z} \\ \mathbf{Z} \end{pmatrix}$$

where $\mathbf{Z} \sim \mathcal{N}\Big(0, \mathbf{I}(\boldsymbol{\theta}_0)\Big)$.

Likelihood Ratio Tests (cont.)

Now we expand the log likelihood itself in a Taylor series giving

$$l_n(\boldsymbol{\theta}) - l_n(\boldsymbol{\theta}_0) \approx \Big[\nabla l_n(\boldsymbol{\theta}_0)\Big]^T(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{\theta}_0)^T\Big[\nabla^2 l_n(\boldsymbol{\theta}_0)\Big](\boldsymbol{\theta}-\boldsymbol{\theta}_0)$$

from which we obtain

$$\begin{aligned}
2\Big[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\mathbf{M}\hat{\boldsymbol{\beta}}_n)\Big] &= 2\Big[n^{-1/2}\nabla l_n(\boldsymbol{\theta}_0)\Big]^T n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad + n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^T\Big[n^{-1}\nabla^2 l_n(\boldsymbol{\theta}_0)\Big]n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\
&\quad - 2\Big[n^{-1/2}\nabla l_n(\boldsymbol{\theta}_0)\Big]^T n^{1/2}\mathbf{M}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
&\quad - n^{1/2}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0)^T\mathbf{M}^T\Big[n^{-1}\nabla^2 l_n(\boldsymbol{\theta}_0)\Big]n^{1/2}\mathbf{M}(\hat{\boldsymbol{\beta}}_n - \boldsymbol{\beta}_0) \\
&\quad + o_p(1)
\end{aligned}$$

Likelihood Ratio Tests (cont.)

Using the results established on slides 34–35 and Slutsky's theorem, we obtain

$$
\begin{aligned}
2\Big[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\mathbf{M}\hat{\boldsymbol{\beta}}_n)\Big] \xrightarrow{\mathcal{D}} & \; 2\mathbf{Z}^T\mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{Z} + \mathbf{Z}^T\mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{I}(\boldsymbol{\theta}_0)^{-1}\mathbf{Z} \\
& - 2\mathbf{Z}^T\mathbf{M}\Big[\mathbf{M}^T\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{M}\Big]^{-1}\mathbf{M}^T\mathbf{Z} \\
& - \mathbf{Z}^T\mathbf{M}\Big[\mathbf{M}^T\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{M}\Big]^{-1}\mathbf{M}^T\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{M}\Big[\mathbf{M}^T\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{M}\Big]^{-1}\mathbf{M}^T\mathbf{Z} \\
= & \; \mathbf{Z}^T\Big(\mathbf{I}(\boldsymbol{\theta}_0)^{-1} - \mathbf{M}\Big[\mathbf{M}^T\mathbf{I}(\boldsymbol{\theta}_0)\mathbf{M}\Big]^{-1}\mathbf{M}^T\Big)\mathbf{Z}
\end{aligned}
$$

where $\mathbf{Z} \sim \mathcal{N}\Big(0, \mathbf{I}(\boldsymbol{\theta}_0)\Big)$.

Likelihood Ratio Tests (cont.)

Let $\mathbf{A}$ be the symmetric matrix square root of $\mathbf{I}(\theta_0)$ (5101, Deck 5, Slide 110), that is, if

$$\mathbf{I}(\boldsymbol{\theta}_0) = \mathbf{O}\mathbf{D}\mathbf{O}^T$$

is a spectral decomposition ($\mathbf{O}$ is orthogonal and $\mathbf{D}$ is diagonal), then

$$\mathbf{A} = \mathbf{O}\mathbf{D}^{1/2}\mathbf{O}^T$$

where $\mathbf{D}^{1/2}$ is diagonal and its diagonal elements are the square roots of the corresponding diagonal elements of $\mathbf{D}$. Assuming the Fisher information matrix is invertible, so is $\mathbf{A}$ and

$$\mathbf{A}^{-1} = \mathbf{O}\mathbf{D}^{-1/2}\mathbf{O}^T$$

Likelihood Ratio Tests (cont.)

Suppose $\mathbf{U}$ is a multivariate standard normal random vector. Then

$$\begin{aligned} E(\mathbf{AU}) &= 0 \\ \operatorname{var}(\mathbf{AU}) &= \mathbf{A} \operatorname{var}(\mathbf{U}) \mathbf{A}^T = \mathbf{A}^2 = \mathbf{I}(\boldsymbol{\theta}_0) \end{aligned}$$

so $\mathbf{Z}$ and $\mathbf{AU}$ have the same distribution. Hence

$$\begin{aligned} 2\Big[l_n(\hat{\boldsymbol{\theta}}_n) - l_n(\mathbf{M}\hat{\boldsymbol{\beta}}_n)\Big] &\xrightarrow{\mathcal{D}} \mathbf{U}^T \mathbf{A} \Big(\mathbf{A}^{-2} - \mathbf{M} \Big[\mathbf{M}^T \mathbf{A}^2 \mathbf{M} \Big]^{-1} \mathbf{M}^T \Big) \mathbf{AU} \\ &= \mathbf{U}^T \Big(\mathbf{I} - \mathbf{AM} \Big[\mathbf{M}^T \mathbf{A}^2 \mathbf{M} \Big]^{-1} \mathbf{M}^T \mathbf{A} \Big) \mathbf{AU} \\ &= \mathbf{U}^T (\mathbf{I} - \mathbf{H}) \mathbf{U} \end{aligned}$$

where

$$\mathbf{H} = \mathbf{AM} \Big[\mathbf{M}^T \mathbf{A}^2 \mathbf{M} \Big]^{-1} \mathbf{M}^T \mathbf{A}$$

is the hat matrix corresponding to the model matrix $\mathbf{AM}$.

Likelihood Ratio Tests (cont.)

We know from the basic theorem for linear models (Deck 5, Slide 31) that $\mathbf{U}^T(\mathbf{I} - \mathbf{H})\mathbf{U}$ has a chi-square distribution with $n - q$ degrees of freedom, where here n is the rank of $\mathbf{I}$ and q is the rank of $\mathbf{H}$. Here n is the dimension of $\boldsymbol{\theta}$ and q is also the rank of $\mathbf{M}$ and the dimension of $\boldsymbol{\beta}$ (assuming $\mathbf{M}$ is full rank).

That proves the theorem about likelihood ratio tests.

Poisson Response

Suppose the data vector $\mathbf{Y}$ has independent Poisson components.

The assumption $\boldsymbol{\mu} = \mathbf{M}\boldsymbol{\beta}$ again seems absurd, because

$$E(Y_i) \geq 0$$

and linear functions are not constrained this way. Moreover

$$\text{var}(Y_i) = \mu$$

so we cannot have constant variance $\text{var}(\mathbf{Y}) = \sigma^2\mathbf{I}$.

Poisson Response (cont.)

The Poisson distribution is an exponential family. The log likelihood is

$$l(\mu) = y \log(\mu) - \mu$$

so the natural statistic is y and the natural parameter is

$$\theta = \log(\mu)$$

Poisson Response (cont.)

The notion of natural affine submodels, suggests we model the natural parameter affinely. If Y_1, ..., Y_n are independent Poisson random variables with

$$Y_i \sim \text{Poi}(\mu_i)$$

let

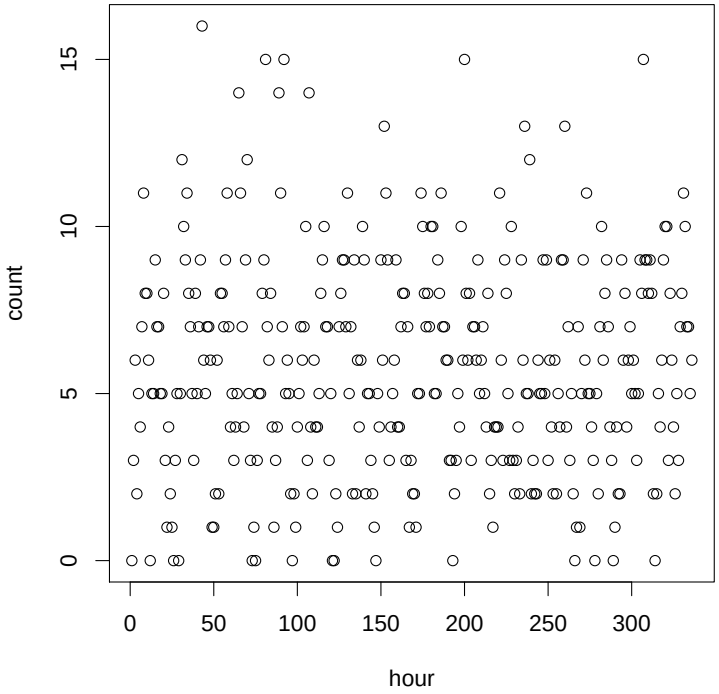
$$\theta_i = \log(\mu_i)$$

and

$$\boldsymbol{\theta} = \mathbf{a} + \mathbf{M}\boldsymbol{\beta}$$

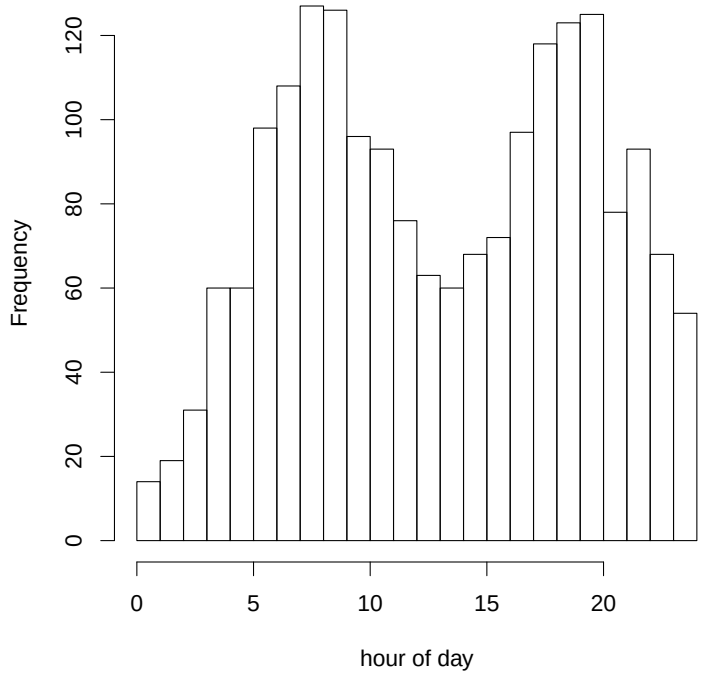
This idea is called *Poisson regression*.

Poisson Response (cont.)



Counts for a non-homogeneous Poisson process in each hour throughout a 14 day period.

Poisson Response (cont.)



Counts for the same process shown on the previous slide aggregated by hour of the day.

Poisson Response (cont.)

The process seems to be periodic with two daily peaks. Hence we model the natural parameter as a Fourier series with terms have frequencies one per day and two per day. The following R statements do this.

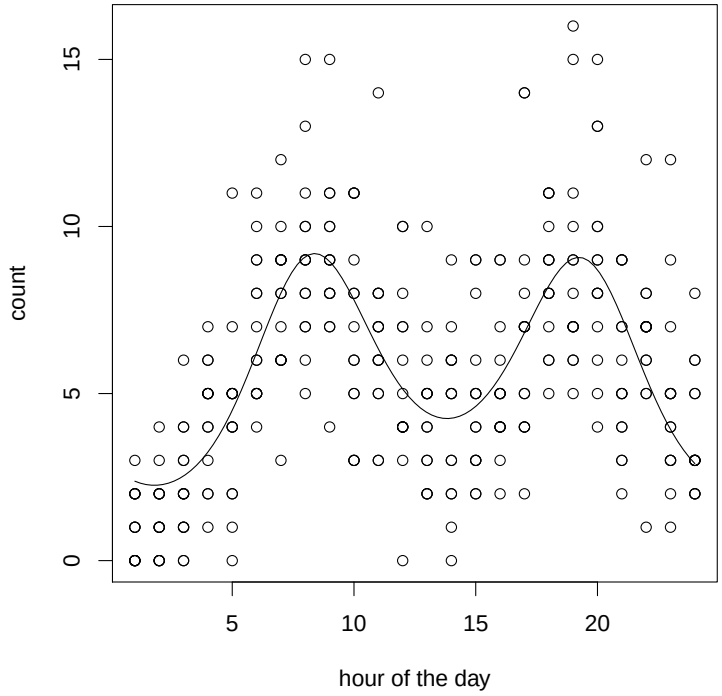
```
Rweb:> w <- hour / 24 * 2 * pi
Rweb:> out2 <- glm(count ~ I(sin(w)) + I(cos(w)) +
+     I(sin(2 * w)) + I(cos(2 * w)), family = poisson)
Rweb:> summary(out2)
Rweb:> plot(hourofday, count, xlab = "hour of the day")
Rweb:> curve(predict(out2, data.frame(w = x / 24 * 2 * pi),
+     type="response"), add=TRUE)
```

Poisson Response (cont.)

Regression coefficients table output by the `summary` command

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   1.65917    0.02494  66.516  < 2e-16 ***
I(sin(w))    -0.13916    0.03128  -4.448 8.66e-06 ***
I(cos(w))    -0.28510    0.03661  -7.787 6.86e-15 ***
I(sin(2 * w)) -0.42974    0.03385 -12.696  < 2e-16 ***
I(cos(2 * w)) -0.30846    0.03346  -9.219  < 2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Poisson Response (cont.)



Counts for the process shown on slides 43–44 plotted against hour of the day with estimated regression function.

Poisson Response (cont.)

Regression coefficients are of no interest at all in a model like this.

Here is an example confidence interval for the natural parameter for the first hour of the day.

```
Rweb:> w1 <- 1 / 24 * 2 * pi
Rweb:> tout <- predict(out2, newdata = data.frame(w = w1),
+     se.fit = TRUE)
Rweb:> tout$fit + c(-1, 1) * qnorm(0.975) * tout$se.fit
[1] 0.7318283 0.9996925
```

Poisson Response (cont.)

And here are corresponding confidence intervals for the mean-value parameter for the first hour of the day.

```
Rweb:> pout <- predict(out2, newdata = data.frame(w = w1),
+    se.fit = TRUE, type = "response")
Rweb:> pout$fit + c(-1, 1) * qnorm(0.975) * pout$se.fit
[1] 2.058481 2.695144
Rweb:> exp(tout$fit + c(-1, 1) * qnorm(0.975) * tout$se.fit)
[1] 2.078878 2.717446
```

Poisson Response (cont.)

The first interval on the preceding slide uses the estimated mean value plus or minus 1.96 standard errors calculated using inverse Fisher information and the delta method. The second uses the interval for θ given on the slide before that, mapping it to the mean value parameter scale.

Unlike the situation on slide 31, these two kinds of intervals closely agree in this example. So asymptotics of maximum likelihood and the delta method seem to be working well here.

Poisson Response (cont.)

Fit three models in which the natural parameter is given by a Fourier series with frequency one per day, two per day or three per day

```
Rweb:> out1 <- glm(count ~ I(sin(w)) + I(cos(w)),
+    family = poisson)
Rweb:> out2 <- glm(count ~ I(sin(w)) + I(cos(w)) +
+    I(sin(2 * w)) + I(cos(2 * w)), family = poisson)
Rweb:> out3 <- glm(count ~ I(sin(w)) + I(cos(w)) +
+    I(sin(2 * w)) + I(cos(2 * w)) + I(sin(3 * w)) +
+    I(cos(3 * w)), family = poisson)
```

Poisson Response (cont.)

Do likelihood ratio tests of model comparison (also called analysis of deviance)

```
Rweb:> anova(out1, out2, out3, test = "Chisq")
Analysis of Deviance Table

Model 1: count ~ I(sin(w)) + I(cos(w))
Model 2: count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w)) +
    I(cos(2 * w))
Model 3: count ~ I(sin(w)) + I(cos(w)) + I(sin(2 * w)) +
    I(cos(2 * w)) + I(sin(3 * w)) + I(cos(3 * w))
  Resid. Df Resid. Dev  Df Deviance P(>|Chi|)
1       333     651.10
2       331     399.58   2   251.52 2.412e-55
3       329     396.03   2     3.55      0.17
```

Poisson Response (cont.)

Model 3 fits no better than model 2 ($P = 0.17$). Model 2 fits much better than model 1 ($P \approx 0$). Hence Model 2 is the simplest model that appears to fit the data.

Link Functions, Latent Variables

The approach to Bernoulli response presented here, now most widely used, was not historically the first.

The first was probit regression.

Here is a story that appeals to some people more than the theory of sufficient statistics and exponential families that leads to logistic regression.

Link Functions, Latent Variables (cont.)

Suppose we really have a linear model but don't get to observe its response. A variable that is not observable but is part of the description of a statistical model is called *latent*. Hence we are imagining a latent linear model with mean vector

$$\boldsymbol{\eta} = \mathbf{M}\boldsymbol{\beta}$$

The latent response for the i-th individual is

$$Z_i \sim \mathcal{N}(\eta_i, \sigma^2)$$

Link Functions, Latent Variables (cont.)

What we get to observe is

$$Y_i = \begin{cases} 1, & Z_i \geq 0 \\ 0, & Z_i < 0 \end{cases}$$

Then

$$Y_i \sim \text{Ber}(\mu_i)$$

where

$$\mu_i = \text{Pr}\left(Z_i \geq 0\right) = \text{Pr}\left(\frac{Z_i - \eta_i}{\sigma} \geq -\frac{\eta_i}{\sigma}\right) = 1 - \Phi\left(-\frac{\eta_i}{\sigma}\right) = \Phi\left(\frac{\eta_i}{\sigma}\right)$$

where Φ is the DF of the standard normal distribution.

Link Functions, Latent Variables (cont.)

Since η_i is a linear function of the regression coefficients, σ is not estimable. Multiply σ by c and divide all components of $\boldsymbol{\beta}$ by c, then the μ_i are unchanged. Hence we might as well set $\sigma = 1$.

To recap

$$
\begin{aligned}
Y_i &\sim \text{Ber}(\mu_i) \\
\mu_i &= \Phi(\eta_i) \\
\boldsymbol{\eta} &= \mathbf{M}\boldsymbol{\beta}
\end{aligned}
$$

This is called *probit regression* and the parameter $\boldsymbol{\eta}$ is called the *linear predictor*. It is also a GLM, fit by the R function `glm`.

Link Functions, Latent Variables (cont.)

How can we know that a variable Z_i that we cannot observe and have no reason to believe really exists is really exactly normal? We can't! Thus it seems plausible to replace the normal DF in

$$\mu_i = \Phi(\eta_i)$$

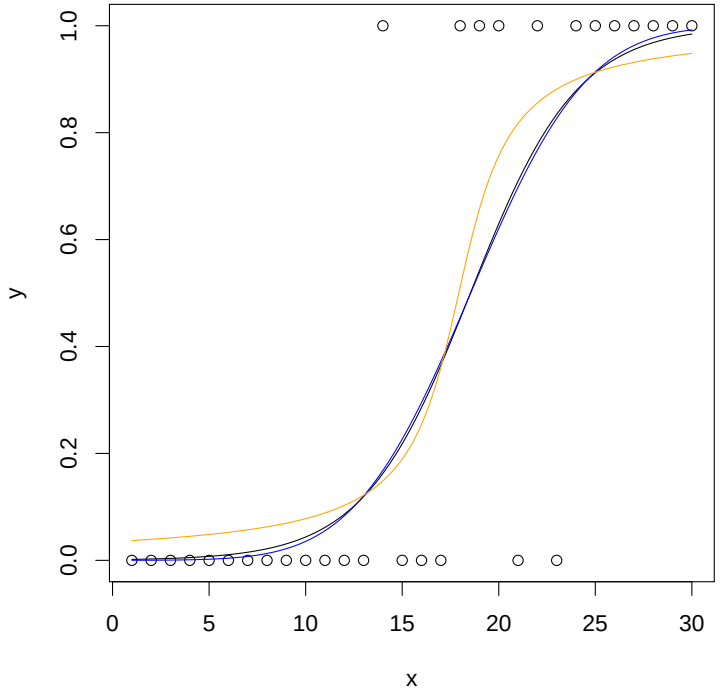
by other DF, Cauchy for example. That is also a GLM, fit by the R function `glm`.

Logistic regression with

$$\mu_i = \frac{1}{1 + \exp(-\eta_i)}$$

also falls in this latent variable scheme, since the right hand side is the DF of a distribution called *logistic*.

Link Functions, Latent Variables (cont.)



Same data as on slide 22 with logistic (black), probit (blue), and cauchit (orange) regression curves.

Link Functions, Latent Variables (cont.)

These GLM's are said to differ in their *link functions*, which map mean value parameter to linear predictor.

Unlike the case of the logit link, GLM using other link functions are not exponential families. They are called *curved exponential families*, which means smooth submodels of an exponential family. The saturated model is an exponential family, but probit or cauchit submodels are not.

This means the log likelihood is not concave and local maxima of the log likelihood are not necessarily global maxima. Nor do these models have low dimensional sufficient statistics (only the whole data vector is sufficient).

Link Functions, Latent Variables (cont.)

So you really have to like the story about unobservable latent variables only the signs of which are observable in order to use probit or cauchit link rather than the default logit link.

The logit link has much stronger statistical properties. It even has a competing story about unobservable unobservable latent variables only the signs of which are observable.

Nevertheless, some people do like these stories and do use these other links.

Categorical Data Analysis

Now we turn to the case where all of the data, response and predictors, are categorical.

We assume individuals classified into categories. The data used in the analysis are the counts of the number of individuals classified into each category.

If the category labels are univariate, the category counts are Y_i, we say we have a one-dimensional contingency table.

If the category labels are bivariate, the category counts are Y_{ij}, we say we have a two-dimensional contingency table.

And so forth.

Categorical Data Analysis

Three sampling models are widely used: Poisson, multinomial, and product multinomial.

In the *Poisson* sampling model, the category counts (the Y_i or Y_{ij} or ...) are assumed to be independent Poisson.

In the *multinomial* sampling model, the category counts (the Y_i or Y_{ij} or ...) are assumed to be jointly multinomial.

5101 Homework problems 9-10 and 9-11. If we start with Poisson sampling, then the multinomial sampling model arises when we condition on the total number of individuals (the sample size). If we start with multinomial sampling, then the Poisson sampling model arises when we make the sample size a Poisson random variable.

If the category counts are Y_{ij}, write

$$Y_{i+} = \sum_j Y_{ij}$$
$$Y_{+j} = \sum_i Y_{ij}$$

and so forth for higher dimensional indices. If we write out the Y_{ij} in an array in the usual way, then Y_{i+} are row sums and Y_{+j} are column sums.

If we start with Poisson sampling, then we get product multinomial sampling when we condition on a marginal. If we condition on the Y_{i+}, then the rows are independent multinomials. If we condition on the Y_{+j}, then the columns are independent multinomials. Similarly for higher dimensional tables.

It's called *product multinomial* because the joint distribution is the product of multinomials (product of multinomials for each row if we condition on Y_{i+} and so forth).

Multinomial sampling can be considered the special case where the indices are univariate Y_i and we condition on Y_+.

Pearson's Chi-Square Statistic

Before continuing with categorical data analysis, we must mention a historical anomaly.

Often, the likelihood ratio test is not used in categorical data analysis, not because there is anything wrong with it, but because of history.

Categorical data analysis was invented before maximum likelihood, hence before likelihood ratio tests. It was also invented before the general notion of linear models and before t and F distributions.

Pearson's Chi-Square Statistic (cont.)

Suppose we have categorical data and suppose multinomial sampling. Denote the data by Y_i, $i \in I$, denote the sample size by n, and denote the cell probabilities by p_i, so

$$E(Y_i) = np_i, \qquad i \in I \qquad (*)$$

Our abstract notation for the index set allows for any dimensions for the contingency table. If we have a four-dimensional contingency table, with data naturally denoted Y_{ijkl}, then we can consider i in $(*)$ to stand for four-tuples of indices.

Pearson's Chi-Square Statistic (cont.)

We wish to compare two nested models. The big model makes no restrictions on the cell probabilities other than that they are nonnegative and sum to one. The small model makes the cell probabilities functions $p_i(\boldsymbol{\theta})$ of a parameter $\boldsymbol{\theta}$. Suppose $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, and suppose $\tilde{\boldsymbol{\theta}}$ is another estimator asymptotically equivalent to the MLE, that is,

$$\tilde{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + o_p(n^{-1/2})$$

We assume the mapping $\boldsymbol{\theta} \mapsto \mathbf{p}(\boldsymbol{\theta})$ is differentiable and its Jacobian $\nabla \mathbf{p}(\boldsymbol{\theta})$ is always full rank.

Pearson's Chi-Square Statistic (cont.)

Assume the little model is correct. Then Pearson's chi-square statistic

$$X_n^2 = \sum_{i \in I} \frac{[Y_i - np_i(\tilde{\boldsymbol{\theta}})]^2}{np_i(\tilde{\boldsymbol{\theta}})}$$

is asymptotically equivalent to the likelihood ratio test statistic

$$G_n^2 = 2 \sum_{i \in I} Y_i \log \left(\frac{Y_i}{np_i(\hat{\boldsymbol{\theta}})} \right)$$

that is

$$X_n^2 - G_n^2 = o_p(1)$$

and the asymptotic distribution of both is chi-square with degrees of freedom the difference in dimensions of the models.

Pearson's Chi-Square Statistic (cont.)

The big model has dimension $k-1$, where k is the number of categories and the cardinality of I. The little model has dimension p, where p is the dimension of $\boldsymbol{\theta}$.

So the asymptotic distribution of X^2 or G^2 is chi-square with $k-1-p$ degrees of freedom.

If the null hypothesis is completely specified, a model containing only a single parameter, then we say $p = 0$, and the degrees of freedom is $k-1$.

Folklore

A bit of silly folklore that has no mathematics behind it says the chi-square approximation is o. k. if and only if

$$np_i(\tilde{\boldsymbol{\theta}}_n) \geq 5, \qquad i \in I$$

(the estimated expected value under the null hypothesis is at least 5 in each cell).

Asymptotics, of course, does not work that way. Adequacy of approximation is not an all or nothing thing. There are examples in the literature of violations of this rule of thumb both ways. Good asymptotic approximation with less than 5 expected in some cells, and bad asymptotic approximation with greater than 5 in all cells.

People like rules, even if they are arbitrary and unscientific.

One-Dimensional Contingency Table

Simulated data about rolls of a die

i	1	2	3	4	5	6
y_i	1038	964	975	983	1035	1005

If the die is fair, then all numbers are equally probable, so the null hypothesis is $p_i = 1/6$, $i = 1, \ldots, 6$.

One-Dimensional Contingency Table (cont.)

The R function `chisq.test` does chi-square tests for one and two dimensional contingency tables.

```
Rweb:>  out <- chisq.test(y)
Rweb:>  print(out)

        Chi-squared test for given probabilities

data:  y
X-squared = 4.904, df = 5, p-value = 0.4277
```

Pearson's chi-squared test statistic is 4.904 on 5 degrees of freedom. The P-value is $P = 0.4277$, which is not statistically significant. We accept the null hypothesis that the die is fair.

One-Dimensional Contingency Table (cont.)

The likelihood ratio test we do by hand

```
Rweb:>  Gsq <- 2 * sum(out$observed *
+    log(out$observed / out$expected))
Rweb:>  print(Gsq)
[1] 4.894725
Rweb:>  pchisq(Gsq, out$parameter, lower.tail = FALSE)
[1] 0.4288628
```

We get almost the same test statistic $X^2 = 4.904$ and $G^2 = 4.895$ and almost the same P-values $P = 0.4277$ for the X^2 test and $P = 0.4289$ for the G^2 test.

One-Dimensional Contingency Table (cont.)

More simulated data about rolls of a die

i	1	2	3	4	5	6
y_i	1047	1017	951	1004	952	1029

If the die is fair, then all numbers are equally probable, so the null hypothesis is $p_i = 1/6$, $i = 1, \ldots, 6$. Here we have a specific alternative hypothesis in mind: that the die is shaved on the six and one faces, making all the other faces smaller, which gamblers call six-ace flats.

One-Dimensional Contingency Table (cont.)

```
Rweb:>  out0 <- chisq.test(y)
Rweb:>  print(out0)

        Chi-squared test for given probabilities

data:  y
X-squared = 8.06, df = 5, p-value = 0.1530
```

The chi-square test with default arguments tests the null hypothesis of equal probabilities and the null hypothesis is accepted $P = 0.15$. The P-value is somewhat low but not statistically significant according to anyone's standards.

But this test is not about the six-ace flats hypothesis!

One-Dimensional Contingency Table (cont.)

We do the likelihood ratio test with equal probabilities for the null hypothesis and six-ace flats for the alternative.

```
Rweb:> nrolls <- sum(y)
Rweb:> phat0 <- rep(1 / 6, 6)
Rweb:> phat1 <- rep(NA, 6)
Rweb:> phat1[c(1, 6)] <- sum(y[c(1, 6)]) / 2 / nrolls
Rweb:> phat1[- c(1, 6)] <- sum(y[- c(1, 6)]) / 4 / nrolls
Rweb:> print(phat1)
[1] 0.1730 0.1635 0.1635 0.1635 0.1635 0.1730
Rweb:> Gsq <- 2 * sum(y * log(phat1 / phat0))
Rweb:> print(Gsq)
[1] 4.305331
Rweb:> pchisq(Gsq, 1, lower.tail = FALSE)
[1] 0.03799309
```

One-Dimensional Contingency Table (cont.)

When we actually do the likelihood ratio test with null hypothesis equal probabilities and alternative hypothesis six-ace flats, the null hypothesis is rejected ($P = 0.038$).

Moral of the story: you can't say anything about a hypothesis until you have done a test involving that hypothesis!

One-Dimensional Contingency Table (cont.)

The same three models can be fit using the R function `glm` assuming Poisson sampling

```
Rweb:> sixace <- factor(num %in% c(1, 6))
Rweb:> num <- factor(num)
Rweb:> out.big <- glm(y ~ num, family = poisson)
Rweb:> out.middle <- glm(y ~ sixace, family = poisson)
Rweb:> out.little <- glm(y ~ 1, family = poisson)
```

One-Dimensional Contingency Table (cont.)

The likelihood ratio tests (a. k. a., analysis of deviance) have the same test statistics and the same P-values for both Poisson and multinomial sampling.

```
Rweb:> anova(out.little, out.middle, out.big, test = "Chisq")
Analysis of Deviance Table

Model 1: y ~ 1
Model 2: y ~ sixace
Model 3: y ~ num
  Resid. Df Resid. Dev Df Deviance P(>|Chi|)
1         5     8.0945
2         4     3.7892  1   4.3053    0.0380
3         0   8.97e-14  4   3.7892    0.4353
```

Two-Dimensional Contingency Table

In a two-dimensional contingency table, the data are Y_{ij} and the cell probabilities p_{ij}. The test that is usually done has null hypothesis that the cell probabilities have multiplicative form

$$p_{ij} = \alpha_i \beta_j, \qquad \text{for all } i \text{ and } j$$

If the α_i are chosen to be nonnegative and sum to one, then the β_j have the same property and these are the marginal distributions of the random variables whose values are the row and column labels.

Two-Dimensional Contingency Table (cont.)

This this hypothesis is one of statistical independence if both row and column labels are random, that is, if we have multinomial sampling.

If we fix one marginal, that is, if we have product multinomial sampling, then we say the test is of homogeneity of proportions.

Either way, the chi-square test statistic is the same and the degrees of freedom is the same.

Two-Dimensional Contingency Table (cont.)

If there are r rows and c columns, then there are $k = rc$ cells in the contingency table. In order to figure the degrees of freedom, we need to know the number of parameters for the null hypothesis. Since we have the two constraints

$$\sum_{i=1}^{r} \alpha_i = 1$$

$$\sum_{j=1}^{c} \beta_j = 1$$

the null hypothesis has $(r-1)+(c-1)$ free parameters. Hence the degrees of freedom for the chi-square test is

$$rc - 1 - (r-1) - (c-1) = (r-1)c - (r-1) = (r-1)(c-1)$$

Two-Dimensional Contingency Table (cont.)

We need to know the MLE for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. The log likelihood is

$$\sum_{i=1}^{r}\sum_{j=1}^{c} y_{ij}\log(\alpha_i\beta_j) = \sum_{i=1}^{r} y_{i+}\log(\alpha_i) + \sum_{j=1}^{c} y_{+j}\log(\beta_j)$$

where, as before, the $+$ subscript indicates summation over an index.

In order to maximize the likelihood we must write it in terms of free parameters

$$\ell = \sum_{i=1}^{r-1} y_{i+}\log(\alpha_i) + y_{r+}\log\left(1 - \sum_{i=1}^{r-1}\alpha_i\right) + \sum_{j=1}^{c-1} y_{+j}\log(\beta_j) + y_{+c}\log\left(1 - \sum_{j=1}^{c-1}\beta_j\right)$$

Two-Dimensional Contingency Table (cont.)

Then

$$\begin{aligned}
\frac{\partial \ell}{\partial \alpha_i} &= \frac{y_{i+}}{\alpha_i} - \frac{y_{r+}}{1 - \sum_{k=1}^{r-1} \alpha_k} \\
&= \frac{y_{i+}}{\alpha_i} - \frac{y_{r+}}{\alpha_r} \\
\frac{\partial \ell}{\partial \beta_j} &= \frac{y_{+j}}{\beta_j} - \frac{y_{+c}}{1 - \sum_{k=1}^{c-1} \beta_k} \\
&= \frac{y_{+j}}{\beta_j} - \frac{y_{+c}}{\beta_c}
\end{aligned}$$

setting equal to zero and solving gives

$$\begin{aligned}
\hat{\alpha}_i &= \hat{\alpha}_r \cdot \frac{y_{i+}}{y_{r+}} \\
\hat{\beta}_j &= \hat{\beta}_c \cdot \frac{y_{+j}}{y_{+c}}
\end{aligned}$$

Two-Dimensional Contingency Table (cont.)

Now we again use the fact that the alphas and betas sum to one.

$$\sum_{i=1}^{r} \widehat{\alpha}_i = \widehat{\alpha}_r \cdot \frac{y_{++}}{y_{r+}} = 1$$

$$\sum_{j=1}^{c} \widehat{\beta}_j = \widehat{\beta}_c \cdot \frac{y_{++}}{y_{+c}} = 1$$

hence

$$\widehat{\alpha}_i = \frac{y_{i+}}{y_{++}}$$

$$\widehat{\beta}_j = \frac{y_{+j}}{y_{++}}$$

An example is on the computer examples web pages.