Accounting for Overdispersion

> ##
> ##### Plot Pearson residuals vs. fitted values
> ##
> nhat <- fitted( fit )
> plot( nhat, presid^2, xlab="Fitted mean response",
>      ylab="Squared Pearson residuals"
>     )
> abline( h=1, col="red", lwd=2 )
> sfit <- loess( presid^2 ~ nhat )
> lines( sort(sfit$x), sfit$fitted[order(sfit$x)], col="blue", lwd=2 )
> abline(h=phihat, lty=2, col="red", lwd=2)

Fitted mean response
Squared Pearson residuals

• Again, it looks as though the smoother is consistently above the y=1 line, indicating overdispersion

IX-21 D. Gillen, STAT 211

Maximum Likelihood
Estimation
Asymptotic Distribution
of the Score and MLE
Asymptotic Distribution
of Test Statistics
Testing a subset of the
department space
Maximum Likelihood Estimation

Why?

- **Note:** One very commonly used form of an estimating equation (and the primary one we will use in this class) is the likelihood equation.

- Many other estimating equations draw from underlying likelihood theory as well, hence it is worthwhile to examine the derivations of likelihood theory.
Maximum Likelihood Estimation

**Definitions**

- **Model:** Suppose we have data \((Y_i, \tilde{X}_i), i = 1, \ldots, n\) with

  1. \(Y_i \sim f_i(y; \tilde{X}_i, \tilde{\theta})\) independent with \(f_i\) known densities (probability mass functions) except for the unknown parameter \(\tilde{\theta} \in \Theta \subset \mathbb{R}^p\), with \(\Theta\) open.

  2. The **likelihood function** for \(\tilde{\theta}\) is

        \[
        L(\tilde{\theta}) = \prod_{i=1}^{n} f_i(Y_i; \tilde{X}_i, \tilde{\theta})
        \]

  3. The **log likelihood function** is

        \[
        \ell(\tilde{\theta}) = \log L(\tilde{\theta}) = \sum_{i=1}^{n} \log f_i(Y_i; \tilde{X}_i, \tilde{\theta}) = \sum \ell_i(\tilde{\theta})
        \]
Maximum Likelihood Estimation

Definitions

4. The (efficient) score function is

\[ \hat{U}(\bar{\theta}) = \left( \frac{\partial \ell(\bar{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ell(\bar{\theta})}{\partial \theta_p} \right)^T = \left( U_1(\bar{\theta}), \ldots, U_p(\bar{\theta}) \right)^T \]

where

\[ \hat{U}(\bar{\theta}) = \sum_{i=1}^{n} \frac{\partial \ell_i(\bar{\theta})}{\partial \theta} = \sum_{i=1}^{n} \hat{U}_i(\bar{\theta}) = \sum_{i=1}^{n} \left( U_{1i}(\bar{\theta}), \ldots, U_{pi}(\bar{\theta}) \right)^T \]
Maximum Likelihood Estimation

Definitions

5. Fisher’s information matrix is

\[ \mathcal{I}(\hat{\theta}) = -E \left[ \frac{\partial^2}{\partial \hat{\theta}_j \partial \hat{\theta}_k} \ell(\hat{\theta}) \right] = -E \left[ \frac{\partial \hat{U}(\hat{\theta})}{\partial \hat{\theta}^T} \right] \]

\[ \mathcal{I}_{jk}(\hat{\theta}) = -E \left[ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell(\hat{\theta}) \right] = -\sum_{i=1}^{n} E \left[ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell_i(\hat{\theta}) \right] \]

\[ = -\sum_{i=1}^{n} \mathcal{I}_{jki}(\hat{\theta}) \]
Maximum Likelihood Estimation

Definitions

6. The observed information matrices are (used most often)

\[
\mathbf{I}(\tilde{\theta}) = - \left[ \frac{\partial^2}{\partial \tilde{\theta} \partial \tilde{\theta}^T} \ell(\tilde{\theta}) \right]
\]

\[
l_{jk}(\tilde{\theta}) = - \left[ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell(\tilde{\theta}) \right] = - \sum_{i=1}^{n} \left[ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \ell_i(\tilde{\theta}) \right]
\]

\[
= - \sum_{i=1}^{n} l_{jki}(\tilde{\theta})
\]

or (used less frequently)

\[
\mathbf{I}^*(\tilde{\theta}) = \sum_{i=1}^{n} \mathbf{\tilde{U}}_i(\tilde{\theta}) \mathbf{\tilde{U}}_i^T(\tilde{\theta})
\]

\[
l_{jk}^*(\tilde{\theta}) = \sum_{i=1}^{n} U_{ji}(\tilde{\theta}) U_{ki}(\tilde{\theta})
\]
Maximum Likelihood Estimation

Definitions

7. The maximum likelihood estimator $\hat{\theta}$ is defined by the estimating equation
   $$ \bar{U}(\hat{\theta}) = \bar{0} $$

8. The likelihood ratio is
   $$ R(\hat{\theta}) = \frac{L(\theta)}{L(\hat{\theta})} $$
Maximum Likelihood Estimation

Regularity Assumptions

- The general theoretical results regarding maximum likelihood estimates condition upon general *regularity assumptions*:

1. The true value of \( \hat{\theta} \) is \( \theta_0 \).

2. (identifiability) \( f_i(y; \tilde{X}_i, \theta_1) = f_i(y; \tilde{X}_i, \theta_2) \) for all \( y \) implies \( \theta_1 = \theta_2 \)

3. (common support) \( A_i = \{ y : f_i(y; \tilde{X}_i, \tilde{\theta}) > 0 \} \) is independent of \( \tilde{\theta} \)
Maximum Likelihood Estimation

Regularity Assumptions

4. (interchange of differentiation with respect to \( \vec{\theta} \) and integration with respect to \( y \))

\[
\frac{\partial}{\partial \vec{\theta}} \int f_i(y; \vec{X}_i, \vec{\theta}) \, dy = \int \frac{\partial}{\partial \vec{\theta}} f_i(y; \vec{X}_i, \vec{\theta}) \, dy
\]

\[
\frac{\partial^2}{\partial \vec{\theta} \partial \vec{\theta}^T} \int f_i(y; \vec{X}_i, \vec{\theta}) \, dy = \int \frac{\partial^2}{\partial \vec{\theta} \partial \vec{\theta}^T} f_i(y; \vec{X}_i, \vec{\theta}) \, dy
\]

5. \( \frac{\partial^3}{\partial \vec{\theta}^3} f_i(y; \vec{X}_i, \vec{\theta}) \) is bounded in some open neighborhood of \( \vec{\theta}_0 \)
Maximum Likelihood Estimation

Regularity Assumptions

6. As \( n \to \infty \), the Fisher’s information matrix \( \mathcal{I}(\vec{\theta}) \) is of constant rank in some open neighborhood of \( \vec{\theta}_0 \), and the eigenvalues of \( \mathcal{I}(\vec{\theta}) \) all grow to \( \infty \). (As \( n \to \infty \), for all \( j, k \in \{1, \ldots, p\} \) and all \( 1 \leq i \leq n \)

\[
\frac{\mathcal{I}_{jki}(\vec{\theta})}{\mathcal{I}_{jk}(\vec{\theta})} \to 0
\]
Asymptotic Distribution of the Score and MLE

Proposition 1

Under the above regularity conditions,

1. As \( n \to \infty \), with probability 1, \( \hat{\theta} \) exists and \( \hat{\theta} \to_p \theta_0 \)

2. \( E \left[ \tilde{U}(\theta_0) \right] = \sum_{i=1}^{n} E[\tilde{U}_i(\theta_0)] = 0 \)

3. \( \text{Cov} \left[ \tilde{U}(\theta_0) \right] = I(\theta_0) \)

4. As \( n \to \infty \), the distribution of \( \tilde{U}(\theta_0) \) is approximately

\[
\tilde{U}(\theta_0) \sim \mathcal{N}_p \left( 0, I(\theta_0) \right)
\]

5. As \( n \to \infty \), the distribution of \( \hat{\theta} \) is approximately

\[
\hat{\theta} \sim \mathcal{N}_p \left( \theta_0, I^{-1}(\theta_0) \right)
\]
Notes on the behavior of the MLE

1. The fact that the maximum likelihood estimates are consistent and have an asymptotic variance related to Fisher’s information can be related to the Cramér-Rao lower bound.

   ▶ In small samples, estimates that have such behavior would in some sense be optimal.

   ▶ In asymptotic distributions, it is possible to beat that bound at a selected alternative, but that is of little practical use. Hence, the maximum likelihood estimates can be viewed as being asymptotically efficient.
Asymptotic Distribution of Test Statistics

Facts about the multivariate normal distribution

- Let \( \mathbf{y} \) be a \( n \times 1 \) random vector with joint density function

\[
f(\mathbf{y}) = (2\pi)^{-1/2} |\Sigma| e^{-(\mathbf{y} - \mu)^T \Sigma^{-1} (\mathbf{y} - \mu)}
\]

where \( |\Sigma| = \text{det}(\Sigma) \), \( \mu \) is a \( n \times 1 \) vector of parameters, and \( \Sigma \) is an \( n \times n \) symmetric positive definite matrix of parameters.

- We say \( \mathbf{y} \) has a multivariate normal distribution with mean \( \mu \) and variance \( \Sigma \)
Accounting for Overdispersion

## Plot Pearson residuals vs. fitted values

```r
nhat <- fitted( fit )
plot( nhat, presid^2, xlab="Fitted mean response",
ylab="Squared Pearson residuals" )
abline( h=1, col="red", lwd=2 )
sfit <- loess( presid^2 ~ nhat )
lines( sort(sfit$x), sfit$fitted[order(sfit$x)], col="blue", lwd=2 )
abline(h=phihat, lty=2, col="red", lwd=2)
```

10 20 30 40
0 1 2 3 4 5 6

Fitted mean response
Squared Pearson residuals

Again, it looks as though the smoother is consistently above the y=1 line, indicating overdispersion

Asymptotic Distribution of Test Statistics

### Facts about the multivariate normal distribution

- **Fact (Quadratic Forms)** If \( y \) has a \( n \) dimensional multivariate normal distribution with mean \( \mu \) and variance \( \Sigma \) then

\[
(y - \mu)^T \Sigma^{-1} (y - \mu) \sim \chi^2_n
\]
Asymptotic Distribution of Test Statistics

Theorem 2

1. (Score or Rao statistic) Under the null hypothesis $H_0 : \vec{\theta} = \vec{\theta}_0$,

   $\bar{U}^T(\vec{\theta}_0) I^{-1}(\vec{\theta}_0) \bar{U}(\vec{\theta}_0) \to_d \chi^2_p$

   $\bar{U}^T(\vec{\theta}_0) I^{-1}(\vec{\theta}_0) \bar{U}(\vec{\theta}_0) \to_d \chi^2_p$

   $\bar{U}^T(\vec{\theta}_0) I^*^{-1}(\vec{\theta}_0) \bar{U}(\vec{\theta}_0) \to_d \chi^2_p$
Asymptotic Distribution of Test Statistics

Theorem 2

2. (Wald statistic) Under the null hypothesis $H_0 : \hat{\theta} = \theta_0$,

$$
(\hat{\theta} - \theta_0)^T \mathcal{I}(\theta_0)(\hat{\theta} - \theta_0) \to_d \chi^2_p
$$

$$
(\hat{\theta} - \theta_0)^T \mathcal{I}(\hat{\theta})(\hat{\theta} - \theta_0) \to_d \chi^2_p
$$

$$
(\hat{\theta} - \theta_0)^T \mathcal{I}(\hat{\theta})^*(\theta_0)(\hat{\theta} - \theta_0) \to_d \chi^2_p
$$
Asymptotic Distribution of Test Statistics

**Theorem 2**

3. (Likelihood ratio statistic) Under the null hypothesis $H_0 : \bar{\theta} = \bar{\theta}_0$,

\[
-2 \log R(\bar{\theta}_0) = -2(\ell(\bar{\theta}_0) - \ell(\hat{\theta})) \rightarrow_d \chi^2_p
\]
Accounting for Overdispersion

> ##
> #### Plot Pearson residuals vs. fitted values
> ##
> nhat <- fitted( fit )
> plot( nhat, presid^2, xlab="Fitted mean response",
> ylab="Squared Pearson residuals" )
> abline( h=1, col="red", lwd=2 )
> sfit <- loess( presid^2 ~ nhat )
> lines( sort(sfit$x), sfit$fitted[order(sfit$x)], col="blue", lwd=2 )
> abline(h=phihat, lty=2, col="red", lwd=2)

• Again, it looks as though the smoother is consistently above the y=1 line, indicating overdispersion

IX-21 D. Gillen, STAT 211

Maximum Likelihood Estimation

Asymptotic Distribution of the Score and MLE

Asymptotic Distribution of Test Statistics

Testing a subset of the parameter space
### Asymptotic Distribution of Test Statistics

#### Notes on the behavior of the tests

1. The parallels between the likelihood ratio test and the Neyman-Pearson lemma’s most powerful test are readily apparent. Hence, tests based on any of the three likelihood based statistics can be expected to be ‘most powerful’ in some sense.
Testing a subset of the parameter space

Note

- Theorem 2 considered only the case of testing the value of the entire $p$ dimensional parameter vector. It is rare that we do this. Hence, we need to modify our statistics for the case of testing a composite hypothesis about a subset of the parameter vector.
Testing a subset of the parameter space

Theorem 3

- Consider a partition of the parameter vector \( \vec{\theta} = (\vec{\theta}_1, \vec{\theta}_2)^T \), where the dimension of \( \vec{\theta}_1 \) is \( r \), and let \( \hat{\vec{\theta}} = (\hat{\theta}_1, \hat{\theta}_2)^T \) be the partitioned maximum likelihood estimate under the alternative.

- Let \( \mathcal{I}_{11}, \mathcal{I}_{12}, \mathcal{I}_{21}, \) and \( \mathcal{I}_{22} \), be the analogous partitions of Fisher's information matrix, with similarly defined partitions of the observed information matrices.

- We consider testing hypotheses of the form \( H_0 : \vec{\theta}_1 = \vec{\theta}_{10} \).
Testing a subset of the parameter space

Theorem 3

1. (MLE) Under $H_0$, the maximum likelihood estimate of $\hat{\theta}$ is $\hat{\theta}_0 = (\hat{\theta}_{10}, \hat{\theta}_{20})^T$, where $\hat{\theta}_{20}$ is found from the estimating equation

$$\frac{\partial \ell(\theta)}{\partial \theta_2} \bigg|_{\theta = \hat{\theta}_0} = 0$$
Testing a subset of the parameter space

**Theorem 3**

2. (Score or Rao statistic) Under the null hypothesis

\[
\tilde{U}^T(\hat{\theta}_0)\mathcal{I}^{-1}(\hat{\theta}_0)\tilde{U}(\hat{\theta}_0) \rightarrow_d \chi^2_r
\]

The observed information matrices can be used in place of Fisher’s information.

- **Note:** We could have expressed these statistics solely in terms of the partitioned information matrices, because \( p - r \) elements of the score vectors are zero. However, I find it easiest to just use the same formulas and only adjust the degrees of freedom in the chi square distribution.
Testing a subset of the parameter space

**Theorem 3**

3. (Wald statistic) Under the null hypothesis

\[
(\hat{\theta}_1 - \hat{\theta}_{10})^T \mathcal{I}_{11.2}(\hat{\theta})(\hat{\theta}_1 - \hat{\theta}_{10}) \rightarrow_d \chi_r^2
\]

where, again, we can use observed information matrices in place of Fisher’s information, and we can use the estimate of \( \hat{\theta}_1 \) in whichever information function we use.

- **Note 1**: \( \mathcal{I}_{11.2} = \mathcal{I}_{11} - \mathcal{I}_{12}\mathcal{I}_{22}^{-1}\mathcal{I}_{21} \) is the inverse of the corresponding partition of the inverse of the Fisher’s information matrix.

- **Note 2**: This statistic is derived solely from the marginal distribution of \( \hat{\theta}_1 \) using the asymptotic distribution of \( \hat{\theta} \) derived in (Theorem 1). (multivariate normal theory)
Testing a subset of the parameter space

**Theorem 3**

4. (Likelihood ratio statistic) Under the null hypothesis

\[-2 \log(R(\hat{\theta}_0)) \to_d \chi^2_r\]
Testing a subset of the parameter space

Notes

1. The score, Wald, and likelihood ratio statistics are asymptotically equivalent, however, they are not equally well behaved in small samples.

   - The Wald statistic is less stable under some parameterizations of $\tilde{\theta}$.

   - The score statistic is invariant under reparameterizations, providing Fisher’s information matrix is used instead of an observed information matrix.

   - The likelihood ratio statistic is invariant under reparameterizations of $\tilde{\theta}$.
2. The three statistics are not all equally easy to compute.

2.1 The score statistic does not require finding the maximum likelihood estimate $\hat{\theta}$, though in the case of a composite hypothesis we do have to find $\hat{\theta}_0$.

Furthermore, if we use the observed information matrix $I^*$, we do not need to find the second derivative of the log likelihood function. As a general rule, however, $I^*$ behaves worse than $I$ in small samples.
Notes

2. The three statistics are not all equally easy to compute.

2.2 The Wald statistic is appealing because it is based on an estimate of the parameter. Hence direct inferential statements can be made about the estimate. It does require finding the estimate, and it tends to be less stable in small samples.

2.3 The likelihood ratio statistic tends to be the most stable of the three statistics in small samples, but, in the case of a composite hypothesis, it does require fitting two models to find $\hat{\theta}$ and $\hat{\theta}_0$. 