PubH 7405 Biostat Regression - Fall 2005

Baolin Wu, Ph.D.
Division of Biostatistics
School of Public Health
University of Minnesota
baolin@biostat.umn.edu

Lecture 1 - September 7, 2005
Instructor: Dr. Baolin Wu, baolin@biostat.umn.edu
TA: Zang Xiong, zangx@biostat.umn.edu; Qun Shi, qunshi@biostat.umn.edu; Yanni Zhu, yannizhu@biostat.umn.edu
Website: http://www.biostat.umn.edu/~baolin/teaching/ph7405-05F
Notes: http://www.biostat.umn.edu/~baolin/teaching/ph7405-05F/ScheduleAndLectureNotes.html
Course Grade

1. Eleven homeworks: 50%

2. Inclass Midterm Exam: 20%

3. Inclass Final Exam: 30%
Other Information

Make sure to check out the CourseWebsite often. Updated lecture notes, developed computer programs in the class, TA and Lab sessions, resources on learning UNIX and SAS/R programming, relevant papers on some interesting topics, and some technical supplements to the course notes will be posted on the website.

I will mainly use R in the class. You can use whatever software you feel comfortable, but I won’t provide any help on them except SAS and R.
Computing Information

Check out the course computing information website: http://www.biostat.umn.edu/~baolin/teaching/ph7405-05F/ComputingInfo.html. Some codes developed in the class will be posted in http://www.biostat.umn.edu/~baolin/teaching/ph7405-05F/ComputerCodes.html.
Course Outline

We will cover Chapter 1 to Chapter 13 of the textbook. Check out the ScheduleAndLectureNotes on the CourseWebsite for tentative course lecture schedule.

If time permits, I will discuss some recent developments in linear regression. These are quite new and very interesting results which have received lots of research interests recently. They won’t be covered in any exam or homeworks.
Some Basic Results in Probability & Statistics

• Linear Algebra

• Probability

• Random Variables

• Common Statistical Distributions

• Statistical Estimation

• Statistical Inference about Normal Distributions
Linear Algebra

• Summation and Product Operators

\[
\sum_{j=1}^{p} X_j = X_1 + X_2 + \cdots + X_p, \quad \prod_{i=1}^{n} y_i = y_1 \cdot y_2 \cdots y_n,
\]

\[
\sum_{i=1}^{n} \sum_{j=1}^{p} x_{ij} = \sum_{i=1}^{n} (x_{i1} + \cdots + x_{ip}) = x_{11} + \cdots + x_{1p} + \cdots + x_{n1} + \cdots + x_{np}.
\]

• Matrix: a rectangular display and organization of data. You can treat matrix as data with two subscripts, e.g. \( x_{ij} \), the first subscript is row index and the second the column index. We note the matrix as \( X_{n \times p} = (x_{ij}) \), and call it a \( n \) by \( p \) matrix.
Matrix Operations

- Identity matrix $I$: square ($n = p$), diagonal equal to 1 and 0 elsewhere.

- Summation: element-wise summation

- Product: for $X_{n \times p} = (x_{ij})$, $B_{p \times K} = (\beta_{jk})$, their product $Y = XB = (y_{ik})$ is a $n$ by $K$ matrix with $y_{ik} = \sum_{j=1}^{p} x_{ij} \beta_{jk}$.

- Inverse: the product of a matrix $X$ and its inverse $X^{-1}$ is identity matrix.

- Trace: for square matrix $X_{n \times n}$, $tr(X) = \sum_{i=1}^{n} x_{ii}$.

- Transpose: reverse the row and column index. So $t(X)_{ij} = x_{ji}$. 
Some Notes about Matrix

- When doing matrix product $XB$, always make sure the number of columns of $X$ and rows of $B$ are equal.

- Matrix product has orders, $XB$ and $BX$ are different. For inverse matrix we have $XX^{-1} = X^{-1}X = I$.

- Only square matrix has inverse and trace, and $tr(XB) = tr(BX)$. If $X^{-1} = t(X)$, we call $X$ an orthogonal matrix.
Probability

- Sample space $\Omega$, events (sets) $A, B$

- Basic rules
  
  \[
  \begin{align*}
  \Pr(\Omega) &= 1, \Pr(\emptyset) = 0 \\
  \Pr(A \cup B) &= \Pr(A) + \Pr(B) - \Pr(A \cap B) \\
  \Pr(A \cap B) &= \Pr(A) \Pr(B|A) = \Pr(B) \Pr(A|B)
  \end{align*}
  \]

- Complementary events: \( \Pr(\bar{A}) = 1 - \Pr(A) \)
Random Variables

- A mapping (function) $Y$ from sample space $\Omega$ to $\mathbb{R}^1$. For continuous random variables, the distribution and density functions are defined as $F(y) = \Pr(Y \leq y)$, $f(y) = \lim_{\epsilon \to 0} \{F(y + \epsilon) - F(y)\} / \epsilon$.

- Quantile: the inverse of distribution function. $F(q_\alpha) = \alpha$, where $q_\alpha$ is the $\alpha$-th quantile.

- Joint, Marginal, and Conditional Probability Distributions
  
  $\Pr(y_i) = \sum_j \Pr(y_i, z_j)$, $\Pr(y_i | z_j) = \Pr(y_i, z_j) / \Pr(z_j)$

- Expectation: $E(Y) = \sum_i y_i \Pr(y_i) = \int y f(y) dy$ (Linear Operator)
• Variance: \( \text{Var}(Y) = E[Y - E(Y)]^2 = E(Y^2) - E(Y)^2 \)

• Covariance: \( \text{Cov}(Y, Z) = E\{[Y - E(Y)][Z - E(Z)]\} = E(YZ) - E(Y)E(Z) \)

• Correlation: \( \rho(Y, Z) = \frac{\text{Cov}(Y, Z)}{\sqrt{\text{Var}(Y)\text{Var}(Z)}} \)

• Independent Random Variables: \( \iff \Pr(y_i, z_j) = \Pr(y_i)\Pr(z_j) \longrightarrow \text{Cov}(Y, Z) = 0. \)

• Central Limit Theorem: If \( Y_1, \cdots, Y_n \) are \( n \) iid (independent and identically distributed) random variables with mean \( \mu \) and variance \( \sigma^2 \), then the sample mean \( \bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i \) is approximately normally distributed when the sample size \( n \) is reasonably large, with mean \( \mu \) and variance \( \sigma^2/n \).
Common Statistical Distributions

- Normal Distribution $N(\mu, \sigma^2)$: density $\frac{1}{\sqrt{2\pi\sigma}} \exp\left\{ -\frac{(y-\mu)^2}{2\sigma^2} \right\}$, where $\mu, \sigma^2$ are the mean and variance for $Y$. We have $E(Y) = \mu$, $E(Y - \mu)^2 = \sigma^2$, $E(Y - \mu)^4 = 3\sigma^4$. More generally

$$E(Y - \mu)^{2k-1} = 0, \ E(Y - \mu)^{2k} = \sigma^{2k}(2k - 1)!!,$$

(1)

where $(2k - 1)!! = (2k - 1) \times (2k - 3) \times \cdots \times 3 \times 1$.

- Linear functions of normal random variables are still normal. $\frac{Y - \mu}{\sigma}$ is standard normal with mean 0 and variance 1. $\phi(\cdot)$ and $\Phi(\cdot)$ are commonly used to code the standard normal density and distribution functions.
• **χ² Random Variable:** \( \chi^2(n) = \sum_{i=1}^{n} z_i^2 \), where \( z_i \) are iid standard normal random variables and \( n \) is called the degree of freedom. We have

\[
E\{\chi^2(n)\} = n, \ Var\{\chi^2(n)\} = 2n.
\]

• **t Random Variable:** \( t(n) = \frac{z}{\sqrt{\chi^2(n)/n}} \), where \( z \) is standard normal and independent of \( \chi^2(n) \).

• **F Random Variable:** \( F(n, m) = \frac{\chi^2(n)/n}{\chi^2(m)/m} \), where \( \chi^2(n) \) and \( \chi^2(m) \) are two independent \( \chi^2 \) random variables.
Common Distribution Density Functions

Standard Normal Density
\[ \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \]

\[ y \]

\[ 0.0 \quad 0.2 \quad 0.4 \]

\[ 0 \quad 1 \quad 2 \quad 3 \]

\[ y \]

\[ \Gamma(n/2) \]

\[ (n+1)/2 \]

\[ \frac{\Gamma((n+1)/2)}{\sqrt{n\pi} \Gamma(n/2)} (1 + y^2/2)^{-(n+1)/2} \]

\[ t_n \text{ Density} \]

\[ y \]

\[ 0.0 \quad 0.2 \]

\[ -4 \quad -2 \quad 0 \quad 2 \quad 4 \]

\[ y \]

\[ F_{n,m} \text{ Density} \]

\[ y \]

\[ 0.0 \quad 0.3 \quad 0.6 \]

\[ 0 \quad 5 \quad 10 \quad 15 \]

\[ y \]

\[ \frac{\Gamma((n+m)/2)(n/m)^{n/2}}{\Gamma(n/2) \Gamma(m/2)} y^{n/2-1} (1 + ny/m)^{-(n+m)/2} \]

\[ \chi^2_n \text{ Density} \]

\[ y \]

\[ 0.0 \quad 0.10 \]

\[ 0 \quad 5 \quad 10 \quad 15 \]
Statistical Estimations

• Estimator Properties: an estimator $\hat{\theta}$ is a function of the sample observations $(y_1, \cdots, y_n)$, which estimates some parameter $\theta$ associated with the distribution of $Y$.

• Maximum Likelihood Estimation

• Least Squares Estimation
Estimator Properties

- **unbiased**: $E(\hat{\theta}) = \theta$

- **consistent**: $\lim_{n \to \infty} \Pr(|\hat{\theta} - \theta| \geq \epsilon) = 0, \forall \epsilon > 0$

- **sufficient**: $\Pr(y_1, \cdots, y_n|\hat{\theta})$ doesn’t dependent on $\theta$.

- **minimum variance estimator**: $\text{Var}(\hat{\theta}) \leq \text{Var}(\hat{\theta}^\star), \forall \hat{\theta}^\star$
Maximum Likelihood Estimators (MLE)

Maximum Likelihood is a general method of finding estimators. Suppose \((y_1, \cdots, y_n)\) are \(n\) iid samples from distribution \(f(y, \theta)\) with parameter \(\theta\). The “probability of observing these samples” is

\[
L(\theta) = \prod_{i=1}^{n} f(y_i, \theta),
\]

which is called the likelihood function. Maximize \(L(\theta)\) with respect to \(\theta\) yields the MLE

\[
\hat{\theta} = \arg \max_{\theta} L(\theta).
\]

Under very general conditions, MLEs are consistent and sufficient.
MLE for Normal Distributions

Suppose \((y_1, \cdots, y_n)\) are iid samples from normal distribution \(N(\mu, \sigma^2)\). What’s the MLE for parameters \(\mu\) and \(\sigma^2\)?

\[
L(\mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(y_i - \mu)^2}{2\sigma^2} \right\} = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp \left\{ -\sum_{i=1}^{n} \frac{(y_i - \mu)^2}{2\sigma^2} \right\}
\]

Maximize \(L\) is equivalent to maximize \(\log(L)\), the “Log Likelihood”, and we can easily get the following MLE:

\[
\hat{\mu} = \frac{\sum_{i=1}^{n} y_i}{n} = \bar{y}, \quad \hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}
\] (2)
**Property of MLE for Normal Distributions**

\[ E(\hat{\mu}) = \mu, \ Var(\hat{\mu}) = \sigma^2/n, \] and \( \hat{\mu} \) is normally distributed, \( N(\mu, \sigma^2/n) \).

For \( \epsilon > 0 \) we have

\[ \Pr(|\hat{\mu} - \mu| > \epsilon) = 2\Phi\left(-\frac{\epsilon}{\sigma/\sqrt{n}}\right) \]

\[ E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2, \ Var(\hat{\sigma}^2) = \frac{2(n-1)}{n^2}\sigma^4, \] and \( n\hat{\sigma}^2/\sigma^2 \) has chi-square distribution \( \chi^2(n-1) \).

From \( \Pr(\chi^2(\alpha/2, n-1) < n\hat{\sigma}^2/\sigma^2 \leq \chi^2(1-\alpha/2, n-1)) = \alpha \), we can
get the $1 - \alpha$ **Confidence Interval** for $\sigma^2$

$$\left( \frac{n\hat{\sigma}^2}{\chi^2(1 - \alpha/2, n - 1)}, \frac{n\hat{\sigma}^2}{\chi^2(\alpha/2, n - 1)} \right).$$

Furthermore

$$\frac{(\hat{\mu} - \mu)/\left(\sigma/\sqrt{n}\right)}{\sqrt{(n\hat{\sigma}^2/\sigma^2)/(n - 1)}} = \sqrt{n - 1} \frac{\hat{\mu} - \mu}{\hat{\sigma}} \sim t_{n-1}. \quad (3)$$

Similarly we can get the $1 - \alpha$ **Confidence Interval** for $\mu$

$$\left( \hat{\mu} - \frac{t(1 - \alpha/2, n - 1)\hat{\sigma}}{\sqrt{n - 1}}, \hat{\mu} + \frac{t(1 - \alpha/2, n - 1)\hat{\sigma}}{\sqrt{n - 1}} \right)$$
Least Squares Estimators (LS)

LS is another general method of finding estimators. The sample observations are assumed to be of the form \( y_i = f_i(\theta) + \epsilon_i, \ i = 1, \cdots, n \), where \( f_i(\theta) \) is a known function of the parameter \( \theta \) and the \( \epsilon_i \) are random variables, usually assumed to have expectation \( E(\epsilon_i) = 0 \).

LS estimators are obtained by minimizing the sum of squares

\[
Q = \sum_{i=1}^{n} [y_i - f_i(\theta)]^2 .
\]

Here \( \mathcal{L}_2 \) distance is used, more generally \( \mathcal{L}_q \) distance can be considered.
Linear Regression: Maximum Likelihood Approach to Linear Models with Normal Errors

Assume $Y$ is normally distributed with mean $\mu$ and variance $\sigma^2$. And suppose we can model $\mu$ as a linear function of covariates $X$ (you can imagine the possibility of extending to more complex models; or treat linear model as the first order Taylor approximation to the underlying true model.)

$$\mu = X\beta, \ Y \sim N(X\beta, \sigma^2),$$

or equivalently we can write the model as

$$Y = X\beta + \epsilon, \ \epsilon \sim N(0, \sigma^2).$$

Here the parameters are $(\beta, \sigma)$. After observing $n$ samples we can obtain
MLEs for parameters

$$\max_{\beta,\sigma} \frac{1}{(2\pi \sigma^2)^{n/2}} \exp \left\{ -\frac{\sum_{i=1}^{n} (y_i - x_i \beta)^2}{2\sigma^2} \right\},$$

which leads to the ordinary linear regression model (using matrix notation)

$$\hat{\beta} = \arg \min_{\beta} ||Y - X\beta||^2,$$

and the MLE for $\sigma^2$ is

$$\hat{\sigma}^2 = \frac{||Y - X\hat{\beta}||^2}{n}.$$

You may guess that this is a biased estimator. More about linear regression inference later in the semester!
Statistical Inference about Normal Distribution

• One Sample: \((y_1, \cdots, y_n) \sim N(\mu, \sigma^2)\)
  
  1. Mean: \(\mu?\) t-test
  2. Variance: \(\sigma^2?\) \(\chi^2\)-test

• Two Sample: \((y_1, \cdots, y_n) \sim N(\mu_1, \sigma_1^2)\) independent of \((x_1, \cdots, x_m) \sim N(\mu_2, \sigma_2^2)\)
  
  1. Mean: \(\mu_1 = \mu_2?\) t-test
  2. Variance: \(\sigma_1 = \sigma_2?\) F-test
Hypothesis Testing

Hypothesis testing is concerned with the state of population, which is usually characterized by some parameters, e.g. we’re interested in testing the mean and variance of a normal distribution. There are several components

1. Null hypothesis $H_0$: the postualted “default” state (value)

2. Alternative hypothesis $H_a$: “abnormal” state

3. Test statistics: the empirical information from observed data (usually some functions of data)

4. Rejection rules: Type-I error $\alpha = \Pr(\text{reject } H_0|H_0 \text{ true})$ and Type-II error $1 - \beta = \Pr(\text{don’t reject } H_0|H_0 \text{ false})$
P-value

P-value for a hypothesis test is defined as

the probability that the sample outcome is more extreme than the observed one when $H_0$ is true.

Large P-values support $H_0$ while small P-values support $H_a$. A test can be carried out by comparing the P-value with the specified type-I error $\alpha$. If P-value $< \alpha$, then $H_0$ is rejected.

Note that the calculation of P-value depends on the rejection rules: the selection of rejection regions, which defines what is “more extreme”.

P-value is usually a function of the test statistic. It is just another test statistic and has uniform distribution when $H_0$ is true.
One Sample Inference about Normal Distribution

- Test $H_0 : \sigma = \sigma_0$ vs $H_a : \sigma \neq \sigma_0$, under $H_0$, $T = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{\sigma_0^2} \sim \chi^2(n - 1)$.

  Control Type-I error at level $\alpha$, rejection regions are constructed as \{\,$\chi^2(\alpha/2, n - 1), \chi^2(1 - \alpha/2, n - 1)$\,\}.

- Test $H_0 : \mu = \mu_0$ vs $H_a : \mu \neq \mu_0$, according to (2) and (3) under $H_0$, $T = \sqrt{n - 1} \frac{\hat{\mu} - \mu_0}{\hat{\sigma}} \sim t_{n-1}$.

  Control Type-I error at $\alpha$, choose rejection regions as \{\,$t(\alpha/2, n - 1), t(1 - \alpha/2, n - 1)$\,\}. This test is commonly known as one-sample t-test.
Two Sample Inference about Normal Distribution

- Test $H_0: \sigma_1 = \sigma_2$ vs $H_a: \sigma_1 \neq \sigma_2$, under $H_0$

  \[ s_1^2 / s_2^2 \sim F_{n-1, m-1}, \quad s_1^2 = \sum_{i=1}^{n} (y_i - \bar{y})^2 / (n - 1), \quad s_2^2 = \sum_{j=1}^{m} (x_j - \bar{x})^2 / (m - 1). \]

  Select rejection regions as \{ $F(\alpha/2, n - 1, m - 1)$, $F(1 - \alpha/2, n - 1, m - 1)$ \}

- Test $H_0: \mu_1 = \mu_2$ vs $H_a: \mu_1 \neq \mu_2$, assuming $\sigma_1 = \sigma_2$. Under $H_0$

  \[ \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{s^2(1/n + 1/m)}} \sim t_{n+m-2}, \quad s^2 = \frac{\sum_{i=1}^{n} (y_i - \bar{y})^2 + \sum_{j=1}^{m} (x_j - \bar{x})^2}{n + m - 2}. \]
Rejection regions are \( \{t(\alpha/2, n + m - 2), t(1 - \alpha/2, n + m - 2)\} \). This is commonly known as two-sample t-test.

- If \( \sigma_1 \neq \sigma_2 \), we can use the test statistic
  \[
  \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}
  \]
  Under \( H_0 \) it has t-distribution with degree of freedom
  \[
  \frac{(s_1^2/n + s_2^2/m)^2}{(s_1^2/n)^2/(n - 1) + (s_2^2/m)^2/(m - 1)}.
  \]
  More details can be found at [http://www.biostat.umn.edu/~baolin/teaching/ph7405-05F/SupplementaryNotes.html](http://www.biostat.umn.edu/~baolin/teaching/ph7405-05F/SupplementaryNotes.html).