Linear Modeling Summary Sheet

Shravan Vasishth (vasishth@uni-potsdam.de)

July 13, 2015

Contents

Maximum likelihood estimation **2**
Binomial . . . 2
Normal . . . 2
Exponential . . . 2
Exponential: . . . 2
Practical implication of MLE . . . 2

Asymptotic properties of MLEs **2**
SE in the Binomial . . . 3
SE in the Normal . . . 3
Practical implication . . . 3
Connection to linear models . . . 3

Basic theory of linear models **3**

Inference **5**
Wald statistic (t-test) . . . 5
Type I, II, error, power, Type S and M errors . . . 6
Likelihood ratio test . . . 7
ANOVA . . . 8
Checking model assumptions . . . 9

Generalized Linear Models **9**
General form of the exponential family and the canonical link . . . 9
Assessing model fit and hypothesis testing in GLMs . . . 9

Linear mixed models **10**
Varying intercepts model . . . 10
Varying intercepts and slopes (with correlation) . . . 10

Linear Modeling Summary Sheet
Compiled by: Shravan Vasishth (vasishth@uni-potsdam.de)
Version dated: July 13, 2015

Maximum likelihood estimation

Binomial

$$L(\theta) = \binom{n}{x}\theta^x(1-\theta)^{n-x} \tag{1}$$

MLE: $\hat{\theta} = \frac{x}{n}$.

Normal

Exponential

$$\begin{aligned} L(\mu;\sigma^2) = & \prod N(x_i;\mu,\sigma) & (2)\\ = & (\tfrac{1}{\sigma\sqrt{2\pi}})^n \exp(-\tfrac{1}{2\sigma^2}\textstyle\sum_{i=1}^n (x_i-\mu)^2) & (3)\\ & & (4)\end{aligned}$$

The MLEs:

$$\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x} \tag{5}$$

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum (x_i-\bar{x})^2 \tag{6}$$

Exponential:

$$f(x;\lambda) = \lambda\exp(-\lambda x) \tag{7}$$

MLE:

$$\frac{n}{\lambda} = \sum x_i \tag{8}$$

Practical implication of MLE

Having decided that some data x_i, $i = 1,\ldots,n$ can be modeled as being generated from the normal distribution (for example), we can obtain estimates of the parameters using the closed-form expressions for the MLEs above.

Asymptotic properties of MLEs

The first essential point here is that under repeated sampling, the sampling distribution of the MLEs is asymptotically normal. This is the Central Limit Theorem.
The second essential point here is that under repeated sampling, we can compute the standard deviation of the sampling distribution of the MLE. This standard deviation is the **standard error**, and the SE is what allows us to do inference (hypothesis testing). We can compute the SE using the closed-form expression (derived in the lecture notes), where $\hat{\sigma}$ is the estimate of the standard deviation, and n is the sample size:

$$SE = \hat{\sigma}/\sqrt{n} \tag{9}$$

SE in the Binomial

$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$

SE in the Normal

$SE = \frac{\hat{\sigma}}{\sqrt{n}}$

Practical implication

Given an MLE for the mean $\bar{x}$, and the standard deviation $\hat{\sigma}$, from a single sample of size n, we can use these asymptotic properties and closed form solutions to derive an estimate of the SE.
That in turn allows us to compute the 95% confidence interval:
$\bar{x} \pm 2 \times SE$
which has the weird interpretation that if we were to repeatedly sample 100 times, 95% of those hypothetical CIs would contain the true value of the parameter (μ, which is a point value).

Connection to linear models

In linear models, we will have models like

$$y = \beta_0 + \beta_1 x + \varepsilon \quad \varepsilon \sim N(0, \sigma^2) \tag{10}$$

We will compute MLEs of the β parameters and of σ, and then we will do hypothesis testing using the estimated SEs of the β.

Basic theory of linear models

We can compute MLEs and SEs of β using these matrix results:

$$\hat{\beta} = (X^T X)^{-1} X^T Y \tag{11}$$

$$Var(\beta) = \sigma^2 [(X^T X)^{-1}] \tag{12}$$

For example, consider the design matrix with only eight data points:

```
(X<-matrix(c(rep(1,8),rep(c(-1,1),each=4),
            rep(c(-1,1),each=2,2)),ncol=3))

##      [,1] [,2] [,3]
## [1,]    1   -1   -1
## [2,]    1   -1   -1
## [3,]    1   -1    1
## [4,]    1   -1    1
## [5,]    1    1   -1
## [6,]    1    1   -1
## [7,]    1    1    1
## [8,]    1    1    1
```

```r
library(Matrix)
## det non-zero, hence full rank, hence invertible:
det(t(X)%*%X)

## [1] 512
```

Generate some hypothetical data with known β:

```r
beta<-matrix(c(2,0.5,0.25),nrow=3)
sigma<-1
n<-8
## some hypothetical data:
Y<-X%*%beta + matrix(rnorm(n,0,sigma),nrow=n)
```

You can compute the estimates of β by hand here:

```r
## inverse of X^T X:
(invXTX<-solve(t(X)%*%X))

##       [,1]  [,2]  [,3]
## [1,] 0.125 0.000 0.000
## [2,] 0.000 0.125 0.000
## [3,] 0.000 0.000 0.125

## (X^T X)^{-1} X^T Y:
(hatbeta<-solve(t(X)%*%X)%*%t(X)%*%Y)

##             [,1]
## [1,]  2.05477396
## [2,]  0.42627516
## [3,] -0.03219928
```

The SEs are the square roots of the diagonals in the variance covariance matrix. We first have to get an estimate of σ, which is $\sum e_i^2/(n-p)$, where p is the number of β parameters (here, 3):

```r
p<-3
e<-Y - X%*%hatbeta
(hatsigma2<-sum(e^2)/(n-p))

## [1] 1.649131

(hatsigma<-sqrt(hatsigma2))

## [1] 1.284185
```

Check that `lm` also gives this σ estimate:

```
m<-lm(Y ~ X[,2]+X[,3])
summary(m)$sigma

## [1] 1.284185
```

Now we are ready to compute the variance covariance matrix of the β, using the closed-form expression in equation 12:

```
round(hatsigma2 * invXTX,digits=3)

##       [,1]  [,2]  [,3]
## [1,] 0.206 0.000 0.000
## [2,] 0.000 0.206 0.000
## [3,] 0.000 0.000 0.206
```

Check that lm also produces this variance-covariance matrix (identical perhaps up to rounding error):

```
round(vcov(m),digits=3)

##             (Intercept) X[, 2] X[, 3]
## (Intercept)       0.206  0.000  0.000
## X[, 2]            0.000  0.206  0.000
## X[, 3]            0.000  0.000  0.206
```

Inference

Having estimated β and their SEs, we are ready to do inference.

Wald statistic (t-test)

$$\frac{\hat{\beta}_j - \beta_j}{\sqrt{Var(\beta_j)}} \sim Normal(0,1) \tag{13}$$

For each β_j, this statistic tests the null hypothesis that $\beta_j = 0$. If the null hypothesis is true, and the sample mean $\beta_j \sim N(0, SE^2_{\beta_j})$. So, if the $\hat{\beta}_j$ we get is extremely far away from 0, we reject the null.

The convention is that if the probability of getting a $\hat{\beta}_j$ that we got, or something more extreme in either direction, is less than 0.05, then we reject the null.

The p-value is the **conditional** probability of getting such a $\hat{\beta}_j$ estimate (or something more extreme), assuming that the null is true. **It is not the probability of the null being true**.

The t-value in `lm` that comes with a Wald test is with reference to the t-distribution with $n-1$ degrees of freedom. The t-distribution is an approximation to the normal. It has more probability mass in the tails for $n < 15$ or so. Beyond $n = 15$, the normal and t-distribution are basically indistinguishable.

The t-value is the number of SEs that the observed $\hat{\beta}_j$ is away from the hypothesized mean β (usually 0):

$$t = \frac{\hat{\beta}_j - \beta_j}{\sqrt{Var(\beta_j)}} \sim t(n-1) \tag{14}$$

R functions analogous to those we have seen for normal and binomial distributions are available for t-distributions:

```
## area to the left of -2 in t(8):
pt(-2,df=8)

## [1] 0.04025812

## critical t-value in t(8):
qt(0.025,df=8)

## [1] -2.306004
```

For large n (basically, anything larger than 15 or so data points), the critical t-value (the smallest t-value that would reject the null) is about 2.

Type I, II, error, power, Type S and M errors

In frequentist statistics, we can also compute Type I and Type II error rates.
Type I error is the probability of incorrectly rejecting the null (when it's actually true); this is typically set at 0.05 by the researcher and is called the α value.
Type II error is defined as the probability of incorrectly "accepting" (more accurately, failing to reject) the null hypothesis when it's false.
(1-Type II) error is called power, and is the probability of correctly rejecting the null.
Gelman adds two more errors:
Type S error: the probability that the sign of the effect is incorrect, given that (a) the result is statistically significant, or (b) the result is statistically non-significant.
Type M error: the expectation of the ratio of the absolute magnitude of the effect to the hypothesized true effect size (conditional on whether the result is significant or not). Gelman and Carlin also call this the exaggeration ratio, which is perhaps more descriptive than "Type M error".
For low power studies, Type S and M errors are embarrassingly huge. The take-home point is that if you are running a low power experiment, then even if you get a significant p value, don't get too excited—you're as likely to get the sign wrong as you are to get it right, and you are very likely to get a hugely exaggerated estimated of the true effect size.
You can get a feel for how bad the situation is by using our simulation above with eight data points, this time increasing sigma to 100.

```
beta<-matrix(c(2,0.5,0.25),nrow=3)
sigma<-100
n<-8
## some hypothetical data:
Y<-X%*%beta + matrix(rnorm(n,0,sigma),nrow=n)
##good luck estimating beta:
m<-lm(Y~X[,2]+X[,3])
summary(m)
```

```
##
## Call:
## lm(formula = Y ~ X[, 2] + X[, 3])
##
## Residuals:
##       1       2       3       4       5       6       7       8
##  -68.02   45.28  -69.22   91.95   62.37  -39.64 -107.08   84.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   -2.728     33.153  -0.082    0.938
## X[, 2]        -9.251     33.153  -0.279    0.791
## X[, 3]         3.996     33.153   0.121    0.909
##
## Residual standard error: 93.77 on 5 degrees of freedom
## Multiple R-squared:  0.01814,Adjusted R-squared:  -0.3746
## F-statistic: 0.0462 on 2 and 5 DF,  p-value: 0.9553
```

Likelihood ratio test

This is an alternative way to do hypothesis testing by explicitly comparing models that are more vs less complex (the simpler model is nested inside the other—in our examples, the simpler model has one predictor less).
Suppose that we have some data $x_1, \ldots, x_n$ from a random variable X whose distribution depends on the parameter θ. Suppose also that we want to test a hypothesis H_0 against H_1.
Define the **likelihood ratio test statistic** as

$$\lambda = 2\{\ell(\theta_1) - \ell(\theta_0)\} \tag{15}$$

where θ_1 and θ_0 are the estimates of θ under the alternative and null hypotheses, respectively. The likelihood ratio test rejects H_0 if λ is sufficiently large. As the sample size approaches infinity,

$$\lambda = \chi_r^2 \tag{16}$$

where r is called degrees of freedom and is the difference in the number of parameters estimated under H_1 and H_0. This is called Wilks' theorem.
Note that sometimes you will see the equivalent form:

$$\lambda = -2\{\ell(\theta_0) - \ell(\theta_1)\} \tag{17}$$

The test has a logic similar to the t-test. If the estimate of λ is highly unlikely given the null, which has the distribution χ_r^2 for λ, then we reject the null. Again, we compute the p-value with reference to some distribution, which is the chi-squared distibution here.

Table 1: default

Source of variance	df	Sum of squares	Mean square
Model with β_0	q	$\beta_0^T X_0^T Y$	
Improvement due to β_1	p-q	$\hat{\beta}_1 X_1^T Y - \hat{\beta}_0^T X_0^T Y$	$\frac{\hat{\beta}_1 X_1^T Y - \hat{\beta}_0^T X_0^T Y}{p-q}$
Residual	n-p	$Y^T Y - \hat{\beta}_1^T X_1^T Y$	$\frac{Y^T Y - \hat{\beta}_1^T X_1^T Y}{n-p}$
Total	n	$y^T y$	

ANOVA

Here, we compute the F-ratio, which is the ratio of between group variance to within-group variance. The basic calculations are summarized below:
In practical terms, you will be using the anova function to compare models. Example:

```
m1<-lm(Y~X[,2])
m2<-lm(Y~X[,2]+X[,3])
anova(m1,m2)

## Analysis of Variance Table
##
## Model 1: Y ~ X[, 2]
## Model 2: Y ~ X[, 2] + X[, 3]
##   Res.Df   RSS Df Sum of Sq      F Pr(>F)
## 1      6 44093
## 2      5 43965  1    127.77 0.0145 0.9087

summary(m)

##
## Call:
## lm(formula = Y ~ X[, 2] + X[, 3])
##
## Residuals:
##       1       2       3       4       5       6       7       8
##  -68.02   45.28  -69.22   91.95   62.37  -39.64 -107.08   84.35
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)   -2.728     33.153  -0.082    0.938
## X[, 2]        -9.251     33.153  -0.279    0.791
## X[, 3]         3.996     33.153   0.121    0.909
##
## Residual standard error: 93.77 on 5 degrees of freedom
## Multiple R-squared:  0.01814,Adjusted R-squared:  -0.3746
## F-statistic: 0.0462 on 2 and 5 DF,  p-value: 0.9553
```

Here, in m2 we have three β parameters $\beta_0, \beta_1, \beta_2$, and we are testing the null hypothesis that the third parameter $\beta_2 = 0$. This test is with reference to the F-distribution, but the logic is the same as in the t-test and the likelihood ratio test. We compare the observed F-score with the distribution of the F-statistic under the null, which has an F-distribution with the appropriate degrees of freedom.

Checking model assumptions

This step is generally omitted by people, leading quite often to ridiculous models that get published and lead to major scientific errors. The main things to know here are: check the distribution of residuals, look for influential values, consider a Box-Cox transform to stabilize variance, check for multicollinearity.

Generalized Linear Models

Logistic regression is the most important and common model:

$$\log \frac{p}{1-p} = \beta_0 + \beta_1 \mathrm{x} \tag{18}$$

$$p = \frac{\exp(\beta_0 + \beta_1 \mathrm{x})}{1 + \exp(\beta_0 + \beta_1 \mathrm{x})} \tag{19}$$

The inference theory remains the same as in linear models.

General form of the exponential family and the canonical link

$$f(y; \theta_i, \phi) = \exp\left[\frac{y\theta_i - b(\theta_i)}{\phi/w} + c(y, \phi)\right] \tag{20}$$

The big thing about the canonical link is that is expresses θ_i as a linear combination of the parameters: $x_i^T \beta$. You can decide which link to use by plotting $g(\mu_i)$ against the predictor (in case we have only a single predictor x).

Assessing model fit and hypothesis testing in GLMs

Deviance for the binomial distribution Deviance is defined as $D = \sum d_i$, where:

$$d_i = -2 \times n_i[y_i \log(\frac{\hat{\mu}_i}{y_i}) + (1 - y_i)\log(\frac{1-\hat{\mu}_i}{1-y_i})] \tag{21}$$

The basic idea here is that if the model fit is good, Deviance will have a χ^2 distribution with $N - p$ degrees of freedom. So that is what we will use for assessing model fit.

We will also use deviance for hypothesis testing. The difference in deviance (residual deviance) between two models also has a χ_2 distribution (this should remind you of ANOVA), with dfs being $p - q$, where q is the number of parameters in the first model, and p the number of parameters in the second.

The anova function for glm does the implicit model comparison for you, delivering the residual deviance and the difference in degrees of freedom between the two models being compared, and the deviance of the model. So the anova function gives you all the output for doing inference and for evaluating model fit.

Distribution	$h(x_i^T\beta) = \mu_i$	$g(\mu_i) = \theta_i$
Binomial logit link	$\frac{\exp[\theta_i]}{1+\exp[\theta_i]}$	$\log \frac{y}{1-y}$
Normal identity	θ	$g = h$
Poisson log	$\exp[\theta]$	$\log[\mu]$
Gamma inverse	$-\frac{1}{\theta}$	$-\frac{1}{\mu_i}$
Cloglog cloglog	$1 - \exp[-\exp[\theta_i]]$	$\log(-\log(1-\mu_i))$
Probit probit	$\Phi(\theta)$	$\Phi^{-1}(\theta)$ (qnorm)

Linear mixed models

Varying intercepts model

The model for a categorical predictor is:

$$Y_{ijk} = \beta_j + b_i + \epsilon_{ijk} \tag{22}$$

$i = 1, \ldots, 10$ is subject id, $j = 1, 2$ is the factor level, k is the number of replicates (here 1). $b_i \sim N(0, \sigma_b^2), \epsilon_{ijk} \sim N(0, \sigma^2)$.
For a continuous predictor:

$$Y_{ijk} = \beta_0 + \beta_1 t_{ijk} + b_{ij} + \epsilon_{ijk} \tag{23}$$

Varying intercepts and slopes (with correlation)

The model for a categorical predictor is:

$$Y_{ij} = \beta_1 + b_{1i} + (\beta_2 + b_{2i})x_{ij} + \epsilon_{ij} \quad i = 1, ..., M, j = 1, ..., n_i \tag{24}$$

with $b_{1i} \sim N(0, \sigma_1^2), b_{2i} \sim N(0, \sigma_2^2)$, and $\epsilon_{ij} \sim N(0, \sigma^2)$.
Another way to write such models is:

$$Y_{ijk} = \beta_j + b_{ij} + \epsilon_{ijk} \tag{25}$$

$b_{ij} \sim N(0, \sigma_b)$. The variance σ_b must be a 2×2 matrix:

$$\begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \tag{26}$$

You should be able to state what the random effects variance-covariance matrix is given model output.