Linear Regression Models

Xi Tan (tan19@purdue.edu)

December 30, 2012

Contents

1 Introduction **3**

2 Simple Linear Regression **4**

2.1 Model 4
2.2 Estimators 4
2.3 Properties of Residuals 4
2.4 Properties of b_1 and b_0 4
2.5 Confidence Interval of b_1 and b_0 5

Preface

TBD

1 Introduction

Generalized linear models include as special cases, linear regression and analysis-of-variance models, logit and probit models for quantal responses, log linear models and multinomial response models for counts and some commonly used models for survival data.

The second-order properties of the parameter estimates are insensitive to the assumed distributional form: the second-order properties depend mainly on the assumed variance-to-mean relationship and on uncorrelatedness or independence.

Data types:

2 Simple Linear Regression

2.1 Model

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{1}$$

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

2.2 Estimators

$$\begin{aligned} b_1 &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \sum_{i=1}^n (X_i - \bar{X}) Y_i & (2) \\ b_0 &= \bar{Y} - b_1 \bar{X} & (3) \\ \hat{\sigma}^2 &= \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} & (4) \end{aligned}$$

Notice, $\sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$, and $b_1 = \rho \cdot \frac{s_Y}{s_X}$, where ρ is the correlation between X and Y and s_Y, s_X are standard error of Y and X, respectively.

2.3 Properties of Residuals

$$\begin{aligned} e_i &= Y_i - \hat{Y}_i & (5) \\ \sum e_i &= 0 & (6) \\ \sum X_i e_i &= 0 & (7) \\ \sum \hat{Y}_i e_i &= 0 & (8) \end{aligned}$$

2.4 Properties of b_1 and b_0

$$\begin{aligned} b_1 &\sim \mathcal{N}\left(\beta_1, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right) & (9) \\ b_0 &\sim \mathcal{N}\left(\beta_0, \frac{\sigma^2}{n} + \frac{\sigma^2 \bar{X}^2}{\sum (X_i - \bar{X})^2}\right) & (10) \end{aligned}$$

where σ^2 can be estimated by the MSE, i.e., $\hat{\sigma}^2 = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2}$

Now, since

$$\begin{aligned} b_1 &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \sum_{i=1}^n (X_i - \bar{X}) Y_i & (11) \\ &= \sum_{i=1}^n k_i Y_i & (12) \end{aligned}$$

where $k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}$, we have

$$\begin{aligned} \sum k_i &= 0 & (13) \\ \sum X_i k_i &= 1 & (14) \\ \sum k_i^2 &= \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} & (15) \end{aligned}$$

The first two identity hold as a requirement for the unbiasness, since

$$E(b_1) = E\left(\sum k_i Y_i\right) = E\left(\sum k_i (\beta_0 + \beta_1 X_i)\right) = E\left(\beta_0 \sum k_i \beta_0 + \beta_1 \sum k_i X_i\right) = \beta_1$$

requires $\sum k_i = 0$ and $\sum X_i k_i = 1$. The third identity ensures the attainment of the minimum variance.

2.5 Confidence Interval of b_1 and b_0

Since $SSE/\sigma^2 \sim \chi^2_{n-2}$, and $\frac{s^2\{b_1\}}{\sigma^2\{b_1\}} \sim \frac{\chi^2_{n-2}}{n-2}$

$$\frac{b_1 - \beta_1}{s\{b_1\}} = \frac{b_1 - \beta_1}{\sigma\{b_1\}} \Big/ \frac{s\{b_1\}}{\sigma\{b_1\}} \sim \frac{z}{\sqrt{\frac{\chi^2_{n-2}}{n-2}}} = t_{n-2} \tag{16}$$

so the confidence interval for b_1, with confidence level α is

$$b_1 \pm t(1 - \alpha/2; n-2) s\{b_1\} \tag{17}$$

or

$$b_1 \mp t(\alpha/2; n-2) s\{b_1\} \tag{18}$$

Similarly, the confidence interval for b_0, with confidence level α is

$$b_0 \pm t(1 - \alpha/2; n-2) s\{b_0\} \tag{19}$$

or

$$b_0 \mp t(\alpha/2; n-2) s\{b_0\} \tag{20}$$

The power of testing $\beta_1 = \beta^{H_0}$ is $Power = P\{|t^*| > t(1-\alpha/2; n-2) | \delta\}$, where $\delta = \frac{|\beta_1 - \beta^{H_0}|}{\sigma\{b_1\}}$. Similar for β_0.

	Estimate	Expectation	Variance
Y_i	$\hat{Y}_i$	$\beta_0 + \beta_1 X_i$	σ^2
b_1	$\frac{\sum(X_i-\bar{X})Y_i}{\sum(X_i-\bar{X})^2}$	β_1	$\sigma^2 \cdot \frac{1}{\sum(X_i-\bar{X})^2}$
b_0	$\bar{Y} - b_1\bar{X}$	β_0	$\sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i-\bar{X})^2}\right]$
$\hat{Y}_i$	$\bar{Y} + b_1(X_i - \bar{X})$	$\beta_0 + \beta_1 X_i$	$\sigma^2 \cdot \left[\frac{1}{n} + \frac{(X_i-\bar{X})^2}{\sum(X_i-\bar{X})^2}\right]$
e_i	$Y_i - \hat{Y}_i$	0	TBD

Table 1: Simple Linear Regression