Qualifying Exam Preparation I
Methods

Xi Tan (tan19@purdue.edu)

October 12, 2013

Contents

1 Simple Linear Regression

1.1 Model

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

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

1.2 Estimated Regression Function

$$\begin{aligned}
b_1 &= \rho_{XY} \cdot \frac{s_Y}{s_X} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \sum_{i=1}^n \left[\frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \right] Y_i & (2) \\
b_0 &= \bar{Y} - b_1 \bar{X} & (3) \\
\hat{\sigma}^2 &= \frac{MSE}{n-2} = \frac{\sum_{i=1}^n (Y_i - \hat{Y}_i)^2}{n-2} & (4)
\end{aligned}$$

Notice, $\sum (X_i - \bar{X})^2 = \sum X_i^2 - n\bar{X}^2$.

The slope of the fitted line is equal to the correlation between y and x corrected by the ratio of standard deviations of these variables. The intercept of the fitted line is such that it passes through the center of mass $(\bar{x}, \bar{y})$ of the data points.

Another way of writing the estimated regression function is

$$\hat{Y}_i = \bar{Y} + b_1 (X_i - \bar{X}) \tag{5}$$

Notice, $\bar{Y}$ and b_1 are uncorrelated (check it using the fact that $b_1 = \sum_{i=1}^n k_i Y_i$).

1.3 Properties of k_i

$$k_i = \frac{X_i - \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{6}$$

$$\sum_{i=1}^n k_i = 0 \tag{7}$$

$$\sum_{i=1}^n k_i X_i = 1 \tag{8}$$

$$\sum k_i^2 = \frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \tag{9}$$

The second and third identities hold as a requirement for the unbiasness, since

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

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

1.4 Properties of e_i

$$e_i = Y_i - \hat{Y}_i \tag{10}$$

$$\sum e_i = 0 \tag{11}$$

$$\sum X_i e_i = 0 \tag{12}$$

$$\sum \hat{Y}_i e_i = 0 \tag{13}$$

1.5 Properties of b_1 and b_0

$$b_1 \sim \mathcal{N}\left(\beta_1, \sigma^2\left[\frac{1}{\sum(X_i - \bar{X})^2}\right]\right) \tag{14}$$

$$b_0 \sim \mathcal{N}\left(\beta_0, \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]\right) \tag{15}$$

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

1.6 Inference About b_1 and b_0

The confidence interval for b_1, with confidence level α is

$$b_1 \pm t(1 - \alpha/2; n-2)s\{b_1\} \tag{16}$$

or

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

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

$$b_0 \pm t(1 - \alpha/2; n-2)s\{b_0\} \tag{18}$$

or

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

	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	$\sigma^2 \cdot \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum(X_i - \bar{X})^2}\right]$
$\hat{Y}_h$	$\bar{Y} + b_1(X_h - \bar{X})$	$\beta_0 + \beta_1 X_h$	$\sigma^2 \cdot \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$
$\hat{Y}_{h(new)}$	$\bar{Y} + b_1(X_h - \bar{X})$	$\beta_0 + \beta_1 X_h$	$\sigma^2 \cdot \left[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$
$\hat{Y}_{h(new_m)}$	$\bar{Y} + b_1(X_h - \bar{X})$	$\beta_0 + \beta_1 X_h$	$\sigma^2 \cdot \left[\frac{1}{m} + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum(X_i - \bar{X})^2}\right]$
e_i	$Y_i - \hat{Y}_i$	0	$1 - h_{ii}$

Table 1: Simple Linear Regression

In particular, when $X_h = 0$ we obtain the formulas for b_0, and when $X_h - \bar{X} = 1$ we obtain the formulas for b_1.

1.7 ANOVA of Simple Linear Regression Model

$$SSTO = SSR + SSE \tag{20}$$

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i) = \sum_{i=1}^{n}(\hat{Y}_i - \bar{y}) + \sum_{i=1}^{n}(\bar{y} - \hat{Y}_i) \tag{21}$$

SSR can also be computed as $SSR = b_1^2 \sum_{i=1}^{n}(X_i - \bar{X})$, so given the same "distribution" of X, the steeper the slope of the regression line, the higher the SSR, and hence the better fit of the model.

To test $H_0 : \beta_1 = 0$, we use $F = \frac{SSR}{SSE}$. There is equivalence between an F test and a t test: $[t(1 - \alpha/2, n - 2)]^2 = F(1 - \alpha, n - 2)$.

2 Survivial Analysis

$$S(t) = \exp\left[-\int_0^t \lambda(u)du\right] \tag{22}$$

$$L(\lambda) = \prod_{i=1}^{n} [\lambda(t_i)]^{\delta_i} [S(t_i)]^{1-\delta_i} \tag{23}$$

where $S(t)$ is the survival function, and $\lambda(t)$ is the hazard function.

	Estimate	Standard Error	NOTE
S	$\hat{S(t)} = \prod \frac{n_j - d_j}{n_j}$	$\hat{S(t)}\sqrt{\sum \frac{d_i}{n_j(n_j-d_j)}}$	
Λ	$-\log \hat{S(t)}$	$\sqrt{\sum \frac{d_i}{n_j(n_j-d_j)}}$	
λ	$\frac{\sum \delta_i}{\sum (X_i - V_i)}$	$\frac{\hat{\lambda}}{\sqrt{\sum \delta_i}}$	

Table 2: Survival Analysis

3 Exponential Family

$$f(y|\theta, \phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y, \phi)\right\} \tag{24}$$

$$E(y) = b'(\theta) \tag{25}$$

$$Var(y) = b''(\theta)a(\phi) \tag{26}$$