Fall 1999 Course 513-697: Applied Linear Models

Highlights / Key Concepts in NKNW4 Chapter 2.1-2.6

[See also "Notes on M&M Chapters 2 and 9", "Bridge" from 607" & "Chapter 5" under Chapter 5 of webpage for course 678]

- Parameters of Interest: β_1, β_0 and derivatives of them; Estimators of these: b_0, b_1 and derivatives of them

- The following are all based on the assumption of Gaussian Error Regression Model

  - Inferences based on t distrn. [non-Gaussian errors => t-based inferences not entirely accurate, but 'close' if n large]

  - Reason for t: - b_0, b_1, and estimates derived from them are all linear combinations of Y's and so all have Gaussian variation

    - variances of b_0, b_1 and estimates derived from them all involve σ^2;

    - if σ^2 known, all inferences would be based on Gaussian distribution but σ^2 has to be estimated, so must use a slightly wider distribution (t) instead

- **§2.1 Inference concerning** β_1 [β_1 usually of far greater interest than β_0]

  - $\beta_1 = 0$ <==> "No linear association b/w Y and X" (Distrn of Y | X identical for all X

  - $\beta_1 \neq 0$ <==> "Linear association b/w Y and X"

  • b_1 is linear combination of Gaussian random variables -- each has a different mean if $\beta_1 \neq 0$

  • $E\{b_1\} = \beta_1$ so b_1 is an unbiased estimator of β_1

  • $\text{var}\{b_1\} = \dfrac{\sigma^2}{\sum[X - \bar{X}]^2}$ See "Notes on M&M Chapters 2 and 9" (under chapter 5 in 678 www page) for discussion of alternative forms for var$\{b_1\}$ and for the factors that affect var$\{b_1\}$

  • $b_1 \sim \text{Gaussian}(ß_1, \text{var}\{b_1\}) \Rightarrow \dfrac{b_1 - \beta_1}{\sqrt{\text{var}\{b_1\}}} \sim \text{Gaussian}(0,1)$

  BUT var$\{b_1\}$ involves σ^2 ... and σ^2 is typically unknown and so must be ESTIMATED... by $\text{MSE} = \dfrac{\sum[Y - \hat{Y}]^2}{n-2}$

  so, we have instead: $\dfrac{b_1 - \beta_1}{\sqrt{\text{ESTIMATED var}\{b_1\}}}$ ~ t(with n-2 degrees of freedom)

  THIS IS THE BASIS FOR INFERENCES CONCERNING β_1

  This is the same concept as when in a first course in statistics, we wished to make inference concerning μ on the basis of n independent observations from a single Gaussian(μ, σ^2). In that case... $\bar{Y}$ is a linear combination of i.i.d. Gaussian random variables -- with mean μ; $E\{\bar{Y}\} = \mu$ so $\bar{Y}$ is an unbiased estimator of μ; $\text{var}\{\bar{Y}\} = \dfrac{\sigma^2}{n}$

  => $\bar{Y} \sim \text{Gaussian}(\mu, \text{var}\{\bar{Y}\}) \Rightarrow \dfrac{\bar{Y} - \mu}{\sqrt{\text{var}\{\bar{Y}\}}} \sim \text{Gaussian}(0,1)$

  BUT var$\{\bar{Y}\}$ involves σ^2 ... but if σ^2 is unknown and must be ESTIMATED... by $\text{MSE} = \dfrac{\sum[Y - \bar{Y}]^2}{n-1}$

  then, we have instead: $\dfrac{\bar{Y} - \mu}{\sqrt{\text{ESTIMATED var}\{\bar{Y}\}}}$ ~ t(n-1 degrees of freedom)

- t variable with ν degrees of freedom = $\dfrac{\text{Gaussian}[0;1] \text{ variable}}{\sqrt{\dfrac{\text{Independent } \chi^2 \text{ variable with } \nu \text{ degrees of freedom}}{\nu}}}$

Highlights / Key Concepts in NKNW4 Chapter 2.1-2.6

- $100(1-\alpha)\%$ 2-sided CI for β_1 : $b_1 \pm t(1-\alpha/2, n-2)\sqrt{\text{ESTIMATED var}\{b_1\}}$

  - $\sqrt{\text{ESTIMATED var}\{b_1\}}$ is often called the Standard Error or "SE" of b_1.

- Test of hypothesis H_0: β_1 = specified value [not necessarily zero]

  vs. H_a: $\beta_1 \neq$ specified value [2-sided] or say $\beta_1 >$ specified value [1-sided]

  based on test statistic $t^* = \dfrac{b_1 - \text{specified value}}{\sqrt{\text{ESTIMATED var}\{b_1\}}}$ vis-a-vis $t(n-2)$

  NOTE the link between 2-sided tests and 2-sided CI's (cf example 1 p 51, next line after 2.16)

  INSTEAD OF "CONCLUDING H_0" (in 2.18 p 51), PREFERABLE TO SAY "DID NOT REJECT H_0"
  (there's a big difference between 'concluding' and 'not ruling out' : if we took the author's wording, then a great way to never conclude anything but H_0 would be to not collect much data, so that the power to detect H_a, even if it were true, was minimal; there is a big difference between "evidence of no relation" and "no evidence of a relation")

- **2.2 Inference concerning β_0** [β_0 usually of lesser interest -- might not even be any data close to X=0]

  Inference via $b_0 = \bar{Y} - b_1 \bar{X}$

  Can rewrite b_0 as a linear combination of Y's, so if errors (and thus Y's) are Gaussian, so will be behaviour of b_0 .

  $$\text{var}\{b_0\} = \frac{\sigma^2}{n} + \sigma^2 \frac{\bar{X}^2}{\sum[X - \bar{X}]^2} = \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum[X - \bar{X}]^2} \right]$$

  note that the further the data are from X=0, the larger the uncertainty in the estimate of the intercept.

  inference via fact that $\dfrac{b_0 - \beta_0}{\sqrt{\text{ESTIMATED var}\{b_0\}}}$ ~ t(with n-2 degrees of freedom)

- **2.3 Notes:**

  - Asymptotic normality: akin to Central Limit Theorem and the fact that a linear combination of a large number of non-identical but INDEPENDENTLY distributed [a key assumption] random variables will have close to a Gaussian distribution even if the random variables do not themselves have Gaussian distributions. A little more complicated here since dealing with ratio of a linear combination of random variables to a separate estimate of variance.

  - Spacing of X levels: see "factors that affect SE of estimate of slope" in other handout (from course 607).

  - Power of Tests... skip for now

- **2.4 Inference concerning E{Y | specified level of X}** [don't know why authors used h in X_h]**:**

  Point Estimator of E{Y | specified level, X_h, of X} : $\hat{Y}_h = b_0 + b_1 X_h$ **Note : b_0 & b_1 negatively correlated***

  - This is a linear combination of the Y's and so has a Gaussian distribution with

Highlights / Key Concepts in NKNW4 Chapter 2.1-2.6

$E\{\hat{Y}_h\} = \beta_0 + \beta_1 X_h$

$$\text{Var}\{\hat{Y}_h\} = \sigma^2 \left[\frac{1}{n} + \frac{[X_h - \bar{X}]^2}{\sum[X - \bar{X}]^2} \right]$$

*** var more easily derived if rewrite $\hat{Y}_h = \bar{Y} + b_1 [X_h - \bar{X}]$ ... 2 components uncorrelated >**

- Again, as in 2.1 and 2.2, we must usually ESTIMATE σ^2 by MSE, so that we instead have

$$\frac{\hat{Y}_h - \{\beta_0 + \beta_1 X_h\}]}{\sqrt{\text{ESTIMATED Var}\{\hat{Y}_h\}}} \sim \text{t(n-2 degrees of freedom)}$$

CI's and tests are as in §2.1 or §2.2.

As can be seen from variance formula, CI's are wider further away from the center, $\bar{X}$, of the X points.

Note also that if $X_h = \bar{X}$, then $\hat{Y}_h = \bar{Y}$ and Var{ $\hat{Y}_h$ } reduces to the familiar var{ $\bar{Y}$ } = $\sigma^2 \left[\frac{1}{n} \right]$.

- **2.5 Inference (prediction) concerning a new Y at a specified level of X}:**

Have to approach in two steps:

1 estimate what the mean (center) of all possible observations would be at $X = X_h$.

2 Overlay the distribution of individual Y's on this estimated mean. Having lots of data to estimate the <u>center</u> quite precisely will not alter the fact that the individuality of the Y values remains unaltered; mind you, we will have to estimate --via the MSE -- this individuality.

The uncertainty about a new individual now contains two components 1. the precision (or lack of it) associated with getting the middle correct and 2. the (unalterable) individuality or individuals

pred observation on individual = true mean + error in estimating this mean + individuality

var{pred observation on individual} = var{estimate of mean} + var{individuals about true mean}

$$= \sigma^2 \left[\frac{1}{n} + \frac{[X_h - \bar{X}]^2}{\sum[X - \bar{X}]^2} \right] + \sigma^2$$

$$= \sigma^2 \left[1 + \frac{1}{n} + \frac{[X_h - \bar{X}]^2}{\sum[X - \bar{X}]^2} \right]$$

CI for individual based on t(n – 2) rather than Z, since σ^2 has to be estimated by MSE.

- **2.6 Confidence Band for ENTIRE Regression Line:**

• This is different from what is usually output, namely the CI given in §2.5

• See especially notes 3 and 4 p69