MATH3805 Regression Analysis

Hong Kong Baptist University

Fall 2021

Textbook:

Mendenhall, W. and Sincich, T. *A Second Course in Statistics Regression Analysis, 7th edn*. Pearson, 2012.

References:

Draper, N.R. and Smith, H. *Applied Regression Analysis, 3rd edn*. Wiley, 1998.

Freund, R.J. and Wilson, W.J. *Regression Analysis*. Academic Press, 1998.

Montgomery, D.C. *Design and Analysis of Experiments, 4th edn*. Wiley, 1997.

Dobson, A.J. *An Introduction to Generalized Linear Models, 2nd edn*. Chapman & Hall, 2002.

Daly, F., Hand, D.J., Jones, M.C., Lunn, A.D. and McConway, K.J. *Elements of Statistics*. Addison-Wesley, 1995.

Moore, D.S. and McCabe, G.P. *Introduction to the Practice of Statistics, 3rd edn*. Freeman, 1999, or 4th edn, Freeman, 2003.

Instructor:

Dr. PENG Heng
Office: FSC1205
Email: hpeng@hkbu.edu.hk

Assessment methods:

final score = 0.4 (mid-term & assignments) + 0.6 (final exam)

Distributions derived from the normal distribution

Definition

If $Z_1, \ldots, Z_\nu$ are i.i.d. with $Z_1 \sim \mathcal{N}(0,1)$, then the distribution of $\sum_{i=1}^{\nu} Z_i^2$ is called the χ^2_ν distribution (ν is called **degrees of freedom**).

Definition

If $Z \sim \mathcal{N}(0,1)$ and $Y \sim \chi^2_\nu$ independent of Z, then the distribution of $Z/\sqrt{Y/\nu}$ is called the t_ν distribution (ν is called the degrees of freedom).

Definition

If $W_1 \sim \chi^2_{k_1}$, $W_2 \sim \chi^2_{k_2}$, and W_1 and W_2 are independent, then the distribution of $\frac{W_1/k_1}{W_2/k_2}$ is called the F_{k_1,k_2} distribution (k_1 and k_2 are the degrees of freedom).

One normal sample

$Y_1, \ldots, Y_n \sim \mathcal{N}(\mu, \sigma^2)$ or

$$Y_i = \mu + \epsilon_i, \ \ \epsilon_i \sim \mathcal{N}(0, \sigma^2), \ \ i = 1, \ldots n, \tag{1}$$

where $Y_1, \ldots, Y_n$ are i.i.d.
Define

$$\begin{aligned} \bar{Y} &= \frac{1}{n}\sum_{i=1}^{n} Y_i, \\ S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \\ T &= \frac{\sqrt{n}(\bar{Y} - \mu)}{S}. \end{aligned} \tag{2}$$

$\bar{Y}$: sample mean
S^2: sample variance

Theorem

If $Y_1, \ldots, Y_n$ are i.i.d. with $Y_1 \sim \mathcal{N}(\mu, \sigma^2)$, then $\bar{Y}$ and S as defined in (2) are independent, and $\bar{Y} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{n-1}$, and $T \sim t_{n-1}$.

Point estimation of parameters in normal distribution

Model: $Y_1, \ldots, Y_n$ are i.i.d with $\mathcal{N}(\mu, \sigma^2)$ distribution.

Since $\bar{Y} \sim \mathcal{N}(\mu, \sigma^2/n)$, $\bar{Y}$ is unbiased for μ ($E(\bar{Y}) = \mu$) and it has variance $Var(\bar{Y}) = \sigma^2/n$.
Also, since $\frac{(n-1)}{\sigma^2} S^2 \sim \chi^2_{n-1}$, S^2 is unbiased for σ^2 ($E(S^2) = \sigma^2$) and it has variance $Var(S^2) = 2\sigma^4/(n-1)$.

Method of moments estimators: The method of moments estimators for μ and σ^2 are the solutions of μ and σ^2 to the following equations:

$$\begin{aligned} \mu &= \bar{Y}, \\ \mu^2 + \sigma^2 &= \frac{1}{n} \sum_{i=1} Y_i^2. \end{aligned}$$

They are $\bar{Y}$ and $\frac{n-1}{n} S^2$.

Maximum likelihood estimators: Likelihood function given $Y_1, \ldots, Y_n$:

$$\ell(\mu, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i-\mu)^2}{2\sigma^2}} = (2\pi\sigma^2)^{-\frac{n}{2}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i-\mu)^2}.$$

Log likelihood function:

$$L(\mu, \sigma^2) = \log \ell(\mu, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2.$$

Solving for μ and σ^2 the following system of equations:

$$\begin{aligned} 0 &= \frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(Y_i - \mu), \\ 0 &= \frac{\partial L}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(Y_i - \mu)^2, \end{aligned}$$

and checking the solutions yield maximum, we obtain the maximum likelihood estimators for μ and σ^2 as $\bar{Y}$ and $\frac{n-1}{n}S^2$ respectively.

Coefficient of Variation

The parameter σ^2 is important to us because the greater the variability of the random term, the greater the errors in the estimation.

The rule of thumb (i.e. solely a working principle based on experience and perhaps wisdom but not on mathematical arguments) is that models with CV no more than 10% usually lead to accurate prediction, where

$$\text{CV} = \text{coefficient of variation} = \sigma/\mu \times 100\%.$$

Confidence interval for normal location

$1-\alpha$: confidence level

Let $t_{\nu;1-\alpha/2}$ be the percentage point of the t_ν distribution that leaves a probability $\alpha/2$ in the upper tail. Since

$$\begin{aligned} 1-\alpha &= P(t_{n-1;\alpha/2} \leq T \leq t_{n-1;1-\alpha/2}) \\ &= P\left(t_{n-1;\alpha/2} \leq \frac{\sqrt{n}(\bar{Y}-\mu)}{S} \leq t_{n-1;1-\alpha/2}\right) \end{aligned}$$

and $t_{n-1;\alpha/2} = -t_{n-1;1-\alpha/2}$, we have

$$P\left(\bar{Y} - t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}} \leq \mu \leq \bar{Y} + t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}\right) = 1-\alpha.$$

Hence

$$\bar{Y} \pm t_{n-1;1-\alpha/2}\frac{S}{\sqrt{n}}$$

is a $(1-\alpha)\times 100\%$ confidence interval for μ.

Confidence interval for normal variance

Let $\chi^2_{\nu,1-\alpha/2}$ be the percentage point of the χ^2_ν distribution that leaves a probability $\alpha/2$ in the upper tail. Since

$$\begin{aligned}1-\alpha &= P(\chi^2_{n-1;\alpha/2} \leq \frac{n-1}{\sigma^2}S^2 \leq \chi^2_{n-1;1-\alpha/2}) \\ &= P\Big(\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\Big),\end{aligned}$$

a $(1-\alpha)\times 100\%$ confidence interval for σ^2 is

$$\Big[\frac{(n-1)S^2}{\chi^2_{n-1;1-\alpha/2}}, \frac{(n-1)S^2}{\chi^2_{n-1;\alpha/2}}\Big].$$

Testing Hypotheses on normal location

Example Work times of a worker: 13.9, 10.8, 13.9, 9.3, 11.7, 9.1, 12.0, 10.4, 13.3, 11.1.

Question: Can the worker perform the task in 10 minutes on average?

Test the null hypothesis $H_0: \ \mu = \mu_0 = 10$ against the alternative hypothesis $H_1: \ \mu > \mu_0 = 10$.

Test statistic is $T = \frac{\sqrt{n}(\bar{Y} - \mu_0)}{S}$, and we would reject H_0 if the observed value of T, denoted as t, is large.

Since distribution of T under H_0 is t_{n-1}, critical value at significance level α is $t_{n-1;1-\alpha}$.

p-value is $P(T > t | H_0) = P(t_{n-1} > t)$, where t is the observed value of T given the sample.

The critical region tells us what values are considered too extreme (i.e. too unlikely to be seen) for the test statistic, if the null hypothesis is true.

Hence, if the observed value of the test statistic happens to be in the critical region, then we believe the null hypothesis is not true.

The p-value is the probability, assuming the null hypothesis is true, of observing what we have observed or something more extreme.

Thus, a small p-value means that what has happened would be in fact unlikely to happen if the null hypothesis is true. However, it really has happened and so we believe that the null hypothesis is not true.

Two normal samples

$Y_{11}, \ldots, Y_{1n_1}, Y_{21}, \ldots, Y_{2n_2}$ independent,
$Y_{1j} \sim \mathcal{N}(\mu_1, \sigma^2),\ j = 1, \ldots, n_1,$
$Y_{2j} \sim \mathcal{N}(\mu_2, \sigma^2),\ j = 1, \ldots, n_2.$

$$\Leftrightarrow Y_{ij} = \mu_i + \epsilon_{ij},\ i = 1, 2,\ j = 1, \ldots n_i,\ \epsilon_{ij}\ i.i.d. \sim \mathcal{N}(0, \sigma^2).$$

Note: equal variances assumption

Example Compare the working times with that of another worker.
Worker 1: 13.9, 10.8, 13.9, 9.3, 11.7, 9.1, 12.0, 10.4, 13.3, 11.1
Worker 2: 14.1, 10.7, 13.2, 10.4, 10.0, 10.1, 10.6, 12.5, 14.5, 10.9

Independent two-sample t-test

Given two independent samples $Y_{11}, \ldots, Y_{1n_1}$ i.i.d. $\sim \mathcal{N}(\mu_1, \sigma^2)$ and $Y_{21}, \ldots, Y_{2n_2}$ i.i.d. $\sim \mathcal{N}(\mu_2, \sigma^2)$.

$H_0: \ \mu_1 - \mu_2 = \mu_0$ (usually $\mu_0 = 0$), $\ H_1: \ \mu_1 - \mu_2 \neq \mu_0$

Test statistic:

$$T = \frac{\bar{Y}_1 - \bar{Y}_2 - \mu_0}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}},$$

where

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2},$$

$\bar{Y}_1$ and $\bar{Y}_2$ are the sample means, and S_1^2 and S_2^2 are the sample variances.

Distribution of T under H_0:

$T \sim t_{n_1+n_2-2}$ when H_0 is true.

p-value is $P(|T| > |t||H_0) = P(|t_{n_1+n_2-2}| > |t|)$, where t is the observed value of T given the two samples.

p-value $< \alpha$ if $t < -t_{n_1+n_2-2;1-\alpha/2}$ or $t > t_{n_1+n_2-2;1-\alpha/2}$.

Level $1-\alpha$ confidence interval for $\mu_1 - \mu_2$ is

$$(\bar{Y}_1 - \bar{Y}_2) \pm S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}\, t_{n_1+n_2-2;1-\alpha/2}$$

Paired two-sample t-test

Use additional information if you know that the two samples consist of paired observations:
$Z_j = Y_{1j} - Y_{2j}$, $j = 1, \ldots, n$, i.i.d. with
$Z = Y_1 - Y_2 \sim \mathcal{N}(\mu_d, \sigma_d^2)$, $\mu_d = \mu_1 - \mu_2$.

Example: Y_{1j} and Y_{2j} are test scores of the jth pairs of slower learners.

Perform one-sample t-test for $H_0: \ \mu_d = 0$ based on the data $Z_1, \ldots, Z_n$.
Test statistic is

$$T = \frac{\sqrt{n}(\bar{Z} - 0)}{S_d},$$

where $\bar{Z} = \bar{Y}_1 - \bar{Y}_2$ and S_d^2 is the sample variance of $Z_1, \ldots, Z_n$, and its distribution is t_{n-1} under the null hypothesis $H_0: \ \mu_d = 0$.