VE401 Probabilistic Methods in Eng.

Final Review

> By TA: DONG Juechu, April. 2021

if you want to edit this note, you can find it here https://github.com/joydddd/VE401-2021SP-notes

Mathematica code is also available on github.

Categorial Test

$$f_{X_1 X_2 \cdots x_k}(x_1, \ldots, x_k) = \frac{n!}{x_1! \cdots x_k!} p_1^{x_1} \cdots p_k^{x_k}$$

$p_1, \ldots, p_k \in (0,1), n \in \mathbb{N} \backslash \{0\}$ is said to have a multinomial distribution with parameters n and $p_1, \ldots, p_k$.

1. The (marginal) expectations of the individual random variables X_i are given by

$$\mathrm{E}[X_i] = np_i, \quad i = 1, \ldots, k.$$

2. $\operatorname{Var}[X_i] = np_i(1 - p_i), i = 1, \ldots, k$,
3. $\operatorname{Cov}[X_i, X_j] = -np_i p_j, 1 \leq i < j \leq k$.

Pearson Chi-squared Goodness of Fit Test

Test if the data follows multinomial distribution with parameters $(p_0, p_1, p_2 \ldots)$

$$H_0 : p_i = p_{i_0}, \quad i = 1, \ldots, k$$
$$X_{k-1}^2 = \sum_{i=1}^{k} \frac{(X_i - np_{i_0})^2}{np_{i_0}}$$

We reject H_0 at significance level α if $X_{k-1}^2 > \chi_{\alpha, k-1}^2$.

Cochran's Rule: make sure the chi-squared approximation is appropriate

$\mathrm{E}[X_i] = np_i \geq 1$ for all $i = 1, \ldots, k$,

$\mathrm{E}[X_i] = np_i \geq 5$ for 80% of all $i = 1, \ldots, k$,
Especially if the p_i are not known roughly beforehand, care needs to be taken to ensure that the sample size n is sufficiently large so that these criteria can apply.

Categorical Test on Discrete Distribution

Test if the data follows a particular discrete distribution with m parameters estimated from the given data

Divide our data into categories, use the discrete distribution to estimate the parameters, calculate the expected count of each category.

$$X_{k-1-m}^2 = \sum_{1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

Independence of Category

Test if the row and column categorizations are independent,

$$H_0 : p_{ij} = p_i \cdot p_{\cdot j}$$

We can now compare the observed frequencies O_{ij} in the (i, j) th cell to the expected frequencies E_{ij}.

$$X^2_{(r-1)(c-1)} = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$
$$E_{ij} = n \cdot \widehat{p_{ij}} = \frac{n_i \cdot n_{\cdot j}}{n}$$

We reject H_0 if the value of $X^2_{(r-1)(c-1)}$ exceeds the critical value of the corresponding chi-squared distribution.

Linear Regression

Simple Linear Regression

$Y \mid x = \beta_0 + \beta_1 x + E$ where $\mathrm{E}[E] = 0$

E: remainder. $E[E] = 0$ is guaranteed because b_0, b_1 are unbiased estimators.

β_0, β_1: true parameter of linear relationship.

b_0, b_1: estimator for β_0, β_1.

B_0, B_1 : statistics for estimators b_0, b_1

Least Square Method

For each measurement y_i find residual e_i

$$Y_i = b_0 + b_1 x_i + e_i$$

error sum of squares:

$$\mathrm{SS_E} := e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^{n} (y_i - (b_0 + b_1 x_i))^2$$

least-squares estimates for β_0 and β_1 is determined by minimizing this sum of squares

$$\begin{aligned} S_{xx} &:= \sum_{i=1}^{n} (x_i - \bar{x})^2, S_{yy} := \sum_{i=1}^{n} (y_i - \bar{y})^2, \\ S_{xy} &:= \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}). \\ b_0 &= \bar{y} - b_1 \bar{x}, \quad b_1 = \frac{S_{xy}}{S_{xx}} \\ \mathrm{SS_E} &= S_{yy} - b_1 S_{xy} \end{aligned}$$

with confidence interval

$$B_1 \pm t_{\alpha/2, n-2} \frac{S}{\sqrt{S_{xx}}}, \quad B_0 \pm t_{\alpha/2, n-2} \frac{S\sqrt{\sum x_i^2}}{\sqrt{n S_{xx}}}$$

Significance Test

We say that a regression is significant if there is statistical evidence that the slope $\beta_1 \neq 0$.

We reject

$$H_0 : \beta_1 = 0$$

at significance level α if the statistic

$$T_{n-2} = \frac{B_1}{S/\sqrt{S_{xx}}}$$

satisfies $|T_{n-2}| > t_{\alpha/2,n-2}$.

Test for Correlation

Coefficient

$$R^2 := \frac{SS_T - SS_E}{SS_T} = \frac{S_x^2}{S_x S_{yy}} = \hat{\rho}_{XY}^2 \text{ (from Paired T-test)}$$

The coefficient R^2 expresses the proportion of the total variation in Y that is explained by the linear model

Let (X, Y) follow a bivariate normal distribution with correlation coefficient $\varrho \in (-1, 1)$. Let R be the estimator for ϱ. Then

$$H_0 : \varrho = 0$$

is rejected at significance level α if

$$\left| \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \right| > t_{\alpha/2,n-2}$$

Lack of Fit Test

Test if the linear regression model is appropriate. Need multiple y data at each x data point!

$$SS_E = SS_{E,pe} + SS_{E,lf}$$

pr: pure error, lf: lack of fitting (X & Y are not linearly related)

$$\mathrm{SS}_{\mathrm{E,pe}} := \sum_{i=1}^{k} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

H_0 : the linear regression model is appropriate
is rejected at significance level α if the test statistic

$$F_{k-2,n-k} = \frac{SS_{E,\text{ lf}}/(k-2)}{SS_{E,pe}/(n-k)}$$

satisfies $F_{k-2,n-k} > f_{\alpha,k-2,n-k}$.

Prediction

$Y \mid x = \beta_0 + \beta_1 x + E$ where $\mathrm{E}[E] = 0$

$100(1-\alpha)\%$ Prediction interval for $Y \mid x$:

$$\widehat{Y \mid x} \pm t_{\alpha/2,n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$

$100(1-\alpha)\%$ prediction interval for $\mu_Y \mid x$:

$$\widehat{\mu_Y \mid x} \pm t_{\alpha/2,n-2} S \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{S_{xx}}}$$