STT 863 Compendium

Kenyon Cavender

October 27, 2019

Baseline Knowledge

<u>Def</u> Boole's inequality

$$P(\cup_{i=1}^{n} A_i) \leq \sum_{i=1}^{n} P(A_i)$$

<u>Def</u> Bonferroni's inequality

$$P(\cap_{i=1}^{n} A_i) \geq 1 - \sum_{i=1}^{n} P(A_i^c)$$

<u>Def</u> A function F is called the cumulative distribution function iff:

a. $lim_{x \to -\infty} F(x) = 0$ and $lim_{x \to \infty} = 1$

b. $F(x)$ is a nondecreasing fn of x

c. $F(x)$ is right-continuous; that is for every number $x_0, lim_{x \downarrow x_0} F(x) = F(x_0)$

<u>Def</u> A PDF of a continuous r.v. x is $f_X(x) = \frac{d}{dx} F_X(x)$:

a. $f_X(x) \geq 0 \ \forall \, x$

b. $\int_{-\infty}^{\infty} f_X(x) dx = 1$

c. $F_X(x) = \int_{-\infty}^{x} f_X(s) ds$

For below, change integrals to sums if the r.v is discrete

<u>Def</u> Expectation

$$\mu := \mathbb{E}(X) = \int_{-infty}^{\infty} x f(x) dx$$

For any function h:

$$\mathbb{E}(h(X)) = \int_{-infty}^{\infty} h(x) f(x) dx$$

<u>Def</u> Variance

$$\sigma^2(X) := \mathbb{E}((X-\mu)^2) = \mathbb{E}(X^2) - \{\mathbb{E}(X)\}^2$$

<u>Def</u> Normal Distribution $X \sim N(\mu, \sigma^2)$
PDF:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \ -\infty < x < \infty$$

<u>Def</u> Covariance is a measure of a linear relationship between X and Y.
$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$

<u>Def</u> Correlation coefficient between X and Y is defined as:
$\rho = Corr(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)Var(Y)}}$

Statistical Inference

<u>**Def**</u> **Sample Mean**

$$\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$$

<u>**Def**</u> **Sample Variance**

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

Remark Properties of $\bar{Y}$ and s^2

a. $\mathbb{E}(\bar{Y}) = \mu$

b. $Var(\bar{Y}) = \frac{\sigma^2}{n}$

c. $\mathbb{E}(s^2) = \sigma^2$

<u>**Def**</u> **Homoscedasticity**: $\sigma^2_{Y|X}$ is fixed across X values.

<u>**Def**</u> **Simple linear regression (SLR)** model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

a. ϵ and X are independent

b. $\mathbb{E}\epsilon_i = 0,\ i = 1, 2, ..., n$

c. $Var(\epsilon_i) = \sigma^2,\ i = 1, 2, ..., n$

d. $Cov(\epsilon_i, \epsilon_j) = 0,\ i \neq j$

<u>**Def**</u> **Normal equations** for the least square method:

$$\sum_{i=1}^{n} Y_i = nb_o + b_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} X_i Y_i = b_o \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2$$

<u>**Def**</u> **Least Square Estimators**

$$b_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

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

$$s^2 = \text{MSE} = \frac{\text{SSE}}{n-2} = \frac{1}{n-2}\sum_{i=1}^{n} e_i^2$$

Theorem Gauss-Markov Under the assumptions of the regression model, the least square estimators b_0 and b_1 are
- linear
- unbiased
- have minimum variance among all unbiased linear estimators of β_0 and β_1

<u>**Def**</u> Some properties of sample estimators:

a. $\sum_{i=1}^{n} e_i = 0$

b. $\sum_{i=1}^{n} \hat{Y}_i = \sum_{i=1}^{n} Y_i$

c. $\sum_{i=1}^{n} X_i e_i = 0$

d. $\sum_{i=1}^{n} \hat{Y}_i e_i = 0$

Def Sampling distribution of b_1

a. $b_1 = \sum_{i=1}^{n} k_i Y_i$ where

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

b. $\mathbb{E}(b_1) = \beta_1$

c. $Var(b_1)$:

$$\sigma^2\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

d. Standard error:

$$s^2\{b_1\} = \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{\text{MSE}}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

Def Sampling distribution of b_0

a. $b_0 = \sum_{i=1}^{n} l_i Y_i$ where

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

b. $\mathbb{E}(b_0) = \beta_0$

c. $Var(b_0)$:

$$\sigma^2\{b_0\} = \sigma^2[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]$$

d. Standard error:

$$s^2\{b_0\} = \text{MSE}[\frac{1}{n} + \frac{\bar{X}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]$$

Prediction in SLR

Remark Estimating $\mathbb{E}(Y_h)$
We estimate $\mathbb{E}(Y_h)$ by $\hat{Y_h} = b_0 + b_1 X_h$
$\mathbb{E}(\hat{Y_h})\beta_0 + \beta_1 X_h = \mathbb{E}(Y_h)$
$Var(Y_h)$:

$$\sigma^2\{Y_h\} = \sigma^2[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]$$

Standard error:

$$s^2\{Y_h\} = \text{MSE}[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]$$

Remark C.I. of $\mathbb{E}(Y_h)$
The $100(1-\alpha)\%$ C.I. of $\mathbb{E}(Y_h)$

$$\hat{Y_h} \pm t_{1-\alpha/2;n-2} s\{\hat{Y_h}\}$$

Where

$$s^2\{\hat{Y_h}\} = \text{MSE}[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]$$

Remark P.I. of $\mathbb{E}(Y_{h(new)})$
The $100(1-\alpha)\%$ C.I. of $\mathbb{E}(Y_h)$

$$\hat{Y_h} \pm t_{1-\alpha/2;n-2} s\{pred\}$$

Where

$$s^2\{pred\} = s^2 + s^2\{\hat{Y_h}\} = \text{MSE}[1 + \frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2}]$$

Brown-Forsythe Test

Test for Heteroskedasticity

a. Let n_1 and n_2 be sample sizes for two groups with $n = n_1 + n_2$

b. Divide the residual sets into two groups with medians $\tilde{e_1}$ and $\tilde{e_2}$

c. Define,

   (a) $d_{i1} = |e_{i1} - \tilde{e_1}|$, $i = 1, ...n_1$
   (b) $d_{j2} = |e_{j2} - \tilde{e_2}|$, $j = 1, ...n_1$

d. Let $\bar{d_1}$ and $\bar{d_2}$ be means of d's from the previous groups, and let

$$s_d^2 = \frac{\sum_{i=1}^{n1}(d_{i1} - \bar{d_1})^2 + \sum_{j=1}^{n2}(d_{j2} - \bar{d_2})^2}{n-2}$$

e. Test Statistic:

$$t_{BF}^* = \frac{\bar{d_1} - \bar{d_2}}{s_d\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

f. Reject the null hypothesis (constant variance) if $|t_{BF}^*| > t_{1-\alpha/2;n-2}$

Breusch-Pagan Test

Assumptions:

a. ϵ_i's are independent and normal

b. If $\sigma_i^2 = Var(\epsilon_i)$, then it satisfies $log\sigma_i^2 = \gamma_0 + \gamma_1 X_i$

Test Procedure:

a. Hypotheses:

   (a) $H_0 : \gamma_1 = 0$
   (b) $H_a : \gamma_1 \neq 0$

b. Regress e_i^2 on X_i's. Let SSR* be the regression SS

c. Test statistic: $\chi_{BP}^2 = \frac{\text{SSR}}{2} \div (\frac{\text{SSE}}{n})^2$, where SSE is the error SS of the original regression.

d. Reject H_0 if $\chi_{BP}^2 > \chi_{1-\alpha;1}^2$