Chapter 7

Ⓐ Correlation tests.

Data are pairs $(x_1, y_1), \ldots, (x_n, y_n)$ and we want to know if there is a (linear) association between them.

Measures of Association

$$Cov(Y, X) = E\{(Y - E(Y))(X - E(X))\}$$

$$Cov(Y, Y) = Var(Y) \quad \uparrow$$

Estimating: $$s_{yx} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})(x_i - \bar{x})$$

Notice $s_{yy} = s_y^2 \equiv s^2$

The problem w/ cov is that it has

both scale (Var) & association at the same time.

$$-1 \leq Cor(Y,X) = \frac{Cov(Y,X)}{\sqrt{Var(Y)\,Var(X)}} \leq 1$$

$$\rho \equiv \rho_{YX}$$

If $\rho = \emptyset$, then there is no (linear) association between Y & X.

<u>Estimating</u>: $r_{YX} = \frac{S_{YX}}{S_Y S_X} \equiv r$

Focus on testing:

$H_0: \rho = \emptyset$
$H_1: \rho \neq \emptyset$

can test $\rho = \rho_0$, see book.

$Y_1, \ldots, Y_n \overset{iid}{\sim} N(\mu_Y, \sigma_Y^2)$
$X_1, \ldots, X_n \overset{iid}{\sim} N(\mu_X, \sigma_X^2)$

also indep between samples b/c $H_0: \rho = \emptyset$

From Chap 5:

$\sigma_y^2 \sim (n-1) s_y^2 / \chi^2_{n-1}$ $\quad$ $\sigma_x \sim (n-1) s_x^2 / \chi^2_{n-1}$

The math is hard. It can be shown that, under $H_0: \rho = 0$

$$u = \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \sim t_{n-2}$$

Chapter 7

(SLR)

Ⓑ Linear models / Simple Linear Regression

Linear <u>model</u>

$$Y_i = \underset{\text{intercept}}{\beta_0} + \underset{\text{slope}}{\beta_1} x_i + \underset{\text{errors}}{\varepsilon_i}, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

or $\quad Y_i \sim N(\underbrace{\beta_0 + \beta_1 x_i}_{\mu_i}, \sigma^2) \quad i = 1, \dots, n$

How to dial in estimates of β_0, β_1 & σ^2?

→ MLE

$$f(y_1, \dots, y_n; \beta_0, \beta_1, \sigma^2) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(y_i - \overbrace{\beta_0 - \beta_1 x_i}^{\mu_i})^2\right\}$$

$$= (2\pi)^{-n/2} (\sigma^2)^{-n/2} \exp\left\{\frac{-1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2\right\}$$

$$l(\beta_0, \beta_1, \sigma^2) = c - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \underbrace{\beta_0 - \beta_1 x_i}_{\mu_i})^2$$

(OLS) ordinary least squares

find MLE

$$\phi \overset{set}{=} \frac{\partial l}{\partial \beta_0} \cancel{\frac{-1}{2\sigma^2}} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

$$= \cancel{2} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) \cancel{(-1)}$$

$$\emptyset = \sum_i y_i - n\beta_0 - \beta_1 \sum x_i$$

$$\beta_0 = \frac{1}{n} \sum y_i - \beta_1 \frac{1}{n} \sum x_i$$

$$\boxed{\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}}$$

"means on the line"

$$\phi \overset{set}{=} \frac{\partial l}{\partial \beta_1} = \frac{\partial l}{\partial \beta_1} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \beta_1 x_i)^2$$

$\vdots$ (see book)

$$\boxed{\hat{\beta}_1 = \frac{s_{yx}}{s_x^2} = r_{yx} \cdot \frac{s_y}{s_x}}$$

connecting SLR to corr ← correlation × rise / run

$\emptyset \overset{set}{=} \frac{\partial l}{\partial \sigma^2}$ exactly as in Ch 3

w/ $\hat{\mu}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$\hat{y}_i \equiv$ fitted value

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (\underbrace{y_i - \hat{\mu}_i}_{\text{residual}})^2$$

more common to use

$$\boxed{s^2 = \frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}$$

I want to test hypotheses about estimated parameters.

$H_0 : \beta_1 = \emptyset$

$H_1 : \beta_1 \neq \emptyset$

} any other value of β_1 (e.g. $\beta_1 = 1$)

Ch 3: $\sigma^2 \sim (n-2) s^2 / \chi^2_{n-1}$

The math way is more work, but doable.

$$\hat{B}_1 = \frac{s_{yx}}{s_x^2} = \frac{\sum_i (x_i-\bar{x})(y_i-\bar{y})}{\sum_i (x_i-\bar{x})^2} = \sum_i \underbrace{\frac{(x_i-\bar{x})}{\sum_i (x_i-\bar{x})^2}}_{w_i} y_i$$

$$\hat{B}_1 = \sum_{i=1}^{n} w_i Y_i$$

$$E\{\hat{B}_1\} = \sum_{i=1}^{n} w_i \overset{B_0+B_1x_i}{E\{Y_i\}}$$

$$= \overset{\phi}{\cancel{B_0 \sum w_i}} + B_1 \overset{1}{\cancel{\sum w_i x_i}} + \sum w_i \overset{\phi}{\cancel{E\{\varepsilon_i\}}}$$

$$= \phi + B_1 + 0$$

$E\{\hat{B}_1\} = B_1$ <u>unbiased</u>

$$Var\{\hat{B}_1\} = \sum_{i=1}^{n} w_i^2 \overbrace{Var\{Y_i\}}^{\sigma^2}$$

$$= \sigma^2 \sum w_i^2 = \boxed{\frac{\sigma^2}{(n-1)\underline{s_x^2}} = Var\{\hat{B}_1\}}$$

$$\boxed{\hat{B}_1 \sim N\left(B_1, \sigma^2/(n-1)s_x^2\right)}$$

use a t-test by replacing σ^2 w/ s^2

$$t_1 = \frac{\hat{\beta}_1 - \beta_1}{s_{b_1}} \quad \text{w/} \quad s_{b_1} = \frac{s^2}{(n-1) s_x^2}$$

& $T_1 \sim t_{n-2}$.

Chapter 7

(C) Prediction for SLR

Suppose you had a new x location where you wtk $Y(x) \mid (x_1, y_1), \ldots, (x_n, y_n)$

Could predict $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$

generic

Understanding uncertainty in SLR is like a big CI calculation.

CI & PI

How do you understand $Y(x)$ w/ math?

Denote $\hat{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$ w/ MLEs

& $\tilde{Y}(x) = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$, $\varepsilon \overset{iid}{\sim} N(0, \sigma^2)$

Consider $\mathbb{E}\{\tilde{Y}(x)\} = \mathbb{E}\{\hat{\beta}_0\} + \mathbb{E}\{\hat{\beta}_1\}x + \underbrace{\cancel{\mathbb{E}\{\varepsilon\}}}_{0}$

(unbiased) $= \beta_0 + \beta_1 x = \mathbb{E}\{Y(x)\}$

$\mathbb{E}\{\hat{Y}(x)\} = \mathbb{E}\{\tilde{Y}(x)\} = \mathbb{E}\{Y(x)\}$ ✓

so unbiased prediction.

$$\mathbb{V}ar\{\hat{Y}(x)\} = \underbrace{\mathbb{V}ar\{\hat{\beta}_0\}}_{\sigma^2_{b_0}} + \underbrace{\mathbb{V}ar\{\hat{\beta}_1\}}_{\sigma^2_{b_1}} x^2 + 2x \underbrace{Cov(\hat{\beta}_0, \hat{\beta}_1)}_{-\sigma^2\left(\frac{\bar{x}}{(n-1)s_x^2}\right)}$$

from Ⓑ

$$\underline{s^2_{fit}(x)} = \sigma^2\left(\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}\right)$$ ✓ Ch13

replace σ^2 with s^2 & use t_{n-2}

Now, with full uncertainty, including ε

$$\mathbb{V}ar\{\tilde{y}(x)\} = \mathbb{V}ar\{\hat{y}(x)\} + \mathbb{V}ar\{\varepsilon\}$$

$$\sigma^2_{pred}(x) = \sigma^2_{fit}(x) + \sigma^2$$

use $\underline{s^2_{pred}}(x) = s^2\left(1 + \overbrace{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}}^{s^2_{fit}}\right)$