Inference about Regression Parameters

Under our multiple linear regression model,

$$\underset{\sim}{b} = (X'X)^{-1}X'\underset{\sim}{y}$$

* are unbiased $E\{\underset{\sim}{b}\} = \underset{\sim}{\beta}$ (equivalent to $\begin{pmatrix} E\{b_0\} \\ E\{b_1\} \\ \vdots \\ E\{b_{p-1}\} \end{pmatrix} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix}$)

* $$\sigma^2\{\underset{\sim}{b}\} = \begin{pmatrix} \sigma^2\{b_0\} & \sigma\{b_0, b_1\} & \cdots & \sigma\{b_0, b_{p-1}\} \\ \sigma\{b_0, b_1\} & \sigma^2\{b_1\} & \cdots & \sigma\{b_1, b_{p-1}\} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma\{b_0, b_{p-1}\} & \sigma\{b_1, b_{p-1}\} & \cdots & \sigma^2\{b_{p-1}\} \end{pmatrix}$$

$$= \sigma^2 (X'X)^{-1}$$

* can get an estimate of $\sigma^2\{\underset{\sim}{b}\}$, $s^2\{\underset{\sim}{b}\} = \underset{(MSE)}{s^2}(X'X)^{-1}$

$$s^2\{\underset{\sim}{b}\} = s^2 (X'X)^{-1} = \begin{pmatrix} s^2\{b_0\} & s\{b_0, b_1\} & \cdots & s\{b_0, b_{p-1}\} \\ s\{b_0, b_1\} & s^2\{b_1\} & \cdots & s\{b_1, b_{p-1}\} \\ \vdots & \vdots & \ddots & \vdots \\ s\{b_0, b_{p-1}\} & s\{b_1, b_{p-1}\} & \cdots & s^2\{b_{p-1}\} \end{pmatrix}$$

Can use $s\{\underset{\sim}{b}\}$ for:
* Interval estimation of β_k
* tests for β_k
* interval estimation for fitted values

* Interval estimation of β_k

under our normal multiple linear regression model

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim t_{n-p} \quad \text{for } k = 0, 1, \ldots, p-1$$

confidence limits for β_k with $1-\alpha$ confidence coefficient are

$$\left(b_k + t(\alpha/2, n-p)\, s\{b_k\},\ b_k + t(1-\alpha/2, n-p)\, s\{b_k\}\right)$$

* Interval estimation of β_k

Under our normal multiple linear regression model

$$\frac{b_k - \beta_k}{s\{b_k\}} \sim t_{n-p} \quad \text{for } k=0,1,\dots,p-1$$

confidence limits for β_k with $1-\alpha$ confidence coefficient are $\left(b_k + t(\alpha/2, n-p)\, s\{b_k\},\ b_k + t(1-\alpha/2, n-p)\, s\{b_k\}\right)$

same as $\left(b_k - t(1-\alpha/2, n-p)\, s\{b_k\},\right.$

$\left. b_k + t(1-\alpha/2, n-p)\, s\{b_k\}\right)$ because t-distribution is symmetric about 0,

$t(\alpha/2, n-p) = -t(1-\alpha/2, n-p)$

* Tests for β_k

To test $H_0: \beta_k = 0$ versus $H_a: \beta_k \neq 0$

we use the test statistic $t^* = b_k / s\{b_k\}$

Our decision rule for a level-α test will be

* If $|t^*| \leq t(1-\alpha/2, n-p)$ then conclude H_0
* If $|t^*| > t(1-\alpha/2, n-p)$ then conclude H_a

We can also use a type of F-test to test H_0 versus H_a

<u>but</u> we're going to wait until chapter 7 for this!

* interval estimation for fitted values

Under our normal multiple linear regression model, given values of $X_1, X_2, \ldots, X_{p-1}$ for a new observation denoted by $X_{h1}, X_{h2}, \ldots, X_{hp-1}$,

$$E\{Y_h\} = \beta_0 + \sum_{k=1}^{p-1} \beta_k X_{hk}$$

average response of observation h

$$= \underset{\sim}{X}_h' \underset{\sim}{\beta} \quad \text{where} \quad \underset{\sim}{X}_h = \begin{pmatrix} 1 \\ X_{h1} \\ X_{h2} \\ \vdots \\ X_{hp-1} \end{pmatrix}$$

$$\hat{Y}_h = \underset{\sim}{X}_h' \underset{\sim}{b} = b_0 + \sum_{k=1}^{p-1} b_k X_{hk}$$

estimated mean response

$$E\{\hat{Y}_h\} = E[Y_h] = \underset{\sim}{X}_h' \underset{\sim}{\beta}$$

$$\sigma^2\{\hat{Y}_h\} = \sigma^2 \underset{\sim}{X}_h' (X'X)^{-1} \underset{\sim}{X}_h$$

$1\times p \times p\times p \times p\times 1 = 1\times 1$

$$= \underset{\sim}{X}_h' \sigma^2\{\underset{\sim}{b}\} \underset{\sim}{X}_h$$

$$S^2\{\hat{Y}_h\} = S^2 \underset{\sim}{X}_h' (X'X)^{-1} \underset{\sim}{X}_h$$

The $1-\alpha$ confidence limits for $E\{\hat{Y}_h\}$ are $(\hat{Y}_h + t(\alpha/2, n-p)\, s\{\hat{Y}_h\},\ \hat{Y}_h + t(1-\alpha/2, n-p)\, s\{\hat{Y}_h\}$

Interval Estimation for future values of the response

$$Y_h = E\{Y_h\} + \varepsilon_h$$

$$\sigma^2\{pred\} = \sigma^2 + \sigma^2\{\hat{Y}_h\}$$

$$s^2\{pred\} = s^2 + s^2\{\hat{Y}_h\} = s^2\left(1 + \underset{\sim}{X}_h'(X'X)^{-1}\underset{\sim}{X}_h\right)$$

The $1-\alpha$ prediction limits for a new observation $Y_{h(new)}$ corresponding to $\underset{\sim}{X}_h$ are

$$\left(\hat{Y}_h + t(\alpha/2, n-p)\, s\{pred\},\ \hat{Y}_h + t(1-\alpha/2, n-p)\, s\{pred\}\right)$$

Interval estimation for average of m future values of the response?

$$\sigma^2\{pred_{mean}\} = \frac{\sigma^2}{m} + \sigma^2\{\hat{Y}_h\}$$

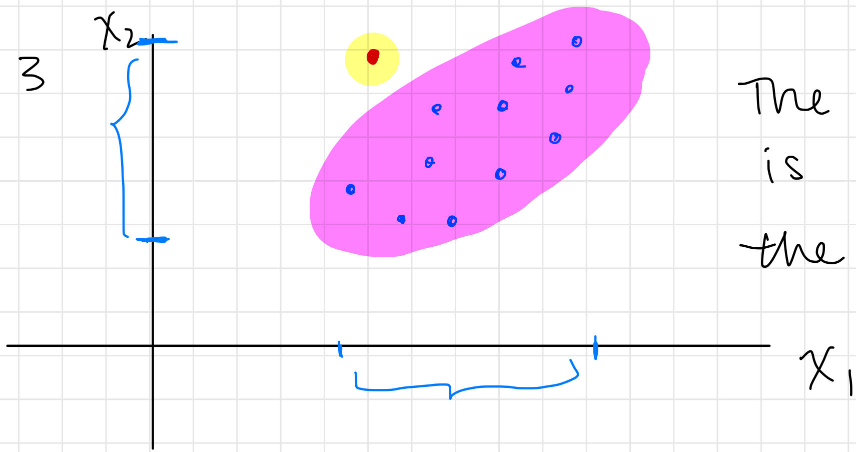
Extrapolation / Defining Scope in Multiple Linear Regression

Say $p=2 \Rightarrow$ simple linear regression

data contains $\underset{\sim}{X} = \begin{pmatrix} 11 \\ 21 \\ 18 \\ 32 \\ 13 \end{pmatrix}$

Is $X_n = 0$ within the scope?

Say $p=3$



The red point is not within the scope of the model

Diagnostics / "Remedial Measures"

Transformations we may need to take and/or additional covariates we may need to include to improve the plausibility of multiple linear regression assumptions

* scatter plots of response and covariates
  - Is linearity reasonable?
  - Is constant variance reasonable?
  - Is our model complete (are the relevant predictors/covariates included)

* scatter plots of residuals from a candidate model and fitted values, observed response values, covariates

Example - Dwaine Photo Studios Example

Y_i : sales in city i

X_{i1}: proportion of population under 17 in city i

X_{i2}: average per capita disposable income in city i

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i \quad , \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$