10/04/16 ALSM 1

ANALYSIS OF VARIANCE (ANOVA IN THE DEPENDENT VARIABLE)

Variation in y can be quantified by error:

$\Sigma (y_i - 0)^2 = \Sigma y_i^2$ Total Uncorrected Sum of Squares SS(uncorrected)

or more usually

$\Sigma (y_i - \bar{y})^2$ Total Corrected Sum of Squares SS(corrected)

$\Sigma (y_i - \bar{y})^2 = \Sigma (y_i^2 - 2 y_i \bar{y} + \bar{y}^2)$

$= \Sigma y_i^2 - 2\bar{y} \Sigma y_i + n\bar{y}^2$

$= \Sigma y_i^2 - 2n\bar{y}^2 + n\bar{y}^2$

$= \Sigma y_i^2 - n\bar{y}^2$

$= SS(\text{uncorrected}) - \text{correction}$

Total corrected $\Sigma (y_i - \bar{y})^2$ = Total Uncorrected Σy_i^2 − Correction $n\bar{y}^2$

Write SS(uncorrected) in terms of fitted value and residual $y_i = \hat{y}_i + \hat{\varepsilon}_i$

$SS(\text{uncorrected}) = \Sigma y_i^2 = \Sigma (\hat{y}_i + \hat{\varepsilon}_i)^2$

$= \Sigma (\hat{y}_i^2 + 2\hat{\varepsilon}_i \hat{y} + \hat{\varepsilon}_i^2)$

$= \Sigma \hat{y}_i^2 + \Sigma \hat{\varepsilon}_i^2 + 2\Sigma \hat{y}_i \hat{\varepsilon}_i$

$\Sigma \hat{y}_i \hat{\varepsilon}_i = \Sigma \hat{y}_i (y_i - \hat{y}_i)$

$= \Sigma \hat{y}_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)$

$= \Sigma \hat{y}_i (y_i - \bar{y} + \hat{\beta}_1 \bar{x} - \hat{\beta}_1 x_i)$

$= \Sigma \hat{y}_i [(y_i - \bar{y}) + \hat{\beta}_1 (x_i - \bar{x})]$

$= \Sigma [\bar{y} + \hat{\beta}_1 (x_i - \bar{x})][(y_i - \bar{y}) + \hat{\beta}_1 (x_i - \bar{x})]$

$= \Sigma [\bar{y}(y_i - \bar{y}) + \bar{y}\hat{\beta}_1 (x_i - \bar{x}) + \hat{\beta}_1 (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 (x_i - \bar{x})(x_i - \bar{x})]$

$= \bar{y} \Sigma (y_i - \bar{y}) + \bar{y}\hat{\beta}_1 \Sigma (x_i - \bar{x}) + \hat{\beta}_1 \Sigma (x_i - \bar{x})(y_i - \bar{y}) + \hat{\beta}_1^2 \Sigma (x_i - \bar{x})^2$

$= \bar{y}(0) + \bar{y}\hat{\beta}_1 (0) + \hat{\beta}_1 S_{xy} - \hat{\beta}_1^2 S_{xx}$

$\frac{S_{xy}}{S_{xx}} S_{xy} - \frac{S_{xy}^2}{S_{xx}^2} S_{xx} = \frac{S_{xy}^2}{S_{xx}} - \frac{S_{xy}^2}{S_{xx}} = 0$

$= \Sigma \hat{y}_i^2 + \Sigma \hat{\varepsilon}_i^2$

$SS(\text{uncorrected}) = SS(\text{Model}) + SSE$

10/04/16

SS(Model) $\sum \hat{y}_i^2 = \sum (\hat{y}_i - \bar{y} + \bar{y})^2$

$= \sum [(\hat{y}_i - \bar{y})^2 + 2\bar{y}(\hat{y}_i - \bar{y}) + \bar{y}^2]$

$= \sum (\hat{y}_i - \bar{y})^2 + 2\bar{y} \sum (\hat{y}_i - \bar{y}) + n\bar{y}^2$

$\bar{y} \sum (\hat{y}_i - \bar{y}) = \bar{y} \sum (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})$

$= \bar{y} \sum (\bar{y} - \hat{\beta}_1 \bar{x}_i + \hat{\beta}_1 x_i - \bar{y})$

$= \bar{y} \sum (x_i - \bar{x}) \hat{\beta}_1$

$= 0$

$\bar{y} \sum (x_i - \bar{x}) = \bar{y} \sum \hat{y}_i - \bar{y}$

$= \bar{y} \sum \hat{y}_i - \bar{y} \sum \bar{y}$

$= \bar{y} \sum y_i - n\bar{y}\bar{y}$

$= 0$

→ $SS(Model) = \sum (\hat{y}_i - \bar{y})^2 + n\bar{y}^2$

$= SS(Regression) + Correction$

The regression sum of squares represents the reduction in variation (around the predicted mean) by adding the $\beta_1 x_i$ term to a model containing only β_0

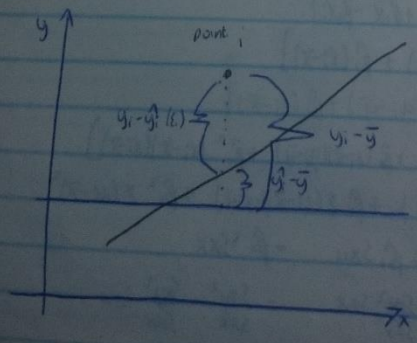
$SS(uncorrected) = SS(Reg) + n\bar{y}^2 + SSE$

$\sum \hat{y}_i^2 = \sum (\hat{y}_i - \bar{y})^2 + n\bar{y}^2 + \sum \varepsilon_i^2$

If we subtract correct $n\bar{y}^2$ from both sides:

$\sum \hat{y}_i^2 - n\bar{y}^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum \varepsilon_i^2$

$SS(corrected) = SS(Reg) + SSE$



$\sum (y_i - \bar{y})^2 = \sum (y - \hat{y})^2 + (\hat{y}_i - \bar{y})^2$

ANOVA

$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$

SS(corrected) = SS(Regression) (model) + SSE (error in general)

Model Comparison

$M_1: y_i = \beta_0 + \varepsilon_i$ $\quad \hat{\beta}_0 = \bar{y}$ $\quad SSE = \sum (y_i - \bar{y})^2$

$M_2: y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $\quad \hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ $\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$ $\quad SSE = \sum (y_i - \hat{y}_i)^2$ $\quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

M_1 less restricted than M_2. $SSE_2 \leq SSE_1$

$SSE_1 - SSE_2$ measure of how much better M_2 is than M_1

$SSE_1 - SSE_2 = \sum (y_i - \bar{y})^2 - \sum (y_i - \hat{y}_i)^2$

$= SS(\text{corrected}) - SSE_2$

$= SS(\text{Regression})$

$SS(\text{Reg})$ indicates importance of $\beta_1 x_i$ term in the model

$SS(\text{Reg}) = R(\beta_1 | \beta_0)$

$M_0: y_i = \varepsilon_i$ $\quad SSE_0 = \sum (y_i - 0)^2 = \sum (y_i^2)$

$M_1: y_i = \beta_0 + \varepsilon_i$ $\quad SSE_1 = \sum (y_i - \bar{y})^2$

$\Rightarrow SSE_0 - SSE_1 \rightarrow \sum (y_i^2) - \sum (y_i - \bar{y})^2$

$= SS(\text{uncorrected}) - SS(\text{corrected}) = n\bar{y}^2$ (correction)

Hence $n\bar{y}^2$ represents importance of β_0 in the mode $R(\beta_0) = n\bar{y}^2$

$M_0: y_i = \varepsilon_i$ $\quad SSE = \sum y_i^2$

$M_2: y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + \varepsilon$ $\quad SSE = \sum (y_i - \hat{y}_i)^2$

$SSE_0 - SSE_2 = \sum y_i^2 - \sum (y_i - \hat{y}_i)^2$

$= SS(\text{uncorrected}) - SSE$

$= SS(\text{Reg}) + n\bar{y}^2$

$= SS(\text{Model})$

- Importance of $\beta_0 + \beta_1$ in model: $R(\beta_1, \beta_0) = SS(\text{Model}) = R(\beta_0) + R(\beta_1 | \beta_0)$

Degrees of Freedom

SS (Uncorrected)	Σy_i^2	n	
SS (Corrected)	$\Sigma(y_i-\bar{y})^2 \rightarrow \Sigma y_i^2 - n\bar{y}^2$	$n-1$	
Correction	$n\bar{y}^2$	1	
SS (Model) $R(\beta_0,\beta_1)$	$\Sigma(\hat{y}_i-\bar{y})^2 + n\bar{y}^2$	2	(1 for each param)
SS (Regression)	$\Sigma(\hat{y}_i-\bar{y})^2$	1	(1 for β_1)
SSE	$\Sigma\varepsilon_i^2 \Rightarrow \Sigma(y_i-\hat{y}_i)^2$	$n-2$	d.f.

SS(corrected) = SS(uncorrected) − correction
SS (uncorrected) = SS(Model) + SSE
SS (Model) = SS(Regression) + correction
SS (uncorrected) = SS(Regression) + correction + SSE
SS (corrected) = SS(Regression) + SSE

$SS(Reg) = \Sigma(\hat{y}_i - \bar{y})^2 = \hat{\beta}_1 S_{xy}$ $\quad \hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

$= \Sigma(\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2$

$= \Sigma(\bar{y} - \hat{\beta}_1\bar{x} + \hat{\beta}_1 x_i - \bar{y})^2$

$= \Sigma(\hat{\beta}_1(x_i - \bar{x}))^2$

$= \hat{\beta}_1^2 \Sigma(x_i - \bar{x})^2$

$= \hat{\beta}_1^2 (S_{xx})$

$= \frac{S_{xy}^2}{S_{xx}^2} S_{xx} = \hat{\beta}_1 S_{xy}$

Source	DF	SS	Mean Square $SS/DF=$
Regression on X	1	$\Sigma(\hat{y}_i-\bar{y})^2$	MS(Reg)
Residual error	$n-2$	$\Sigma(y_i-\hat{y}_i)^2$	MSE
Total corrected	$n-1$	$\Sigma(y_i-\bar{y})^2$	

15/04/16 MSM1 SAMPLE PAPER

Q1 A. $y_i = \mu + \varepsilon_i \quad i = 1, \ldots, N$

$Q(\mu) = \sum (y_i - \mu)^2$ For SLR: $\sum (y_i - \beta_0 - \beta_1 x_i)^2$

B. LS estimator of μ

$$\frac{dQ}{d\mu} = \sum_{i=1}^{N} (-2)(y_i - \mu)$$

$$= -2 \sum (y_i - \mu) = 0$$

$$\sum y_i - n\hat{\mu} = 0$$

$$n\hat{\mu} = \sum y_i$$

$$\hat{\mu} = \sum y_i / N$$

$$\hat{\mu} = \bar{y}$$

C. SSE: Sum of Squared Errors $= \sum (y_i - \hat{y}_i)^2$ deviations around fitted mean

Usually SSE $= \sum (y_i - \hat{y})^2$ $\hat{y}_i$ here is $y_i = \mu + \varepsilon_i$ $\hat{y}_i = \mu$ $\hat{y} = \bar{y}$

MSE = SSE / d.f. Here we will have $n-1$ d.f. because we have only estimated one parameter μ of this model

D Model with a "day effect"

→ n observations from the First day

→ m observations from the Second day

"Day Effect" introduce a variable $x_i = \begin{cases} 0 & 1 \le i \le n \\ 1 & n+1 \le i \le n+m \end{cases}$

Use the model: $y_i = \mu + \mu_d x_i + \varepsilon_i \quad 1 \le i \le n+m$

First day: $y_i = \mu + \varepsilon_i \quad i = 1, \ldots, N$

Second day: $y_i = \mu + \mu_d + \varepsilon_i \quad n+1 \le i \le n+m$

- There will be a day effect if $\mu_d \neq 0$
- We want to test $H_0: \mu_d = 0$ v $H_1: \mu_d \neq 0$ For a day effect

$Z4 = \begin{bmatrix} -1 & \\ 0 & 0 \end{bmatrix}$ β_1 a) increase

- H_0 can be carried out using a t-test

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

- All quantities from LS SLR model - compare $|t|$ with $t_{n+m-2, \alpha/2}$ and reject H_0 if $|t| > t_{n+m-2, \alpha/2}$

15/04/16 ALSM 1

Exam Questions

Q2 A. $E[y_i] = \beta_0 + \beta_1 x_i$

$\hat{y}_i$ at x_i

$\hat{y}_i$ is the predicted mean value of y at x_i

$\hat{y}_i$ is a predictor of $E[y|x=x_i] = \beta_0 + \beta_1 x_i$

It's given by $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$

B. Show $\Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2$

$SS(\text{corrected}) = SS(\text{Reg}) + SSE$

$$\Sigma(y_i - \bar{y})^2 = \Sigma(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$$

$$= \Sigma\left((y_i - \hat{y}_i)^2 + (\hat{y}_i - \bar{y})^2 + 2(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})\right)$$

$$= \Sigma(y_i - \hat{y}_i)^2 + \Sigma(\hat{y}_i - \bar{y})^2 + 2\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y})$$

↑ need to show that this $= 0$

$$\Sigma(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \Sigma(y_i - b_0 - b_1 x_i)(b_0 + b_1 x_i - \bar{y})$$

$$= \Sigma(y_i - \bar{y} + b_1\bar{x} - b_1 x_i)(\bar{y} - b_1\bar{x} + b_1 x_i - \bar{y})$$

$$= \Sigma\left[(y_i - \bar{y}) - b_1(x_i - \bar{x})\right]\left[b_1(x_i - \bar{x})\right]$$

$$= b_1 \Sigma(x_i - \bar{x})(y_i - \bar{y}) - b_1^2 \Sigma(x_i - \bar{x})^2$$

$$= b_1 S_{xy} - b_1^2 S_{xx} = \frac{S_{xy}}{S_{xx}} S_{xy} - \frac{S_{xy}^2}{S_{xx}^2} S_{xx} = 0$$

$$\therefore \Sigma(y_i - \bar{y})^2 = \Sigma(\hat{y}_i - \bar{y})^2 + \Sigma(y_i - \hat{y}_i)^2$$

C $SS(\text{Corrected})$ has $n-1$ d.f. since this is like the SSE of a simple mean model

$SSE = \Sigma(y_i - b_0 - b_1 x_i)^2$ has $n-2$ d.f, one for each b_1, b_2

The remaining 1 d.f. is for $SS(\text{Reg})$

D

Soure	DM SS	D.F.	MS	F
Regression	$\Sigma(y_i - \bar{y})^2$	1	SS/D.F	SSReg/[illegible] MSE
Error	$\Sigma(y_i - \hat{y}_i)^2$	$n-2$	SS/D.F	
Total	$\Sigma(y_i - \bar{y})^2$	$n-1$		

Aside: Why 1 D.F for $SS(Reg)$?

$$\sum(\hat{y}_i - \bar{y})^2 = \sum(b_0 + b_1 x_i - \bar{y})^2$$

$$= \sum(\bar{y} - \hat{b}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2$$

$$= \hat{\beta}_1^2 \sum(x_i - \bar{x})^2 = S_{xy}^2 / S_{xx}$$

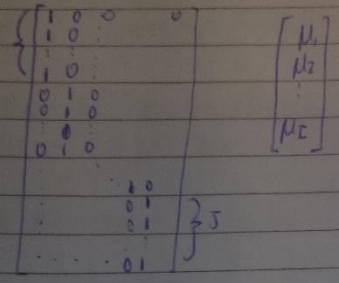
↳ only comes as an end product of other processes and therefore is the single amalgamation of other processes which is why it has 1 d.f

15/04/16 ALSM I

EXAM QUESTION 1

Q3 A. One way classification model

Design Matrix



LS estimates of $\mu_1, \ldots, \mu_I$

$$Q(\mu_1, \ldots, \mu_I) = \sum_i \sum_j (y_{ij} - \mu_i)^2$$

$$\frac{dQ}{d\mu_i} = \sum_{j=1}^{J} (-2)(y_{ij} - \mu_i) = 0$$

$$\sum_j y_{ij} - J\mu = 0$$

$$\hat{\mu}_i = \sum y_{ij} / J$$

$$= \bar{y}_{i\cdot}$$

ANOVA for Model?

SS(corr) for model? $\sum_i \sum_j y_{ij}/IJ = \bar{y}_{\cdot\cdot}$

$$SS(\text{corr}) = \sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2.$$

Within a population LS estimate of μ_i is $\bar{y}_{i\cdot}$

$$\sum_i \sum_j (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

$$= \sum_i \sum_j [(y_{ij} - \bar{y}_{i\cdot})^2 + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + \underbrace{2(y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})}_{=0}$$

$$= \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2 + J \sum (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$$

B. $H_0: L\mu = 0$ I different intercepts, one for each population

$H_1: \mu_1 = \mu_2 = \ldots = \mu_J$ implies that $\mu_i - \mu_j = 0$ for $i \neq j$

Write L as $\begin{bmatrix} 1 & -1 & & \\ 1 & 0 & -1 & \\ 1 & 0 & 0 & -1 \\ & & & \ddots \end{bmatrix}$ $J-1$

C. An F-Test

$$F = \frac{MS(Reg)}{MSE} = \frac{R(\beta_1, \beta_2 | \beta)/2(J-1)}{MSE}$$

If H_0 is true, i.e. $\mu_1 = \mu_2 = 0$ then F follows an F-distribution with I and $2(J-1)$ D.F.

If $F > F_{2,2(J-1)}$ then reject H_0 at $100(1-\alpha)\%$ significance level.

15/04/16 ALSM 1

EXAM QUESTIONS

Q4 A $\underline{y}$ $n \times 1$ response

X $n \times (p+1)$ Design Matrix

$\underline{\beta}$ $(p+1) \times 1$

$\underline{\beta} = (X^TX)^{-1}X^T\underline{y}$

$\mathbb{E}[\underline{y}] = X\underline{\beta}$ or $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$

Where $\underline{\varepsilon}$ is 0 $n \times 1$ error vector with 0 mean and uncorrelated entries

B The Hat matrix is such that $\underline{\hat{y}} = H\underline{y}$ Here $\underline{\hat{y}} = X\underline{\hat{\beta}}$

$\underline{\hat{y}} = X(X^TX)^{-1}X^T\underline{y}$

$H = X(X^TX)^{-1}X^T$

C Vector of residuals

$\underline{\hat{\varepsilon}} = \underline{y} - \underline{\hat{y}} = \underline{y} - H\underline{y}$

$= \underline{y} - X\underline{\hat{\beta}}$

D SSE $= \sum_{i=1}^{N} \hat{\varepsilon}_i^2 = \underline{\hat{\varepsilon}}^T\underline{\hat{\varepsilon}}$

$= (\underline{y} - X\underline{\hat{\beta}})^T(\underline{y} - X\underline{\hat{\beta}})$

$= \underline{y}^T\underline{y} - \underline{y}^TX\underline{\hat{\beta}} - \underline{\hat{\beta}}^TX^T\underline{y} + \underline{\hat{\beta}}^TX^TX\underline{\hat{\beta}}$

$= \underline{y}^T\underline{y} - \underline{y}^TH\underline{y} - \underline{y}^TH\underline{y} + \underline{y}^TH^TH\underline{y}$ $\quad H^TH = H$

$= \underline{y}^T\underline{y} + \underline{y}^TH\underline{y} = \underline{y}^T(I-H)\underline{y}$

$H = X(X^TX)^{-1}X^T$

$H^T = (X^T)^T(X^TX)^{-1}X^T = X(X^TX)^{-1} = H \Rightarrow$ Symetric Matrix

$H^TH = HH = X(X^TX^{-1})X^TX(X^TX)^{-1}X^T$

$= X(X^TX)^{-1}X^T = H$

Q5 Algae grown y concentration of Chemical x

P value for x=0.15
This is the p-value for testing the hypothesis, Slope of LR line: [H_0: $\beta=0$, H_1: $\beta \neq 0$]

Concentration decrease algae growth?
Correlation of Sample, Correlation between x and y, of 0.5, large?

- Correlation suggests y and x are positively related, is 0.5 actually correct correlation?
- What statistical software are you using?
- Does it have default settings for hypothesis testing? i.e. SLR?

- Assuming the default test (relating to x above, p=0.15) is whether the slope is equal to zero, a p-value of 0.15 may suggest you make tentative conclusions without further investigation
- You should carry out regression diagnostics, have you considered standardised residuals or normality plots?

17/05/16 ALSM I EXAM NOTES : SLR, ANOVA

SLR

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ $\quad \mathbb{E}[y|x] = \hat{\beta_0} + \hat{\beta} x_i$

1. For a fixed value of x, y is a random variable with a finite mean and variance
   $|\mathbb{E}[y|x]| < \infty$ $\quad Var[y|x] < \infty$ (i.e. they exist)

2. The values of y are uncorrelated

3. The model is linear i.e $\mathbb{E}[y|x] = \beta_0 + \beta_1 x_i$ $\quad \mu_{y|x} = \beta_0 + \beta_1 x_i$

4. Homoscedacity - Variance of y is not dependent on x, ie variance is constant
   $Var[y|x] = Var[y] = \sigma^2$

Statement of model:

- Data $(x_1, y_1) \ldots (x_n, y_n)$
- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
- $\mathbb{E}[\varepsilon_i] = 0 \quad \forall i$
- $Var[\varepsilon_i] = \sigma^2 \ \forall i$ (Homoscedastic)
- $Cov[\varepsilon_i, \varepsilon_j] = 0$ if $i \neq j$ y_i's uncorrelated

Least Squares

$SS[error] = SSE = \sum \hat{\varepsilon_i}^2 = \sum (y_i - \hat{y_i})^2 = Q(\hat{\beta_0}, \hat{\beta})$

$\frac{dSSE}{d\beta_0} = -2 \sum (y_i - \hat{\beta_0} - \beta_1 x_i)$ $\quad \hat{\beta_0} = \bar{y} - \hat{\beta_1} \bar{x}$

$\frac{dSSE}{d\beta_1} = -2 \sum (x_i)(y_i - \beta_0 - \beta_1 x_i)$ $\quad \hat{\beta_1} = \frac{S_{xy}}{S_{xx}}$

$S_{xy} = \sum (x_i)(y_i - \bar{y}) \ / \ \sum (x_i - \bar{x})(y_i - \bar{y}) \ / \ \sum (y_i)(x_i - \bar{x}) \ / \ \sum x_i y_i - \frac{\sum x_i \sum y_i}{n}$

$S_{xx} = \sum (x_i)(x_i - \bar{x}) \ / \ \sum (x_i - \bar{x})(x_i - \bar{x}) \ / \ \sum (x_i)^2 - \frac{\bar{x}^2}{n}$

$SSE = \sum (y_i - \hat{\beta_0} - \hat{\beta_1} x_i)$ can be written: $S_{yy} - S_{xy}^2 / S_{xx}$

ANOVA

$SS(\text{Uncorrected}) = \sum y_i^2$

$SS(\text{Corrected}) = \sum (y_i - \bar{y})^2$ $\qquad SS(\text{Corrected}) = SS(\text{Uncorrected}) - \text{Correction}$

$= \sum (y_i^2 - 2y_i\bar{y} - \bar{y}^2)$

$= \sum y_i^2 - 2n\bar{y}^2 - n\bar{y}^2$

$= \sum y_i^2 - n\bar{y}^2$

$SS(\text{Corrected}) - \text{Correction}$

$SS(\text{Uncorrected}) = \sum y_i^2 = \sum (\hat{y}_i + \hat{\varepsilon}_i)^2$

$= \sum \hat{y}_i^2 + \sum \hat{\varepsilon}_i^2 + \sum \hat{y}_i \hat{\varepsilon}_i$ $\quad \sum \hat{y}_i \hat{\varepsilon}_i \to 0$

$= \sum \hat{y}_i^2 + \sum \hat{\varepsilon}_i^2$

$SS(\text{Model}) \quad SSE$

$SS(\text{Model}) = \sum \hat{y}_i^2$

$SSE = \sum (y_i - \hat{y}_i)^2$

$SS(\text{Model}) = \sum \hat{y}_i^2 = \sum [\hat{y}_i - \bar{y} + \bar{y}]^2$

$= \sum (\hat{y}_i - \bar{y})^2 + 2\bar{y} \sum (\hat{y}_i - \bar{y}) + n\bar{y}^2$

$= 0$

$= \sum (\hat{y}_i - \bar{y})^2 + n\bar{y}^2$

$SS(\text{Regression}) + \text{Corr}^n$

$SS(\text{Regression}) = \sum (\hat{y}_i - \bar{y})^2$

Source	SS	DF	
SS(Uncorrected)	$\sum y_i^2$	n	
SS(Corrected)	$\sum (y_i - \bar{y})^2 = \sum y_i^2 - n\bar{y}^2$	$n-1$	
Correction	$n\bar{y}$	1	importance of β_0 in model
SS(Model) $R(\beta_0, \beta_1)$	$\sum (\hat{y}_i - \bar{y})^2 + n\bar{y}^2$	2	(1 for each parameter) importance of β_0 and β_1
SS(Regression)	$\sum (\hat{y}_i - \bar{y})^2$	1	1 for β_1 importance of β_1 term in model
SSE	$= \sum \hat{\varepsilon}_i^2 \quad \sum (y_i - \hat{y}_i)^2$	$n-2$	

Regression on X	1	$\sum (\hat{y}_i - \bar{y})^2$	MS(Reg)
Residual Error	$n-2$	$\sum (y_i - \hat{y}_i)^2$	MSE
Total (Uncorrected	$n-1$	$\sum (y_i - \bar{y})^2$	

10/5/16 ALSM I EXAM NOTES: ANOVA

SS(Corrected) = SS(Uncorrected) − Correction $R^2 = \frac{SS(Reg)}{SS(Corrected)}$

SS(Uncorrected) = SS(Model) + SSE

SS(Model) = SS(Reg) + Correction

SS(Uncorrected) = SS(Reg) + Correction + SSE

SS(Corrected) = SS(Reg) + SSE

Show $\underset{SS(Corrected)}{\sum(y_i-\bar{y})^2} = \underset{SSE}{\sum(y_i-\hat{y}_i)^2} + \underset{SS(Reg)}{\sum(\hat{y}_i-\bar{y})^2}$

$\sum(y_i-\bar{y})^2 = \sum(y_i-\hat{y}_i+\hat{y}_i-\bar{y})^2$

$= \sum(y_i-\hat{y}_i)^2 + (\hat{y}_i-\bar{y})^2 + 2(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$

$= \sum(y_i-\hat{y}_i)^2 + \sum(\hat{y}_i-\bar{y})^2 + 2\sum(y_i-\hat{y}_i)(\hat{y}_i-\bar{y})$

$\sum(y_i-\hat{\beta}_0-\hat{\beta}_1x_i)(\hat{\beta}_0+\hat{\beta}_1x_i-\bar{y})$ $b_0 = \bar{y}-b_1\bar{x}$

$\sum(y_i-\bar{y}+b_1\bar{x}-\beta_1x_i)(\bar{y}-b_1\bar{x}+\beta_1x_i-\bar{y})$

$\sum[(y_i-\bar{y})+b_1(x_i-\bar{x})][b_1(x_i-\bar{x})]$

$= b_1\sum(y_i-\bar{y})(x_i-\bar{x}) + b_1^2\sum(x_i-\bar{x})^2$

$\frac{S_{xy}}{S_{xx}}S_{xy} - \frac{S_{xy}^2}{S_{xx}^2}S_{xx} = 0$

$= \sum(y_i-\hat{y}_i)^2 + \sum(\hat{y}_i-\bar{y})^2$ ✓

Properties

$E[\hat{\beta}_1] = \beta_1$ $Var[\hat{\beta}_1] = \sigma^2/S_{xx}$ $SE(\hat{\beta}_1) = \sqrt{\frac{MSE}{S_{xx}}}$

$E[\hat{\beta}_0] = \beta_0$ $Var[\hat{\beta}_0] = \sigma^2\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)$ $SE(\hat{\beta}_0) = \sqrt{MSE\left(\frac{1}{n}+\frac{\bar{x}^2}{S_{xx}}\right)}$

$Cov[\hat{\beta}_0,\hat{\beta}_1] = \frac{-\bar{x}\sigma^2}{S_{xx}}$

If p value below 0.05 ⇒ reject H_0 $P[-t_{n-2,1-\alpha/2} \le t_{n-2} \le t_{n-2,1-\alpha/2}] = 1-\alpha$

CI for σ^2? $\frac{(n-2)MSE}{\sigma^2} \sim \chi^2_{n-2}$ use this (χ^2 not a symmetric distribution)

$\sigma^2 \pm \frac{(n-2)MSE}{\chi^2_{n-2,1-\alpha/2}}$

$Z\Sigma Z' = \begin{vmatrix} 0 & k & 0 \\ -1 & -k & 1 \end{vmatrix}$ β_0 included?

F test

-Non directional

$F = \frac{SS(Reg)/\sigma^2 / 1}{MSE(n-1)/\sigma^2 / n-2} = \frac{SS(Reg)}{MSE} \sim F_{1,n-2}$

Ratio of two chi-square distributions

When H_0 is true $(\beta_1 = 0)$ $SS(Reg)/\sigma^2 \sim \chi^2_1$

If Fobs < Ftable accept H_0

$F \approx t^2$

INFERENCE CONCERNING $E[y|x']$

Point estimate: $\hat{y}' = \hat{\beta}_0 + \hat{\beta}_1 x'$ (mean value)

$Var[\hat{y}'] = \sigma^2 \left(\frac{1}{n} + \frac{(x'-\bar{x})^2}{S_{xx}} \right)$ (important into CI)

Prediction

$y' - \hat{y}'$ denote between predicted and

$E[y' - \hat{y}'] = 0$

$Var[y' - \hat{y}'] = \sigma^2 \left[1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{S_{xx}} \right]$

$\Rightarrow \frac{(y' - \hat{y}') - 0}{\sqrt{MSE\left(1 + \frac{1}{n} + \frac{(x'-\bar{x})^2}{S_{xx}}\right)}} \sim t_{n-2}$

18/05/16 ALSM1 EXAM NOTES: PROBLEM SHEET

1

$y_i = \mu + \varepsilon_i$

$Q(\mu) = \sum (y_i - \mu)^2$

LS: $\sum -2(y_i - \mu)$

$\sum y_i - n\mu = 0$

$\mu = \bar{y}$

SSE = Sum of Squared error = $\sum (y_i - \hat{y})^2 = \sum (y_i - \bar{y})^2$ which is the corrected sum of squares

MSE = SSE / D.F D.F = n-1 because we have only estimated 1 parameter

Day from n First day, m second day $= \begin{cases} 0 & 1 \le i \le n \\ 1 & n+1 \le i \le n+m \end{cases}$

include a variable x_i

We make $y_i = b_0 + b_1 x_i + \varepsilon_i$ $1 \le i \le n+m$

First day: $y_i = \mu + \varepsilon_i$ $i = 1 \ldots n$ Second day: $y_i = \mu + b_1 x_i + \varepsilon_i$ $n+1 \le i \le n+m$

- Need to test $H_0: \beta_1 = 0$ vs $H_1 \neq 0$ for a day effect
- Use t-test $t_0 = b_1 / \sqrt{s^2 \{b_1\}}$
- Compare $|t_0|$ with $t_{n+m-2, \alpha/2}$ and decide if $|t_0| < t_{n+m-2, \alpha/2}$ accept H_0

$E[MSR] = E[\sum (\hat{y}_i - \bar{y})^2] = E[\hat{\beta}_1^2 S_{xx}]$

$= S_{xx} E[\hat{\beta}_1^2]$

$= S_{xx} [Var(\hat{\beta}_1) + [E(\hat{\beta}_1)]^2]$

$= S_{xx} [\sigma^2 / S_{xx} + \beta_1^2]$

$= \sigma^2 + \beta_1^2 S_{xx}$

Q3 $y_i = \alpha_0 + \alpha_1 (x_i - \bar{x}) + \varepsilon_i$

$= \alpha_0 + \alpha_1 x_i - \alpha_1 \bar{x}$

$\beta_0 = \alpha_0 - \alpha_1 \bar{x}$ $\beta_1 = \alpha_1$

L) $\sum (y_i - \alpha_0 - \alpha_1 (x_i - \bar{x}))$

$\frac{dQ}{d\alpha_0} = \sum -2 (y_i - \alpha_0 - \alpha_1 (x_i - \bar{x})) = 0$ $\frac{dQ}{d\alpha_1} = -2 \sum (x_i - \bar{x})(y_i - \alpha_0 - \alpha_1 (x_i - \bar{x}))$

$\sum y_i - n\alpha_0 - \alpha_1 \sum (x_i - \bar{x}) = 0$ $\sum (x_i - \bar{x}) y_i - \alpha_0 \sum (x_i - \bar{x}) - \alpha_1 \sum (x_i - \bar{x})^2 = 0$

$\alpha_0 = \bar{y}$ $\alpha_1 = S_{xy} / S_{xx}$

$$Cov[y_i, \frac{S_{xy}}{S_{xx}}] = \frac{1}{S_{xx}} Cov[y_i, \Sigma(x_i-\bar{x})y_i]$$

$$\frac{1}{S_{xx}} Cov[y_i, (x_1-\bar{x})y_1 + \dots + (x_n-\bar{x})y_n]$$

$$= \frac{1}{S_{xx}} \Sigma(x_i-\bar{x}) Cov[y_i, y_i]$$

$$= \frac{\sigma^2(x_i-\bar{x})}{S_{xx}}$$

p2: Q5 (x_1, y_1) (x_2, y_2) First day (x_{n+1}, y_{n+1}) (x_n, y_n) second day

be can vary day to day b. can't

Day 1: $a_0 + b_1 x_i$ $i = 1, \dots, n_1$

Day 2: $a_2 + b_2 x_i$ $i = n_1+1, \dots, n$ ⋮

$$Q(a_1, a_2, b) = \Sigma \hat{\varepsilon}_i^2 = \sum_{i=1}^{n_1} \hat{\varepsilon}_i^2 + \sum_{i=n_1+1}^{n} \hat{\varepsilon}_i^2$$

$$\sum_{i=1}^{n_1} |y_i - a_1 - bx_i|^2 + \sum_{i=n_1+1}^{n} (y_i - a_2 - bx_i)$$

$$\frac{\partial Q}{\partial b} = -2\Sigma(y_i - a_1 - bx_i)2 \quad ① \qquad \frac{\partial Q}{\partial b} = -2\Sigma(y_i - a_2 - bx_i)^2 \quad ②$$

$$\frac{\partial Q}{\partial b} = -2\Sigma(x_i y_i - b_2 - \bar{y}_i)(x_i)2 - 2\Sigma(x_i y_i - a_1 - \bar{y}_i)(x_i)2 \quad ③$$

① $0 = \Sigma x_i \hat{\varepsilon}_i - n\hat{a}_1 - \hat{b}\Sigma x_i$

$$\hat{a}_1 = \bar{y}_1 - \hat{\beta}\bar{x}_1 \qquad \Rightarrow \hat{a}_2 = \bar{y}_2 - \hat{\beta}\bar{x}_2$$

③ $0 = \Sigma x_i^2 \hat{\beta} - \Sigma x_i y_i + \Sigma x_i \hat{a}_1 + \Sigma x_i^2 \hat{b}$

$$\Sigma x_i^2 \hat{\beta} - \Sigma x_i(x_i\hat{\beta} - \bar{y}_1) - \Sigma x_i^2 \hat{\beta} + \Sigma x_i(\bar{x}_1\hat{\beta} - \bar{y}_1)$$

$$0 = \Sigma x_i^2 \hat{\beta} - \Sigma x_i(x_i\hat{\beta} - \bar{y}_2) - \Sigma x_i^2 \hat{\beta} + \Sigma x_i(\bar{x}_2\hat{\beta} - \bar{y}_2)$$

$$0 = \hat{\beta}(\Sigma x_i^2 - \Sigma x_i^2 - \Sigma x_i \bar{x}_2 + \Sigma x_i \bar{x}_1) - \Sigma x_i \bar{y}_2 + \Sigma x_i \bar{y}_1$$

$$0 = \hat{\beta}(S_{xx} - S_{xx}) - S_{xy}$$

$$\hat{\beta}(S_{xx}+S_{xx}) = S_{xy}+S_{xy}$$

$$\hat{\beta} = \frac{S_{xy}+S_{xy}}{S_{xx}+S_{xx}}$$

18/05/16 ALSM I EXAM NOTES: PROBLEM SHEET

$$Var[\hat{\beta}] = Var\left[\frac{w_1\hat{\beta}_1 + w_2\hat{\beta}_1}{w_1 + w_2}\right]$$

$$= \frac{1}{(w_1+w_2)^2} Var[w_1\hat{\beta}_1 + w_2\hat{\beta}_1]$$

$$= \frac{1}{(w_1+w_2)^2}[w_1^2 Var[\hat{\beta}_1] + w_2^2 Var[\hat{\beta}_1] + 2w_1w_2 Cov[\hat{\beta}_1, \hat{\beta}_1]]$$

$Var[\hat{\beta}_1] = \sigma^2/S_{xx}$ $Var[\hat{\beta}_1] = \sigma^2/S_{xx}$

$Cov[\hat{\beta}_1, \hat{\beta}_1] = Cov\left[\sum_0 c_i y_i + \sum_0 d_i y_i\right]$ all Cov will be $0 \Rightarrow$ Diff var

$$Var[\hat{\beta}] = \frac{1}{w_1+w_2^2}\left[\frac{[w_1]\sigma^2}{S_{xx}^1} + \frac{[w_2]^2\sigma^2}{S_{xx}^2}\right]$$

18/5/16 ALSM1 EXAM NOTES: MATRIX FORMULATION

$Var[a^T x] = a^T \Sigma a$

$y = X\beta + \varepsilon$

Response vector, Design matrix, Coefficient, Error vector

$E[y] = X\beta$ $\quad E[\varepsilon] = 0$ (zero vector)

$Var[y] = Var[\varepsilon] = \begin{bmatrix} \sigma^2 & & 0 \\ & \ddots & \\ 0 & & \sigma^2 \end{bmatrix} = \sigma^2 I$ ← identity matrix; independent $\neq 0$ in off diagonals

Multiple Regression

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & & x_{p2} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$n \times 1$ $\quad n \times (p+1)$ $\quad (p+1) \times 1$ $\quad n \times 1$

Least Squares

Minimise SSE $\hat{\varepsilon}^T \hat{\varepsilon}$

$= (y - X\hat{\beta})^T (y - X\hat{\beta}) = y^T y - y^T X\beta - y X^T \beta^T + X^T \beta^T X\beta$

$\frac{d\text{SSE}}{d\beta} = 0 - X^T y - X^T y + 2X^T X\hat{\beta} = 0$

$X^T X\hat{\beta} = X^T y$

Assuming $X^T X$ is non singular (i.e. invertible) then $(X^T X)^{-1} X^T X\beta = (X^T X)^{-1} X^T y$

$\hat{\beta} = (X^T X)^{-1} X^T y$

Predicted mean of y is $\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy$ $\quad H = X(X^T X)^{-1} X^T$

SSE

$SSE = \hat{\varepsilon}^T \hat{\varepsilon} = (y - X\hat{\beta})^T (y - X\beta)$

$= y^T y - y^T X\hat{\beta} - \beta^T y X^T + \hat{\beta} X^T X\beta$

$= y^T y - \beta^T X^T y - \beta^T y X^T + \beta^T X^T X\beta$

$= y^T y - 2\beta^T X^T y + \beta^T X^T X[(X^T X)^{-1} X^T y]$

$= y^T y - 2\beta^T X^T y + \beta^T X^T y$

$= y^T y - \beta^T X^T y$ $\quad y^T y = \sum y_i^2$

$= y^T y - (X^T X)^{-1} X^T y X^T y = y^T [I - H] y$ $\quad H = X(X^T X)^{-1} X^T$

18/05/16

$\hat{\beta} = (X^TX)^{-1}X^Ty$

$E[\hat{\beta}] = E[(X^TX)^{-1}X^Ty] = (X^TX)^{-1}X^T E[y]$

$= (X^TX)^{-1}X^T [X\beta]$

$= \beta$ = unbiased

$Var[\hat{\beta}] = Var[(X^TX)^{-1}X^Ty]$

$= A Var[y] A^T$

$= A\sigma^2 I$

$= \sigma^2 [(X^TX)^{-1}X^T][(X^TX^{-1})X]$

$= \sigma^2 [(X^TX)^{-1}]$

ANOVA

Source	DF	SS	MS
Regression	p	$\hat{\beta}^T X^T y - n\bar{y}^2$	$SS(Reg)/p$
Residual	$n-p-1$	$(y-X\beta)^T(y-X\beta)$	$SSE/(n-p-1)$
Total (corrected)	$n-1$	$y^Ty - n\bar{y}^2$	

$H_0: \beta_1 = 0$ v $H_1: \beta_1 \neq 0$ need estimate for S.E for $[\hat{\beta}_1 - \beta_1]$

$Var[\hat{\beta}_1 - \beta_1] = Var[\beta_1] + Var[\beta_1] - 2Cov[\beta_0, \beta_1]$ used in one-way classification models

Confidence Intervals

$\hat{\beta}$ is a linear estimator hence $\hat{\beta} \sim N_{p+1}(\beta, \sigma^2(X^TX)^{-1})$

$\hat{\beta}_j \sim N(\beta_j, c_{jj}\sigma^2)$

Where c_{jj} is the $j+1^{st}$ diagonal entry of $(X^TX)^{-1}$ for $j = 0, 1, \ldots, p$

$\Rightarrow \dfrac{\hat{\beta}_j - \beta_j}{\sqrt{(MSE)(c_{jj})}} \sim t_{n-p-1}$ $\quad \pm \hat{\beta}_j\, t_{n-p-1, \alpha/2} \sqrt{(MSE)(c_{jj})}$

Confidence

Define $X_0 = \begin{bmatrix} 1 \\ x_{01} \\ \vdots \\ x_{0p} \end{bmatrix}$ $(p+1) \times 1$ column vector

The mean value of y at this point X_0 is $X_0^T\beta$ which is estimated by $X_0^T\hat{\beta} = \hat{y}_0$

18/05/16 ALSM I EXAM NOTES: MATRIX FORMULATION

$E[\hat{y}] = E[x_0^T \hat{\beta}] = x_0^T \beta = \mu_0$

$Var[\hat{y}] = x_0^T Var(\hat{\beta}) x_0$

$= \sigma^2 x_0^T (X^T X)^{-1} x_0$

$\hat{y} \sim N(\mu_0, \sigma^2 x_0^T (X^T X)^{-1} x_0)$ Replace σ^2 by MSE

CI for μ_0: $\frac{\hat{y} - \mu_0}{\sqrt{\sigma^2 x_0^T (X^T X)^{-1} x_0}} \sim N(0,1)$

$\frac{\hat{y}_0 - \mu_0}{\sqrt{MSE(x_0^T (X^T X)^{-1} x_0)}} \sim t_{n-p-1}$ Distribution

$100(1-\alpha)\%$ CI: $\hat{y}_0 \pm t_{n-p-1, \alpha/2} \sqrt{MSE\, x_0^T (X^T X)^{-1} x_0}$

Prediction

PI for a new observation y_0 is the value of a future observation x_0, estimated by $\hat{y}_0 = x_0^T \hat{\beta}$ and the PI is:

$\hat{y}_0 \pm t_{n-p-1, \alpha/2} \left(\sqrt{MSE(1 + x_0^T (X^T X)^{-1} x_0)}\right) = 1-\alpha$

Hypothesis Testing

Overall test of significance given by:

$H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0$ vs $H_1: \beta_j \neq 0$ for at least one j.

Use F-test = $MS(Reg)/MSE$

If $f < F_{p, n-p-1, \alpha/2}$ accept H_0

Individual test

$H_0: \beta_j = 0$ vs $H_1: \beta_j \neq 0$ t-test $\frac{\hat{\beta}_j}{\sqrt{MSE\, c_{jj}}}$ $n-p-1$ d.f.

$(j+1)^{st}$ entry in $(X^T X)^{-1}$

Partial F-Test Extra Sum of Squares

- Create full model (all terms) and reduced model (remove β_j)
- $R(\beta_j | \beta_0, \beta_1, \ldots, \beta_{j-1}, \beta_{j+1}, \ldots, \beta_p)$ = Extra Sum of Squares
- ESS is the partial SS for x_j and represents the contribution of x_j, adjusted for all other independent variables in the model

$F = \frac{R(\beta_s | \beta_0 \ldots) / 1}{MSE}$ D.F. $1, n-p-1$

is the test statistic for $H_0: \beta_s = 0$ vs $H_1: \beta_s \neq 0$
Called the partial F-test for β_s

Tests for a Subset of Regression Coefficients
Reduced model: $E[y_i] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_q x_{qi}$ $q < p$
Full Model: $E[y_i] = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$

$H_0: \beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$
H_A: At least one of these β's is $\neq 0$

$R(\beta_{q+1}, \ldots, \beta_p | \beta_0, \ldots, \beta_q) = ESS$ (How much more info you explain by including them in model)
Use F-Test: $\frac{R(\beta_{q+1}, \ldots, \beta_p | \beta_0, \ldots, \beta_q) / (p-q)}{MSE}$
If H_0 is true, then F follows an $F_{p-q, n-p-1}$ Distribution

Source	DF
$R(\beta_1, \ldots, \beta_q \mid \beta_0)$	q
$R(\beta_{q+1}, \ldots, \beta_p \mid \beta_0, \beta_1, \ldots, \beta_q)$	$p-q$
$R(\beta_1, \ldots, \beta_p \mid \beta_0)$	p
Residual	$n-p-1$
Total	$n-1$

or

Source	DF
$R(\beta_1 \mid \beta_0)$	1
$R(\beta_2 \mid \beta_0, \beta_1)$	1
$R(\beta_p \mid \ldots)$	1
$R(\beta_1, \ldots, \beta_p \mid \beta_0)$	p
Residual	$n-p-1$
Total	$n-1$

Sequential SS - Add one predictor at a time and find reduction in SSR not contained in $x_1, \ldots, x_{s-1}$
Partial SS - Find info contained in x_s which is not contained in x_k for $k \neq s$.
$R(\beta_1 | \beta_0) + R(\beta_2 | \beta_0, \beta_1) + R(\beta_3 | \beta_0, \beta_1, \beta_2) = R(\beta_1, \beta_2, \beta_3 | \beta_0) = SS(Reg)$

General Linear Hypothesis
$H_0: L\beta = c$ $H_A: L\beta \neq c$
L: $R \times (p+1)$ matrix of coefficients c $R \times 1$ vector of constants
β: $(p+1) \times 1$ vector of parameters

18/05/16 ALSM I EXAM NOTES: MATRIX FORMULATION

eg $H_0: \beta_2 = 0$ v $H_a: \beta_2 \neq 0$

$$\begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \end{bmatrix}$$

$L \quad \beta = L$

$H_0: \beta_1 = \beta_2 = 0$

$$\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$L \quad \beta \quad L$

- The extra sum of squares method can be used to test $H_0: L\beta = L$

$F = \frac{\text{Extra SS}/M}{MSE}$ DF: $M, n-p-1$

PROBLEM SET 4

18/05/16 ALSM 1 : EXAM NOTES PROBLEM SHEETS

PS 1

$X^T X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \Sigma x_i \\ \Sigma x_i & \Sigma x_i^2 \end{bmatrix}$

$|X^T X|$ - ad-bc $= n\Sigma x_i^2 - (\Sigma x_i)^2 = n S_{xx}$

$(X^T X)^{-1} = \frac{1}{|X^T X|} X^T X = \frac{1}{n S_{xx}} \begin{bmatrix} \Sigma x_i^2 & -\Sigma x_i \\ -\Sigma x_i & n \end{bmatrix}$

Q3 $\mathbb{E}[y_t] = x_t^T \beta_1 \quad t = 1, \ldots, \tau$

$\mathbb{E}[y_t] = x_t^T \beta_2 \quad t = \tau+1, \ldots, n$

$Q(\beta_1, \beta_2)$ $y_1 = \begin{bmatrix} y_1 \\ y_\tau \end{bmatrix}$ $X_1 = \begin{bmatrix} 1 & x_1 \\ 1 & x_\tau \end{bmatrix}$ Similarly y_2 and X_2

$Q: (y_1 - X_1\beta_1)^T (y_2 - X_1\beta) + (y_2 - X_2\beta_2)^T (y_t - X_1\beta_1)$

LS estimators: $\hat{\beta}_1 = (X_1^T X_1)^{-1} X_1^T y_1$

$\hat{\beta}_2 = (X_2^T X_2)^{-1} X_2^T y_2$

$\hat{\sigma}^2$ = pooled MSE = $\frac{Q(\hat{\beta}_1, \hat{\beta}_2)}{n-2}$ (SSE)

B If variance can be different, would have to use a weighted LSE

$\hat{\beta}_1, \hat{\beta}_2$ are as before

$\hat{\sigma}_1^2 = \frac{Q(\hat{\beta}_1)}{\tau - 2}$ Use MSE for first segment $\hat{\sigma}_2^2 = \frac{Q(\hat{\beta}_2)}{n - \tau - 2}$

4. $H^T H = X(X^T X)^{-1} X^T \; X^T (X^T X)^{-1} X \qquad H = X(X^T X)^{-1} X^T \qquad H^T = X^T (X^T X)^{-1} X$

$= X(X^T X)^{-1} X^T = H$

5. SSE $= \sum \hat{\varepsilon}_i^2 = (y - \hat{y})^T (y - \hat{y}) = (y - Hy)^T (y - Hy)$

$= y^T y - y^T y H - y H^T y^T + H^T H y^T y \qquad H^T H = H$

$= y^T y - 2 y^T y H + y^T y H$

$= y^T y - y^T y H$

$= y^T (I - H) y$

B4

$y_{ij} = \mu_i + \varepsilon_{ij}$ $\quad i = 1, \dots, I$ $\quad$ I different populations

$j = 1, \dots, J$ $\quad$ equal sample size - called balanced design

LS $Q(\hat{\mu}_1, \dots, \hat{\mu}_I) = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \hat{\mu}_i)^2$

$= \sum_{i} \sum_{j} (y_{ij}^2 - 2\hat{\mu}_i y_{ij} + \hat{\mu}_i^2)$

$= \sum\sum y_{ij}^2 - 2 \sum \hat{\mu}_i \sum y_{ij} + \sum J \hat{\mu}_i^2$

$\frac{dQ}{d\hat{\mu}_i} = -2 \sum_{j} y_{ij} + 2J\hat{\mu}_i = 0$

$\hat{\mu}_i = \sum_j y_{ij} / J = \bar{y}_{i\cdot}$ Sample mean for i^{th} population

'dot notation, summing over J

B Estimate I parameters $SSE = \sum_i \sum_j (y_{ij} - \bar{y}_{i\cdot})^2$ has DF:

$DF = n - I = IJ - I = I(J-1)$

$\hat{\sigma}^2 = MSE = SSE / I(J-1)$

C Design Matrix $y = X\mu + \varepsilon$

$\begin{bmatrix} 1 & 0 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \mu_1 \\ \mu_2 \\ \vdots \\ \mu_I \end{bmatrix} + \begin{bmatrix} \varepsilon_{11} \\ \vdots \\ \varepsilon_{IJ} \end{bmatrix}$ $\quad IJ \times I$ matrix

D $H_0: L\mu = 0$ vs $H_1: L\mu \neq 0$

$H_0: \mu_1 - \mu_2 = \mu_2 - \mu_3 = 0$ vs H_1 Not H_0

Implies that $\mu_i - \mu_3 = 0$ for $i \neq 3$

Write L as

$\begin{bmatrix} 1 & -1 & 0 & \cdots & 0 \\ 1 & 0 & -1 & \cdots & \\ \vdots & & & & \\ 1 & \cdots & & & -1 \\ 0 & 1 & \cdots & & \\ & & & & \\ & & & 1 & -1 \end{bmatrix}$ $\quad L$ is $\left(\frac{I}{2}\right) \times I$ matrix

Use F-test $\frac{MS(Reg)}{MSE} \sim F_{I,\, I(J-1)}$ compare to F critical value, if small accept H_0

PROBLEM SET 4

18/05/16 ALSM1 : EXAM NOTES EXAM QUESTIONS

Q3 A LS : $Q = \sum_i \sum_j (y_{ij} - \mu_i)^2$

$\partial Q / \partial \mu_i = \sum_{j=1}^{J} (-2)(y_{ij} - \mu_i) = 0$

$\sum_j y_{ij} - J\hat{\mu}_i = 0$

$\hat{\mu}_i = \sum y_{ij} / J$

$\hat{\mu}_i = \bar{y}_{i\cdot}$ (summing over J)

ANOVA for model?

$SS(Corr) = \sum_i \sum_j y_{ij} / IJ = \bar{y}_{\cdot\cdot}$

$SS(Tot) = \sum_{i=1}^{I} \sum_{j=1}^{J} (y_{ij} - \bar{y}_{\cdot\cdot})$

Within a population LS estimate of μ_i is $\bar{y}_{i\cdot}$

$\sum \sum (y_{ij} - \bar{y}_{\cdot\cdot})^2 = \sum \sum (y_{ij} - \bar{y}_{i\cdot} + \bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2$

$= \sum \sum (y_{ij} - \bar{y}_{i\cdot})^2 + (\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})^2 + 2(y_{ij} - \bar{y}_{i\cdot})(\bar{y}_{i\cdot} - \bar{y}_{\cdot\cdot})$

$= 0$

F test $F = MSR(Reg)/MSE = \dfrac{R(\beta_1 | \beta_0)/I(J-1)}{MSE}$

If H_0 is true $\mu_1 = \mu_2 = 0$ then $F \sim$ an F-dist with I and $I(J-1)$ d.f.

If $F > F_{2,2}$ then reject H_0 at $100(1-\alpha)\%$ significance

$Var[\beta] = Var\left[\dfrac{w_1\hat{\beta}_1 + w_2\hat{\beta}_2}{w_1 + w_2}\right] = \dfrac{1}{(w_1+w_2)^2} Var[w_1\hat{\beta}_1 + w_2\hat{\beta}_2]$

$= \dfrac{1}{(w_1+w_2)^2} [w_1^2 Var\hat{\beta}_1 + w_2^2 Var\hat{\beta}_2 + 2w_1w_2 Cov[\hat{\beta}_1, \hat{\beta}_2]]$ (Cov term $= 0$)

$Var[\mu_1] = \sigma^2/n_1 \quad Var[\mu_2] = \sigma^2/n_2$

$Var[\mu] = \dfrac{1}{(w_1+w_2)^2}\left[\dfrac{(w_1)^2\sigma^2}{n_1} + \dfrac{(w_2)^2\sigma^2}{n_2}\right]$

Q4 $\underline{y} = X\underline{\beta} + \underline{\varepsilon}$

$\underline{y}$ $n \times 1$ vector of response

X $n \times (p+1)$ matrix - design matrix

$\underline{\beta}$ $(p+1) \times 1$ vector of estimated parameters

$\hat{\underline{\beta}} = (X^TX)^{-1}X^T\underline{y}$ estimate of parameter

$E[y] = X\beta$ or $y = X\beta + \varepsilon$

where ε is $n \times 1$ error vector with 0 mean and uncorrelated errors

B The Hat matrix is such that $\hat{y} = Hy$ Here $\hat{y} = X\hat{\beta}$

$\hat{y} = (X^T X)^{-1} X^T y$

$H = X(X^T X)^{-1} X^T$

C Vector of residuals

$\hat{\varepsilon} = y - \hat{y} = y - Hy$

$= y - X\hat{\beta}$

D SSE $= \hat{\varepsilon}^T \hat{\varepsilon} = \hat{\varepsilon}_1 \hat{\varepsilon}$

$= (y - X\beta)^T (y - X\beta)$

$= y^T y - \beta^T X^T y - y^T X \beta + \beta^T X^T X \beta$

$= y^T y - y^T H y - y^T H y + y^T H^T H y$ $\quad H^T H = H$

$= y^T y - y^T H y$

$= y^T (I - H) y$

$H = X(X^T X)^{-1} X^T \quad H^T = X[(X^T X)^{-1}]^T X^T$ (symmetric)

$H^T H = X[(X^T X)^{-1}]^T X^T X[(X^T X)^{-1}] X^T$

$= X(X^T X)^{-1} X^T = H$ ✓ $H^T H = H$

Coláiste na Trionóide, Baile Átha Cliath
Trinity College Dublin
Oliscoil Átha Cliath | The University of Dublin

Faculty of Engineering, Mathematics and Science

School of Computer Science & Statistics
Statistics

Trinity Term 2016

Sophister Mathematics

ST3451 : Applied linear statistical methods 1

19/05/2016 Sports Centre 9:30-11:30

Dr. Jason Wyse

Instructions to Candidates:
Full marks will be awarded for complete solutions to any **three** questions. Each full question is worth **20 marks**, and marks for component parts are indicated in brackets. Non-programmable calculators are permitted. ~~Some useful distributions and~~ their properties are provided at the end of the exam.

Materials permitted for this examination:
Non-programmable calculators are permitted for this examination — please indicate the make and model of your calculator on each answer book used.

1. The simple linear regression model is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \qquad i = 1, 2, \ldots, n$$

where ε_i are the *error terms*.

(a) What assumptions are made about the error terms? [2 marks]

(b) Derive the least squares estimators $\widehat{\beta}_0$ and $\widehat{\beta}_1$ of β_0 and β_1. [6 marks]

(c) Show that

$$\widehat{\beta}_1 = \sum_{i=1}^{n} c_i Y_i \qquad \widehat{\beta}_0 = \sum_{i=1}^{n} d_i Y_i$$

giving the explicit expressions for the c_i, d_i. [6 marks]

(d) Using the results of part (c) find $\mathrm{Var}\{\widehat{\beta}_1\}$. [6 marks]

2. A one-way classification model assumes that J observations are taken from each of I normal populations, that is

$$Y_{ij} = \mu_i + \varepsilon_{ij} \qquad (i = 1, 2, \ldots, I; j = 1, 2, \ldots, J)$$

where the ε_{ij} are iid $N(0, \sigma^2)$.

(a) Show that the least squares estimates of $\mu_1, \ldots, \mu_I$ are given by $\widehat{\mu}_i = \bar{Y}_{i\cdot} = \sum_{j=1}^{J} Y_{ij}/J$ for $i = 1, \ldots, I$. [5 marks]

(b) Explain why

$$\mathrm{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{i\cdot})^2.$$

What are its associated degrees of freedom? Explain. [3 marks]

(c) Show that the SSE partitions into two sums of squares as

$$\mathrm{SSE} = \sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij} - \bar{Y}_{\cdot\cdot})^2 + J\sum_{i=1}^{I}(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot})^2$$

where $\bar{Y}_{\cdot\cdot} = \sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}/(IJ)$ is the overall mean. [7 marks]

(d) Describe how one would carry out the test of hypothesis $H_0 : \mu_1 = \mu_2 = \cdots = \mu_I = \mu$ versus $H_A : H_0$ not true. [5 marks]

3. Consider a model where observations are time ordered

$$(x_1, Y_1), (x_2, Y_2), \ldots, (x_t, Y_t), \ldots, (x_n, Y_n),$$

where t denotes time. There is a changepoint at time $t = \tau$, so that up to time τ it is reasonable to assume that

$$\mathrm{E}\{Y_t\} = \alpha_1 + \beta_1 x_t, \qquad t = 1, \ldots, \tau$$

while after time τ

$$\mathrm{E}\{Y_t\} = \alpha_2 + \beta_2 x_t, \qquad t = \tau + 1, \ldots, n$$

such that $\alpha_1 \neq \alpha_2$ and $\beta_1 \neq \beta_2$.

(a) Make a sketch indicating the main features of the model. [2 marks]

(b) Write this model in the form $\mathrm{E}\{\mathbf{Y}\} = \mathbf{X}\theta$ where $\theta = (\alpha_1, \beta_1, \alpha_2, \beta_2)^\mathsf{T}$ is a column vector and $\mathbf{X}$ is a design matrix. [6 marks]

(c) Assuming a constant error variance σ^2 before and after the changepoint τ, suggest an estimator of σ^2 giving reasons for your answer. [6 marks]

(d) Suggest an approach for determining an estimate of τ if it is not known and must be estimated. [6 marks]

4. Consider the matrix formulation of the simple linear regression model $\mathrm{E}\{\mathbf{Y}\} = \mathbf{X}\beta$, where $\mathbf{Y} = (Y_1, \ldots, Y_n)^\mathsf{T}$ is an $n \times 1$ column vector of responses, $\beta = (\beta_0, \beta_1)^\mathsf{T}$, and $\mathbf{X}$ is the $n \times 2$ design matrix with row i equal to $(1, x_i)$.

(a) Find $\mathbf{X}^\mathsf{T}\mathbf{X}$, $|\mathbf{X}^\mathsf{T}\mathbf{X}|$ and $(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}$ in terms of the x_i. [5 marks]

(b) Compute the least squares estimator $\widehat{\beta}$ of β in terms of the x_i, Y_i. You may assume $\widehat{\beta} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\mathbf{Y}$. [5 marks]

(c) Write down an expression for the vector of model residuals $\widehat{\varepsilon}$ in terms of $\mathbf{Y}, \mathbf{X}$ and $\widehat{\beta}$. [5 marks]

(d) Show that sum of squared errors (SSE) can be written as

$$\mathrm{SSE} = \mathbf{Y}^\mathsf{T}\left[\mathbf{I} - \mathbf{X}(\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\right]\mathbf{Y}$$

[5 marks]

5. A practitioner (who has little experience with statistics) has asked for a consult to explain some of the terminology associated with regression analysis and regression model diagnostics. They send a list of questions by email before the meeting in order to help you (as the consultant) prepare. The questions follow.

1. What is a studentized residual? I think I understand what a residual is, but I'm not sure if this is the same as a studentized residual.

2. I obtained a 95% confidence interval for b1 (slope). Does this mean that 95% of my response variable observations will be the between these two numbers?

3. What is heteroscedasticity and should I be worried about it?

4. I know the normal curve is the bell curve, but I can't see how this is related to my analysis. How would it be related to my analysis?

Write notes addressing the questions in the email, explaining how you would answer each of them in a non-technical manner (remember, the practitioner is not statistically literate).

[20 marks]