16/11/15 ALSM1

1

CHAPTER 2
MULTIPLE REGRESSION

2.1 REVIEW OF VECTOR AND MATRIX

* Matrix Cookbook (Petersen and Pedersen) Book

Differentiation: Let $\underline{x}$ be a column vector in $\mathbb{R}^d$

$$\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix} = [x_1\ x_2\ \dots\ x_d]^T$$

← Differentiate each individually

Then for a function $F: \mathbb{R}^d \to \mathbb{R}$, $\frac{d\,F(\underline{x})}{d\underline{x}} = \begin{bmatrix} dF/d(x_1) \\ \vdots \\ dF/d(x_d) \end{bmatrix} = \nabla_{\underline{x}} f(\underline{x})$ Gradient of $\underline{x}$

① IF $F(\underline{x}) = c$, a constant, then $dF/d\underline{x} = 0$ is a $d \times 1$ vector of 0's

linear combination of x's

② IF $F(\underline{x}) = \underline{a}^T\underline{x} = a_1x_1 + a_2x_2 + \dots + a_dx_d$

$$dF/d\underline{x} = \begin{bmatrix} a_1 \\ \vdots \\ a_d \end{bmatrix} = \underline{a}$$

③ Similarly if $F(\underline{x}) = \underline{x}^T\underline{a} = a_1x_1 + a_2x_2 + \dots + a_dx_d = \underline{a}^T\underline{x}$

$$\frac{dF}{d\underline{x}} = \underline{a}$$

a $d \times d$ matrix

④ Consider $F(\underline{x}) = \underline{x}^T A \underline{x}$ $\quad \frac{dF}{d\underline{x}} = ?$

Look at $A\underline{x}$: $d \times 1$ vector with j^{th} entry $[A\underline{x}]_j = \sum_{k=1}^{d} A_{jk} x_k$

j^{th} row of A by corresponding row of x

Then $\underline{x}^T A \underline{x} = [x_1, \dots x_d] A \underline{x}$

$$= \sum_{j=1}^{d} x_j [A\underline{x}]_j$$

Sum over x_j

$$= \sum_{j=1}^{d} x_j \sum_{k=1}^{d} A_{jk} x_k$$

$$= \sum_{j=1}^{d} \sum_{k=1}^{d} x_j A_{jk} x_k$$

$$= F(\underline{x})$$

$$\frac{df}{dx_L} = \frac{d}{dx}\left[\sum_{J=1}^{d}\sum_{K=1}^{d} x_J A_{JK} x_K\right] \qquad \underline{x}^T A \underline{x} = f(\underline{x})$$

$$= \frac{d}{dx_L}\left[\sum_{J=1}^{d}\left(A_{JJ} x_J^2 + x_J \sum A_{JK} x_K\right)\right]$$

$$= \frac{d}{dx_L}\left[A_{LL} x_L^2 + x_L \sum A_{LK} x_K\right] + \frac{d}{dx_L}\left[\sum_{J\neq L} x_J A_{JL} x_L\right]$$

$$= \left[2 A_{LL} x_L^2 + \sum_{K\neq L} A_{LK} x_K + \sum_{J\neq L} A_{JL} x_J\right]$$

$$= 2 A_{LL} x_L + \sum_{K\neq L}\left[A_{LK} + A_{KL}\right] x_K$$

Mostly we are concerned with a Symetric $\Rightarrow A_{LK} = A_{KL}$ $\quad (A = A^T)$

$\hookrightarrow \; 2 A_L x_L + \sum_{K\neq L} (2 A_{LK}) x_K$

$= 2 \sum_{K=1}^{d} A_{LK} x_K \quad = df/dx_L$ w.r.t L^{th} x

$$\frac{df}{d\underline{x}} = 2A\underline{x}$$

Moments of Random Variables

$\underline{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$ is a d-dimensional random vector (or d-dimensional random variable)

$$\mathbb{E}[\underline{x}] = \begin{bmatrix} \mathbb{E}[x_1] \\ \vdots \\ \mathbb{E}[x_d] \end{bmatrix}$$

If a and b are constants and $\underline{x}$ $\underline{y}$ are random vectors (of some dimension) then
$\mathbb{E}[a\underline{x} + b\underline{y}] = a\,\mathbb{E}[\underline{x}] + b\,\mathbb{E}[\underline{y}]$

Covariance matrix of $\underline{x}$

$Var[\underline{x}]$
$Cov[\underline{x}]$ $\begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1d} \\ \vdots & & & \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_{dd} \end{bmatrix}$ How they vary with each other. Symetric matrix

- Where $\sigma_{ii} = Var[x_i]$ $\quad \sigma_{ij} = Cov[x_i, x_j]$ when $i \neq j$
- The Covariance matrix is Symetric $\Rightarrow \Sigma = \Sigma^T$

ALSM1

If the off diagonal entries are all zero, then the elements of x are uncorrelated
$$\begin{bmatrix} \sigma_1^2 & & 0 \\ & \sigma_2^2 & \\ 0 & & \sigma_n^2 \end{bmatrix}$$

$\text{Var}[\underline{a}^T\underline{x}] = \underline{a}^T \Sigma \underline{a}$

$\text{Cov}[\underline{a}^T\underline{x}, \underline{b}^T\underline{x}] = \underline{a}^T \Sigma \underline{b}$

2.2 MATRIX FORMULATION OF THE SLR MODEL

- The SLR model is written $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i \quad i = 1, \dots, n$
- We can write this in matrix terms using random vectors.

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

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

$\underline{y}$: Response Variable
X : Design Matrix
$\underline{\beta}$: Parameter / Coefficient Vector
$\underline{\varepsilon}$: Error Vector

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

$\mathbb{E}[\underline{\varepsilon}] = \underline{0}$ zero vector → $n \times 1$ vector of zeros

$\text{Var}[\underline{y}] = \text{Var}[\underline{\varepsilon}] = \begin{bmatrix} \sigma^2 & & 0 \\ & \sigma^2 & \\ 0 & & \sigma^2 \end{bmatrix} = \sigma^2 I$ independent ⇒ 0 in off diagonals

↖ Identity matrix $\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$

2.3 MULTIPLE REGRESSION

Consider extending the SLR model such that the dependent variable (response) has mean depending on a number of predictors / independent variables $x_1 \dots x_m$

$$y_i = \underbrace{\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_p x_{ip}}_{\mathbb{E}[y \mid (x_{1i}, x_{i2}, \dots, x_{pi})]} + \varepsilon_i$$

The model can be written in matrix notation $y = X\beta + \varepsilon$

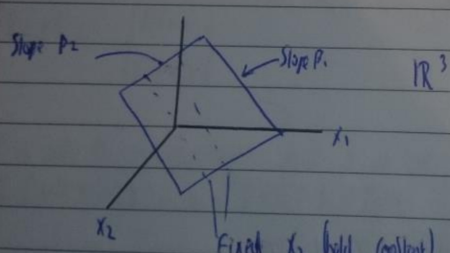
$$\underset{n \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} \quad \underset{n \times (p+1)}{\begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{p1} \\ 1 & x_{12} & x_{22} & & x_{p2} \\ \vdots & & & & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{pn} \end{bmatrix}} \quad \underset{(p+1) \times 1}{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}} + \underset{n \times 1}{\begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}}$$

β is a vector of unknown parameters to be estimated from the data

β_j is the change in the mean value of y per unit change in x_j assuming all other independent variables are held constant

$\mathbb{E}[y] = X\beta$ $\quad \mathbb{E}[\varepsilon] = 0 \quad Var[y] = \sigma^2 I = Var[\varepsilon]$ (independent errors)

When $p = 2$ the model is 2 dimensional plane in a 3d space

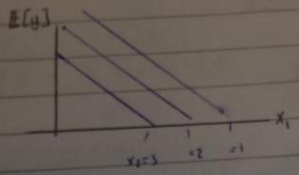


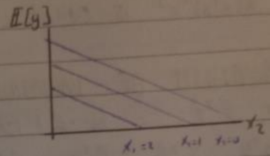
$\mathbb{R}^3$

23/11/15 ALSM I

Previous Diagram, would have $\beta = \begin{bmatrix}\beta_0\\ \beta_1\\ \beta_2\end{bmatrix}$

Could write model as $\underline{y} = X\beta + \underline{\varepsilon}$





When x_2 held constant, all lines have

Slope β_1

- The model $\underline{y} = X\beta + \underline{\varepsilon}$ is known as the general linear model - not to be confused with GLM's
- Includes: SLR, multiple regression, analysis of variance (one way classification), others.

2.4 LEAST SQUARES ESTIMATORS

The sum of squared errors can be written in matrix form

$$SSE = \sum \hat{\varepsilon}_i^2 = [\hat{\varepsilon}_1, \dots, \hat{\varepsilon}_n]\begin{bmatrix}\hat{\varepsilon}_1\\ \vdots \\ \hat{\varepsilon}_n\end{bmatrix} = \underline{\hat{\varepsilon}}^T\underline{\hat{\varepsilon}}$$

$(y^T - \beta x)$

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

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

We want to minimise SSE WRT $\hat{\beta}$: i.e. least squares estimates

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

$$X^T X\hat{\beta} = X^T\underline{y} \quad \text{(normal equations)}$$

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

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

2.5 LEAST SQUARES PLANE

Once we have the least squares estimates of β, then the predicted mean of y is $\hat{\underline{y}} = X\hat{\underline{\beta}} = X(X^TX)^{-1}X^Ty = H\underline{y}$

$H = X(X^TX)^{-1}X^T$ is called the Hat Matrix

2.6 ANALYSIS OF VARIATION IN $\underline{y}$

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

$= \underline{y}^T\underline{y} - \underbrace{\underline{y}^TX\hat{\underline{\beta}}}_{\text{scalar}} - \hat{\underline{\beta}}^T\underline{y} + \hat{\underline{\beta}}^TX^TX\hat{\underline{\beta}}$

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

$= \underline{y}^T\underline{y} - 2\hat{\underline{\beta}}^TX^T\underline{y} + \hat{\underline{\beta}}^T\underbrace{X^TX[(X^TX)^{-1}}_{\text{cancel each other}}X^T\underline{y}]$

$= \underline{y}^T\underline{y} - 2\hat{\underline{\beta}}^TX^T\underline{y} + \hat{\underline{\beta}}^TX^T\underline{y}$

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

$\underline{y}^T\underline{y} = [y_1 \ldots y_n]\begin{bmatrix} y_1 \\ \vdots \\ y_N \end{bmatrix} = y_1^2 + \ldots + y_N^2 = \sum y_i^2 = SS(\text{Uncorrected})$

$SSE = SS(\text{Uncorrected}) - SS(\text{Model})$

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

$n\bar{y}^2 = \text{correction}$

$SS(\text{Uncorrected}) - \text{Correction} = SS(\text{Reg}) + SSE$

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

23/11/15 ALSM 1

2.7 PROPERTIES OF THE ESTIMATORS

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

$\mathbb{E}[\hat{\beta}_1] = \beta_1 \qquad Var[\hat{\beta}_1] = \sigma^2/S_{xx}$

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

$$\begin{aligned}
\mathbb{E}[\hat{\beta}] &= \mathbb{E}[(X^TX)^{-1}X^T y] \\
&= (X^TX)^{-1}X^T\,\mathbb{E}[y] \quad \text{x's are fixed quantities} \Rightarrow \text{not random} \\
&= (X^TX)^{-1}X^T[X\beta] \\
&= \underbrace{(X^TX)^{-1}X^T(X^TX)}_{I}\beta \\
&= \beta \quad \text{LS is unbiased estimator}
\end{aligned}$$

For SLR model:

$$Var[\hat{\beta}] = \begin{bmatrix} Var[\hat{\beta}_0] & Cov[\hat{\beta}_0, \hat{\beta}_1] \\ Cov[\hat{\beta}_0, \hat{\beta}_1] & Var[\hat{\beta}_1] \end{bmatrix}$$

In general,

$$\begin{aligned}
Var[\hat{\beta}] &= Var[\overbrace{(X^TX)^{-1}X^T}^{A}y] \\
&= A\,Var[\underline{y}]\,A^T \\
&= A\,[\sigma^2]\,A^T \\
&= \sigma^2 AA^T \\
&= \sigma^2[(X^TX)^{-1}X^T][X(X^TX)^{-1}] \\
&= \sigma^2(X^TX^{-1})
\end{aligned}$$

In multiple regression, Gauss-Markov theorem holds: LS estimates are Best linear unbiased estimators (BLUE)

Example: Cereals Dataset

- Nutritional info and shelf location for 77 breakfast cereals
- A rating was calculated from consumer reports
- "Middle shelf" cereals tended to have lowest rating
- R Script: investigate relationship b/w rating, sugar content (per 100g) and fat content (per 100g)

$$\underset{\text{Rating}}{y_i} = \underset{\text{Intercept}}{\beta_0} + \underset{\text{Sugar /100g}}{\beta_1 x_{1i}} + \underset{\text{Fat /100g}}{\beta_2 x_{2i}} + \varepsilon_i$$

$$\beta = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} 61.084 \\ -2.213 \\ -3.066 \end{bmatrix}$$

- Mean rating decreasing by 2.2 for every extra gram of sugar per 100g, keeping the fat content fixed
- Similarly, the mean rating decreases by ~ 3 for every extra gram of fat, keeping sugar fixed

ANOVA Table $n = 77$

Source	DF	SS	MS
X_1	1	86547	86547
X_2	1	670.5	670.5
Residual	74	5671.5	76.6

Add together for prev table

MSE = 76.6

$$\text{Var}[\hat{\beta}] = \text{MSE}\,(X^T X)^{-1} = \begin{bmatrix} 3.813 & -0.315 & -0.632 \\ -0.315 & 0.055 & -0.066 \\ -0.632 & -0.066 & 1.074 \end{bmatrix}$$

$S.E.[\hat{\beta}_0] = 3.815$

$S.E.[\hat{\beta}_1] = 0.055$

$S.E.[\hat{\beta}_2] = 1.074$

ANOVA Table

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

SLR: $SSE = \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \rightarrow n-2$ df

Here we have p predictors (independent variables). We estimate $p+1$ parameters

So the DF associated with SSE is $n-(p+1)$

$y^T y - n\bar{y}^2 = \sum y_i^2 - n\left(\frac{\sum y_i}{n}\right)^2 = \sum y_i^2 - \frac{(\sum y_i)^2}{n} = \sum (y_i - \bar{y})^2 \quad df = n-1$

$R^2 = \frac{SS(Reg)}{SS(corrected)}$ $\quad 0 < R^2 < 1$ The square of the multiple correlation between y and the predictors $(x_1, \ldots, x_p)$

- Gives proportion of ~~variance~~ variation in y explained by the predictors

$\hat{\underline{\varepsilon}} = \underline{y} - \hat{\underline{y}}$ vector of residuals

$= \underline{y} - H\underline{y} \quad = I_{n \times n}\, y - Hy$

$= (I - H)\underline{y}$

$Cov[\hat{\beta}_0, \hat{\beta}_1] = -0.315$

$Cov[\hat{\beta}_1, \hat{\beta}_2] = -0.066$

$H_0: \beta_0 - \beta_1 = 0 \quad v \quad H_1: \beta_0 - \beta_1 \neq 0$ need estimate for S.E. for $[\hat{\beta}_0 - \hat{\beta}_1]$ ↵

$Var[\hat{\beta}_0 - \hat{\beta}_1] = Var[\hat{\beta}_0] + Var[\hat{\beta}_1] - 2Cov[\hat{\beta}_0, \hat{\beta}_1]$

Used in one-way classification models.

2.8 INFERENCE FOR MULTIPLE REGRESSION

In order to make inferences we again have to make an assumption about the distribution of $\underline{y}$. We usually assume normality

30/11/1

ASIDE: y has a multivariate normal distribution with parameters μ (mean vector) and Σ (Covariance matrix)

$$\underset{n\times 1}{y} \sim N_n(\underset{n\times 1}{\mu}, \underset{n\times m}{\Sigma})$$

Density function of y

$$f(y) = \frac{1}{(2\pi)^{n/2}} |\Sigma|^{-1/2} \exp\left[-\tfrac{1}{2}(y-\mu)^T \Sigma^{-1} (y-\mu)\right]$$

multivariate distribution

$\mu = X\beta$ $\quad f(y) = \frac{1}{(2\pi)^{n/2}} |\sigma^2 I|^{-1/2} \exp\left[-\tfrac{1}{2}(y-X\beta)^T(\sigma^2 I)^{-1}(y-X\beta)\right]$

$\Sigma = \sigma^2 I$

$$\frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} \exp\left[\frac{-1}{2\sigma^2}(y-X\beta)^T(y-X\beta)\right]$$

When $n=1$, $y_i \sim N(\mu, \sigma^2)$

$$f(y) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left[\frac{1}{2\sigma^2}(y-\mu)^2\right]$$

CONFIDENCE INTERVALS

Confidence Intervals on the regression coefficients: $\hat{\beta}$ is a linear estimator hence

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

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

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

Consequently $\dfrac{\hat{\beta}_j - \beta_j}{\sqrt{MSE\, C_{jj}}} \sim t_{n-p-1}$

The degrees of freedom of the t dist is that associated with the MSE using a similar argument to Section 1.2

$$P\left[\hat{\beta}_j - t_{n-p-1,\alpha/2}\sqrt{MSE\, C_{jj}} \leq \beta_j \leq \hat{\beta}_j + t_{n-p-1,\alpha/2}\sqrt{MSE\, C_{jj}}\right] = 1-\alpha$$

This is a $100(1-\alpha)\%$ confidence interval for β_j

/11/15 ALSM I

$\hat{\beta} \sim N_{p+1} \left(\beta , \sigma^2 (X^T X)^{-1} \right)$

c_{jj} $(j+1)$th diagonal element

Confidence Interval For the mean response

Define $x_0 = \begin{bmatrix} 1 \\ x_{10} \\ x_{20} \\ \vdots \\ x_{p0} \end{bmatrix}$ $(p+1) \times 1$ col vector

The mean value of y at this point x_0 is $x_0^T \beta$ which is estimated by $x_0^T \hat{\beta} = \hat{y}_0$

$E[\hat{y}_0] = E[x_0^T \hat{\beta}] = x_0^T \beta = \mu_0$

$Var[\hat{y}_0] = x_0^T Var[\hat{\beta}] x_0$

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

$\hat{y}_0 \sim N \left(\mu_0, \sigma^2 x_0^T (X^T X)^{-1} x_0 \right)$

Replace σ^2 by MSE

CI for μ_0 : $\dfrac{\hat{y}_0 - \mu_0}{\sqrt{\sigma^2 x_0^T (X^T X)^{-1} x_0}} \sim N(0,1)$

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

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

Prediction

Prediction Interval for a new observation y_0 is the value of a future observation of x_0, it is estimated by $y_0 - x_0^T \hat{\beta}$ and the prediction interval is

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

Hypothesis Testing

An overall test of significance of regression is given by:

$H_0: \beta_1 = \beta_2 = \; = \beta_p = 0$ vs $H_A: \beta_J \neq 0$ for atleast one J

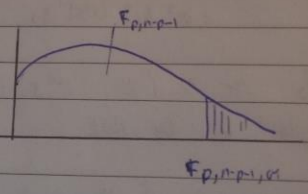
$E[y_i] = \beta_0 + \beta_1 x_{1i} + \; + \beta_p x_{pi}$ if all β_J, $J \geq 0$ ~~[illegible]~~ are zero, $E[y_i] = \beta_0$

Test Statistic

$$F = \frac{MS(Reg)}{MSE} = \frac{R(\beta_1, \; , \beta_p \mid \beta_0)/p}{MSE}$$

If H_0 is true, ~~[illegible]~~ (i.e. $\beta_1 = \; = \beta_p = 0$), then F follows an F-Dist with p and $n-p-1$ D.F.

If $F > F_{p,n-p-1,\alpha}$ then reject H_0 at $100\alpha\%$ significance level with α controlling probability of type I error.



Testing On Individual Regression Coefficient

$H_0: \beta_J = 0$ v $H_A: \beta_J \neq 0$ (can be directional)

$$t\text{-test} = \frac{\hat{\beta}_J}{\sqrt{MSE\, C_{JJ}}}$$ $n-p-1$ D.F.

C_{JJ}: $J+1^{st}$ diagonal element of $(X^TX)^{-1}$

This is the test of the contribution of X_J given all the other independent variables are the model. Compare test statistic with critical value and conclude

"Extra Sum of Squares"

Reduced model: $E[y_i] = \beta_0 + \beta_1 X_{1i} + \; + \beta_{J-1} X_{J-1i} + \beta_{J+1i} + \; + \beta_p X_{pi}$

Just removed the J^{th} term

Full Model: $E[y_i] = \beta_0 + \beta_1 X_{1i} + \; + \beta_p X_{pi}$

/11/15 ALSM1 3

$R(\beta_J | \beta_0, \ldots, \beta_{J-1}, \beta_{J+1}, \ldots, \beta_P)$ = Extra Sum of Squares

Here, the extra sum of squares is the partial SS for X_J and represents the contribution of X_J, adjusted for all other independent variables in the model

$$F = \frac{R(\beta_J | \beta_0, \beta_1, \ldots)/1}{MSE} \rightarrow DF\ 1, n-p-1$$

is the test statistic for $H_0: \beta_J = 0$ vs $H_A: \beta_J \neq 0$ (can't do directional testing).

This is called the partial F-test for β_J. Equivalent to the two tailed t-test [$F = t^2$]

How to do a joint test for regression coefficients to measure overall usefullness of the regression model

$H_0: \beta_1 = \beta_2 = \ldots = \beta_P = 0$ (Model Utility Test)
H_A: At least one of the $\beta_J \neq 0$

Test 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: $\beta_{q+1} = \beta_{q+2} = \ldots = \beta_p = 0$ (testing a subset)
H_A: At least one of these β's is $\neq 0$

$R(\beta_{q+1}, \ldots, \beta_p | \beta_0, \ldots, \beta_q)$ = Extra Sum of Squares (how much more info you explain by including them in model)

Use F test: $F = \frac{R(\beta_{q+1}, \ldots, \beta_p | \beta_0, \ldots, \beta_q)/(p-q)}{MSE}$ ← Difference

If H_0 is true, then F follows an $F_{p-q, n-p-1}$ distribution
Compute F and compare it to a critical value

2.9 SEQUENTIAL SUM OF SQUARES

The extra sum of squares partitions $SS(Reg)$ in the ANOVA Table

$$SS(Reg) = R(\beta_1, \beta_2, \ldots, \beta_p \mid \beta_0)$$
$$= R(\beta_1, \ldots, \beta_q \mid \beta_0) + R(\beta_{q+1}, \ldots, \beta_p \mid \beta_0, \ldots, \beta_q)$$

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

We can extend this partitioning as follows:

Source	DF
$R(\beta_1 \mid \beta_0)$	1
$R(\beta_2 \mid \beta_0, \beta_1)$	1
$\vdots$	$\vdots$
$R(\beta_p \mid \beta_0, \ldots, \beta_{p-1})$	1
$SS(Reg) = R(\beta_1, \ldots, \beta_p \mid \beta_0)$	p
Residual	$n-p-1$
Total	$n-1$

$R(\beta_J \mid \beta_0, \ldots, \beta_{J-1})$ is known as the sequential sum of squares for X_J → The amount by which you will reduce the residual sum of squares by adding X_J to the model given $X_1, \ldots, X_{J-1}$ → and represent the contribution of X_J in the model adjusted for $X_1, \ldots, X_{J-1}$ but NOT $X_{J+1}, \ldots, X_p$. It also depends on the order.

We can see $\sum_{J=1}^{p} R(\beta_J \mid \beta_0, \ldots, \beta_{J-1}) = R(\beta_1, \ldots, \beta_p \mid \beta_0)$

Example:

$$\mathbb{E}[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$$

30/11/15 ALSM 1

Var	Partial SS	Sequential SS
X_1	$R(\beta_1 \mid \beta_0, \beta_2, \beta_3)$	$R(\beta_1 \mid \beta_0)$
X_2	$R(\beta_2 \mid \beta_0, \beta_1, \beta_3)$	$R(\beta_2 \mid \beta_0, \beta_1)$
X_3	$R(\beta_3 \mid \beta_0, \beta_1, \beta_2)$ ← equal →	$R(\beta_3 \mid \beta_0, \beta_1, \beta_2)$

- Partial SS - Find the info contained in X_J which is not contained in X_K for $K \neq J$
- Sequential SS - Info in X_J not contained in $X_1, \dots, X_{J-1}$

$R(\beta_1|\beta_0) + R(\beta_2|\beta_0,\beta_1) + R(\beta_3|\beta_0,\beta_1,\beta_2) = R(\beta_1,\beta_2,\beta_3|\beta_0) = SS(Reg)$

In general, sum of Partial SS $\neq$ SS(Reg)

Example: Cereals

1. 95% CI: β_0: $\hat{\beta_0} \pm t_{94,0.025}$ S.E.$[\hat{\beta_0}]$ $= 61.084 \pm 1.993\sqrt{3.813}$

   β_1: $\hat{\beta_1} \pm t_{94,0.025}$ S.E.$[\hat{\beta_1}]$ $= -2.213 \pm 1.993\sqrt{0.055}$

   β_2: $\hat{\beta_2} \pm t_{94,0.025}$ S.E.$[\hat{\beta_2}]$ $= -3.066 \pm 1.993\sqrt{1.074}$

2. H_0: $\beta_1 = \beta_2 = 0$ v H_A: Either β_1 or $\beta_2 \neq 0$

   $F = \frac{9325.2/2}{76.6} = 60.869$ df = 2, 94

   5% critical value: 3.12 ⇒ test is highly significant, reject H_0

3. H_0: $\beta_1 = 0$ (given β_2 is in the model)

   H_A: $\beta_1 \neq 0$

   Partial SS ⇒ $R(\beta_1|\beta_0,\beta_2)$

   $t = \frac{\hat{\beta_1} - 0}{SE[\hat{\beta_1}]} = \frac{-2.213}{\sqrt{0.055}} = -9.4362$

   Critical value 5% ⇒ ± 1.993, highly significant, reject H_0

2.10 THE GENERAL LINEAR HYPOTHESIS

H_0: $L\beta = c$ H_A: $L\beta \neq c$

$\underline{L}$: A $k \times (p+1)$ matrix of coefficients
$\underline{\beta}$: $(p+1) \times 1$ vector of parameters
$\underline{c}$: $k \times 1$ vector of constants

All the hypotheses we discussed are special cases of this
Example: $E[y_i] = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i}$
$H_0: \beta_2 = 0$ v $H_1: \beta_2 \neq 0$

$$\underbrace{\begin{bmatrix} 0 & 0 & 1 & 0 \end{bmatrix}}_{\underline{L}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}}_{\underline{\beta}} = \underbrace{\begin{bmatrix} 0 \end{bmatrix}}_{\underline{c}}$$

$H_0: \beta_1 = \beta_2 = 0$ $$\underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix}}_{\underline{L}} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix}}_{\underline{\beta}} = \underbrace{\begin{bmatrix} 0 \\ 0 \end{bmatrix}}_{\underline{c}}$$

- The extra sum of squares method can be used to test $H_0: \underline{L}\underline{\beta} = \underline{c}$
- Reduced model (H_0) is constrained so that H_0 is true. Full model is not constrained

$F = \dfrac{\text{extra SS}/k}{MSE}$ with $k, n-p-1$ df

02/12/15 ALSM I

2.11 ORTHOGONALITY

Consider for example : $y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \varepsilon_i$

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

$X \Rightarrow$ Design matrix $n \times 3$

$$X = \begin{bmatrix} 1 & x_{11} & x_{21} \\ 1 & x_{12} & x_{22} \\ \vdots & & \\ 1 & x_{1n} & x_{2n} \end{bmatrix} = [\underline{x}_0 \vdots \underline{x}_1 \vdots \underline{x}_2]$$

Consider the situation where the columns of X are orthogonal

$\underline{x}_0^T \underline{x}_1 = 0$

$\underline{x}_1^T \underline{x}_2 = 0$ Orthogonality

$\underline{x}_0^T \underline{x}_2 = 0$

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

$$X^T = \begin{bmatrix} \underline{x}_0^T \\ \underline{x}_1^T \\ \underline{x}_2^T \end{bmatrix} \qquad X^T X = \begin{bmatrix} \underline{x}_0^T \\ \underline{x}_1^T \\ \underline{x}_2^T \end{bmatrix} [\underline{x}_0 \ \underline{x}_1 \ \underline{x}_2]$$

$$= \begin{bmatrix} \underline{x}_0^T \underline{x}_0 & \underline{x}_0^T \underline{x}_1 & \underline{x}_0^T \underline{x}_2 \\ \underline{x}_1^T \underline{x}_0 & \underline{x}_1^T \underline{x}_1 & \underline{x}_1^T \underline{x}_2 \\ \underline{x}_2^T \underline{x}_0 & \underline{x}_2^T \underline{x}_1 & \underline{x}_2^T \underline{x}_2 \end{bmatrix}$$

$$= \begin{bmatrix} \underline{x}_0^T \underline{x}_0 & 0 & 0 \\ 0 & \underline{x}_1^T \underline{x}_1 & 0 \\ 0 & 0 & \underline{x}_2^T \underline{x}_2 \end{bmatrix}$$ because of orthogonality

$$(X^T X)^{-1} = \begin{bmatrix} (\underline{x}_0^T \underline{x}_0)^{-1} & 0 & 0 \\ 0 & (\underline{x}_1^T \underline{x}_1)^{-1} & 0 \\ 0 & 0 & (\underline{x}_2^T \underline{x}_2)^{-1} \end{bmatrix}$$

$$X^T y = \begin{bmatrix} x_0^T \\ x_1^T \\ x_2^T \end{bmatrix} [\underline{y}] = \begin{bmatrix} x_0^T y \\ x_1^T y \\ x_2^T y \end{bmatrix}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y = \begin{bmatrix} (x_0^T x_0)^{-1} x_0^T y \\ (x_1^T x_1)^{-1} x_1^T y \\ (x_2^T x_2)^{-1} x_2^T y \end{bmatrix}$$

SS(model) $\hat{\beta}^T X^T \underline{y}$

$$= [\hat{\beta}_0 \ \hat{\beta}_1 \ \hat{\beta}_2] \begin{bmatrix} x_0^T \\ x_1^T \\ x_2^T \end{bmatrix} \underline{y}$$

$$= [\hat{\beta}_0 \ \hat{\beta}_1 \ \hat{\beta}_2] \begin{bmatrix} x_0^T y \\ x_1^T y \\ x_2^T y \end{bmatrix}$$

$= \hat{\beta}_0 x_0^T y + \hat{\beta}_1 x_1^T y + \hat{\beta}_2 x_2^T y$ Patitions according to each colum of design matrix

So there is a contribution (clean) for each colum of the design matrix

Consider a reduced model: $y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$

Can be shown as above:

$\hat{\beta}_0 = (x_0^T x_0)^{-1} x_0^T y$

$\hat{\beta}_1 = (x_1^T x_1)^{-1} x_1^T y$

So the full and reduced models have the same parameter estimates for β_0 and β_1

The reduced model has:

$SS_{Red}(\text{Model}) = \hat{\beta}_0 x_0^T y + \hat{\beta}_1 x_1^T y$

Orthogonally gives us separation we don't normally get in design matrix

Therefore: $SS(\text{Model}) - SS_{Red}(\text{Model}) = \hat{\beta}_2 x_2^T y$

$R(\beta_2 | \beta_0, \beta_1) = \hat{\beta}_2 x_2^T y$

Which is independent of $\hat{\beta}_0, \hat{\beta}_1$ and it's regression sum of squares we'd get if we regress y on X_2 Alone

$R(\beta_2 | \beta_0, \beta_1) = R(\beta_2 | \beta_0)$

In orthogonal model: Partial SS = Sequential SS

Example: Regression Through The Origin

Consider problem 3 on problem sheet 2

M1: Full: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

Reduced: $y_i = \beta_1 x_i + \varepsilon_i$

M2: Full: $y_i = a_0 + a_1(x_i - \bar{x}) + \varepsilon_i$

Reduced: $y_i = a_1(x_i - \bar{x}) + \varepsilon_i$

M1 Full: $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$ $\quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

Reduced: $Q = \sum (y_i - \beta_1 x_i)^2$

$dQ/d\hat{\beta} = 2 \sum (y_i - \beta_1 x_i)(x_i) = 0$

$\hat{\beta}_1 = \sum x_i y_i / \sum x_i^2$ (Different from full model)

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

M2: Full $a_1 = \frac{S_{xy}}{S_{xx}} = \hat{\beta}_1$ (full) $\quad \hat{a}_0 = \bar{y}$

Reduced: $y_i = a_1(x_i - \bar{x}) + \varepsilon_i$

$Q = \sum (y_i - a_1(x_i - \bar{x}))^2$

$\frac{dQ}{da_1} = \sum (x_i - \bar{x})(y_i - a_1(x_i - \bar{x}))$

$= \sum (x_i - \bar{x}) y_i - \hat{a}_1 \sum (x_i - \bar{x})^2$

$\hat{a}_1 = \frac{\sum (x_i - \bar{x}) y_i}{\sum (x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$ Orthogonal

So the values of $\hat{a}_1$ are the same as the full and reduced model

Making the transformation $z_i = x_i - \bar{x}$ gives an orthogonality transform

Write down design matrix for full M2

$$\begin{bmatrix} 1 & x_1-\bar{x} \\ 1 & x_2-\bar{x} \\ \vdots & \vdots \\ 1 & x_n-\bar{x} \end{bmatrix} \qquad x_0^T x_1 = [1 \; \cdots \; 1] \begin{bmatrix} x_1-\bar{x} \\ x_2-\bar{x} \\ \vdots \\ x_n-\bar{x} \end{bmatrix} = \Sigma(x_i-\bar{x}) = 0$$

x_0 x_1

2.12 MULTICOLLINEARITY

Recall the normal equations in matrix form $(X^TX)\hat{\beta} = X^Ty$

We made the assumption that X^TX is invertible

For $(X^TX)^{-1}$ to exist, the colums of X^TX have to be linearly independent

If the colums are not linearly independent it can suggest that there is some redundancy of information in the predictors (model) $\Rightarrow$ i.e. the predictors are giving the same information about the mean of y

If x_1 and x_2 are the proportion of water and solids in beer, then $x_1+x_2=1$ and we have a linear dependency

- The correct approach would be to fit an intercept and either x_1 or x_2

- If there is linear dependence in the colums of X, we can't invert X^TX
- This is called multicollinearity

- One could have approximate or close to linear dependence between columns. leads to large values in $(X^TX)^{-1}$ which in turn leads to large standard errors for $\hat{\beta}$
- Unreasonably large CI for $\hat{\beta}$

$$X = \begin{bmatrix} 1 & x_{11} & 1-x_{11} \\ 1 & x_{21} & 1-x_{21} \\ \vdots & \vdots & \vdots \\ 1 & x_{n1} & 1-x_{n1} \end{bmatrix}$$