STAT 332 Regression Analysis

Ch.2 Linear regression with one independent variable:

→ Simple linear regression model:
(with distribution of error terms unspecified)

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i \quad , \quad i = 1, 2, \ldots, n$$

Y_i -- value of response variable Y (observations)

X_i ... value of independent variable (constant)

ξ_i -- random error term

$$E(\xi_i) = 0 \quad , \quad Var(\xi_i) = \sigma^2 \quad , \quad cov(\xi_i, \xi_j) = 0 \quad \forall\, i \neq j$$

β_0, β_1 ... regression coefficients

(β_1) → slope of regression line
(change in mean of the prob. distribution of Y per unit increase in X)

(β_0) → Y-intercept of the regression line
(mean of the prob. distribution of Y at level X=0)

$$E(Y) = \beta_0 + \beta_1 X \quad , \text{ mean response}$$

$$Y_i = \underbrace{\beta_0 + \beta_1 X}_{\text{constr. term}} + \underbrace{\xi_i}_{\text{random term}}$$

$$Var(Y_i) = \sigma^2$$

* Alternative versions of model:

1- $Y_i = \beta_0 X_0 + \beta_1 X_i + \xi_i \quad , X_0 = 1$

2- $Y_i = \beta_0^* + \beta_1 (X_i - \bar{X}) + \xi_i \quad , \beta_0^* = \beta_0 + \beta_1 \bar{X}$

Estimation of regression function:

→ Method of least squares:

$$\varepsilon_i = Y_i - (\beta_0 + \beta_1 X_i)$$

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2 \quad \text{sum of squared deviations}$$

Values of β_0, β_1 that min. Q $\quad (b_0, b_1)$

$$\frac{\partial Q}{\partial \beta_0} = -2 \sum (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\frac{\partial Q}{\partial \beta_1} = -2 \sum X_i (Y_i - \beta_0 - \beta_1 X_i) = 0$$

$$\sum (Y_i - b_0 - b_1 X_i) = 0 \quad , \quad \sum X_i (Y_i - b_0 - b_1 X_i) = 0$$

$$\sum Y_i = n b_0 + b_1 \sum X_i \quad (1)$$

$$\sum X_i Y_i = b_0 \sum X_i + b_1 \sum X_i^2 \quad (2)$$

normal equations

Calculate

$$\sum X_i, \ \sum X_i^2, \ \sum Y_i, \ \sum X_i Y_i$$

then solving eqn's (1), (2)

multiply (1) * $\sum X_i$ $\qquad \sum X_i \sum Y_i = n b_0 \sum X_i + b_1 (\sum X_i)^2$

(2) * n $\qquad n \sum X_i Y_i = n b_0 \sum X_i + n b_1 \sum X_i^2$

$$n \sum X_i Y_i - \sum X_i \sum Y_i = b_1 \left[n \sum X_i^2 - (\sum X_i)^2 \right]$$

$$b_1 = \frac{n \sum X_i Y_i - \sum X_i \sum Y_i}{n \sum X_i^2 - (\sum X_i)^2} = \frac{\sum X_i Y_i - \frac{1}{n} \sum X_i \sum Y_i}{\sum X_i^2 - \frac{1}{n} (\sum X_i)^2}$$

$$b_0 = \frac{1}{n} \left[\sum Y_i - b_1 \sum X_i \right] = \bar{Y} - b_1 \bar{X}$$

Estimated regression function

$$\hat{Y} = b_0 + b_1 X$$

Residuals:

ith residual $e_i = Y_i - \hat{Y}_i = Y_i - b_0 - b_1 X_i$

note that:

1- $\sum_{i=1}^{n} e_i = 0$ from normal eqn. (1)

2- $\sum_{i=1}^{n} e_i^2$ is a minimum.

3- $\sum_{i=1}^{n} X_i e_i = 0$ from normal eqn. (2)

4- $\sum_{i=1}^{n} Y_i e_i = 0$

5- Regression line always goes though point $(\bar{X}, \bar{Y})$

proof: from alternative regression model:

$Y_i = \beta_0^* + \beta_1 (X_i - \bar{X}) + \xi_i$, then

$b_0^* = b_0 + b_1 \bar{X} = \bar{Y} - b_1 \bar{X} + b_1 \bar{X} = \bar{Y}$

$\hat{Y} = \bar{Y} + b_1 (X - \bar{X})$ $\quad X = \bar{X} \Rightarrow \hat{Y} = \bar{Y}$

Estimation of error term ξ_i variance: (σ^2)

sum of squares $SSE = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2$

$= \sum (Y_i - b_0 - b_1 X_i)^2$

$= \sum e_i^2$

mean square: $MSE = \frac{SSE}{n-2} = \frac{\sum e_i^2}{n-2}$

$(df = n-2)$

MSE is an unbiased estimator of σ^2

$E(MSE) = \sigma^2$

note: point estimate for σ is $\sqrt{MSE}$

Ch.3 Inferences in regression analysis:

Normal error regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad , \quad i = 1, 2, \ldots, n$$

$$\varepsilon_i \sim N(0, \sigma^2) \quad \text{- independent}$$

$$Y_i \sim N(\beta_0 + \beta_1 X_i \, , \, \sigma^2)$$

$\longrightarrow$ Inferences concerning β_1

* point estimator

$$\hat{\beta}_1 = b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

* Sampling distribution of b_1 is normal

proof

$$b_1 = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sum (X_i - \bar{X})^2}$$

$$= \frac{\sum (X_i - \bar{X}) Y_i}{\sum (X_i - \bar{X})^2} - \frac{\sum (X_i - \bar{X}) \bar{Y}}{\sum (X_i - \bar{X})^2}$$

$\to 0$ since $\sum (X_i - \bar{X}) = 0$

let $k_i = \dfrac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}$

$$\sum (X_i - \bar{X})(Y_i - \bar{Y}) = \sum (X_i - \bar{X}) Y_i$$

then $b_1 = \sum k_i Y_i$ is a linear combination of Y_i

since Y_i independent normally distributed r.v.'s

then b_1 is normally distributed.

* $E(b_1) = \beta_1$ prove ??

b_1 is an unbiased estimator of β_1

$$b_1 = \frac{\sum (x_i - \bar{x})(Y_i - \bar{Y})}{\sum (x_i - \bar{x})^2}$$

$$k_i = \frac{x_i - \bar{x}}{\sum (x_i - \bar{x})^2}$$, since $\sum (x_i - \bar{x}) = 0$

then $\sum k_i = 0$

$$\therefore b_1 = \sum k_i (Y_i - \bar{Y}) = \sum k_i Y_i - \bar{Y} \sum k_i = \sum k_i Y_i$$

$$\therefore E(b_1) = E(\sum k_i Y_i) = \sum k_i E(Y_i)$$

$$= \sum k_i (\beta_0 + \beta_1 x_i)$$

$$= \beta_0 \sum k_i + \beta_1 \sum k_i x_i$$

$$\because \sum k_i = 0 \text{ , } \sum k_i x_i = \frac{\sum (x_i - \bar{x}) x_i}{\sum (x_i - \bar{x})^2}$$

$$= \frac{\sum (x_i - \bar{x})(x_i - \bar{x})}{\sum (x_i - \bar{x})^2} = 1$$

$$\therefore E(b_1) = \beta_1 \quad \#$$

* $Var(b_1) = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$ prove ??

$$\sigma^2(b_1) = \sigma^2 (\sum k_i Y_i) = \sum k_i^2 \sigma^2(Y_i)$$

$$= \sigma^2 \sum k_i^2 = \frac{\sigma^2}{\sum (x_i - \bar{x})^2}$$

where $\sum k_i^2 = \frac{1}{\sum (x_i - \bar{x})^2}$.

Estimated variance of b_1

$$S^2(b_1) = \frac{MSE}{\sum (X_i - \bar{X})^2}$$

note: $\sum (X_i - \bar{X})^2 = \sum X_i^2 - \frac{1}{n}(\sum X_i)^2$

$$\frac{b_1 - \beta_1}{S(b_1)} \sim t(n-2)$$

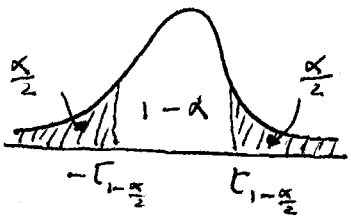
* $(1-\alpha)100\%$ Confidence interval for β_1

$$\left[b_1 - t_{1-\frac{\alpha}{2}} S(b_1) \;,\; b_1 + t_{1-\frac{\alpha}{2}} S(b_1) \right]$$

* Tests concerning β_1

Test statistic $T = \frac{b_1}{S(b_1)}$, α level of signific.

Two-sided test



$H_0: \beta_1 = 0 \quad (\beta_1 = \beta_{1,0})$

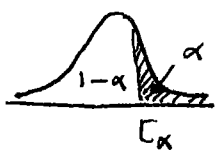
$H_1: \beta_1 = 0 \quad (\beta_1 \neq \beta_{1,0})$

$$|T| = \left| \frac{b_1 - \beta_{1,0}}{S(b_1)} \right|$$

if $|T| \leq t_{1-\frac{\alpha}{2}}$ accept H_0

$|T| > t_{1-\frac{\alpha}{2}}$ reject H_0

one-sided test



$H_0: \beta_1 = \beta_{1,0}$

$H_1: \beta_1 > \beta_{1,0}$

$H_0: \beta_1 = \beta_{1,0}$

$H_1: \beta_1 < \beta_{1,0}$

→ Inferences concerning β_0

* point estimator:

$$\hat{\beta}_0 = b_0 = \bar{Y} - b_1 \bar{X}$$

* sampling distribution of b_0 is normal

since b_0 is a linear combination of Y_i

* mean $\quad E(b_0) = \beta_0$

* variance $\quad \sigma^2(b_0) = \sigma^2 \dfrac{\sum X_i^2}{n \sum (X_i - \bar{X})^2}$

$$= \sigma^2 \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right]$$

* Estimated variance of b_0

$$s^2(b_0) = MSE \left[\frac{1}{n} + \frac{\bar{X}^2}{\sum (X_i - \bar{X})^2} \right]$$

$$\frac{b_0 - \beta_0}{s(b_0)} \sim t(n-2)$$

* $(1-\alpha)100\%$ confidence interval for β_0

$$\left[b_0 - t_{1-\frac{\alpha}{2}} s(b_0) \;,\; b_0 + t_{1-\frac{\alpha}{2}} s(b_0) \right]$$

* Interval estimation of $E(Y_h)$

$E(Y_h)$ mean of Y (Y_h) when $X = X_h$

$\hat{Y}_h = b_0 + b_1 X_h$ $\hat{Y}_h$ point estimate of $E(Y_h)$

→ sampling distribution of $\hat{Y}_h$ is normal

mean: $E(\hat{Y}_h) = \beta_0 + \beta_1 X_h$

variance: $\sigma^2(\hat{Y}_h) = \sigma^2 \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$

* $\text{Cov}(\bar{Y}, b_1) = 0$ prove ??

$\bar{Y} = \frac{1}{n} \sum Y_i$, $b_1 = \sum k_i Y_i$, $k_i = \frac{X_i - \bar{X}}{\sum (X_i - \bar{X})^2}$

since Y_i's are independent random variables

$$\sigma(\bar{Y}, b_1) = \sum \left(\frac{1}{n}\right) k_i \sigma^2(Y_i)$$

$$= \frac{\sigma^2}{n} \sum k_i$$

since $\sum k_i = \frac{\sum (X_i - \bar{X})}{\sum (X_i - \bar{X})^2} = 0$

then $\sigma(\bar{Y}, b_1) = 0$ #

note: theorem: if Y_i's are independent r.v.'s, then the covariance of two linear combinations of Y_i's $\sum a_i Y_i$, $\sum c_i Y_i$: equals $\sum a_i c_i \sigma^2(Y_i)$

Estimated variance of $\hat{Y}_h$

$$s^2(\hat{Y}_h) = MSE \left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\Sigma (X_i - \bar{X})^2} \right]$$

$(1-\alpha)100\%$ confidence interval for $E(Y_h)$

$$\left[\hat{Y}_h - t_{1-\frac{\alpha}{2}} s(\hat{Y}_h) , \hat{Y}_h + t_{1-\frac{\alpha}{2}} s(\hat{Y}_h) \right]$$

→ Prediction of new observations:

* when regression parameters β_0, β_1 are known

$(1-\alpha)100\%$ prediction interval for $Y_{h(new)}$ are:

$$\left[E(Y_h) - z_{1-\frac{\alpha}{2}} \sigma , E(Y_h) + z_{1-\frac{\alpha}{2}} \sigma \right]$$

where $E(Y_h) = \beta_0 + \beta_1 X_h$

* when β_0, β_1 are unknown:

$(1-\alpha)100\%$ prediction interval for $Y_{h(new)}$ are:

$$\left[\hat{Y}_h - t_{1-\frac{\alpha}{2}} s(Y_{h(new)}) , \hat{Y}_h + t_{1-\frac{\alpha}{2}} s(Y_{h(new)}) \right]$$

where, $s^2(Y_{h(new)}) = s^2(\hat{Y}_h) + MSE$

$(1-\alpha)100\%$ prediction interval for $\bar{Y}_{h(new)}$:

($\bar{Y}_{h(new)}$ ↓ mean of m new observations)

$$\left[\hat{Y}_h - t_{1-\frac{\alpha}{2}} s(\bar{Y}_{h(new)}) , \hat{Y}_h + t_{1-\frac{\alpha}{2}} s(\bar{Y}_{h(new)}) \right]$$

where, $s^2(\bar{Y}_{h(new)}) = s^2(\hat{Y}_h) + \frac{MSE}{m}$

* Analysis of variance approach to regression analysis :

→ sum of squares of total deviation of Y_i

$$SSTO = \sum (Y_i - \bar{Y})^2$$

→ sum of squares (Error) $SSE = \sum (Y_i - \hat{Y}_i)^2$

→ sum of squares (Regression) $SSR = \sum (\hat{Y}_i - \bar{Y})^2$

$$SSTO = SSR + SSE \qquad \text{prove ??}$$

proof: $\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2$??

↓ SSTO

$$\sum (Y_i - \bar{Y})^2 = \sum (\hat{Y}_i - \bar{Y} + Y_i - \hat{Y}_i)^2$$

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

$$= \sum (\hat{Y}_i - \bar{Y})^2 + \sum (Y_i - \hat{Y}_i)^2 + 2 \underbrace{\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i)}_{\to 0}$$

$$= SSR + SSE$$

since, $\sum (\hat{Y}_i - \bar{Y})(Y_i - \hat{Y}_i) = \sum \hat{Y}_i (Y_i - \hat{Y}_i) - \bar{Y} \sum (Y_i - \hat{Y}_i)$

$$= \underbrace{\sum \hat{Y}_i e_i}_{\to 0} - \bar{Y} \underbrace{\sum e_i}_{\to 0} = 0$$

#

note: $SSTO = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = \sum Y_i^2 - n\bar{Y}^2$

$$SSR = b_1 \left(\sum x_i Y_i - \frac{1}{n} \sum x_i \sum Y_i \right)$$

$$= b_1^2 \sum (x_i - \bar{x})^2 \qquad \text{(prove ??)}$$

* $SSR = b_1^2 \sum (x_i - \bar{x})^2$

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

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

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

$$= \sum [b_1 (x_i - \bar{x})]^2$$

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

* ANOVA table for simple linear regression:

source of variation	SS	df	MS	F^*
regression	$SSR = b_1^2 [\sum x_i^2 - \frac{(\sum x_i)^2}{n}]$	1	$MSR = \frac{SSR}{1}$	$\frac{MSR}{MSE}$
error	$SSE = SSTO - SSR$	$n-2$	$MSE = \frac{SSE}{n-2}$	
Total	$SSTO = \sum y_i^2 - n\bar{y}^2$	$n-1$		

F-test of $H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$, test statistic $F^* = \frac{MSR}{MSE}$

if $F^* \leq F_{1-\alpha}$ accept H_0

$F^* > F_{1-\alpha}$ reject H_0

$F_{1-\alpha}(df_1 = 1; df_2 = n-2)$

note: for given α F-test and t-test ($H_0: \beta_1 = 0$, $H_1: \beta_1 \neq 0$) are equivalent.

→ Coefficient of determination: مدى إمكانية التنبؤ بقيمة Y

$$r^2 = \frac{SSTO - SSE}{SSTO} = \frac{SSR}{SSTO}$$

* uncertainty in predicting Y

$$0 \leq r^2 \leq 1$$

- $r^2 = 1$ (i.e. $SSE = 0$, X accounts for all) variations in Y

- $r^2 = 0$ (there's no linear association between X and Y in the sample data)

- as r^2 increases, the more is total variation of Y reduced by X.
  i.e. the larger is r^2, the more is the degree of linear association between X and Y.

⟶ Coefficient of correlation

$$r = \pm\sqrt{r^2} \qquad -1 \leq r \leq 1$$

r doesn't have a clear interpretation as r^2.

$$r = \frac{\sum x_i y_i - \frac{1}{n}\sum x_i \sum y_i}{\sqrt{\left[\sum x_i^2 - \frac{1}{n}(\sum x_i)^2\right]\left[\sum y_i^2 - \frac{1}{n}(\sum y_i)^2\right]}}$$

Ch.4 Aptness of model & Remedial measures:

between 12

) ple linear regression model is appropriate

riation of the application data ??

$= \beta_0 + \beta_1 X_i + \xi_i$

degree of

$= Y_i - E(Y_i)$ observed error

$E(\xi_i) = 0 \;,\; \sigma^2(\xi_i) = \sigma^2$

$= Y_i - \hat{Y}_i$ ith residual

$E(e_i) = \frac{\sum e_i}{n} = 0$

is r^2. variance $\frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = MSE$

standardized residuals:

$$\frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

nfunes from model to be studied by residuals

raphical analysis of residuals)

Nonlinearity of regression function.

Non-constancy of error term's variance.

Presence of outliers (extreme observations).

- Non-normality of error terms.
- Non-independence of error terms.
- Omission of important independent variables.

- $r^2 = 0$ (there's no linear association between X and Y in the sample data)

- as r^2 increases, the more is total variation of Y reduced by X.
  i.e. the larger is r^2, the more is the degree of linear association between X and Y.

→ coefficient of correlation

$$r = \pm\sqrt{r^2} \qquad -1 \leq r \leq 1$$

r doesn't have a clear interpretation as r^2.

$$r = \frac{\Sigma x_i y_i - \frac{1}{n}\Sigma x_i \Sigma y_i}{\sqrt{\left[\Sigma x_i^2 - \frac{1}{n}(\Sigma x_i)^2\right]\left[\Sigma y_i^2 - \frac{1}{n}(\Sigma y_i)^2\right]}}$$

Ch.4 Aptness of model & Remedial measures:

→ simple linear regression model is appropriate for the application data ??

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

$\xi_i = Y_i - E(Y_i)$ observed error

$E(\xi_i) = 0$, $\sigma^2(\xi_i) = \sigma^2$

$e_i = Y_i - \hat{Y}_i$ ith residual

$E(e_i) = \frac{\sum e_i}{n} = 0$

variance $\frac{\sum (e_i - \bar{e})^2}{n-2} = \frac{\sum e_i^2}{n-2} = MSE$

* standandized residuals:

$$\frac{e_i - \bar{e}}{\sqrt{MSE}} = \frac{e_i}{\sqrt{MSE}}$$

Departures from model to be studied by residuals (graphical analysis of residuals)

1- Nonlinearity of regression function.

2- Non-constancy of error term's variance.

3- Presence of outliers (extreme observations).

4- Non-normality of error terms.

5- Non-independence of error terms.

6- Omission of important independent variables.

4-3 Tests involving residuals:

Test for randomness / constancy of variance / Outliers / Normality.

4-4 F-test for lack of fit:

whether or not a regression fun. adequately fits data. Assumes that the observations Y for a given X are: (1) independent, (2) normally distributed and (3) distribution of Y has variance σ^2.

Replications: repeated trials for the same level of X.

Replicates: resulting observations.

Decomposition of SSE

Denote different levels of X as: $X_1, X_2, \ldots, X_c$

no. of observations for jth level of X as n_j

then $n = \sum_{j=1}^{c} n_j$

observations	$X_1 = 75$	$X_2 = 100$	$X_3 = 125$	$X_4 = 150$	$X_5 = 175$	$X_6 = 200$
$i = 1$	28	112	160	152	156	124
$i = 2$	42	136	150		124	104
Mean $\bar{Y}_j$	35	124	155	152	140	114

pure error sum of squares:

$$SSPE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 \qquad df = n - c$$

Pure error mean square:

$$MSPE = \frac{SSPE}{n - c}$$

* F-test for lack-of-fit :

→ replications : repeated trials for the same level of X.

→ replicates : resulting observations

Denote different levels of X as: $X_1, X_2, \ldots, X_c$

no. of observations for X_j is n_j , $n = \sum_{j=1}^{c} n_j$

pure-error sum of squares:

$$SSPE = \sum_{j=1}^{c} \sum_{i=1}^{n_j} (Y_{ij} - \bar{Y}_j)^2 \qquad df = n - c$$

$$MSPE = \frac{SSPE}{n-c}$$

Lack-of-fit sum of squares:

$$SSLF = SSE - SSPE$$

* ANOVA table for testing lack-of-fit :

source of variation	SS	df	MS	F^*
regression	SSR	1	MSR	$F^* = \frac{MSR}{MSE}$
error	SSE	$n-2$	MSE	
lack of fit	SSLF	$c-2$	MSLF	$F^* = \frac{MSLF}{MSPE}$
pure error	SSPE	$n-c$	MSPE	
Total	SSTO	$n-1$		

Test statistic $F^* = \frac{MSLF}{MSPE}$ $\qquad F_{1-\alpha}(c-2, n-c)$ (df_1, df_2)

H_0: $E(Y) = \beta_0 + \beta_1 X$

H_1: $E(Y) \neq \beta_0 + \beta_1 X$

if $F^* \leq F_{1-\alpha}$ accept H_0

$F^* > F_{1-\alpha}$ reject H_0

4-6 Transformations:

Transformation of one or both of the original variables X, Y, before carrying out regression analysis.

Useful transformations:

$Y' = \sqrt{Y}$, $X' = \sqrt{X}$

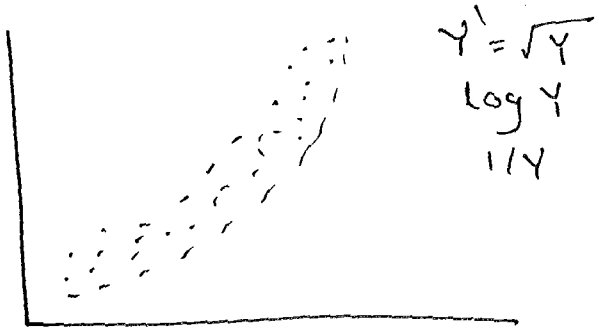
$Y' = \log Y$, $X' = \log X$

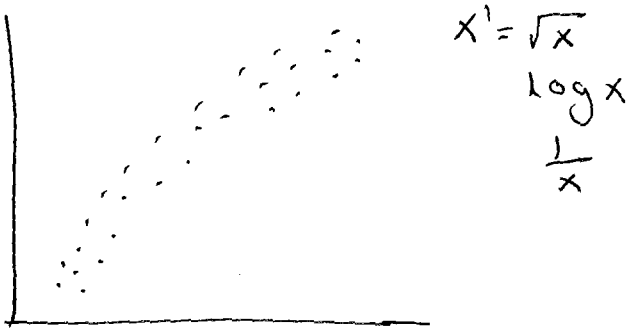
$Y' = \frac{1}{Y}$, $X' = \frac{1}{X}$

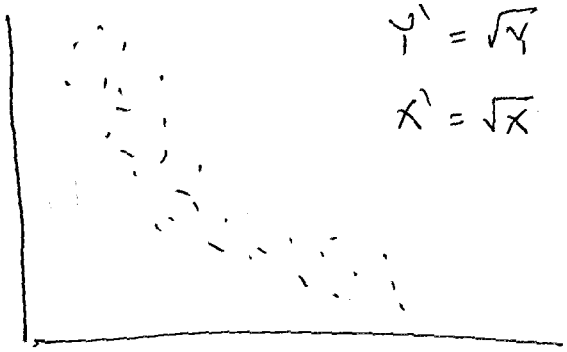
$\sigma_i^2 \sim E(Y_i)$ $\quad Y' = \sqrt{Y}$

$\sigma_i \sim E(Y_i)$ $\quad Y' = \log Y$

$\sqrt{\sigma_i} \sim E(Y_i)$ $\quad Y' = \frac{1}{Y}$







$Y' = \sqrt{Y}$ or $\log Y$ or $\frac{1}{Y}$

$X' = \sqrt{X}$ or $\log X$ or $\frac{1}{X}$

Ch.6 Matrix approach to simple regression:

→ Simple linear regression model in matrix terms:

$$Y_i = \beta_0 + \beta_1 X_i + \xi_i \quad , \quad i = 1, 2, \ldots, n$$

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} \quad , \quad X = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix} \quad , \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} \quad , \quad \xi = \begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix}$$

model in matrix terms:

$$\underset{n\times 1}{Y} = \underset{n\times 2}{X}\ \underset{2\times 1}{\beta} + \underset{n\times 1}{\xi}$$

$$E(\xi) = E\begin{bmatrix} \xi_1 \\ \xi_2 \\ \vdots \\ \xi_n \end{bmatrix} = \begin{bmatrix} E(\xi_1) \\ E(\xi_2) \\ \vdots \\ E(\xi_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} = 0$$

$$\sigma^2(\xi) = \sigma^2 I = \begin{bmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ 0 & 0 & \sigma^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & & \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{bmatrix}$$

Normal error model in matrix terms:

$$Y = X\beta + \xi \quad , \quad \xi \text{ vector of independent normal r.v.'s, with } E(\xi) = 0, \ \sigma^2(\xi) = \sigma^2 I$$

Least-square estimation of regression parameters:

normal eqn.'s in matrix terms:

$$X'X\ b = X'Y \quad , \qquad X'X = \begin{bmatrix} n & \Sigma X_i \\ \Sigma X_i & \Sigma X_i^2 \end{bmatrix}$$

$$b = (X'X)^{-1} X'Y \qquad X'Y = \begin{bmatrix} \Sigma Y_i \\ \Sigma X_i Y_i \end{bmatrix}$$

ANOVA results:

$$SSTO = \sum Y_i^2 - \frac{(\sum Y_i)^2}{n} = \sum Y_i^2 - n\bar{Y}^2$$

$$= Y'Y - \left(\frac{1}{n}\right) Y' 1\, 1\, Y$$

$$Y'Y = \sum Y_i^2$$

$$\left(\frac{1}{n}\right) Y' 1\, 1\, Y = \frac{(\sum Y_i)^2}{n}$$

$$SSE = \sum e_i^2 = e'e = (Y - Xb)'(Y - Xb)$$

$$= Y'Y - b' X'Y$$

$$SSR = SSTO - SSE$$

* Quadratic form example: (Sum of squares)

$$5 Y_1^2 + 6 Y_1 Y_2 + 4 Y_2^2$$

$$[Y_1 \;\; Y_2] \begin{bmatrix} 5 & 3 \\ 3 & 4 \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix}$$

In general: A $(n \times n)$ symmetric matrix

$$Y'AY = \sum_{i=1}^{n} \sum_{j=1}^{n} a_{ij} Y_i Y_j$$

Inferences in regression analysis:

1. Regression coefficients:

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

$$s^2(b) = MSE\,(X'X)^{-1}$$

$$\begin{bmatrix} s^2(b_0) & \cdots \\ \cdots & s^2(b_1) \end{bmatrix}$$

2. Joint confidence region for β_0 and β_1

$$\frac{(b-\beta)' X' X (b-\beta)}{2\,MSE} \sim F(1-\alpha; 2; n-2)$$

3. Mean response at X_h:

$$X_h = \begin{bmatrix} 1 \\ X_h \end{bmatrix}, \quad X_h' = [1 \quad X_h]$$

fitted value $\hat{Y}_h = X_h' b = [1 \quad X_h] \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$

$$= b_0 + b_1 X_h = \hat{Y}_h$$

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

$$s^2(\hat{Y}_h) = MSE\,(X_h' (X'X)^{-1} X_h) = X_h' s^2(b) X_h$$

4. Prediction of new observation:

$$s^2(Y_{h(new)}) = MSE + s^2(\hat{Y}_h)$$

$$= MSE + X_h' s^2(b) X_h$$

$$= MSE\,[1 + X_h' (X'X)^{-1} X_h]$$

Ch.7 Multiple regression:

Example: with two independent variables:

Model: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$

$, \varepsilon_i \sim N(0, \sigma^2)$

Consider the following data for the regression function of Y on X_1, X_2

i	Y	X_1	X_2
1	6.40	1.32	1.15
2	15.05	2.69	3.40
3	18.75	3.56	4.10
4	30.25	4.41	8.75
5	48.95	6.20	14.82
Sums	119.4	18.18	32.22
Sum squares		79.5402	325.8874

$$X'X = \begin{bmatrix} n & \Sigma X_1 & \Sigma X_2 \\ \Sigma X_1 & \Sigma X_1^2 & \Sigma X_1 X_2 \\ \Sigma X_2 & \Sigma X_1 X_2 & \Sigma X_2^2 \end{bmatrix} \qquad X'Y = \begin{bmatrix} \Sigma Y \\ \Sigma X_1 Y \\ \Sigma X_2 Y \end{bmatrix}$$

(a) obtain SSTO with accuracy

$$SSTO = \Sigma Y_i^2 - \frac{(\Sigma Y_i)^2}{n} = 3930.19 - \frac{(119.4)^2}{5}$$

$$= 1078.92$$

(b) $$X'X = \begin{bmatrix} 5 & 18.18 & 32.22 \\ 18.18 & 79.5402 & 155.732 \\ 32.22 & 155.732 & 325.887 \end{bmatrix}$$

$$X'Y = \begin{bmatrix} 119.4 \\ 552.58 \\ 1125.53 \end{bmatrix}$$

(c) Find $b = (X'X)^{-1} X'Y$

You can take $(X'X)^{-1} = \begin{bmatrix} 3.31174 & -1.79982 & 0.53265 \\ -1.79982 & 1.17542 & -0.38280 \\ 0.53265 & -0.38280 & 0.13333 \end{bmatrix}$

$$b = \begin{bmatrix} 0.40143 \\ 2.65613 \\ 2.14477 \end{bmatrix}$$

(d) Write down regression equation:

$$Y = 0.40143 + 2.65613 X_1 + 2.14477 X_2$$

(e) let $SSR = 1078.38$

$$SSE = SSTO - SSR$$

$$= 1078.92 - 1078.38 = 0.54$$

ANOVA table

SV	SS	df	MS	F^*
Regression	1078.38	2 (p-1)	539.19	1992.19
Error	0.54	2 (n-p)	0.27	
Total	1078.92	4 (n-1)		

(p = 3 regression parameters)

(f) Test $H_0: \beta_1 = \beta_2 = 0$ at $\alpha = 0.05$

$$F^* = \frac{MSR}{MSE} \qquad F_{1-\alpha}(p-1; n-p)$$

$$F_{0.95}(2; 2) = 19.0$$

Since $F^* > F_{1-\alpha}$ reject H_0

(g) find 95% C.I. of $E(Y_h)$ when,

$$X_h = \begin{bmatrix} 1 \\ 6 \\ 15 \end{bmatrix}$$

$X_{h1} = 6$, $X_{h2} = 15$??

$$s^2(b) = MSE\ (X'X)^{-1}$$

$$= 0.27\ (X'X)^{-1}$$

باستخدام البرنامج

$$= \begin{bmatrix} 0.894169 & & \\ & & 0.316823 \\ & & \\ & & 0.036 \end{bmatrix}$$

$s^2(b_0)$ ← $s(b_0)$

$s^2(b_1)$ → $s(b_1)$

$s^2(b_2)$ → $s(b_2)$

باستخدام البرنامج

$$s^2(\hat{Y}_h) = X_h'\ s^2(b)\ X_h$$

$$= 0.2789$$

$$s(\hat{Y}_h) = 0.5281$$

$$\hat{Y}_h = X_h'\ b$$

$$= 48.5097$$

95% C.I. for $E(Y_h)$

$$\left[48.5097 - 4.303\,(0.5281),\ 48.5097 + 4.303\,(0.5281) \right]$$

note: $t_{1-\frac{\alpha}{2}}(n-p)$ $t_{0.975}(2)$

=

Minitab خطوات الحل باستخدام

1- ادخل البيانات X1, X2, Y

2- Stat → Regression

Response select Y

Predictors select X1, X2

Storage: ✓ Coefficient

✓ X'X inverse

3- `print m1` (#$(X'X)^{-1}$)

4- `name C5 as Y2`

5- `let Y2 = Y * Y`

6- `sum Y k11`

7- `sum Y2 k12`

8- `n X1 k1`

ΣY^2	ΣY	n
3930.19	119.4	5

↓ SSTO

9- `let k10 = k12 - k11**2/k1`

10- `print k10`

11- `Ser C6` `5(1)` `end`

or أو ادخل 5 1's في عمود C6 باليد

12- `copy C6 C1 C2 m11` # X

13- `Transpose m11 m12` # X_p

14- `multiply m12 m11 m22` # X_pX

15- `print m22`

16- `let k13 = sum(X1*Y)`

17- `print k13`

18- let K14 = sum (X2*Y)

19- print K14

20- read 3 L m3

copy استخدام
paste
لادخال المصفوفة

119.40
552.58
1125.53

21- print m3 لطباعة

22- multiply m1 m3 m4 لحساب $b = (X'X)^{-1}(X'Y)$

23- print m4

24- multiply 0.27 m1 m5 لحساب $S^2(b) = 0.27 (X'X)^{-1}$
↑ MSE

25- print m5

26- read 3 1 m6 لادخال X_h

27- transpose m6 m7 X'_h

28- multiply m7 m5 m8 X'_h $S^2(b)$

29- multiply m8 m6 m9 X'_h $S^2(b)$ X_h
$= S^2(\hat{Y}_h)$

30- multiply m7 m4 m10 $\hat{Y}_h = X'_h b$

Exercises:

ch. 4	4.13 , 4.14
ch. 6	6-5 , 6.18 , 6.19 , 6.23 , 6.24
ch. 7	7.8 , 7.9 , 7.10