MULTIPLE REGRESSION

Model: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$

$$\sum \hat{u}_i^2 = \sum (Y_i - \alpha_0 - \alpha_1 X_{1i} - \alpha_2 X_{2i})^2$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\alpha}_0} = -2 \sum (Y_i - \alpha_0 - \alpha_1 X_{1i} - \alpha_2 X_{2i}) \Rightarrow \sum \hat{u}_i = 0$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\alpha}_1} = -2 \sum X_{1i} (Y_i - \alpha_0 - \alpha_1 X_{1i} - \alpha_2 X_{2i}) \Rightarrow \sum \hat{u}_i X_{1i} = 0$$

$$\frac{\partial \sum \hat{u}_i^2}{\partial \hat{\alpha}_2} = -2 \sum X_{2i} (Y_i - \alpha_0 - \alpha_1 X_{1i} - \alpha_2 X_{2i}) \Rightarrow \sum \hat{u}_i X_{2i} = 0$$

Normal eq^n's:

$\boxed{\sum \hat{u}_i = 0}$, $\boxed{\sum \hat{u}_i X_{1i} = 0}$, $\boxed{\sum \hat{u}_i X_{2i} = 0}$

$$\sum \hat{u}_i X_{1i} = 0$$

$$\Rightarrow \sum \hat{u}_i (x_{1i} + \bar{X}_1) = 0$$

$$\Rightarrow \sum \hat{u}_i x_{1i} + \bar{X}_1 \sum \hat{u}_i = 0 \quad \Rightarrow \boxed{\sum \hat{u}_i x_{1i} = 0}$$

Similarly, $\sum \hat{u}_i x_{2i} = 0$

$$\sum \hat{u}_i \hat{Y}_i = \sum \hat{u}_i (\hat{\alpha}_0 + \hat{\alpha}_1 X_{1i} + \hat{\alpha}_2 X_{2i}) = \hat{\alpha}_0 \sum \hat{u}_i + \hat{\alpha}_1 \sum \hat{u}_i X_{1i} + \hat{\alpha}_2 \sum \hat{u}_i X_{2i} = 0$$

$$\sum \hat{u}_i \hat{y}_i = \sum \hat{u}_i (\hat{Y}_i - \bar{Y}) = \sum \hat{u}_i \hat{Y}_i - \bar{Y} \sum \hat{u}_i = 0$$

$$\boxed{\sum \hat{u}_i \hat{Y}_i = \sum \hat{u}_i \hat{y}_i = 0}$$

$\sum \hat{u}_i y_i \neq$ ~~$\sum \hat{u}_i (Y_i - \bar{Y}) \neq \sum \hat{u}_i Y_i - \bar{Y} \sum \hat{u}_i$~~

$$= \sum \hat{u}_i (\hat{y}_i + \hat{u}_i) = \sum \hat{u}_i \hat{y}_i + \sum \hat{u}_i^2 = \sum \hat{u}_i^2 = RSS.$$

$$RSS = \sum \hat{u}_i^2 = \sum \hat{u}_i y_i$$

$$\hat{\alpha}_1 = \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

$$\hat{\alpha}_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

$$\hat{\alpha}_0 = \bar{Y} - \hat{\alpha}_1 \bar{X}_1 - \hat{\alpha}_2 \bar{X}_2$$

$$E(Y_i | X_{1i}, X_{2i}) = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i}$$

• Implications of Multicollinearity

a) If there is perfect multicollinearity between X_1 and X_2, i.e., $X_{2i} = \lambda X_{1i}$

$X_{2i} = \lambda X_{1i}$ and $\bar{X}_2 = \lambda \bar{X}_1 \Rightarrow x_{2i} = \lambda x_{1i}$

$$\hat{\alpha}_1 = \frac{(\sum y_i x_{1i})(\sum \lambda^2 x_{1i}^2) - (\sum y_i \lambda x_{1i})(\sum \lambda^2 x_{1i}^2)}{(\sum x_{1i}^2)(\sum \lambda^2 x_{1i}^2) - (\sum \lambda x_{1i} x_{2i})^2} = \frac{0}{0}$$

Similarly,

$$\hat{\alpha}_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

$$= \frac{\lambda(\sum y_i x_{1i})(\sum x_{1i}^2) - \lambda(\sum y_i x_{1i})(\sum x_{1i}^2)}{\lambda^2(\sum x_{1i}^2)(\sum x_{1i}^2) - \lambda^2(\sum x_{1i}^2)^2} = \frac{0}{0}$$

∴ Slope coefficients come out to be undefined.

(b) If there is no multicollinearity between X_1 and X_2, i.e.

$$r_{12} = 0$$

Consider the following models:

(1) $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$

(2) $Y_i = \beta_0 + \beta_1 X_{1i} + v_i$

(3) $Y_i = \gamma_0 + \gamma_1 X_{2i} + w_i$

(4) $X_{2i} = \delta_0 + \delta_1 X_{1i} + \varepsilon_i$

(5) $X_{1i} = \theta_0 + \theta_1 X_{2i} + \eta_i$

$$\hat{\alpha}_1 = \frac{\left(\sum y_i x_{1i}\right)\left(\sum x_{2i}^2\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right) - \left(\sum x_{1i} x_{2i}\right)^2}$$

$$\Rightarrow \hat{\alpha}_1 = \frac{\frac{\sum y_i x_{1i}}{\sum x_{1i}^2} - \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{1i} x_{2i}\right)}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right)}}{1 - \frac{\left(\sum x_{1i} x_{2i}\right)^2}{\left(\sum x_{1i}^2\right)\left(\sum x_{2i}^2\right)}}$$

$$\Rightarrow \hat{\alpha}_1 = \frac{\left(\frac{\sum y_i x_{1i}}{\sum x_{1i}^2}\right) - \left(\frac{\sum y_i x_{2i}}{\sum x_{2i}^2}\right)\frac{\sum x_{1i} x_{2i}}{\sum x_{1i}^2}}{1 - \left[\frac{\sum x_{1i} x_{2i}}{\sqrt{\left(\sum x_{1i}^2\right)}\sqrt{\left(\sum x_{2i}^2\right)}}\right]^2}$$

$$\Rightarrow \hat{\alpha}_1 = \frac{\hat{\beta}_1 - \hat{\gamma}_1 \hat{\delta}_1}{1 - r_{12}^2} = \frac{\hat{\beta}_1 - \hat{\gamma}_1\left(r_{12}\frac{\sigma_2}{\sigma_1}\right)}{1 - r_{12}^2}$$

$$\hat{\alpha}_1 = \frac{\hat{\beta}_1 - \hat{\gamma}_1 \hat{\delta}_1}{1 - r_{12}^2} = \frac{\hat{\beta}_1 - \hat{\gamma}_1 \left(r_{12} \frac{\sigma_2}{\sigma_1} \right)}{1 - r_{12}^2}$$

σ_1 = Standard deviation of X_1

σ_2 = Standard deviation of X_2

If $r_{12} = 0$, $\hat{\alpha}_1 = \hat{\beta}_1$

$$\hat{\alpha}_2 = \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}$$

$$= \frac{\dfrac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2)}}{\dfrac{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2}{(\sum x_{1i}^2)(\sum x_{2i}^2)}}$$

$$= \frac{\dfrac{\sum y_i x_{2i}}{\sum x_{2i}^2} - \dfrac{\sum y_i x_{1i}}{\sum x_{1i}^2} \dfrac{\sum x_{1i} x_{2i}}{\sum x_{2i}^2}}{1 - \left[\dfrac{\sum x_{1i} x_{2i}}{\sqrt{(\sum x_{1i}^2)} \sqrt{(\sum x_{2i}^2)}} \right]^2}$$

$$\hat{\alpha}_2 = \frac{\hat{\gamma}_1 - \hat{\beta}_1 \hat{\theta}_1}{1 - r_{12}^2} = \frac{\hat{\gamma}_1 - \hat{\beta}_1 \left(r_{12} \frac{\sigma_1}{\sigma_2} \right)}{1 - r_{12}^2}$$

$r_{12} = 0 \quad \hat{\alpha}_2 = \hat{\gamma}_1$

Composition of Goodness-of-Fit

$$ESS = \sum \hat{y}_i^2 = \sum \hat{y}_i (\hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i}) = \hat{\alpha}_1 \sum \hat{y}_i x_{1i} + \hat{\alpha}_2 \sum \hat{y}_i x_{2i}$$

$$= \hat{\alpha}_1 \sum (y_i - \hat{u}_i) x_{1i} + \hat{\alpha}_2 \sum (y_i - \hat{u}_i) x_{2i}$$

$$= \hat{\alpha}_1 \sum y_i x_{1i} - \hat{\alpha}_1 \sum \hat{u}_i x_{1i} + \hat{\alpha}_2 \sum y_i x_{2i} - \hat{\alpha}_2 \sum \hat{u}_i x_{2i}$$

($\sum \hat{u}_i x_{1i} \downarrow 0$, $\sum \hat{u}_i x_{2i} \downarrow 0$)

$$= \hat{\alpha}_1 \sum y_i x_{1i} + \hat{\alpha}_2 \sum y_i x_{2i}$$

$$= \left\{ \frac{(\sum y_i x_{1i})(\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \right\} (\sum y_i x_{1i})$$

$$+ \left\{ \frac{(\sum y_i x_{2i})(\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \right\} (\sum y_i x_{2i})$$

$$= \left[\frac{(\sum y_i x_{1i})^2 (\sum x_{2i}^2) - (\sum y_i x_{2i})(\sum y_i x_{1i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \right]$$

$$+ \left[\frac{(\sum y_i x_{2i})^2 (\sum x_{1i}^2) - (\sum y_i x_{1i})(\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \right]$$

$$= \left[\frac{(\sum y_i x_{1i})^2 (\sum x_{2i}^2) + (\sum y_i x_{2i})^2 (\sum x_{1i}^2) - 2(\sum y_i x_{1i})(\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i} x_{2i})^2} \right]$$

$$= \frac{(\sum y_i x_{1i})^2 (\sum x_{2i}^2)}{(\sum x_{2i}^2)(\sum x_{1i}^2)} + \frac{(\sum y_i x_{2i})^2 (\sum x_{1i}^2)}{(\sum x_{1i}^2)(\sum x_{2i}^2)} - 2 \frac{(\sum y_i x_{1i})(\sum y_i x_{2i})(\sum x_{1i} x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2)}$$

$$1 - \left[\frac{\sum x_{1i} x_{2i}}{\sqrt{(\sum x_{1i}^2)(\sum x_{2i}^2)}} \right]^2$$

$$= \left[\frac{1}{1-r_{12}^2}\right]\left[\frac{(\sum y_i x_{1i})^2(\sum x_{2i}^2) + (\sum y_i x_{2i})^2(\sum x_{1i}^2) - 2(\sum y_i x_{1i})(\sum y_i x_{2i})(\sum x_{1i}x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2)(\sum y_i^2)}\right]$$

$$R^2 = \frac{ESS}{TSS} = \left(\frac{1}{1-r_{12}^2}\right)\left[\frac{(\sum y_i x_{1i})^2(\sum x_{2i}^2) + (\sum y_i x_{2i})^2(\sum x_{1i}^2) - 2(\sum y_i x_{1i})(\sum y_i x_{2i})(\sum x_{1i}x_{2i})}{(\sum x_{1i}^2)(\sum x_{2i}^2)(\sum y_i^2)}\right]$$

$$= \frac{1}{1-r_{12}^2}\left(r_{31}^2 + r_{32}^2 - 2r_{31}r_{32}r_{12}\right)$$

$$\boxed{R^2 = \frac{r_{31}^2 + r_{32}^2 - 2r_{31}r_{32}r_{12}}{1-r_{12}^2}}$$

$$\boxed{\text{var}(\hat{\alpha}_0) = \left[\frac{1}{n} + \frac{\bar{X}_1^2\sum x_{2i}^2 + \bar{X}_2^2\sum x_{1i}^2 - 2\bar{X}_1\bar{X}_2\sum x_{1i}x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - (\sum x_{1i}x_{2i})^2}\right]\sigma^2}$$

$$\text{var}(\hat{\alpha}_1) = \frac{\sum x_{2i}^2}{(\sum x_{1i}^2)(\sum x_{2i}^2) - (\sum x_{1i}x_{2i})^2}\sigma^2$$

$$= \frac{\sigma^2}{\sum x_{1i}^2 - \frac{(\sum x_{1i}x_{2i})^2}{\sum x_{2i}^2}} = \frac{\sigma^2}{\sum x_{1i}^2(1-r_{12}^2)}$$

$$\boxed{\text{var}(\hat{\alpha}_1) = \frac{\sigma^2}{\sum x_{1i}^2(1-r_{12}^2)}}$$

$$\boxed{\text{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1-r_{12}^2)}}$$

Adjusted R^2 ∞:

$$\bar{R}^2 = 1 - (1 - R^2)\left[\frac{n-1}{n-k}\right]$$

$$\text{cov}(\hat{\alpha}_1, \hat{\alpha}_2) = \frac{-r_{12}\sigma^2}{(1 - r_{12}^2)\sqrt{\sum x_{1i}^2}\sqrt{\sum x_{2i}^2}}$$

$$\hat{\sigma}^2 = \frac{\sum \hat{u}_i^2}{n-3}$$

Example:

Cobb-Douglas Production function

$$Y_i = \beta_1 X_{1i}^{\beta_2} X_{2i}^{\beta_3} e^{u_i}$$

$Y \rightarrow$ output, $X_1 \rightarrow$ labour input, $X_2 \rightarrow$ capital input,

$u \rightarrow$ stochastic disturbance term.

$$\ln Y_i = \ln \beta_1 + \beta_2 \ln X_{1i} + \beta_3 \ln X_{2i} + u_i$$

let $\beta_0 = \ln \beta_1$

$\beta_2 \rightarrow$ partial elasticity of output w.r.t. labour input.

$\beta_3 \rightarrow$ partial elasticity of output w.r.t. capital input.

$\beta_2 + \beta_3 \rightarrow$ returns to scale.

★ Partial Correlation Coefficients

$r_{1,2.3}$ = partial correlation coefficient between X_1 and X_2, holding Y constant.

$r_{1,3.2}$ = partial correlation coefficient between X_1 and Y, holding X_2 constant

$r_{2,3.1}$ = partial correlation coefficient between Y and X_2, holding X_1 constant

$$r_{1,2.3} = \frac{r_{12} - r_{13} r_{23}}{\sqrt{(1-r_{13}^2)(1-r_{23}^2)}}$$

$$r_{1,3.2} = \frac{r_{13} - r_{12} r_{23}}{\sqrt{(1-r_{12}^2)(1-r_{23}^2)}}$$

$$r_{2,3.1} = \frac{r_{23} - r_{12} r_{13}}{\sqrt{(1-r_{12}^2)(1-r_{13}^2)}}$$

$r_{1,2.3}^2$ → Coefficient of partial determination and is interpreted as the proportion of the variation in X_1 not explained by Y that has been explained by the inclusion of X_2 in the model.

• Comparing Two Coefficients of a Multiple Regression Model

Model : $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$

Test Hypothesis : $H_0 : \alpha_1 = \alpha_2$ or $H_0 : \alpha_1 - \alpha_2 = 0$

(a) Through t-test

Test Statistic : $$\theta = \frac{(\hat{\alpha}_1 - \hat{\alpha}_2) - (\alpha_1 - \alpha_2)}{SE(\hat{\alpha}_1 - \hat{\alpha}_2)} \sim t_{(n-k)}$$

• $$SE(\hat{\alpha}_1 - \hat{\alpha}_2) = \sqrt{var(\hat{\alpha}_1) + var(\hat{\alpha}_2) - 2\,cov(\hat{\alpha}_1, \hat{\alpha}_2)}$$

$$\theta = \frac{\hat{\alpha}_1 - \hat{\alpha}_2}{\sqrt{var(\hat{\alpha}_1) + var(\hat{\alpha}_2) - 2\,cov(\hat{\alpha}_1, \hat{\alpha}_2)}} \sim t_{(n-k)}$$

$$var(\hat{\alpha}_1) = \frac{\sigma^2}{\sum x_{1i}^2 (1-r_{12}^2)} \qquad var(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1-r_{12}^2)}$$

$$cov(\hat{\alpha}_1, \hat{\alpha}_2) = \frac{-r_{12}\sigma^2}{(1-r_{12}^2)\sqrt{\sum x_{1i}^2}\sqrt{\sum x_{2i}^2}}$$

(b) Through Restricted F-test

Unrestricted Model: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$

Restriction: $\alpha_1 = \alpha_2$

Restricted model: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_1 X_{2i} + u_i$

$= \alpha_0 + \alpha_1 (X_{1i} + X_{2i}) + u_i$

Test statistic:

$$\theta = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} \sim F_{m,(n-k)}$$

$m \rightarrow$ number of restrictions

Example:

$Y_i = \beta_1 X_{1i}^{\beta_2} X_{2i}^{\beta_3} e^{u_i}$

$\ln Y_i = \beta_0 + \beta_2 \ln X_{1i} + \beta_3 \ln X_{2i} + u_i$

$H_0: \beta_2 + \beta_3 = 1$

$$t = \frac{(\hat{\beta}_2 + \hat{\beta}_3) - 1}{\sqrt{var(\hat{\beta}_2) + var(\hat{\beta}_3) + 2cov(\hat{\beta}_2, \hat{\beta}_3)}}$$

F-test:

$\beta_2 = 1 - \beta_3$

Restricted Model:

$$\ln Y_i = \beta_0 + \ln X_{1i} + \beta_3 \ln\left(\frac{X_{2i}}{X_{1i}}\right) + u_i$$

$$\Rightarrow \ln\left(\frac{Y_i}{X_{1i}}\right) = \beta_0 + \beta_3 \ln\left(\frac{X_{2i}}{X_{1i}}\right) + u_i$$

$$F = \frac{(R^2_{UR} - R^2_R)/m}{(1 - R^2_{UR})/(n-k)}$$

$R^2_{UR} \geq R^2_R$ and $\sum \hat{u}^2_{UR} \leq \sum \hat{u}^2_R$

- **Significance of the regression coefficients**

$t = \frac{\hat{\alpha}_0 - \alpha_0}{SE(\hat{\alpha}_0)}$ $t = \frac{\hat{\alpha}_1 - \alpha_1}{SE(\hat{\alpha}_1)}$ $t = \frac{\hat{\alpha}_2 - \alpha_2}{SE(\hat{\alpha}_2)} \sim t_{(n-k)}$

$H_0: \alpha_0 = 0$

$t = \frac{\hat{\alpha}_0}{SE(\hat{\alpha}_0)}$

- **Overall Significance of the Sample Regression**

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$$

$$Y_i = \hat{\alpha}_0 + \hat{\alpha}_1 X_{1i} + \hat{\alpha}_2 X_{2i} + \hat{u}_i = \hat{y}_i + \hat{u}_i$$

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

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

$$\sum y_i^2 = \hat{\alpha}_1 \sum y_i x_{1i} + \hat{\alpha}_2 \sum y_i x_{2i} + \sum \hat{u}_i^2$$

$$F = \frac{(\hat{\alpha}_1 \sum y_i x_{1i} + \hat{\alpha}_2 \sum y_i x_{2i})/2}{\sum \hat{u}_i^2/n-3} = \frac{ESS/k-1}{RSS/n-k}$$

Source of Variation	SS	df	MSS
Regression (ESS)	$\hat{\alpha}_1 \sum y_i x_{1i} + \hat{\alpha}_2 \sum y_i x_{2i}$	$k-1$	$\frac{\hat{\alpha}_1 \sum y_i x_{1i} + \hat{\alpha}_2 \sum y_i x_{2i}}{k-1}$
Residual (RSS)	$\sum \hat{u}_i^2$	$n-k$	$\frac{\sum \hat{u}_i^2}{n-k} = \hat{\sigma}^2$
TSS	$\sum y_i^2$	$n-1$	

Decision Rule:

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \dots + \alpha_k X_{ki} + u_i$$

$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_k = 0$

H_1: Not all slope ~~co~~ coefficients are simultaneously zero.

$$F = \frac{ESS/df}{RSS/df} = \frac{ESS/k-1}{RSS/n-k}$$

If $F > F_\alpha(k-1, n-k) \rightarrow$ reject H_0.

$$F = \frac{n-k}{k-1}\frac{ESS}{TSS-ESS} = \frac{n-k}{k-1}\frac{ESS/TSS}{1-(ESS/TSS)} = \frac{n-k}{k-1}\frac{R^2}{1-R^2}$$

$$\boxed{F = \frac{R^2/k-1}{(1-R^2)/n-k.}}$$

Source of variation	SS	df	MSS
ESS	$R^2(\sum y_i^2)$	$k-1$	$R^2(\sum y_i^2)/k-1$
RSS	$(1-R^2)(\sum y_i^2)$	$n-k$	$(1-R^2)(\sum y_i^2)/n-k$
TSS	$\sum y_i^2$	$n-1$	

- Incremental or Marginal Contribution of an Explanatory variable

$$F = \frac{Q_2/df}{Q_4/df}$$

$$\boxed{F = \frac{(ESS_{new} - ESS_{old})/\text{number of new regressors}}{RSS_{new}/df}}$$

↓

(n - number of parameters in the new model)

Let

Old Model: $Y_i = \alpha_0 + \alpha_1 X_{1i} + u_i$

New Model: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + u_i$

Source of Variation	SS	df	MSS
ESS due to X_1 alone	$Q_1 = \hat{\alpha}_1^2 \sum x_{1i}^2$	1	$\frac{Q_1}{1}$
ESS due to the addition of X_2	$Q_2 = Q_3 - Q_1$	1	$\frac{Q_2}{1}$
ESS due to both X_1, X_2	$Q_3 = \hat{\alpha}_1 \sum x_{1i}^2 + \hat{\alpha}_2 \sum x_{2i}^2$	2	$\frac{Q_3}{2}$
RSS	$Q_4 = Q_5 - Q_3$	$n-3$	$\frac{Q_4}{n-3}$
TSS	$\sum y_i^2$		

- **Testing for Inclusion/Exclusion of Variables**

Case I:

Unrestricted Model: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + \alpha_5 X_{5i} + u_i$

Restriction: $\alpha_4 = \alpha_5 = 0$

$H_0: \alpha_4 = \alpha_5 = 0$

H_1: atleast α_4 or $\alpha_5 \neq 0$.

Restricted Model: $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i$

Restricted F Test:

$$\theta = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} \sim F_{m,(n-k)}$$

$m \rightarrow$ no. of restrictions

$k \rightarrow$ no. of coefficients

Rejection of the Null Hypothesis indicates that the unrestricted model should be ~~the~~ selected.

Case II :

Unrestricted Model :

$$Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + \alpha_5 X_{5i} + u_i$$

Restriction : $\alpha_5 = 0$

$H_0 : \alpha_5 = 0$

$H_1 : \alpha_5 \neq 0$

Restricted Model : $Y_i = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 X_{4i} + u_i$

Restricted F Test :

$$\theta = \frac{(RSS_R - RSS_{UR})/m}{RSS_{UR}/(n-k)} \sim F_{(m, n-k)}$$

$m \rightarrow$ no. of restrictions

$k \rightarrow$ no. of coefficients

Through t Test :

Test statistic : $\lambda = \dfrac{\hat{\alpha}_5}{SE(\hat{\alpha}_5)} \sim t_{(n-k)}$

Rejection of the Null Hypothesis in either case indicates that the unrestricted Model should be selected.

★ TESTS FOR STABILITY

Testing constancy of parameters

Two tests:

(i) Analysis of Variance Test (Chow Test)

(ii) Predictive Test

Two regression equations for two data sets

First set: $Y_t = \alpha_1 + \beta_1 X_t + u_{1t}$; $u_{1t} \sim IN(0, \sigma_1^2)$

Second set: $Y_t = \alpha_2 + \beta_2 X_t + u_{2t}$; $u_{2t} \sim IN(0, \sigma_2^2)$

$$H_0: \alpha_1 = \alpha_2 ; \beta_1 = \beta_2$$

If H_0 is not rejected - Stability of parameters

Restricted Model: Estimation of single equation for the entire dataset by pooling the two.

Test Statistic:

$$F = \frac{(RRSS - URSS)/k}{URSS/(n_1 + n_2 - 2k)} \sim F_{k,(n_1+n_2-2k)}$$

$k \rightarrow$ No. of parameters estimated (incl. intercept)

RRSS $\Rightarrow$ Restricted residual sum of squares

URSS $\Rightarrow$ Unrestricted residual sum of squares (Sum of RSS of all sets)

$$URSS = RSS_1 + RSS_2$$

RSS_1 = RSS for first data set

RSS_2 = RSS for second data set

$$\frac{RSS_1}{\sigma^2} \sim \chi^2_{(n_1-k)} \quad ; \quad \frac{RSS_2}{\sigma^2} \sim \chi^2_{(n_2-k)}$$

$$\frac{USSR}{\sigma^2} \sim \chi^2_{(n_1+n_2-2k)}$$

RRSS $\rightarrow$ For regression with pooled data $\frac{RRSS}{\sigma^2} \sim \chi^2_{(n_1+n_2-k)}$

Assumptions of Chow Test

1. $u_{1t} \sim N(0, \sigma^2)$
   $u_{2t} \sim N(0, \sigma^2)$ } Same (homoscedastic) variance

$H_0: \sigma_1^2 = \sigma_2^2$

Test for Equality of Variance:

$$\hat{\sigma}_1^2 = \frac{RSS_1}{n_1 - k} \qquad \hat{\sigma}_2^2 = \frac{RSS_2}{n_2 - k}$$

Test statistic: $\frac{\hat{\sigma}_1^2}{\hat{\sigma}_2^2} \sim F_{(n_1-k),(n_2-k)}$

By convention, we put the larger of the two estimated variances in the numerator.

Limitations:

1) Chow Test can only be carried out if σ_1^2 & σ_2^2 are not statistically different.

2) The Chow test will tell us only if the two regressions are different without telling whether the difference is in slope or intercept.

3) Sensitive to the choice of structural breaks: It is not possible to determine when the structural change actually took place.

~~DUMMY VARIABLE REGRESSION MODEL~~

The Predictive Test

Can be used when n_1 or n_2 is less than k.

Test statistic: $$F = \frac{(RRSS - RSS_1)/n_2}{RSS_1/n_1 - k} \sim F_{n_2, n_1 - k}$$

If computed value of F-stat is greater than the critical value for the given degrees of freedom and at the chosen

level of significance, the null hypothesis is rejected. Indicates structural change in the relationship.

DUMMY VARIABLE REGRESSION MODELS

If we use m dummy variables for m categories:

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + \beta_3 D_{3i} + u_i$$

→ Perfect collinearity

If a qualitative variable has m categories, introduce only (m-1) dummy variables.

If we do not follow this rule, we would fall into what is called the dummy variable trap, i.e., the situation of perfect collinearity or perfect multicollinearity if there is more than one exact relationship among the variables.

For each qualitative ~~variable~~ regressor, the number of dummy variables introduced must be one less than the categories of that variable.

The category for which no dummy variable is assigned is known as the base, benchmark and all comparisons are made in relation to the benchmark category.

The intercept value represents the mean value of the benchmark category.

The coefficients attached to the dummy variables are known as the differential intercept coefficients because they tell by how much the value of the category that ~~recieves~~ receives the value 1 differs from the intercept coefficient of the benchmark category.

If the no. of dummy variables introduced = no. of categories of that variable, then intercept must be dropped.

Analysis of Variance (ANOVA)

Regression model with only dummy or qualitative variables. Compares the mean of two or more categories.

Analysis of Covariance (ANCOVA)

Regression model with mix of qualitative and ~~quanta~~ quantitative variables. Examines the main and interaction effects of categorical variables on a continuous dependent variable, controlling the effects of other continuous variables. In addition to examining impact of qualitative aspects or attributes, dummy (independent) variables are also used for seasonality analysis and examining structural breaks/differences.

The Dummy Variable Alternative to the Chow Test

Model: $Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 D_i X_{2i} + u_i$

Y_i = Wage Rate ; X_2 = Productivity

D_i = 1 for male, 0 for female – Dummy variable.

For male : $E(Y_i | X_i) = \alpha_1 + (\alpha_2 + \alpha_3) X_{2i} + u_i$

For female : $E(Y_i | X_i) = \alpha_1 + \alpha_2 X_{2i} + u_i$

$\alpha_3 \rightarrow$ change in mean wage rate between male and female with change in X_2.

If α_3 is statistically significant, there is a difference in change in the mean wage rate between male and female with change in X_2.

Coincident Regressions

Two regression lines with same intercept and slope.

Parallel Regressions

Two regression lines with different intercept but same slope.

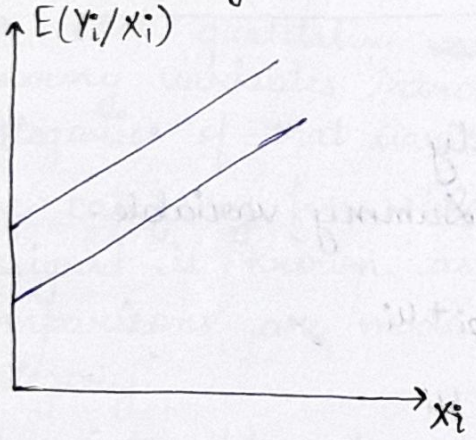
Concurrent Regressions

Two regression lines with different slope but same intercept.

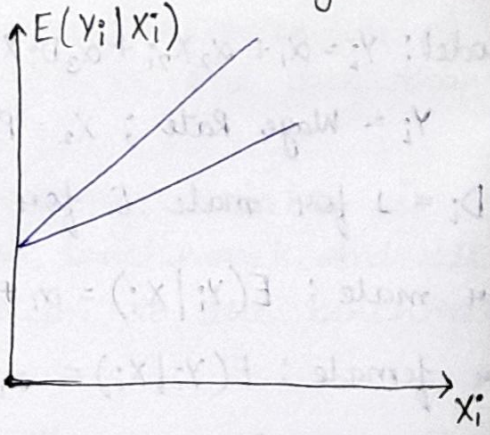
Dissimilar Regressions

Two regression lines with different intercept as well as slope.

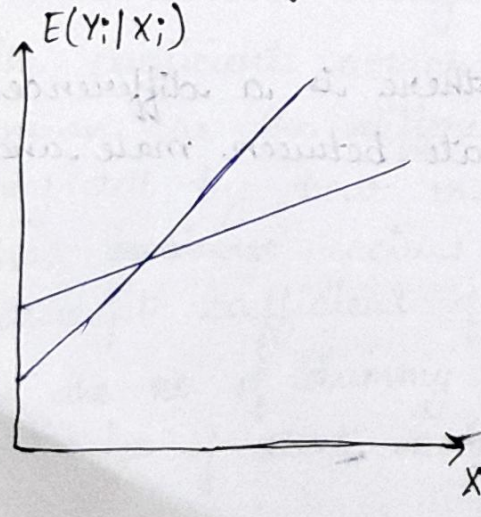
Parallel Regression



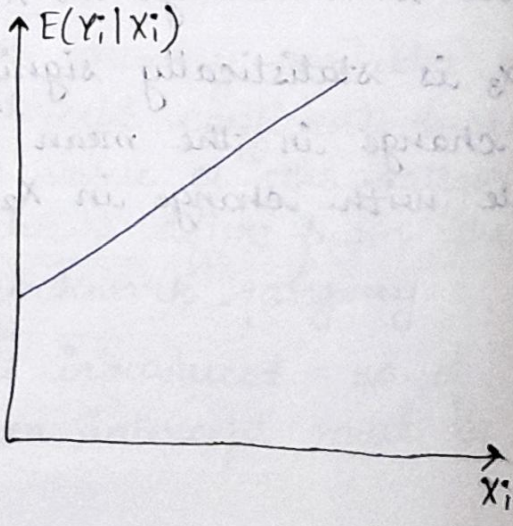
Concurrent Regression



Dissimilar Regression



~~Con~~ Coincident Regression



Advantages of using Dummy variable

1. We need to run only a single regression.
2. The single regression can be used to test a variety hypothesis like whether the change is due to slope intercept.
3. Since pooling (including all the observations in a regression) increases the degrees of freedom, it may improve the relative precision of the estimated parame

Every addition of a dummy variable will consume one degree of freedom.

ANOVA Examples

Example 1:

To examine if average monthly per capita consumption expenditure (MPCE) varies depending on whether the households belong to rural, urban and semi-urban areas, i.e.,

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 D_{2i} + u_i$$

Y_i = Monthly per capita consumption expenditure.

$D_{1i} = 1$ (Urban) $D_{2i} = 1$ (semi-urban)

$= 0$ (ow) $= 0$ ow.

Base category = Rural area.

1) $D_{1i} = 0, D_{2i} = 0$; $E(Y_i) = \alpha$

2) $D_{1i} = 1, D_{2i} = 0$; $E(Y_i) = \alpha + \beta_1$

3) $D_{1i} = 0, D_{2i} = 1$; $E(Y_i) = \alpha + \beta_2$

Comparison between rural and urban households.

β_1 → statistically not significant → average MPCE of urban households is not significantly different from that of rural households.

β_1 → statistically significant & +ve/(-ve)
average MPCE of urban households is significantly higher/(lower) than that of rural households

Comparison between rural and semi-urban households

β_2 → statistically not significant
average MPCE of semi-urban households is not statistically significantly different from that of rural households.

β_2 → statistically significant & +ve/(-ve)
average MPCE of semi-urban households is significantly higher/(lower) than that of rural households.

Comparison between urban and semi-urban households

β_1 not statistically significantly different from β_2
→ average MPCE of urban households is not significantly different from that of semi-urban households

$\beta_1 > \beta_2$ → average MPCE of urban households is significantly higher than that of semi-urban households.

$\beta_1 < \beta_2$ → average MPCE of urban households is significantly lower than that of semi-urban households.

ANCOVA Example

To examine if monthly per capita consumption expenditure varies depending on income and whether the households belong to rural and urban areas, i.e.,

$$Y_i = \alpha + \beta_1 D_{1i} + \beta_2 X_i + \beta_3 (D_{1i} * X_{2i}) + u_i$$

Y_i = Monthly per capita consumption expenditure

X_i = Monthly income of the household

D_{2i} = 1 (urban) Base category: Rural area
= 0 (otw)

Given X_i, $D_{2i} = 0$, $E(Y_i) = \alpha + \beta_2 X_i$

Given X_i, $D_{2i} = 1$, $E(Y_i) = (\alpha + \beta_1) + (\beta_2 + \beta_3) X_i$

1) If both β_1 & β_3 are ~~stat~~ not significant, the two PRFs will coincide.

2) only β_1 significant: The two PRFs will be parallel (difference will be only in respect of intercept - autonomous consumption)

$\beta_1 \to +ve$, Urban households will have higher/(lower) ($-ve$) intercept.

3) Only β_3 is significant: The two PRFs will be concurrent (difference will only be in respect of slope-induced consumption)

$\beta_3 \to +ve / (-ve)$

PRF for urban households will be steeper/(flatter)

1) Both $\beta_1, \beta_3 \to$ significant: The two PRFs will be dissimilar

$\beta_1 \to +ve$, $\beta_3 \to +ve$: PRF for urban households will be steeper with a higher intercept.

$\beta_1 \to +ve$, $\beta_3 \to$ ~~+ve~~ $-ve$: flatter with higher intercept

$\beta_1 \to -ve$, $\beta_3 \to +ve$: Steeper with lower intercept.

$\beta_1 \to -ve$, $\beta_3 \to -ve$: flatter with lower intercept.

J-TEST

Applied to compare 2 different models with exactly the same dependent variable and atleast one different explanatory variable.

Helps in examining if any of the models is superior to the other.

Model I: $Y_i = \alpha + \beta X_{1i} + u_{1i}$

Model II: $Y_i = \alpha + \beta X_{2i} + u_i$

Steps:

Estimate Model-I and obtain fitted value of Y

Use fitted value of Y as an explanatory variable in estimation of Model-II.

Examine the t-statistic for the fitted value.

Repeat the steps for Model-I.

		Fitted Y of Model-I in Model-II	
		$P<0.10$	$P>0.10$
Fitted Y of Model-II in Model-I	$P<0.10$	Combined Model	Model-II
	$P>0.10$	Model-I	Either Model.

TESTING FOR VALIDITY OF ASSUMPTIONS

1) Normality assumption of u_i

→ Jarque-Bera test of normality.

2) Fixed vs Stochastic Regressors.

If X's are random or stochastic

Relaxing the assumption that X is nonstochastic and replacing ~~with~~ it by the assumption that X is stochastic but independent of [u] does not change the ~~desi~~ desirable properties and ~~feasibile~~ feasibility of least squares estimation.

3) Zero mean value of u_i

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki} + u_i$$

$$E(u_i | X_{2i}, X_{3i}, \ldots X_{ki}) = w$$

$$E(Y_i | X_{2i}, X_{3i}, \ldots, X_{ki}) = (\beta_1 + w) + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}$$

$$= \alpha + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots + \beta_k X_{ki}$$

If $E(u_i) = w_i$, then both intercept & slope coefficients are different.

If the disturbances $[u_i]$ are independently and identically distributed with zero mean and [constant] variance σ^2 and if the explanatory variables are constant in repeated samples, the OLS coefficient estimators are asymptotically normally distributed with means equal to the corresponding β's.

Therefore, the usual test procedures - the t and F tests - are still valid asymptotically, that is, in the large sample, but not in the finite or small samples.

For testing the validity of assumptions :-

(i) Identify the nature of the problem.

(ii) Examine its consequences

(iii) Suggest methods of detecting it

(iv) Consider remedial measures so that they may lead to estimators that possess the desirable statistical properties.