Regression Diagnostics and Model Selection

Specification Error	Broad Consequences
Omission of Important Variables Example: True Model: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$ Estimated Model: $Y_i = \gamma + \delta X_{1i} + v_i$	Biased and Inconsistent Estimators; Incorrect Variances and Standard Errors – Misleading Conclusions on Testing of Hypotheses
Inclusion of Irrelevant Variables Example: True Model: $Y_i = \gamma + \delta X_{1i} + v_i$ Estimated Model: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$	Unbiased, Consistent but Inefficient Estimators – Wider Confidence Intervals
Adoption of Wrong Functional Form Example: True Model: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{1i}^2 + u_i$ Estimated Model: $Y_i = \gamma + \delta X_{1i} + v_i$	Specification errors – Bias in inferences on the causal relationships
Incorrect Specification of the Random Disturbance Term Example: True Model: $Y_i = \alpha X_{1i}^{\beta} u_i$ Estimated Model: $Y_i = \gamma + \delta X_{1i} + v_i$	Changes in distribution of the random disturbance term and hence its properties.
Errors in Measurement of the Variables True Model: $Y_i = \gamma + \delta X_{1i} + v_i$ Estimated Model: $Y_i^* = \alpha + \beta X_{1i}^* + u_i$ Here, $Y_i^* = Y_i + \varepsilon_i$ and $X_i^* = X_i + \eta_i$	In Case of the Dependent Variables: Unbiased, Consistent but Inefficient Estimators In Case of the Independent Variables: Biased and Inconsistent Estimators

(1) Omission of Important Variables:

True Model: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$; Estimated Model: $Y_i = \gamma + \delta X_{1i} + v_i$

Hence, the OLS estimator of the slope coefficient, $\hat{\delta} = \dfrac{\sum x_{1i} y_i}{\sum x_{1i}^2}$ (1)

Now, for the true model, $\overline{Y} = \alpha + \beta_1 \overline{X}_1 + \beta_2 \overline{X}_2 + \overline{u}$ (2)

Subtracting (2) from the true model we get, $y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + u_i - \overline{u}$ (3)

Substituting (3) in (1) we get,

$$\hat{\delta} = \frac{\sum x_{1i}(\beta_1 x_{1i} + \beta_2 x_{2i} + u_i - \overline{u})}{\sum x_{1i}^2} = \beta_1 + \beta_2 \frac{\sum x_{1i} x_{2i}}{\sum x_{1i}^2} + \frac{\sum x_{1i} u_i}{\sum x_{1i}^2} - \frac{\overline{u} \sum x_{1i}}{\sum x_{1i}^2}$$

Or, $\hat{\delta} = \beta_1 + \beta_2 \hat{\theta}_{21} + \dfrac{\sum x_{1i} u_i}{\sum x_{1i}^2}$ [as $\sum x_{1i} = 0$]

Here, $\hat{\theta}_{21}$ is the OLS estimator of the slope coefficient while regressing X_{2i} on X_{1i}

Hence, $E(\hat{\delta}) = \beta_1 + \beta_2\theta_{21}$ [as $E(u_i) = 0$ by assumption]

Thus, $\hat{\delta}$ is a biased estimator of β_1 with the extent of bias being $\beta_2\theta_{21}$. This means that exclusion of important variables results in biased estimators of the coefficients.

Further, $var(\hat{\delta}) = \frac{\sigma_v^2}{\sum x_{1i}^2}$ and $var(\tilde{\beta}_1) = \frac{\sigma_u^2}{\sum x_{1i}^2(1-r_{12}^2)}$

If $\sigma_v^2 = \sigma_u^2$, $\text{var}(\hat{\delta}) < \text{var}(\hat{\beta}_1)$

However, since both σ_u^2 and σ_v^2 are unknown, their OLS estimators are used for the empirical purpose.

Now, $\tilde{\sigma}_u^2 = \frac{\sum \tilde{u}_i^2}{n-3}$ and $\hat{\sigma}_v^2 = \frac{\sum \hat{v}_i^2}{n-2}$.

Hence, $var(\hat{\delta}) = \frac{\frac{\sum \hat{v}_i^2}{(n-2)}}{\sum x_{1i}^2} = \frac{\sum \hat{v}_i^2}{(n-2)\sum x_{1i}^2}$ and

$$var(\tilde{\beta}_1) = \frac{\frac{\sum \tilde{u}_i^2}{(n-3)}}{\sum x_{1i}^2(1-r_{12})^2} = \frac{\sum \tilde{u}_i^2}{(n-3)\sum (1-r_{12})^2}$$

If $\sum \tilde{u}_i^2 < \sum \hat{v}_i^2$ (which is likely with exclusion of important variables), comparison between $\text{var}(\hat{\delta})$ and $var(\hat{\beta}_1)$ becomes an empirical issue. Nonetheless, exclusion of important variables is likely to alter the estimated variance of the regression coefficients and hence the respective confidence intervals, leading to misleading conclusions on testing of hypotheses.

(2) Inclusion of Irrelevant Variables:

True Model: $Y_i = \gamma + \delta X_{1i} + v_i$ Estimated Model: $Y_i = \alpha + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$

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

Now, for the true model we have, $\overline{Y} = \gamma + \delta \overline{X}_1 + \overline{v}$ (2)

Subtracting (2) from the true model we get, $y_i = \delta x_{1i} + v_i - \overline{v}$ (3)

Substituting (3) in (1) we get,

Hence, $\hat{\beta}_1 = \frac{\sum x_{1i}(\delta x_{1i} + v_i - \overline{v}) \sum x_{2i}^2 - \sum x_{2i}(\delta x_{1i} + v_i - \overline{v}) \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i}\right)^2}$

Or, $\hat{\beta}_1 = \dfrac{\delta \sum x_{1i}^2 \sum x_{2i}^2 + \sum x_{1i} v_i \sum x_{2i}^2 - \delta \left(\sum x_{1i} x_{2i} \right)^2 - \sum x_{2i} v_i \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i} \right)^2}$

[As $\sum x_{1i} = \sum x_{2i} = 0$]

Or, $\hat{\beta}_1 = \dfrac{\delta \left\{ \sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i} \right)^2 \right\} + \left\{ \sum x_{1i} v_i \sum x_{2i}^2 - \sum x_{2i} v_i \sum x_{1i} x_{2i} \right\}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i} \right)^2}$

Or, $\hat{\beta}_1 = \delta + \dfrac{\sum x_{1i} v_i \sum x_{2i}^2 - \sum x_{2i} v_i \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i} \right)^2}$

Hence, $E(\hat{\beta}_1) = \delta$ [As $E(v_i) = 0$ by assumption]

Thus, $\hat{\beta}_1$ is an unbiased estimator of δ.

On the other hand, $\hat{\beta}_2 = \dfrac{\sum x_{2i} y_i \sum x_{1i}^2 - \sum x_{1i} y_i \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i} \right)^2}$ (4)

Substituting (3) in (4) we get,

$\hat{\beta}_2 = \dfrac{\sum x_{2i} (\delta x_{1i} + v_i - \bar{v}) \sum x_{1i}^2 - \sum x_{1i} (\delta x_{1i} + v_i - \bar{v}) \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i} \right)^2}$

Or, $\hat{\beta}_2 = \dfrac{\delta \sum x_{1i} x_{2i} \sum x_{1i}^2 + \sum x_{2i} v_i \sum x_{1i}^2 - \delta \sum x_{1i}^2 \sum x_{1i} x_{2i} - \sum x_{1i} v_i \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i} \right)^2}$

[As $\sum x_{1i} = \sum x_{2i} = 0$]

Or, $\hat{\beta}_2 = \dfrac{\sum x_{2i} v_i \sum x_{1i}^2 - \sum x_{1i} v_i \sum x_{1i} x_{2i}}{\sum x_{1i}^2 \sum x_{2i}^2 - \left(\sum x_{1i} x_{2i} \right)^2}$

Hence, $E(\hat{\beta}_2) = 0$ [as $E(v_i) = 0$ by assumption]

Thus, $\hat{\beta}_2$ is also an unbiased estimator [as β_2 does not exist in the true model and $E(\hat{\beta}_2) = 0$].

This means that the OLS estimators of the regression coefficients continue to be unbiased despite inclusion of irrelevant variables.

Now, $var(\tilde{\delta}) = \dfrac{\sigma_v^2}{\sum x_{1i}^2}$ and $var(\hat{\beta}_1) = \dfrac{\sigma_u^2}{\sum x_{1i}^2 (1 - r_{12}^2)}$

If $\sigma_v^2 = \sigma_u^2$, $\text{var}(\hat{\delta}) < \text{var}(\hat{\beta}_1)$ and this makes $\hat{\beta}_1$ and inefficient estimators of β_1

However, since σ_u^2 and σ_v^2 are unknown, their OLS estimators are used for empirical purpose.

Now, $\hat{\sigma}_u^2 = \frac{\sum \hat{u}_i^2}{n-3}$ and $\tilde{\sigma}_v^2 = \frac{\sum \tilde{v}_i^2}{n-2}$.

Hence, $var(\tilde{\delta}) = \frac{\frac{\sum \tilde{v}_i^2}{(n-2)}}{\sum x_{1i}^2} = \frac{\sum \tilde{v}_i^2}{(n-2)\sum x_{1i}^2}$ and

$$var(\hat{\beta}_1) = \frac{\frac{\sum \hat{u}_i^2}{(n-3)}}{\sum x_{1i}^2(1-r_{12})^2} = \frac{\sum \hat{u}_i^2}{(n-3)\sum (1-r_{12})^2}$$

If $\sum \hat{u}_i^2$ is not different from $\sum \tilde{v}_i^2$ (which is very likely), $var(\tilde{\sigma}) < var(\hat{\beta}_1)$

Thus, although comparison between $\text{var}(\hat{\delta})$ and $var(\hat{\beta}_1)$ is an empirical issue, inclusion of irrelevant variables is likely to increase the estimated variance of the regression coefficients and make these coefficients inefficient. Further, inclusion of irrelevant variables will also result in loss of degrees of freedom.

Review Question: What will be the consequences if the true model is $Y_i = \delta X_i + v_i$, but the model $Y_i = \gamma + \lambda X_i + w_i$ is estimated? How will the consequence change in a reverse situation?

Answer:

(a) True Model: $Y_i = \delta X_i + v_i$ Estimated Model: $Y_i = \gamma + \lambda X_i + w_i$

Here, $\hat{\lambda} = \frac{\sum x_i y_i}{\sum x_i^2}$ (1)

For the true model, $\overline{Y} = \delta \overline{X}_1 + \overline{v}$ (2)

Subtracting (2) from (1) we get, $y_i = \delta x_i + (v_i - \overline{v})$ (3)

Substituting (3) in (1),

$$\hat{\lambda} = \frac{\sum x_i\{\delta x_i + (v_i - \overline{v})\}}{\sum x_i^2} = \delta + \frac{\sum x_i(v_i - \overline{v})}{\sum x_i^2} = \delta + \frac{\sum x_i v_i - \overline{v}\sum x_i}{\sum x_i^2} = \delta + \frac{\sum x_i v_i}{\sum x_i^2} \text{ [as } \sum x_i = 0 \text{]}$$

Hence, $E(\hat{\lambda}) = \delta$ (as $E(v_i) = 0$ by assumption)

This indicates that $\hat{\lambda}$ is an unbiased estimator of δ.

Further, $var(\hat{\lambda}) = \frac{\sigma_w^2}{\sum (X_i - \overline{X})^2}$ and $var(\tilde{\delta}) = \frac{\sigma_v^2}{\sum X_i^2}$

If $\sigma_w^2 = \sigma_v^2$, $\text{var}(\hat{\lambda}) > \text{var}(\hat{\delta})$. Thus, the variance of the estimated slope coefficient may increase with inclusion of the intercept in the estimated model.

(b) True Model: $Y_i = \gamma + \lambda X_{1i} + w_i$ Estimated Model: $Y_i = \delta X_{1i} + v_i$

Here, $\hat{\delta} = \dfrac{\sum X_i Y_i}{\sum X_i^2}$ (1)

Substituting the true model in (1) we get

$$\hat{\delta} = \frac{\sum X_i(\gamma + \lambda X_i + u_i)}{\sum X_i^2} = \frac{\gamma \sum X_i + \lambda \sum X_i^2 + \sum X_i u_i}{\sum X_i^2} = \lambda + \frac{\gamma \sum X_i}{\sum X_i^2} + \frac{\sum X_i u_i}{\sum X_i^2}$$

Hence, $E(\hat{\delta}) = \lambda + \dfrac{\gamma \sum X_i}{\sum X_i^2} \neq \lambda$ [as $E(u_i) = 0$ by assumption]

Further, $var(\hat{\delta}) = \dfrac{\sigma_v^2}{\sum X_i^2}$ and $var(\tilde{\lambda}) = \dfrac{\sigma_w^2}{\sum (X_i - \bar{X})^2}$

If $\sigma_w^2 = \sigma_v^2$, $\text{var}(\hat{\lambda}) < \text{var}(\hat{\delta})$ indicating that variance of the estimated slope coefficient may decrease with exclusion of the intercept from the estimated model.

(3) Errors in Variables:

True Model: $Y_i = \alpha + \beta X_i + u_i$

Estimated Model: $Y_i^* = \alpha + \beta X_i^* + v_i$ with $Y_i^* = Y_i + \varepsilon_i$ and $X_i^* = X_i + \eta_i$

Assumptions: $E(\varepsilon_i) = E(\eta_i) = 0$; $var(\varepsilon_i) = \sigma_\varepsilon^2$; $var(\eta_i) = \sigma_\eta^2$;
$cov(\varepsilon_i, \eta_i) = cov(\varepsilon_i, X_i) = cov(\varepsilon_i, Y_i) = cov(\eta_i, X_i) = cov(\eta_i, Y_i) = 0$

Hence, the model can be rewritten as,

$Y_i^* - \varepsilon_i = \alpha + \beta(X_i^* - \eta_i) + u_i$ Or, $Y_i^* = \alpha + \beta X_i^* + v_i$ with $v_i = \varepsilon_i + u_i - \beta \eta_i$ (1)

Now, $cov(v_i, X_i^*) = cov(\varepsilon_i + u_i - \beta \eta_i, X_i + \eta_i) = E((\varepsilon_i + u_i - \beta \eta_i)(X_i + \eta_i - X_i)) = -\beta \sigma_\eta^2 \neq 0$

Thus, there is endogeneity problem and the basic assumption of the method of OLS is violated.

However, if there is measurement error only in case of Y ,i.e., if $\eta_i = 0$,
$\text{cov}(v_i, X_i^*) = \text{cov}(v_i, X_i) = 0$

In such a case, the model can be estimated by applying the method of OLS.

But, $var(\hat{\beta}) = \dfrac{\sigma_v^2}{\sum x_i^2} = \dfrac{\sigma_\varepsilon^2 + \sigma_u^2}{\sum x_i^2} > \dfrac{\sigma_u^2}{\sum x_i^2}$

Thus, the random disturbance term has larger variance now and this brings in inefficiency.

If model (1) is estimated by applying the method of OLS,

$$\hat{\beta} = \frac{\sum x_i^* y_i^*}{\sum x_i^{*2}} = \frac{\sum \{x_i + (\eta_i - \overline{\eta})\}\{\beta x_i + (u_i - \overline{u}) + (\varepsilon_i - \overline{\varepsilon})\}}{\sum \{x_i + (\eta_i - \overline{\eta})\}^2}$$

$$= \frac{\beta \sum x_i^2 + \beta \sum x_i(\eta_i - \overline{\eta}) + \sum x_i(u_i - \overline{u}) + \sum (\eta_i - \overline{\eta})(u_i - \overline{u}) \sum x_i(\varepsilon_i - \overline{\varepsilon}) + \sum (\eta_i - \overline{\eta})(\varepsilon_i - \overline{\varepsilon})}{\sum x_i^2 + 2\sum x_i(\eta_i - \overline{\eta}) + \sum (\eta_i - \overline{\eta})^2}$$

Dividing both the numerator and the denominator by n,

$$\hat{\beta} = \frac{\left\{\beta \sum x_i^2 + \beta \sum x_i(\eta_i - \overline{\eta}) + \sum x_i(u_i - \overline{u}) + \sum (\eta_i - \overline{\eta})(u_i - \overline{u}) \sum x_i(\varepsilon_i - \overline{\varepsilon}) + \sum (\eta_i - \overline{\eta})(\varepsilon_i - \overline{\varepsilon})\right\}\left(\frac{1}{n}\right)}{\left\{\sum x_i^2 + 2\sum x_i(\eta_i - \overline{\eta}) + \sum (\eta_i - \overline{\eta})^2\right\}\left(\frac{1}{n}\right)}$$

Hence,

$$p\lim \hat{\beta} = \frac{p\lim\left\{\beta \sum x_i^2 + \beta \sum x_i(\eta_i - \overline{\eta}) + \sum x_i(u_i - \overline{u}) + \sum (\eta_i - \overline{\eta})(u_i - \overline{u}) \sum x_i(\varepsilon_i - \overline{\varepsilon}) + \sum (\eta_i - \overline{\eta})(\varepsilon_i - \overline{\varepsilon})\right\}\left(\frac{1}{n}\right)}{p\lim\left\{\sum x_i^2 + 2\sum x_i(\eta_i - \overline{\eta}) + \sum (\eta_i - \overline{\eta})^2\right\}\left(\frac{1}{n}\right)}$$

Or, $p\lim \hat{\beta} = \frac{\beta \sigma_X^2}{\sigma_X^2 + \sigma_\eta^2} = \beta\left(\frac{1}{1 + \frac{\sigma_\eta^2}{\sigma_X^2}}\right) \neq \beta$

Thus, $\hat{\beta}$ is a biased and inconsistent estimator of β. This means that errors in measurement of the independent variables(s) result in biased and inconsistent estimators of the regression coefficients.

(4) Choice of Functional Form:

Functional Form	Mathematical Expression	Slope	Elasticity
Linear	$y = a + bx$	b	$b\frac{x}{y}$
Quadratic	$y = a + bx + cx^2$	$b + 2cx$	$(b + 2cx)\frac{x}{y} = \frac{bx + 2cx^2}{y}$
Reciprocal	$y = a + \frac{b}{x}$	$-\frac{b}{x^2}$	$-\left(\frac{b}{x^2}\right)\left(\frac{x}{y}\right) = \frac{b}{xy}$
Semi-Logarithmic	$ln(y) = a + bx$ or $y = e^{a+bx}$	by	bx
	$y = a + b\,ln(x)$ or $e^y = cx^b$	$\frac{b}{x}$	$\frac{b}{y}$
Log-log	$ln(y) = a + b\,ln(x)$	$\frac{by}{x}$	b
Interaction	$y = a + bx + cxz + dz$	$b + cz$	$(b + 2cz)\frac{x}{y} = \frac{bx + 2cxz}{y}$

Logistic	$ln\left(\frac{y}{1-y}\right)=a+bx$	$by(1-y)$	$bx(1-y)$

Note: Derivation for the Logistic Function

$y=\frac{e^{a+bx}}{1+e^{a+bx}}$ or $1-y=1-\frac{e^{a+bx}}{1+e^{a+bx}}=\frac{1}{1+e^{a+bx}}$

Hence, $\frac{y}{1-y}=e^{a+bx}$ or $ln\left(\frac{y}{1-y}\right)=a+bx$

Hence, $\left(\frac{1-y}{y}\right)\left(\frac{(1-y)\frac{dy}{dx}+y\frac{dy}{dx}}{(1-y)^2}\right)=b$ or, $\left(\frac{1}{y(1-y)}\right)\frac{dy}{dx}=b$ or, $\frac{dy}{dx}=by(1-y)$ (slope)

Further, $\frac{dy}{dx}\frac{x}{y}=by(1-y)\frac{x}{y}=bx(1-y)$ (elasticity)

(5) Some Important Statistical Tests for Model Selection:

(a) Special Wald Test for Goodness-of-Fit

Model I: $Y_i=\alpha_0+\sum_{j=1}^{p}\alpha_j X_i+u_i$ (Unrestricted Model)

Model II: $Y_i=\beta_0+v_i$ (Restricted Model with p restrictions, i.e., $\alpha_1=\alpha_2=..=\alpha_p=0$)

Null Hypothesis: $\alpha_1=\alpha_2=..=\alpha_p=0$

For the restricted model, $TSS=RSS_R=\sum(Y_i-\overline{Y})^2$ (as $\hat{Y}_i=\hat{\beta}_0=\overline{Y}$)

Hence, the test statistic $F=\frac{(RSS_R-RSS_{UR})/p}{RSS_{UR}/(n-p-1)}=\left(\frac{TSS-RSS_{UR}}{RSS_{UR}}\right)\left(\frac{p}{n-p-1}\right)=F_{UR}$

Rejection of the null hypothesis suggests that the unrestricted model is valid.

(b) Restricted F Test for Linearity

Linear Function: $Y_i=\alpha+\beta X_i+u_i$ versus Cubic Function: $Y_i=\theta_0+\theta_1 X_i+\theta_2 X_i^2+\theta_3 X_i^3+v_i$

Null Hypothesis: $\theta_2=\theta_3=0$

Test Statistic: $F=\frac{(RSS_L-RSS_C)/(C-L)}{RSS_C/(n-C)}=\left(\frac{R_C^2-R_L^2}{1-R_C^2}\right)\left(\frac{n-C}{C-L}\right)\sim F_{[(C-L),(n-C)]}$

Rejection of the null hypothesis suggests that the linear specification is not valid.

(c) Lagrange Multiplier Test for Model Selection

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

Unrestricted Model:
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_3 X_{mi} + \beta_{m+1} X_{(m+1)i} + \beta_{m+2} X_{(m+2)i} + .. + \beta_k X_{ki} + v_i$

Null Hypothesis: $\beta_{m+1} = \beta_{m+2} = .. = \beta_k = 0$

Steps to be Followed:

1. Estimate the restricted model
2. Obtain the estimated residual as $\hat{u}_i = Y_i - \hat{\alpha}_0 - \hat{\alpha}_1 X_{1i} - \hat{\alpha}_2 X_{2i} - .... - \hat{\alpha}_3 X_{mi}$
3. Estimate the model:
   $\hat{u}_i = \theta_0 + \theta_1 X_{1i} + \theta_2 X_{2i} + ... + \theta_3 X_{mi} + \gamma_{m+1} X_{(m+1)i} + \gamma_{m+1} X_{(m+2)i} + .. + \gamma_k X_{ki} + w_i$
4. Obtain R^2 from this auxiliary regression
5. Test Statistic $\theta = nR^2 \sim \chi^2_{k-m}$ (for large sample)
6. Rejection of the null hypothesis indicates that the additional variables should be added into the model

(d) Likelihood Ratio Test for Model Selection

For the model $Y_i = \alpha + \beta X_i + u_i$ with $u_i \sim IIN(0, \sigma^2)$, we have the likelihood function

$$L(\alpha, \beta, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^n e^{-\frac{\sum (Y_i - \alpha - \beta X_i)^2}{2\sigma^2}}$$

Taking natural logarithm in both the sides,

$$ln(L) = -n\,ln(\sigma) - n\,ln(\sqrt{2\pi}) - \frac{\sum (Y_i - \alpha - \beta X_i)^2}{2\sigma^2}$$

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

Unrestricted Model:
$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_3 X_{mi} + \beta_{m+1} X_{(m+1)i} + \beta_{m+2} X_{(m+2)i} + .. + \beta_k X_{ki} + v_i$

Null Hypothesis: $\beta_{m+1} = \beta_{m+2} = .. = \beta_k = 0$

Now, for the unrestricted model the likelihood value is

$$\hat{L} = \left(\frac{1}{\hat{\sigma}\sqrt{2\pi}} \right)^n e^{-\frac{\sum \hat{u}_i^2}{2\hat{\sigma}^2}} = \left(\frac{1}{\hat{\sigma}\sqrt{2\pi}} \right)^n e^{-\left(\frac{\sum \hat{u}_i^2}{2\left(\sum \hat{u}_i^2 / n\right)} \right)} = \left(\frac{1}{\hat{\sigma}\sqrt{2\pi}} \right)^n e^{-\left(\frac{n}{2}\right)}$$

Similarly, for the restricted model we have,

$$\tilde{L} = \left(\frac{1}{\tilde{\sigma}\sqrt{2\pi}}\right)^n e^{-\left(\frac{\sum \tilde{u}_i^2}{2\tilde{\sigma}^2}\right)} = \left(\frac{1}{\tilde{\sigma}\sqrt{2\pi}}\right)^n e^{-\left(\frac{\sum \tilde{u}_i^2}{2\sum \tilde{u}_i^2/n}\right)} = \left(\frac{1}{\tilde{\sigma}\sqrt{2\pi}}\right)^n e^{-\left(\frac{n}{2}\right)}$$

Hence, the likelihood ratio (LR), $\lambda = \frac{\tilde{L}}{\hat{L}} = \left(\frac{\hat{\sigma}}{\tilde{\sigma}}\right)^n = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right)^{\frac{n}{2}}$

The LR test statistic is, $LR = -2\,ln(\lambda) = -n \times ln\left(\frac{\hat{\sigma}^2}{\tilde{\sigma}^2}\right) \sim \chi^2_{k-m}$

(e) Nested versus Non-Nested Models: Davidson-Mackinnon J Test for Model Selection

Case I

Model I: $Y_i = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_i + u_i$; Model II: $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i + \sum_{j=1}^{q} \gamma_j Z_i + v_i$

Model I (Restricted) is nested in Model II (Unrestricted) and hence the Restricted F Test, the Likelihood Ratio Test and the Lagrange Multiplier Test can be applied for model selection

Case II

Model I: $Y_i = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_i + \sum_{j=1}^{r} \delta_j M_i + u_i$; Model II: $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i + \sum_{j=1}^{q} \gamma_j Z_i + v_i$

Model I is not nested in Model II and hence Restricted F Test, Likelihood Ratio Test and Lagrange Multiplier Test cannot be applied for model selection. In such cases, the Davidson-Mackinnon J Test can be applied for selection of the appropriate model.

Steps to be Followed:

1. Estimate Model I and compute the fitted values of the dependent variable $\hat{Y}_i^I$
2. Estimate the model $Y_i = \beta_0 + \sum_{j=1}^{p} \beta_j X_i + \sum_{j=1}^{q} \gamma_j Z_i + \lambda \hat{Y}_i^I + v_i$ and test the statistical significance of the coefficient λ
3. Estimate Model II and compute the fitted values of the dependent variable $\hat{Y}_i^{II}$
4. Estimate the model $Y_i = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_i + \sum_{j=1}^{q} \delta_j M_i + \theta \hat{Y}_i^{II} + u_i$ and test the statistical significance of the coefficient θ
5. Select the model on the basis of the following decision matrix

Decision Matrix

	Non-Rejection of $H_0 : \lambda = 0$	Rejection of $H_0 : \lambda = 0$
Non-Rejection of $H_0 : \theta = 0$	Both the models stand individually (they cannot be combined)	Model I

Rejection of $H_0 : \theta = 0$	Model II	Combined model should be estimated

Comparing R^2 Values between Models

The coefficient of determination of a linear regression model or the R^2 value is considered as a measure of its *goodness-of-fit*. The model with higher value of R^2 is considered as a better fit as the estimated value of the dependent variable in such a case will be closer to its actual value. Thus, often there is a tendency to add more variables into a model and increase the value of R^2. However, there are following important concerns in using R^2 value as a criterion for model selection.

(i) The value of R^2 indicates *in-sample* goodness-of-fit and hence the model selected based on only the R^2 value may not forecast *out-of-sample* observations accurately.

(ii) While the value of R^2 does not fall despite inclusion of more explanatory variables into a regression model, degrees of freedom decline. Although adjusted R^2 penalizes for inclusion of more independent variables, rejection of a model on the basis of low value of adjusted R^2 only may be misleading.

(iii) More importantly, when the dependent variables are different, the TSS are not the same for the two models. As a result, the goodness-of-fit (R^2) of the two models are not directly comparable. What should be done in such cases to make the R^2 values comparable?

Consider the following two models:

Model I: $Y_i = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_i + u_i$; Model II: $ln(Y_i) = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_i + u_i$

Here, the dependent variable of Model I and Model II are Y_i and $ln(Y_i)$ respectively. Thus, the dependent variables of the two models are different and hence their R^2 values are not directly comparable. The following steps may be followed to make the R^2 values comparable:

<u>Steps to Followed:</u>

1. Estimate Model II and compute the fitted values of the dependent variable
2. Generate $\widehat{Y''}_i = exp\left(\widehat{\ln(Y_i)}\right)$
3. Compute the square of the correlation coefficient between Y_i and $\widehat{Y_i}$ which is comparable with the unadjusted R^2 of the linear model
4. Alternatively, estimate Model I and compute the fitted values of the dependent variable
5. Make logarithmic transformation of the fitted values of the dependent variable as $\ln(\widehat{Y'}_i)$
6. Compute the square of the correlation coefficient between $\ln(Y_i)$ and $\ln(\widehat{Y'}_i)$ which is comparable with the unadjusted R^2 of Model II

(f) Mackinnon-White-Davidson (MWD) Test:

Model I: $Y_i = \alpha_0 + \sum_{j=1}^{p} \alpha_j X_i + u_i$; Model II: $ln(Y_i) = \alpha_0 + \sum_{j=1}^{p} \alpha_j ln(X_i) + u_i$

The hypotheses can be written in the following way:

Model I: The linear function is valid;

Model II: The log-log function is valid

Steps to be followed:

1. Estimate the linear model and obtain the estimated values of the dependent variable (Y_i) as $\hat{Y}_i'$
2. Estimate the log-log model and obtain the estimated values of the dependent variable $ln(Y_i)$ as $\hat{Y}_i''$
3. Compute $Z_{1i} = \hat{Y}_i'' - ln(\hat{Y}_i')$ and regress Y on the independent variables (X) and Z_1
4. Reject the null hypothesis for Model I if the coefficient of Z_1 is statistically significant
5. Again, compute $Z_{2i} = exp(\hat{Y}_i'') - \hat{Y}_i'$ and regress $ln(Y)$ on the independent variables $ln(X)$ and Z_2
6. Reject the null hypothesis for Model II if the coefficient of Z_2 is statistically significant

(g) Information Criterion for Model Selection:

1. Akaike Information Criterion (AIC): $AIC = -2\,ln(L) + 2k$
2. Bayesian Information Criterion (BIC): $BIC = -2\,ln(L) + k \times ln(n)$

Here, k = number of coefficients; L = the log-likelihood value computed as:

$$ln(L) = -\frac{n}{2} ln\left(\frac{\sum \hat{u}_i^2}{n}\right) - \frac{n}{2} ln(2\pi) - \frac{n}{2}$$

This is based on $\hat{L} = \left(\frac{1}{\hat{\sigma}\sqrt{2\pi}}\right)^n e^{-\left(\frac{n}{2}\right)}$ or, $ln(\hat{L}) = -\frac{n}{2} ln(\hat{\sigma}^2) - -\frac{n}{2} ln(2\pi) - \frac{n}{2}$

The model with the least value of the AIC and the BIC should be selected.

(6) Interaction Effects in Econometric Models:

Example 1: Interaction between two dummy variables

Model: $Y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 D_{2i} + \alpha_3 (D_{1i} \times D_{2i}) + u_i$

Here, Y_i = Annual household income; $D_{1i} = 0$ for the rural the households, and $D_{1i} = 1$ for the urban households; $D_{2i} = 0$ for the BPL households, and $D_{2i} = 1$ for the APL households

Population Regression Function	Interpretation
$E\left(Y_i \mid_{D_{1i}=0, D_{2i}=0}\right) = \alpha_0$	Average income of the BPL households in rural areas
$E\left(Y_i \mid_{D_{1i}=1, D_{2i}=0}\right) = \alpha_0 + \alpha_1$	Average income of the BPL households living in urban areas; α_1 measures the effect of living in urban areas for the BPL households
$E\left(Y_i \mid_{D_{1i}=0, D_{2i}=1}\right) = \alpha_0 + \alpha_2$	Average income of the APL households living in rural areas; α_2 measures the effect of being APL for the rural households
$E\left(Y_i \mid_{D_{1i}=1, D_{2i}=1}\right) = \alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$	Average income of the APL households living in urban areas; $\alpha_1 + \alpha_3$ measures the effect of living in urban areas for APL households; $\alpha_2 + \alpha_3$ measures the effects of being APL for urban households

Example 2: Interaction between a dummy and a continuous variable

Model: $Y_i = \alpha_0 + \alpha_1 D_{1i} + \alpha_2 X_i + \alpha_3 \left(D_{1i} \times X_i\right) + u_i$

Here, Y_i = Monthly per capita consumption expenditure (MPCE); X_i = Monthly per capita income; $D_{1i} = 0$ for the rural households, and $D_{1i} = 1$ for the urban households

Population Regression Function	Interpretation
$E\left(Y_i \mid_{D_{1i}=0, X_i}\right) = \alpha_0 + \alpha_2 X_i$	Average monthly per capita expenditure by households living in rural areas
$E\left(Y_i \mid_{D_{1i}=1, X_i}\right) = \left(\alpha_0 + \alpha_1\right) + \left(\alpha_2 + \alpha_3\right) X_i$	Average monthly per capita expenditure by households living in urban areas

Example 3: Interaction between two continuous variables

Model: $Y_i = \alpha_0 + \alpha_1 X_i + \alpha_2 Z_i + \alpha_3 \left(X_i \times Z_i\right) + u_i$

Here, Y_i = MPCE; X_i = Monthly per capita income; Z_i = Wealth of the household

Partial Effect	Interpretation
$\frac{\partial Y_i}{\partial X_i} = \alpha_1 + \alpha_3 Z_i$	Partial effect of X on Y depends on Z (i.e., Partial effect of X on Y varies with Z)
$\frac{\partial Y_i}{\partial Z_i} = \alpha_2 + \alpha_3 X_i$	Partial effect of Z on Y depends on X (i.e., Partial effect of Z on Y varies with X)