# Basic Econometrics in Transportation

## Model Specification

Amir Samimi

Civil Engineering Department
Sharif University of Technology

Primary Source: Basic Econometrics (*Gujarati*)

## Outline

- How does one go about finding the “correct” model?
- What are the consequences of specification errors?
- How does one detect specification errors?
- What remedies can one adopt and with what benefits?
- How does one evaluate the performance of competing models?

## Model Selection Criteria

1. Be data admissible
   - Predictions made from the model must be logically possible.
2. Be consistent with theory
   - Must make good economic sense.
3. Have weakly exogenous regressors
   - The explanatory variables, must be uncorrelated with the error term.
4. Exhibit parameter constancy
   - In the absence of parameter stability, predictions will not be reliable.
5. Exhibit data coherency
   - Estimated residuals must be purely random (technically, white noise).
6. Be encompassing
   - Other models cannot be an improvement over the chosen model.

## Types of Specification Errors

- The model that we accept as a good model:
  - $Y_i = \beta_1 + \beta_2 X_{1i} + \beta_3 X_{2i} + \beta_4 X_{3i} + u_{1i}$

1. Omission of a relevant variable(s)
   - $Y_i = \alpha_1 + \alpha_2 X_{1i} + \alpha_3 X_{2i} + u_{2i}$ $u_{2i} = u_{1i} + \beta_4 X_{3i}$
2. Inclusion of an unnecessary variable(s)
   - $Y_i = \lambda_1 + \lambda_2 X_{1i} + \lambda_3 X_{2i} + \lambda_4 X_{3i} + \lambda_5 X_{4i} + u_{3i}$ $u_{3i} = u_{1i} - \lambda_5 X_{4i}$
3. Adopting the wrong functional form
   - $\text{Ln}Y_i = \gamma_1 + \gamma_2 X_{1i} + \gamma_3 X_{2i} + \gamma_4 X_{3i} + u_{4i}$
4. Errors of measurement
   - $Y^*_i = \beta^*_1 + \beta^*_2 X^*_{1i} + \beta^*_3 X^*_{2i} + \beta^*_4 X^*_{3i} + u^*_i$ $Y^*_i = Y_i + \varepsilon_i$ and $X^*_i = X_i + w_i$
5. Incorrect specification of the stochastic error term
   - Multiplicative versus additive error term: $Yi = \beta X_i u_i$ and $Y_i = \alpha X_i + u_i$

## Specification and Mis-specification Errors

- The first four types of error are essentially in the nature of model specification errors.
  - We have in mind a “true” model but somehow we do not estimate the correct one.
- In model mis-specification errors, we do not know what the true model is to begin with.
  - The controversy between the Keynesians and the monetarists.
  - The monetarists give primacy to money in explaining changes in GDP.
  - The Keynesians emphasize the role of government expenditure to explain changes in GDP.
  - There are two competing models.
- We will first consider model specification errors and then examine model mis-specification errors.

## Consequences of Model Specification Errors

- To keep the discussion simple,
  - We will answer this question in the context of the three-variable model and consider the first two types of specification errors discussed earlier:
- Underfitting a model
  - Omitting relevant variables,
- Overfitting a model
  - Including unnecessary variables.

## Underfitting a Model

- Suppose the true model is:
  - $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
- For some reason we fit the following model:
  - $Y_i = \alpha_1 + \alpha_2 X_{2i} + v_i$
- The consequences of omitting variable $X_3$ are as follows:
  1. If $X_2$ and $X_3$ are correlated, estimated $\alpha_1$ and $\alpha_2$ are biased and inconsistent.
  2. If $X_2$ and $X_3$ are not correlated, only estimated $\alpha_1$ is biased.
  3. The disturbance variance is incorrectly estimated.
  4. The conventionally measured variance of estimated $\alpha_2$ is biased.
  5. In consequence, the hypothesis-testing procedures are likely to give misleading conclusions.
  6. As another consequence, the forecasts will be unreliable.

## Underfitting a Model

- It can be shown that: $E(\hat{\alpha}_2) = \beta_2 + \beta_3 b_{32}$
  - $b_{32}$ is the slope in the regression of $X_3$ on $X_2$.
  - $b_{32}$ will be zero if $X_3$ on $X_2$ are uncorrelated.
- It can also be shown that:

$$\operatorname{var}(\hat{\alpha}_2) = \frac{\sigma^2}{\sum x_{2i}^2} \qquad \operatorname{var}(\hat{\beta}_2) = \frac{\sigma^2}{\sum x_{2i}^2(1 - r_{23}^2)} = \frac{\sigma^2}{\sum x_{2i}^2}\text{VIF}$$

  - Although estimated $\alpha_2$ is biased, its variance is smaller.
  - There is a tradeoff between bias and efficiency involved here.
  - Estimated $\sigma^2$ are not the same because RSS and df of the models are different.
  - If X has a strong impact on Y, it may reduce RSS more than the loss in df.
  - Thus, inclusion of such variables will reduce bias and standard error.
- Once a model is formulated on the basis of the relevant theory, one is ill-advised to drop a variable from such a model.

# Overfitting a Model

- Suppose the true model is:
  - $Y_i = \beta_1 + \beta_2 X_{2i} + u_i$
- For some reason we fit the following model:
  - $Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + v_i$
- The consequences of adding variable $X_3$ are as follows:
  1. The OLS estimators of the parameters of the “incorrect” model are all unbiased and consistent.
  2. The error variance $\sigma^2$ is correctly estimated.
  3. The usual hypothesis-testing procedures remain valid.
  4. The estimated α’s will be generally inefficient (larger variance): $\frac{\operatorname{var}(\hat{\alpha}_2)}{\operatorname{var}(\hat{\beta}_2)} = \frac{1}{1 - r_{23}^2}$
- Inclusion of unnecessary variables makes the estimations less precise.

# Tests of Specification Errors

## Detecting the Presence of Unnecessary Variables

- Suppose we develop a k-variable model, but are not sure that $X_k$ really belongs in the model: $Y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i$
  - One simple way to find this out is to test the significance of the estimated $\beta_k$ with the usual t test.
  - If we are not sure whether $X_3$ and $X_4$ legitimately belong in the model, we can be easily ascertained by the F test discussed before.
  - It is important to remember that in carrying out these tests of significance we have a specific model in mind.
  - **Given that model**, then, we can find out whether one or more regressors are really relevant by the usual t and F tests.

# Tests of Specification Errors

## Detecting the Presence of Unnecessary Variables

- **Note**: we should not use t and F tests to build a model iteratively.
  - We should not say Y is related to $X_2$ only because estimated $\beta_2$ is statistically significant and then expand the model to include $X_3$ and decide to keep that variable in the model, and so on.
- This strategy of building a model is called the **bottom-up approach** or by the somewhat critical term, data mining.
- Nominal significance in data mining approach:
  - For c candidate Xs out of which k are finally selected on the basis of data mining, the true level of significance ($\alpha^*$) is related to the nominal level of significance ($\alpha$) as : $\alpha^* = 1 - (1 - \alpha)^{c/k}$
  - The art of the applied econometrician is to allow for data-driven theory while avoiding the considerable dangers in data mining.

# Tests of Specification Errors

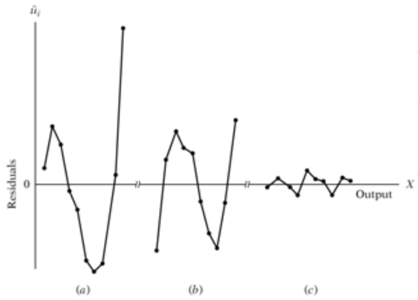
## Omitted Variables and Incorrect Functional Form

- On the basis of theory or prior empirical work, we develop a model that we believe captures the essence of the subject under study.
  - Then we look at some broad features of the results, such as adjusted $R^2$, t ratios, signs of the estimated coefficients, Durbin–Watson statistic, …
  - If diagnostics do not look encouraging, then we begin to look for remedies:
    - Maybe we have omitted an important variable,
    - Maybe we have have used the wrong functional form,
    - Maybe we have not first-differenced (to remove autocorrelation)
- To determine whether model inadequacy is on account of one of these problems, we can use some of the coming methods.

# Omitted Variables and Incorrect Function

## Examination of Residuals

- Omission of an important variable or incorrect functional form, causes a plot of residuals to exhibit distinct patterns.



a) A linear function:
   $Y_i = \lambda_1 + \lambda_2 X_i + u_{3i}$
b) A quadratic function:
   $Y_i = \alpha_1 + \alpha_2 X_i + \alpha_3 X^2_i + u_{2i}$
c) The true total cost function:
   $Y_i = \beta_1 + \beta_2 X_i + \beta_3 X^2_i + \beta_4 X^3_i + u_i$

- As we approach the truth, the residuals are smaller and they do not exhibit noticeable patterns.

# Omitted Variables and Incorrect Function

## Durbin–Watson d Statistic

1. From the assumed model, obtain the OLS residuals.
2. If you believe the model is misspecified because it excludes Z, order the residuals according to increasing values of Z.
   - The Z variable could be one of the X variables included in the assumed model or it could be some function of that variable.
3. Compute the d statistic from the residuals thus ordered.
4. If the estimated d value is significant, one can accept the hypothesis of model misspecification.

$$\begin{array}{lll} \hat{Y}_i = & 166.467 & +\ 19.933X_i \\ & (19.021) & \quad (3.066) \\ & (8.752) & \quad (6.502) \end{array} \qquad d = 0.716$$

$$\begin{array}{llll} \hat{Y}_i = & 222.383 & -\ 8.0250X_i & +\ 2.542X_i^2 \\ & (23.488) & \quad (9.809) & \quad (0.869) \\ & (9.468) & \quad (-0.818) & \quad (2.925) \end{array} \qquad d = 1.038$$

$$\begin{array}{lllll} \hat{Y}_i = & 141.767 & +\ 63.478X_i & -\ 12.962X_i^2 & +\ 0.939X_i^3 \\ & (6.375) & \quad (4.778) & \quad (0.9856) & \quad (0.0592) \\ & (22.238) & \quad (13.285) & \quad (-13.151) & \quad (15.861) \end{array} \qquad d = 2.70$$

# Omitted Variables and Incorrect Function

## Ramsey's RESET Test

1. From the chosen model obtain the estimated $Y_i$
   - $Y_i = \lambda_1 + \lambda_2 X_i + u_{3i}$
2. Rerun the model introducing the estimated $Y_i$ in some form as an additional regressor. (get idea from the plot of residuals and estimated Y)
   - $Y_i = \beta_1 + \beta_2 X_i + \beta_3 \hat{Y}_i^2 + \beta_4 \hat{Y}_i^3 + u_i$
3. Use F test to find out if $R^2$ is significantly improved. $F = \dfrac{\left(R^2_{new} - R^2_{old}\right)/DF}{\left(1 - R^2_{new}\right)/DF}$
   - $F = \dfrac{(0.9983 - 0.8409)/2}{(1.0.9983)/(10 - 4)} = 284.4035$
4. If F value is significant, one can accept that the model is mis-specified.

# Omitted Variables and Incorrect Function

## Lagrange Multiplier Test for Adding Variables

1. Estimate the restricted regression by OLS and obtain residuals.
   - $Y_i = \lambda_1 + \lambda_2 X_i + u_{3i}$
2. If the unrestricted regression ($Y_i = \beta_1 + \beta_2 X_i + \beta_3 X^2_i + \beta_4 X^3_i + u_i$) is the true one, the residuals should be related to $X^2_i$ and $X^3_i$.
3. Regress the estimated $u_i$ from Step 1 on all the regressors
   - $\hat{u}_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + \beta_4 X_i^3 + v_i$
4. n times $R^2$ from the auxiliary regression follows the chi square distribution: $nR^2 \underset{asy}{\sim} \chi^2_{(\text{number of restrictions})}$
5. If the chi-square value exceeds the value, we reject the restricted regression.

## Errors of Measurement

- We assume that the data is “accurate”.
  - Not guess estimates, extrapolated, rounded off, etc in any systematic manner.
- Unfortunately, this ideal is not usually met in practice!
  - Non-response errors, reporting errors, and computing errors.
- Whatever the reasons, it is a potentially troublesome problem.
  - It forms another example of specification bias.
- Will be discussed in two parts:
  - Errors of measurement in the dependent variable Y
  - Errors of measurement in the explanatory variable X

## Errors of Measurement in Y

- Consider the following model:
  - $Y^*_i = \alpha + \beta X_i + u_i$
  - $Y^*_i$ = permanent consumption expenditure; $X_i$ = current income
- $Y^*_i$ is not directly measurable, we may use $Y_i$ such that
  - $Y_i = Y^*_i + \varepsilon_i$
- Therefore, we estimate
  - $Y_i = (\alpha + \beta X_i + u_i) + \varepsilon_i = \alpha + \beta X_i + v_i$
  - $v_i$ is a composite error term: population disturbance and measurement error.
- For simplicity assume that:
  - $E(u_i) = E(\varepsilon_i) = 0$, $\text{cov}(X_i, u_i) = 0$ (which is the assumption of CLRM);
  - $\text{cov}(X_i, \varepsilon_i) = 0$: errors of measurement in $Y^*_i$ are uncorrelated with $X_i$;
  - $\text{cov}(u_i, \varepsilon_i) = 0$: equation error and the measurement error are uncorrelated.

## Errors of Measurement in Y

- With these assumptions, it can be seen that
  - β estimated from either equations will be an unbiased estimator of the true β.
  - The standard errors of β estimated from the two equations are different:

    First model: $\text{var}(\hat{\beta}) = \dfrac{\sigma_u^2}{\sum x_i^2}$ Second model: $\text{var}(\hat{\beta}) = \dfrac{\sigma_v^2}{\sum x_i^2} = \dfrac{\sigma_u^2 + \sigma_\varepsilon^2}{\sum x_i^2}$
  - The latter variance is larger than the former.

- Therefore, although the errors of measurement in Y still give unbiased estimates of the parameters and their variances, the estimated variances are now larger than in the case where there are no such errors of measurement.

## Errors of Measurement in X

- Consider the following model:
  - $Y_i = \alpha + \beta X^*_i + u_i$
- $X^*_i$ is not directly measurable, we may use $X_i$ such that
  - $X_i = X^*_i + w_i$
- Therefore, we estimate
  - $Y_i = \alpha + \beta(X_i - w_i) + u_i = \alpha + \beta X_i + z_i$
  - $z_i = u_i - \beta w_i$ is a compound of equation and measurement errors.
- Assumptions:
  - Even if we assume that $w_i$ has zero mean, is serially independent, and is uncorrelated with $u_i$,
  - We can no longer assume $\text{cov}(z_i, X_i) = 0$: we can show $\text{cov}(z_i, X_i) = -\beta\sigma^2_w$
- A crucial CLRM assumption is violated!

## Errors of Measurement in X

- As the explanatory variable and the error term are correlated, the OLS estimators are biased and inconsistent.
  - It is shown that: $\operatorname{plim} \hat{\beta} = \beta \left[ \frac{1}{1 + \sigma_w^2 / \sigma_{X^*}^2} \right]$
- What is the solution?
  - The answer is not easy.
  - If $\sigma^2_w$ is small compared to $\sigma^2_{X^*}$ , we can "assume away" for practical purposes.
    - o The rub here is that there is no way to judge their relative magnitudes.
  - One other remedy is the use of instrumental or proxy variables.
    - o Although highly correlated with X, are uncorrelated with the equation and measurement error terms.
- Measure the data as accurately as possible.

## Incorrect Specification of Error Term

- Since the error term is not directly observable, there is no easy way to determine the form in which it enters the model.

- Consider $Y_i = \beta X_i u_i$ and $Y_i = \alpha X_i + u_i$
- Assume the multiplicative is the "correct" model ($\ln u_i \sim N(0, \sigma^2)$) but we estimated the other one.
- It can be shown that $u_i \sim$ log normal $[e^{\sigma^2/2}, e^{\sigma^2}(e^{\sigma^2} - 1)]$, and thus E(estimated $\alpha$) = $\beta\ e^{\sigma^2/2}$
- Estimated $\alpha$ is a biased, as its average is not equal to the true $\beta$.

## Nested Versus Non-Nested Models

- Two models are nested, if one can be derived as a special case of the other:
  - Model A: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \beta_4 X_{4i} + \beta_5 X_{5i} + u_i$
  - Model B: $Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i$
  - We already have looked at tests of nested models.
- Tests of non-nested hypotheses:
  - The **discrimination** approach:
    - o Given competing models, one chooses a model based on some criteria of goodness of fit ($R^2$, adjusted $R^2$, Akaike's information, etc).
  - The **discerning** approach:
    - o We take into account information provided by other models, in investigating one model. (**non-nested F test**, **Davidson–MacKinnon J test**, Cox test, JA test, P test, Mizon–Richard encompassing test, etc).

## The Non-Nested *F* Test

- Consider:
  - Model C: $Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + u_i$
  - Model D: $Y_i = \beta_1 + \beta_2 Z_{2i} + \beta_3 Z_{3i} + v_i$
- Estimate the following nested or *hybrid* model and use the *F* test:
  - $Y_i = \lambda_1 + \lambda_2 X_{2i} + \lambda_3 X_{3i} + \lambda_4 Z_{2i} + \lambda_5 Z_{3i} + w_i$
  - If Model C is correct, $\lambda_4 = \lambda_5 = 0$, whereas Model D is correct if $\lambda_2 = \lambda_3 = 0$.
- Problems with this procedure:
  - If X's and Z's are highly correlated, we have no way of deciding which one is the correct model.
  - To test the significance of an incremental contribution, one should choose Model C or D as the reference. Choice of the reference hypothesis could determine the outcome of the choice model.
  - The artificially nested model F may not have any economic meaning.

## Davidson–MacKinnon *J* Test

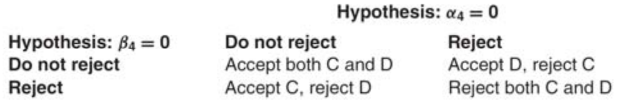
1. Estimate Model D and obtain the estimated Y values.
2. Add estimated Y as a regressor to Model C and estimate:

   $$Y_i = \alpha_1 + \alpha_2 X_{2i} + \alpha_3 X_{3i} + \alpha_4 \hat{Y}_i^D + u_i$$

   This model is an example of the encompassing principle.
3. Using the t test, test the hypothesis that $\alpha_4 = 0$.
4. If this is not rejected, we accept Model C as the true model.
   - This is because the influence of variables not included in Model C, have no additional explanatory power beyond that contributed by Model C.
   - If the null hypothesis is rejected, Model C cannot be the true model.
5. Reverse the roles of hypotheses, or Models C and D.

## Davidson–MacKinnon *J* Test

- Some problems of the J test:
  - No clear answer if the test leads to acceptance or rejection of both models.



| | **Hypothesis: $\alpha_4 = 0$** | |
|---|---|---|
| **Hypothesis: $\beta_4 = 0$** | **Do not reject** | **Reject** |
| **Do not reject** | Accept both C and D | Accept D, reject C |
| **Reject** | Accept C, reject D | Reject both C and D |

  - The J test may not be very powerful in small samples.

## Model Selection Criteria

- We distinguish between in-sample and out-of-sample forecasting.
  - In-sample forecasting tells us how the model fits the data in a given sample.
  - Out-of-sample forecasting tries to determine how a model forecasts future.
- Several criteria are used for this purpose:
  1. $R^2$
  2. Adjusted $R^2$
  3. Akaike information criterion (AIC)
  4. Schwarz Information criterion (SIC)
  5. Mallow's $C_p$ criterion
  6. forecast $\chi^2$ (chi-square)
- There is a tradeoff between goodness of fit and a model's complexity (as judged by the number of X's) in criteria 2 to 5.

## The $R^2$ Criterion

- $R^2$ is defined as ESS/TSS
  - Necessarily lies between 0 and 1.
  - The closer it is to 1, the better is the fit.
- Problems with $R^2$
  1. It measures in-sample goodness of fit.
     - There is no guarantee that it will forecast well out-of-sample observations.
  2. The dependent variable must be the same, in comparing two or more $R^2$'s.
  3. An $R^2$ cannot fall when variables are added to the model.
     - There is every temptation to play the game of "maximizing the $R^2$".
     - Adding variables may increase $R^2$ but it may also increase the variance of forecast error.

## Adjusted $R^2$

- The adjusted $R^2$ is defined as $\bar{R}^2 = 1 - \frac{\text{RSS}/(n-k)}{\text{TSS}/(n-1)}$
- Adjusted $R^2$ is a better measure than $R^2$, as it penalizes for adding more X's.
- Again keep in mind that Y must be the same for the comparison to be valid.

## Akaike Information Criterion (AIC)

- AIC is defined as

$$\text{AIC} = e^{2k/n}\frac{\sum \hat{u}_i^2}{n} = e^{2k/n}\frac{\text{RSS}}{\text{n}} \quad \text{or} \quad \ln \text{AIC} = \left(\frac{2k}{n}\right) + \ln\left(\frac{\text{RSS}}{n}\right)$$

- AIC imposes a harsher penalty than adjusted $R^2$ for adding X's.
- In comparing two or more models, the model with the lowest value of AIC is preferred.
- One advantage of AIC is that it is useful for not only in-sample but also out of-sample forecasting performance of a regression model.
- Also, it is useful for both nested and non-nested models.

## Schwarz Information Criterion (SIC)

- SIC is defined as

$$\text{SIC} = n^{k/n}\frac{\sum \hat{u}^2}{n} = n^{k/n}\frac{\text{RSS}}{n} \quad \text{or} \quad \ln \text{SIC} = \frac{k}{n}\ln n + \ln\left(\frac{\text{RSS}}{\text{n}}\right)$$

- SIC imposes a harsher penalty than AIC.
- Like AIC, the lower the value of SIC, the better the model.
- Like AIC, SIC can be used to compare in-sample or out-of-sample forecasting performance of a model.

## Mallows's $C_p$ Criterion

- Mallows has developed a criterion for model selection:

$$C_p = \frac{\text{RSS}_p}{\hat{\sigma}^2} - (n - 2p)$$

  - $\text{RSS}_p$ is the residual sum of squares using the p regressors.
- If the model with p regressors does not suffer from lack of fit, it can be shown that $E(\text{RSS}_p) = (n - p)\sigma^2$.
  - It is true approximately that $E(C_p) \approx p$.
  - In practice, one usually plots $C_p$ against p and look for a model that has a low $C_p$ value, about equal to p.

## Forecast Chi-Square

- Suppose we have a regression model based on *n* observations and suppose we want to use it to forecast for an additional t observations. Now the forecast $\chi^2$ test is defined as follows:

$$\text{Forecast, } \chi^2 = \frac{\sum_{n+1}^{n+t} \hat{u}_i^2}{\hat{\sigma}^2}$$

  - If we hypothesize that the parameter values have not changed between the sample and post-sample periods, it can be shown that the statistic above follows the chi-square distribution with t degrees of freedom.
  - The forecast $\chi^2$ test has a weak statistical power, meaning that the probability that the test will correctly reject a false null hypothesis is low and therefore the test should be used as a signal rather than a definitive test.

## Ten Commandments

1. Thou shalt use common sense and economic theory.
2. Thou shalt ask the right questions.
3. Thou shalt know the context (No ignorant statistical analysis).
4. Thou shalt inspect the data.
5. Thou shalt not worship complexity.
6. Thou shalt look long and hard at thy results.
7. Thou shalt beware the costs of data mining.
8. Thou shalt be willing to compromise.
9. Thou shalt not confuse statistical with practical significance.
10. Thou shalt anticipate criticism.

## Homework 7

Basic Econometrics (Gujarati, 2003)

1. Chapter 13, Problem 21 [35 points]
2. Chapter 13, Problem 22 [25 points]
3. Chapter 13, Problem 25 [40 points]

Assignment weight factor = 1