Introductory Econometrics

Jeffrey M. Wooldridge

Chapter 1 The Nature of Econometrics and Economic Data

I. The goal of any econometric analysis is to estimate the parameters in the model and to test hypotheses about these parameters; the values and signs of the parameters determine the validity of an economic theory and the effects of certain policies.

II. Panel data - advantages:

1. Having multiple observations on the same units allows us to control certain unobserved characteristics of individuals, firms, and so on. The use of more than one observation can facilitate causal inference in situations where inferring causality would be very hard if only a single cross section were available.
2. They often allow us to study the importance of lags in behavior or the result of decision making.

Part 1 Regression Analysis with Cross-Sectional Data

Chapter 2 The Simple Regression Model

I. Model: Y = b0 + b1x + u

1. **Population regression function (PRF)**: E(y|x) = b0 + b1x
2. systematic part of y: b0 + b1x
3. unsystematic part: u

II. **Sample regression function (SRF)**: yhat = b0hat + b1hat*x

1. PRF is something fixed, but unknown, in the population. Since the SRF is obtained for a given sample of data, a new sample will generate a different slope and intercept.

III. **Correlation**: it is possible for u to be uncorrelated with x while being correlated with functions of x, such as x^2.

E(u|x) = E(u) → Cov(u, x) = 0. not vice versa.

IV. Algebraic properties of OLS statistics

1. The sum of the OLS residuals is zero.
2. The sample covariance between the (each) regressors and the residuals is zero. Consequently, the sample covariance between the fitted values and the residuals is zero.
3. The point $(\bar{x}, \bar{y})$ is on the OLS regression line.
4. the goodness-of-fit of the model is **invariant** to changes in the units of y or x.
5. The homoskedasticity assumption plays no role in showing OLS estimators are unbiased.

V. Variance

1. $Var(b1) = var(u)/SST_x$

a. more variation in the unobservables (u) affecting y makes it more difficult to precisely estimate b1.

b. More variability in x is preferred, since the more spread out is the sample of independent variables, the easier it is to trace out the relationship between E(y|x) and x. That is, the easier it is to estimate b1.
2. **standard error of the regression**, standard error of the estimate and the root mean squared error = $\sqrt{\frac{1}{(n-2)}\sum \bar{u}^2}$

Chapter 3 Multiple Regression Analysis: Estimation

I. The power of multiple regression analysis is that is allows us to do in nonexperimental environments what natural scientists are able to do in a controlled laboratory setting: keep other factors fixed.

II. Model: $Y = b_0 + b_1x_1 + b_2x_2 + u$

$\vec{b_1} = (\sum_{i=1}^{n} v_{i1}y_i)/(\sum_{i=1}^{n} v_{i1}^2)$, where v is the OLS residuals from a simple regression of x_1 on x_2.

1. v is the part of x1 that is uncorrelated with x2, or v is x1 after the effects of x2 have been partialled out, or netted out. Thus, b1 measures the sample relationship between y and x1 after x2 has been partialled out.

III. Goodness-of-fit

1. **R^2 = the squared correlation coefficient between the actual y and the fitted values yhat.**
2. R^2 never decreases because the sum of squared residuals never increases when additional regressors are added to the model.

IV. **Regression through the origin**:

1. OLS residuals no longer have a zero sample average.
2. R^2 can be negative. This means that the sample average "explains" more of the variation in the y than the explanatory variables.

V. MLR Assumptions:

A1: linear in parameters.
A2: random sampling.
A3: Zero conditional mean: E(u|x1, x2, …, xk) = 0
When A3 holds, we say that we have Exogenous explanatory variables. If xj is correlated with u for any reason, then xj is said to be an endogenous explanatory variables.
A4: No perfect collinearity.
A1 – A4 → unbiasedness of OLS

VI. Overspecifying the model:

1. Including one or more irrelevant variables, does not affect the unbiasedness of the OLS estimators.

VII. Variance of OLS estimators:

A5: homoskedasticity

1. **Gauss – Markov assumptions: A1 – A5**
2. $Var(b_j) = \frac{\sigma^2}{SST_j(1-R_j^2)}$, where R^2 is from regressing xj on all other independent variables (and including an intercept).

a. The error variance, σ2, is a feature of the population, it has nothing to do with the sample size.

b. SSTj: the total sample variation in xj: a small sample size → small value of SSTj → large var(bj)

c. R_j^2: high correlation between two or more independent variables is called multicollinearity.

3. **A high degree of correlation between certain independent variables can be irrelevant as to how well we can estimate other parameters in the model**:

$Y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + u$, where x2 and x3 are highly correlated.

The var(b2) and var(b3) may be large. But the amount of correlation between x2 and x3 has no direct effect on var(b1). In fact, if x1 is uncorrelated with x2 and x3, then R_1^2=0 and $var(b1) = \frac{\sigma^2}{SST_1}$, regardless of how much correlation there is between x2 and x3.

If b1 is the parameter of interest, we do not really care about the amount of correlation between x2 and x3.

4. **The tradeoff between bias and variance**.

If the true model is $Y = b_0 + b_1x_1 + b_2x_2 + u$, instead, we estimate $Y = b_0 + b'_1x_1 + u$

a. when b2 is nonzero, b'1 is biased, b1 is unbiased, var(b'1)<var(b1);

b. when b2 is zero, b'1 is unbiased, b1 is unbiased, var(b'1)<var(b1) → **a higher variance for the estimator of b1 is the cost of including an irrelevant variable in a model;**

VIII. Estimating: standard errors of estimators.

1. **Under A1-A5**: $E(\sigma'^2) = \sigma^2$, where $\sigma'^2 = \frac{1}{(n-k-1)}\sum \bar{u}^2$ (σ' is σhat)
2. **Standard deviation** of bj', $sd(bj') = \frac{\sigma}{\sqrt{SST_j(1-R_j^2)}}$
3. **Standard error** of bj': $se(bj') = \frac{\sigma hat}{\sqrt{SST_j(1-R_j^2)}}$

Standard error of bj' is not a valid estimator of sd(bj') if the errors exhibit heteroskedasticity. Thus, while the presence of heteroskedasticity does not cause bias in the bj', it does lead to bias in the usual formula for Var(bj'), which then invalidates the standard errors.

Chapter 4 Multiple Regression Analysis: Inference

I. Classical Linear model (CLM) assumptions:

1. Even under Gauss-Markov assumptions (A1-5), the distribution of estimators can have virtually any shape.
2. A6: Normality→ $u \sim N(0, \sigma^2)$.
3. **Under CLM, the OLS estimators are the minimum variance unbiased estimators**; we no longer have to restrict our comparison to estimators that are linear in the y.
4. **THM 4.1** under A1-A6: $b'j \sim N(bj, Var(b'j))$; → $(b'j - bj)/sd(b'j) \sim N(0,1)$.
5. **THM 4.2** under A1-A6: → $(b'j - bj)/se(b'j) \sim t(n-k-1)$.

II. Test – T-test:

1. It is a key to remember that we are testing hypothesis about the population parameters. We are not testing hypothesis about the estimates from a particular sample. Thus, it never makes sense to state a null hypothesis as "H0: $b'j = 0$"
2. There is no "correct" significance level.
3. P-value:
   a. Given the observed value of the t statistic, what is the smallest significance level at which the null hypothesis would be rejected?
   b. P-value is the probability of observing a t statistic as extreme as we did if the null hypothesis is true. This means that small p-values are evidence against the null; large p-values provide little evidence against H0.
   c. **To obtain the one-sided p-value: just divide the two-sided p-value by 2**.
4. We should say "**we fail to reject H0 at the x% level**," rather than "H0 is accepted at the x% level".

III. Economic vs. statistic significance.

1. The **statistical significance** of a variable xj is determined entirely by the size of t_{bj}, whereas the **economic significance** or practical significance of a variable is related to the **size and sign of bj.**
2. With large sample sizes, parameters can be estimated very precisely: standard errors are often quite small relative to the coefficient estimates, which usually results in statistical significance.
3. Some researchers insist on using smaller significance levels as the sample size increases, partly as a way to offset the fact that standard errors are getting smaller. Most researchers are also willing to entertain larger significance levels in applications with small sample sizes, reflecting the fact that it is harder to find significance with smaller sample sizes.

IV. Confidence intervals (CI)

1. meaning of a CI: if random samples were obtained over and over again, with the CI computed each time, then the unknown population value bj would lie in the CI for 95% of the samples. Unfortunately, for the single sample that we use to construct the CI, we do not know whether bj is actually contained in the interval. We hope we have obtained a sample that is one of the 95% of all samples where the interval estimate contains bj, but we have no guarantee.

2. Use a rule of thumb for a 95% CI: bhat plus or minus two of its standard errors. For small degrees of freedom, the exact percentiles should be obtained from the t table.

V. Testing hypothesis about a single linear combination of the parameters.
$Y = b_0 + b_1x_1 + b_2x_2 + u$
Test: H0: b1=b2.
Let a = b1-b2, then the new model is:
Y = b0 + (a + b2)x1 + b2x2 + u = b0 + ax1 + b2(x1+x2) + u
Then the new model can deliver the standard error of interest.

VI. Testing multiple linear restrictions: F test.
1. $F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$: think of F as measuring the relative increase in SSR when moving from the unrestricted to the restricted model.
2. the denominator of F is just the unbiased estimator of var(u) in the unrestricted model.
3. F test is often useful for testing exclusion of a group of variables when the variables in the group are highly correlated (t-test for each single variable may be insignificant, but the joint test is significant).
4. the F statistic for testing exclusion of a single variable is equal to the square of the corresponding t statistic. The t statistic is more flexible for testing a single hypothesis because it can be used to test against one-sided alternatives.
5. It is also possible that, in a group of several explanatory variables, one variable has a significant t statistic, but the group of variable is jointly insignificant at the usual significant levels. → **F statistic is intended to detect whether any combination of a set of coefficients is different from zero, but it is never the best test for determining whether a single coefficient is different from zero**.
6. $F = \frac{(R_{ur}^2 - R_r^2)/q}{(1-R_{ur}^2)/(n-k-1)}$, Be careful, when y is different, we cannot use this.

Chapter 5 Multiple Regression Analysis: OLS Asymptotics

I. Consistency

1. Under A1-A4, the OLS estimators are consistent.
2. **A3'**: zero mean and zero correlation: E(u) and cov(xj, u) = 0, for all j. This assumption **is weaker than A3**.
3. **OLS is consistent if we use A3' instead of A3**.
4. If the error is correlated with any of the independent variables, then OLS is biased and inconsistent because any bias persists as the sample size grows.

II. Asymptotic normality and large sample inference

1. Normality plays no role in the unbiasedness of OLS, nor does it affect the conclusion that OLS is the BLUE under the Gauss-Markov assumptions. But exact inference based on t and F statistics requires normality assumption (A6).

2. **with large sample, we don't need A6 to arrive at normality**
3. **THM 5.2:**

Under the Gauss-Markov assumptions A1-A5

a. $\sqrt{n}(b_j^{'} - b_j) \xrightarrow{a} N(0, \sigma^2 / a_j^2)$, where $\sigma^2 / a_j^2 > 0$ is the asymptotic variance of $\sqrt{n}(b_j^{'} - b_j)$; for the slope coefficients, $a_j^2 = p\lim(n^{-1}\sum_{j=1}^{n} r_{ij}^2)$, where r are the residuals from regressing xj on the other independent variables.

b. σ '^{2} is a consistent estimator of σ^2= var(u);

c. for each j, (b'j – bj)/se(b'j) → N(0,1) where se(b'j) is the usual OLS standard error.

Note: We **don't need normality assumption** here, but we **do need to assume** the error has finite variance, **zero conditional mean and homoskedasticity.**

4. THM 5.3: **Asymptotic efficiency** of OLS: under A1-A5, the OLS estimators have the smallest asymptotic variances.

III. Other large sample tests: The Lagrange Multiplier statistic

LM requires estimation of the restricted model only.

1. Regress y on the **restricted set** of independent variables and save the residuals, uhat.
2. Regress uhat on **all** of the independent variables and obtain the R2
3. Compute **LM=nR^2** (n is the original sample size).
4. LM ~ χ_q^2

As with the F statistic, we must be sure to use the same observations in steps 1 and 2. If data are missing for some of the independent variables that are excluded under the null hypothesis, the residuals from step 1 should be obtained from a regression on the reduced data set.

Chapter 6 Multiple Regression Analysis: Further Issues

I. **Beta coefficients**:

1. Compute the z-score for every variable in the sample.
2. $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$
   $z = b_0 + b_1 x_1 + ... + b_k x_k + v$, where $b_j = (\sigma_j / \sigma_y)\beta_j$
3. If x1 increases by one standard deviation, then y changes by b1 standard deviations. This **makes the scale of the regressors irrelevant**.
4. Whether we use **standardized or unstandardized variables does not affect statistical significance: the t statistics are the same** in both cases.

II. Models with interaction terms.

$Y = b_0 + b_1x_1 + b_2x_2 + b3x_1*x_2 + u$, or

$Y = a_0 + a_1x_1 + a_2x_2 + a3(x_1 - \mu_1)(x_2 - \mu_2) + u$

III. Adjusted R2

1. The **population R2** is defined as 1 - σ_u^2 / σ_y^2, R2 estimates σ_u^2 / σ_y^2 by (SSR/n)/(SST/n). These are biased.
2. adjusted R2 = 1 - [SSR/(n-k-1)]/[SST/(n-1)].
3. It is tempting to think that adjusted R2 corrects the bias in R2 for estimating the population R2, but it does not: **the ratio of two unbiased estimators is not an unbiased estimator.**
4. If we add a new independent variable to a regression equation, **adjusted R2 increases if, and only if, the t statistic on the new variable is greater than one in absolute value.** (an extension of this is that adjusted R2 increases when a group of variables is added to a regression, iff the F statistic for joint significance of the new variables is greater than unity.) Thus, we see that using adjusted R2 to decide whether a certain independent variable (or set of variables) belongs in a model gives us a different answer than standard t or F testing.
5. Comparing adjusted R2 to choose among different nonnested sets of independent variables can be valuable when these variables represent different functional forms.
   $Y = b_0 + b_1 \log(x) + u$, or
   $Y = b_0 + b_1 x + b_2 x^2 + u$, or
6. There is an important limitation in using adjusted R2 to choose between nonnested models: we cannot use it to choose between different functional forms for the dependent variables.

IV. Controlling for too many factors in regression

1. Adding a new independent variable to a regression can exacerbate the multicollinearity problem. On the other hand, since we are taking something out of the error term, adding a variable generally reduces the error variance. Generally, we cannot know which effect will dominate.
2. We **should always include independent variables that affect y and are uncorrelated with all of the independent variables of interest**. → smaller sampling variance, though we obtain an unbiased and consistent estimator with/without the controlling variables.

V. Prediction and residual analysis

1. **Confidence intervals for predictions**: CI for the average value of y for the subpopulation with a given set of covariates.

$Y = \theta + b1(x1 - c1) + b2(x2 - c2) \ldots + bk(xk - ck) + u.$

Where θ is exactly the predicted value of y when xj = cj. We can get the CI of θ from the OLS.

- Because the **variance of the intercept estimator is smallest when each explanatory variables has zero sample mean**, it follows from the regression above that the variance of the prediction is smallest at the mean values of the xj.

2. prediction interval: CI for a particular unit from the population.

$$y^0 = \beta_0 + \beta_1^0 x_1 + \ldots + \beta_k^0 x_k + u^0$$

$$e'^0 = y^0 - y'^0 = (\beta_0 + \beta_1^0 x_1 + \ldots + \beta_k^0 x_k) + u^0 - y'^0$$

$se(e'^0) = \sqrt{[se(y'^0)^2 + \sigma'^2}$,We can get standard errors of y' from step 1.

VI. Predicting y when log(y) is the dependent variable

$Log(y) = b_0 + b_1x_1 + b_2x_2 + \ldots + bkx_k + u$ --- (1)

If the model follows CLM assumptions A1-A6,

$E(y|x) = \exp(\sigma^2)*\exp(b_0 + b_1x_1 + b_2x_2 + \ldots + bkx_k)$

Using the sample: $y' = \exp(\sigma'^2)*\exp(\log(y)')$.

1. **The prediction above is biased, but consistent**. There are no unbiased predictions of y, and in many cases, the above prediction works well. However, **it does rely on the normality of the error term, u**.
2. It is useful to have a prediction that does not rely on normality. If we just assume that u is independent of xj, then

E(y|x) =a*exp(b0 + b1x1 + b2x2 + … + bkxk)

Where **a is the expected value of exp(u)**, which must be greater than unity. Given an estimate a', we can predict y as

Y' = a'* exp(log(y)') ---------- (2)

3. **Predicting y when the dependent variable is log(y)**:

a. Run OLS in (1)

b. for given values of x1,…, xk, obtain logy' from (1), create $m_i = \exp(logy_i')$

c. regress y on the single variable m without an intercept; the coefficient on m, the only coefficient there is, is the estimate of a in (2). Get the prediction y' from (2).

4. Find a goodness-of-fit measure in the log(y) model that can be compared with an R2 from original model→ get the square of correlation between y and y' from (2), compare this square with R2 from the original model.

Chapter 7 Multiple Regression Analysis with Qualitative Information: Binary variables

I. Dummy independent variables

1. If the regression model is to have different intercepts for g groups, we need to include g-1 dummy variables along with an intercept. The intercept for the base group is the overall intercept in the model, and **the dummy variable coefficient for a particular group represents the estimated difference in intercepts between that group and the base group.**
2. in some cases, the ordinal variable takes on **too many values** so that a dummy variable cannot be included for each value. If we do not wish to put the rank directly in the equation, we can **break it down into categories**.
3. run log(wage) on edu, female and female*edu → the coefficient on female is not significant. Before we add the interaction term, it was highly significant. → The coefficient on female is now estimated much less precisely than it was without interaction term. The reason for this is that female and female*edu are highly correlated in the sample.

II. Testing for differences in Regression functions across groups.

1. **Chow test**: just an F test, **only valid under homoskedasticity, but normality is not needed** for asymptotic analysis.

H0: two groups: g1 and g2: the intercept and all slopes are the same across the two groups.

a. Run OLS for g1 and g2 separately, get SSR1 and SSR2 → SSRur = SSR1 + SSR2.
b. The restricted SSR is just the SSR from pooling the groups and estimating a single equation, say SSRp.
c. $F = \frac{[SSR_p - (SSR_1 + SSR_2)]}{SSR_1 + SSR_2} \bullet \frac{[N - 2(k+1)]}{k+1}$

2. Note: **one important limitation of the Chow test is that the null hypothesis allows for no differences at all between the groups**. → two ways to allow the intercepts to differ under H0.

a. one is to include the group dummy and all interaction terms and then test joint significance of the interaction terms only.
b. Second: to form F statistic just like Chow test, but include a dummy group variable in the restricted pooling regression.

III. A binary dependent variable → **Linear probability model (LPM)**.

Model: $P(y=1|x) = E(y|x) = b_0 + b_1x_1 + b_2x_2 + \ldots + bkx_k$

Limitations:

1. predictions can be either >1 or <0.
2. Heteroskedasticity: $Var(y|x) = p(x)[1-p(x)]$ → except in the case where the probability does not depend on any of the independent variables, **there must be heteroskedasticity in a LPM**. This does not cause bias in the OLS estimators. But homoskedasticity is crucial for justifying the usual t and F statistics, even in large samples.

Even with these problems, the LPM is useful and often applied in economics. It usually works well for values of the independent variables that are near the averages in the sample.

IV. Self-selection

The term is used when a binary indicator of participation might be systematically related unobserved factors. E.g.

$$Y = b0 + b1* \text{participate} + u$$

Where participate is a binary variable equal to 1 if the individual participates in a program. Then we are worried that the average value of u depends on participation: $E(u|participate=1) \neq E(u|participate=0)$ → OLS estimators are biased. The self-seleciton problem is another way that an explanatory variable can be endogenous.

Chapter 8 Heteroskedasticity

I. Consequences of heteroskedasticity:

1. Heteroskedasticity does not cause bias or inconsistency in the OLS estimators, whereas something like omitting an important variable would have this effect.

2. R^2 and adjusted R^2 are also unaffected by the presence of heteroskedasticity because population R^2 is simply 1 - σ_u^2 / σ_y^2. Both variances in the population R^2 are unconditional variances, it is unaffected by the presence of heteroskedasticity in Var(u|x). Further, SSR/n consistently estimates σ^2u, and SST/n consistently estimates σ^2y, whether or not Var(u|x) is constant.
3. However, the usual OLS standard errors are no longer valid for constructing CIs and t statistics. The usual OLS t statistic do not have t distributions in the presence of heteroskedasticity, and the problem is not solved by using large sample sizes. → OLS is no longer BLUE.

II. Heteroskedasticity-robust procedures

1. One regressor:

Var(b'j) = $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sigma_i^2}{SST_x^2}$, The valid estimator of b'j, for heteroskedasticity of any form (including homoskedasticity), is $\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2 u'^2_i}{SST_x^2}$

2. multiple regressors: Var(b'j) = $\frac{\sum_{i=1}^{n} r_{ij}^2 u'^2_i}{SSR_j^2}$, ---------- (1). where rij denotes the ith residual form regressing xj on all other independent variables, and SSRj is the sum of squared residuals from this regression. The square root of the quantity is called the **heteroskedasticity-robust standard error for b'j.**

a. Sometimes, as a degree of freedom correction, (1) is multiplied by n/(n-k-1) before taking the square root. Typically, we use whatever form is computed by the regression package at hand.
b. **Heteroskedasticity-robust t statistic**: t = (OLS estimator – hypothesized value)/ heteroskedasticity-robust standard error.

3. Heteroskedasticity-robust standard errors can be either larger or smaller than the usual standard errors. As an empirical matter, the robust standard errors are often found to be larger than the usual standard errors.
4. **Why do we still use usual standard errors?** – If the homoskedasticity assumption holds and the errors are normally distributed, then the usual t statistics have **exact** t distributions, regardless of the sample size. **The robust standard errors and robust t statistics are justified only as the sample size becomes large**.
5. **In large sample sizes**, we can make a case for **always reporting only the heteroskedasticity-robust standard errors in cross-sectional applications**.

III. A Heteroskedasticity-robust LM statistic:

1. Obtain the residual u' from the restricted model.
2. Regress each of the independent variables excluded under the null on all of the included independent variables; if there are q excluded variables, this leads to q sets of residuals (r1, r2,…, rq).

3. Find the products between each rj and u' (for all observations).
4. Run the regression of **1** on r1u', r2u', …, rqu', **without an intercept**. The heteroskedasticity-robust LM statistic is n – SSR1, where SSR1 is just the usual sum of squared residuals form this final regression. Under H0, LM is distributed approximately as χ_q^2

IV. **Testing for heteroskedasticity**.

1. **Breusch-Pagan test**:

a. Estimate the orginal model by OLS. Obtain the squared residuals, u'^2.
b. Run $u'^2 = a_0 + a_1x_1 + a_2x_2 + \ldots + akx_k$ + error. Keep the R2 from this regression.
c. Form either the F statistic or the LM statistic and compute the p-value (using the $F_{k,n-k-1}$ distribution in the former case and the χ_k^2 distribution in the latter case). If the p-value is sufficiently small, → reject H0 of homoskedasticity.

Note: using the **logarithmic functional form for the dependent variable can often reduce heteroskedasticity**.

2. The **White test** for heteroskedasticity: → run OLS of u'^2 on all independent variables, their quadratic forms and all cross products.
3. **A special case of the White test**:

a. Obtain the OLS residuals u' and the fitted values yhat.
b. Run $u'^2 = a_0 + a_1y' + a_2y'^2$ + error. Keep the R2 from this regression
c. Form F or LM statistic.

Note: if A3 is violated – in particular, if the functional form of E(y|x) is misspecified – then a test for heteroskedasticity can reject H0, even if Var(y|x) is constant.

V. **Weighted Least Squares estimation**.

1. If heteroskedasticity is detected, **one possible response** is to use heteroskedasticity-robust statistics after estimation by OLS. **The other response** is to model and estimate the specific form of heteroskedasticity. This leads to a more efficient estimator than OLS, and it produces t and F statistics that have t and F distributions.
2. Heteroskedasticity is known up to a multiplicative constant.
   Model: $\mathrm{Var}(u|x) = \sigma^2 h(x)$

$$y = \beta_0 + \beta_1 x_1 + \ldots + \beta_k x_k + u \quad \text{---------- (1)} \rightarrow$$

$$y_i / \sqrt{h_i} = \beta_0 / \sqrt{h_i} + \beta_1 (x_{i1} / \sqrt{h_i}) + \ldots + \beta_k (x_{ik} / \sqrt{h_i}) + (u_i / \sqrt{h_i})$$

$$y^* = \beta_0 x_0^* + \beta_1 x_1^* + \ldots + \beta_k x_k^* + u^* \quad \text{---------- (2).}$$

a. (2) doesn't have intercept.
b. We must remember to interpret the estimates in light of the original equation (1)
c. R2 from (2), while useful for computing F statistic, is not especially informative as a goodness-of-fit measure: it tells us how much variation in y* is explained by x*j, and this is seldom very meaningful.

3. Special case: we rarely know how the variance depends on a particular independent variable in a simple form. But if the equation at the individual level satisfies the homoskedasticty assumption, then the firm-level must have heteroskedasticity. In

fact, if Var($u_{i,e}$) = σ^2 for all I and e, then $Var(\bar{u}_i) = \sigma^2 / m_i$, where mi is the number of workers in firm i.

VI. The heteroskedasticity function must be estimated: **FGLS** → estimate the function h(xi).

Var(u|x) = $\sigma^2 \exp(b_0 + b_1 x_1 + ... + b_k x_k)$.

1. Run the regression of y on x1, x2, …, xk and obtain the residuals, u'.
2. Create log(u'^{2}).
3. Run the regression of log(u'^{2}) on x1, …, xk and obtain the fitted values, g'.
4. let h' = exp(g').
5. Estimate the equation $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$, by WLS, using weights 1/h'.

Note:

- Having to estimate hi using the same data means that the **FGLS estimator is no longer unbiased**. Nevertheless, the **FGLS estimator is consistent and asymptotically more efficient than OLS**. At any rate, for large sample size, FGLS is an attractive alternative to OLS when there is evidence of heteroskedasticity that inflates the standard errors of the OLS estimates.
- While computing F statistic, it is important that the same weights be used to estimate the unrestricted and restricted models.
- OLS and WLS estimates can be substantially different. When this happens, typically, this indicates that one of the other Gauss-Markov assumptions is false, particularly the zero conditional mean assumption on the error. The **Hauseman test can be used to formally compare the OLS and WLS estimates to see if they differ by more than the sampling error suggests**.

Chapter 9 More on Specification and Data problems

I. Functional form misspecification

1. Definition: the omitting variable is a function of an explanatory variable in the model.
2. Ramsey's **Regression specification error test (RESET)**:

Original model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u$, if it satisfies A3, then no nonlinear functions of the independent variables should be significant when added to equation.

Testing model: $y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \delta_1 y'^2 + \delta_2 y'^3 + error$

RESET is the F statistic for testing H0: δ 1= δ 2=0.

Note:

- RESET provides no real direction on how to proceed if the model is rejected.
- It can be shown that RESET has no power for detecting omitted variables whenever they have expectations that are linear in the included independent variables in the model. Further, if the functional form in properly specified, RESET has no power for

detecting heteroskedasticity. The bottom line is that **RESET is a functional form test, and nothing more.**

3. Tests against nonnested alternatives:

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + u$ ---------- (1) vs.

$y = \beta_0 + \beta_1 \log(x_1) + \beta_2 \log(x_2) + u$.---------- (2)

Two ways:

a. $y = b_0 + b_1 x_1 + b_2 x_2 + b_3 \log(x_1) + b_4 \log(x_2) + u$, can first test H0: b1=b2=0, can also test H0: b3=b4=0.

b. **Davidson – MacKinnon** test: if (1) is true, then the fitted values from (2), should be insignificant in (1). T-test in the equation $y = b_0 + b_1 x_1 + b_2 x_2 + \theta y'' + error$,w here y'' is the fitted values from (2).

II. Lagged dependent variables as proxy variables.

1. Using a lagged dependent variable can account for historical factors that cause current differences in the dependent variable that are difficult to account for in other ways.
2. Using lagged dependent variables as a general way of controlling for unobserved variables is hardly perfect. But it can aid in getting a better estimate of the effects of policy variables on various outcomes.

III. Measurement error:

1. **Measurement error in the dependent variables**: y*: true. y: observed. e = y – y*.

$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + u + e$, Assume e is independent of each explanatory variable.

→ **OLS estimators are unbiased and consistent**.

→ Var(u+e) = $\sigma_u^2 + \sigma_e^2 > \sigma_u^2$ → **larger error variance than when no error occurs**.

Bottom line: if the measurement error in the dependent variable is systematically related to one or more of the regressors, the OLS estimators are biased. If it is just a random reporting error that is independent of the explanatory variables, as is often assumed, then OLS is perfectly appropriate.

2. Measurement error in the independent variables: x*: true. x: observed. e = x – x*.

a. cov(e,x) = 0 → y =b0 + b1x* + u = b0 + b1x + (u-b1*e). → OLS is consistent with larger variance.

b. **Classical errors-in-variables (CEV): cov(e,x*) = 0** →cov(e, x) = var(e), cov(x, u-b1*e) = -b1Var(e) → OLS is biased and inconsistent estimator.

c. **Under CEV: plim(b'1) =b1*(** $\frac{\sigma_{r_1^*}^2}{\sigma_{r_1^*}^2 + \sigma_e^2}$), where r1* is the population error in the equation x1* = a0 + a1x2 +… + akxk + r1* → plim(b'1) is also closer to zero than is b1. This is called the **attenuation bias** in OLS due to classical errors-in-variables: on average (or in large samples), the estimated OLS effect will be attenuated.

d. There are less clear-cut for estimating the bj on the variables not measured with error.

IV. **Nonrandom samples**:

1. **Exogenous sample selection**: sample selection based on the independent variables → no problem.

2. **Endogenous sample selection**: sample selection based on the dependent variable → bias always occurs in OLS. Because E(y>constant|x) ≠E(y|x).
3. Stratified sampling: the population is divided into nonoverlapping, exhaustive groups or strata. Then, some groups are sampled more frequently than is dictated by their population representation, and some groups are sampled less frequently. E.g, some surveys purposely oversample minority groups or low-income groups. Whether special methods are needed hinges on whether the stratification is exogenous or endogenous (based on the dependent variable).

V. **Least absolute deviations (LAD)**: minimizes the sum of the absolute deviation of the residuals, rather than the sum of squared residuals.

1. The estimates got by LAD are resilient to outlying observations.
2. **Drawbacks**:

a. there are no formulas for the estimators; they can only be found by using iterative methods on a computer.
b. All statistical inference involving LAD estimators is justified only asymptotically. With OLS, under A1-A6, t, F statistics have exact t and F distributions.
c. LAD does not always consistently estimate the parameters appearing in the conditional mean function, E(y|x). LAD is intended to estimate the effects on the conditional median. Generally, mean and median are the same only when the distribution of y given the covariates x1, …, xk is symmetric about $b_0 + b_1x_1 + b_2x_2 + \dots + bkx_k$. OLS produces unbiased and consistent estimators in the conditional mean whether or not the error distribution is symmetric.
d. If u is independent of (x1,…xk), the OLS and LAD slopes estimates should differ only by sampling error whether or not the distribution of u is symmetric.
e. Therefore, either the distribution of u given (x1,…xk) has to be symmetric about zero, or u must be independent of (x1,…xk), in order for LAD to consistently estimate the conditional mean parameters.

Part 2 Regression Analysis with Time Series Data

Chapter 10 Basic Regression analysis with Time Series Data

I. **Finite Distributed Lag Models** – allow one or more variables to affect y with a lag.

$$y_t = a + b_0 x_t + b_1 x_{t-1} + \dots b_q x_{t-q} + u_t$$

1. The **impact propensity** is the coefficient on the contemporaneous x, b0.
2. The **long-run propensity** is the sum of all coefficients on the variables x: LRP = b0 + b1 +…+bq
3. Due to multicollinearity among x and its lags, it can be difficult to obtain precise estimates of the individual bj. However, **even when the bj cannot be precisely estimated, we can often get good estimates of the LRP**.

Note: if both y and x are in logarithmic form, then bo is called the short-run elasticity and b0 +…+bq is called the long-run elasticity.

II. **Finite Sample** Properties of OLS under Classical Assumptions.

1. **TS.A1: Linear in parameters**:

The stochastic process {(xt1, …, xtk, yt): t=1,2,…,n} follows the linear model:

$y_t = \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + u_t$

2. **TS.A2: Zero conditional mean** – $E(u_t|\mathbf{X})=0$, t=1,2,…,n → Explanatory variables are **strictly exogenous**.

a. The error at time t, ut, is uncorrelated with each explanatory variable in every time period.
b. If ut is independent of X and E(ut) =0, then TS.A2 automatically holds.
c. If $E(u_t|\mathbf{x_t})=0$, we say that the x_{tj} are **contemporaneously exogenous**. → Sufficient for consistency, but strict exogeneity needed for unbiasedness.
d. TS.A2 puts no restriction on correlation in the independent variables or in the u_t across time. It only says that the average value of u_t is unrelated to the independent variables in all time periods.
e. If TS.A2 fails, omitted variables and measurement error in regressors are two leading candidates for failure.
f. Other reasons: X can have no lagged effect on y. if X does have a lagged effect on y, then we should estimate a distributed lag model. A2 also excludes the possibility that changes in the error term today can cause future changes in X. e.g. higher ut → higher yt → higher Xt+1. Explanatory variables that are strictly exogenous cannot react to what has happened to y in the past.

3. **TS.A3: no perfect collinearity**.
4. **THM 10.1 – Unbiasedness of OLS:** Under A1-A3, the OLS estimators are unbiased conditional on **X**, and therefore unconditionally as well: E(b'j) = bj, j = 0,1,…,k.
5. **TS.A4: Homoskedasticity:** $Var(u_t|\mathbf{X})=Var(ut)$, t=1,2,…,k
6. **TS.A5: No serial correlation**: $Corr(u_t, u_s|\mathbf{X})=0$, for all $t \neq s$. → A5 assumes nothing about temporal correlation in the independent variable.
7. **THM 10.2 – (OLS Sampling Variance)**: Under the TS Gauss-Markov A1-A5, the variance of b'j, conditional on X, is $Var(b'j|X)=\frac{\sigma^2}{SST_j(1-R_j^2)}$, j =1,…,k, where SSTj is the total sum of squares of Xtj and R^2j is the R^2 from the regression of Xj on the other independent variables.
8. **THM 10.3 – Unbiased estimation of σ^2**: under A1-5, the estimator σ'2=SSR/df is an unbiased estimator of σ^2, where df = n – k -1
9. **THM 10.4 – Gauss-Markov Theorem:** Under A1-5, OLS estimators are the BLUE.
10. **TS.A6**: ut are independent of X and are iid ~ N(0, σ^2).
11. **THM 10.5 – Normal sampling distributions:** under A1-6, the CLM assumptions for time series, the OLS estimators are normally distributed, conditional on X. Further, under the null hypothesis, each t statistic has a t distribution, and each F statistic has an F distribution. The usual construction of confidence intervals is also valid.

III. Trends and seasonality

1. Phenomenon: in many cases, two time series appear to be correlated only because they are both trending over time for reasons related to other unobserved factors – **spurious regression**.

2. **Linear time trend**: yt = a0 + a1*t + et, t=1,2,…
3. Many economic time series are better approximated by an **exponential trend**, which follows when a series has the same average growth rate from period to period.

Log(yt) = a0 + a1*t + et, t=1,2,…

4. Accounting for explained or explanatory variables that are trending is straightforward **– just put a time trend in the regression.**

$y_t = \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + \alpha t + u_t$ ---------- (1)

5. In some cases, adding a time trend can make a key explanatory variable more significant. This can happen if the dependent and independent variables have different kinds of trends (say, one upward and one downward), but movement in the independent variable about its trend line causes movement in the dependent variable away from its trend line.
6. Interpretations: betas from (1) are just the same betas from:

a. linearly detrend y, all x (residual from y = a +bt + u)
b. run linearly detrended y on all linearly detrended x.
c. This interpretation of betas shows that **it is a good idea to include a trend in the regression if any independent variable is trending, even if y is not.** If y has no noticeable trend, but, say, x is growing over time, then excluding a trend from the regression may make it look as if x has no effect on y, even though movements of x about its trend may affect y.

7. The usual and adjusted R^2s for time series regressions can be artificially high when the dependent variable is trending. **We can get R^2 from the regression: time detrended y on x and t.**
8. For seasonality: just include a set of seasonal dummy variables to account for seasonality in the dependent variable, the independent variables, or both.

Chapter 11 Further Issues in Using OLS with Time Series Data

I. Stationary and weak dependence

1. **Stationary stochastic process**: $\{x_t: t = 1,2,\ldots\}$ is stationary if for every collection of time indices $1 \leqq t_1 < t_2 < \ldots < t_m$, the joint distribution of $(x_{t1}, x_{t2}, \ldots, x_{tm})$ is the same as the joint distribution of $(x_{t1+h}, x_{t2+}h, \ldots, x_{tm+h})$ for all integers $h \geqq 1$

a. The sequence $\{x_t: t = 1,2,\ldots\}$ is identically distributed, when m=1 and t1=1.
b. **A process with a time trend is clearly nonstationary**: at a minimum, its mean changes over time.

2. C**ovariance stationary process**: E(xt), Var(xt) are constant and for any t, $h \geqq 1$, $Cov(x_t, x_{t+h})$ depends only on h and not on t.

→ if a stationary process has a finite second moment, then it must be covariance stationary, but the converse is not true.

→ **Stationarity simplifies statements of the LLN and CLT**. If we allow the relationship between two variables (say, y_t and x_t) to change arbitrarily in each time period, then we cannot hope to learn much about how a change in one variable affects the other variable if we only have access to a single time series realization.

3. **Weakly dependent**: a stationary time series process is said to be weakly dependent if x_t and x_{t+h} are "almost independent" as h increases without bound. A similar

statement holds true is the sequence is nonstationary, but then we must assume that the concept of being almost independent does not depend on the starting point, t.

- Covariance stationary process is weakly dependent if the correlation between xt and xt+h goes to zero "sufficiently quickly" as h→ ∞
- Covariance stationary sequences where Corr(xt, xt+h) → 0, as h→∞ are said to be **asymptotically uncorrelated**.
- Why is weak dependence important? → **essentially, it replaces the assumption of random sampling in implying that the LLN and CLT hold**.
- **A trending series, while certainly nonstationary, can be weakly dependent**. A series that is stationary about its time trend, as well as weakly dependent, is often called a trend-stationary process. Such process can be used in regression analysis, provided appropriate time trends are included in the model.

II. Asymptotic Properties of OLS

1. **TS.A1'**: the same TS.A1, but we add the assumption that {(xt, yt)} is stationary and weakly dependent. In particular, the LLN & CLT can be applied to sample averages.

- But stationarity is not at all critical for OLS to have its standard asymptotic properties. The important extra restriction in TS.A1' is the weak dependence assumption.

2. **TS.A2'**: Zero conditional mean: the regressors are **contemporaneously exogenous**. $E(u_t|\mathbf{x}_t)=0$.

- this is much weaker than TS.A2 because it puts no restrictions on how u_t is related to the explanatory variables in other time periods.
- By stationaity, if contemporaneous exogeneity holds for one time period, it holds for them all. Relaxing stationarity would simply require us to assume the condition holds for all t = 1, 2, …
- Unlike A2: A2' doesn't exclude the possibility that changes in the error term today can cause future changes in X.

3. **TS.A3'** = TS.A3 (no perfect collinearity)
4. **THM 11.1 (consistency of OLS)**: under TS.A1' – TS.A3', the OLS estimators are consistent: plimb'j = bj, j = 0, 1, .., k.

- Difference between Thm 11.1 and Thm 10.1: In Thm 11.1, we conclude that the OLS estimators are consistent, but not necessarily unbiased. Second, in Thm 11.1, we have weakened the sense in which X must be exogenous, but weak dependence is required in the underlying time series.

5. TS.A4': homoskedasticity: the errors are contemporaneously homoskedastic, that is, $Var(u_t|\mathbf{xt})= \sigma^2$, t=1,2,…,k
6. TS.A5': no serial correlation: $Corr(u_t, u_s|\mathbf{x_t, x_s})=0$, for all $t \neq s$
7. **THM 11.2 – Asymptotic normality of OLS:** under A1'-5', the OLS estimators are asymptotically normally distributed. Further, the usual OLS standard errors, t statistics, F statistics and LM statistics are asymptotically valid.

- Under A1'-5', OLS is asymptotically efficient.
- Models with trending explanatory variables can effectively satisfy A1'5', provided they are trend stationary. As long as time trends are included in the equations when needed, the usual inference procedures are asymptotically valid.

III. Using **highly persistent (or strong dependent) Time series**

1. The previous section shows that, provided the time series are weakly dependent, usual OLS inference procedures are valid under assumptions weaker than the CLM assumptions.
2. Random walk: yt = yt-1 + et, where, et is iid with zero mean and variance σ^2. Assume the initial value, y0, is independent of et, for any t.

- yt = et + et-1 + … + e1 + y0 → **E(yt) = E(y0) → does not depend on t**.
- **Var(yt) = σ^2t**
- A random walk displays highly persistent behavior because the value of y today is significant for determining the value of y in the very distant future.
  Yt+h = et+h + et+h-1 +…+ et+1 + yt → $E(y_{t+h}|y_t)$ = yt, for all h≧1
- Corr(yt, yt+h) = sqrt(t/(t+h)). The correlation depends on the starting point, t. Therefore, **a random walk does not satisfy the requirement of an asymptotically uncorrelated sequence**.

3. A random walk is a special case of a unit root process: yt = yt-1 + et, where, et is allowed to be general, weakly dependent series. {et} can follow an MA or AR process. When {et} is not an iid sequence, the properties of the random walk no longer hold. But the key feature of {yt} is preserved: the value of y today is highly correlated with y even in the distant future.
4. **Not to confuse trending and highly persistent behaviors**: a series can be trending but not highly persistent. Factors such as interest rates, inflation rates, and unemployment rates are thought to be highly persistent, but they have no obvious upward or downward trend. However, it is often the case that a highly persistent series also contains a clear trend. One model that leads to this behavior is the **random walk with drift**.

- yt = a + yt-1 + et, if y0=0 then E(yt)=at, $E(y_{t+h}|y_t)$ =ah+ yt

5. Transformations on highly persistent time series:

a. Weakly dependent processes are said to be **integrated of order zero, or I(0)**. This means that nothing needs to be done to such series before using them in regression analysis: averages of such sequences already satisfy the standard limit theorems. Unit root processes are said to be **integrated of order one, or I(1).** This means that the first difference of the process is weakly dependent (and often stationary)
b. Many time series yt that are strictly positive are such that log(yt) is I(1).
c. **Differencing time series before using them in regression analysis has another benefit: it removes any linear time trend.** → rather than including a time trend in a regression, we can instead difference those variables that show obvious trends.

IV. Dynamically complete models:

1. $y_t = \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + u_t$, let **xt** = $(x_{1t}, ..., x_{kt})$, may or may not included lags of y or z. if we assume that E(ut|**xt**, yt-1, **xt-1**,…) = 0 or, equivalently E(yt|**xt**, yt-1, **xt-1**,…) = E(yt| **xt**).
2. The above model is dynamically complete model → whatever is in **xt**, enough lags have been included so that further lags of y and the explanatory variables do not matter for explaining yt.
3. Dynamically complete model must satisfy TS.A5'. → errors are serially uncorrelated.

Chapter 12 Serial Correlation and Heteroskedasticity in Time Series Regression

I. Chapter 11 shows that when in an appropriate sense, the dynamics of a model have been completely specified, the errors will not be serially correlated. Thus, testing for serial correlation can be used to detect dynamic misspecification. Furthermore, static and finite distributed lag models often have serially correlated errors even if there is no underlying misspecification of the model.

II. OLS with serially correlated errors.

1. Unbiasedness: **as long as the explanatory variables are strictly exogenous, the b'j are unbiased, regardless of the degree of serial correlation in the errors (Thm 10.1 assumed nothing about serial correlation).** This is analogous to the observation that heteroskedasticity in the errors does not cause bias in the b'j.
2. Efficiency and Inference: **OLS is no longer BLUE. The usual OLS standard errors and test statistics are not valid, even asymptotically**. → if assume error term is AR(1), errors are positively correlated, and x is positively correlated. OLS estimators are underestimated.
3. Goodness-of-fit: **our usual R2 and adjusted R2 are still valid**, provided the data are stationary and weakly dependent.

- The population R2 in a cross-sectional context is $1 - \sigma_u^2 / \sigma_y^2$. This definition is still appropriate in time series with stationary, weakly dependent data: the variances of both the errors and y do not change over time. By LLN, R2 consistently estimate the population R2.
- The argument is essentially the same as in the cross-sectional case, whether or not there is heteroskedasticity.
- Since there is never an unbiased estimator of the population R2, it makes no sense to talk about bias in R2 caused by serial correlation. This argument does not go through if {yt} is an I(1).

4. Serial correlation in the presence of lagged dependent variable:

a. Wrong statement: "OLS is inconsistent in the presence of lagged dependent variables and serially correlated errors". → Assume yt = a + byt-1 + Ut and E(Ut|yt-1) = 0 → satisfies TS.A2' → OLS estimators are consistent. But {ut} can be serially correlated because even though ut is uncorrelated with yt-1, it can be correlated with yt-2 → cov(ut, ut-1) is nonzero. Therefore, the serial correlation in the errors will cause the OLS statistics to be invalid for testing purposes, but it will not affect consistency.

b. But if we assume {ut} follows AR(1) → OLS inconsistent. But in this case, yt can be expressed by AR(2).

c. The bottom line is that you **need a good reason for having both a lagged dependent variable in a model and a particular model of serial correlation in the errors**. Often, serial correlation in the errors of a dynamic model simply indicates that the dynamic regression function has not been completely specified.

III. Testing for serial correlation

1. Case 1: t – test for AR(1) serial correlation with **strictly exogenous regressors**.

a. Run OLS of yt on xt and obtain residuals, ut;
b. Run ut on ut-1, for all t=2,…,n, obtaining the coefficient rho on ut-1 and its t-statistic. (this regression may or may not contain an intercept; the t-statistic for rho will be slightly different, but it is asymptotically valid either way
c. Use t-statistic to test H0: rho=0 in the usual way.

- the test can detect other kinds of serial correlation. Any serial correlation that causes adjacent errors to be correlated can be picked up by this test. However, it does not detect serial correlation where adjacent errors are uncorrelated.
- To use the usual t statistic, we must assume errors in ut=rho*ut-1 + et, satisfy the appropriate homoskedasticity assumption. But it is easy to use heteroskedasticity robust t statistic from chapter 8.

2. Case 2: Durbin-Watson Test under **classical Assumptions**

$$\mathbf{DW} = \frac{\sum_{t=2}^{n}(u_t - u_{t-1})^2}{\sum_{t=2}^{n} u_t^{\,2}}$$

DW ≈ 2(1-rho) → tests based on DW and the t test based on rho are conceptually the same.

- The fact that an exact sampling distribution for DW can be tabulated is the only advantage that Dw has over the t-test. Given that the tabulated critical values are exactly valid only under the full set of CLM assumptions and that they can lead to a wide inconclusive region, the practical disadvantages of the DW statistic are substantial.
- T-statistic is simple to compute and asymptotically valid without normally distributed errors. T-statistic is also valid in the presence of heteroskedasticity that depends on the xt → just use heteroskedasticity adjusted form.

3. Testing for AR(1) serial correlation **without strictly exogenous regressors**.

- When xt are not strictly exogenous, neither t test nor the DW are valid, even in large samples.
    a. Run OLS of yt on $x_{1t},...,x_{kt}$ and obtain the residuals, ut for all t.
    b. Run OLS of ut on $x_{1t},...,x_{kt}$, ut-1, for all t=2,…,n to get coefficient rho on ut-1 and its t-statistic (if there exists heteroskedasticity, use heteroskedasticity robust t statistic).
    c. Use t-rho to test H0: rho=0.

4. Testing for AR(q) serial correlation **without strictly exogenous regressors**.
    a. Run OLS of yt on $x_{1t},...,x_{kt}$ and obtain the residuals, ut for all t.
    b. Run OLS of ut on $x_{1t},...,x_{kt}$, ut-1, ut-2,…, ut-q, for all t=q+1,…,n
    c. Compute F test for joint significance of coefficients on ut-1,…,ut-q (if there exists heteroskedasticity, use heteroskedasticity robust F statistic).

- can also use Lagrange multiplier (LM) statistic – **Breusch-Godfrey test**:
    LM = (n-q)R2 from step b.

IV. Correcting for serial correlation with **strictly exogenous regressors (→ at a minimum, we should not use these corrections when the explanatory variables include lagged dependent variables)**

$$u_t = \rho u_{t-1} + e_t$$

$$\text{var}(u_t) = \sigma_e^2/(1-\rho^2)$$

Model: $y_t - \rho y_{t-1} = (1-\rho)\beta_0 + \beta_1(x_t - \rho x_{t-1}) + e_t, t \geq 2$

$$(1-\rho^2)^{1/2} y_1 = (1-\rho^2)^{1/2}\beta_0 + \beta_1(1-\rho^2)^{1/2} + (1-\rho^2)^{1/2} u_1$$

This gives the BLUE estimators under TS.A1-TS.A4 and the AR(1) for ut.

1. GLS estimators are BLUE, t and F statistics from the transformed equation are valid.
2. **FGLS**:

a. Run OLS of yt on xt and obtain residuals ut, t=1,…,n
b. Run ut on ut-1 and obtain rho.
c. Apply OLS to the above model, the usual standard errors, t and F statistics are asymptotically valid (no constant, x11=$(1-\rho^2)^{1/2}$, xt1=$(1-\rho)$.

3. FGLS:

- the cost of using estimated rho in place of real rho is that **FGLS estimators has no tractable finite sample properties**. It is not unbiased, although it is consistent when the data are weakly dependent.
- It is asymptotically more efficient than the OLS estimator when the AR(1) model for serial correlation holds (and the explanatory variables are strictly exogenous).
- **Cochrane-Orcutt** estimation: omits the first observation.
- **Prais-Winsten (PW)** estimation: uses the first observation.

4. weaker assumptions:

a. For OLS: cov(xt,ut)=0 → consistent.
b. For FGLS: cov(xt,ut)=0 + cov[(xt-1 + xt+1)] =0 → consistent. If the latter fails, OLS is still consistent, but FGLS is not → OLS is preferred to FGLS.
c. Consistency and asymptotic normality of OLS and FGLS **rely heavily on the time series processes yt and xt being weakly dependent**. Strange things can happen if we apply either of them when process has unit roots.

V. Serial correlation-robust inference after OLS

1. Recently, it's been more popular to estimate models by OLS but to correct the standard errors for fairly arbitrary forms of serial correlation (and heteroskedasticity). Why, even knowing OLS is inefficient?

- xt may not be strictly exogenous. In this case, FGLS is not even consistent, let alone efficient.
- In most applications of FGLS, the errors are assumed to follow an AR(1) model. It may be better to compute standard errors that are robust to more general forms of serial correlation.

2. Serial correlation-robust standard error for beta1:

a. estimate $y_t = \beta_0 + \beta_1 x_{1t} + \ldots + \beta_k x_{kt} + u_t$ by OLS to get "se(b'1)", σ and residuals ut.

b. Compute the residuals rt from the auxiliary regression:
$x_{t1} = \delta_0 + \delta_2 x_{t2} + ... + \delta_k x_{tk} + r_t$, then form at = rt*ut for each t

c. For your choice of g, compute v as (for annual data, choosing a small g, such as g=1 or 2. Newey and West (1987) recommend taking g to be the integer part of $4(n/100)^{2/9}$.):

$$\nu = \sum_{t=1}^{n} a_t^2 + 2\sum_{t=1}^{n}[1 - h/(g+1)](\sum_{t=h+1}^{n} a_t a_{t-h})$$

d. Compute se(b'1) from: $se(\beta_1) = [" se(\beta_1)"/\sigma]^2\sqrt{\nu}$

3. Empirically, the SC-robust standard errors are typically larger than the usual OLS when there is serial correlation.

- **Reasons of the relatively unpopularity of SC-robust standard errors:**
  a. SC-robust se can be poorly behaved when there is substantial correlation and the sample size is small.
  b. Have to choose integer g.
  c. In the presence of sever serial correlation, OLS can be very inefficient, especially in small sample sizes. After performing OLS and correcting the se for serial correlation, the coefficients are often insignificant, or at least less significant than they were with the usual OLS se.
- **SC-robust se is most useful** when we have doubts about some of the xj being strictly exogenous, so that methods such as Cochrane-Orcutt are not even consistent. It is valid to use the SC-robust se in models with lagged dependent variables, assuming, of course, that there is good reason for allowing serial correlation in such models.

VI. Heteroskedasticity in Time series Regressions

1. The presence of heteroskedasticity, while not causing bias or inconsistency, does invalidate the usual se, t and F.
2. Testing for heteroskedasticity: tests in chapter 8 can be applied directly, but:

- **the errors ut should not be serially correlated;** any serial correlation will generally invalidate a test for heteroskedasticity.
- it makes sense to test for serial correlation first, using a heteroskedasticity-robust test if heteroskedasticity is suspected. Then, after sth has been done to correct for serial correlation, we can test for heteroskedasticity.

3. **Autoregressive Conditional Heteroskedasticity**:

a. Assume that the Gause-Markov assumptions hold → OLS are BLUE and even with Var(ut|x) is constant, there are other ways that heteroskedasticity can arise:

$E(u_t^2 \mid u_{t-1}, u_{t-2}, ...) = E(u_t^2 \mid u_{t-1}) = \alpha_0 + \alpha_1 u_{t-1}^2$ → ARCH(1) ----------(1)

Since conditional variances must be positive, this model only makes sense if alpha0>0 and alpha1 ≧ 0.

b. It is instructive to write (1) as $u_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \nu_t$, where E(vt|ut-1,…)=0 and vt are not independent of past ut because of the constraint that $\nu_t \geq -\alpha_0 - \alpha_1 u_{t-1}^2$

c. If OLS still has desirable properties under ARCH, why should we care about ARCH forms of heteroskedasticity in static and distributed lag models?

- it is possible to get consistent (but not unbiased) estimators of the bj that are asymptotically more efficient than the OLS estimators.

VII. Both Heteroskedasticity and serial correlation

$$y_t = \beta_0 + \beta_1 x_{1t} + ... + \beta_k x_{kt} + u_t$$

Model: $u_t = \sqrt{h_t}\, v_t$

$$v_t = \rho v_{t-1} + e_t$$

FGLS:

1. Estimate the model by OLS and save the residuals, ut.
2. Regress log(ut2) on x1,…xk and obtain the fitted values, say gt.
3. obtain the estimates of ht: ht=exp(gt).
4. estimate the transformed equation:
   $$\sqrt{h_t}\, y_t = \sqrt{h_t}\,\beta_0 + \beta_1 \sqrt{h_t}\, x_{1t} + ... + \beta_k \sqrt{h_t}\, x_{kt} + error$$
   by standard Cochrane-Orcutt or Prais-Winsten methods.
5. These FGLS estimators are asymptotically efficient. All standard errors and test statistics from the CO or PW methods are asymptotically valid.

Part 3 Advanced Topics

Chapter 13 Pooling Cross Sections across Time. Simple Panel Data Methods

I. <u>Independent pooled cross section</u>

1. Definition: Obtained by sampling randomly from a large population at different points in time.
2. **<u>Chow test (simply an F test)</u>** – can be used to determine whether a regression differs across two groups or two time periods. **<u>Two forms</u>**:

a. Obtains SSR from the pooled estimation as the restricted SSR. The unrestricted SSR is the sum of the SSRs for the two separately estimated time periods.
b. Interacting each variable with a year dummy for one of the two years and testing for joint significance of the year dummy and all of the interaction terms.
c. Chow test can be computed for more than two time periods. It is usually more interesting to allow the intercepts to change over time and then test whether the slope coefficients have changed over time. We can test the constancy of slope coefficients generally by interacting all of the time period dummies with one, several, or all of the explanatory variables and test the joint significance of the interaction terms.

3. a **<u>natural experiment (or a quasi-experiment)</u>** occurs when some exogenous event – often a change in government policy – changes the environment in which individuals, families or firms operate.

a. A natural experiment always has a control group, which is not affected by the policy change and a treatment group.

b. In order to control for systematic differences between the **control and treatment** groups, we need two years of data → sample is usefully broken down into four groups: the control/treatment group before/after the change.
c. Let A: control group. B: treatment group. 1: before change. 2: after change.
y = b0 + b1*d2 + b2*dB + b3*d2*dB + other factors:
- without other factors in the regression, b3 id the **difference-in-differences** estimator: b3 = (avg $y_{2,B}$ – avg $y_{2,A}$) – (avg $y_{1,B}$ – avg $y_{1,A}$)
- with other regressors, b3 no longer has the simple form of the above, but its interpretation is similar.

II. **Two period Panel Data analysis**

1. Panel data can **help reduce omitted variable problems**.
2. $y_{it} = \beta_0 + \delta_0 d2_t + \beta_1 x_{it} + a_i + u_{it}$, we can use the **unobserved factors** affecting the dependent variable as consisting of two types: **constant and vary over time**.
- in this model, we allow the intercept to change over time, which is important in most applications.
- ai captures all unobserved, time-constant factors that affect yit. It is called an unobserved effect or **fixed effect or unobserved heterogeneity**.
- this model is called fixed effect model.
- the error uit is called the **idiosyncratic error or time-varying error**.
- the main reason for collecting panel data is to allow for the unobserved effect, ai, to be correlated with the explanatory variables.
3. **First-differenced equation**: $\Delta y_i = \delta_0 + \beta_1 \Delta x_i + \Delta u_i$ → a single cross-sectional equation, but each variable, including dummies, is differenced over time.
- the most important assumption is that Δu_i **is uncorrelated with the explanatory** variable in both time periods. This is another version of the strict exogeneity assumption in chapter 10 for time series models. In particular, this assumption rules out the case where xit is the lagged dependent variable.
- second crucial assumption is that Δx_i **must have some variation across i.**
- Costs: differencing can greatly reduce the variation in x → larger OLS standard errors.

III. Differencing with more than two time periods

1. the model: $\Delta y_i = \alpha_0 + \alpha_3 d3_t ... + \alpha_T dT_t + \beta_1 \Delta x_{it1} + ... + \beta_k \Delta x_{itk} + \Delta u_{it}$, t=2,3,…,T, where we have T-1 time periods on each unit I for the first-differenced equation. The total number of observations is N(T-1).
- In addition to homoskedasticity, we must assume that Δu_i **is uncorrelated over time** for the usual standard errors and test statistics to be valid.
2. **First-differencing Assumptions**:
a. FD.1: linear model.
b. FD.2: a random sample from the cross section at t=1.
c. FD.3: let **Xi** denote the regressors for all time periods for cross-sectional observation i. For each t, the expected value of the idiosyncratic error given the

explanatory variables in all time periods and the unobserved effect is zero: E(uit|**Xi**, ai) = 0.When this holds, we often say that the **xitj are strictly exogenous** conditional on the unobserved effect.

d. FD.4: each regressor changes over time (for at least some i), and no perfect linear relationships exist among the regressors.

Under FD.1-4 → FD estimators are unbiased and consistent with a fixed T and as N→infinity.

e. FD.5 (homoskedastic): the variance of the differenced errors, conditional on Xi, is constant:
$Var(\Delta u_{it} \mid X_i) = \sigma^2$, t=2,…T

f. FD.6 (serially uncorrelated): for all t≠s, the differences in the idiosyncratic errors are uncorrelated (conditional on all regressors): $Cov(\Delta u_{it}, \Delta u_{is} \mid X_i) = 0$, t≠s.

Under FD.1-6 → FD estimators are the best linear unbiased estimators.

g. FD.7: conditional on Xi, the Δu_{it} are iid normal random variables.

Under FD.1-7 → FD estimators are normally distributed, and the t and F test have exact distributions.

Without 7 → rely on the usual asymptotic approximations.

Chapter 14 Advanced Panel Data Methods

I. Fixed effects estimation

1. The model:
$$y_{it} = \beta_0 + \beta_1 x_{it1} + \ldots + \beta_k x_{itk} + a_i + u_{it}, ---(1)$$
$$\overline{y_i} = \beta_1 \overline{x_{i1}} + \ldots + \beta_k \overline{x_{ik}} + a_i + \overline{u_i}, ---(2)$$
$$\ddot{y}_{it} = \beta_1 \ddot{x}_{it1} + \ldots + \beta_k \ddot{x}_{itk} + \ddot{u}_{it}, ---(3)$$

2. Data in (3) are the **time-demeaned data**. The fixed effects transformation is also called the **within transformation**. We should estimate (3) be pooled OLS (**no intercept in (3)**). A pooled OLS estimator that is based on the time-demeaned variable is called the **fixed effects estimator or the within estimator**.

- df = NT – k – N(because of demeaning).

3. **Assumptions for fixed effects**:

a. FE.1: the model in (1).

b. FE.2: have a random sample in the cross-sectional dimension

c. FE.3: let **Xi** denote the regressors for **all** time periods for cross-sectional observation i. For each t, the expected value of the idiosyncratic error given the explanatory variables in **all** time periods and the unobserved effect is zero: E(uit|**Xi**, ai) = 0.

d. FE.4: each regressor changes over time (for at least some i), and no perfect linear relationships exist among the regressors.

Under FE.1-4 → FE estimators are unbiased and consistent with a fixed T and as N→infinity. Those assumptions are identical to FD.1-4.

e. FE.5: $Var(u_{it} \mid X_i, a_i) = \sigma_u^2$, t=2,…T

f. FE.6 (serially uncorrelated): for all t≠s, the idiosyncratic errors are uncorrelated (conditional on all regressors): $Cov(u_{it}, u_{is} \mid X_i, a_i) = 0$, t≠s.

Under FE.1-6 → FE estimators are the best linear unbiased estimators. Since the FD estimator is linear and unbiased, it is necessarily worse than the FE estimator. The assumption that makes FE better than FD is FE.6, which implies that the idiosyncratic errors are serially uncorrelated.

g. FD.7: conditional on Xi and ai, the uit are iid normal random variables.

Under FD.1-7 → FD estimators are normally distributed, and the t and F test have exact distributions.

Without 7 → rely on the usual asymptotic approximations.

4. Although **time-constant variables cannot be included by themselves in a fixed effects model, they can be interacted with variables that change over time** and, in particular, with year dummies.
5. **Dummy variable regression**: it gives us exactly the same estimates that we would obtain from the regression on time-demeaned data, and the standard errors and other major statistics are identical. One benefit is that is properly computes the df directly (different from (3)).
6. **FE or FD**?

a. When T=2, they are the same.
b. When T>2, they are different. Both are unbiased under FE.1-4.
c. When the uit are serially uncorrelated, FE is more efficient than FD. Since the unobserved effects model is typically stated with serially uncorrelated idiosyncratic errors, the FE estimator is used more than the FD estimator. But if uit is serially correlated, the FD may be better.
d. When T is large and especially when N is not very large, we must be cautious about FE. FD has the advantage of turning an integrated time series process into a weakly dependent process. Therefore, if we apply FD, we can appeal to the CLT even in the cases where T is larger than N.

II. Random Effects Models

1. The model: $y_{it} = \beta_0 + \beta_1 x_{it1} + ... + \beta_k x_{itk} + a_i + u_{it}$, unobserved effect ai is uncorrelated with each regressor at any time period.
2. βj can be consistently estimated by using a single cross section. But this disregards much useful info in the other time periods. If we run OLS using a **pooled OLS**, we can also get consistent estimators. But it ignores a key feature of the model: **the composite error term is serially correlated (vit = ai + uit)**.
3. therefore, the best way is to use FGLS:

$$\lambda = 1 - [\sigma_u^2 / (\sigma_u^2 + T\sigma_a^2)]^{1/2}$$

The transformed equation is:

$$y_{it} - \lambda \overline{y_i} = \beta_0 (1-\lambda) + \beta_1 (x_{it1} - \lambda \overline{x_{i1}}) + ... + \beta_k (x_{itk} - \lambda \overline{x_{ik}}) + (v_{it} - \lambda \overline{v_i})$$

This involves **quasi-demeaned** data on each variable. → Allows for explanatory variables that are constant over time.

4. Assumptions: **RE.1 = FE.1, RE.2 = FE.2**

RE.3 = FE.3+, the expected value of ai given all explanatory variables is constant: E(ai|Xi) = 0. → this rules out correlation between the unobserved effect and regressors, and **it is the key distinction between fixed effects and random effects**. **Because of this, we can include time-constant regressors**.
RE.4: no perfect linear relationships among regressors.
RE.5 = FE.5 + the variance of ai, given all regressors is constant: $Var(a_i \mid X_i) = \sigma_a^2$
RE.6 = FE.6.
Under RE.1-4 → Random effects estimator is consistent as N gets large for fixed T. RE estimator is biased unless we know λ, which keeps us from having to estimate it. Under RE.1-6, the RE estimator is approximately normally distributed with large N, and usual t and F statistics obtained from the quasi-demeaned regression are valid with large N.

5. the panel data methods can be used when working with **matched pairs or cluster samples**. FD and FE can eliminate the cluster effect. If the cluster effect is uncorrelated with the regressors, pooled OLS can be used, but the standard errors and test statistics should be adjusted for cluster correlation. Random effects estimation is also a possibility.

Chapter 15 Instrumental Variables Estimation and Two Stage Least Squares

I. Motivation: omitted variables in a regression model
When faced with the prospect of omitted variables bias (or unobserved heterogeneity):

a. Find a proxy → not always possible to find a good proxy.
b. Assume the omitted variable does not change over time and use FE or FD.
c. **Instrumental variables: leaves the unobserved variable in the error term**, but rather than estimating the model by OLS, it uses an estimation method that recognizes the presence of the omitted variable.

II. Requirements:
Y = b0 + b1*x + u and cov(x, u) ≠ 0
1. **IV.1**: an IV for x – z: z is uncorrelated with u: **cov(u,z)=0.** → we cannot generally hope to test this assumption: in the vast majority of cases, we must maintain cov(z,u)=0 by appealing to economic behavior or introspection.
2**. IV.2**: z is correlated with x: **cov(z,x) ≠ 0**
- Differences b/w an IV and a proxy. Suppose ability is unobservable: IQ is correlated with ability, therefore, IQ can be a proxy. But it cannot be an IV for education, since it violates assumption 1. An IV approach may not be necessary at all if a good proxy exists for ability.
3. b1 = cov(z,y)/cov(z,x) and use sample to get b'1, → plim(b'1) = b1. If either one of the assumption fails, IV is not consistent. **In small samples, IV estimator can have a substantial bias, which is one reason why large samples are preferred**.
4. **IV.3**: homoskedasticity: **$E(u^2|z) = \sigma^2 = var(u)$**

Under IV.1-3 → the asymptotic variance of b'1 is $\frac{\sigma^2}{n\sigma_x^2\rho_{x,z}^2}$, estimated asymptotic variance is $\frac{\sigma_{hat}^2}{SST_x R_{x,z}^2}$, the estimated asymptotic variance of OLS is $\frac{\sigma_{hat}^2}{SST_x}$. → **2SLS variance is always larger than the OLS variance when OLS Is valid** (under Gauss-Markov assumptions)

- **IV and regressors can be binary variables**.

III. **Properties of IV with a poor IV**.

1. Weak correlation between z and x can have even more serious consequences: the IV estimator can have a large asymptotic bias even if z and u are only moderately correlated.
2. $p\lim(b_{1hat}) = b1 + \frac{corr(z,u)}{corr(z,x)} * \frac{\sigma_u}{\sigma_x}$ → even if corr(z,u) is small, the inconsistency in the IV can be very large if corr(z,x) is also small.
3. We should always check to see if the endogenous explanatory variable is correlated with the IV candidate.
4. Most software compute R^2 after IV estimation, using R^2 = 1-SSR/SST, where SSR is the sum of squared IV residuals. **R^2 from IV estimation be negative because SSR for IV can actually be larger than SST**.
5. **OLS R^2 will always be larger** because OLS minimizes the sum of squared residuals.

IV. IV estimation of the multiple regression model:

1. **Structural equation**: the equation is supposed to measure a causal relationship.

$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$, where z_1 is exogenous and y_2 is endogenous.

2. **Reduced form equation**: we have written an endogenous variable in terms of exogenous variables.

$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + v_2$, where z2 is an IV and $\pi 2 \neq 0$.

V. 2SLS.

1. One endogenous, two excluded exogenous variables (z_2, z_3):

$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + u_1$

$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + v_2$, → $\pi 2 \neq 0$ or $\pi 3 \neq 0$

2. Multicollinearity and 2SLS:

a. 2SLS estimator of β1 can be approximated as: $\frac{\sigma^2}{SST_2(1-R_2^2)}$, where, σ2 =var(u1), SST_2 is the total variation in y_2 hat, and R^2 is from a regression of y_2 hat on all other exogenous variables appearing in the structural equation.

b. Two reasons why the variance of 2SLS estimator is larger than for OLS:

- y2 hat has less variation than y2, because total sum of squares > explained sum of squares.
- The correlation b/w y2hat and the exogenous variables is often much higher than the correlation between y2 and these variables.

3. **Order condition** for identification of an equation: need at least as many excluded exogenous variables as there are included endogenous explanatory variables in the structural equation. The sufficient condition for identification is called – **rank condition.**
4. IV and errors-in-variables problems:

a. the model: $y = \beta_0 + \beta_1 x_1^* + \beta_2 z_2 + u$, $x_1 = x_1^* + e_1$ → correlation b/w x1 and e1 causes OLS to be biased and inconsistent. If the classical errors-in-variables (CEV) assumptions hold, the bias in the OLS estimator of β1 is toward zero.

b. If we assume u is uncorrelated with x1, x1*, and x2; e1 is uncorrelated with x1* and x2 → x2 is exogenous and x1 is endogenous → need an IV for x1. Such an IV must be correlated with x1, uncorrelated with u and e1.

5. **Testing for endogeneity**.

2SLS is less efficient than OLS when x are exogenous because 2SLS estimates can have very large se.

Model:
$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + u_1, ---(1)$$
$$y_2 = \pi_0 + \pi_1 z_1 + \pi_2 z_2 + \pi_3 z_3 + \pi_4 z_4 + v_2, -(2)$$

a. Hauseman (1978) suggested directly comparing the OLS and 2SLS estimates and determining whether the differences are significant. After all, both OLS and 2SLS are consistent if all variables are exogenous. If 2SLS and OLS differ significantly, we conclude that y2 must be endogenous.

b. **Regression test**:

- Estimate the reduced form for y2 (2) by regressing it on all exogenous variables (including those in the structural equation and the additional IVs). Obtain the residuals, v2hat. Corr(y2, u1)=0, iff, corr(v2,u1)=0. we can run OLS of u1 on v2, or:
- Add v2hat to the structural equation (which includes y2) and test for significance of v2hat using an OLS. Using a heteroskedasticity-robust t test to test v2hat = 0.

$$y_1 = \beta_0 + \beta_1 y_2 + \beta_2 z_1 + \beta_3 z_2 + \delta v_{2hat} + error, (3)$$

- The estimates on all of the variables (except v2hat) are identical to the 2SLS estimates. → **including v2hat in the OLS clears up the endogeneity of y2**.
- Test for endogeneity of multiple explanatory variables: for each suspected variable, get the reduced form residuals, then test for joint significance of these residuals in the structural equation, using an F test.

6. testing overidentifying restrictions

a. Estimate the structural equation by 2SLS and obtain the 2SLS residuals, u1hat.
b. Regress u1hat on all exogenous variables. Obtain the R2.
c. Under the Ho that all IVs are uncorrelated with u1, $nR^2 \sim \chi_q^2$, where q is # of IVs from outside the model - # of endogenous variables. If nR^2 exceeds the 5%

critical value in the chi-sq distribution, we reject Ho and conclude that at least some of the IVs are not exogenous.

VI. Other issues:

1. 2SLS with heteroskedasiticity – the same as in OLS.
2. applying 2sls to TS equations:

- The mechanics of 2sls are identical for TS or CS data, but for TS the statistical properties of 2sls depend on the trending and correlation properties of the underlying sequences. We must include trends if we have trending dependent or explanatory variables. Since a time trend is exogenous, it can always serve as its own IV.
- Unit root process must be used with care. Often, differencing the equation is warranted before estimation, and this applies to the instruments as well.

Chapter 16 Simultaneous Equations Models

I. **The nature of SEMs**:

1. The most important point is that **each equation in the system should have a ceteris paribus, causal interpretation**.
2. The model (structural equations: labor supply and demand function are derivable from economic theory and have causal interpretations):

Labor supply: $h_i = \alpha_1 w_i + \beta_1 z_{i1} + u_{i1}$

Labor demand: $h_i = \alpha_2 w_i + \beta_2 z_{i2} + u_{i2}$

- Hours and wage are endogenous; z1 and z2 are exogenous. Z1 and z2 are both uncorrelated with the supply and demand errors, u1 and u2.
- Without including z1 and z2 in the model, there is no way to tell which equation is the supply function and which is the demand function.
- Each equation should have a behavioral, ceteris paribus interpretation on its own.

3. Simultaneity bias in OLS: a regressor (w) that is determined simultaneously with the dependent variable is generally correlated with the error term → OLS is biased and inconsistent.

II. Identifying and estimating a structural equation.

1. The model:

Supply: $q = \alpha_1 p + \beta_1 z_1 + u_1$

Demand: $q = \alpha_2 p + u_2$

Demand equation is identified → use z1 as an IV for price in demand equation.

- Intuitively, since we observe z1, that shifts the supply equation while not affecting the demand equation. Given the variation in z1 and no errors, we can trace out the demand curve.
- It is the presence of an exogenous variable in the supply equation that allows us to estimate the demand equation.

2. **Rank condition**: the first equation in a two-equation simultaneous equations model is identified iff the second equation contains at least one exogenous variable (with a nonzero coefficient) that is excluded from the first equation.

III. Estimation by 2SLS: the IVs consist of the exogenous variables appearing in **either** equation.

IV. Systems with more than two equations:

1. Each identified equation can be estimated by 2SLS. The IVs for a particular equation consist of the exogenous variables appearing **anywhere** in the system. Tests for endogeneity, heteroskedasticity, serial correlation, and overidentifying restrictions can be obtained, just as in chapter 15.
2. Any system with two or more equations is correctly specified and certain additional assumptions hold, system estimation methods are generally more efficient than estimating each equation by 2SLS. The most common system estimation method in the context of SEMs is 3SLS.

V. Others:

1. SEMs with TS: Using growth rates of trending or I(1) variables in SEMs is fairly common in TS applications. If a structural model contains a Time trend – which may capture exogenous, trending factors that are not directly modeled – then the trend acts as its own IV.
2. SEMs with panel data: two steps:

- eliminate the unobserved effects from the equations of interest using the fixed effects transformation or first differencing, and
- Find IVs for the endogenous variables in the transformed equation. We need to find IVs that change over time.

Chapter 17 Limited Dependent Variable Models and Sample Selection Corrections

I. Logit and Probit models for binary response.

1. Model:

$P(y=1 \mid x) = G(\beta_0 + \beta_1 x_1 + ... + \beta_k x_k) = G(\beta_0 + x\beta)$, where 0<G(z)<1.

2. **logit model**: G(z) = exp(z)/[1 + exp(z)]

3. **probit model**: G(z) =Φ(z) = $\int_{-\infty}^{z} \phi(v)dv$ -- standard normal density.

Both models increase most quickly at z = 0, and as z → -∞, G(z) → 0, and as z → ∞, G(z) → 1.

4. **Latent variable model**: both models can be derived from an underlying latent variable model. **Let y* be an unobserved, or latent, variable**.

$y^* = \beta_0 + x\beta + e, y = 1[y^* > 0]$

- e is independent of x
- e either has the standard logistic distribution or the standard normal distribution. → e is symmetrically distributed about zero, which means that 1 – G(-z) = G(z).
- P(y=1|x) = P(y*>1|x) =

$P[e > -(\beta_0 + x\beta) \mid x] = 1 - G[-(\beta_0 + x\beta)] = G(\beta_0 + x\beta)$

- When xj is a roughly continuous variable, partial effect on p(x) = P(y=1|x) is: $\frac{\partial p(x)}{\partial x_j} = g(\beta_0 + x\beta)\beta_j$ → the partial effect of xj on p(x) depends on x through the positive quantity g(.), → **partial effect always has the same sign as βj**.
- **Relative effects of any two continuous regressors do not depend on x**: the ratio of the partial effects for xj and xh is βj/ βh
- If xj is a binary variable: then the partial effect is: G(xj=1) – G(xj=0).

5. **MLE**:
   a. MLE is based on the distribution of g given x, the heteroskedasticity in Var(y|x) is automatically accounted for.
   b. **MLE is consistent, asymptotically normal and asymptotically efficient**.
   c. **Testing multiple hypothesis**:

- **Lagrange multiplier**: estimate the restricted model under Ho.
- **Wald test**: estimate unrestricted model – essentially a F test.
- **Likelihood ratio**: use both restricted and unrestricted models.

LR = 2(Lur – Lr) ~ χ_q^2

Lur > Lr (like R2). And for binary response models, **L is always negative**.

6. **Goodness-of-fit measures**:
   percent correctly predicted: for each I, compute the estimated probability that yi takes on the value one, if G(zhat) > 0.5, the prediction of yi =1, if ≤0, yi is predicted to be zero. The percentage of times the predicted yi matches the actual yi is the percent correctly predicted.
   Pseudo R2 (1): McFadden (1974): R2 = 1 – Lur/Lo, where Lur is the log-likelihood function for the estimated model, and L0 is the long-likelihood function in the model with only an intercept.
   Pseudo R2 (2): get yihat, compute the squared correlation between yi and yihat.

II. **Tobit model for corner solution responses**: y = 0 for a nontrivial fraction of the population but is roughly continuously distributed over positive values (the amount of beers: under 21=0…).

$y^* = \beta_0 + x\beta + u, u \mid x \sim N(o, \sigma^2)$ ---------- (Tobit.1)

y = max(0,y*).

$P(y_i = 0 \mid x) = 1 - \Phi(x_i\beta / \sigma)$, when y=0.

The density of y given x is: $(1/\sigma)\phi[(y - x_i\beta)/\sigma]$, when y>0.

1. MLE requires numerical methods.
2. Testing multiple exclusion restrictions is easily done using the Wald test for Likelihood ratio test.
3. Two expectations: E(y|y>0, x) and E(y|x)

$E(y \mid x) = P(y > 0 \mid x) \bullet E(y \mid y > 0, x) = \Phi(x\beta / \sigma) \bullet E(y \mid y > 0, x)$

- **if z~ N(0,1), then E(z|z>c) =** $\phi(c)/[1 - \Phi(c)]$ **for any constant c**.
- $E(y \mid y > 0, x) = x\beta + E(u \mid u > -x\beta) = x\beta + \sigma\phi(x\beta / \sigma) / \Phi(x\beta / \sigma)$

Let $\lambda(c) = \phi(c)/\Phi(c)$, is called **the inverse Mills ratio** → a ratio between the standard normal pdf and cdf, each evaluated at c. →

$E(y \mid y > 0, x) = x\beta + \sigma\lambda(x\beta/\sigma)$ ----------(1)

- $E(y \mid x) = \Phi(x\beta/\sigma) \bullet [x\beta + \sigma\lambda(x\beta/\sigma)] = \Phi(x\beta/\sigma)x\beta + \sigma\phi(x\beta/\sigma)$ ----------(2)
- **E(y|x) is a nonlinear function of x and beta**, the right-hand side of the above equation is always positive.
- **Partial effects of xj on both expectations have the same sign as the coefficient βj,** but the magnitude of the effects depends on the values of all explanatory variables and parameters, including σ.
- **Partial effects:**

$$\frac{\partial E(y \mid y > 0, x)}{\partial x_j} = \beta_j + \beta_j \frac{\partial \lambda(x\beta/\sigma)}{\partial c} = \beta_j \{1 - \lambda(x\beta/\sigma)[x\beta/\sigma + \lambda(x\beta/\sigma)]\}$$

$$\frac{\partial E(y \mid x)}{\partial x_j} = \beta_j \Phi(x\beta/\sigma)$$

using $d\phi(c)/dc = -c\phi(c)$ **and** $d\Phi(c)/dc = \phi(c)$

- **R^2: the square of the correlation coefficient between yi and yihat,** where yihat is got from (2), which is the estimate of E(y|x=xi).
- **Tobit estimates are not chosen to maximize an R^2** – they maximize the log-likelihood function – **whereas the OLS estimates are the values that do produce the highest R2 given the linear functional form**.
- The Tobit model and the formulas for the expectations in (1) and (2), **rely crucially on normality and homoskedasticity** in the underlying latent variable model (Tobit.1). If any of the assumptions in Tobit.1 fail, then it is hard to know what the Tobit MLE is estimating. Nevertheless, for moderate departures from the assumptions, the Tobit model is likely to provide good estimates of the partial effects on the conditional means.

III. Poisson regression model

1. the normal distribution for **count data is the Poisson distribution**, which is entirely determined by its mean, so we only need to specify E(y|x).
2. $P(y = h \mid x) = \exp[-\exp(x\beta)][\exp(x\beta)]^h / h!$
3. quasi-MLE: we do not assume that the Poisson distribution is entirely correct.

IV. Censored and truncated regression models

1. Typically, the **censoring** is due to survey design or institutional constraints. Essentially, the problem solved by censored regression models is a missing data problem, but where we have some info about the missing variable, namely, whether the outcome of the variable is above or below a known threshold.
2. a **truncated** regression model arises when we exclude, on the basis of y, a subset of the population in our sampling scheme. But we know the rule that was used to

include units in the sample. This rule is determined by whether y is above or below a certain threshold.

3. **Censored normal regression model**:

$$y_i = \beta_0 + x_i\beta + u_i, u_i \mid x_i, c_i \sim N(0,\sigma^2)$$

- the model: $w_i = \min(y_i, c_i)$
- **Right censoring**: we only observe yi if it is less than a censoring value, ci. The censoring threshold, ci, can change with individual or family characteristic.
- An OLS regression of wi on xi, using all observations, is not consistent.
- In the Tobit model, we are modeling economic behavior, which often yields zero outcomes; Tobit model is supposed to reflect this.
- With censored regression, we have a data collection problem, because, for some reason, the data are censored.
- MLE: the log-likelihood for observation I is obtained by taking the natural log of the density for each i. We maximize the sum of these across all I, with respect to beta and sigma, to obtain the MLEs.

$$P(w_i = c_i \mid x_i) = P(y_i \geq c_i \mid x_i) = P(u_i \geq c_i - x_i\beta \mid x_i) = 1 - \Phi[(c_i - x_i\beta)/\sigma]$$

the density of wi, given xi and ci:

$$f(w \mid x_i, c_i) = 1 - \Phi[(c_i - x_i\beta)/\sigma], w = c_i$$

$$= (1/\sigma)\phi[(w - x_i\beta)/\sigma], w < c_i$$

- We can interpret **the betas just as in the linear regression model** under random sampling. This is **much different than the Tobit model, where the expectations are nonlinear functions of betas**.
- An important application of censored models is **duration analysis**. A duration is a variable that measures the time before a certain event occurs.
- If any of the assumptions of the censored normal regression model are violated – in particular, if there is heteroskekasticity or nonnormality – the MLEs are generally inconsistent. This shows that the censoring is potentially very costly, as OLS using an uncensored sample requires neither normality nor homoskedasticity for consistency.

4. **Truncated models**:

a. the model: $y = \beta_0 + x\beta + u, u \mid x \sim N(0,\sigma^2)$

b. a random draw (xi,yi) is observed only if yi≤ci. This differs from the censored model: where yi can be larger than ci; we simply do not observe yi if yi>ci. In a censored model, we observe xi for any randomly drawn observation from the population; in the truncated model, we only observe xi if yi≤ci.

c. The density function: $g(y \mid x_i, c_i) = \frac{f(y \mid x_i\beta, \sigma^2)}{F(c_i \mid x_i\beta, \sigma^2)}, y \leq c_i$, where f(.) denotes the normal density with mean $\beta_0 + x\beta$ and variance σ^2, and F(.) is the normal cdf with the same mean and variance, evaluated at ci.

d. MLE: take the log of density function, sum across all i. → consistent, approximately normal estimators.

e. If the underlying homoskedastic normal assumption is violated, the truncated normal MLE is biased and inconsistent.

V. Sample selection corrections

Chapter 18 Advanced Time Series Topics

I. Infinite distributed Lag Models (IDL):

1. the model:

$y_t = \alpha + \delta_0 z_t + \delta_1 z_{t-1} + ... + u_t$ ---------- (1).

- In order for (1) to make sense, the lag coefficients, δj, must tend to zero as j→ ∞
- The **long run propensity** is the sum of all of the lag coefficients:
  LRP = δ0 + δ1 + δ2 +…

2. **Geometric (or Koyck) distributed lag**:

δj = $\gamma\rho^j$, |ρ| <1, j = 0,1,2,…

- the impact propensity (IP) is δ0 = γ
- LRP = γ/(1-ρ)
- (1) can be simplified as (by multiplying yt-1 by ρ):

$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + v_t$, ----------(2) where $\alpha_0 = (1-\rho)\alpha, v_t = u_t - \rho u_{t-1}$

a. but vt can be correlated with y-1 and it is serially correlated. Under strict exogenous assumption: E(u_t|…, z_{t-2},z_{t-1},z_t,z_{t+1},…) = 0 → zt is uncorrelated with ut and ut-1, and therefore with v_t. Thus, we only need an IV for y_{t-1}. Since v_t is uncorrelated with z_{t-1}. If γ ≠ 0, z_{t-1} and y_{t-1} are correlated, even after partialling out zt. Therefore, we can use IVs (zt, z_{t-1}) to estimate (2). Generally, the standard errors need to be adjusted for serial correlation in the {vt}.

b. An alternative to IV is assume $u_t = \rho u_{t-1} + e_t$, E(et|z_t,y_{t-1},z_{t-1},…) = 0 →

$y_t = \alpha_0 + \gamma z_t + \rho y_{t-1} + e_t$ is a dynamically complete model → use OLS directly. But to assume {u_t} follows an AR(1) process with the same ρ appearing in (2) is too strong sometimes. Need to test.

3. Rational distributed lag models – most easily described by adding a lag of z to (2):

$y_t = \alpha_0 + \gamma_0 z_t + \rho y_{t-1} + \gamma_1 z_{t-1} + v_t$, LRP = $(\gamma_0 + \gamma_1)/(1-\rho)$

- **it is important to see whether we need to add a lag of z_t.**

II. Testing for unit roots:

1. the model: $y_t = \alpha + \rho y_{t-1} + e_t$, $E(e_t \mid y_{t-1}, y_{t-2}, ...y_0) = 0$ {et} is said be a **martingale difference sequence** with respect to {yt-1, yt-2,…}.
2. Test: H0: ρ=1. **the test model:** $\Delta y_t = \alpha + \theta y_{t-1} + e_t$, where, $\theta = \rho - 1$, under the assumption about θ, this is a dynamically complete model.
3. **The asymptotic distribution of t-statistic under H0 (I(1) process) is Dickey-Fuller distribution**. The critical value of Dickey-Fuller is negative. We reject Ho if t <c

4. **Augmented Dickey-Fuller test**:

$$\Delta y_t = \alpha + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_p \Delta y_{t-p} + e_t \quad \text{----------***}$$

The critical values and rejection rule are the same as ordinary Dickey-Fuller test. The inclusion of the lagged changes in *** is intended to clean up any serial correlation in Δyt. If we include too many lags, the small sample power of the test generally suffers. But **if we include too few lags, the size of the test will be incorrect, even asymptotically, because the validity of the critical values relies on the dynamics being completely modeled.** For annual data, one or two lags usually suffice. For monthly data, we might include twelve lags.

5. **Dickey-Fuller test with time trend:**

a. A time-stationary process – which has a linear trend in its mean but is I(0) about its trend – can be mistaken for a unit root process if we do not control for a time trend in the Dickey-Fuller regression. If we carry out the usual DF or augmented DF test on a trending but I(0) series, we will probably have little power for rejecting a unit root.

b. To **allow for series with time trends**, we need a new model:

$$\Delta y_t = \alpha + \delta t + \theta y_{t-1} + \gamma_1 \Delta y_{t-1} + \ldots + \gamma_p \Delta y_{t-p} + e_t$$

it is common to only test Ho: θ=0 using a t test.

c. The critical values of the test change. Intuitively, this is because detrending a unit root process tends to make it look more like an I(0) process. Therefore, we require a larger magnitude for the t statistic in order to reject Ho.

d. The t-statistic on the trend does not have an asymptotic standard normal distribution. Typically, we rely on intuition or plots of the TS to decide whether to include a trend in the DF test.

III. **Spurious regression**.

1. Even if the two TS have means that are not trending, a simple regression involving two independent I(1) TS will often result in significant t statistic.
2. Including a time trend does not really change the conclusion.
3. The same considerations arise with multiple independent variables, each of which may be I(1) or some of which may be I(0). If {yt} is I(1) and at least some of the regressors are I(1), the regression results may be spurious.

IV. Cointegration and Error correction models

1. Definition: if {yt} and {xt} are I(1), it is possible that yt – bxt is an I(0) → constant mean, variance, autocorrelations that depends only on the time distance between any two variables in the series, and it is asymptotically uncorrelated. If such nonzero b exists, y and x are **conintegrated** and call b the **cointegration parameter**.

- e.g. three month interest rate vs. 6month rate: cointegration has an economic interpretation. If r6 and r3 were not cointergrated, the difference b/w two rates could become very large, with no tendency for them to come back together.

2. Testing for cointegration is hard when the (potential) cointegration parameter b is unknown.

- if y and x are cointegrated, OLS bhat from y = a + bx +u is consistent for b. The problem is under Ho, the two series are not cointegrated, which means that under Ho, we are running a spurious regression.
- Fortunately, even when b is estimated, when we apply DF to the residual, uhat = y – ahat – bhat*x. the only difference is the Asymptotic critical values for cointegration test of DF are different. Run Δut on ut-1 and lags of Δut → if t-statistic < c, y-bx is I(0).

3. estimation:

a. the model: $y_t = \alpha + \beta x_t + u_t$, the problem is ut may correlated with xt.

b. Let $u_t = \eta + \phi_0 \Delta x_t + \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \gamma_1 \Delta x_{t+1} + \gamma_2 \Delta x_{t+2} + e_t$ →

$y_t = \alpha_0 + \beta x_t + \phi_0 \Delta x_t + \phi_1 \Delta x_{t-1} + \phi_2 \Delta x_{t-2} + \gamma_1 \Delta x_{t+1} + \gamma_2 \Delta x_{t+2} + e_t$ → xt is strictly exogenous now. Whether we need to include leads and lags of the changes, and how many, is an empirical issue.

The OLS estimator of beta is called **the leads and lags estimator of beta**. The possible serial correlation of et can be dealt with by computing a serial correlation-robust se.

4. **Error correction model**:

$$\Delta y_t = \alpha_0 + \alpha_1 \Delta y_{t-1} + \gamma_0 \Delta x_t + \gamma_1 \Delta x_{t-1} + \delta(y_{t-1} - \beta x_{t-1}) + u_t$$

where δ(y-bx) is the error correction term. δ <0, if yt-1 > bxt-1, then y in the previous period has overshot the equilibrium; because δ <0, the error correction term works to push y back towards the equilibrium.

If we have to estimate b first, → **Engle-Granger two-step procedure**.

V. Forecasting

1. Set-up
   a. let f_t = the forecast of y_{t+1} made at t → ft is a one-step-ahead forecast.
   b. Forecast error $e_{t+1} = y_{t+1} - f_t$
   c. Loss function: $E(e_{t+1}^2 \mid I_t) = E[(y_{t+1} - f_t)^2 \mid I_t]$ → conditional expectation $E(y_{t+1}|It)$ minimizes this. That is, if we wish to minimize the expected squared forecast error given info at t, our forecast should be the expected value of yt+1 given info at t.
   d. **Martingale**: {yt} is a martingale if $E(y_{t+1} \mid y_t, y_{t-1}, ..., y_0) = y_t$, for all t ≥0. **if {yt} is a martingale then {Δyt} is a martingale difference sequence**. → for a martingale process, the predicted value of y for the next period is always the value of y for this period.
   e. **Exponential smoothing**: $E(y_{t+1} \mid I_t) = \alpha y_t + \alpha(1-\alpha) y_t + ... + \alpha(1-\alpha)^t y_0$, if $f_0 = y_0$, then for t>0 → $f_t = \alpha y_t + (1-\alpha) f_{t-1}$: suitable for only very specific TS and requires choosing alpha.

2. One-step-ahead forecasting:

$y_t = \delta_0 + \alpha_1 y_{t-1} + \gamma_1 z_{t-1} + u_t, E(u_t \mid I_{t-1}) = 0$ →

$f_n = \delta_{0hat} + \alpha_{1hat} y_n + \gamma_{1hat} z_n$

- the forecast f_n of y_{n+1} is called **point forecast**.
- Forecast interval: we can obtain fn and se(fn) as the intercept and its se from the regression of y_t on $(y_{t-1} - y_n)$ and $(z_{t-1} - z_n)$; that is, we subtract the time n value of y from each lagged y, and similarly for z. → $se(e_{n+1}) = \{[se(f_n)]^2 + \sigma^2\}^{1/2}$ and the approximate 95% forecast interval is: $f_n \pm 1.96 * se(e_{n+1})$

3. **Granger: z Granger causes** y if $E(y_t \mid I_{t-1}) \neq E(y_t \mid J_{t-1})$, ------(a) where It-1 contains past info on y and z, and Jt-1 contains only info on past y. When it holds, past z is useful, in addition to past y, for predicting y.
- Caution: The term "causes" in "Granger causes" should be interpreted with caution. The only sense in which z "causes" y is given in (a). It has nothing to say about contemporaneous causality b/w y and z, so it does not allow us to determine whether Zt is an exogenous or endogenous variable in an equation relating yt to zt → the notion of Granger causality does not apply in pure cross-sectional contexts.
- **How many lags of y and z to include**? First, estimate an autoregressive model for y and performing t and F tests to determine how many lags of y should appear. Then we can test for lags of z.

4. Comparing one-step-ahead forecasts
   a. **In-sample criteria**: R2's.
   b. **Out-sample criteria**: have n+m observations, use the first n to estimate the parameters and save the last m for forecasting.

- root mean squared error: $RMSE = (m^{-1}\sum_{h=0}^{m-1} e_{n+h+1}^2)^{1/2}$
- mean absolute error: $MAE = m^{-1}\sum_{h=0}^{m-1} |e_{n+h+1}|$

5. **Multiple-step-ahead forecast**: Generally, we can build up multiple-step-ahead forecasts of y by using the recursive formula:
   $f_{n,h} = \delta_{0hat} + \alpha_{1hat} f_{n,h-1} + \gamma_{1hat} z_{n,h-1}$
6. **Forecasting trending, seasonal, and intergrated processes**:
   a. y/log(y) = a + bt + ut, E(ut|It-1)=0.
   For n+h, just plug n+h into the trend equation. But, if we use log(y), exp(a+b(n+h)) is not a consistent estimate of y, because of the error term.

- run yt on exp(logyt_hat) without an intercept, let γ be the slope coefficient on exp(logyt_hat), then the forecast of y in period n+h is:
  $f_{n,h} = \gamma \exp[\alpha + \beta(n+h)]$

   b. we can combine trends with other models, say AR(1) model.
   c. For unit-root: two ways
   - impose a unit root: estimate Δyt+1, since yt+1 = Δyt+1 + yt, E(yt+1|It) = E(Δyt+1|It) + yt → forecast of yn+1 at time n is just fn = gn + yn, where gn is the forecast of Δyn+1 at time n. Similarly, we can easily get yn+h in terms of Δyn+h,… Δyn+1.

Chapter 19 Carrying Out an Empirical Project

I. Posing a question

1. For a question to be interesting, it does not need to have broad-based policy implications;
2. When choosing a topic, you should be reasonably sure that data sources exist that will allow you to answer your question in the allotted time.
3. While you are formulating your question, it is helpful to discuss your ideas with your classmates and instructor.

II. **Literature review:** some like to have a separate section called "literature review," while others like to include the literature as part of the introduction.

III. Data collection

1. Deciding on the appropriate data set

Deciding on which kind of data to collect often depends on the nature of the analysis. To answer questions at the individual level, we often only have access to a single cross section. However, in most cases, it is not obvious that we will be able to do a ceteris paribus analysis with a single cross section.

2. Inspecting, cleaning, and summarizing your data:

If we are using time-series data, we must know which variables, if any, have been seasonally adjusted. When we try to get lagged, leading or differenced data, we should short by ID and year (month).

IV. Econometric analysis

1. OLS: in order to justify OLS, you must make a convincing case that the key OLS assumptions are satisfied for your model.

- The first issue is whether the error term is uncorrelated with the explanatory variables.
- When dealing with individual, family, or firm-level cross-sectional data, the **self-selection problem** is often relevant (Chapter 7 and 15).
- You should also be able to argue that the other potential sources of endogeneity – namely, **measurement error and simultaneity** – are not a serious problem.
- Cross-sectional analysis: heteroskedasticity problem. The simplest way is to compute heteroskedasticity-robust statistics.
- If your model has some potential misspecification, such as omitted variables, and you use OLS, you should use **misspecification analysis** (chapter 3 and 5).

2. Making functional form decisions:

- Should some variables appear in logarithmic form?
- Should some variables be included in levels and squares, to possibly capture a diminishing effect?
- How should qualitative factors appear?
- Is it enough to just include binary variables for different attributes or groups?
- Or, do these need to be interacted with quantitative variables? (chapter 7 for detail).

3. Ordinal response as dependent variable: use ordered probit or ordered logit.

4. Use IV to solve various forms of endogeneity, including omitted variables (chap 15), errors-in-variables (chap 15), and simultaneity (chap 16).
5. **Good papers contain sensitivity analysis**. If some observations are much different from the bulk of the sample – say, you have a few firms in s sample that are larger than the other firms – do your results change much when those observations are excluded from the estimation? If so, you may have to alter functional forms to allow for these observations or argue that they follow a completely different model.
6. To use "stepwise regression" to decide which variables should be included in the model. But, the final model often depends on the order in which variables were dropped or added. However, in most applications, one or two explanatory variables are of primary interest, and then the goal is to see how robust the coefficients on those variables are to either adding or dropping other variables, or to changing functional form.

V. Writing an empirical paper

1. Introduction:

The introduction states the basic objectives of the study and explains why it is important. It generally entails a review of the literature, indicating what has been done and how previous work can be improved upon (an extensive literature review can be put in a separate section). Presenting simple statistics or graphs that reveal a seemingly paradoxical relationship is a useful way to introduce the paper's topic. Most researchers like to summarize the findings of their paper in the introduction. This can be a useful device for grabbing the reader's attention.

2. Conceptual (or theoretical) framework

This is the section where you describe the general approach to answering the question you have posed. It can be formal economic theory, but in may cases, it is an intuitive discussion about what conceptual problems arise in answering your question.

3. Economic models and estimation methods

It is very useful to have a section that contains a few equations of the sort you estimate and present in the results section of the paper. This allows you to fix ideas about what the key explanatory variable is and what other factors you will control for. **The distinction between a model and an estimation method** should be made in this section.

- **A model represents a population relationship**.
- After specifying a model, it is appropriate to discuss estimation methods. In most cases, this will be OLS, but, you might use FGLS to do a serial correlation correction. However, the method for estimating a model is quite different from the model itself (OLS, Weighted least square, Cochrane-Orcutt, etc.).
- Any assumptions that are used in obtaining an estimable economic model from an underlying economic model should be clearly discussed.
- We always have to **make assumptions about functional form** whether or not theoretical model has been presented. There are no hard rules on how to choose functional form, but the guidelines discussed in chapter 6 seem to work well in practice.

- If you are using a more advanced estimation method, such as 2SLS, you need to provide some reasons for why you are doing so. You must provide a careful discussion on why your IV choices for the endogenous explanatory variable are valid.

4. The data

- Along with a discussion of the data sources, be sure to discuss the units of each of the variables. Including a table of variable definitions is very useful to the reader.
- It is also informative to present a table of summary statistics, such as minimum and maximum values, means, and standard deviations for each variable. Having such a table makes it easier to interpret the coefficient estimates in the next section, and it emphasizes the units of measurement of the variables.
- For trending variables, things like means are less interesting. It is often useful to compute the average growth rate in a variable over the years in your sample.
- You should always clearly state how many observations you have. For time series data sets, identify the years that you are using in the analysis, including a description of any special periods in history. If you use a pooled cross section or a panel data set, be sure to report how many cross-sectional units.

5. Results

- The results section should include your estimates of any models formulated in the models section. You might start with a very simple analysis.
- The most important thing is to discuss the interpretation and strength of your empirical results. Do the coefficients have the expected signs? Are they statistically significant?
- Be sure to describe the magnitudes of the coefficients on the major explanatory variables. Often one or two policy variables are central to the study. Their signs, magnitudes, and statistical significance should be treated in detail.
- Remember to distinguish between economic and statistical significance. If a t statistic is small, is it because the coefficient is practically small or because its standard error is large?

6. Conclusions

This can be a short section that summarizes what you have learned. The conclusion should also discuss caveats to the conclusions drawn, and it might even suggest directions for further research. It is useful to imagine readers turning first to the conclusion in order to decide whether to read the rest of the paper.

Appendix: Some fundamentals of probability

4. If X and Y are independent, then Cov(X, Y) = 0. The converse is not true. E.g. Y = X2, Cov(X,Y) = 0 (any RV with E(X) = E(X3) = 0 has this property), but Y and X are not independent. → covariance is useful in contexts when relationships are at least approximately linear.
5. Conditional expectations:

- E[c(X)] = c(X), for any function c(X).
- $E[a(X)Y + b(X) \mid X] = a(X)E[Y \mid X] + b(X)$
- if **X and Y are independent, then E(Y|X) = E(Y)**
- $E[E(Y \mid X)] = E(Y)$ → if we first obtain E(Y|X) as a function of X and take the expected value of this with respect to the distribution of X (E(Y|X) is a function of X), then we end up with E(Y).
- $E(Y \mid X) = E[E(Y \mid X, Z) \mid X]$
- **IF** $E(Y \mid X) = E(Y)$**, then Cov(X,Y) = 0.** and every function of X is uncorrelated with Y. The converse is not true. → conditional expectation captures the nonlinear relationship between X and Y that correlation analysis would miss entirely. → if Y and X are RV such that E(Y|X) =0, then E(Y) = 0, and X and Y are uncorrelated.
- Let $E(Y \mid X) = \mu(X)$, if $E(Y^2) < \infty, E[g(X)^2] < \infty$ for some function g, then

$$E\{[Y - \mu(X)]^2 \mid X\} \le E\{[Y - g(X)]^2 \mid X\}$$

$$E\{[Y - \mu(X)]^2\} \le E\{[Y - g(X)]^2\}$$

→ if we measure prediction inaccuracy as the expected squared prediction error, conditional on X, then the conditional mean is better than any other function of X for predicting Y. The conditional mean also minimizes the unconditional expected squared prediction error.