Book Summary: *Econometrics for Dummies*

Yan Zeng

Version 1.0.3, last revised on 2016-08-04.

Abstract

Summary of Pedace [3] and [4].

Contents

Part I
Getting Started with Econometrics

1 Econometrics: The Economist's Approach to Statistical Analysis

2 Getting the Hang of Probability

3 Making Inferences and Testing Hypotheses

Applicability of the Central Limit Theorem.

• When the probability distribution of X is normal, the distribution of $\overline{X}$ is exactly normally distributed regardless of sample size.

• When the probability distribution of X is symmetrical, the CLT applies very well to small sample sizes (often as small as $10 \leq n \leq 25$).

• When the distribution of X is asymmetrical, the approximation to a normal distribution becomes more accurate as n becomes large.

Generally, a good convergence of the sample mean distribution to a normal distribution can be achieved with a sample size of 25 or more.

The chi-squared distribution. The chi-squared distribution is typically used with *variance* estimates and rests on the idea that you begin with a normally distributed random variable, such as $X \sim N(\mu_X, \sigma_X^2)$. With sample data, you estimate the variance of this random variable with

$$s_X^2 = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{n-1}.$$

The chi-squared distribution is obtained by

$$\frac{(n-1)s_X^2}{\sigma_X^2} = \frac{\sum_{i=1}^n (X_i - \overline{X})^2}{\sigma_X^2} \sim \chi_{n-1}^2.$$

The chi-squared distribution takes only nonnegative values and tends to be right-skewed. The extent of its skewness depends on the degrees of freedom or number of observations. The higher the degrees of freedom (more observations), the less skewed (more symmetrical) the chi-squared distribution.

The t-distribution. The t-distribution is derived from a ratio of a standard normal random variable and the square root of a χ^2 random variable. It's bell-shaped symmetrical around zero, and approaches a normal distribution, as the degrees of freedom (number of observations) increases. When you take the ratio of the standard normal to the square root of your chi-squared distribution, you end up with a t-distribution:

$$\frac{(\overline{X} - \mu_X)/\frac{\sigma_X}{\sqrt{n}}}{\sqrt{\frac{s_X^2}{\sigma_X^2}}} = \frac{\overline{X} - \mu_X}{\frac{\sigma_X}{\sqrt{n}}} \cdot \frac{\sigma_X}{s_X} = \frac{\overline{X} - \mu_X}{\frac{s_X}{\sqrt{n}}} \sim t_{n-1}.$$

The F-distribution. The F-distribution is used to compare variances of two different normal distributions. It is derived from a ratio of two χ^2 distributions divided by their respective degrees of freedom. The F-distribution tends to be right-skewed, with the amount of skewness depending on the degrees of freedom. As the degree of freedom in the numerator and denominator increase, the F-distribution approaches a normal distribution.

$$\frac{\sum_{i=1}^n (X_i - \overline{X})^2/(n-1)}{\sum_{i=1}^m (Y_i - \overline{Y})^2/(m-1)} = \frac{s_X^2}{s_Y^2} \sim F_{(n-1),(m-1)}$$

Part II
Building the Classical Linear Regression Model

4 Understanding the Objectives of Regression Analysis

Model specification consists of selecting an outcome of interest or dependent variable and one or more independent factors, as well as choosing an appropriate functional form.

Spurious correlation occurs when two variables coincidentally have a statistical relationship but one doesn't cause the other.

Causation cannot be proven by statistical results. Your results can be used to support a hypothesis of causality, but only after you've developed a model that is well grounded in economic theory and/or good common sense.

In regression analysis, the random error term represents the difference between the observed value of your dependent variable and the conditional mean of the dependent variable derived from your model:

$$\varepsilon = Y - E[Y|X_1, \cdots, X_n] = Y - E[Y|\mathbf{X}].$$

The random error can result from one or more of the following factors:

✓Insufficient or incorrectly measured data.

✓A lack of theoretical insights to fully account for all the factors that affect the dependent variable.

✓Applying an incorrect functional form; for example, assuming the relationship is linear when it's quadratic.

✓Unobservable characteristics.

✓Unpredictable elements of behavior.

The equation $\varepsilon = Y - E[Y|\mathbf{X}]$ can be written more explicitly as

$$Y(\mathbf{x};\omega) = E[Y|\mathbf{X}=\mathbf{x}] + \varepsilon(\mathbf{x};\omega), \ \textit{or} \ \varepsilon(\mathbf{x};\omega) = Y(\mathbf{x};\omega) - E[Y|\mathbf{X}=\mathbf{x}]$$

where $Y(\mathbf{x};\omega)$ is sampled according to the conditional distribution of Y at $\mathbf{X}=\mathbf{x}$. We typically assume $(\varepsilon(\mathbf{x};\cdot))_{\mathbf{x}}$ has identical variance or even i.i.d.. Note this is truly a strong assumption.

Cross-sectional data contains measurements for individual observations at a given point in time.

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k X_{ik} + \varepsilon_i.$$

Time-series data contains measurements on one or more variables over time in a given space.

$$Y_t = \beta_0 + \sum_{k=1}^{p} \beta_k X_{tk} + \varepsilon_t.$$

Panel data (also referred to as *longitudinal data*) contains a time series for each cross-sectional unit in the sample.

$$Y_{it} = \beta_0 + \sum_{k=1}^{p} \beta_k X_{itk} + \varepsilon_{it}.$$

Pooled cross-sectional data. Simply because your dataset contains both a cross-sectional and time-series component doesn't make it a panel dataset. It isn't a panel dataset unless the same individual units are observed in each subsequent time period.

5 Going Beyond Ordinary with the Ordinary Least Squares Technique

Regression coefficients in a model with one independent variable:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} = \frac{\hat{s}_{XY}^2}{\hat{s}_X^2},\ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Intercept term is usually ignored in applied work, because situations where all of the explanatory variables equal zero are unlikely to occur.

Justifying the least squares principle. In most situations, OLS remains the most popular technique for estimating regressions for the following three reasons:

- Using OLS is easier than the alternatives. Other techniques require more mathematical sophistication and more computing power.
- OLS is sensible. You can avoid positive and negative residuals canceling each other out and find a regression line that's as close as possible to the observed data points.
- OLS results have desirable characteristics.

✓The regression line always passes through the sample means of Y and X, or $\overline{Y} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X}$ (the point $(\overline{X}, \overline{Y})$ falls on the line $y = \hat{\beta}_0 + \hat{\beta}_1 x$): by the definition of $\hat{\beta}_0$ and $\hat{\beta}_1$.

✓The mean of the estimated (predicated) Y value is equal to the mean value of the actual Y, or $\overline{\hat{Y}} = \overline{\hat{\beta}_0 + \hat{\beta}_1 X} = \hat{\beta}_0 + \hat{\beta}_1 \overline{X} = \overline{Y}$.

✓The mean of the residuals is zero, or $\overline{\hat{\varepsilon}} = \overline{Y - (\hat{\beta}_0 + \hat{\beta}_1 X)} = \overline{Y} - (\hat{\beta}_0 + \hat{\beta}_1 \overline{X}) = 0$.

✓The residuals are uncorrelated with the predicted Y, or $\sum_{i=1}^n (\hat{Y}_i - \overline{Y})\hat{\varepsilon}_i = 0$.

✓The residuals are uncorrelated with observed values of the independent variable, or $\sum_{i=1}^n \hat{\varepsilon}_i X_i = 0$.

Standardizing regression coefficients. Comparing coefficient values is not as straightforward as you may first think. Here are a few reasons why:

- In standard OLS regression, the coefficient with the largest magnitude is not necessarily associated with "the most important" variable.
- Coefficient magnitudes can be affected by changing the units of measurement; in other words, scale matters.
- Even variables measured on similar scales can have different amounts of variability.

If you want to compare coefficient magnitudes in a multiple regression, you need to calculate the *standardized regression coefficients*. You can do so in two ways:

- Calculating a Z-score for every variable of every observation and then performing OLS with the Z values rather than the raw data.
- Obtaining the OLS regression coefficients using the raw data and then multiplying each coefficient by $\left(\frac{\hat{\sigma}_{X_k}}{\hat{\sigma}_Y}\right)$.

Mathematically, you transform the original regression equation $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i$ to

$$\frac{Y_i - \overline{Y}}{\hat{\sigma}_Y} = \beta_1 \left(\frac{X_{i1} - \overline{X_1}}{\hat{\sigma}_{X_1}}\right)\left(\frac{\hat{\sigma}_{X_1}}{\hat{\sigma}_Y}\right) + \beta_2 \left(\frac{X_{i2} - \overline{X_2}}{\hat{\sigma}_{X_2}}\right)\left(\frac{\hat{\sigma}_{X_2}}{\hat{\sigma}_Y}\right) + \cdots + \beta_p \left(\frac{X_{ip} - \overline{X_p}}{\hat{\sigma}_{X_p}}\right)\left(\frac{\hat{\sigma}_{X_p}}{\hat{\sigma}_Y}\right) + \frac{\hat{\varepsilon}_i}{\hat{\sigma}_Y}.$$

where we have taken advantage of one of the desirable OLS properties, namely that the average residual is zero.

Note regular OLS coefficients and standardized regression coefficients do not have the same meaning. The standardized regression coefficient estimates the standard deviation change in your dependent variable for a 1-standard-deviation change in the independent variable, holding other variables constant.

Measuring goodness of fit.

- *Explained sum of squares* (ESS), *residual sum of squares* (RSS), and *total sum of squares* (TSS):

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2,\ RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2,\ TSS = \sum_{i=1}^n (Y_i - \overline{Y})^2 = ESS + RSS.$$

• *Coefficient of determination* (*R-squared*) and *adjusted R-squared* (adjusted by degrees of freedom):

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS},\ R^2_{adj} = 1 - \frac{\frac{RSS}{n-p-1}}{\frac{TSS}{n-1}}.$$

where n is the number of observations, and p is the number of independent variables in the model.

If you increase the number of explanatory variables in a regression model, your R-squared value increases or remains the same, but it can never cause your R-squared value to decrease. When you add more variables, you lose *degree of freedom* (the number of observations above and beyond the number of estimated coefficients). **Fewer degrees of freedom make your estimates less reliable** (for more on this topic, turn to Chapter 6). In order to compare two models on the basis of R-squared (adjusted or not), the dependent variable and sample size must be the same.

Here are a few reasons why you shouldn't use R-squared (adjusted or not) as the only measure of your regression's quality:

• A regression may have a high R-squared but have no meaningful interpretation because the model equation is not supported by economic theory or common sense.

• Using a small data set or one that includes inaccuracies can lead to a high R-squared value but deceptive results.

• Obsessing over R-squared may cause you to overlook important econometric problems.

In economic settings, a high R-squared (close to 1) is more likely to indicate that something is wrong with the regression instead of showing that it's of high quality. High R-squared values may be associated with regressions that violate assumptions and/or have nonsensical results (coefficients with the wrong sign, unbelievable magnitudes, and so on.). When evaluating regression quality, give these outcomes more weight than the R-squared.

6 Assumptions of OLS Estimation and the Gauss-Markov Theorem

The OLS/CLRM assumptions and their intuition.

• *The model is linear in parameters and has an additive error term.* Other techniques, such as *maximum likelihood* (ML) estimation, can be used when the function you need to estimate is not linear in parameters.

• *The value for the independent variables are derived from a random sample of the population and contain variability.*

Strictly speaking, the CLRM assumes that the values of the independent variables are fixed in repeated random samples. The more common version of the assumption is that the values of the independent variable are random from sample to sample but independent of the error term. The weaker version is equivalent asymptotically (with large samples).

This assumption isn't likely to hold when you use lagged values of your dependent variable as an independent variable (*autoregression*) or when the value of your dependent variable simultaneously affects the value of one (or more) of your independent variables (*simultaneous equations*). Therefore, OLS is inappropriate in these situations.

In practice, for each random sample X_i we often observe Y only once. So we either assume a simple parametric model, e.g. linear regression, or use points in a neighborhood of X_i for averaging, e.g. K-nearest neighbor regression (KNN regression). See James et al. [2, page 104] for details.

• *No independent variable is a perfect linear function of any other independent variable(s) (no perfect collinearity).*

If you have perfect collinearity, the software program you use to calculate regression results cannot estimate the regression coefficients, since perfect collinearity causes you to lose linear independence and the computer can't identify the unique effect of each variable. In applied cases, high collinearity is much more common than perfect collinearity.

• *The model is correctly specified and the error term has a zero conditional mean.*

$E[\varepsilon|X=x]=0$ means for given x, the residuals $\varepsilon(x)=y-(\beta_0+\beta_1 x)$ oscillate around 0 with average equal to 0. Graphically, this means the values of the dependent variable oscillate around the regression line with averages falling on the regression line. This assumption may fail if you have *misspecification* (you fail to include a relevant independent variable or you use an incorrect functional form) or a *restricted dependent variable* (namely, a qualitative or limited dependent variable).

- *The error term has a constant variance (no heteroskedasticity).*

Graphically, this means the "scatteredness" of the values of the independent variable around the regression line is approximately the same everywhere. Heteroskedasticity is a common problem for OLS regression estimation, espcially with cross-sectional and panel data.

- *The values of the error term aren't correlated with each other (no autocorrelation or no serial correlation).*

Graphically, no autocorrelation means the scatter plot of $(\varepsilon_{i-k},\varepsilon_i)_{i=k+1}^{\infty}$ spreads out homogeneously in all directions, for any $k\geq 1$. Autocorrelation can be quite common when you are estimating models with time-series data, because when observations are collected over time, they are unlikely to be independent from one another.

The Gauss-Markov Theorem. This theorem states that the ordinary least squares (OLS) estimators are the best linear unbiased estimators (BLUE) given the assumptions of the CLRM.

- *Linearity of OLS* (as a function of the observed Y values):

$$\hat{\beta}_1=\sum_{i=1}^{n}c_i(Y_i-\overline{Y}),\ \hat{\beta}_0=\overline{Y}-\left[\sum_{i=1}^{n}c_i(Y_i-\overline{Y})\right]\overline{X},$$

where $c_i=\frac{X_i-\overline{X}}{\sum_{i=1}^{n}(X_i-\overline{X})^2}$, $i=1,\cdots,n$.

- *Unbiasedness*: $E[\hat{\beta}_1]=\beta_1$, $E[\hat{\beta}_0]=\beta_0$.
- *Best* means achieving the smallest possible variance among all similar estimators.

$$\text{Var}(\hat{\beta}_1)=\frac{\sigma_\varepsilon^2}{\sum_{i=1}^{n}(X_i-\overline{X})^2}.$$

When judging how good or bad an estimator is, econometricians usually evaluate the amount of bias and variance of that estimator. The BLUE property of OLS estimators is viewed as the gold standard.

Econometricians have devised methods to deal with failures of the CLRM assumptions, but they aren't always successful in proving that the alternative method produces a BLUE. In those cases, they usually settle for an *asymptotic* property known as *consistency.* Estimators are consistent if, as the sample size approaches infinity, the variance of the estimator gets smaller and the value of the estimator approaches the true population parameter value.

Also refer to Table 6-1: Summary of Gauss-Markov Assumptions [3], page 19.

7 The Normality Assumption and Inference with OLS

The normality assumption. The normality assumption in econometrics states that, for any given X value, the error term follows a normal distribution with a zero mean and constant variance: $\varepsilon|X\sim N(0,\sigma_\varepsilon^2)$.

The normality assumption isn't required for performing OLS estimation. It's necessary only when you want to produce confidence intervals and/or perform hypothesis tests with your OLS estimates.

In some applications, the assumption of a normal distribution for the error term may be difficult to justify. These situations typically involve a dependent variable Y that has limited or highly skewed values. Econometricians have shown that with *large* sample sizes, normality is not a major issue because the OLS estimators are approximately normal even if the errors are not normal.

The sampling distribution of OLS coefficients. All OLS coefficients are a linear function of the error term. If you assume that the error term has a normal distribution, you're also assuming that the sampling distribution of the coefficients is normal:

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^n (X_i-\overline{X})^2} = \frac{\sum_{i=1}^n (X_i-\overline{X})(\beta_0+\beta_1 X_i+\varepsilon_i-\overline{Y})}{\sum_{i=1}^n (X_i-\overline{X})^2} = \beta_1 + \frac{\sum_{i=1}^n \varepsilon_i (X_i-\overline{X})}{\sum_{i=1}^n (X_i-\overline{X})^2} \\ &\sim N\left(\beta_1, \sigma^2_{\hat{\beta}_1}\right),\end{aligned}$$

where $\sigma^2_{\hat{\beta}_1} = \frac{\sigma^2_\varepsilon}{\sum_{i=1}^n (X_i-\overline{X})^2}$ and

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} = \beta_0 - \frac{\sum_{i=1}^n \varepsilon_i (X_i-\overline{X})}{\sum_{i=1}^n (X_i-\overline{X})^2}\overline{X} \sim N\left(\beta_0, \sigma^2_{\hat{\beta}_0}\right).$$

where $\sigma^2_{\hat{\beta}_0} = \frac{(\sum_{i=1}^n X_i^2)\sigma^2_\varepsilon}{n\sum_{i=1}^n (X_i-\overline{X})^2}$.

OLS standard errors and the t-distribution. In practice, the true variance of the error σ^2_ε isn't known, but you can estimate it by calculating the *mean square error* (MSE):

$$\hat{\sigma}^2_\varepsilon = \frac{RSS}{n-p-1} = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-p-1}.$$

This provides the *standard errors of the coefficients*:

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\varepsilon}{\sqrt{\sum_{i=1}^n (X_i-\overline{X})^2}},\ \hat{\sigma}_{\hat{\beta}_0} = \sqrt{\frac{\sum_{i=1}^n X_i^2}{n\sum_{i=1}^n (X_i-\overline{X})^2}} \cdot \hat{\sigma}_\varepsilon.$$

The assumption that the error is normally distributed implies that the MSE and the estimated variances of the coefficients have a chi-squared (χ^2) distribution with $n-p-1$ degrees of freedom. Therefore, for $k=0,1$,

$$\frac{\hat{\beta}_k - \beta_k}{\hat{\sigma}_{\hat{\beta}_k}} \sim t_{n-p-1}.$$

Significance of individual regression coefficients. You an report the statistical significance of your coefficients with either the *confidence interval approach* or the *test of significance approach.* The former provides you with a range of possible values for your estimator in repeated sampling, and the latter gives you a test statistic that's used to determine the likelihood of your hypothesis.

A *type I error* is rejecting a hypothesis that's true, and a *type II error* is failing to reject a hypothesis that's false. If you choose a higher level of significance, you increase the chances of committing a type I error. And if you choose a lower level of signifiance, you increase the chances of committing a type II error.

Overall/joint significance. The explained and unexplained variations from a regression model have a chi-squared distribution under the assumption that the conditional distribution of Y is normal ($Y|X \sim N(\beta_0+\beta_1 X_{i1}+\beta_2 X_{i2}+\cdots, \sigma^2_Y)$), which is equivalent to assuming that the error term is normally distributed ($\varepsilon|X \sim N(0,\sigma^2_\varepsilon)$).

The R-squared value is a measure of overall fit for a regression model, but it doesn't tell you whether the amount of explained variation is statistically significant. Despite a low R-squared value, your model may explain a significant amount of variation in your dependent variable. The opposite may also be true; a high R-squared value may not be statistically significantly different from zero. In models with numerous independent variables, many of the variables can be individually statistically insignificant, yet they are collectively significant.

The null and alternative hypotheses to test for a regression model's overall significant are

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0,\ H_1 : H_0 \text{ is not true.}$$

Overall significance only examines the impact of the slope coefficients and is tested using the following F-statistic:

$$F = \frac{\frac{ESS}{p}}{\frac{RSS}{n-p-1}} = \frac{\frac{R^2}{p}}{\frac{(1-R^2)}{(n-p-1)}} \sim F_{p,n-p-1}.$$

For given R-squared value, smaller p yields bigger F, which has the interpretation of "same R-squared value with less explanatory variables has more significance".

The F-test can also be used to examine the joint significance of a subset of variables in a regression model that includes several independent variables:

$$F = \frac{\frac{RSS_r - RSS_{ur}}{q}}{\frac{RSS_{ur}}{n-p-1}} = \frac{\frac{ESS_{ur} - ESS_r}{q}}{\frac{RSS_{ur}}{n-p-1}} \sim F_{q,n-p-1}$$

where RSS_r is the RSS for the *restricted* model (the model with fewer independent variables), RSS_{ur} is the RSS for the *unrestricted* model (the model with more independent variables), n is the number of sample measurements, p is the number of independent variables in the unrestricted model, and q is the number of independent variables contained in your unrestricted model that are not contained in your restricted model.

The F-test of overall significance is a special case of the more general test. In that case, $q = p$ because the restricted model contains no independent variables in a test of overall significance.

The intuition of F-test is explained by Brooks [1] as follows.

"To see why the test centres around a comparison of the residual sums of squares from the restricted and unrestricted regressions, recall that OLS estimation involved choosing the model that minimised the residual sum of squares, with no constraints imposed. Now if, after imposing constraints on the model, a residual sum of squares results that is not much higher than the unconstrained model's residual sum of squares, it would be concluded that the restrictions were supported by the data. On the other hand, if the residual sum of squares increased considerably after the restrictions were imposed, it would be concluded that the restrictions were not supported by the data and therefore that the hypothesis should be rejected.

It can be further stated that $RSS_r \geq RSS_{ur}$. Only under a particular set of very extreme circumstances will the residual sums of squares for the restricted and unrestricted models be exactly equal. This would be the case when the restriction was already present in the data, so that it is not really a restriction at all."

Forecasting. Under the normality assumption, for a given value X_0 of the independent variable, the forecasted value of dependent variable $\hat{Y}_0 = \hat{\beta}_0 + \hat{\beta}_1 X_0$ is normally distributed:

$$\hat{Y}_0 \sim N\left(\beta_0 + \beta_1 X_0, \sigma^2_{\hat{Y}_0}\right)$$

where

$$\sigma^2_{\hat{Y}_0} = \sigma^2_\varepsilon \left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right].$$

In practice, we don't know the true variance of the error, so we use

$$\hat{\sigma}^2_{\hat{Y}_0} = \hat{\sigma}^2_\varepsilon \left[\frac{1}{n} + \frac{(X_0 - \overline{X})^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right],$$

which has a chi-squared distribution with $n - p - 1$ degrees of freedom.

Consequently, the confidence interval for the prediction is $\hat{Y}_0 \pm t_{\alpha/2,n-p-1}\hat{\sigma}_{\hat{Y}_0}$. A unique characteristic of this confidence interval is the changing standard error of the prediction; smallest at the mean value of X and increasing exponentially as X deviates from the mean.

Part III
Working with the Classical Regression Model

8 Functional Form, Specification, and Structural Stability

Functional Form.

• *Dimension/unit/scale.* Change in absolute amount or in percentage?

✓Log-log model (elasticity, i.e. the estimated percentage change in the dependent variable for a percentage change in the independent variable).

✓Log-linear model (the estimated percentage change in the dependent variable for a unit change in the independent variable).

✓Linear-log model (the estimated unit change in the dependent variable for a percentage change in the independent variable).

• *Graph of the dependent-independent variable chart.*

✓Quadratic function (best for finding minimums and maximums).

✓Cubic function (good for inflexion).

✓Inverse function (limiting the value of the dependent variable).

✓Linear-log model (the impact of the independent variable on the dependent variable decreases as the value of the independent variable increases).

Misspecification.

• *Omitting relevant variables.* You have an omitted variable bias if an excluded variable has some effect on your dependent variable and it's correlated with at least one of your independent variables. The intuition is best illustrated by projection in Hilbert space.

• *Including irrelevant variable.* The estimated coefficients remain unbiased but the standard errors are increased–the estimated standard error for any given regression coefficient is given by

$$\hat{\sigma}_{\hat{\beta}_k} = \sqrt{\frac{\hat{\sigma}_\varepsilon^2}{\sum_{i=1}^n (X_i - \overline{X})^2 (1 - R_k^2)}}$$

where R_k^2 is the R-squared from the regression of X_k on the other independent variables. Including irrelevant variables does not change $\hat{\sigma}_\varepsilon^2$ while increasing R_k^2.

Just because your estimated coefficient isn't statistically significant doesn't make it irrelevant. A well-specified model usually includes some variables that are statistically significant and some that aren't. Additionally, variables that aren't statistically significant can contribute enough explained variation to have no detrimental impact on the standard errors.

Structural Stability.

• *Perform a RESET to test the severity of specification issues.* Ramsey's *regression specification error test* (RESET) is conducted by adding a quartic function of the fitted values of the dependent variable ($\hat{Y}_i^2$, $\hat{Y}_i^3$, and $\hat{Y}_i^4$) to the original regression and then testing the joint significance of the coefficients for the added variables. The logic of using a quartic of your fitted values is that they serve as proxies for variables that may have been omitted. Because the proxies are essentially nonlinear functions of your Xs, RESET is also testing misspecification from functional form.

1. Estimate the model you want to test for specification error. E.g. $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \varepsilon_i$.

2. Obtain the fitted values after estimating your model and estimate $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \alpha \hat{Y}_i^2 + \gamma \hat{Y}_i^3 + \delta \hat{Y}_i^4 + \varepsilon_i$.

3. Test the joint significance of the coefficients on the fitted values of Y_i terms (α, γ, and δ) using an F-statistic.

A RESET allows you to identify whether misspecification is a serious problem with your model, but it doesn't allow you to determine the source.

• *Use the Chow test to determine structural stability.* Sometimes specification issues arise because the parameters of the model either aren't stable or they change. We can conduct a Chow test for structural stability between any two groups (A and B) in just three steps:

1. Estimate your model combining all data and obtain the residual sum of squares (RSS_r) with degrees of freedom $n-p-1$.

2. Estimate your model separately for each group and obtain the residual sum of squares for group A, $RSS_{ur,A}$, with degrees of freedom n_A-p-1 and the residual sum of squares for group B, $RSS_{ur,B}$, with degrees of freedom n_B-p-1.

3. Compute the F-statistic by using this formula:

$$F=\frac{\frac{RSS_r-(RSS_{ur,A}+RSS_{ur,B})}{p+1}}{\frac{RSS_{ur,A}+RSS_{ur,B}}{n-2p-2}}.$$

The null hypothesis for the Chow test is structural stability. The larger the F-statistic, the more evidence you have against structural stability and the more likely the coefficients are to vary from group to group. Note the result of the F-statistic for the Chow test assumes homoskedasticity. A large F-statistic only informs you that the parameters vary between the groups, but it doesn't tell you which specific parameter(s) is (are) the source(s) of the structural break.

• *Robustness/sensitivity analysis.* If the coefficients of your core variables aren't sensitive (maintain the same sign with similar magnitudes and levels of significance), then they are considered *robust.* Some variables, despite not being of primary interest (that is, despite not being core), are likely to be essential control variables that would be included in any analysis of your outcome of interest (you should rely on economic theory and your common sense here).

9 Regression with Dummy Explanatory Variables

Interpretation.

• The coefficient for your dummy variables(s) in a regression containing a quantitative variable shifts the regression function up or down. The same holds true when there's more than one dummy variable.

• The inclusion of an interaction term in your econometric model allows the regression function to have a different intercept and slope for each group identified by your dummy variables. The coefficient for your dummy variable(s) in a regression shifts the intercept, and the coefficient for your interaction term changes the slope (which is the impact of your quantitative variable).

• The inclusion of interacted dummy variables in your econometric model allows the regression function to have different intercepts for each combination of qualitative attributes. The coefficients for your dummy variables and their interaction shift the intercept by the estimated magnitude.

Testing for significance.

• Testing the joint significance of a group of dummy variables in a gression model is accomplished by generalizing the F-test of overall significance to

$$F=\frac{\frac{RSS_r-RSS_{ur}}{q}}{\frac{RSS_{ur}}{n-p-1}}=\frac{\frac{ESS_{ur}-ESS_r}{q}}{\frac{RSS_{ur}}{n-p-1}}\sim F_{q,n-p-1}$$

where RSS_r is the residual sum of squares for the *restricted* model (the model excluding the dummy variables), RSS_{ur} is the residual sum of squares for the *unrestricted* model (the model including the dummy variables), n is the number of sample measurements, p is the number of independent variables in the unrestricted model, and q is the number of dummy variables added in your unrestrictd model that are not contained in your restricted model.

• Using a dummy variable and interaction terms, a test of joint significance can be equivalent to performing a Chow test.

1. Create a dummy variable (D) that identifies any two groups suspected of a structural break.

2. Create interaction variables with your dummy variable and every other variable in your model.
3. Estimate the regression model that includes the quantitative, dummy, and interaction variables.
4. Test the joint significance of the dummy variable identifying the two groups and all the interaction terms that include this dummy variable.

The advantage of the dummy variable approach to testing for structural stability is that it allows you to identify the source of the difference between the groups. The disadvantage of the dummy variable approach is that it may not be practical if you're working with numerous independent variables.

Part IV
Violations of Classical Regression Model Assumptions

10 Multicollinearity

Multicollinearity refers to a linear relationship between two or more independent variables in a regression model. There are two types of multicollinearity:

Perfect multicollinearity. When perfectly collinear variables are included as independent variables, you can't use the OLS technique to estimate the value of the parameters. Your regression coefficients are indeterminate and their standard errors are infinite.

High multicollinearity. It's much more common than its perfect counterpart and can be equally problematic when it comes to estimating an econometric model. Technically, the presence of high multicollinearity doesn't violate any CLRM assumptions. Consequently, OLS estimates can be obtained and are BLUE with high multicollinearity. The larger variances (and standard errors) of the OLS estimators are the main reason to avoid high multicollinearity.

When econometricians point to a multicollinearity issue, they're typically referring to *high* multicollinearity rather than *perfect* multicollinearity. Most econometric software programs identify perfect multicollinearity and drop one (or more) variables prior to providing the estimation results.

- *Causes of multicollinearity include*

  ✓You use variables that are lagged values of one another.

  ✓You use variables that share a common time trend component.

  ✓You use variables that capture similar phenomena.

- *Consequences of high multicollinearity include*

  ✓Larger standard errors and insignificant t-statistics:

$$\sigma^2_{\hat{\beta}_k} = \frac{\hat{\sigma}^2_{\varepsilon}}{\sum(X_{ik} - \overline{X}_k)^2(1 - R^2_k)},$$

where $\hat{\sigma}^2_{\varepsilon}$ is the mean squared error (MSE) and R^2_k is the R-squared value from regressing X_k on the other Xs. Higher multicollinearity results in a larger R^2_k, which increases the standard error of the coefficient. Because the t-statistic associated with a coefficient is $t_k = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}}$, high multicollinearity also tends to result in insignificant t-statistics.

✓Coefficient estimates that are sensitive to changes in specification. If the independent variables are highly collinear, the estimates must emphasize small differences in the variables in order to assign an independent effect to each of them.

✓Nonsensical coefficient signs and magnitudes. With higher multicollinearity, the variance of the estimated coefficients increases, which in turn increases the chances of obtaining coefficient estimates with extreme values.

When two (or more) variables exhibit high multicollinearity, there's more uncertainty as to which variables should be credited with explaining variation in the dependent variable. For this reason, a high R-squared value combined with many statistically insignificant coefficients is a common consequence of high multicollinearity.

Rule of thumb for identifying multicollinearity. Because high multicollinearity doesn't violate a CLRM assumption and is a sample-specific issue, researchers typically don't use formal statistical tests to detect multicollinearity. Instead, they use two sample measurements as indicators of a potential multicollinearity problem.

• **Pairwise correlation coefficients**. The sample correlation of two independent variables, X_k and X_j, is calculated as

$$r_{kj} = \frac{s_{kj}}{s_k s_j}.$$

As a rule of thumb, correlation coefficients around 0.8 or above may signal a multicollinearity problem. Other evidence you should also check include insignificant t-statistics, sensitive coefficient estimates, and nonsensical coefficient signs and values.

Note the pairwise correlation coefficients only identify the linear relationship of two variables. It does not check linear relationship among more than two variables.

• **Auxiliary regression and the variance inflation factor (VIF)**. A VIF for any given independent variable is calculated by

$$VIF_k = \frac{1}{1 - R_k^2}$$

where R_k^2 is the R-squared value obtained by regressing independent variable X_k on all the other independent variables in the model.

As a rule of thumb, VIFs greater than 10 signal a highly likely multicollinearity problem, and VIFs between 5 and 10 signal a somewhat likely multicollinearity issue. Remember to check also other evidence of multicollinearity (insignificant t-statistics, sensitive or nonsensical coefficient estimates, and nonsensical coefficient signs and values). A high VIF is only an indicaotr of potential multicollinearity, but it may not result in a large variance for the estimator if the variance of the independent variable is also large.

Resolving multicollinearity issues. If the primary purpose of your study is to estimate a model for prediction or forecasting, then the best solution may be to do nothing. If you want to obtain reliable estimates of the individual parameters in the model, you need to be more concerned with multicollinearity. (But you shouldn't modify your model if the t-statistics of the suspect variables(s) are greater than 2 *and* the coefficient signs and magnitudes make economic sense.)

We should take a holistic approach that considers the benefits of eliminating high correlation between the independent variables against the costs of addressing an issue that's specific to the sample you're using rather than the population of interest. Once we decide to resolve the multicollinearity issue, we have several options:

• **Acquire more data**. High multicollinearity may be unique to your sample, so the acquisition of additional data is a potential solution. But don't automatically assume a "more is better" mentality when building your database, since the collection of additional data may be costly or could inadvertently result in a change of your population.

• **Use a new model**.

✓ *First-differencing.* Its use is limited to models utilizing time-series or panel data. It also has its cost: 1) losing observations; 2) losing variation in your independent variables (resulting in insignificant coefficients); 3) changing the specification (possibly resulting in misspecification bias).

✓ *The composite index variable.* But never combine variables into an index that would, individually, be expected to have opposite signs.

• **Expel the problem variables(s)**. In case of severely high multicollinearity (correlation coefficients greater than 0.9), you don't have to follow any statistical rationale for choosing to drop one variable over another. If you're using VIFs to detect multicollinearity, a variable with a VIF greater than 10 is usually the most likely to be dropped. A smaller MSE usually signals that the statistical benefits of dropping the variable exceed the costs of specification bias. Save this method as a last resort and place theoretical considerations above purely statistical justifications.

11 Heteroskedasticity

Homoskedasticity is expressed as $Var(\varepsilon|\mathbf{X}_i) = \sigma_\varepsilon^2 =$ a constant for all i $(i = 1, 2, \cdots, N)$, where $\mathbf{X}_i$ represents a vector of values for each individual and for all the independent variables. **Heteorskedasticity** is expressed as $Var(\varepsilon|\mathbf{X}_i) = \sigma_{i\varepsilon}^2$ $(i = 1, 2, \cdots, N)$.

The consequences of heteroskedasticity. In the presence of heteroskedasticity, the OLS estimators may not be efficient (achieve the smallest variance). In addition, the estimated standard errors of the coefficients will be biased, which results in unreliable hypothesis tests (t-statistics). The OLS estimates, however, remain unbiased.

Under the assumption of homoskedasticity, for model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$,

$$Var(\hat{\beta}_1) = \frac{\sigma_\varepsilon^2}{TSS_X}$$

where $TSS_X = \sum(X_i - \overline{X})^2$. Without the homoskedasticity assumption, the variance of β_1 is

$$Var(\hat{\beta}_1) = \frac{\sum(X_i - \overline{X})^2 \sigma_{i\varepsilon}^2}{TSS_X^2}$$

where $\sigma_{i\varepsilon}^2$ is the heteroskedastic variance of the error. The t-statistic for coefficients is calculated with

$$t = \frac{\text{estimated } \beta - \text{hypothesized } \beta}{\text{std error}}.$$

Therefore, any bias in the standard error estimate is passed on to your t-statistics and conclusions about statistical significance.

Heteroskedasticity is a common problem for OLS regression estimation, especially with cross-sectional and panel data. You usually have no way to know in advance if it's going to be present, and theory is rarely useful in anticipating its presence.

Detecting heteroskedasticity with residual analysis. The challenge to identifying heteroskedasticity is that you can only know $\sigma_{i\varepsilon}^2$ if you have the entire population corresponding to the chosen independent variables (Xs). In practice, you'll be using a sample with only a limited number of observations for a particular X. Consequently, in applied situations the detection of heteroskedasticity relies on your intuition, prior empirical work, educated guesswork, or even sheer speculation.

- **Examining the residuals in graph form**.

- **The Breusch-Pagan test**. This test assumes that heteroskedasticity may be a linear function of all the independent variables in the model: $\varepsilon_i^2 = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + u_i$. The values for ε_i^2 aren't known in practice, so the $\hat{\varepsilon}_i^2$ are calculated from the residuals and used as proxies for ε_i^2. Generally, the BP test is based on the estimation of $\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + u_i$. Alternatively, a BP test can be performed by estimating $\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i$, where $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$. Here's how to perform a BP test:

1. Estimate your model, $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$, using OLS.
2. Obtain the predicted Y values ($\hat{Y}_i$) after estimating the model.
3. Estimate the auxiliary regression, $\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i$, using OLS.
4. Retain the R-squared value $R_{\hat{\varepsilon}^2}^2$, from this auxiliary regression.
5. Calculate the F-statistic, $F = \frac{\frac{R_{\hat{\varepsilon}^2}^2}{1}}{\frac{(1 - R_{\hat{\varepsilon}^2}^2)}{n-2}}$, or the chi-squared statistic, $\chi^2 = nR_{\hat{\varepsilon}^2}^2$. If either of these test statistics is significant, then you have evidence of heteroskedasticity.

- **The White test**. The White test assumes that heteroskedasticity may be a linear function of all the independent variables, a function of their squared values, and a function of their cross products:

$$\hat{\varepsilon}_i^2 = \alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip} + \alpha_{p+1} X_{i1}^2 + \cdots + \alpha_{2p} X_{ip}^2 + \alpha_{2p+1}(X_{i1} X_{i2}) + \cdots + u_i,$$

where $\hat{\varepsilon}_i^2$ are calculated from the residuals and used as proxies for ε_i^2. Alternatively, a White test can be performed by estimating $\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i + \delta_2 \hat{Y}_i^2$ where $\hat{Y}_i$ represents the predicted values from $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \cdots + \hat{\beta}_p X_{ip}$. Here's how to perform a White test:

1. Estimate your model, $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$, using OLS.
2. Obtain the predicted Y values ($\hat{Y}_i$) after estimating your model.
3. Estimate the model $\hat{\varepsilon}_i^2 = \delta_0 + \delta_1 \hat{Y}_i + \delta_2 \hat{Y}_i^2$ using OLS.
4. Retain the R-squared value ($R^2_{\hat{\varepsilon}^2}$) from this regression.
5. Calculate the F-statistic, $F = \frac{\frac{R^2_{\hat{\varepsilon}^2}}{2}}{\frac{(1-R^2_{\hat{\varepsilon}^2})}{n-3}}$, or the chi-squared statistic, $\chi^2 = nR^2_{\hat{\varepsilon}^2}$. If either of these test statistics is significant, then you have evidence of heteroskedasticity.

• **The Goldfeld-Quandt test**. The Goldfeld-Quandt (GQ) test begins by assuming that a defining point exists and can be used to differentiate the variance of the error term. Sample observations are divided into two groups, and evidence of heteroskedasticity is based on a comparison of the residual sum of squares (RSS) using the F-statistic.

1. Estimate your model separately for each group and obtain the residual sum of squares for Group A (RSS_A) and the residual sum of squares for Group B (RSS_B).
2. Compute the F-statistic by

$$F = \frac{\frac{RSS_A}{n-p-1}}{\frac{RSS_B}{n-p-1}}.$$

The null hypothesis for the GQ test is homoskedasticity. The larger the F-statistic, the more evidence you'll have against the homoskedasticity assumption.

• **The Park test**. The Park test assumes that the heteorskedasticity may be proportional to some power of an independent variable (X_k) in the model: $\sigma^2_{i\varepsilon} = \sigma^2_\varepsilon X^\alpha_{ik}$.

1. Estimate the model $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$ using OLS.
2. Obtain the squared residuals, $\hat{\varepsilon}_i^2$, after estimating your model.
3. Estimate the model $\ln \hat{\varepsilon}_i^2 = \gamma + \alpha \ln X_{ik} + u_i$ using OLS.
4. Examine the statistical significance of α using the t-statistic: $t = \frac{\hat{\alpha}}{\hat{\sigma}_{\hat{\alpha}}}$. If the estimate of α coefficient is statistically significant, then you have evidence of heteroskedasticity.

Correcting your regression model for the presence of heteroskedasticity.

• **Weighted least squares (WLS)**. The goal of the WLS transformation is to make the error term in the original econometric model homoskedastic. First, you assume that the heteroskedasticity is determined proportionally from some function of the independent variables: $Var(\varepsilon|\mathbf{X}_i) = \sigma^2_\varepsilon h(\mathbf{X}_i)$. Then you use knowledge of this relationship to divide both sides of the original model by the component of heteroskedasticity that give the error term a constant variance. More specifically, the objective of OLS is

$$\min \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip} \right)^2 .$$

The objective of WLS is

$$\min \sum \frac{\left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_{i1} - \cdots - \hat{\beta}_p X_{ip} \right)^2}{h(\mathbf{X}_i)} .$$

In practice, knowing the exact functional form of $h(\mathbf{X}_i)$ is impossible. In applied settings, the exponential function is the most common approach to modeling heteroskedasticity: $Var(\varepsilon|\mathbf{X}_i) = \sigma^2_\varepsilon \exp(\alpha_0 + \alpha_1 X_{i1} + \cdots + \alpha_p X_{ip})$.

1. Estimate the original model, $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$, and obtain the residuals, $\hat{\varepsilon}_i$.
2. Square the residuals and take their natural log to generate $\ln \hat{\varepsilon}_i^2$.
3. Estimate the regression $\ln \hat{\varepsilon}_i^2 = \gamma + \delta_1 X_{i1} + \cdots + \delta_p X_{ip} + v_i$ or $\ln \hat{\varepsilon}_i^2 = \gamma + \phi_1 \hat{Y}_i + \phi_2 \hat{Y}_i^2 + u_i$ and obtain the fitted values: $\hat{g}_i = \hat{\gamma} + \hat{\phi}_1 \hat{Y}_i + \hat{\phi}_2 \hat{Y}_i^2$.
4. Take the inverse natural log of the fitted residuals $\exp(\hat{g}_i)$ to obtain $\hat{h}_i$.
5. Estimate the regression $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$ by WLS using $\hat{h}_i$ as weights.

If the proposed model of heteroskedasticity is misspecified, then WLS may not be more efficient than OLS. The problem is that misspecificaiton of the heteroskedasticity is difficult to identify. A large difference between OLS and WLS coefficients is more likely to imply that the model suffers from functional form specification bias than to suffer from heteroskedasticity.

• **Robust standard errors (White-corrected standard errors, heteroskedasticity-corrected standard errors)**. In a model with one independent variable and homoskedasticity, the variance of the estimator can be reduced to $Var(\hat{\beta}_1) = \sigma_\varepsilon^2 \sum c_i^2$; with heteroskedasticity, the variance of the estimator is $Var(\hat{\beta}_i) = \sum c_i^2 \sigma_{i\varepsilon}^2$. In applied settings, the squared residuals ($\hat{\varepsilon}_i^2$) are used as estimates of $\sigma_{i\varepsilon}^2$.

In a model with one independent variable, the robust standard error is

$$se(\hat{\beta}_i)_{HC} = \sqrt{\frac{\sum (X_i - \overline{X})^2 \hat{\varepsilon}_i^2}{\left(\sum (X_i - \overline{X})^2\right)^2}}.$$

Generalizing this result to a multiple regression model, the robust standard error is

$$se(\hat{\beta}_k)_{HC} = \sqrt{\frac{\sum \hat{\omega}_{ik}^2 \hat{\varepsilon}_i^2}{\left(\sum \hat{\omega}_{ik}^2\right)^2}}$$

where the $\hat{\omega}_{ik}^2$ are the residuals obtained from the auxiliary regression of X_j on all the other independent variables. Here's how to calculate robust standard errors:

1. Estimate your original multivariate model, $Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$, and obtain the squared residuals, $\hat{\varepsilon}_i^2$.

2. Estimate p auxiliary regressions of each independent variable on all the other independent variables and retain all p squared residuals ($\hat{\omega}_{ik}^2$).

3. For any independent variable, calculate the robust standard errors:

$$se(\hat{\beta}_k)_{HC} = \sqrt{\frac{\sum \hat{\omega}_{ik}^2 \hat{\varepsilon}_i^2}{\left(\sum \hat{\omega}_{ik}^2\right)^2}}.$$

Numerous versions of robust standard errors exist for the purpose of improving the statistical properties of the heteroskedasticity correction; no form of robust standard error is preferred above all others.

12 Autocorrelation

Patterns of autocorrelation. The CLRM assumes there's no autocorrelation: $Cov(\varepsilon_t, \varepsilon_s) = 0$ or $Corr(\varepsilon_t, \varepsilon_s) = 0$ for all $t \neq s$. When the error term exhibits no autocorrelation, the positive and negative error values are random.

If autocorrelation is present, positive autocorrelation is the most likely outcome. *Positive autocorrelation* occurs when an error of a given sign tends to be followed by an error of the same sign, which is called *sequencing. Negative autocorrelation* occurs when an error of a given sign tends to be followed by an error of the opposite sign, which is called *switching.*

When you're drawing conclusions about autocorrelation using the error pattern, all other CLRM assumptions must hold, especially the assumption that the model is correctly specified. If a model isn't correctly specified, you may mistakenly identify the model as suffering from autocorrelation. Misspecification is a more serious issue than autocorrelation.

Effect of autoregressive errors. In the presence of autocorrelation, the OLS estimators may not be efficient. In addition, the estimated standard errors of the coefficients are biased, which results in unreliable hypothesis tests (t-statistics). The OLS estimates, however, remain unbiased.

Typically, autocorrelation is assumed to be represented by a *first-order autoregression*:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ti} + \varepsilon_t$$

with

$$\varepsilon_t = \rho\varepsilon_{t-1} + u_t,$$

where $-1 < \rho < 1$ and u_t is a random error that satisfies the CLRM assumptions; namely $E[u_t|\varepsilon_{t-1}] = 0$, $Var(u_t|\varepsilon_{t-1}) = \sigma_u^2$, and $Cov(u_t, u_s) = 0$ for all $t \neq s$.

By repeated substitution, we obtain

$$\varepsilon_t = u_t + \rho u_{t-1} + \rho^2 u_{t-2} + \rho^3 u_{t-3} + \cdots .$$

Therefore

$$E[\varepsilon_t] = 0,\ Var(\varepsilon_t) = \sigma_u^2 + \rho^2\sigma_u^2 + \rho^4\sigma_u^2 + \cdots = \frac{\sigma_u^2}{1-\rho^2}.$$

The stationarity assumption ($|\rho| < 1$) is necessary to constrain the variance from becoming an infinite value. OLS assumes no autocorrelation; that is, $\rho = 0$ in the expression $\sigma_\varepsilon^2 = \frac{\sigma_u^2}{1-\rho^2}$. Consequently, in the presence of autocorrelation, the estimated variances and standard errors from OLS are underestimated.

Test for autocorrelation.

• **Graphical inspection of residuals**. Look for *sequencing* or *switching* of residual errors if autocorrelation is present.

• **The run test (the Geary test)**. You want to use the run test if you're uncertain about the nature of the autoregressive process (no assumptions about the ρ values).

A *run* is defined as a sequence of positive or negative residuals. The hypothesis of no autocorrelation isn't sustainable if the residuals have too many or too few runs.

The most common version of the test assumes that runs are distributed normally. If the assumption of no autocorrelation is sustainable, with 95% confidence, the number of runs should be between

$$\mu_r \pm 1.96\sigma_r$$

where μ_r is the expected number of runs and σ_r is the standard deviation. These values are calculated by

$$\mu_r = \frac{2T_1T_2}{T_1+T_2} + 1,\ \sigma_r = \sqrt{\frac{2T_1T_2(2T_1T_2 - T_1 - T_2)}{(T_1+T_2)^2(T_1+T_2-1)}}$$

where r is the number of observed runs, T_1 is the number of positive residuals, T_2 is the number of negative residuals, and T is the total number of observations.

If the number of observed runs is below the expected interval, it's evidence of positive autocorrelation; if the number of runs exceeds the upper bound of the expected interval, it provides evidence of negative autocorrelation.

• **The Durbin-Watson test for $AR(1)$ processes**. The Durbin-Watson (DW) test begins by assuming that if autocorrelation is present, then it can be described by an $AR(1)$ process:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ti} + \varepsilon_t,\ \varepsilon_t = \rho\varepsilon_{t-1} + u_t.$$

The value produced by the DW test is called *d statistic* and is calculated as follows:

$$d = \frac{\sum_{t=2}^{T}(\hat\varepsilon_t - \hat\varepsilon_{t-1})^2}{\sum_{t=1}^{T}\hat\varepsilon_t^2} = \frac{\sum_{t=2}^{T}\hat\varepsilon_t^2}{\sum_{t=1}^{T}\hat\varepsilon_t^2} + \frac{\sum_{t=2}^{T}\hat\varepsilon_{t-1}^2}{\sum_{t=1}^{T}\hat\varepsilon_t^2} - \frac{2\sum_{t=2}^{T}\hat\varepsilon_t\hat\varepsilon_{t-1}}{\sum_{t=1}^{T}\hat\varepsilon_t^2} \approx 1 + 1 - \frac{2\frac{\hat\rho\hat\sigma_u^2}{1-\hat\rho^2}}{\frac{\hat\sigma_u^2}{1-\hat\rho^2}} \approx 2(1-\hat\rho).$$

where T represents the last observation in the time series.

From the approximate formula $d \approx 2(1-\hat\rho)$, the closer d is to 2, the stronger the evidence of no autocorrelation; the closer d is to 0, the more likely positive autocorrelation. If d is closer to 4, then no autocorrelation is rejected in favor of negative autocorrelation.

The DW test has no unique critical value defining the point at which you reject the null hypothesis of no autocorrelation. However, it does have a zone of indecision defined by a lower bound (d_L) and upper bound (d_u) that depend on the number of observations and the number of estimated coefficients ($p+1$) in the original model:

$$\begin{cases} \text{Reject } H_0\text{: } \rho > 0 & 0 < d < d_L \\ \text{indecision} & d_L \leq d \leq d_U \\ \text{Fail to reject } H_0\text{: No autocorrelation} & d_U < d < 4 - d_L \\ \text{indecision} & 4 - d_U \leq d \leq 4 - d_L \\ \text{Reject } H_0\text{: } \rho < 0 & 4 - d_L < d < 4. \end{cases}$$

The DW d-statistic is the most popular test for autocorrelation, but it's limited to identifying $AR(1)$ autocorrelation. It's a good initial test, but additional testing may be required to rule out other forms of autocorrelation. Furthermore, a d-statistic that ends up in the indecision zone requires an alternative test to achieve a more conclusive result.

• **The Breusch-Godfrey test for $AR(q)$ processes**. The Breusch-Godfrey (BG) test begins by assuming that if autocorrelation is present, then it can be described by an $AR(q)$ process:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ti} + \varepsilon_t, \ \varepsilon_t = \sum_{j=1}^{q} \rho_j \varepsilon_{1-j} + u_t$$

where $1 \leq q < T$. A special case of this test with $q = 1$ is known as *Durbin's alternative statistic.*

You can perform a BG test by following these steps:

1. Estimate the model $Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ti} + \varepsilon_t$ using OLS.
2. Obtain the residual values, $\hat{\varepsilon}_t$, after estimating your model.
3. Estimate the auxiliary regression $\hat{\varepsilon}_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i X_{ti} + \sum_{j=1}^{q} \rho_j \hat{\varepsilon}_{t-j} + u_t$ using OLS.
4. Retain the R-squared value, $R^2_{\hat{\varepsilon}}$, from this regression.
5. Calculate the F-statistic for joint significance of $\hat{\rho}_1$, $\hat{\rho}_2$, $\cdots$, and $\hat{\rho}_q$ or the chi-squared statistic $\chi^2 = (n-q)R^2_{\hat{\varepsilon}}$ with q degrees of freedom.

If the F or chi-squared test statistics are significant, then you have evidence of autocorrelation. If not, you fail to reject the null hypothesis of no autocorrelation, which is $H_0 : \rho_1 = \rho_2 = \cdots = \rho_q = 0$.

Remedying harmful autocorrelation.

• **Feasible generalized least squares (FGLS)**. There are two FGLS techniques: the Cochrane-Orcutt (CO) transformation and the Prais-Winsten (PW) transformation. They transform the original model with autocorrelation into one without autocorrelation by *quasi-differencing.* If the proposed $AR(1)$ model of autocorrelation, $\varepsilon_t = \rho \varepsilon_{t-1} + u_t$, isn't correct, then you have no guarantee of getting more accurate standard errors with FGLS thanOLS.

Here's how to apply either the CO or PW technique:

1. Estimate your original model, $Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ti} + \varepsilon_t$, and obtain the residuals $\hat{\varepsilon}_t$.
2. Use the residuals to estimate ρ by performing one of the following calculations:

✠ $\hat{\rho} = \frac{\sum_{t=2}^{T} \hat{\varepsilon}_t \hat{\varepsilon}_{t-1}}{\sum_{t=1}^{T} \hat{\varepsilon}_t^2}$. This calculation can be used in large samples but may have significant error in smaller samples.

✠ $\hat{\rho} = 1 - \frac{d}{2}$. This calculation, known as *Thiel's estimator*, can be used with smaller samples.

✠ Estimate $\hat{\varepsilon}_t = \rho \hat{\varepsilon}_{t-1} + u_t$ and obtain $\hat{\rho}$ from the regression. This method is the most common for estimating ρ but is recommended only with larger samples.

In practice, knowing the exact value of ρ is impossible. In applied settings, you use the estimated value for ρ (that is, $\hat{\rho}$) to transform the model.

3. Estimate the quasi-differenced CO or PW regression using $\hat{\rho}$ in place of ρ:

$$Y_t - \rho Y_{t-1} = \beta_0(1-\rho) + \sum_{i=1}^{p} \beta_i (X_{ti} - X_{(t-1)i}) + u_t$$

The CO transformation sets $Y_t^* = Y_t - \rho Y_{t-1}$ and $\varepsilon_t^* = u_t$ so that the regression equation becomes

$$Y_t^* = \beta_0^* + \sum_{i=1}^{p} \beta_i^* X_{ti}^* + \varepsilon_t^*.$$

The PW transformation maintains the CO structure with the exception of the first observation:

$$Y_1^* = (\sqrt{1-\rho^2})Y_1,\ X_1^* = (\sqrt{1-\rho^2})X_1,\ \varepsilon_t^* = (\sqrt{1-\rho^2})\varepsilon_1.$$

In large samples, the difference between the CO and PW estimates is usually small. In small samples, however, the difference can be significant.

• **Serial correlation robust standard errors**. Estimating the model using OLS and adjusting the standard errors for autocorrelation has become more popular than other correction methods. There are two reasons for this: (1) The serial correlation robust standard errors can adjust the results in the presence of a basic $AR(1)$ process or a more complex $AR(q)$ process, and (2) only the biased portion of the results (the standard errors) are adjusted, while the unbiased estimates (the coefficients) are untouched, so no model transformation is required.

Adjusting the OLS standard errors for autocorrelation produces *serial correlation robust standard errors.* These are also referred to as *Newey-West (NW) standard errors* and can be calculated by applying the following steps:

1. Estimate your original model $Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ti} + \varepsilon_t$ and obtain the residuals: $\hat{\varepsilon}_t$.
2. Estimate the auxiliary regression $X_{t1} = \alpha_0 + \sum_{i=2}^{p} \alpha_i X_{ti} + r_t$ and retain the residuals: $\hat{r}_t$.
3. Find the intermediate adjustment factor, $\hat{\alpha}_t = \hat{r}_t \hat{\varepsilon}_t$, and decide how much serial correlation (the number of lags) you're going to allow. A Breusch-Godfrey test can be useful in making this determination.
4. Obtain the error variance adjustment factor, $\hat{v} = \sum_{t=1}^{T} \hat{\alpha}_t^2 + 2\sum_{h=1}^{g} \left[1 - \frac{h}{g+1}\right] \left(\sum_{t=h+1}^{T} \hat{\alpha}_t \hat{\alpha}_{t-h}\right)$, where g represents the number of lags determined in Step 3.
5. Calculate the serial correlation robust standard error, which is also known as the *heteroskedasticity-autocorrelation-corrected* (HAC) standard error because the calculation simultaneously adjusts the standard error for heteroskedasticity and autocorrelation. For variable X_1,

$$se(\hat{\beta}_1)_{HAC} = \left(\frac{se(\hat{\beta}_1)}{\hat{\sigma}_\varepsilon}\right)^2 \sqrt{\hat{v}}.$$

6. Repeat Steps 2 through 5 for independent variables X_2 through X_p.

Part V
Discrete and Restricted Dependent Variables in Econometrics

13 Qualitative Dependent Variables

The Linear Probability Model (LPM). A basic LPM can be expressed as $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, where Y is a dummy variable that is equal to 1 if a particular outcome is observed and 0 otherwise. As usual, the predicted value $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is interpreted as an estimate of $E[Y|X]$.

Although OLS estimation always produces the typical R-squared measure of fit, its interpretation is less meaningful when all the values of the dependent variable are at 0 or 1. In the case of an LPM, more appropriate measures of fit capture the fraction of times the model predicts accurately:

• Accurate prediction defined as (a) $\hat{P}_i \geq 0.5$ and $Y = 1$ or (b) $\hat{P}_i < 0.5$ and $Y = 0$. Accurate predictions aggregated by calculating the total number of accurate predictions as a percentage of the total number of observations.

- Accurate prediction defined as (a) $\hat{P}_i \geq \overline{Y}$ and $Y = 1$ or (b) $\hat{P}_i < \overline{Y}$ and $Y = 0$. Accurate predictions aggregated by calculating the total number of accurate predictions as a percentage of the total number of observations.
- Accurate prediction defined as $\hat{P}_i \geq 0.5$ and $Y = 1$ or $\hat{P}_i < 0.5$ and $Y = 0$. Accurate predictions aggregated by calculating the percent of accurte predictions in each group (for $Y = 0$ and $Y = 1$) and weighting the percent of observations in each group.
- Accurate prediction defined as $\hat{P}_i \geq \overline{Y}$ and $Y = 1$ or $\hat{P}_i < \overline{Y}$ and $Y = 0$. Accurate predictions aggregated by calculating the percent of accurate predictions in each group (for $Y = 0$ and $Y = 1$) and weighting the percent of observations in each group.

The three main LPM problems.

- **Non-normality of the error term**. The assumption that the error is normally distributed is critical for performing hypothesis tests. The error term of an LPM has a binomial distribution instead of a normal distribution. It implies that the traditional t-tests for individual significance and F-test for overall significance are invalid.
- **Heteroskedasticity**. The assumption of homoskedasticity is required to prove that the OLS estimators are efficient. The presence of heteroskedasticity can cause the Gauss-Markov theorem to be violated and lead to other undesirable characteristics for the OLS estimators. The error term in an LPM is heteroskedastic because its variance isn't constant:

$$Var(\varepsilon_i) = (\beta_0 + \beta_1 X_i)(1 - \beta_0 - \beta_1 X_i).$$

- **Unbounded predicted probabilities**.

The probit and logit models. In a probit or logit model, we estimate

$$E[Y|X_i] = P(Y = 1|X_i) = F(\beta_0 + \beta_1 X_i),$$

where F is a monotone increasing function with range $(0, 1)$. For a probit model, F is the CDF of a standard normal: $F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\xi^2/2} d\xi$; for a logit model, $F(x) = \frac{e^x}{1+e^x}$.

Probit and logit functions are both nonlinear in parameters, so OLS can't be used to estimate the βs. Instead, we use maximum likelihood estimation: we solve for

$$\arg\max_{\beta_0,\beta_1} (\text{probability of observing } Y_1, \cdots, Y_n) = \arg\max_{\beta_0,\beta_1} \prod_{i=1}^{n} F(\beta_0 + \beta_1 X_i)^{Y_i} [1 - F(\beta_0 + \beta_1 X_i)]^{1-Y_i}.$$

Finding the optimal values for the $\hat{\beta}$ terms requires solving the following first-order conditions

$$\begin{cases} \frac{\partial \ln \hat{L}}{\partial \hat{\beta}_0} = \sum_{i=1}^{n} \left[\frac{Y_i F'(\hat{\beta}_0 + \hat{\beta}_1 X_i)}{F(\hat{\beta}_0 + \hat{\beta}_1 X_i)} - \frac{(1-Y_i) F'(\hat{\beta}_0 + \hat{\beta}_1 X_i)}{1-F(\hat{\beta}_0 + \hat{\beta}_1 X_i)} \right] = 0 \\ \frac{\partial \ln \hat{L}}{\partial \hat{\beta}_1} = \sum_{i=1}^{n} \left[\frac{Y_i F'(\hat{\beta}_0 + \hat{\beta}_1 X_i)}{F(\hat{\beta}_0 + \hat{\beta}_1 X_i)} - \frac{(1-Y_i) F'(\hat{\beta}_0 + \hat{\beta}_1 X_i)}{1-F(\hat{\beta}_0 + \hat{\beta}_1 X_i)} \right] X_i = 0 \end{cases}$$

Probit and logit estimation always produces a *Pseudo R-squared* measure of fit: $\widetilde{R}^2 = 1 - \frac{\ln \hat{L}_{ur}}{\ln \hat{L}_0}$, where $\ln L_{ur}$ is the log likelihood for the estimated model and $\ln L_0$ is the log likelihood in the model with only an intercept.

You can obtain more appropriate measures of fit for probit and logit models by comparing the model's predicted probabilities to the observed Y values. Appropriate measures of fit typically capture the fraction of times the model accurately predicts the outcome, e.g. the four measures of fit used for the LPM.

14 Limited Dependent Variable Models

Limited dependent variables.

- **Censored dependent variables**. With a *censored dependent variable*, some of the actual values for the dependent variable are limited to a minimum and/or maximum threshold value. This leads to nonzero conditional mean of the error and correlation between the value of the error and the value of the independent variable.

• **Truncated dependent variables**. With a *truncated dependent variable*, some of the values for the variables are missing (meaning they aren't observed if they are above or below some threshold). Sometimes observations included in the sample have missing values for both the independent and dependent variables, and in other cases only the values for the dependent variable are missing. Common scenarios resulting in truncation include *nonrandom sample selection* and *self-selection*.

Truncated data leads to nonzero conditional mean of the error and correlation between the value of the error and the value of the independent variable.

The primary difference between a truncated and a censored variable is that the value of a truncated variable isn't observed at all. However, a value is observed for a censored variable, but it's suppressed for some observations at the threshold point.

Regression analysis for limited dependent variables.

• **Tobin's Tobit for censored dependent variables**. If you use OLS estimation with the observed data as if they're all uncensored values, you get biased coefficients. To avoid them, the estimation procedure must properly account for the censoring of the dependent variable. Maximum likelihood (ML) estimation does so.

Suppose you have the following model with upper-limit censoring (the most common type):

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \varepsilon \sim N(0, \sigma_\varepsilon^2),\ Y_i = \begin{cases} Y_i^* & Y_i^* < b \\ b & Y_i^* \geq b. \end{cases}$$

Using the probability of censorship, estimation is accomplished with ML, where the log likelihood function to be maximized is

$$\ln L = \sum_{i=1}^{n} \left\{ \ln F\left(\frac{\beta_0 + \beta_1 X_i - b}{\sigma_\varepsilon}\right) + \ln\left[\frac{1}{\sigma_\varepsilon} F'\left(\frac{Y_i - \beta_0 - \beta_1 X_i}{\sigma_\varepsilon}\right)\right]\right\}$$

where F denotes the standard normal CDF

Tobit estimation produces a likelihood ratio chi-squared statistic. It's analogous to the F-statistic in OLS, and it tests the null hypothesis that the estimated model doesn't produce a higher likelihood than a model with only a constant term.

• **Truncated regression for truncated dependent variables with unobserved independent variables**. In this case, you can't apply OLS estimation to the observed data as if it's representative of the entire population. If you do, you'll wind up with biased coefficients. Instead, you need to use maximum likelihood (ML) estimation so you can properly account for the truncation by rescaling the normal distribution so that the cumulative probabilities add up to one over the restricted area.

Consider the following model

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \varepsilon \sim N(0, \sigma_\varepsilon^2),\ Y_i = \begin{cases} Y_i^* & Y_i^* < b \\ \cdot & Y_i^* \geq b. \end{cases}$$

The dot $(\cdot)$ represents a missing value at and above the truncation point. Using a rescaling of the normal distribution, estimation is accomplished with ML, where the log likelihood function to be maximized is

$$\ln L = -\frac{n}{2}\ln(2\pi\sigma_\varepsilon^2) - \frac{1}{2\sigma_\varepsilon^2}\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i)^2 - \sum_{i=1}^{n} \ln F\left(\frac{b - \beta_0 - \beta_1 X_i}{\sigma_\varepsilon}\right)$$

where F denotes the standard normal CDF.

Truncated normal estimation also produces a chi-squared statistic, which is like the F-statistic in OLS. It confirms or rejects the null hypothesis that the estimated model doesn't produce a higher likelihood than a model with only a constant term.

Ignoring the truncation and estimating the model using OLS will produce coefficients biased toward finding no relationship (smaller coefficients/effects).

• **Heckman's selection bias correction for truncated dependent variables with observed independent variables**. Assume we work with the following model:

$$Y_i^* = \beta_0 + \beta_1 X_i + \varepsilon_i,\ \varepsilon \sim N(0, \sigma_\varepsilon^2)$$

with self-selection defined by

$$S_i = \gamma_0 + \gamma_1 W_{i1} + \gamma_2 W_{i2} + \cdots + u_i,\ S_i = \begin{cases} 1 & \text{if } Y_i^* \text{ observed} \\ 0 & \text{if } Y_i^* \text{ not observed,} \end{cases}\ u \sim N(0,1),\ Corr(\varepsilon, u) = \rho.$$

The log likelihood function that's maximized is

$$\begin{aligned} \ln L &= \sum_{i=1}^{n} \left\{ \ln F \left[\frac{((\gamma_0 + \gamma_1 W_{i1} + \gamma_2 W_{i2} + \cdots) + (Y_i^* - \beta_0 - \beta_1 X_i)\rho)/\sigma_\varepsilon}{\sqrt{1-\rho^2}} \right] \right. \\ &\left. -\frac{1}{2} \left(\frac{Y_i^* - \beta_0 - \beta_1 X_i}{\sigma_\varepsilon} \right)^2 - \ln(\sqrt{2\pi}\sigma_\varepsilon) + \ln F(-\gamma_0 - \gamma_1 W_{i1} - \gamma_2 W_{i2} - \cdots) \right\} \end{aligned}$$

where F denotes the standard normal CDF. In a Heckman model, the variables that influence truncation usually aren't identical to those that influence the value of the dependent variable (in contrast to the Tobit model, where they're assumed to be the same).

Sometimes the ML estimation fails to converge, and an alternative is to use the Heckit model. It can be accomplished by following these steps:

1. Estimate the selection equation $S_i = \gamma_0 + \gamma_1 W_{i1} + \gamma_2 W_{i2} + \cdots + u$ with a probit model.
2. Compute the inverse Mills ratio:

$$\hat{\lambda}_i = \frac{F'(\hat{\gamma}_0 + \hat{\gamma}_1 W_{i1} + \hat{\gamma}_2 W_{i2} + \cdots)}{F(\hat{\gamma}_0 + \hat{\gamma}_1 W_{i1} + \hat{\gamma}_2 W_{i2} + \cdots)}$$

where F is the standard normal CDF.

3. Estimate the model $Y_i = \beta_0 + \beta_1 X_i + \beta_2 \hat{\lambda}_i + \varepsilon_i$ using the selected sample.

Estimation of a Heckman selection model also produces a chi-squared statistic, which is similar to the F-statistic in OLS and tests the null hypothesis that esttimated model doesn't produce a higher likelihood than a model with only a constant term.

Part VI
Extending the Basic Econometric Model

15 Static and Dynamic Models

Using contemporaneous and lagged variables in regression analysis.

• **Problems with dynamic models**. When you're using time-series data, you can assume that the independent variables have a contemporaneous (static) or lagged (dynamic) effect on our dependent variable. A generic dynamic model is a *distributed lag model*. You can specify it as

$$Y_t = \alpha + \delta_0 X_t + \delta_1 X_{t-1} + \delta_2 X_{t-2} + \cdots + \delta_r X_{t-r} + \varepsilon_t.$$

In practice, distributed lag models can be plagued by estimation problems. The two most common issues are high multicollinearity and the loss of degrees of freedom: high multicollinearity usually causes the coefficient estimates to display erratic behavior, while loss of degrees of freedom increases the standard errors and reduces the chances of finding statistically significant coefficients.

A common solution to the estimation issues is to replace the lagged values of the independent variable with a lagged value of the dependent variable - an autoregressive model like $Y_t = \alpha + \delta X_t + \gamma Y_{t-1} + \varepsilon_t$. Using recursive substitution, we can show that the autoregressive model is equivalent to the distributed lag model.

The distributed lag estimates suffer from unpredictable shifts in the parameter estimates because they're plagued by high collinearity. Therefore, when estimating dynamic models, applied econometricians prefer the autoregressive model to the distributed lag model.

• **Testing and correcting for autocorrelation in dynamic models**. Autocorrelation in a dynamic model causes the OLS coefficients to be biased. Because econometricians view biased coefficients to be more problematic than biased standard errors, testing for autocorrelation is essential if you're estimating a dynamic model. Turn to the Breusch-Godfrey test in this scenario. Avoid using the Durbin-Watson d statistic when you're estimating a dynamic time-series model since in a dynamic model, the Durbin-Watson d statistic is biased toward 2 (that is, finding no autocorrelation).

If you find evidence of autocorrelation, you can perform the preferred method of autocorrelation correction with dynamic models: feasible generalized least squares (FGLS).

Projecting time trends with OLS. If the dependent variable has a relatively steady increase over time, your best bet is to model the relationship with a linear time trend $Y_t = \alpha_0 + \alpha_1 t + \varepsilon_t$; if the growth rate is fairly steady, then you need to model the relationship with an exponential time trend $\ln Y_t = \alpha_0 + \alpha_1 t + \varepsilon_t$.

• **Spurious correlation and time series**. If your regression model contains dependent and independent variables that are trending, then you end up with a *spurious correlation problem.* This is because if time significantly explains variation in the dependent variable and is also correlated with your independent variable, then you've excluded a relevant variable from your model and you'll overstate the explanatory power of your independent variables.

Adding some form of time trend component to your regression takes care of the spurious correlation problem.

• **Detrending time-series data**. The main point of estimating a regression model with detrended data is to derive the explanatory power of the other independent variables. Here's how to obtain the goodness-of-fit, or R-squared, net of trend effects:

1. Regress your dependent variable on the trend variable to obtain the estimated function $Y_t = \hat{\alpha}_0 + \hat{\alpha}_1 t + \hat{\varepsilon}_{tY}$ and retain the residuals from this regression.

2. Regress each of your independent variables on the trend variable to obtain the estimated functions $X_{tk} = \hat{\alpha}_{0k} + \hat{\alpha}_{1k} t + \hat{\varepsilon}_{tX_k}$, where k represents a specific independent variable, and retain the residuals from all k of these regressions.

3. Regress the residuals obtained in Step 1 on the residuals obtained in Step 2 to estimate $\hat{\varepsilon}_{tY} = \beta_0 + \beta_1 \hat{\varepsilon}_{tX_k} + u_t$.

The R-squared from this regression provides a better measure of fit when the time series exhibits extensive trending.

Using OLS for seasonal adjustments. The higher the frequency of an economic time series, the more likely it is to display seasoned patterns. The most common models capturing seasonal patterns include dummy variables representing the frequency with which the data were collected (usually quarter or month dummies): $Y_t = \alpha_0 + \alpha_1 S_1 + \alpha_2 S_2 + \cdots + \varepsilon_t$, where S variables are your season dummy variables.

• **Estimating seasonality effects**. Seasonally effects can be correlated with both your dependent and independent variables. If you include dummy variables for seasons along with the other relevant independent variables, you can simultaneously obtain better estimates of both seasonality and the effects of the other independent variables, and make more convincing arguments about the causal relationship between your independent variables and dependent variables:

$$Y_t = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ti} + \sum_{j=1}^{q} \lambda_i S_i + \varepsilon_t.$$

• **Deseasonalizing time-series data**.

1. Regress your dependent variable on the seasonal dummy variables to obtain the estimated function $Y_t = \hat{\alpha}_0 = \hat{\alpha}_0 + \sum_{j=1}^{q} \hat{\alpha}_j S_j + \hat{\varepsilon}_{tY}$ and retain the residuals from this regression.

2. Regress each of your independent variables on the seasonal dummy variables to obtain the estimated functions $X_{tk} = \hat{\alpha}_{0k} + \sum_{j=1}^{q} \hat{\alpha}_{jk} S_j + \hat{\varepsilon}_{tX_k}$, where k represents a specific independent variable, and retain the residuals from all k of these regressions.

3. Regress the residuals obtained in Step 1 on the residuals obtained in Step 2 to estimate $\hat{\varepsilon}_{tY} = \beta_0 + \beta_1 \hat{\varepsilon}_{tX_k} + u_t$.

The R-squared from this regression provides a better measure of fit when the time series exhibits considerable seasonality.

Econometricians mainly estimate the regression model with deseasonalized data to derive the explanatory power of the other independent variables. Your primary econometric results, however, should report the estimates from the model with the raw data and season dummy variables.

16 Diving into Pooled Cross-Section Analysis

A pooled cross section combines independent cross-sectional data that has been collected over time. Typically, pooled cross sections contain many more cross-sectional observations than the number of time periods being pooled. Consequently, the models usually resemble cross-sectional analysis with possible heteroskedasticity corrections. Because the time gap between the collection of cross-sectional units is usually large, autocorrelation and other time-series issues tend to be ignored.

Do not confuse a pooled cross section with a panel dataset. In a panel dataset the same cross-sectional units are included in each time period rather than being randomly selected in each period.

Including dummy variables in your model for each time period, except the *reference period*, allows you to identify changing parameter values:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \delta_1 R_{i1} + \delta_2 R_{i2} + \cdots + \varepsilon_i.$$

By examining the statistical significance of the estimated δ (or $\hat{\delta}$) terms, you can identify any shifts (whether up or down) in the relationship for a given period.

Adding time-period dummy variables interacted with the other independent variables allows you to identify both changing intercepts and slopes:

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \delta_0 R_i + \delta_1 (X_1 \cdot R)_i + \delta_2 (X_2 \cdot R)_i + \cdots + \varepsilon_i.$$

If you're interested in any distributional change that may have occurred in your population of interest between time periods, you can perform an F-test of joint significance for all the $\delta(\delta_0, \delta_1, \delta_2, \cdots)$ parameters. Essentially, this test identifies whether the time period has a collective influence on the intercept and/or impact of the independent variables. It's equivalent to performing a Chow test for structural stability.

17 Panel Econometrics

Examples of well-known panel datasets include the National Longitudinal Surveys (NLS), the Panel Study of Income Dynamics (PSID), and the Survey of Income and Program Participation (SIPP).

Estimating the uniqueness of each individual unit. Suppose the model that explains your outcome of interest is

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 w_{it} + \varepsilon_{it}$$

where X is an observable independent variable, and ω is an unobservable independent variable.

The danger with combining panel data and OLS estimation is that you may end up with results containing *heterogeneity bias*. The existence of unobservable factors that consistently impact your outcome of interest (Y variable) is likely with panel data, which means you need to consider using one of three estimation methods:

✓First difference (FD) transformation.

✓Dummy variable (DV) regression.
✓The fixed effects (FE) estimator (the method most commonly used by applied econometricians).

First difference (FD) transformation. In order to use the FD approach, we rely on a couple of assumptions. First, we assume that the values for the unobserved variable remain constant through time for a given subject, but vary across subjects; $\omega_{it} = \omega_i \ \forall t$. Second, we assume that the model doesn't change over time. Under these two assumptions, we can take the first difference (FD) of individual observations over time: $Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 w_{it} + \varepsilon_{it}$ and $Y_{it-1} = \delta_0 + \beta_1 X_{it} + \beta_2 w_{it} + \varepsilon_{it}$, and obtain

$$\Delta Y_i = Y_{it} - Y_{it-1} = (\beta_0 - \delta_0) + \beta_1 (X_{it} - X_{it-1}) + \beta_2 (\omega_{it} - \omega_{it-1}) + (\varepsilon_{it} - \varepsilon_{it-1}) = \alpha_0 + \beta_1 \Delta X_i + \Delta \varepsilon_i.$$

Dummy variable (DV) regression. A DV model can be represented as

$$Y_{it} = \sum_{i=1}^{n} \alpha_{i0} A_i + \sum_{k=1}^{p} \beta_k X_{it,k} + \varepsilon_{it}$$

where $A = 1$ for any observation that pertains to individual i and 0 otherwise.

Fixed effects (FE) estimator. FE estimation is applied by *time demeaning* the data. Demeaning deals with unobservable factors because it takes out any component that is constant over time. By assumption, that would be the entire amount of the unobservable variable. Typically, FE model also include *time effect* controls. You can add them by adding dummy variables for each time period in which cross-sectional observations were obtained.

Increasing the efficiency of estimation with random effects. If you have panel data, your econometric model can explicitly estimate the unobserved effects associated with your cross-sectional unit using the fixed effects (FE) model:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 \omega_{it} + \varepsilon_{it},$$

where $\omega_{it} = \omega_i$ are unobserved characteristics for each cross-sectional unit that don't vary over time. On the other hand, your econometric model can allow all unobserved effects to be relegated to the error term by specifying the model as

$$Y_{it} = \beta_0 + \beta_1 X_{it} + v_{it}$$

where $v_i t = \omega_{it} + \varepsilon_{it}$. This approach is known as the *random effects (RE) model.*

With panel data, the advantage of the RE model over the FE model is more efficient estimates of the regression parameters. The RE technique doesn't estimate the fixed effects separately for each cross-sectional unit, so you get fewer estimated parameters, increased degrees of freedom, and smaller standard errors. A critical assumption of the RE model is that the unobserved individual effect (ω_i) isn't correlated with the independent variable(s). In addition, for the homoskedasticity assumption to hold, we must also impose a constant variance on the individual effects.

Although ε_{it} satisfies the classical linear regression model (CLRM) assumptions, the inclusion of ω_i in the composite error $v_{it} = \omega_i + \varepsilon_{it}$ results in a CLRM assumption violation. If you relegate the individual effects (ω_i) to the error term, you create positive serial correlation in the composite error. As a result, RE estimation requires feasible generalized least squares (FGLS) rather than OLS to appropriately eliminate serial correlation in the error term and to produce the correct standard errors and test statistics.

Testing efficiency against consistency with the Hausman test. The RE model produces more efficient estimates than the FE model. However, if individual fixed effects are correlated with the independent variable(s), then the RE estimates will be biased. In that case, the FE estimates would be preferred. The Hausman test checks the RE assumptions and helps you decide between RE and FE estimation. Note if heteroskedasticity is present, the Hausman test results could be misleading.

In a model with one independent variable, the Haussman test statistic is defined as

$$H = \frac{(\hat{\beta}_{1(FE)} - \hat{\beta}_{1(RE)})^2}{\sigma^2_{\hat{\beta}_{1(FE)}} - \sigma^2_{\hat{\beta}_{1(RE)}}} \sim \chi^2_1$$

Part VII
The Part of Tens

18 Ten Components of a Good Econometrics Research Project

- *Introducing Your Topic and Posing the Primary Question of Interest.*
- *Discussing the Relevance and Importance of Your Topic.*
- *Reviewing the Existing Literature.* Sources for references include

✓Google Scholar (scholar.google.com) lets you search by keyword.

✓Social Science Research Network (www.ssrn.com) contains a repository of working papers with the latest research findings.

✓Economic Journals on the web (http://www.oswego.edu/∼economic/journals.htm) provides a list of economic journals.

✓EconLit (www.aeaweb.org/econlit/) lists sources of economic research and is available through most electronic resources of university libraries.

- *Describing the Conceptual or Theoretical Framework.* One of the characteristics that differentiates applied research in econometrics from other applications of statistical analysis is a theoretical structure supporting the empirical work, rather than focus only on the statistical fit between variables.

- *Explaining Your Econometric Model.* You should explain and justify any specification characteristics of the econometric model (logs, quadratic functions, qualitative dependent variables, and so on) that aren't directly addressed by the conceptual framework. This can be achieved with intuition, scatter plots, and/or conventions derived by researchers in previously published work.

If there are contesting theories, then you should explain whether this implies that you could end up with different estimates of the relationship between the variables in a single model or if you should estimate more than one model.

- *Discussing the Estimation Method(s).* Estimation problems arising from a failure of the CLRM assumptions are common in applied econometric research. It's usually a good idea to estimate your model using OLS to obtain baseline results, even if you ultimately decide to use a different estimation technique. You may find that the results are similar and OLS is the easiest to interpret.

- *Providing a Detailed Description of Your Data.*

✓How the dataset was acquired and its source(s)
✓The nature of the data (cross sectional, time series, or panel)
✓The time span covered by the data
✓How and with what frequency the data was collected
✓The number of observations present
✓Whether any observations were thrown out and why
✓Summary statistics for any variables used in your econometric model(s)

- *Constructing Tables and Graphs to Display Your Results.*
- *Interpreting the Reported Results.* Reporting your econometric results is not enough; you also need to decipher the results for your readers. The most important element is the evaluation of statistical significance and magnitude for the primary variables of interest. The discussion should include an explanation of magnitude, directionality (positive/negative effects), statistical significance, and the relationship with the research question and theoretical hypotheses posed earlier in your paper.

- *Summarizing What You Learned.* Synthesize your results and explain how they're connected to your primary question. Avoid

✓focusing on variables with coefficients that are statistically significant even when the magnitude of their effect on the dependent variable is negligible (nearly no effect);

✓ignoring variables with statistically insignificant coefficients–finding no-relationship between variables is important when economic theory or the prevailing wisdom says differently.

19 Ten Common Mistakes in Applied Econometrics

- *Failing to Use Your Common Sense and Knowledge of Economic Theory.* One of the characteristics that differentiate applied research in econometrics from other applications of statistical analysis is the use of economic theory and common sense to motivate the connection between the independent and dependent variables.

- *Asking the Wrong Questions First.* Conceptual questions are more important to ask than technical ones.

- *Ignoring the Work and Contributions of Others.*

- *Failing to Familiarize Yourself with the Data.* Do some exploratory work that includes descriptive statistics, line charts (for time-series data), frequency distributions, and even listing of some individual data values. Notable issues include
  - ✓Variables you thought were measured continuously are actually in categories or groups.
  - ✓Measurements that you believed were real values are actually missing values.
  - ✓Data values that appear perfectly legitimate are actually censored values.

- *Making It Too Complicated.* The art of econometrics lies in finding the appropriate specification or functional form to model your particular outcome of interest. Given the uncertainty of choosing the "perfect" specification, many applied econometricians make the mistake of overspecifying their models or favor complicated estimation methods over more straightforward techniques. If theory and common sense aren't fairly conclusive about the hypothesized effect of a variable, it's probably best to refrain from including it. Consequently, additional sophistication in your model should be introduced as necessary and not simply to exhibit your econometric skills.

- *Being inflexible to Real-World Complications.* The *ceteris paribus* assumption often does not hold. Use proxies that seem appropriate and that others would find acceptable. Avoid forcing a particular dataset into estimation that isn't appropriate for the research question.

- *Looking the Other Way When You See Bizarre Results.* If some results don't pass a common-sense test, then the statistical tests are likely to be meaningless and may even indicate that you've made a mistake with your variables, the estimation technique, or both.

- *Obsessing over Measures of Fit and Statistical Significance.* The importance of your results shouldn't be determined on the basis of fit (R-squared values) or statistical significance alone. The primary finding in many of the best papers using econometrics involves findings of statistical insignificance.

- *Forgetting about Economic Significance.* The most important element in the discussion of your results is the evaluation of statistical significance *and* magnitude for the primary variables of interest. If a variable has a statistically significant coefficient but the magnitude is too small to be of any importance, then you should be clear about its lack of economic significance.

- *Assuming Your Results Are Robust.* You want to perform robustness (or sensitivity) analysis to show that your model estimates aren't sensitive (are robust) to slight variations in specification.

Part VIII

Appendices

A Specifying Your Econometrics Regression Model

As you define your regression model, you need to consider several elements:

- Economic theory, intuition, and common sense should all motivate your regression model.
- The most common regression estimation technique, ordinary least squares (OLS), obtains the best estimates of your model if the classical linear regression model (CLRM) assumptions hold.
- Assuming a normal distribution of the error term is important for hypothesis testing and prediction/forecasting.

When a regression model is estimated, prior to obtaining results, you need to provide a sound justification for the variables you've chosen.

The characteristics of the error term are of critical importance in econometrics. The assumption that the error term is normally distributed isn't required for performing OLS estimation, but it is necessary when you want to produce *confidence intervals* and/or perform *hypothesis tests* with your OLS estimates.

B Choosing the Functional Form of Your Regression Model

- Take the time to think through specification issues methodically.
- Explain why you've chosen specific independent variables for your model.
- Justify the functional form you've chosen for the model.
- Test the assumptions of the classical linear regression model (CLRM) and make change sot the model as necessary.
- Spend some time examining the sensitivity of your results by making slight modifications to the variables and the functional form of the relationship.

C Working with Special Dependent Variables in Econometrics

Like qualitative variables, the limited (censored or truncated) values cause the distributional assumptions of the classical linear regression model to fail. Fortunately, econometricians have developed techniques to handle restricted/limited dependent variables that are similar to those used for qualitative dependent variables.

The following list contains special dependent variable situation and the names of the techniques econometricians have developed to handle them:

- **Dichotomous or binary response dependent variable:** A discrete variable with two outcomes, usually 0 or 1. Handled with *Probit/Logit models.*
- **Censored dependent variable:** A continuous variable where some of the actual values have been limited to some predetermined minimum or maximum value. Handled with the *Tobit (censored normal) model.*
- **Truncated dependent variable:** A continuous variable where some of the actual values aren't observed if they are less than some predetermined minimum value or more than some predetermined maximum value. Handled with the *truncated normal model.*
- **Self-selected sample:** Missing values for the dependent variable due to nonrandom participation decisions from population of interest. Handled with the *Heckman selection model.*
- **Polychotomous or a multiple response dependent variable:** A discrete variable with more than two outcomes. Handled with a *multinomial Probit/Logit model* or *ordered Probit/Logit model* (covered in more advanced econometrics courses).
- **Discrete dependent variable:** A nonnegative, discrete count variable that assumes integer values (0, 1, 2, $\cdots$). Handled with a *Poisson model* or *negative binomial model* (covered in more advanced econometrics courses).

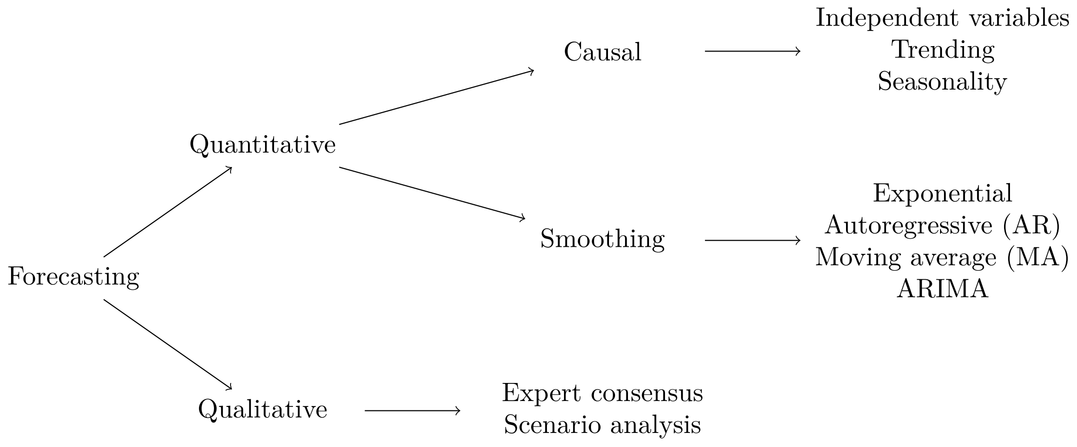
D Choose a Forecasting Method in Econometrics

E *Econometrics for Dummies* Cheat Sheet

E.1 The CLRM assumptions

Assumptions of the classical linear regression model (CLRM):

- The model parameters are linear.
- The values for the independent variables are derived from a random sample of the population, and they contain variability.
- The explanatory variables don't have perfect collinearity.



- The error term has zero conditional mean.
- The model has no heteroskedasticity.
- The model has no autocorrelation.

Under the above assumptions, the ordinary least squares (OLS) generates the optimal results (Gauss-Markov theorem).

E.2 Useful formulas in econometrics

Regression coefficients in a model with one independent variable:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2},\ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Standard error of the estimate or mean squared error:

$$\hat{\sigma}_\varepsilon = \sqrt{\frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n-p-1}}.$$

Standard error of regression coefficients in a model with one independent variable:

$$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}_\varepsilon}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})^2}},\ \hat{\sigma}_{\hat{\beta}_0} = \sqrt{\frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \overline{X})^2}} \cdot \hat{\sigma}_\varepsilon.$$

Explained sum of squares (ESS), residual sum of squares (RSS), and total sum of squares (TSS):

$$ESS = \sum_{i=1}^n (\hat{Y}_i - \overline{Y})^2,\ RSS = \sum_{i=1}^n (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^n \hat{\varepsilon}_i^2,\ TSS = \sum_{i=1}^n (Y_i - \overline{Y})^2 = ESS + RSS.$$

Coefficient of determination; R-squared:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}.$$

t-statistic for regression coefficients:

$$t = \frac{\hat{\beta}_k}{\hat{\sigma}_{\hat{\beta}_k}}.$$

Confidence interval for regression coefficients:

$$\hat{\beta}_k \pm t_{\alpha/2} \cdot \hat{\sigma}_{\hat{\beta}_k}$$

E.3 Common functional forms for regression

Quadratic functions:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \varepsilon_i.$$

Cubic functions:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + \beta_3 X_i^3 + \varepsilon_i.$$

Inverse functions:

$$Y_i = \beta_0 + \beta_1 \frac{1}{X_i} + \varepsilon_i.$$

Log-log functions:

$$\ln Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i.$$

Log-linear functions:

$$\ln Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i.$$

Linear-log functions:

$$Y_i = \beta_0 + \beta_1 \ln X_i + \varepsilon_i.$$

E.4 Typical problems estimating econometric models

High multicollinearity.

- Definition: two or more independent variables in a regression model exhibit a close linear relationship.
- Consequences: large standard errors and insignificant t-statistics, coefficient estimates sensitive to minor changes in model specification, and nonsensical coefficient signs and magnitudes.
- Detection: pairwise correlation coefficients and variance inflation factor (VIF).
- Solution: 1. collect additional data; 2. re-specify the model; 3. drop redundant variables.

Heteroskedasticity.

- Definition: the variance of the error term changes in response to a change in the value of the independent variables.
- Consequences: inefficient coefficient estimates, biased standard errors, and unreliable hypothesis tests.
- Detection: Park test, Goldfeld-Quandt test, Breusch-Pagan test, and White test.
- Solution: 1. weighted least squares (WLS); 2. robust standard errors.

Autocorrelation.

- Definition: an identifiable relationship exists between the values of the error in one period and the values of the error in another period.
- Consequences: inefficient coefficient estimates, biased standard errors, and unreliable hypothesis tests.
- Detection: Geary or runs test, Durbin-Watson test, and Breusch-Godfrey test.
- Solution: 1. Cochrane-Orcutt transformation; 2. Prais-Winstein transformation; 3. Newey-West robust standard errors.

References

[1] Chris Brooks. *Introductory Econometrics for Finance*, 2ed.. New York, Cambridge University Press, 2008. 9

[2] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning : with applications in R.* New York, Springer, 2014. 6

[3] Roberto Pedace. *Econometrics for dummies.* Hoboken, John Wiley & Sons Inc., 2013. 1, 7

[4] Roberto Pedace. *Econometrics for Dummies* extras. www.dummies.com/extras/econometrics. 1