Testing Joint Hypotheses

The F-test

FORDHAM THE JESUIT UNIVERSITY OF NEW YORK | Gabelli School of Business

Testing Joint Hypotheses: The F-Test, Part I

- A null hypothesis with multiple conjectures, expressed with more than one equal sign, is called a **joint hypothesis**
  1. Example: Should a group of explanatory variables be included in a particular model?
  2. Example: Does the quantity demanded of a product depend on the prices of substitute goods, or only on its own price?

Testing Joint Hypotheses: The F-Test, Part II

- Both examples are of the form:
  - (6.1) $H_0 : \beta_4 = 0, \beta_5 = 0, \beta_6 = 0$
  - The joint null hypothesis in equation 6.1 contains three conjectures (three equal signs): $\beta_4 = 0$, $\beta_5 = 0$, and $\beta_6 = 0$
  - A test of H_0 is a joint test for whether all three conjectures hold simultaneously

Testing Joint Hypotheses: The F-Test, Part III

- The *F*-statistic determines what constitutes a large reduction or a small reduction in the sum of squared errors (SSE):
  - (6.4) $$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)}$$
  - Where *J* is the number of restrictions, *N* is the number of observations, and *K* is the number of coefficients in the unrestricted model
- **If the null hypothesis is true**, then the statistic *F* has what is called an *F*-distribution with *J* numerator degrees of freedom and $N - K$ denominator degrees of freedom
- **If the null hypothesis is not true**, then the difference between SSE_R and SSE_U becomes large

Example: Testing the Effect of Advertising, Part I

- Consider the model:
  - (6.2) $SALES = \beta_1 + \beta_2 PRICE + \beta_3 ADVERT + \beta_4 ADVERT^2 + e$
- Test whether or not advertising has an effect on sales—but advertising is in the model as two variables
  1. Advertising will have no effect on sales if $\beta_3 = 0$ and $\beta_4 = 0$
  2. Advertising will have an effect if $\beta_3 \neq 0$ or $\beta_4 \neq 0$ or if both β_3 and β_4 are nonzero

Example: Testing the Effect of Advertising, Part II

- The null hypotheses are:

$$H_0: \beta_3 = 0, \beta_4 = 0$$

$$H_1: \beta_3 \neq 0 \text{ or } \beta_4 \neq 0 \text{ or both are nonzero}$$

- Relative to the null hypothesis H_0:$\beta_3 = 0$, $\beta_4 = 0$, the model in equation 6.2 is called the **unrestricted model**
- The restrictions in the null hypothesis have not been imposed on the model
- It contrasts with the restricted model, which is obtained by assuming the parameter restrictions in H_0 are true

Example: Testing the Effect of Advertising, Part III

- When H_0 is true, $\beta_3 = 0$ and $\beta_4 = 0$, and *ADVERT* and $ADVERT^2$ drop out of the model

  (6.3) $SALES = \beta_1 + \beta_2 PRICE + e$

- The *F*-test for the hypothesis $H_0{:}\beta_3 = 0, \beta_4 = 0$ is based on a comparison of the sums of squared errors (sums of squared least squares residuals) from the unrestricted model in equation 6.2 and the restricted model in equation 6.3
- Shorthand notation for these two quantities is SSE_U and SSE_R, respectively

Example: Testing the Effect of Advertising, Part IV

- Adding variables to a regression reduces the SSE
- More of the variation in the dependent variable becomes attributable to the variables in the regression and less of its variation becomes attributable to the error
- In terms of our notation, SSER − SSEU ≥ 0
- We find that SSE_U = 1532.084 and SSE_R = 1896.391; adding *ADVERT* and $ADVERT^2$ to the equation reduces the SSEs from 1896.391 to 1532.084

Example: The F-Test Procedure, Part I

1. Specify the null and alternative hypotheses
   - The joint null hypothesis is H_0:$\beta_3 = 0$, $\beta_4 = 0$; the alternative hypothesis is H_0:$\beta_3 \neq 0$ or $\beta_4 \neq 0$ both are nonzero
2. Specify the test statistic and its distribution if the null hypothesis is true
   - Having two restrictions in H_0 means $J = 2$
   - Also, recall that $N = 75$

$$F = \frac{\left(SSE_R - SSE_U\right)/2}{SSE_U/\left(75-4\right)}$$

Example: The F-Test Procedure, Part II

3. Set the significance level and determine the rejection region
4. Calculate the sample value of the test statistic and, if desired, the *p*-value

$$F = \frac{\left(SSE_R - SSE_U\right)/J}{SSE_U/\left(N-K\right)} = \frac{\left(1896.391 - 1532.084\right)/2}{1532.084/\left(75-4\right)} = 8.44$$

- The corresponding *p*-value is $p = P(F_{(2, 71)} > 8.44) = 0.0005$

Example: The F-Test Procedure, Part III

5. State your conclusion
   - Because $F = 8.44 > F_c = 3.126$, we reject the null hypothesis that both $\beta_3 = 0$ and $\beta_4 = 0$, and conclude that at least one of them is not zero
   - Advertising does have a significant effect upon sales revenue
   - The same conclusion is reached by noting that p-value = 0.0005 < 0.05

Testing the Significance of the Model, Part I

- Consider again the general multiple regression model with (K – 1) explanatory variables and K unknown coefficients
  - (6.5) $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_K x_K + e$
- To examine whether we have a viable explanatory model, we set up the following null and alternative hypotheses:
  - (6.6) $H_0 : \beta_2 = 0, \beta_3 = 0, \ldots, \beta_K = 0$

    $H_1 :$ *At least one* of the β_k is nonzero for $k = 2, 3, \ldots K$

Testing the Significance of the Model, Part II

- Because we are testing whether or not we have a viable explanatory model, the test for equation 6.6 is sometimes referred to as a **test of the overall significance of the regression model**
  - Given that the *t*-distribution can only be used to test a single null hypothesis, we use the *F*-test for testing the joint null hypothesis in equation 6.6
- The restricted model, assuming the null hypothesis is true, becomes $y_i = \beta_1 + e_i$
- The least squares estimator of β_1 in this restricted model is:

$$b_1^* = \sum_{i=1}^{N} y_i \Big/ N = \bar{y}$$

Testing the Significance of the Model, Part III

- The restricted SSEs from the hypothesis (equation 6.6) is:

$$SSE_R = \sum_{i=1}^{N}\left(y_i - b_1^*\right)^2 = \sum_{i=1}^{N}\left(y_i - \bar{y}\right)^2 = SST$$

- Thus, to test the overall significance of a model, *but not in general*, the *F*-test statistic can be modified and written as:
  - (6.8) $$F = \frac{(SST - SSE)/(K-1)}{SSE/(N-K)}$$

The Relationship Between t-Tests and F-Tests

- What happens if we have a null hypothesis that is not a joint hypothesis? It only has one equality in H_0
- When testing a single "equality" null hypothesis (a single restriction) against a "not equal to" alternative hypothesis, either a t-test or an *F*-test can be used
- Two-tail t-tests are equivalent to *F*-tests when there is a single hypothesis H_0

The Relationship Between t-Tests and F-Tests (cont.)

- The elements of an *F*-test
  1. The null hypothesis H_0 consists of one or more equality restrictions on the model parameters β_k
  2. The alternative hypothesis states that one or more of the equalities in the null hypothesis is not true
  3. The test statistic is the *F*-statistic in equation 6.4
  4. If the null hypothesis is true, *F* has the *F*-distribution with *J* numerator degrees of freedom and $N - K$ denominator degrees of freedom
  5. When testing a single equality null hypothesis, it is perfectly correct to use either the *t*-test or *F*-test procedure: they are equivalent

More General F-Tests

- The conjectures made in the null hypothesis were that particular coefficients are equal to zero.
- The *F*-test can also be used for much more general hypotheses.
- Any number of conjectures ($\leq K$) involving linear hypotheses with equal signs can be tested.

Example: Testing Optimal Advertising, Part I

- Consider the issue of testing:
  - (6.11) $\beta_3 + 2\beta_4 ADVERT_0 = 1$
- If $ADVERT_0$ = \$1,900 per month, then:

$$H_0: \beta_3 + 2 \times \beta_4 \times 1.9 = 1 \qquad H_1: \beta_3 + 2 \times \beta_4 \times 1.9 \neq 1$$

Or

$$H_0: \beta_3 + 3.8\beta_4 = 1 \qquad H_1: \beta_3 + 3.8\beta_4 \neq 1$$

- Note that, when H_0 is true, $\beta_3 = 1 - 3.8\beta_4$ so that:

$$SALES = \beta_1 + \beta_2 PRICE + \left(1 - 3.8\beta_4\right) ADVERT + \beta_4 ADVERT^2 + e$$

Example: Testing Optimal Advertising, Part II

- Or (6.12) $\left(SALES - ADVERT\right) = \beta_1 + \beta_2 PRICE + \beta_4 \left(ADVERT^2 - 3.8ADVERT\right) + e$
- The calculated value of the *F*-statistic is:

$$F = \frac{\left(1552.286 \ - \ 1532.084\right)/1}{1532.084/71} = 0.9362$$

- For α = 0.05, the critical value is $F_c = 3.976$. Because $F = 0.9362 < F_c = 3.976$, we do not reject H_0
- We conclude that an advertising expenditure of $1,900 per month is optimal and compatible with the data

Example: Testing Optimal Advertising, Part III

- The *t*-value is $t = 0.9676$
  - $F = 0.9362$ is equal to $t^2 = (0.9676)^2$
  - The *p*-values are identical:

$$
\begin{aligned}
p-value &= P\left(F_{(1,71)} > 0.9362\right) \\
&= P\left(t_{(71)} > 0.9676\right) + P\left(t_{(71)} < -0.9676\right) \\
&= 0.3365
\end{aligned}
$$

Using Computer Software

- Though it is possible and instructive to compute an *F*-value by using the restricted and unrestricted sums of squares, it is often more convenient to use the power of econometric software.
- Most software packages have commands that will automatically compute *t*-values and *F*-values and their corresponding *p*-values when provided with a null hypothesis.
- These tests belong to a class of tests called **Wald tests**.

Large Sample Tests, Part I

- There are two key requirements for the F-statistic to have the F-distribution in samples of all sizes
  1. Assumptions MR1–MR6 must hold
  2. The restrictions in H_0 must be linear functions of the parameters β_1, β_2, …, β_k a
- In this section, we are concerned with what test statistics are valid in large samples when the errors are no longer normally distributed or when the strict exogeneity assumption is weakened to $E(e_i) = 0$ and $cov(e_i, x_{jk}) = 0 (i \neq j)$

Large Sample Tests, Part II

- An F random variable is defined as the ratio of two independent chi-square (χ^2) random variables, each divided by their degrees of freedom

$$(6.13) \quad F = \frac{V_1/J}{V_2/(N-K)} = \frac{\dfrac{SSE_R - SSE_U}{\sigma^2} \Big/ J}{\dfrac{(N-K)\hat{\sigma}^2}{\sigma^2} \Big/ (N-K)} = \frac{(SSE_R - SSE_U)/J}{\hat{\sigma}^2} \sim F_{(J,N-K)}$$

- Note that $\sigma^2 = SSE_U/(\mathrm{N}-\mathrm{K})$, and so the result in equation 6.13 is identical to the F-statistic first introduced in equation 6.4
- When we drop the normality assumption or weaken the strict exogeneity assumption, the argument becomes slightly different

Large Sample Tests, Part III

- We can go one step further and say that replacing σ^2 by its consistent estimator $\widehat{\sigma^2}$ does not change the asymptotic distribution of $V_{1;}$ that is:

$$(6.14) \quad \hat{V}_1 = \frac{SSE_R - SSE_U}{\hat{\sigma}^2} \,\tilde{a}\, \chi^2_{(J)}$$

- This statistic is a valid alternative for testing joint linear hypotheses in large samples under less restrictive assumptions
- Test statistics for joint hypotheses that are nonlinear functions of the parameters s can typically be carried out by your software with relative ease

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

The Use of Nonsample Information

The Use of Nonsample Information, Part I

- In many estimation problems, we have information over and above the information contained in the sample observations
- This nonsample information may come from many places, such as economic principles or experience
- When it is available, it seems intuitive that we should find a way to use it

The Use of Nonsample Information, Part II

- Consider the log-log functional form for a demand model for beer:
  - (6.17) $\ln(Q) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) + e$
- This model is a convenient one because it precludes infeasible negative prices, quantities, and income, and because the coefficients β_2, β_3, β_4, and β_5 are elasticities
- A relevant piece of nonsample information can be derived by noting that, if all prices and income go up by the same proportion, we would expect there to be no change in quantity demanded

The Use of Nonsample Information, Part III

- Having all prices and income change by the same proportion is equivalent to multiplying each price and income by a constant, say λ:
  - (6.18) $\ln(Q) = \beta_1 + \beta_2 \ln(\lambda PB) + \beta_3 \ln(\lambda PL) + \beta_4 \ln(\lambda PR) + \beta_5 \ln(\lambda I) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) + (\beta_1 + \beta_2 + \beta_3 + \beta_4 + \beta_5) \ln(\lambda) + e$

The Use of Nonsample Information, Part IV

- To have no change in ln(Q) when all prices and income go up by the same proportion, it must be true that:

$$\beta_2 + \beta_3 + \beta_4 + \beta_5 = 0$$

- Start with $\ln(Q) = \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + \beta_4 \ln(PR) + \beta_5 \ln(I) + e$
- Solve the restriction for one of the parameters, say β_4: $\beta_4 = -\beta_2 - \beta_3 - \beta_5$
- Substituting gives:

$$\begin{aligned}
\ln(Q) &= \beta_1 + \beta_2 \ln(PB) + \beta_3 \ln(PL) + (-\beta_2 - \beta_3 - \beta_5)\ln(PR) + \beta_5 \ln(I) + e \\
&= \beta_1 + \beta_2 \left[\ln(PB) - \ln(PR)\right] + \beta_3 \left[\ln(PL) - \ln(PR)\right] \\
&\quad + \beta_5 \left[\ln(I) - \ln(PR)\right] + e \\
&= \beta_1 + \beta_2 \ln\left(\frac{PB}{PR}\right) + \beta_3 \ln\left(\frac{PL}{PR}\right) + \beta_5 \ln\left(\frac{I}{PR}\right) + e
\end{aligned}$$

The Use of Nonsample Information, Part V

- Properties of this restricted least squares estimation procedure
    1. The restricted least squares estimator is biased, unless the constraints we impose are exactly true
    2. The restricted least squares estimator is that its variance is smaller than the variance of the least squares estimator, whether the constraints imposed are true or not

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Model Specification

Model Specification

- In any econometric investigation, choice of the model is one of the first steps
  - What are the important considerations when choosing a model?
  - What are the consequences of choosing the wrong model?
  - Are there ways of assessing whether a model is adequate?
- Three essential features of model choice are:
  1. Choice of functional form
  2. Choice of explanatory variables (regressors) to be included in the model
  3. Whether the multiple regression assumptions MR1–MR6, listed in Chapter 5, hold

Causality vs. Prediction

- With causal inference, we are primarily interested in the effect of a change in a regressor on the conditional mean of the dependent variable.
- We wish to be able to say that a one-unit change in an explanatory variable will cause a particular change in the mean of the dependent variable with other factors held constant.
- This type of analysis is important for policy work.

Causality vs. Prediction (cont.)

- On the other hand, if the purpose of a model is to predict the value of a dependent variable, then, for regressor choice, it is important to choose variables that are highly correlated with the dependent variable and that lead to a high R^2.
- Predictive analysis using variables from the increasingly popular field of big data is an example of where variables are chosen for their predictive ability, rather than to examine causal relationships.

Omitted Variables

- It is possible that a chosen model may have important variables omitted.
- Our economic principles may have overlooked a variable, or lack of data may lead us to drop a variable even when it is prescribed by economic theory.
- There are four observations for the omitted variable bias proof:
    1. Omitting a relevant variable is a special case of using a restricted least squares estimator where the restriction $\beta 3 = 0$ is not true. It leads to a biased estimator for $\beta 2$, but one with a lower variance.

Omitted Variables (cont.)

2. Knowing the sign of β3 and the sign of the covariance between x and z tells us the direction of the bias.
3. The $bias(b_2^*|\boldsymbol{X}) = E(b_2^*|\boldsymbol{X}) - \beta_2 = \beta_3 \frac{\widehat{cov}(x,z)}{\widehat{var}(x)}$ can also be written as $\beta_3 \widehat{\gamma^2}$, where $\widehat{\gamma^2}$ is the least squares estimate of γ^2 from the regression equation E(z|x) = γ_1+ γ_1x.
4. The importance of the assumption E(e_i|x, **z**) = 0 becomes clear.

In the equation y_i = β_1 + $\beta_2 x_i$ + v_i, we have E(v_i| x_i) = $\beta_3 E(z_i|x_i)$. It is the nonzero value for E(z_i |x_i) that leads to the biased estimator for β_2.

Irrelevant Variables

- You think that a good strategy is to include as many variables as possible in your model
- Doing so will not only complicate your model unnecessarily, but may also inflate the variances of your estimates because of the presence of **irrelevant variables**
  - Those whose coefficients are zero because they have no direct effect on the dependent variable

Control Variables

- Variables included in the equation to avoid omitted variable bias in the coefficient of interest are called **control variables**.
- For a control variable to serve its purpose and act as a proxy for an omitted variable, it needs to satisfy a **conditional mean independence assumption**.
- Labor economists are interested in the question:
  - What is the causal relationship between more education and higher wages?
- One variable that is clearly relevant, but difficult to include because it cannot be observed, is ability.

Control Variables (cont.)

- More able people are likely to have more education, and so ability and education will be correlated.
- If we look at the equation $E(ABILITY|EDUCATION, IQ) = E(ABILITY|IQ)$.
- IQ is correlated with both education and ability.
- Once we know somebody's IQ, knowing their level of education does not add any extra information about their ability.
- We can proceed to use IQ as a control variable or a proxy variable to replace ABILITY.

Choosing a Model, Part I

1. Is the purpose of the model to identify causal effects or prediction? Careful selection of control variables is necessary if the goal is to predict, not find, causality then using variables that have high predictive power is the major concern.
2. Having theoretical knowledge and understanding of the relationship are important for choosing the variables and functional form.
3. If an estimated equation has coefficients with unexpected signs, or unrealistic magnitudes, they could be caused by a misspecification.

Choosing a Model, Part II

4. Patterns in least squares residuals can be helpful for uncovering problems caused by an incorrect functional form.
5. One method for assessing whether a variable or a group of variables should be included in an equation is to perform significance tests.
6. Have the leverage, studentized residuals, and DFBETAS and DFFITS measures identified any influential observations?
7. Are the estimated coefficients robust with respect to alternative specifications?

Choosing a Model, Part III

8. A test known as RESET (Regression Specification Error Test) can be useful for detecting omitted variables or an incorrect functional form.
9. Various model selection criteria, based on maximizing R^2, or minimizing the sum of squared errors, subject to a penalty for too many variables, have been suggested.
10. A more stringent assessment of a model's predictive ability is to use a "hold-out" sample.
11. Following the guidelines in the previous 10 points can almost inevitably lead to revisions of originally proposed model.

RESET

- RESET (**RE**gression **S**pecification **E**rror **T**est) is designed to detect omitted variables and incorrect functional form
- Suppose we have the model. $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + e$
- Let the predicted values of *y* be: (6.32) $\hat{y} = b_1 + b_2 x_2 + \mathrm{b}_3 x_3$
- Now consider the following two artificial models:
    1. (6.33) $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + e$
    2. (6.34) $y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \gamma_1 \hat{y}^2 + \gamma_1 \hat{y}^3 + \mathrm{e}$

RESET (cont.)

- In equation 6.34, a test for misspecification is a test of $H_0{:}\gamma_1 = 0$ against the alternative $H_1{:}\gamma_1 \neq 0$
- In equation 6.36, testing $H_0{:}\gamma_1 = \gamma_2 = 0$ against $H_1{:}\gamma_1 \neq 0$ and/or $\gamma_2 \neq 0$ is a test for misspecification
- H0 implies that the original model is inadequate and can be improved; a failure to reject H0 says that the test has not been able to detect any misspecification

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Prediction

FORDHAM
THE JESUIT UNIVERSITY OF NEW YORK
Gabelli School of Business

Prediction, Part I

- In this section, we describe that extension, reinforce earlier material, and provide some more general background
- Consider the equation 6.35:

$$E(y_0|\boldsymbol{x}_0) = \beta_1 + \beta_2 x_{02} + \beta_3 x_{03} + \cdots + \beta_K x_{0K}$$

- Defining $e_0 = y_0 - E(y_0|\boldsymbol{x}_0)$, we can write:
  - (6.36) $y_0 = \beta_1 + \beta_2 x_{02} + \beta_3 x_{03} + \cdots + \beta_K x_{0K} + e_0$
- Define $e_i = y_i - E(y_i|\boldsymbol{x}_i)$, so the model used to estimate is:
  - (6.38) $y_\mathrm{i} = \beta_1 + \beta_2 x_{\mathrm{i}2} + \beta_3 x_{\mathrm{i}3} + \cdots + \beta_K x_{\mathrm{i}K} + e_\mathrm{i}$

Prediction, Part II

- These equations make up the **predictive model**
- Note that equations 6.36 and 6.38 do not have to be causal models
- The error term *e* is simply the difference between the realized value y and its conditional expectation; it is the forecasting error that would occur if $(\beta_1, \beta_2, \ldots, \beta_K)$ were known and did not have to be estimated
- An extra assumption that we need is that $(e_0|\boldsymbol{x}_0)$ is uncorrelated with $(e_i|\boldsymbol{x}_i)$ for i = 1, 2,…, N and i ≠ 0 and $\text{var}(e_0|\boldsymbol{x}_0) = \text{var}(e_i|\boldsymbol{x}_i) = \sigma^2$

Prediction, Part III

- The forecast error is given by:

$$f = y_0 - \widehat{y_0} = (\beta_1 \text{-}\ b_1) + (\beta_2 \text{-}\ b_2)x_{02} + (\beta_3 \text{-}\ b_3)x_{03} + \cdots + (\beta_K -$$

$$\begin{aligned}\operatorname{var}(f \mid \mathbf{x}_0, \mathbf{X}) &= \operatorname{var}\left[\left(\sum_{k=1}^{K}(\beta_k - b_k)\,x_{0k}\right)\middle|\,\mathbf{x}_0, \mathbf{X}\right] + \operatorname{var}(e_0 \mid \mathbf{x}_0, \mathbf{X}) \\ &= \operatorname{var}\left[\left(\sum_{k=1}^{K} b_k x_{0k}\right)\middle|\,\mathbf{x}_0, \mathbf{X}\right] + \sigma^2 \\ &= \sum_{k=1}^{K} x_{0k}^2 \operatorname{var}(b_k \mid \mathbf{x}_0, \mathbf{X}) + 2\sum_{k=1}^{K}\sum_{j=k+1}^{K} x_{0k}x_{0j}\operatorname{cov}(b_k, b_j \mid \mathbf{x}_0, \mathbf{X}) + \sigma^2\end{aligned}$$

Predictive Model Selection Criteria

- In this section, we consider three model selection criteria:
  1. R^2 and $\bar{R}^2$
  2. AIC
  3. SC
- These criteria should be treated as devices that provide additional information about the relative merits of alternative models

Predictive Model Selection Criteria (cont.)

- An alternative measure of goodness of fit called the adjusted-R^2, denoted $\bar{R}^2$
  - Computed as $\bar{R}^2 = 1 - \frac{SSE/(N-K)}{SST/(N-1)}$
- Selecting variables to maximize $\bar{R}^2$ can be viewed as selecting variables to minimize SSE

$AIC = \ln\left(\frac{SSE}{N}\right) + (\frac{2K}{N})$ and SC = $\ln\left(\frac{SSE}{N}\right) + (\frac{K ln(N)}{N})$

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Poor Data, Collinearity, and Insignificance

Poor Data, Collinearity, and Insignificance

- When data are the result of an uncontrolled experiment, many of the economic variables may move together in systematic ways.
  - Such variables are said to be **collinear**, and the problem is labeled **collinearity**.
- It is not just relationships between variables in a sample of data that make it difficult to isolate the separate effects of individual explanatory variables.
- If the values of an explanatory variable do not vary much within a sample of data, then it is difficult to use that data to estimate a coefficient that describes the effect of change in that variable.

The Consequences of Collinearity, Part I

- Exact or extreme collinearity exists when x_2 and x_3 are perfectly correlated, in which case $r_{23} = 1$ and var(b_2) goes to infinity.
- Similarly, if x_2 exhibits no variation, $\sum(x_i - \bar{x}_2)^2$ equals zero and var(b_2) again goes to infinity.
- In this case, x_2 is collinear with the constant term.

The Consequences of Collinearity, Part II

- In general, whenever there are one or more exact linear relationships among the explanatory variables, then the condition of exact collinearity exists.
  - In this case, the least squares estimator is not defined.
  - We cannot obtain estimates of β_k using the least squares principle.

The Consequences of Collinearity, Part III

- The effects of this imprecise information are:
  1. When estimator standard errors are large, it is likely that the usual *t*-tests will lead to the conclusion that parameter estimates are not significantly different from zero.
  2. Estimators may be very sensitive to the addition or deletion of a few observations, or to the deletion of an apparently insignificant variable.
  3. Accurate forecasts may still be possible if the nature of the collinear relationship remains the same within the out-of-sample observations.

Identifying and Mitigating Collinearity

- One simple way to detect collinear relationships is to use sample correlation coefficients between pairs of explanatory variables.
  - These sample correlations are descriptive measures of linear association.
  - However, in some cases in which collinear relationships involve more than two of the explanatory variables, the collinearity may not be detected using auxiliary regression.

Identifying and Mitigating Collinearity (cont.)

- If R^2 from this artificial model is above 0.80, say, the implication is that a large portion of the variation in x_2 is explained by variation in the other explanatory variables.
- The collinearity problem is that the data do not contain enough "information" about the individual effects of explanatory variables to permit us to estimate all the parameters of the statistical model precisely.
- A second way of adding new information is to introduce nonsample information in the form of restrictions on the parameters.

TABLE 6.8 Statistics for Identifying Influential Observations

Influence Statistic	Formula	Investigative Threshold
Leverage	$h_i = \dfrac{\widehat{\text{var}}(\hat{e}_i) - \hat{\sigma}^2}{\hat{\sigma}^2}$	$h_i > \dfrac{2K}{N}$ or $\dfrac{3K}{N}$
Studentized residual	$\hat{e}_i^{stu} = \dfrac{\hat{e}_i}{\hat{\sigma}(i)(1 - h_i)^{1/2}}$	$\lvert \hat{e}_i^{stu} \rvert > 2$
DFBETAS	$\text{DFBETAS}_{ki} = \dfrac{b_k - b_k(i)}{(\hat{\sigma}(i)/\hat{\sigma}) \times \text{se}(b_k)}$	$\lvert \text{DFBETAS}_{ki} \rvert > \dfrac{2}{\sqrt{N}}$
DFFITS	$\text{DFFITS}_i = \left(\dfrac{h_i}{1 - h_i}\right)^{1/2} \hat{e}_i^{stu}$	$\lvert \text{DFFITS}_i \rvert > 2\left(\dfrac{K}{N}\right)^{1/2}$

Example: Influential Observations in the House Price Equations

- The preferred equation for predicting house prices was:

  $$\ln(PRICE) = \beta_1 + \beta_2 SQFT + \beta_3 AGE + \beta_4 AGE^2 + e$$

- In a sample of 900 observations, it is not surprising to find a relatively large number of data points where the various influence measures exceed the recommended thresholds
- The observations with the three largest DFFITS also have the large values for the influence measures

Example: Influential Observations in the House Price Equations (cont.)

- In parentheses next to each of the values is the rank of its absolute value; when we check the characteristics of the three unusual observations, we find that observation 540 is the newest house in the sample and observation 150 is the oldest house; observation 411 is both old and large

TABLE 6.9 Influence Measures for House Price Equation

Observation	h_i (rank)	$\hat{e}_i^{stu}$ (rank)	DFFITS_i (rank)	DFBETAS_{ki} (rank)		
Threshold	$\frac{2.5K}{N} = 0.011$	2	$2\left(\frac{K}{N}\right)^{1/2} = 0.133$	$\frac{2}{\sqrt{N}} = 0.067$		
				SQFT	*AGE*	AGE^2
411	0.0319 (10)	−4.98 (1)	0.904 (1)	−0.658 (1)	0.106 (17)	−0.327 (3)
524	0.0166 (22)	−4.31 (3)	0.560 (2)	0.174 (9)	0.230 (2)	−0.381 (2)
150	0.0637 (2)	1.96 (48)	−0.511 (3)	−0.085 (29)	−0.332 (1)	0.448 (1)

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Nonlinear Least Squares

FORDHAM | Gabelli School of Business
THE JESUIT UNIVERSITY OF NEW YORK

Nonlinear Least Squares

- In this section, we discuss estimation of models that are nonlinear in the parameters
- Consider the model in equation 6.50: $y_i = \beta x_{i1} + \beta^2 x_{i2} + e_i$
- This example differs from the conventional linear model because the coefficient of x_i2 is equal to the square of the coefficient of x_i1, and the number of parameters is not equal to the number of variables

Nonlinear Least Squares (cont.)

- To estimate β, we can use equation 6.51:

$$S(\beta) = \sum_{i=1}^{N} (y_i - \beta x_{i1} - \beta^2 x_{i2})^2$$

- When we have models that are nonlinear in the parameters, we cannot in general derive formulas for the parameter values that minimize the sum of squared errors function
- However, for a given set of data, we can ask the computer to search for the parameter values that take us to the bottom of the bowl
- Those minimizing values are known as the **nonlinear least squares estimates**

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business

Key Words

- χ^2-test
- AIC
- Auxiliary regressions
- BIC
- Causal model
- Collinearity
- Control variables
- F-test
- Influential observations
- Irrelevant variables
- Nonlinear least squares
- Nonsample information
- Omitted variable bias
- Overall significance
- Prediction
- Predictive model
- RESET
- Restricted least squares
- Restricted model
- Restricted SSE
- SC
- Single and joint null hypotheses
- Unrestricted model
- Unrestricted SSE

Copyright

FORDHAM

THE JESUIT UNIVERSITY OF NEW YORK

Gabelli School of Business