ACE 564
Spring 2006

Lecture 5

The Multiple Regression Model: Hypothesis Testing for a Single Parameter, Goodness of Fit and Reporting the Results

by
Professor Scott H. Irwin

Readings:

Griffiths, Hill and Judge. "One-Sided Hypothesis Testing for a Single Coefficient," Section 10.3; "Testing a Zero Null Hypothesis for a Single Coefficient," Section 10.4; "Inference and Reporting the Results," Section 10.5 in *Learning and Practicing Econometrics*

Kennedy. "Interval Estimation and Hypothesis Testing," Chapter 4 in *A Guide to Econometrics*

The Problem

We will continue to work with the data for the Bay Area Rapid Food hamburger chain

To review, each week they must decide:

- How much to spend on advertising
- Whether to lower prices as special promotions

Need to know the relationship between total chain revenue, advertising and prices

The Statistical Model

$$y_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t} + e_t$$

where

y_t is total chain revenue for week t
$x_{2,t}$ is average price of chain products in week t
$x_{3,t}$ is advertising expenditures for week t

Hypothesis Tests about Individual Regression Parameters

In mechanical terms, hypothesis tests regarding individual regression parameters are performed the same way for simple regressions and multiple variable regressions

Hypothesis tests continue to have four basic elements,

1. A null hypothesis, H_0
2. An alternative hypothesis, H_1
3. A test statistic
4. A rejection region

Null Hypothesis

- Belief we maintain until convinced by the sample evidence that it is not true
- Denoted H_0 ("h-naught")

Alternative Hypothesis

- Logical alternative to H_0 that is accepted if the null hypothesis is rejected
- Denoted H_1 or H_a

Test Statistic

Sample information about the validity of the null hypothesis is embodied in the sample value of a test statistic

- Test statistic is a random variable
- Based on sample value of test statistic, will reject or not reject null hypothesis
- The sampling distribution of the test statistic is "known" under the null hypothesis
- The sampling distribution of the test statistic under the alternative hypothesis is "something else"

Rejection Region

A rejection region for the test statistic is,

- A set of test statistic values that have a low probability of occurring when the null hypothesis is true
- If a sample value of the test statistic falls in the rejection region we reject the null hypothesis
- If a sample value of the test statistic does not fall in the rejection region we do not reject the null hypothesis

The null hypothesis will be rejected only if an unusually "small" or "large" value of the test statistic is observed

"Small" and "large" are determined by α

- Tail probabilities that correspond to "small" or "large" values of the test statistic
- Level of significance of the test
- α typically set to 0.01, 0.05, or 0.10

Type I and II Errors

Whenever we reject or don't reject a null hypothesis, there is the chance of making a mistake

For comparison, there are two ways of making a correct decision,

- The null hypothesis is false and we decide to reject it
- The null hypothesis is true and we decide not to reject it

There are two ways of making an incorrect decision,

- The null hypothesis is true and we decide to reject it (Type I error)
- The null hypothesis is false and we decide not to reject it (Type II error)

Whenever we reject a null hypothesis we risk a Type I error

- The probability of a Type I error is α, the level of significance of the test
- We will reject a true null hypothesis with a probability equal to α
- We control the probability of a Type I error by choosing the level of significance
- If a Type I error is costly, we will want to make α small (e.g. 0.05, 0.01)

We risk a Type II error whenever we do not reject a null hypothesis

- Probability of a Type II error is not a single value, as it depends on the true but unknown value of the parameter in question
- Probability of a Type II error is not under our control

In one important respect, hypothesis tests regarding individual regression parameters are different for multiple variable regressions compared to simple regressions

- In multiple regression, a hypothesis about one parameter is stated assuming all other variables are held constant

- Even if the same hypothesis is tested (separately) for each parameter, test result still interpreted in terms of holding all other variables constant

Two-Tailed Hypothesis Testing for a Single Parameter

Let's begin by again considering the estimates for the Bay Area Rapid Food data,

$$\hat{y}_t = 104.79 - 6.642 x_{2,t} + 2.984 x_{3,t} \qquad R^2 = 0.867$$
$$(16.17) \quad (-2.081) \quad (17.868) \qquad (t-stat.)$$
$$[0.000] \quad [0.043] \quad [0.000] \qquad [p-value]$$

where

y_t is total chain revenue for week t
$x_{2,t}$ is average price of chain products in week t
$x_{3,t}$ is advertising expenditures for week t

When we set up the multiple regression model for this problem, we believed that both explanatory variables influenced the dependent variable

To confirm this, we need to test whether our belief is supported by the data

To find whether the data contain any evidence suggesting y is related to x_k we test the null hypothesis

$$H_0 : \beta_k = 0$$
$$H_1 : \beta_k \neq 0$$

To carry out the test we use the following t-statistic

$$t_k = \frac{b_k}{se(b_k)}$$

- Follows a t-distribution with $T\text{-}3$ df
- Use a two-tailed test because the alternative hypothesis is "not equal to"
- Reject H_0 if the computed t-value is greater than or equal to t_c, or less than or equal to $-t_c$

Sample Regression Output from Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.93117**
R Square	**0.86708**
Adjusted R Square	**0.86166**
Standard Error	**6.06961**
Observations	**52**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**2**	**11776.1839**	**5888.0919**	**159.8280**	**0.0000**
Residual	**49**	**1805.1684**	**36.8402**		
Total	**51**	**13581.3523**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**104.7855**	**6.4827**	**16.1638**	**0.0000**	**91.7580**	**117.8130**
X Variable 1	**-6.6419**	**3.1912**	**-2.0813**	**0.0427**	**-13.0549**	**-0.2290**
X Variable 2	**2.9843**	**0.1669**	**17.8769**	**0.0000**	**2.6488**	**3.3198**

Is revenue is related to price?

1. Hypotheses

$$H_0 : \beta_2 = 0$$
$$H_1 : \beta_2 \neq 0$$

2. Test statistic

$$t_2 = \frac{b_2}{\hat{se}(b_2)} = \frac{-6.642}{3.191} = -2.081$$

3. Rejection Region

- With 49 degrees of freedom and a 5% significance level, the critical values that lead to a probability of 0.025 in each tail of the distribution are $t_c = 2.01$ and $-t_c = -2.01$
- Thus we reject the null hypothesis if $t_2 \geq 2.01$ or if $t_2 \leq -2.01$
- In shorthand notation, we reject the null hypothesis if $|t_2| \geq 2.01$

4. Decision

- Since $(t_2 = -2.08) < (-t_c = -2.01)$ we reject the null hypothesis and conclude that the alternative hypothesis is more consistent with the sample data
- The p-value in this case is given by $P[|t_{(49)}| > 2.08] = 2 \times 0.021 = 0.042$
- Using this procedure we reject H_0 because $0.042 < 0.05$.
- Conclude that sample evidence supports the proposition that total revenue is related to price

Is revenue related to advertising?

1. Hypotheses

$$H_0 : \beta_3 = 0$$
$$H_1 : \beta_3 \neq 0$$

2. Test statistic

$$t_3 = \frac{b_3}{\hat{se}(b_3)} = \frac{2.984}{0.1669} = 17.88$$

3. Rejection Region

- With 49 degrees of freedom and a 5% significance level, the critical values that lead to a probability of 0.025 in each tail of the distribution are $t_c = 2.01$ and $-t_c = -2.01$
- Thus we reject the null hypothesis if $t_3 \geq 2.01$ or if $t_3 \leq -2.01$
- In shorthand notation, we reject the null hypothesis if $|t_3| \geq 2.01$

4. Decision

- Since $(t_3 = 17.88) > (t_c = 2.01)$ we reject the null hypothesis and conclude that the alternative hypothesis is more consistent with the sample data
- The *p*-value in this case is given by $P[|t_{(49)}| > 17.88] = 2 \times 0.000] = 0.000$
- Using this procedure we reject H_0 because $0.000 < 0.05$.
- Conclude that sample evidence supports the proposition that total revenue is related to advertising

One-Tailed Hypothesis Testing for a Single Parameter

Testing for Elastic Demand

With respect to demand elasticity we wish to know if

- $\beta_2 \geq 0$: a decrease in price leads to a decrease in total revenue (demand is price inelastic)
- $\beta_2 < 0$: a decrease in price leads to an increase in total revenue (demand is price elastic)

1. Hypotheses

$$H_0 : \beta_2 \geq 0 \quad \textit{demand is inelastic}$$
$$H_1 : \beta_2 < 0 \quad \textit{demand is elastic}$$

2. Test statistic

$$t_2 = \frac{b_2}{\hat{se}(b_2)} = \frac{-6.642}{3.191} = -2.081$$

Note that to create the test statistic we act as if the null hypothesis were the equality $\beta_2 = 0$

3. Rejection Region

- With 49 degrees of freedom and a 5% significance level, the critical value that lead to a probability of 0.05 in the left tail of the distribution are $-t_c = -1.68$
- Thus we reject the null hypothesis if $t_3 \leq -1.68$

4. Decision

- Since $(t_2 = -2.08) < (t_c = -1.68)$ we reject the null hypothesis and conclude that the alternative hypothesis is more consistent with the sample data
- The *p*-value in this case is given by $P[t_{(49)} < -2.08] = 0.021$
- Using this procedure we reject H_0 because $0.021 < 0.05$.
- Sample evidence supports the proposition that a reduction in price will increase total revenue
- Conclude demand is elastic

Testing Advertising Effectiveness

The other hypothesis of interest is whether an increase in advertising expenditure will bring an increase in total revenue that is sufficient to cover the increased cost of advertising

- Will occur if $\beta_3 > 1$
- An increase of one dollar in advertising increases total revenue more than a dollar

1. Hypotheses

$$H_0 : \beta_3 \leq 1$$
$$H_1 : \beta_3 > 1$$

2. Test statistic

$$t_3 = \frac{b_3 - 1}{\hat{se}(b_3)} = \frac{2.984 - 1}{0.1669} = 11.89$$

Note that to create the test statistic we act as if the null hypothesis were the equality $\beta_3 = 1$

3. Rejection Region

- With 49 degrees of freedom and a 5% significance level, the critical value that lead to a probability of 0.05 in the right tail of the distribution are $t_c = 1.68$

- Thus we reject the null hypothesis if $t_3 \geq 1.68$

4. Decision

- Since $(t_3 = 11.89) > (t_c = 1.68)$ we reject the null hypothesis and conclude that the alternative hypothesis is more consistent with the sample data

- The *p*-value in this case is essentially zero and we reject H_0 because 0.000 < 0.05.

- Sample evidence supports the proposition that an increase in advertising expense will be justified by the increase in total revenue

Goodness of Fit

Just as in the case of simple regression, we need to develop a measure of the proportion of the variation in y_t that is explained by the independent variables

$\Rightarrow$ Note that we are now interested in the variation in y_t that is jointly explained by the independent variables

The formal derivation of the measure of "fit" for the multiple regression model follows the same procedure as in the case of simple regression

We can decompose the value of y_t into two components,

$$y_t = \hat{y}_t + \hat{e}_t$$

where $\hat{y}_t = \beta_1 + \beta_2 x_{2,t} + \beta_3 x_{3,t}$ and $\hat{e}_t = y_t - \hat{y}_t$

Subtract the mean of y from both sides of the equation,

$$(y_t - \bar{y}) = (\hat{y}_t - \bar{y}) + \hat{e}_t$$

Since we are interested in "variation" and not "deviation", let's square both sides of the above equation,

$$(y_t - \bar{y})^2 = [(\hat{y}_t - \bar{y}) + \hat{e}_t]^2$$

Which can be expanded as follows,

$$(y_t - \bar{y})^2 = (\hat{y}_t - \bar{y})^2 + \hat{e}_t^2 + 2(\hat{y}_t - \bar{y})\hat{e}_t$$

Finally, sum both sides of the previous equation,

$$\sum_{t=1}^{T}(y_t - \bar{y})^2 = \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{T}\hat{e}_t^2 + 2\sum_{t=1}^{T}(\hat{y}_t - \bar{y})\hat{e}_t$$

and reduce to,

$$\sum_{t=1}^{T}(y_t - \bar{y})^2 = \sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 + \sum_{t=1}^{T}\hat{e}_t^2$$

This is an important relationship, which shows the decomposition of total sample variation in y_t into explained and unexplained components

Now, define the following terms,

$$\sum_{t=1}^{T}(y_t - \bar{y})^2 = \text{Sum of Squares Total } (SST)$$

$$\sum_{t=1}^{T}(\hat{y}_t - \bar{y})^2 = \text{Sum of Squares Regression } (SSR)$$

$$\sum_{t=1}^{T}\hat{e}_t^2 = \text{Sum of Squares Error } (SSE)$$

Hence,

$$SST = SSR + SSE$$

Divide the previous equation by SST to obtain the relationship in proportionate form,

$$\frac{SST}{SST} = \frac{SSR}{SST} + \frac{SSE}{SST}$$

or,

$$1 = \frac{SSR}{SST} + \frac{SSE}{SST}$$

Now, we can define,

$$R^2 = \frac{SSR}{SST}$$

Shows that R^2 measures the total sample variation in y_t jointly explained by the variation in $x_{2,t}$ and $x_{3,t}$

Further Points

By substituting into the SST equation in proportionate form, we obtain,

$$1 = R^2 + \frac{SSE}{SST}$$

or,

$$R^2 = 1 - \frac{SSE}{SST}$$

Just as before, two important limits can be placed on R^2,

$$0 \leq R^2 \leq 1$$

R^2 for The Bay Area Rapid Food Example

For the Bay Area Rapid Food data, the relevant calculation is,

$$R^2 = \frac{11,776.18}{13,581.35} = 0.867$$

Indicates we are able to explain 86.7% of the variation in total revenue by the variation in price and advertising

This leaves 13.3% of the variation unexplained, suggesting the "explanatory power" of the model is high

Sample Regression Output from Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	**0.93117**
R Square	**0.86708**
Adjusted R Square	**0.86166**
Standard Error	**6.06961**
Observations	**52**

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	**2**	**11776.1839**	**5888.0919**	**159.8280**	**0.0000**
Residual	**49**	**1805.1684**	**36.8402**		
Total	**51**	**13581.3523**			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	**104.7855**	**6.4827**	**16.1638**	**0.0000**	**91.7580**	**117.8130**
X Variable 1	**-6.6419**	**3.1912**	**-2.0813**	**0.0427**	**-13.0549**	**-0.2290**
X Variable 2	**2.9843**	**0.1669**	**17.8769**	**0.0000**	**2.6488**	**3.3198**

A Potential Problem with R^2

A property of R^2 is that it can be made large simply by adding more and more variables, even if the added variables do not have an economic justification

- Algebraically it is a fact that as variables are added the sum of squared errors *SSE* goes down (it can remain unchanged but this is rare) and thus R^2 goes up
- *SSE* cannot go up because least squares always has the option to "ignore" the added variable if it would increase *SSE* (assign a zero parameter estimate)
- R^2 could increase, not due to economic importance of added variables, but simply due to a spurious relationship between added variables and dependent variable
- If the model contains $T-1$ variables, then $R^2 = 1$
- Game of "maximizing R^2"

We can see the problem with R^2 by noting that it is based on the unexplained and explained variation in y_t,

$$R^2 = 1 - \frac{SSE}{SST}$$

Hence, there is no adjustment for the number of variables in the model through the degrees of freedom calculation

A natural solution is to use variances instead of variation

Adjusted R^2 is,

$$\bar{R}^2 = 1 - \frac{\hat{\sigma}^2}{\hat{\sigma}^2_{y_t}}$$

or,

$$\bar{R}^2 = 1 - \frac{SSE/(T-K)}{SST/T-1}$$

This alternative measure of fit takes into account the number of variables in the model by dividing the numerator and denominator by their respective degrees of freedom

For the Bay Area Burger data the value of this descriptive measure is

$$\bar{R}^2 = 1 - \frac{11{,}776.18 \,/\, 49}{13{,}581.35 \,/\, 51} = 0.862$$

There is a direct relationship between $\bar{R}^2$ and R^2,

$$\bar{R}^2 = 1 - \left[(1 - R^2) \frac{T-1}{T-K} \right]$$

Three important results,

1. If $K=1$ (intercept only model), then $\bar{R}^2 = R^2$
2. If $K > 1$, $\bar{R}^2 < R^2$
3. $\bar{R}^2$ may be negative

Divergence of opinions about the use of $\bar{R}^2$ vs. R^2 in practice,

Hill, Griffiths and Judge:

"While solving one problem, this corrected measure of goodness of fit unfortunately introduces another one. It loses its interpretation: $\bar{R}^2$ is no longer the percent of variation explained…Let's concentrate on the unadjusted R^2 and think of it as a descriptive device for telling us about the "fit" of the model."

Pindyck and Rubinfeld:

"$\bar{R}^2$ has a number of properties which make it a more desirable goodness-of-fit measure than R^2. When new variables are added to a regression model, R^2 always increases, while $\bar{R}^2$ may rise or fall. The use of $\bar{R}^2$ eliminates at least some of the incentive for researchers to include numerous variables in a model without much thought about why they should appear."

Comparing R^2's Across Regressions

It is strictly valid to compare R^2 across regression equations ONLY if the functional form of the dependent variable is the same

This is easy to see if we go back to the original definition of R^2,

$$R^2 = \frac{SSR}{SST}$$

If functional form of dependent variable is different for two regressions, SST will differ

$\Rightarrow$ If denominator (base) is different then comparison of resulting ratio is inappropriate

Even if the functional form of the dependent variable is not the same, it is possible to compare R^2 across regressions if appropriate computations are performed

Earlier we noted that,

$$r^2_{\hat{y},y} = R^2$$

where,

$$r_{\hat{y},y} = \frac{\hat{\text{cov}}(\hat{y}, y)}{\sqrt{\hat{\text{var}}(\hat{y})\,\hat{\text{var}}(y)}}$$

is the simple correlation between the predicted y and the actual y

By appropriately transforming $\hat{y}_t$ and y_t for a given functional form, and applying the above formula, we can generate an R^2 that can be validly compared to the R^2 for a different functional form

Example: linear versus double-log regressions

Assume we estimate two regressions using the same data sample, one using the linear functional form and the other using the double-log functional form

We would like to make a valid comparison of the fit across the two functional forms

1. Obtain fitted values $\hat{y}_t$ for linear regression
2. Take logarithm of fitted values for linear regresssion, $\ln(\hat{y}_t)$
3. Take logarithm of actual values, $\ln(y_t)$
4. Compute correlation coefficient between $\ln(\hat{y}_t)$ and $\ln(y_t)$
5. The square of the correlation coefficient obtained in the previous step can be compared to the R^2 of the double-log regression

TABLE 7.2
Raw data for comparing two R^2 values

Year	Y_t	$\hat{Y}_t$	$\widehat{\ln Y_t}$	Antilog of $\widehat{\ln Y_t}$	$\ln Y_t$	$\ln (\hat{Y}_t)$
	(1)	(2)	(3)	(4)	(5)	(6)
1970	2.57	2.321887	0.843555	2.324616	0.943906	0.842380
1971	2.50	2.336272	0.853611	2.348111	0.916291	0.848557
1972	2.35	2.345863	0.860544	2.364447	0.854415	0.852653
1973	2.30	2.341068	0.857054	2.356209	0.832909	0.850607
1974	2.25	2.326682	0.846863	2.332318	0.810930	0.844443
1975	2.20	2.331477	0.850214	2.340149	0.788457	0.846502
1976	2.11	2.173233	0.757943	2.133882	0.746688	0.776216
1977	1.94	1.823176	0.627279	1.872508	0.662688	0.600580
1978	1.97	2.024579	0.694089	2.001884	0.678034	0.705362
1979	2.06	2.115689	0.731282	2.077742	0.722706	0.749381
1980	2.02	2.130075	0.737688	2.091096	0.703098	0.756157

Notes: Column (1): Actual Y values from Table 3.4
Column (2): Estimated Y values from the linear model (3.7.1)
Column (3): Estimated Y values from the double-log model (6.4.5)
Column (4): Antilog of values in column (3)
Column (5): Log values of Y in column (1)
Column (6): Log values of $\hat{Y}_t$ in column (2)

Gujarati, D.N. Basic Econometrics. Third Edition. McGraw-Hill, Inc, New York, NY, 1995.

Reporting the Results of Multiple Regression Analysis

Similar formats for reporting multiple regression results and simple regression results are used

One common form is,

$$\hat{y}_t = 104.79 - 6.642x_{2,t} + 2.984x_{3,t} \qquad R^2 = 0.867$$
$$(6.48) \quad (3.191) \quad (0.167) \qquad (s.e.)$$

where *s.e.* stands for standard error

Another is to replace the standard errors by *t*-statistics for a zero null,

$$\hat{y}_t = 104.79 - 6.642x_{2,t} + 2.984x_{3,t} \qquad R^2 = 0.867$$
$$(16.17) \quad (-2.081) \quad (17.868) \qquad (t-stat.)$$

Finally, it has become commonplace in recent years to also report *p*-values in either reporting format,

$$\hat{y}_t = 104.79 - 6.642x_{2,t} + 2.984x_{3,t} \qquad R^2 = 0.867$$
$$(6.48) \quad (3.191) \quad (0.167) \qquad (s.e.)$$
$$[0.000] \quad [0.043] \quad [0.000] \qquad [p-value]$$

$$\hat{y}_t = 104.79 - 6.642x_{2,t} + 2.984x_{3,t} \qquad R^2 = 0.867$$
$$(16.17) \quad (-2.081) \quad (17.868) \qquad (t-stat.)$$
$$[0.000] \quad [0.043] \quad [0.000] \qquad [p-value]$$