

EC 413
Economic Forecast and Analysis
(Professor Lee)

Lecture 3
Regression and Forecasting

Read:

Any econometrics textbook

This lecture is about:

- Review of Regression Analysis (highlights only)
- Review of Testing Hypothesis
- How Regression Analysis is used in forecasting

1. Overview of Regression Analysis

Three goals of Regression Analysis

- Prediction (Forecasting)
- Marginal Effects
- Testing Hypothesis

Example) Hedonic Pricing Models

[Econometric model]

$$\text{Price} = \alpha + \beta_1 \text{SQFT} + \beta_2 \text{YEAR} + \beta_3 \text{POOL} + e$$

Parameters (coefficients): α , β_1 , β_2 , β_3

Error term: e

[Estimation Results]

(A) Simple Regression

$$\text{Predicted_Price} = 52,404 + 61.16 \text{ SQFT}$$

(i) Prediction

If SQFT = **2,850**,

$$\begin{aligned} \text{Predicted_Price} &= 52,404 + 61.16 * \mathbf{2850} \\ &= \$226,708 \end{aligned}$$

(ii) Marginal effect of SQFT = \$61.16

$$\Delta \text{ Price} / \Delta \text{ SQFT} = 61.16$$

“One more unit of SQFT (1 square foot) will lead to \$61.16 increase in price.”

(iii) Testing Hypothesis (on *parameters*)

$$H_0: \beta_1 = 0$$

$$H_a: \beta_1 \neq 0$$

“Significance of *the coefficient of SQFT*”

Or

$$H_0: \beta_1 = \mathbf{100}$$

$$H_a: \beta_1 \neq \mathbf{100}$$

(t-test)

(B) Multiple Regression

$$\text{Predicted_Price} = -7,434,369 + 63.38 \text{ SQFT} + 3,753 \text{ YEAR}$$

(i) Prediction

If SQFT = 2,850, YEAR = 1991, then

$$\text{Predicted_Price} = -7,434,369 + 63.38 * 2850 + 3,753 * 1991 = \$219,171.60$$

(ii) Marginal effect of SQFT = 63.38

$\Delta \text{ Price} / \Delta \text{ SQFT} = 63.38$ (partial effect after controlling the effect of YEAR)

“One more unit of SQFT will lead to \$63.38 increase in price.”

(iii) Testing Hypothesis

- One restriction:

$$H_0: \beta_1 = 0 \quad H_a: \beta_1 \neq 0$$

“Significance of the coefficient of SQFT or on β_2 (coefficient of YEAR)” (t-test)

- More than one restrictions (**F-test**)

$$H_0: \beta_1 = 0, \beta_2 = 0 \quad H_a: H_0 \text{ is not true}$$

“Joint Significance of the coefficients of
SQFT and YEAR”

General Notation:

Population regression Model

$$y = \alpha + \beta X + u \quad (\text{Price} = \alpha + \beta \text{SQFT} + u)$$

or

$$y_t = \alpha + \beta X_t + u_t \quad (\text{Price}_i = \alpha + \beta \text{SQFT}_i + u_i)$$

where u_t = error term

Sample regression Model

$$y_t = \hat{\alpha} + \hat{\beta} X_t + \hat{u}_t$$

where \hat{u}_t = residual

or

$$\hat{y}_t = \hat{\alpha} + \hat{\beta} X_t$$

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

Estimation:

Find $\hat{\alpha}$ and $\hat{\beta}$ such that SSR (Sum of squared residuals) is at the minimum. (Handout)

Two Important Issues:

(A) Coefficient of Dummy variables

= “Difference” between two groups

Ex) Wage equation (\$/hour)

$$(1) \text{ WAGE} = 10.93 - 2.73 \text{ SEX}$$

$$(2) \text{ WAGE} = -1.97 - 2.31 \text{ SEX} + 0.96 \text{ ED}$$

where SEX = 1 for Female workers, and 0 for male workers

How do you interpret the coefficient of SEX in each equation?

(B) Logged data

=> “percentage changes or percentage difference”

Ex) Log-Wage equation

$$\mathbf{\text{Log(WAGE)} = .947 - 0.211 \text{ SEX} + .097 * \text{EDU}}$$

How do you interpret the coefficient of SEX in each equation?

How about the coefficient of EDU?

Trend and Seasonality in Regression

Consider :

$$y_t = \alpha + \beta \cdot t + u_t$$

where $t = 1, 2, 3, \dots, T$. (t is a trend function)

Here, β denotes the average change of y_t .

$$\beta = \Delta y / \Delta t = \Delta y / 1 = \Delta y \quad (\Delta t = t - (t-1) = 1)$$

= average change over time

= **trend coefficient**

Ex) How much did the salary increase per year?

$$\text{Salary}_t = \alpha + \beta \cdot t + u_t$$

The estimated coefficient of β gives the answer:
(average salary increase per year = β).

Consider :

$$\text{Log}(y_t) = \alpha + \beta \cdot t + u_t$$

$$\beta = \Delta \log(y) / \Delta t = \Delta \log(y) / 1 = \Delta y / y$$

= average *percentage change* over time
(average growth rate)

2. Testing Hypothesis in Regression

Example)

$$\text{Predicted_Price} = 52,404 + 61.16 \text{ SQFT}$$

Question: Is the coefficient of SQFT significant?

- Is the coefficient different from zero?
- Is the variable SQFT an important factor to explain the housing price?
- Is SQFT associated with Price?

We denote:

$$H_0: \beta = 0 \quad (\text{null hypothesis})$$

$$H_a: \beta \neq 0 \quad (\text{alternative hypothesis})$$

$$\text{Price} = \alpha + \beta \text{ SQFT} + u$$

[Note: $H_a: \beta > 0$ or $\beta < 0$... one tailed test]

Three approaches

(a) t-statistic

$$t^* = (\hat{\beta} - 0) / s(\hat{\beta})$$

where $s(\hat{\beta})$ is S.E of $\hat{\beta}$.

Find the Critical value (t_c).

- Degree of freedom ($df = T - k - 1$)
- Significance level (α): $\alpha = 5\%$, typically

Decision Rule

Two tailed test

“Reject H_0 if $|t^| > t_c$.”*

One tailed test

“Reject H_0 if $t^ > t_c$ for the right-tailed test.”*

“Reject H_0 if $t^ < -t_c$ for the left-tailed test.”*

Ex) $T = 33$, (Standard errors in parentheses)

$$Y = 102,192 - 9,075 C + .357 P + 1.288 I$$

$$(2,035) \quad (.0727) \quad (.543)$$

$$df = T - k - 1 = 33 - 3 - 1 = 29$$

$$\alpha = 5\%$$

$$t_c = 2.045 \text{ (for two tailed test)}$$

We may have three different t-tests for each coefficient.

$$Y = \alpha + \beta_1 C + \beta_2 P + \beta_3 I + u$$

(i) $H_0: \beta_1 = 0, H_a: \beta_1 \neq 0$

$$t^* = (-9075 - 0) / 2035 = 4.42$$

Since $|t^*| > t_c$, we reject the null at $\alpha = 5\%$.

(The coefficient of C is significant.)

$$(ii) \quad H_0: \beta_2 = 0, \quad H_a: \beta_2 \neq 0$$

$$t^* = (0.357 - 0) / 0.0727 = 4.88$$

Since $|t^*| > t_c$, we reject the null at $\alpha = 5\%$.

(The coefficient of P is significant.)

(iii) Left as an exercise.

(b) Confidence interval

Find $(100-\alpha)\%$ C.I.

$$= \hat{\beta} \pm t_c * (\text{std error of } \hat{\beta})$$

Decision Rule:

“Reject H_0 if \mathbf{b}_0 (value under the null) lies outside the C.I.”

Back to the previous example,

$$Y = 102,192 - 9,075 C + .357 P + 1.288 I$$

$$(2,035) \quad (.0727) \quad (.543)$$

$$(i) \quad H_0: \beta_1 = 0, \quad H_a: \beta_1 \neq 0$$

Here, $\beta_0 = 0$. (Note: $t_c = 2.045$)

$$\begin{aligned}
 95\% \text{ C.I.} &= \hat{\beta}_1 \pm t_c * (\text{std error of } \hat{\beta}_1) \\
 &= -9075 \pm 2.045 * 2035 \\
 &= -9075 \pm 4161 \\
 &= -13236 \sim -4911
 \end{aligned}$$

This C.I. does NOT include **0**. Thus, we reject the null at the 5% level.

(ii) (iii) Left as an exercise.

(c) P-value

- P-value = “the minimum significance level (α) at which H_0 is rejected.”

Point: If p-value < 5%, reject H_0 at the 5% level.

Point: Smaller p-value implies “*rejection*” of H_0 .
Smaller p-value implies “large” t-statistic.

Ex) Back to the previous example,

	$Y = 102,192 - 9,075 C + .357 P + 1.288 I$		
Std. Err →	(2,035)	(.0727)	(.543)
t-stat. →	(-4.42)	(4.88)	(2.37)
p-value →	(.0001)	(.0000)	(.0246)

Thus, each coefficient is viewed as significant. The null hypothesis of insignificant coefficient ($\beta = 0$) is rejected for each.

- Computing p-values (if df is big).
 - Two tailed test
 - $P\text{-value} = 2 * P(t > |t^*|)$
 - (eg) if $t^* = -1.96$, $p\text{-value} = 2 * P(t > 1.96)$
 - $= 2 * P(Z > 1.96)$ (if df is big) $= 0.05$
 - One tailed test
 - $P\text{-value} = P(t > |t^*|)$
 - (eg) if $t^* = -1.96$, $p\text{-value} = P(t > 1.96)$
 - $= P(Z > 1.96)$ (if df is big) $= 0.25$

Example) # of Extra-marital Affairs

Questions: Which variables are determinants?

Dependent Variable: AFFAIRS

Included observations: 601

Variable	Coeff.	Std. Err	t-Stat.	Prob.
AGE	-0.02920	0.011652	-2.50598	0.0125
EDU	0.000466	0.033040	0.014103	0.9888
KIDS	0.006239	0.180400	0.034584	0.9724
OCCU	0.047777	0.045787	1.043458	0.2972
RATING_M	-0.39259	0.061821	-6.35053	0.0000
RELIGION	-0.24994	0.057556	-4.34253	0.0000
SEX	0.065137	0.154795	0.420794	0.6741
YRS_MARR	0.085060	0.021237	4.005333	0.0001
C	3.177421	0.585972	5.422475	0.0000

- (a) How do you interpret each coefficient?
- (b) Which coefficients are significant?
- (c) Test whether the coefficient of SEX is significant.

Joint Restrictions

Consider a regression model,

$$(1) \quad Y = \alpha + \beta_1 C + \beta_2 P + \beta_3 I + u$$

We wish to test the joint hypothesis.

$$H_0: \beta_1 = 0, \beta_2 = 0 \quad H_a: H_0 \text{ is not true}$$

Note: One may consider two different t-tests on each. But, rejections or non-rejections from both tests do not necessarily imply that this joint restriction will or will not hold.

If the null is true ($\beta_1 = 0, \beta_2 = 0$), the model (1) becomes:

$$(2) \quad Y = \alpha + \beta_3 I + e$$

Important Fact:

The value of R-square from (1) is always higher than the value of R-square from (2).

$$R_1^2 > R_2^2$$

The increase in R-square by adding two variables (C and P) is $R_1^2 - R_2^2$. If this increase is big enough, then we can say that both variables are important and β_1 or β_2 or

both are significant. The F-test was driven from this fact.

Due to the same reason, the residual sum of squares (RSS) is smaller in (1) than in (2).

$$RSS_1 < RSS_2$$

F-statistic

$$F^* = [(R_1^2 - R_2^2)/m] / [(1 - R_1^2)/(n-k-1)]$$

Or

$$F^* = [(RSS_2 - RSS_1)/m] / [RSS_1/(n-k-1)]$$

Decision Rule

Find the Critical value (F_c) first.

$$df = (m, n-k-1)$$

"Reject H_0 if $F^* > F_c$."

; Always one (right, upper) tailed test.

Our Questions:

"Is there a trend in the data? How about seasonality?"

(1) Trend

$$y_t = \alpha + \beta \cdot t + u_t$$

Run this regression and
Test whether $\beta = 0$ or not.

$$H_0: \beta_1 = 0, \quad H_a: \beta_1 \neq 0$$

Example) Gap sales data

(2) Seasonality

Define $D_{1t} = 1$ for Spring or 0 otherwise

$D_{2t} = 1$ for Summer or 0 otherwise

$D_{3t} = 1$ for Fall or 0 otherwise

$D_{4t} = 1$ for Winter or 0 otherwise

$$y_t = \alpha + d_1 D_{1t} + d_2 D_{2t} + d_3 D_{3t} + u_t$$

(D_{4t} is dropped. Then, it is used as a reference for comparison.)

d_1 = difference between Spring and Winter

d_2 = difference between Summer and Winter

d_3 = difference between Fall and Winter

Run this regression and test

$$H_0: d_1 = d_2 = d_3 = \mathbf{0}, \quad H_a: H_0 \text{ is not true}$$

Use F-test (joint hypothesis.)

Example) Gap sales data

Using RATS for Regression

Exercise Questions

1. The following model was fitted, to explain the selling prices of houses, to a sample of 815 sales. (standard errors in parentheses)

$$Y = -1264 + 48.18 x_1 + 3382 x_2 - 1859 x_3 + 3219 x_4 + 2005 x_5,$$

$$(0.91) \quad (515) \quad (488) \quad (947) \quad (768)$$

y = Selling price of house, in dollars

x_1 = Square footage of living area

x_2 = Size of garage, in number of cars

x_3 = Age of house, in years

x_4 = Dummy variable taking the value of 1 if the house has a fireplace, and 0 otherwise

x_5 = Dummy variable taking the value 1 if the house has brick siding and 0 if it has vinyl siding

- Carefully interpret the coefficient of x_1 .
- Carefully interpret the coefficient of x_4 .
- Carefully interpret the coefficient of x_5 .
- Test the null hypothesis that type of siding has no impact on selling price, against the alternative that, all other things equal, houses with brick siding have a different selling price than houses with vinyl siding. Use 5% significance level.
- Using the C.I. method, test the same hypothesis in part (d). Use 5% significance level.
- Is the p-value of the test in part (d) higher than 5%?
- Test the null hypothesis that type of siding has no impact on selling price, against the alternative that, all other things equal, houses with brick siding have a higher selling price than houses with vinyl siding.

2. Use the SP500 data (file: SP500_holt.xls).

- Using a regression model, find the average change of the data. Then, test the existence of a (linear) trend.
- Using a regression model Find the average *percentage* change of the data.