Contents

Quantitative Methods

Quantitative Methods

Multiple Regression

Basics of Multiple Regression
and Underlying Assumptions

KAPLAN SCHWESER

Multiple Regression

- **Basic idea**: Linear relationship with more than one independent variable

Dependent variable

Independent variables

$$Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + \ldots + b_k X_{ki} + \varepsilon_i$$

Intercept

Slope coefficients

Error term

Interpreting *p*-Values

- Another way to test H_0 (always agrees with *t*-test)
- If *p*-value < α, then reject H_0
    - **Example**: If α = 5% (i.e., a 95% confidence level) and the calculated *p*-value = 0.07, we fail to reject H_0
- *p*-value can be viewed as smallest significance level (α) at which we can reject H_0

Example: Interpreting *p*-Values

Variable	b_i	S_{bi}	*t-stat*	*p-value*
b_0	–11.6%	1.657%	–7.0	<0.0001
PR	0.25	0.032	7.8	<0.0001
YCS	0.14	0.28	0.5	0.62

- b_0 and PR are significant (high *t*-stat, low *p*-value) while YCS is insignificant (low *t*-stat and high *p*-value)

Formulate Regression Equation

- Suppose we think S&P 500 10-year real earnings growth (EG10) is explained by:*
  - Trailing dividend payout ratio of index (PR)
  - Yield curve slope (YCS)
- Example: Formulate a multiple regression equation to describe this relationship

*Adapted from Arnott and Asness (2003)

Formulate Regression Equation

$$EG10 = b_0 + b_1 \times PR + b_2 \times YCS + \varepsilon$$

where:

EG10 = 10-year real earnings growth in S&P 500

PR = trailing dividend payout ratio of index

YCS = yield difference between 10-year T-bond and 3-month T-bill

Formulate Regression Equation

- Regress EG10 on independent variables
  - 46 annual observations (n = 46)
  - Two independent "X" variables (k = 2)

Note: You will need n and k for several formulae

Regression Output: Frequent Starting Point for Exam Questions

Variable	b_i	S_{bi}
b_0	–11.6%	1.657%
PR	0.25	0.032
YCS	0.14	0.28

Interpreting Coefficients

- EG10 will equal –11.6% if the two independent variables are both zero.
- All else constant:
  - 1% increase in PR = 0.25% increase in EG10.
  - 1% increase in YCS = 0.14% increase in EG10.

Assumptions of Multiple Regression

- Linear relationship between Y and X
- No exact linear relationship among X's
- Expected value of error term = 0
- Variance of error term is constant
- Errors not serially correlated
- Error term normally distributed

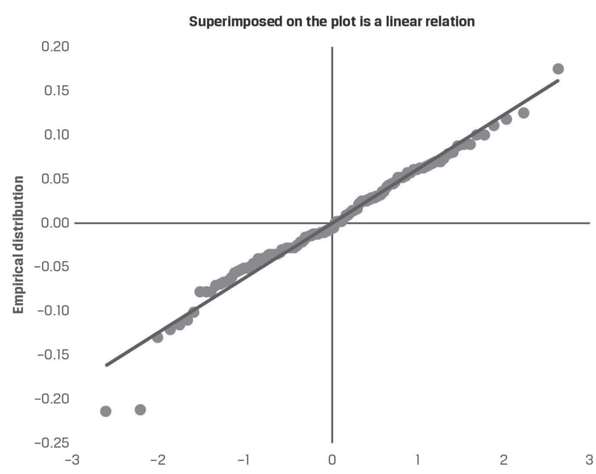
These are related to violations (see later)

Normal Q-Q Plot

- A normal Q-Q plot (usually called a Q-Q plot) compares a variable's distribution to a normal distribution.
  - This helps us to visualize the distribution.
- The Q-Q plot is useful for exploring whether the residuals are normally distributed.
  - This is a key assumption of linear regression.

Normal Q-Q Plot Example

- In this example, the regression model error term is close to being normally distributed
- Though a handful of outliers lie above or below the theoretical range



CFA® Program Curriculum 2023 Level 2, Volume 1, page 17

Quantitative Methods

Multiple Regression

Evaluating Regression Model Fit
and Interpreting Model Results

ANOVA Table

Source of Variation	*Df*	*Sum of Squares*	*Mean Square*
Regression (explained)	k	RSS	MSR
Error (unexplained)	n – k – 1	SSE	MSE
Total	n – 1	SST	

These columns vertically sum

Example: ANOVA

- An analyst runs a regression of monthly value-stock returns on five independent variables over 60 months. The total sum of squares for the regression is 460, and the sum of squared errors is 170.

- Prepare an ANOVA table.
- Use information from the ANOVA table to calculate *F*-test and R^2 *(see later)*.

Example: ANOVA Table

Source of Variation	*Df*	*Sum of Squares*	*Mean Square*
Regression (explained)	5	290	58
Error (unexplained)	54	170	3.15
Total	59	460	

These columns vertically sum

Coefficient of Determination (a.k.a. R^2)

- R^2 measures percentage of total variation in dependent "Y" variable explained by independent "X" variable
- R^2 ranges between 0 and 1
- An R^2 of 0.25 means "X" explains 25% of the variation in "Y"

Calculating R^2

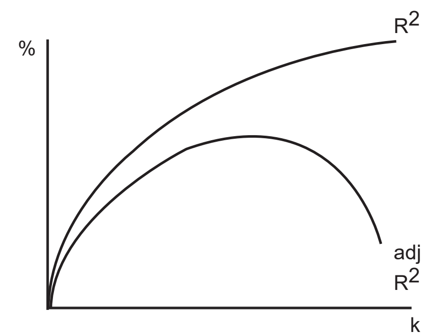
$$R^2 = \frac{\text{explained variation}}{\text{total variation}} = \frac{RSS}{SST}$$

- For simple linear regression (i.e., k = 1): $R^2 = (r_{XY})^2$
- For multiple linear regression: $R^2 = \left(r_{Y,\hat{Y}}\right)^2$
- $r_{y,\hat{y}}$ is called the *multiple r*

Adjusted R^2

- Unadjusted R^2 increases when new variables are added.
- Adjusted R^2 applies a penalty factor to reflect quality of added variables.

$$R^2_{adj} = 1 - \left[\left(\frac{n-1}{n-k-1}\right) \times \left(1 - R^2\right)\right]$$



 7

Calculating R^2 and Adjusted R^2

- Given regression with five independent variables, n = 60, SST = 460, and SSE = 170, calculate R^2 and adjusted R^2:

$$R^2 = \frac{460 - 170}{460} = 63.0\%$$

$$R^2_{adj} = 1 - \left[\left(\frac{60-1}{60-5-1}\right) \times \left(1 - 0.63\right)\right] = 59.6\%$$

 8

Interpreting Adjusted R^2

- Suppose we add four more independent variables. R^2 increases from 63% to 65%, but adjusted R^2 falls from 59.6% to 58.7%. Which model should we choose?

- Answer: Choose the original model because it has a higher adjusted R^2. It has better explanatory power and uses five rather than nine variables.

 9

AIC and BIC Information Criteria

- Akaike's information criterion (AIC) and Schwarz's Bayesian information criteria (BIC) can also be used to evaluate models.
- AIC and BIC are used to select the "best" model among a group with the same dependent variable.

Lower values of both measures are better.

 10

Akaike's Information Criterion (AIC)

Akaike's information criterion (AIC):

$$AIC = n \times \ln\left(\frac{SSE}{n}\right) + 2(k+1)$$

- AIC is preferred if the purpose is prediction (i.e., the goal is to have a better forecast).

Schwarz's Bayesian Information Criteria (BIC)

Schwarz's Bayesian information criteria (BIC):

$$BIC = n \times \ln\left(\frac{SSE}{n}\right) + \ln(n) \times (k+1)$$

BIC is preferred if goodness of fit is the goal.

- BIC imposes a higher penalty for overfitting.

Goodness of Fit for the Rent Model

- Consider a model to explain variation in office rent using age of the property, distance from nearest metro station, and number of restaurants in walking distance.
- This table shows the results for different models:

Variables	K	SSR	SSE	R^2	R^2-Adj	AIC	BIC
Age	1	3318.9	32627.3	9.23%	8.75%	985.9	992.4
Age + Distance	2	20946.1	15000.2	58.27%	57.8%	839.4	849.2
Age + Dist. + Rest.	3	22395.6	13550.5	62.30%	61.7%	822.0	835.0

Goodness of Fit for the Rent Model (continued)

Question: Which model is the most appropriate for use in generating forecasts?

Answer: The model with all three independent variables has the lowest AIC and, hence, is the most appropriate model for generating forecasts.

Variables	K	SSR	SSE	R^2	R^2-Adj	AIC	BIC
Age	1	3318.9	32627.3	9.23%	8.75%	985.9	992.4
Age + Distance	2	20946.1	15000.2	58.27%	57.8%	839.4	849.2
Age + Dist. + Rest.	3	22395.6	13550.5	62.30%	61.7%	822.0	835.0

Goodness of Fit for the Rent Model (continued)

Question: Which model has a better goodness of fit??

Answer: The model with all three independent variables has the lowest BIC and, hence, the best goodness of fit.
Also, the adjusted R^2 for the three-factor model is the highest of the three models.

Variables	K	SSR	SSE	R^2	R^2-Adj	AIC	BIC
Age	1	3318.9	32627.3	9.23%	8.75%	985.9	992.4
Age + Distance	2	20946.1	15000.2	58.27%	57.8%	839.4	849.2
Age + Dist. + Rest.	3	22395.6	13550.5	62.30%	61.7%	822.0	835.0

Joint Hypothesis Test

- The **joint *F*-test** is used to jointly test a subset of variables in a multiple regression.
  - The "restricted" model is based on a narrower set of independent variables nested in the broader "unrestricted" model.

Unrestricted model: $Y_i = b_0 + b_1 X_1 + b_2 X_2 + b_3X_3 + \varepsilon_i$

Restricted model: $Y_i = b_0 + b_1 X_1 + \varepsilon_i$

Joint Hypothesis Test

We want to test the following hypothesis:

H_0: $b_2 = b_3 = 0$. vs. H_a: b_2 or $b_3 \neq 0$.

The null hypothesis is that the slope coefficients of all independent variables outside the restricted model are zero.

Decision rule: reject H_0 if F (test-statistic) > F_C (critical value)

- *F*-test evaluates whether the relative decrease in SSE due to inclusion of *q* additional variables is statistically justified.

Joint Hypothesis Test

- We calculate the *F*-statistic to test this hypothesis as:

$$F = \frac{(SSE_R - SSE_U)/q}{(SSE_U)/(n-k-1)}$$

with *q* and (n–k–1) degrees of freedom.

where: R → represents the restricted model
U → represents the unrestricted model
q = number of excluded variables in the restricted model
k = independent variables in the full model

Example: Joint Hypothesis Test

- Continuing our previous rental model data, and using the information below, conduct a joint test of hypotheses for the slope coefficients 2 and 3 at a 5% level of significance.

Model	k	SSE	n
Unrestricted	3	13,550.64	191
Restricted	1	32,627.29	191

Example: Joint Hypothesis Test

- Partial F-table:

	Numerator df		
Denominator df	1	2	3
186	3.89	3.04	2.65
187	3.89	3.04	2.65
188	3.89	3.04	2.65
189	3.89	3.04	2.65
190	3.89	3.04	2.65
191	3.89	3.04	2.65

Example: Joint Hypothesis Test

Answer:

$$F=\frac{(SSE_R - SSE_U)/q}{(SSE_U)/(n-k-1)}=\frac{(32{,}627.29-13{,}550.64)/2}{(13{,}550.64)/(191-3-1)}=131.63$$

Critical *F*-value with 2 and 187 degrees of freedom is 3.04.

We reject the null hypothesis and conclude that at least one of the two slope coefficients of the excluded variables is statistically different from 0.

F-test

- The general linear *F*-test is an extension of the joint *F*-test.
- The null hypothesis is that the slope coefficients on all independent variables in the unrestricted model = zero.

H_0: $b_1 = b_2 = b_3 = 0$ versus H_a: at least one $b_j \neq 0$

The *F*-statistic then becomes:

$$F=\frac{(RSS_U)/k}{(SSE_U)/(n-k-1)}$$