品职教育
PZACADEMY.COM

Quantitative Methods

CFA二级知识框架图

Quantitative Methods

讲师：何　旋

Framework of CFA II Quantitative Methods

Regression Analysis →
- Correlation and Regression
- Multiple Regression and Issues In Regression Analysis

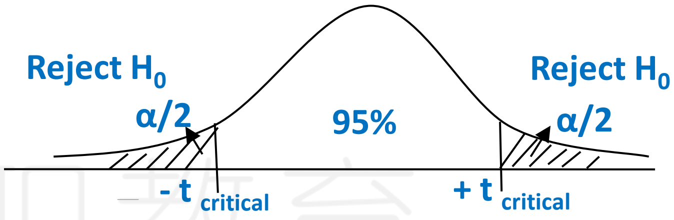
Time-series Analysis →
- Time-series Analysis

Probabilistic Approaches →
- Excerpt From “Probabilistic Approaches: Scenario Analysis, Decision Trees, And Simulations”

Reading 9

CORRELATION AND REGRESSION

Correlation Analysis

Significance test (Hypothesis test) ★★ **计算**	H_0: ρ=0; t-test $t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$, $df = n-2$  结论：例如***reject the null hypothesis*** →the correlation coefficient between X and Y is significantly different from zero.
Limitation	① Outliers (异常值)：few extreme values ② Spurious correlation：no economic explanation ③ Nonlinear relationships：Correlation only measures the linear relationship

Simple Linear Regression

建模 → ANOVA Table分析 → 检验模型 → 预测

1.建模

$$Y_i = b_0 + b_1 X_i + \varepsilon_i, i = 1,\dots,n$$

- Y_i = ***dependent variable***, explained variable, predicted variable
- X_i = ***independent variable***, explanatory variable, predicting variable.

Assumption ★	① A linear relationship exists between X and Y ② X is uncorrelated with the error term. ③ The expected value of the error term is zero (i.e., $E(\varepsilon_i)=0$) ***④ The variance of the error term is constant (homoskedastic)*** ***⑤ The error term is uncorrelated across observations ($E(\varepsilon_i\varepsilon_j)=0$ for all $i \neq j$)*** ⑥ The error term is normally distributed.
Coefficient估计 ★	解释： ✓ An estimated slope coefficient of 2： Y will change two units for every 1 unit change in X. ✓ Intercept term of 2%： the X is zero, Y is 2%.
	$b_1 = \frac{Cov(X,Y)}{Var(X)}$ $b_0 = \bar{Y} - b_1\bar{X}$ **计算**

2. ANOVA Table分析 ★★

	df	SS	MSS
Regression	k=1	RSS	MSR=RSS/k
Error	n-k-1	SSE	MSE=SSE/(n-k-1)
Total	n-1	SST	-

Coefficient Determination (R^2)	**计算：** $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
	$R^2 = r^2_{Y\hat{Y}}$ (多元都成立) $R^2 = r^2_{XY}$ (一元)
	解释： R^2 of 0.90 indicates that the variation of the independent variable explains 90% of the variation in the dependent variable.
SEE	**计算：** $SEE = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$
	性质： ✓ The SEE gauges the “fit” of the regression line. The smaller the standard error, the better the fit. ✓ The SEE is the standard deviation of the error terms in the regression.

3. 检验模型：回归分析相当于抽样估计

考试时给定条件 →

	Coefficient	Standard deviation	t-statistic	p-value
Intercept	$\hat{b}_0$	$s_{\hat{b}_0}$	?	0.18
Slope	$\hat{b}_1$	$s_{\hat{b}_1}$	?	<0.001

参数估计★★ (confidence interval)	$\hat{b}_1 \pm t_c s_{\hat{b}_1}$ → t_c: 查表 { t分布; Confidence level (置信度) } As SEE rises, $s_{\hat{b}_1}$ also increases
假设检验★★ (significance test)	✓ H_0: b_1=0 (没有特殊说明，题目中假设检验都是检验是否为0) $$t = \frac{\hat{b}_i - 0}{s_{\hat{b}_i}}$$ df=n-2 ✓ Decision rule: reject H_0 if $+t_{critical} < t$, or $t < -t_{critical}$ ✓ Rejection of the null means that the slope coefficient is different from zero

4.预测(Predicted Value of Y)

Point estimate★	$\hat{Y} = \hat{b}_0 + \hat{b}_1 X'$
Confidence interval estimate	$\hat{Y} \pm (t_c \times s_f)$ 了解

- When predicting Y using linear regression model, we encounter **two types of uncertainty**:
  - ***Uncertainty in the regression model itself***, as reflected in the standard error of estimate;
  - Uncertainty about the estimates of the regression model's ***parameter***.

Reading 10

MULTIPLE REGRESSION AND ISSUES IN REGRESSION ANALYSIS

Multiple Regression & Simple Linear Regression

Difference	要点
Interpreting the coefficient	Each slope coefficient is the estimated change in Y for a one unit change in X_i, ***holding the other independent variables constant.***
Significance test ★★	**单个检验**(t-test)： H_0: $b_j=0$ $t=\frac{\hat{b}_j}{s_{\hat{b}_j}}$ df = n-k-1
	联合检验(F-test)： The test assesses the effectiveness of the model as a whole in explaining the dependent variable H_0: $b_1 = b_2 = b_3 = \ldots = b_k = 0$ H_a: at least one $b_j \neq 0$ (j = 1 to k) reject H_0 : if F (test-statistic) > F_c (critical value) The F-test here is always a ***one-tailed test*** $F=\frac{MSR}{MSE}=\frac{SSR/k}{SSE/(n-k-1)}$ Analysts typically ***do not use ANOVA and F-tests in simple linear regression*** because the ***F-statistic is the square of the t-statistic*** for the slope coefficient.

Multiple Regression & Simple Linear Regression (Cont.)

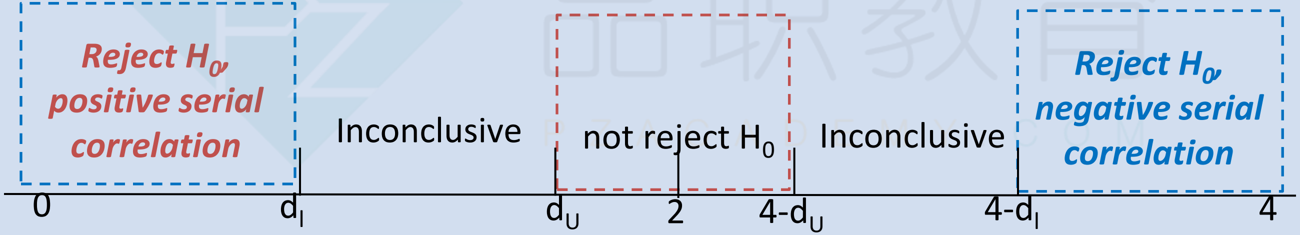
Difference	要点
R^2 ★★	**解释：** R^2 of 0.90 indicates that the model, ***as a whole***, explains 90% of the variation in the dependent variable.
	缺点： R^2 almost always increases as variables are added to the model, even if the ***marginal contribution*** of the new variables is not statistically significant.
	(计算) **Adjusted R^2：** $adjusted\ R^2 = 1 - \left[\left(\frac{n-1}{n-k-1}\right)\left(1-R^2\right)\right]$ $adjusted\ R^2 = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$
	$R^2 = r^2_{Y\hat{Y}}$ $R^2 \neq r^2_{XY}$

Multiple Regression Assumption Violations★★

Heteroskedasticity

Impact	Unconditional: no major problems	
	Conditional: significant problems ✓ Not affect the consistency of parameter estimators ✓ Coefficient estimates are not affected ✓ Standard errors are usually unreliable estimates • too small—>Type I error • too large—>Type II error	
Detection	Breusch-Pagen χ^2 test H_0: ***No heteroskedasticity*** $BP = n \times R_{residual}^2$, df=k, one-tailed test	
Correction	✓ robust, or White-corrected standard errors ✓ generalized least squares	

Serial Correlation (Autocorrelation)

Impact	✓ Serial correlation is often found in time series data ✓ Not affect the consistency of estimated regression coefficients and coefficient estimates ✓ ***Positive serial correlation is much more common***: Positive serial correlation → ***coefficient standard errors that are too small***→ Type I error & F-test unreliable
Detection	***Durbin-Watson test*** (看下图) → H_0: No serial correlation, ***DW ≈ 2×(1−r)*** 
Correction	✓ adjusting the coefficient standard errors (e.g., Hansen method): the Hansen method also corrects for conditional heteroskedaticity. ✓ incorporate the time-series nature

Multicollinearity

Impact	• The situation that two or more independent variables are highly (but not perfectly) correlated with each other. • Not affect the consistency of regression coefficient estimate • Estimates become extremely imprecise and unreliable. • Impossible to distinguish the individual impacts of the independent variable
Detection	① ***t-tests*** indicate that none of the individual coefficients is significantly different than zero, while the ***F-test*** indicates overall significance and the ***R^2*** is high ② $\lvert r_{X1X2} \rvert > 0.7$
Correction	Remove one or more independent variables

Dummy Variables ★★

模型	Qualitative variable： 0 and 1 n categories → n−1 dummy variables 例： $EPS_t = b_0 + b_1Q_{1t} + b_2Q_{2t} + b_3Q_{3t} + \varepsilon_t$ • EPS_t = a quarterly observation of earnings per share • Q_{1t} =1 if period t is the first quarter, Q_{1t} =0 otherwise • Q_{2t} =1 if period t is the second quarter, Q_{2t} =0 otherwise • Q_{3t} =1 if period t is the third quarter, Q_{3t} =0 otherwise
Interpreting the coefficients	✓ b_0： average value of EPS for the fourth quarter ✓ Slope coefficient： difference in EPS (on average) between the respective quarter (i.e., quarter 1, 2, or 3) and the omitted quarter. 比如， $b_1 = EPS_1 – EPS_4$

Model Misspecification

① The functional form can be misspecified.
- Important variables are omitted.
- Variables should be transformed.
- Data is improperly pooled.

② Time series misspecification.
- A lagged Y is used as an X with serially correlated errors.
- A function of the Y is used as an X (forecasting the past).
- Independent variables are measured with error.

③ Time-series data: nonstationarity.

Reading 11

TIME-SERIES ANALYSIS

Trend Models

Linear trend: $y_t=b_0+b_1t+\varepsilon_t$

- The data points appear to be ***equally*** distributed above and below the regression line
- 每期增长量是***constant amount***

Log-linear trend model: $Ln(y_t)=b_0+b_1t+\varepsilon_t$

- If the data plots with a non-linear (curved) shape, then the residuals from a linear trend model will be ***persistently positive or negative for a period of time***
- 每期的增长率是***constant rate***

用DW test检验ε是否有serial correlation

No → 使用trend model

Yes ↓

AR Model

以AR(1)开始模型的估计

$$AR(P) \to x_t=b_0+b_1x_{t-1}+b_2x_{t-2}+\ldots+b_px_{t-p}+e_t$$

Chain rule of forecasting★★	$\hat{x}_{t+1}=\hat{b}_0+\hat{b}_1x_t$ 计算
Assumption (具体看后面)★★	✓ No autocorrelation ✓ No Conditional Heteroskedasticity ✓ Covariance-stationary series
检验是否有Seasonality (具体看后面) ★	$x_t=b_0+b_1x_{t-1}+b_2x_{t-4}+\varepsilon_t$ (检验有无seasonality方法和autocorrelation类似)
Compare forecasting power	smallest RMSE for out-of-sample →最好 **RMSE计算**

AR Model assumption★★

1、No autocorrelation: 针对residual term

Detection	$H_0: r_{e_t,e_{t-k}} = 0$ → No autocorrelation $t\text{-}statistics = \dfrac{r_{e_t,e_{t-k}}}{1/\sqrt{n}}$ standard error $= 1/\sqrt{n}$ Reject H_0: $t > +t_{critical}$, or $t < -t_{critical}$
Correction	Reject H_0: **(add lagged values) AR(1) → AR(2)**

→ 考试时给的表格

Lag	Autocorrelation	Standard Error	*t*-Statistic
Autocorrelations of the Residual			
1	−0.1538	0.0528	−2.9142
2	0.1097	0.0528	2.0782
3	0.0657	0.0528	1.2442
4	0.0920	0.0528	1.7434

检验是否有Seasonality

Detection	$H_0: r_{e_t,e_{t-4}} = 0$ → No seasonality Reject H_0: $t > +t_{critical}$, or $t < -t_{critical}$
Correction	**add lagged values：** $x_t=b_0+b_1 x_{t-1}+ b_2 x_{t-4}+\varepsilon_t$

2、No Conditional Heteroskedasticity: 针对residual term (用ARCH)

含义	Heteroskedasticity refers to the situation that the variance of the error term is not constant
Detection	Test Conditional Heteroskedasticity = Test whether a time series is ARCH(1) $\varepsilon_t^2 = a_0 + a_1\varepsilon_{t-1}^2 + u_t$ **a_1 is significantly different from 0** → Conditional Heteroskedasticity exist
Correction	**Generalized least squares**

3、Covariance-stationary series: 针对X_t $x_t = b_0 + b_1 x_{t-1} + e_t$

含义

Covariance-stationary ⟺ **Mean reversion**

① Constant and finite expected value of the time series
② Constant and finite variance of the time series
③ Constant and finite covariance with leading or lagged values

Mean-reverting: $\frac{b_0}{1-b_1}$

- x_t >mean—> x_{t+1} <x_t
- x_t <mean—> x_{t+1} >x_t

相反

Random walk ⟺ b_1=1 (undefined mean) ⟺ unit root ⟺ nonstationary

Simple random walk: $x_t = x_{t-1} + \varepsilon_t$
Random walk with a drift: $x_t = b_0 + b_1 x_{t-1} + \varepsilon_t$ ($b_0 \neq 0$)

检验

The unit root test of nonstationarity： ***Dickey-Fuller test (DF test)***

$x_t = b_0 + b_1 x_{t-1} + \varepsilon_t \rightarrow x_t - x_{t-1} = b_0 + g x_{t-1} + \varepsilon_t$ ($g = b_1 - 1$)
H_0: g=0 (has a unit root and is nonstationary); H_a: g<0
Reject $H_0 \rightarrow$ the time series does not have a unit root and is stationary

修正

First differencing

Define y_t as $y_t = x_t - x_{t-1} \rightarrow$ AR(1) model $y_t = b_0 + b_1 y_{t-1} + u_t$

$$Y_t = b_0 + b_1 X_t + e_t$$

X_t，Y_t都是time series data

Regression with More Than One Time Series★★

Scenarios		是否可做多元回归
None of the time series has a unit root		✔
At least one time series has a unit root while at least one time series does not		✖
Each time series has a unit root: whether the time series are cointegrated?	conintegrated	✔
	no cointegration	✖

Test the cointegration: Dickey-Fuller Engle-Granger test (DF-EG test)

H_0: no cointegration H_a: cointegration

- If we cannot reject the null, we cannot use multiple regression
- If we can reject the null, we can use multiple regression

Reading 12

EXCERPT FROM “PROBABILISTIC APPROACHES: SCENARIO ANALYSIS, DECISION TREE, AND SIMULATION”

Overall Assessment	

1. Selective & full risk analysis	• Scenario analysis: ***Selective*** risk analysis
	• Decision trees and simulations: ***full*** risk analysis
	    • Decision trees: ***a manageable set of possible outcomes.***
	    • Simulations: ***probability distributions***

2. Type of risk	• Scenario analysis and decision trees: ***discrete*** outcomes in risky events
	    • Scenario analysis: easier to use when risks occur ***concurrently***
	    • Decision trees: better suited for ***sequential risks***
	• Simulations: better suited for ***continuous risks.***

3. Correlation across risks	• Simulations allow for ***explicitly*** modeling these correlations. • In scenario analysis, we can deal with correlations ***subjectively*** by creating scenarios that allow for them. • Correlated risks are ***difficult to model in decision trees.***
4. the quality of the information	Simulations work best in cases where there is ***substantial historical and cross sectional data available***
5. Complement or replacement for risk-adjusted value	• Both decision trees and simulations are approaches that can be used as either ***complements to or substitutes for risk-adjusted value.*** • Scenario analysis will always be a ***complement to risk-adjusted value***, since it does not look at the full spectrum of possible outcomes.


Thank You!