# Quantitative Methods

## Study Session 3 Quantitative Methods

### Reading 6 Fintech in Investment Mgmt

#### Data Science

Data processing: capture/ curation (data quality and accuracy assurance)/ storage/ search/ transfer (from data source/ storage to analytical tool)

Data visualization: heat map, interactive 3-D, tag cloud, network graphs, tree diagrams

#### Selected Applications of Fintech to IM

Text analytics and NLP: analyze annual report, transcripts, policy statements, voices etc.

Robo-Advisory Services: fully automated/ advisor-assisted

Risk Analysis: stress testing and risk assessment

Algorithmic Trading: HFT

#### DLT

Not fully secure, breaches in privacy and data protection are possible.

DLT: a type of database that may be shared among entities in a network.

Cryptography: an algorithmic process to encrypt data, making data unusable if received by unauthorized parties.

Smart contracts: self-execute programs

Blockchain: a type of digital ledger in which info is recorded sequentially within blocks, that are then linked or “chained” together, and secured using cryptographic methods.

Permissioned and Permission-less Networks

DLTG Applications:

1. Cryptocurrencies: ICO
2. Tokenization: representing ownership rights to physical assets on a blockchain or DLT. A single, digital record of ownership with which to verify ownership title and authenticity, including all historical activity.
3. Post-trade clearing and settlement: confirm, clear and settle transactions among counterparties and intermediaries.
4. Compliance: shared information, communications, transparency within and between firms, exchanges, custodians and regulators. Security and privacy in closed or permissioned networks.

### Reading 7 Correlation and Regression

Sample covariance：协方差 | Sample correlation coefficient：相关系数

Limits to correlation analysis

Formulate a hypothesis test – put correlation coefficient to zero and reject at a give level of significance to prove they are related

Dependent and independent variables in a linear regression

Assumptions under linear regression and explain regression coefficient

Calc and interpret standard error estimate | coefficient of determination | a confidence interval for a regression coefficient

A null and alternative hypothesis for a regression coefficient, proper test stats, reject null hypothesis at a given significance level

Calc predicated value for the dependent variable, given an estimated regression model and value of independent variable.

A confidence interval for the dependent variable predicated value.

ANOVA (analysis of variance) in regression analysis, interpret ANOVA results, calc and interpret F-stats

Limits of regression analysis

#### Correlation analysis

Measure how two variables vary in relation to each other.

##### Scatter plots

##### Correlation analysis

Correlation coefficient vary between 1 and -1. 1和-1代表完全在一条直线上而没有散发，0代表完全不在一条直线。

##### Calc and interpret correlation coefficient

计算相关系数要用到协方差和样本方差 – 两个变量分别和他们样本均值差的乘积的平均值（除以n-1而不是n）。

C:\Users\Li\AppData\Local\Temp\ksohtml16636\wps1.jpg

n: sample size

C:\Users\Li\AppData\Local\Temp\ksohtml16636\wps2.jpg: ith observation on variable X

C:\Users\Li\AppData\Local\Temp\ksohtml16636\wps3.jpg: mean of the variable X observations

C:\Users\Li\AppData\Local\Temp\ksohtml16636\wps4.jpg: ith observation on variable Y

C:\Users\Li\AppData\Local\Temp\ksohtml16636\wps5.jpg: mean of the variable Y observations

**协方差表示的是两个变量****的总体[误差](https://baike.baidu.com/item/%E8%AF%AF%E5%B7%AE/738024)**，这与只表示一个变量误差的[方差](https://baike.baidu.com/item/%E6%96%B9%E5%B7%AE/3108412)不同。 如果两个[变量](https://baike.baidu.com/item/%E5%8F%98%E9%87%8F/5271)的变化趋势一致，也就是说如果其中一个大于自身的期望值，另外一个也大于自身的期望值，那么两个变量之间的协方差就是正值。 如果两个变量的变化趋势相反，即其中一个大于自身的期望值，另外一个却小于自身的期望值，那么两个变量之间的协方差就是负值。

A measure of the strength of the linear relationship (correlation) between two variables.

C:\Users\Li\AppData\Local\Temp\ksohtml16636\wps6.jpg

C:\Users\Li\AppData\Local\Temp\ksohtml16636\wps7.jpg

##### Limits of correlation analysis

Not reliable in case of nonlinear relation

Impact of outlier

Does not imply causation（因果论）

Spurious correlation caused by:

1. A particular dataset
2. A calculation mixed each of two variables with a third
3. Arising not from a direct relation but from their relation to a third variable

##### Uses of correlation analysis

Evaluating forecast

Style analysis

Correlations are always symmetrical

##### Testing correlation coefficient significance

用t分布做双尾检验，如果显著性水平上落在区间外就拒绝。显著性水平α要除以2得到t表查表概率p。

检验估计的相关系数的显著性。

Test whether the correlation between the populations of two variables is equal to zero. (If it is zero, it means there is no correlation, otherwise there is.) Use the sample correlation to compute the test statistic.

Two tailed test:双尾检验

C:\Users\lz\AppData\Local\Temp\enhtmlclip\Image.png

假设两个样本集合都是正态分布，使用t检验来决定是否拒绝空假设（即支持两个样本之间有关系）。T的值用样本相关系数r来计算。T检验用在样本数小于30，总体方差未知的情况。如果样本数大于30，且方差已知，则使用Z检验。

Assuming the two populations are normally distributed, we can use a *t-*test to determine whether the null hypothesis should be rejected. The test statistic is computed using the sample correlation, *r*, with **n-2 degrees of freedom (*df*).**

C:\Users\lz\AppData\Local\Temp\enhtmlclip\Image(1).png

Compare the test statistic with the critical *t*-value for the appropriate degrees of freedom and level of significance.  The decision rule can be stated as:

Reject C:\Users\lz\AppData\Local\Temp\enhtmlclip\Image(2).pngif C:\Users\lz\AppData\Local\Temp\enhtmlclip\Image(3).png

Z检验用于大样本（大于30）平均值差异性检验

Z=\frac{\bar{X}-\mu_0}{\frac{S}{\sqrt{n}}}

**T检验用于小样本（小于30），总体方差未知的正态分布**

F检验用于两样本检验

#### Linear Regression

Linear regression assumes a linear relationship between the dependent and independent variables.

线性回归主要使用两类数据：cross sectional 和time series

Linear regression aka linear least squares, computes a line that best fits the observations. It chooses values for the intercept b0 and slope b1, that minimize the sum of the squared vertical distances between observations and regression lines.

=*i*th observation of the dependent variable, Y

=*i*th observation of the independent variable, X

=regression intercept term（回归截距）

=residual for the *i*th observation (also refer to as the disturbance term – 扰动项,or error term – 错误项)

回归过程就是从一个散点图中估计出一条直线的等式，能最好解释从观察到Y值对应到观察到的X值。

线性等式，或者叫最佳匹配线，或者叫回归线，等式如下：

=给定，的估计值

=估计得截距

=估计的斜率系数

帽子符号^代表预计的值

回归线是估计值和真实值之间差的平方的和最小的线。估计值和真实Y值（垂直距离）的差的平方和，也叫平方误差和（sum of squared errors – SSE ）。回归线就是让SSE最小的线。这也解释了为什么简单**线性回归经常被叫做最小二乘（ordinary least squares - OLS）回归**，而估计回归公式得到的值也叫作最小二乘估计。

**斜率系数等于，截距用平均值加上斜率来求解。**

##### Linear regression model assumptions

1. Relationship between dependent variable and independent variable is linear: if the dependent variable is nonlinear, linear regression can be used as long as the regression is linear in the parameters.
2. Independent variable is not random
3. Expected value of the error term is 0: ;
4. The variance of the error term is the same for all observations: . homoskedasticity
5. Error term ε is uncorrelated across observations. for all
6. Error term ε is normally distributed

##### Standard error of estimate

How well a given linear regression model captures the relationship between the dependent and independent variables.

SSE is sum of standard error

计算要考的！

##### The Coefficient of Determination

Two methods:

1. Simpler: only has 1 independent variable. **Simply use square of the correlation coefficient to show the determination percentage;**
2. General: multiple independent variables.

衡量因变量的总体变化程度，真实Y值和平均值之间的差的平方和。和方差不尽相同，**方差variance=SST/(n-1)；**

预测值和平均值之间的差的平方和, 衡量由独立变量所揭示的**因变量的变化程度**。

真实值和预测值之间的差的平方和。**因变量中未被解释的部分**。

Standard error of estimate SEE, is **the standard deviation of the regression residuals.**

Sample variance of the **dependent** variable = SST/ (n-1)

Degrees of freedom for the **regression**: the number of slope parameters in the regression.

RSS Regression of sum squares: the part of total sum of squares explained by the regression.

##### Hypothesis testing

The most common way to test a hypothesis using a regression model is with a t-test of significance.

Solution 1:

t检验可以被用来检验真是斜率系数是否等于某个假设值。 给定的点估计，n-2自由度下的检验统计量是：

Reject if

拒绝假设代表着斜率系数与假设值不一样。

有时候不会直接报是否拒绝关于回归参数的特定假设，他们会报告p-value。这个p-value是空假设能被拒绝的最小显著性水平。

If p-value is 0.005, we can reject the hypothesis at the true parameter is equal to 0 at 0.5 percent significance level (99.5 percent confidence).

Regression analysis: Examine the ability of one or more variables (independent variables) to explain or predict dependent variable.

Solution 2:

Any particular hypothesized value of β confidence interval is:

If the estimated value falls into this interval, you can’t reject the hypothesis of estimated value.

##### Analysis of variance in a regression with one independent variable.

ANOVA is a statistical procedure for dividing the total variability of a variable into components that can be attributed to different sources.

F-statistic tests whether all the slop coefficients in a linear regression are equal to 0.

The ratio of the average regression sum of squares to the average sum of the squared errors.

Regression sum of squares – RSS. Average RSS is RSS/ (number of slope parameters estimated).

Sum of the squared errors – SSE. Average SSE is SSE/ df.

ANOVA typically is not used in regressions with just one independent variable. In such regression, F-statistic is the square of the t-statistic for the slope efficient.

##### Prediction intervals

Two sources of uncertainty when using regression model to make a prediction.

* + - 1. The error term itself. The standard deviation of the error term can be estimated from the standard error of estimate.
      2. Uncertainty in the estimated parameters b0 and b1.
    1. The estimated variance of the prediction error, of Y given X, is:

Prediction interval:

##### Limitations of regression analysis

1. Parameter instability
2. Public knowledge of regression relationships may negate future usefulness
3. Violations of regression assumptions

T检验用来判断两个变量是否相关。

T检验用来判断是否相关系数等于一个预测的值。

The degrees of freedom for the regression is the number of slope parameters in the regression, which is the same as the number of independent variables in the regression.

The degrees of freedom are the number of observations minus the number of parameters estimated.

统计学上，**自由度**是指当以样本的[统计量](https://baike.baidu.com/item/%E7%BB%9F%E8%AE%A1%E9%87%8F)来估计[总体](https://baike.baidu.com/item/%E6%80%BB%E4%BD%93/19951273)的[参数](https://baike.baidu.com/item/%E5%8F%82%E6%95%B0)时，**样本中独立或能自由变化的数据的个数，称为该统计量的自由度**。一般来说，自由度等于独立变量减掉其衍生量数。

去掉残差项会降低模型的预测能力，导致偏差变大，解释能力减小。

### Reading 8 Multiple Regression and Machine Learning

**多重回归就是一个因变量多个独立变量。多重回归估计的假设是方差之和最小，标准差就是真实值和估计值之间的差。**

Image(13)

Image(14)

Image(15)

Image(16)

Image(17)

Image(18)

Image(19)

**n=number of observations**

**k=number of independent variables**

A slope coefficient bj, measures how much the dependent variable Y changes when the independent variable Xj changes by one unit, **holding all other independent variables constant.**

The lower the p-value reported for a test, the more significant the result.

For each regression coefficient, the p value would be the smallest level of significance at which we can reject a null hypothesis that the population value of the coefficient is 0 in a two-sided test. **The lower the p-value, the stronger the evidence against that null hypothesis.**

##### Assumptions of the multiple linear regression model

Image(13)

1. The relationship between the dependent variable, Y, and the independent variables, X1, X2,…,Xk, is linear as described in equation;
2. The independent variables (X1, X2,…,Xk) are not random. Also, no exact linear relation exists between two or more of the independent variables.
3. The expected value of the error term, conditioned on the independent variables, is 0: .
4. The variance of the error term is the same for all observations: 。
5. The error term is uncorrelated across observations:
6. The error term is normally distributed.

##### Predicting the dependent variable in a multiple regression model

2 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 parameters.

##### Testing whether all population regression coefficients equal zero

ANOVA with regard to a regression’s explanatory power.

Use F-test for null hypothesis. Used to test whether at least one of the slope coefficients on the independent variables is significantly different from 0.

One tail F-test

##### Adjusted

Increase by including many additional independent variables that explain even a slight amount of the previously unexplained variation, even if the amount they explain is not statistically significant.

K≥1的时候，肯定比大。

当k增加的时候（新增自变量，并且自变量仅仅引起很小的增长），会减小。

可能为负，虽然永远非负。

很大的时候，并不一定意味着回归包括了正确的变量集合。因为有可能它只是反映了特定回归估计的数据集合。

#### Using dummy variables in regressions

定性的变量，取值只是0或者1.

#### Violations of regression assumptions

##### Heteroskedasticity:异方差性

偏离回归模型的距离变化不一。

###### Consequences

不影响回归参数估计的一致性。

t-test和F-test都不可靠了，因为引入了回归系数标准差估计偏差。

导致standard error变小，t检验变大。

忽略异方差性会，会导致发现显著相关性但实际上并没有。

无条件异方差性：对统计推理没有主要影响

条件异方差性：误差的变化和自变量的值条件相关。

###### Testing

B-P检验用来检测是否存在异方差性。

空假设是无条件相关，（从原本的回归自变量中，计算regression of the squared residuals）是一个卡方随机变量，自由度等于回归中的自变量数。

###### Correcting for heteroskedasticity

Robust standard errors: account for the conditional heteroskedasticity

Generalized least squares: modifies original regression equation to eliminate the heteroskedasticity.

##### Serial correlation

Regression errors correlate across observations, they are serially correlated or automrelated.

###### Consequences

Incorrect estimate of the regression coefficient standard errors

As long as none of the independent variables is a lagged value of the dependent variable (a value of the dependent variable from a previous period), then the estimated parameters themselves will be consistent and need not be adjusted for the effects of serial correlation.

Positive serial correlation: 一个观察值上的positive error会增加另一个观察值上的positive error的机会。反之，negative error也会增加另一个观察值上的negative error的机会。

Standard errors变小，t检验估计变大。

###### Testing

Uses Durbin and Watson.

* DW的取值根据协方差COV的不同，在0和4之间变化。
* DW的值小于,拒绝不存在正序列相关的空假设；
* DW的值大于,接收不存在正序列相关的空假设；
* ,和之间则不确定

###### Correcting

1. Adjust the coefficient standard errors for the linear regression parameter estimates to account for the serial correlation.
2. Modify the regression equation itself to eliminate the serial correlation.

第一种方法更受青睐，因为第二种方法可能导致不一致的参数估计。

第一种方法里面有两种常见手段：

Hanse

Newsey and West.

##### Multicollinearity

完美共线性：One of the independent variables is an exact linear combination of other independent variables, it becomes mechanically impossible to estimate the regression.

和多重共线性比较起来不是一个特别大的问题。

Multicollinearity: when two or more independent variables (or combinations of independent variables) are highly (but not perfectly) correlated with each other.

###### Consequences

Standard error变大，t检验变小，has little power.

###### Detecting

A matter of degree rather than of absence or presence.

High and significant F-test, even though t-test on the estimated slope coefficients are not significant.

###### Correcting

Excluding one or more of the regression variables.

##### Summary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Group | Sub-group name | Consequences | Testing | Correcting |
| Violations of regression assumptions | Heteroskedasticity | On average, the regression residuals grow much larger as the size of the independent variable increases.  F-test and t-test are unreliable due to bias introduced into estimator of the standard error of regression coefficient.  **Estimated standard errors will be underestimated and t-stat is inflated.**  When ignore heteroskedasticity, **find significant relationships where none exists.**  Conditional heteroskedasticity in the error variance that is correlated with the values of the independent variables in the regression.  不影响consistency | Breusch-Pagan.  (from the regression of the squared residuals on the independent variables from the original regression) will be a random variable with the number of degrees of random equal to the number of independent variables in the square residuals..  It breaks the assumption that the variance of error in a regression is constant across observations. 。 | Robust standard errors: corrects the standard errors of the linear regression model’s estimated coefficients to account for the conditional heteroskedasticity.  Generalized least squares: modifies the original equation in an attempt to eliminate the heteroskedasticity. |
| Serial Correlation | Incorrect estimate of the regression coefficient standard errors.  If none of the independent variables is a lagged value of the dependent variable, no adjustment required. Otherwise serial correlation in the error term will cause all parameter estimates inconsistent.  Usually lead to artificially small standard errors for the regression coefficient, and then inflated t-stat, then incorrectly reject null hypothesis.  会变得inconsistent. | Durbin and Watson test.    r is the sample correlation between the regression residuals from one period and those from the previous period.  It breaks the assumption that regression errors are uncorrelated across observations. | Adjust the coefficient standard errors for the linear regression parameter estimates to account for serial correlation. Hansen , Newey and West, 这些方法同时对同向异方差性正确。  Modify regression equation to eliminate serial correlation. |
| Multicollinearity | 完美的共线性反而不是一个问题，因为总是可以通过合并什么处理掉。  **Inflated OLS standard errors for the regression coefficients. T-tests on the coefficient have little power (to reject the null hypothesis).**  **不影响OLS regression coefficient的consistency** | A matter of degree rather than yes or no.  A high and a significant F-stat, but t-stat are not significant.  It breaks the assumption that no exact linear relationship exists between two or more independent variables. | Excluding one or more colinear variables. |
| Misspecified Functional Form | Omitted variable(s) | Estimated values of regression coefficients will be biased and inconsistent.  Estimates of the standard errors of those coefficients will be inconsistent. |  |  |
| Variable(s) may need to be transformed |  |  |  |
| Pools data from different samples that should not be pooled |  |  |  |
| Time-series misspecification (independent variables correlated with errors) |  | Included lagged dependent variables as independent variables.  Included a function of a dependent variable as an independent variable  Independent variables are measured with error. |  |

#### Model specification and errors in specification

##### Model specification principles

1. The model should be grounded in cogent economic reasoning
2. The functional form chosen for the variables in the regression should be appropriate given the nature of the variables
3. The model should be parsimonious. Each variable included in a model plays an essential role.
4. The model should be examined for violations of regression assumptions before being accepted.
5. The model should be tested and be found useful out of sample before being accepted.

##### Misspecified functional form

1. One or more important variables could be omitted from regression.
2. One or more of the regression variables may needed to be transformed.
3. The regression model pools data that should not be pooled.

##### Time-series misspecification (independent variables correlated with errors)

#### Machine Learning

1. Including lagged dependent variables as independent variables in regression with serially correlated errors;
2. Including a function of a dependent variable as an independent variable, some-times as a result of the incorrect dating of variables;
3. Independent variables that are measured with error.

##### Other types of time-series misspecification

Most frequent source of misspecification in linear regressions, use time series from two or more different variables is non-stationarity.