FIN4104/4911
Quantitative Analysis for Financial Decisions

Chapter 6: Correlation and Regression

MSME BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

Our Schedule

Week	Date	Subject
1	07/06/68	Course Introduction
2	14/06/68	TVOM
3	21/06/68	Statistical Concepts & Probability Concepts
4	28/06/68	Sampling and Estimation
5	05/07/68	Hypothesis Testing
6	12/07/68	No Class (Long Holiday)
7	19/07/68	Correlation Analysis and Regression
8	26/07/68	No Class (Long Holiday)
	02/08/68	Mid-Term
9	09/08/68	No Class (Long Holiday)
10	16/08/68	Multiple Linear Regression Analysis
11	23/08/68	Time-Series Analysis
12	30/08/68	Modern Quantitative Finance
13	06/09/68	Technical Analysis + Data Visualization and Presentation
14	13/09/68	Group Presentation
15	20/09/68	Programme close and Revision
16	27/09/68	End
		Final Exam Week

Mid-term Exam

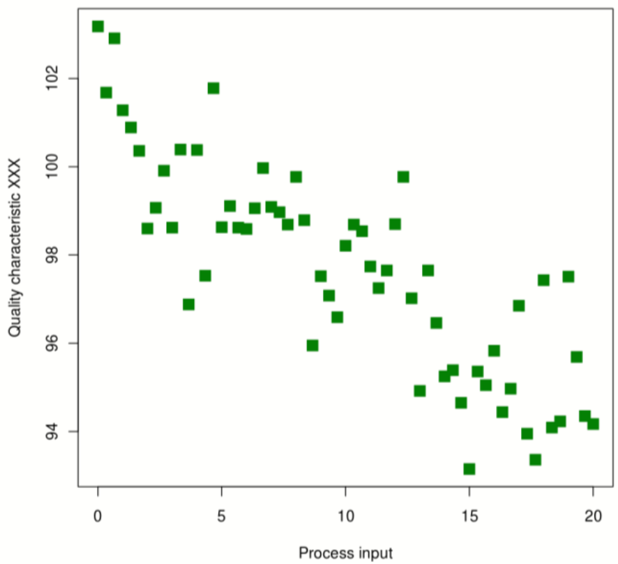
- There are 12 questions for mid-term exam.
  - 10 Short answers
  - Essay Questions: Select **2 questions out of 3 questions** to answer
- Total raw score is 25 scores.
- Scope of exam as following:
  - TVOM, Discounted cash flow
  - Statistical Concepts
  - Probability Concepts
  - Common Probability Distribution
  - Sampling and Estimation
  - Hypothesis Testing

Course Outline

- Correlation
- Regression
- Classification

Correlation

- The Graph that shows the relationship between the observation of two series of data.



- Correlation coefficient is a measure of how closely related two data series are.
- Correlation coefficient measures the direction and extent of linear association.

Correlation

- Correlation can be used in measurement of level of risk diversification.

- When we invest in two stocks without perfectly positive correlation, the risk of combined portfolio will be lower than average risk of individual stock

Correlation

Testing the significance of a correlation coefficient

- We want to test if a non-zero correlation between two variables is the result of chance
- In this test we assume that **both variables are distributed normally** and test whether or not the correlation is significantly different from zero
- The test statistic here is calculated using the following formula, the t tables and n-2 degrees of freedom:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

Where r is the calculated correlation coefficient from the samples

Formula

As n increases we are more likely to reject a false NULL:

1. Degrees of freedom increases and critical statistic falls
2. Numerator increases and test statistic rises

Correlation

Example: Testing the significance of a correlation coefficient

- Sample of 82 observations and a correlation coefficient (r) of 0.7. Test whether the correlation coefficient is significant at a 5% significance level.
  - H_0: r = 0
  - H_a: r ≠ 0 (Note: This is a 2-sided test)
- Test statistic is:

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.7\sqrt{82-2}}{\sqrt{1-0.49}} = \frac{6.26}{0.71} = 8.82$$

- Critical statistic is:
  - Using Student-t tables with degrees of freedom = n - 2 = 82 - 2 = 80 and p = 0.025 we get 1.99

- Given that: **Test statistic > Critical statistic** we would reject H0
- The correlation coefficient is statistically significant

MSME BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

Other Issues: Nonparametric Inference

Parametric tests

- Tests of parameters or tests that make assumptions about the distribution of the population
- E.g., z-test, t-test, chi-square test, or F-test

Non-parametric tests are used in three situations when:

Data does not meet distributional assumptions

- It is not normally distributed
- May want to test hypothesis concerning a population mean when we are dealing with a small sample from a non-normal population

Data is given in ranks

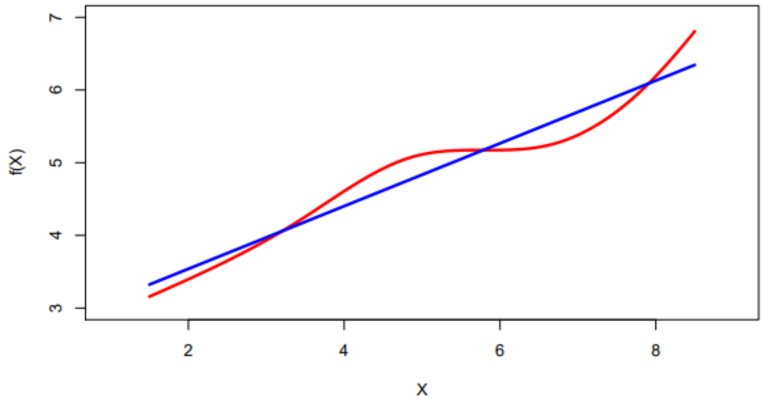
- E.g. ranking of investment managers

Characteristics being tested is not a population parameter

- E.g. testing whether a sample is random or not

Linear Regression

- Linear regression is a simple approach to supervised learning. It assumes that the dependence of Y on $X_1, X_2, \ldots X_p$ is linear.
- True regression functions are never linear!



- although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.

MSME BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

Linear Regression

- We assume a model

$$Y = \beta_0 + \beta_1 X + \epsilon,$$

where β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*, also known as *coefficients* or *parameters*, and ϵ is the error term.

- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ for the model coefficients, we predict future sales using

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. The *hat* symbol denotes an estimated value.

Assumptions of Linear Regression

- The relationship between dependent variable and independent variable is linear.
- The independent variable is not random
- The expected value of the error term is 0
- The variance of the error term is the same for all observation
- The error term is uncorrelated
- The error term is normally distributed

Estimation of the parameters

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i = y_i - \hat{y}_i$ represents the ith *residual*
- We define the *residual sum of squares* (RSS) as

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

or equivalently as

$$\mathrm{RSS} = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \ldots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2.$$

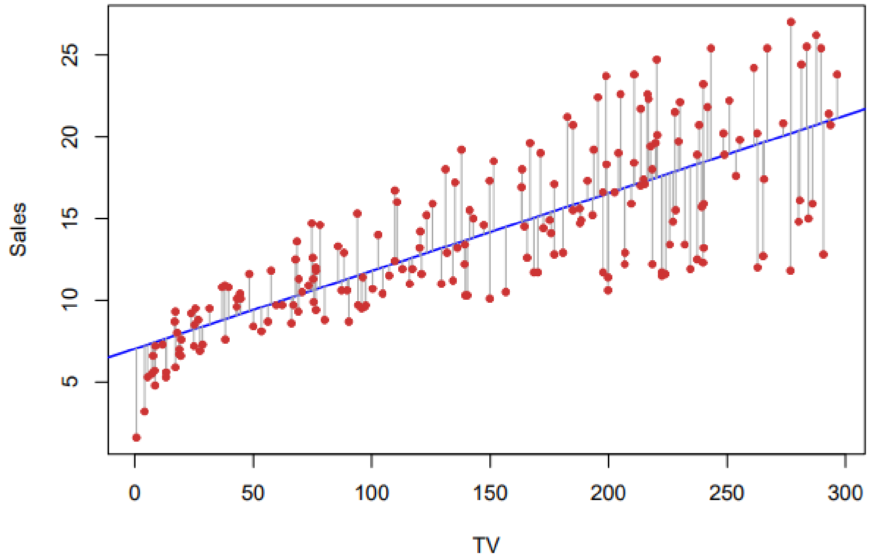
- The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

where $\bar{y} \equiv \frac{1}{n} \sum_{i=1}^n y_i$ and $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$ are the sample means.

Example



The least squares fit for the regression of `sales` onto `TV`.
In this case a linear fit captures the essence of the relationship, although it is somewhat deficient in the left of the plot.

Assessing the accuracy of coefficient estimates

- The standard error of an estimator reflects how it varies under repeated sampling. We have

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \mathrm{SE}(\hat{\beta}_0)^2 = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right]$$

where $\sigma^2 = \mathrm{Var}(\epsilon)$

- These standard errors can be used to compute *confidence intervals*. A 95% confidence interval is defined as a range of values such that with 95% probability, the range will contain the true unknown value of the parameter. It has the form

$$\hat{\beta}_1 \pm 2 \cdot \mathrm{SE}(\hat{\beta}_1).$$

Confidence Intervals

That is, there is approximately a 95% chance that the interval

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

will contain the true value of β_1 (under a scenario where we got repeated samples like the present sample)

Hypothesis Testing

- Standard errors can also be used to perform *hypothesis tests* on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of

  H_0 : There is no relationship between X and Y

  versus the *alternative hypothesis*

  H_A : There is some relationship between X and Y.

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

Hypothesis Testing

- To test the null hypothesis, we compute a *t-statistic*, given by

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

- This will have a t-distribution with $n - 2$ degrees of freedom, assuming $\beta_1 = 0$.
- Using statistical software, it is easy to compute the probability of observing any value equal to $|t|$ or larger. We call this probability the *p-value*.

Results

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Assessing overall accuracy

- We compute the *Residual Standard Error*

$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2},$$

  where the *residual sum-of-squares* is $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.
- *R-squared* or fraction of variance explained is

$$R^2 = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}} = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

  where $\mathrm{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2$ is the *total sum of squares*.
- It can be shown that in this simple linear regression setting that $R^2 = r^2$, where r is the correlation between X and Y:

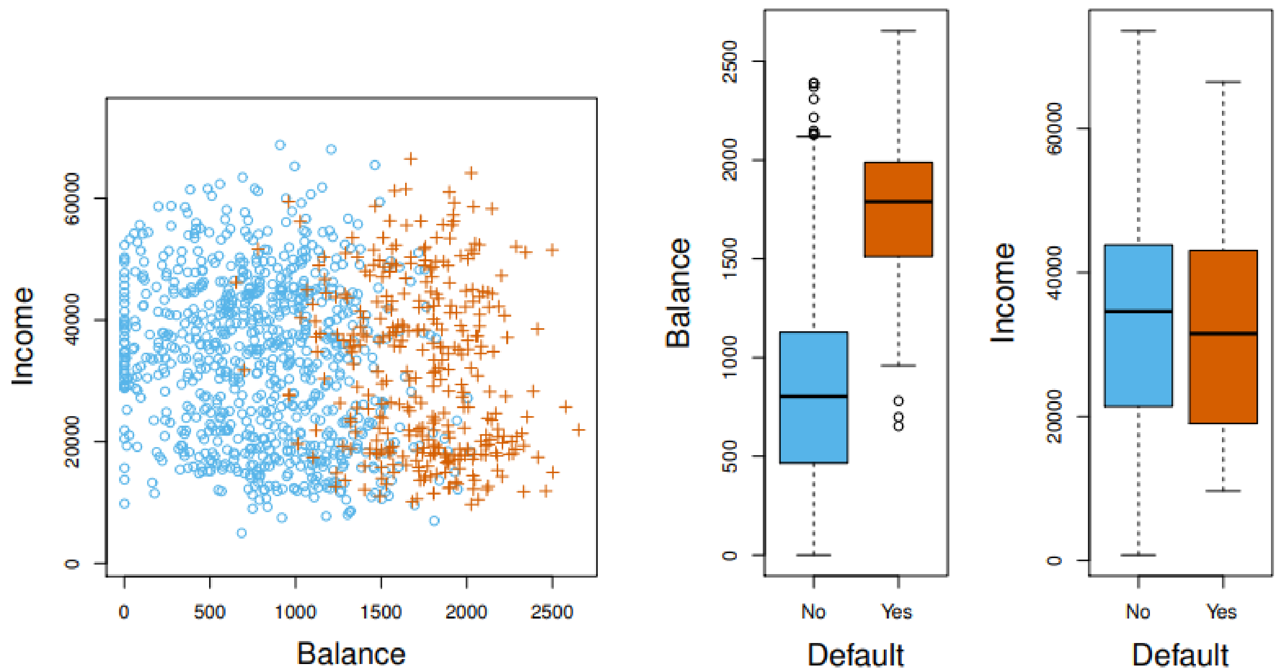
$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}.$$

Quantity	Value
Residual Standard Error	3.26
R^2	0.612
F-statistic	312.1

Classification

- Qualitative variables take values in an unordered set $\mathcal{C}$, such as:
  $\texttt{eye color} \in \{\texttt{brown}, \texttt{blue}, \texttt{green}\}$
  $\texttt{email} \in \{\texttt{spam}, \texttt{ham}\}$.
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.
- Often we are more interested in estimating the *probabilities* that X belongs to each category in $\mathcal{C}$.

Example: Credit Card Default



MSME BUSINESS SCHOOL
ASSUMPTION UNIVERSITY

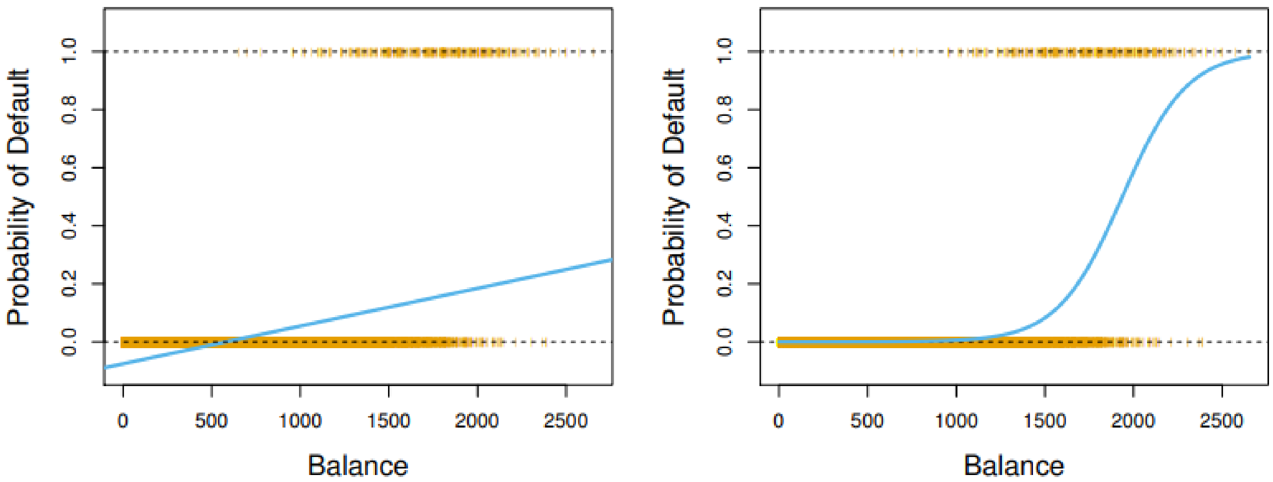
Can we use Linear Regression?

Suppose for the `Default` classification task that we code

$$Y = \begin{cases} 0 & \text{if } \texttt{No} \\ 1 & \text{if } \texttt{Yes}. \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as `Yes` if $\hat{Y} > 0.5$?

Linear VS Logistic Regression



The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

Logistic Regression

Let's write $p(X) = \Pr(Y = 1|X)$ for short and consider using `balance` to predict `default`. Logistic regression uses the form

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

($e \approx 2.71828$ is a mathematical constant [Euler's number.])
It is easy to see that no matter what values β_0, β_1 or X take, $p(X)$ will have values between 0 and 1.

Maximum Likelihood

We use maximum likelihood to estimate the parameters.

$$\ell(\beta_0, \beta) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

This *likelihood* gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximize the likelihood of the observed data.

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.6513	0.3612	-29.5	< 0.0001
`balance`	0.0055	0.0002	24.9	< 0.0001

Making Predictions

What is our estimated probability of `default` for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

Making Predictions

What is our estimated probability of default for someone with a balance of \$1000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}} = 0.006$$

With a balance of \$2000?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}} = 0.586$$

Measure the accuracy

- The results from logit model range from 0-1 (0%-100%).
- However, the results we actually need should be exactly 1 or 0.
- We need to set the cut-off.
- For example at cut-off of 0.5
  – Any result less than 0.5 will be treated as 0
  – Any result more than or equal to 0.5 will be treated as 1