Intro to credit risk modelling/management

African Institute for Mathematical Sciences

Limbe - Cameroon - Dec 2024

Viani Djeundje Biatat: University of Edinburgh
Marc Gaudart & Paul Randall: Trend Advisory Services

Intro to credit risk modelling/management

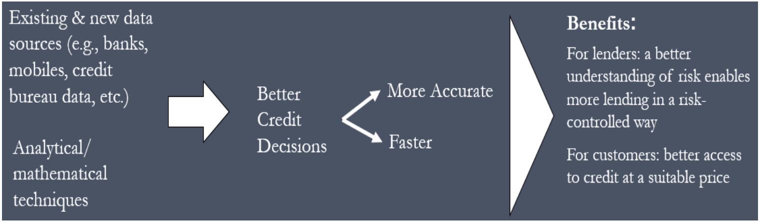
Credit risk products play a major role in all economies.

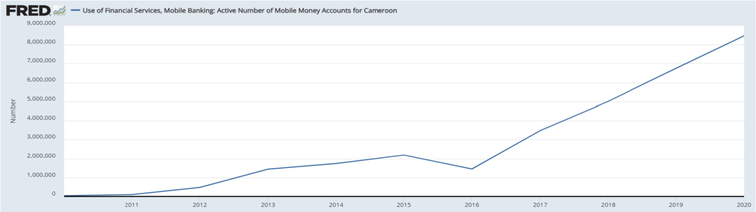
- Economic growth: Credit products provide individuals and small businesses with access to capital, enabling them to invest in growth opportunities, start businesses, and expand operations, driving economic development.

- Poverty reduction: Access to credit allows low-income individuals to cover essential expenses, such as healthcare, education, and housing, helping to reduce poverty and improve living standards.

⚠ BUT ..... there is risk !!!

Intro to credit risk modelling/management





- ▶ Data abound
- Those institutions that introduce new processes and methods will be the ones that win the future expansion of the credit market.
- We are in the early days of digital transformation in the lending business (especially in central Africa).

Intro to credit risk modelling/management

CHANGE IS STARTING; IMPROVEMENT IN DATA AVAILABILITY THROUGH IMPLEMENTATION OF A PRIVATE CREDIT BUREAU

Atelier de sensibilisation des établissements de crédit et de microfinance sur les Bureaux d'Information sur le Crédit (BIC), 2024

Source: La Banque Centrale des Etats de l'Afrique Centrale (BEAC) et IFC – International Finance Corporation organise une atelier de sensibilisation des établissements de crédit et de microfinance sur les Bureaux d'Information sur le Crédit (BIC) - Creditinfo West Africa

Intro to credit risk modelling/management

BUT ..... there is risk!

Objective: This course looks into

(i) The analytical methodologies required to build effective credit scoring system for better credit/loan decisions, and

(ii) The key steps for effective implementation and management to ensure adequate controls.

Intro to credit risk modelling/management: Outline

- Part 1: Intro to credit risk modelling
  - Correlation and regression
  - Generalised linear models
  - Hidden patterns (semiparametric)
  - Boosting methods & machine learning?
  - Dynamic credit scoring (survival models)
- Part 2: Intro to credit risk management
  - Benefits of digital transformation of credit lending
  - Theory of scorecards and the scorecard development process
  - Implementation and achieving business improvement
  - Operational structure to maximise the benefits of scorecards
  - Continual improvement, control and management.

Part 1: Intro to credit risk modelling

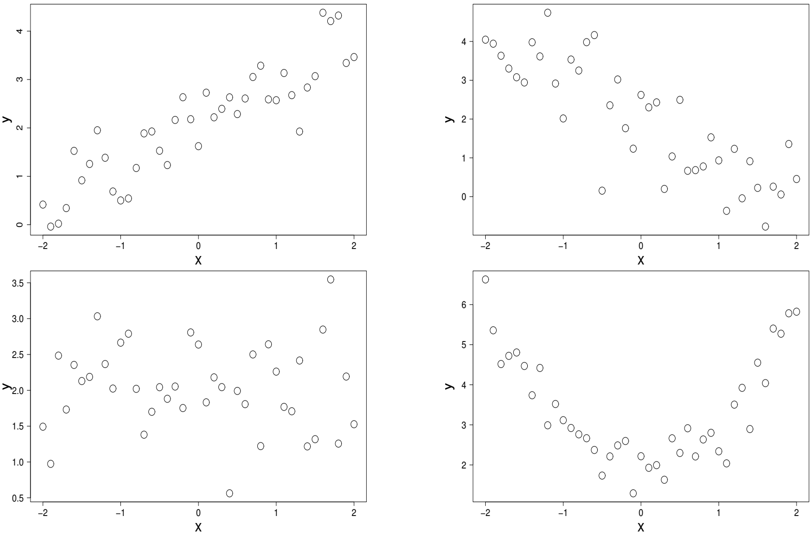
- Objective: This first part of course covers the analytical (statistical) methodologies required to build effective credit scoring systems.

- Data abound:

  Summarize data, explore potential associations between data items, make predictions.
  - Descriptive statistics
  - Regressions
  - Pattern detection
  -
  -
  -
- Tool: **R**

$\hookrightarrow$ Assessment: Weekly quiz + Assignment

Correlation and regression



Can we measure correlation?

- Let X and Y be 2 random variables with probability (density/mass) functions f_X and f_Y, and let g be a function.
- The expected value and variance of $g(X)$ are given by:

$$\mathbb{E}[g(X)] = \begin{cases} \int g(x) f_X(x)\, dx & \text{if } X \text{ is continuous} \\ \sum_x g(x) f_X(x) & \text{if } X \text{ is discrete} \end{cases}$$

$$Var(g(X)) = \mathbb{E}\left[(g(X) - \mathbb{E}[g(X)])^2\right]$$

- Properties:
  - $\mathbb{E}[a\,X + b] = a\,\mathbb{E}[X] + b$
  - $Var[a\,X + b] = a^2\,Var[X]$
  - $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
  - $Var[X + Y] = Var[X] + Var[Y] + 2\,Covar(X, Y)$.
  - $\mathbb{E}[X\,Y] = \mathbb{E}[X] \times \mathbb{E}[Y]$ if X and Y are independent.
  - $Var(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$
- $Covar(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]$

Can we measure correlation?

- Consider 2 random variables X and Y.

  A measures of the strength of their linear relationship is:

$$\rho(X, Y) \quad = \quad \frac{Covar(X,Y)}{\sqrt{var(X) \times Var(Y)}}$$

  This is known as *Population correlation coefficient.*

$- - - - -$

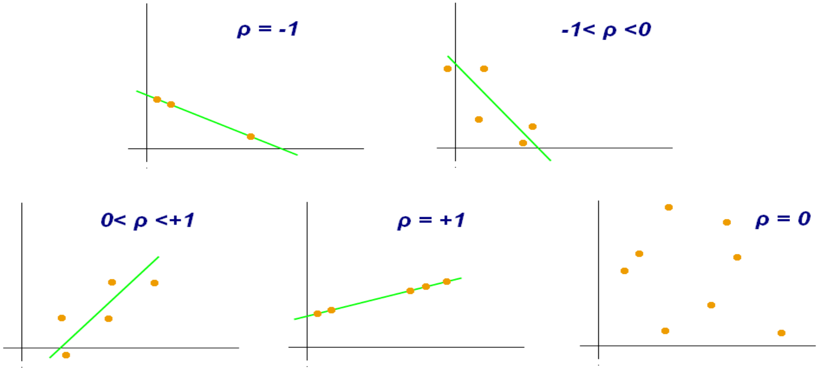
- We observe paired data points $(x_1, y_1),\ (x_2, y_2), ..., (x_n, y_n)$.

  Sample $Var(X) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})^2}{n-1}$

  Sample $Covar(X, Y) = \sum_{i=1}^{n} \frac{(x_i - \overline{x})(y_i - \overline{y})}{n-1}$

  $\Rightarrow$ *Sample correlation coefficient.*

Can we measure correlation?

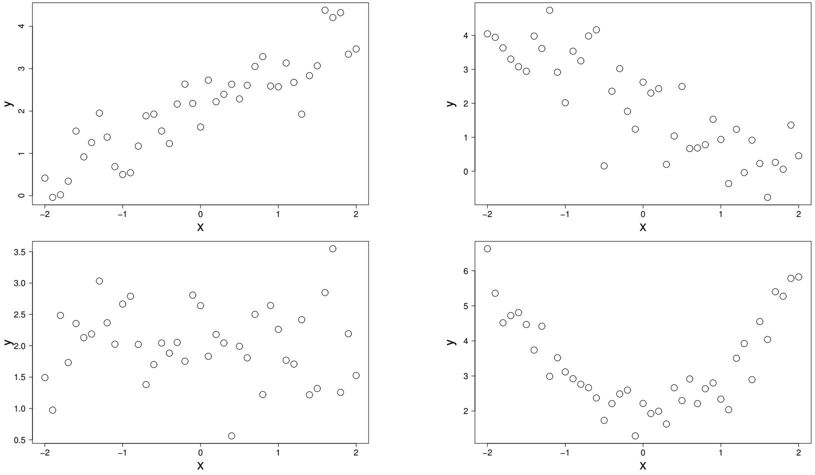


- Correlation coefficient takes values between -1 and $+1$.
- $+1$: Perfect positive (direct) relationship.
- -1: Perfect negative (inverse) linear relationship.
- 0: No linear relationship.

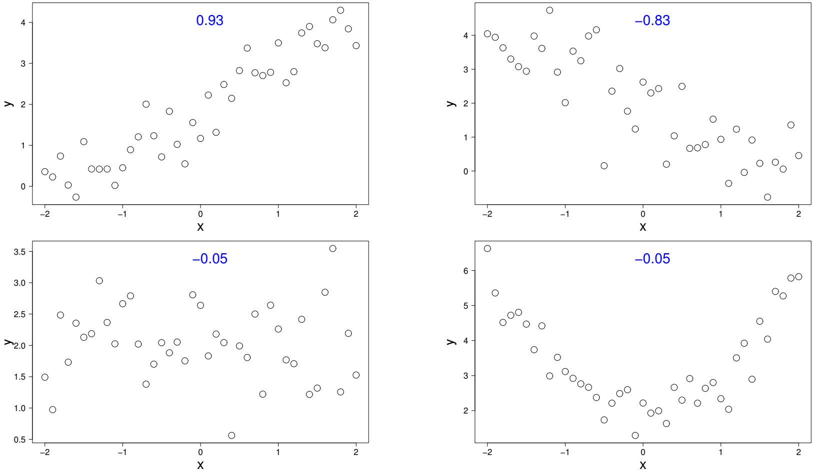
Can we measure correlation?

```
data_simulated <- read.csv( ".../data_simulated.csv" )

colnames(data_simulated)
x <- data_simulated$x
y <- data_simulated$y1

n      <- length(x)
x_bar  <- mean(x)
y_bar  <- mean(y)


#
var_x      <- sum( (x-x_bar)^2 ) / (n-1)
var_y      <- sum( (y-y_bar)^2 ) / (n-1)
covar_x_y  <- sum( (x-x_bar)*(y-y_bar) ) / (n-1)
cor_x_y    <- covar_x_y / ( var_x*var_y  )^0.5
```

Can we measure correlation?



Can we measure correlation?



- Warning: Correlation does not mean causation. Two variables can be highly correlated, but that does not mean that changes in one variable causes the other variable to change. Quite possible that third variable is actually the cause of changes in the first two.

Lab 1.1 (Correlations)

a) Load the dataset `data_simulated.csv` into R.

b) Calculate the correlation coefficients between each pair of columns in the dataset.

c) Comment on the magnitude of the association in each case.

Do not use `cor()` function.

Correlations: Limitations

- Correlation coefficient quantifies magnitude of linear associations.
- What about non-linear associations?
- What about categorical variables?
- What about 3+ variables?
- What about predictions?
-
-

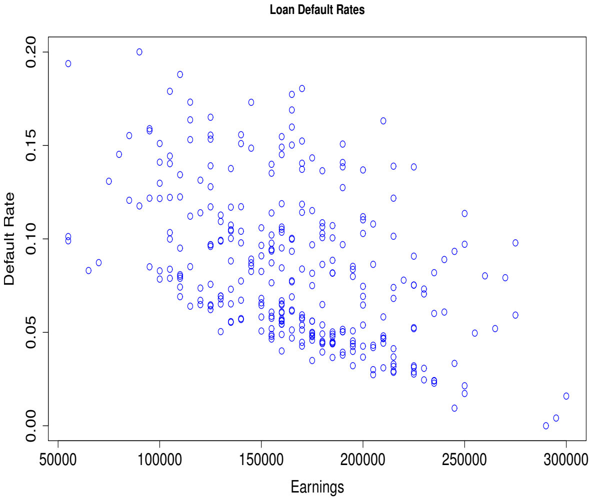
$\hookrightarrow$ Regression allow to overcome these limitations and to do more.

Simple Linear Regression

Aggregated loan counts by earnings:

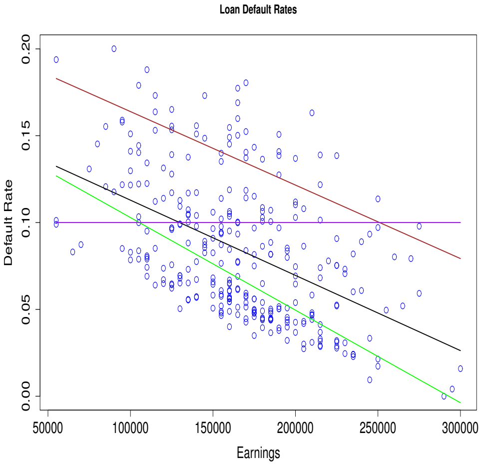
Accounts	Defaults	Earnings
236	46	55000
58	6	55000
81	8	55000
358	30	65000
215	19	70000
398	52	75000
120	17	80000
216	26	85000
346	54	85000
128	26	90000
.		
.		
.		

Simple Linear Regression



- Objective: summarise data by a straight line.

Simple Linear Regression: Formulation



- $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$

  $\mathbb{E}(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$
- y is the response or dependent variable (Default Rates).
- x is the covariate or independent variable (Earnings).
- β_1 is the slope of the line

  β_0 is the intercept
- ε_i is the noise/error

- Task: estimate the parameters β_0, β_1 and σ using data.

Simple Linear Regression: Estimation

- Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\mathbb{E}(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$
- Individual errors: $\varepsilon_i = y_i - (\beta_0 + \beta_1 x_i)$
- Sum of squared errors: $\texttt{SSE} = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$
- Find $(\hat{\beta}_0, \hat{\beta}_1)$ the value of (β_0, β_1) that minimises $\texttt{SSE}$.

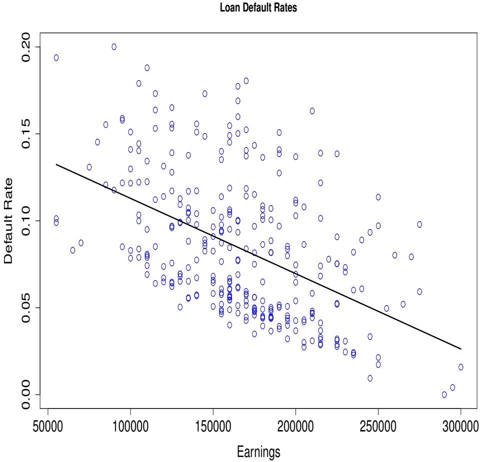
  i.e. solve $\dfrac{\partial \texttt{SSE}}{\partial \beta_0} = 0 = \dfrac{\partial \texttt{SSE}}{\partial \beta_1}, \cdots$
- $\hat{\beta}_1 = \dfrac{\sum_i (x_i - \bar{x})\, y_i}{\sum_i (x_i - \bar{x})^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- For a given value x_{new} of x, the predicted value $\hat{y}_{new}$ of y is

  obtained as $\hat{y}_{new} = \hat{\beta}_0 + \hat{\beta}_1 x_{new}$

Simple Linear Regression: Estimates

- $\hat{\beta}_0 = 0.156$

  $\hat{\beta}_1 = -4.33e \times 10^{-7}$
- Predictions: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$
- e.g. Predicted probability of default for a new borrower earning 286,510: $\hat{\beta}_0 + \hat{\beta}_1 \times 286510 = 0.032$



Simple Linear Regression: Relation with correlation coef

$$\rho(X,Y) = \frac{Covar(X,Y)}{\sqrt{Var(X)} \times \sqrt{Var(Y)}}$$

$$= \hat{\beta}_1 \times \frac{\sqrt{Var(X)}}{\sqrt{Var(Y)}}$$

Simple Linear Regression: Uncertainty

- Model: $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\mathbb{E}(\varepsilon_i) = 0$ and $Var(\varepsilon_i) = \sigma^2$

$$
\begin{aligned}
Var(\hat{\beta}_1) &= Var\left(\frac{\sum_i (x_i - \bar{x})\, y_i}{\sum_i (x_i - \bar{x})^2}\right) \\
&= Var\left(\frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \varepsilon_i)}{\sum_i (x_i - \bar{x})^2}\right) \\
&= Var\left(\frac{\sum_i (x_i - \bar{x})\, \varepsilon_i}{\sum_i (x_i - \bar{x})^2}\right) \\
&= \frac{\sum_i (x_i - \bar{x})^2\, Var(\varepsilon_i)}{\left(\sum_i (x_i - \bar{x})^2\right)^2}; \quad Var(a\, Z) = a^2\, Var(Z) \\
&= \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2}
\end{aligned}
$$

Simple Linear Regression: Uncertainty

- $Var(\hat{\beta}_1) = \dfrac{\sigma^2}{\sum_i (x_i - \overline{x})^2}$
- $Var(\hat{\beta}_0) = \dfrac{\sum_i x_i^2}{n \sum_i (x_i - \overline{x})^2} \sigma^2$
- $Covar(\hat{\beta}_0, \hat{\beta}_1) = \dfrac{-\overline{x}}{\sum_i (x_i - \overline{x})^2} \sigma^2$

$$
\begin{aligned}
Var(\hat{y}_{new}) &= Var(\hat{\beta}_0 + \hat{\beta}_1 x_{new}) \\
&= Var(\hat{\beta}_0) + x_{new}^2 Var(\hat{\beta}_1) + 2 x_{new} Covar(\hat{\beta}_0, \hat{\beta}_1)
\end{aligned}
$$

- Note: The value of σ is still unknown!

Simple Linear Regression: Uncertainty

- We want to be able to decide whether the true value of β_1 is *significantly* different from zero.
- Hypothesis Testing for the slope.
  $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$
- H_0 is the null hypothesis, and H_1 is the alternative.
- Central Limit Theorem $\Rightarrow Z := \dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}} \sim \mathcal{N}(0, 1)$
  Z is refereed to as the test statistics.
- Let α be a small positive number, $(\alpha = 0.05)$,
  z_α denotes the $(1 - \alpha)$ quantile of $\mathcal{N}(0, 1)$, and
  $\mathring{z}$ the value of the *test statistics* Z calculated under H_0.
- $|\mathring{z}| > z_{\frac{\alpha}{2}} \Rightarrow$ strong evidence against the null hypothesis at significance level α.
  Otherwise there is not enough evidence against H_0.
- Smaller $\alpha \Rightarrow$ more rigorous/selective test.

Simple Linear Regression: Uncertainty

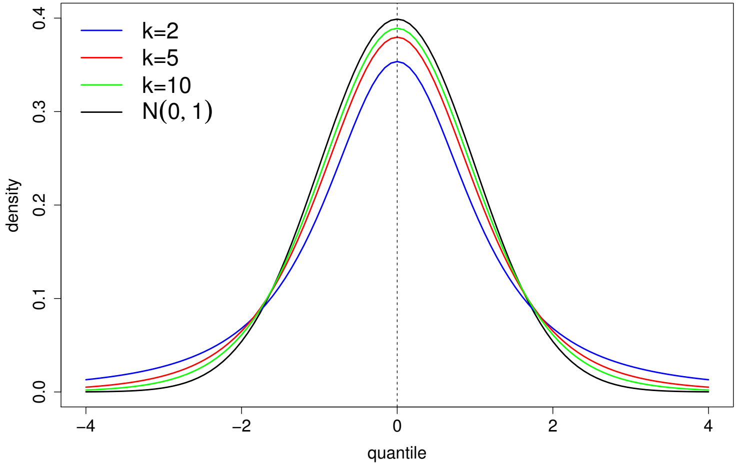
- Note that $|\mathring{z}| > z_{\frac{\alpha}{2}} \Leftrightarrow p := Pr\{|Z| > \mathring{z}\} < \alpha$.
- Definition: p is the *p-value* of the test. It is interpreted as the probability, under the null hypothesis, of obtaining a result at least as extreme as what was actually observed.
- The $(1-\alpha)$ confidence interval for β_1 is
  C.I. $= \left[\hat{\beta}_1 - z_{\frac{\alpha}{2}} \times \sqrt{Var(\hat{\beta}_1)}, \quad \hat{\beta}_1 + z_{\frac{\alpha}{2}} \times \sqrt{Var(\hat{\beta}_1)}\right]$
- Interpretation: The confidence interval can be expressed in terms of samples (or repeated samples). Indeed, "Were this procedure to be repeated on numerous samples, the fraction of calculated confidence intervals (which would differ for each sample) that encompass the true population parameter would tend toward $(1-\alpha)$.
- Warning: A 95% confidence interval does not mean that for a given realized interval there is a 95% probability that the population parameter lies within the interval.

Simple Linear Regression: Uncertainty

- Note: The two previous slides assume that the true value of σ is know. In practice, it is unknown and must be estimated.
- $\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{n-2}$

  This has some implications on the formulas.
- $Var(\hat{\beta}_1) = \frac{\sigma^2}{\sum_i (x_i - \bar{x})^2} \Rightarrow \widehat{Var(\hat{\beta}_1)} = \frac{\hat{\sigma}^2}{\sum_i (x_i - \bar{x})^2}$
- $Z := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{Var(\hat{\beta}_1)}} \sim \mathcal{N}(0,1)$ becomes $T := \frac{\hat{\beta}_1 - \beta_1}{\sqrt{\widehat{Var(\hat{\beta}_1)}}} \sim t_{n-2}$
- t_{n-2} is the t-distribution with $n-2$ degrees of freedom.
- $f_\nu(x) = \frac{\frac{\Gamma(\nu+1)}{2}}{\sqrt{\pi\nu}\,\Gamma(\frac{\nu}{2})} \left(1 + \frac{x^2}{\nu}\right)^{-\frac{\nu+1}{2}}$

Simple Linear Regression: Uncertainty



- t_k tends toward $\mathcal{N}(0,1)$ as k increases.
- Hypothesis testing and confidence intervals as before but using t-distribution instead of $\mathcal{N}(0,1)$.

Lab 1.2 (Simple Linear Regression)

The dataset `DataDefaults.csv` contains information on loan default counts by group from a loan provider. Accounts, Defaults and Earnings are the number of accounts, number of accounts that defaulted, and average earning/salary for the group, respectively.

a) Load `DataDefaults.csv` into R.

b) Estimate the correlation between the loan default rates and earnings, and check your answer using the function `cor()` in R.

c) Estimate the intercept and slope of the simple linear regression model of loan default rate against Earnings, and comment on their values. Do not use `lm()`/`glm()` functions.

d) Does the expected relationship between your estimated slope and correlation coefficient hold?

e) Estimate the variance parameter σ of the error terms.

f) Estimate the covariance matrix for the regression parameters.

g) One AIMS alumni earning 286,510 has just applied for a loan to the Company. Estimate:

   i) The expected default rate for this applicant.

   ii) The standard error around this expected default rate.

   iii) Comment on your result.

Lab 1.3

a) Let X be a random variable, and a and b two constants.

   Show that:

   $\mathbb{E}[aX+b] = a\,\mathbb{E}[X]+b$ and $Var[aX+b] = a^2\,Var[X]$

b) Let $X_1, ..., X_n$ be $i.i.d.$ random variables with expected value μ and variance σ^2.

   Let $\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ and $S^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{(n-1)}$

   Prove that $\bar{X}$ and S^2 are unbiased estimators of μ and σ^2.

c) Derive a relationship between the least square estimator $\hat{\beta}_1$ and the sample correlation coefficient of X and Y. Comment on this relationship.