Intermediate Econometrics

29th October 2025 - Vincenzo Gioia

Regression Models: Examples

Some uses of regression models:

1. To predict an individual's income based on gender, holding other conditions constant (such as education level, age, etc.)
2. To predict the number of exams taken by a first-year student based on demographic data, income, school background, etc.
3. To predict the number of claims made by an insured person based on their individual characteristics and past history
4. To assess whether blood pressure decreases following the administration of a drug, taking individual characteristics into account
5. To evaluate how mortality in the population varies according to the concentration of air pollutants
6. To decide whether a credit card payment is fraudulent

Reg. Models: A Common Framework

Framework:

- a quantity of interest (income, number of exams, number of claims, blood pressure, mortality, fraudulence): **the response (or outcome or dependent variable)**
- other quantities: **the explanatory (independent) variables (also called covariates or regressors)**

Question: **the first is related to the latter, and, if so, how?**

Note

The latter (covariates) are of practical interest only insofar as they are connected to the first (outcome)

Reg. Models: Goals

Predictive

Obtain a tool for predicting the value of the variable of interest given the values of the explanatory variables (e.g., because these are easier to measure or can be observed in advance with respect to the response)

- Example 3: when the goal is to determine the insurance premium that the policyholder should pay
- Example 6: to block a transaction before it is carried out

Interpretative

The main interest is to determine which explanatory variables have the strongest relationship with the response, and in which direction that relationship goes

- Example 1: when the goal is to determine whether there is gender-based discrimination
- Example 4: when the goal is to determine whether the drug is effective

Reg. Models: General Form

The probability distribution of the outcome depends on the covariates

$$([\text{outcome}]) \sim f(y; [\text{covariates}])$$

Note

- Asymmetric relationship
- $f(\cdot; \cdot)$ is specified up to a parameter

General structure depending on

1. **The type of outcome variable**
2. **The functional form of the relationship**

Reg. Models: Type of Outcome

Different models

- Binary: logistic/probit/. . . regression
- (Qualitative) Categorical: multinomial regression
- Counts (Quantitative discrete): Poisson (Negative Binomial) regression
- (Quantitative) Continuous: **Linear regression model**

Note

Under certain conditions the linear regression model can be used for quantitative discrete variables

Reg. Models: Data

Data Matrix

General structure including outcome and covariates

Unit	y	x_1	x_2	$\cdots$	x_p
1	y_1	x_{11}	x_{12}	$\cdots$	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	y_i	x_{i1}	x_{i2}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_n	x_{n1}	x_{n2}	$\cdots$	x_{np}

Reg. Models: Linear Regression Model

$$Y_i \sim f(y_i; x_{i1}, \ldots, x_{ip}), \quad i = 1, \ldots, n$$

Additive structure

$$h(Y_i) = g(x_{i1}, \ldots, x_{ip}) + \varepsilon_i$$

- $h(\cdot)$: known function
- $g(x_{i1}, \ldots, x_{ip})$: systematic component
- ε_i: error component

Reg. Models: Linear Regression Model

Linearity of $g(\cdot)$

$$g(x_{i1}, \ldots, x_{ip}) = \beta_1 g_1(x_{i1}) + \ldots + \beta_p g_p(x_{ip})$$

- $g_j(\cdot)$: known function

Linear model

$$h(Y_i) = \beta_1 g_1(x_{i1}) + \ldots + \beta_p g_p(x_{ip}) + \varepsilon_i$$

- $h(\cdot)$ and $g_j(\cdot)$: known functions
- ε_i: random variables with mean zero (whose distribution is specified up to a parameter)
- $\beta_1, \ldots, \beta_p$: parameters to estimate

Reg. Models: Linear Regression Model

Examples

$$Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

$$Y_i = \beta_1 + \beta_2 x_{i2}^2 + \beta_3 \sqrt{x_{i3}} + \varepsilon_i$$

$$\log(Y_i) = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + \varepsilon_i$$

$$\log(Y_i) = \beta_1 + \beta_2 \log(x_{i2}) + \beta_3 \log(x_{i3}) + \varepsilon_i$$

Reg. Models: Interpretation - Be Careful!

Relationship should not be interpreted as Cause-and-Effect

When we write a model in which one variable (y) is a function of another (x), it is very tempting to interpret it as x causes y

- **A statistical relationship — even a strong one — between y and x does not imply a cause-and-effect relationship**
- For example, both variables might be related to a third variable that causes them both
- There are statistical methods for making inferences about cause-and-effect relationships, but they require greater sophistication or a sample constructed in a specific way

Reg. Models: Linear Model

Linear model: $Y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$

Matrix representation $Y = X\beta + \varepsilon$

- Y: n-dimensional outcome vector
- X: $n \times p$ model matrix

Unit	x_1	x_2	$\cdots$	x_p
1	1	x_{12}	$\cdots$	x_{1p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
i	1	x_{i2}	$\cdots$	x_{ip}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	1	x_{n2}	$\cdots$	x_{np}

Linear Model: Error Term

Linear model: $Y = X\beta + \varepsilon$

- ε: n-dimensional error vector (**this component introduces casuality in the model,** Y is random beacuse ε is random)

Assumptions

1. **Linearity**
2. **Errors having mean 0, homoschedastic, and uncorrelated**
3. **Linear indepencence between explanatory variables**

Note

We do not make distributional assumptions on the error components: these are called **second-order hypotheses**

Linear Model: Assumption 1

Linearity

1. Linearity: $Y = X\beta + \varepsilon$

$$\begin{pmatrix} Y_1 \\ \vdots \\ Y_i \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{ip} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_i \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$(n \times 1) \qquad (n \times p) \qquad (p \times 1) \qquad (n \times 1)$$

Linear Model: Assumption 2

Second-order hypothesis for the error term

2. Errors having mean 0, homoschedastic, and uncorrelated

$$\mathbb{E}(\varepsilon) = 0 \qquad V(\varepsilon) = \begin{pmatrix} \sigma^2 & 0 & 0 & \cdots & 0 \\ 0 & \sigma^2 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \vdots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \sigma^2 \end{pmatrix} = \sigma^2 I$$

Note

- $E[Y|X] = X\beta \qquad (Y|X) = \sigma^2 I$

Linear model: Assumption 3

Linear indepencence between explanatory variables

3. The vectors $x_j, j = 1, \ldots, p$, are linearly independent

- **This guarantees the identifiability of the model**
- **It translates into a matrix** X (which is non-stochastic because we are working conditional to the values observed for the covariates) **that is of full rank** ($\text{rank}(X) = p$)

Linear Model: Least Square Estimator

Least Square (LS) Estimator (OLS: ordinary least square)

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y$$

LS estimator is obtained by minimizing the residual sum of squares

$$\mathrm{RSS}(\beta) = (Y - X\beta)^\top (Y - X\beta) = Y^t Y - 2\beta^\top X^\top Y + \beta^\top X^\top X\beta$$

$$\frac{\partial}{\partial \beta} \mathrm{RSS}(\beta) = -2X^\top Y + 2X^\top X\beta$$

$$\frac{\partial}{\partial \beta} \mathrm{RSS}(\beta) = 0 \implies \hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top Y$$

Linear Model: Important quantities

LS estimate

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top y$$

Predicted values

$$\hat{y} = X\hat{\beta}_{OLS} = X(X^\top X)^{-1} X^\top y = Py$$

where $P = X(X^\top X)^{-1} X^\top$ is called projection matrix (symmetric and idempotent)

Residuals

$$e = y - \hat{y} = (I - P)y$$

Linear Models: OLS properties

Properties of $\hat{\beta}_{OLS}$

- Unbiasedness: $E(\hat{\beta}_{OLS}) = \beta$
- $V(\hat{\beta}_{OLS}) = \sigma^2 (X^\top X)^{-1}$

We need an estimate of σ^2 (which is unknown)

- **The idea is to use the residuals as substitutes for the errors and to use their variance as an estimator of σ^2**
- $\hat{\sigma}^2 = \frac{1}{n} e^\top e$, which is biased
- A consistent estimate is given by

$$S^2 = \frac{1}{n-p} e^\top e$$

- This implies that the variance/covariance of the $\hat{\beta}_{OLS}$ estimator is

$$\hat{V}(\hat{\beta}_{OLS}) = S^2 (X^\top X)^{-1}$$

Linear Models: R^2 coefficient

How well the pedicted values $\hat{y}$ are able to represent the observed data y

- Measure of goodness of fit: R^2 coefficient

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- $R^2 \in [0,1]$
- It represents the fraction of variability of Y explained by the model

Deviance decomposition

$$\text{Total deviance} = \text{Model deviance} + \text{Residual deviance}$$

- $\text{Total deviance} = \sum_{i=1}^{n}(y_i - \bar{y})^2$
- $\text{Model deviance} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$
- $\text{Residual deviance} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

Linear Models: In practice

Credit Card Balance Data

- Description: A simulated data set containing information on ten thousand customers. The aim here is to predict which customers will default on their credit card debt.
- Outcome: Balance
- Available Covariates: Income, Limit, Rating, Cards, Age, Education, Gender, Student, Married, Ethnicity

We just considered a simple example regressing Balance on Student and Limit

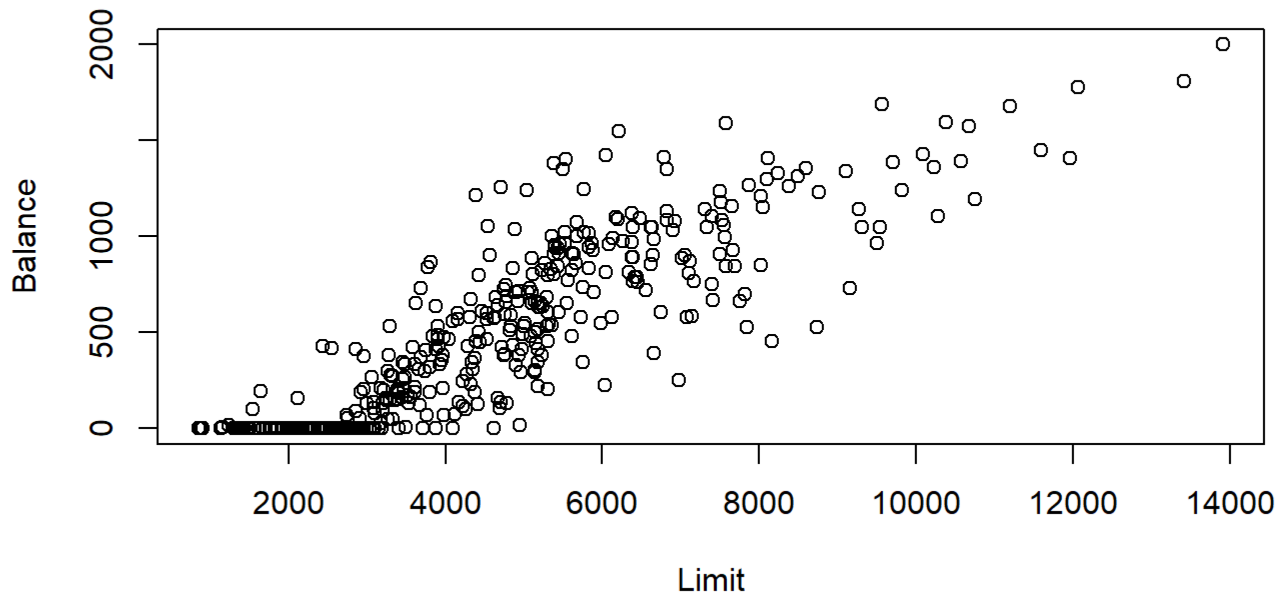
```
library(ISLR)
data("Credit")
#help(Credit) you can see the help
```

Linear Models: In practice

Exploratory purposes

- Scatterplot

```
with(Credit, plot(Limit, Balance))
```

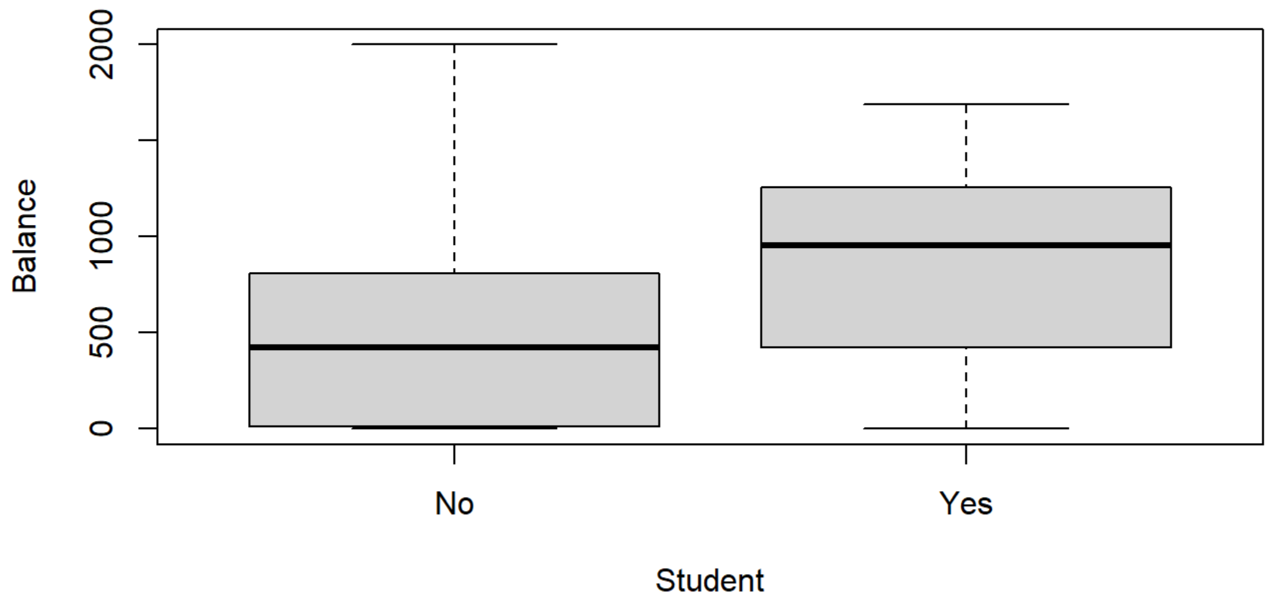


Linear Models: In practice

Exploratory purposes

- Boxplots

```
with(Credit, plot(Balance ~ Student))
```

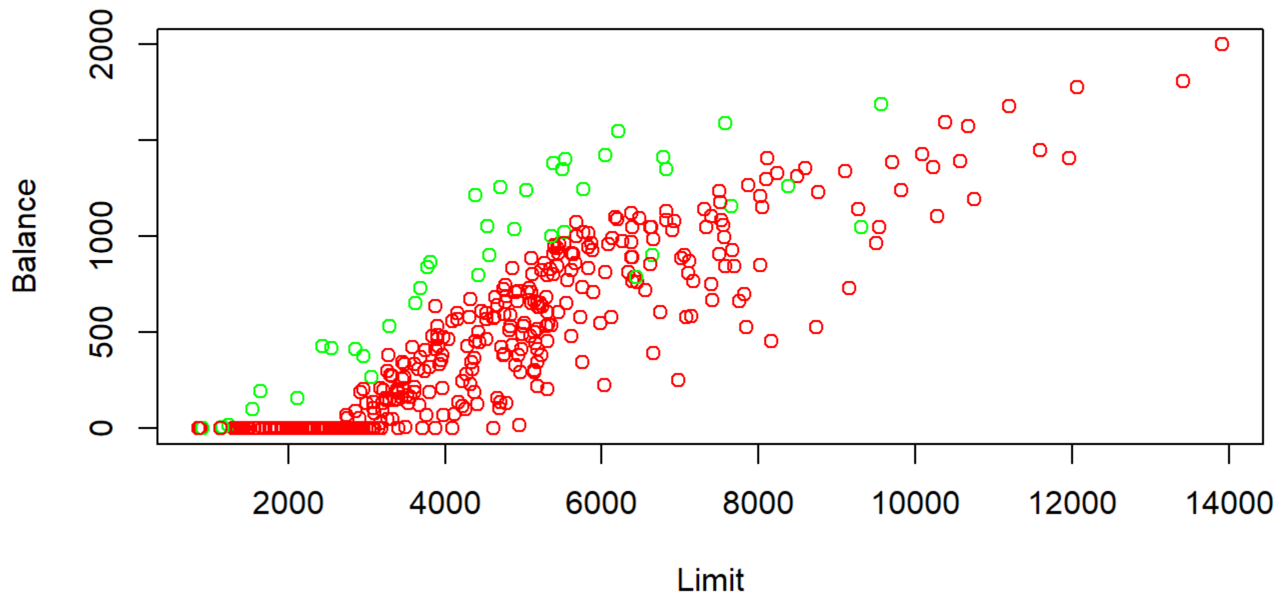


Linear Models: In practice

Exploratory purposes

- Combining both information

```r
with(Credit, plot(Limit, Balance, col = ifelse(Student == "Yes", "green", "red")))
```



Linear Models: In practice

The lm() function: fit a linear model on your dataset

- Model formula
- Specifying the dataset (where the variables can be found)
- Assign it to an object
- By simply digitizing the object you can see only the parameter estimates (in addition ti the call)

```r
lmFit <- lm(Balance ~ Limit + Student, data = Credit)
lmFit
```

```
Call:
lm(formula = Balance ~ Limit + Student, data = Credit)

Coefficients:
(Intercept)        Limit   StudentYes
   -334.730        0.172      404.404
```

Linear Models: In practice

Obtaining the parameter LS estimates by hand

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top y$$

```r
X <- model.matrix(Balance ~ Limit + Student, data = Credit)
y <- Credit$Balance
beta_OLS <- solve(t(X)%*%X)%*%t(X)%*%y
beta_OLS
```

```
                    [,1]
(Intercept) -334.7299372
Limit          0.1719538
StudentYes   404.4036438
```

```r
lmFit$coefficients
```

```
 (Intercept)        Limit   StudentYes 
-334.7299372    0.1719538  404.4036438 
```

Linear Models: An exhaustive summary

```
summary(lmFit)
```

```
Call:
lm(formula = Balance ~ Limit + Student, data = Credit)

Residuals:
    Min      1Q  Median      3Q     Max 
-637.77 -116.90    6.04  130.92  434.24 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) -3.347e+02  2.307e+01  -14.51   <2e-16 ***
Limit        1.720e-01  4.331e-03   39.70   <2e-16 ***
StudentYes   4.044e+02  3.328e+01   12.15   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 199.7 on 397 degrees of freedom
```

Linear Models: Table of coefficients

```
summary(lmFit)$coefficients
```

```
                Estimate   Std. Error   t value      Pr(>|t|)
(Intercept) -334.7299372 23.069301674 -14.50976  1.417949e-38
Limit          0.1719538  0.004330826  39.70463 2.558391e-140
StudentYes   404.4036438 33.279679039  12.15167  4.181612e-29
```

Quantities: first and second column

$$\hat{\beta}_{OLS} = (X^\top X)^{-1} X^\top y$$

$$\sqrt{[\hat{V}(\hat{\beta}_{OLS}]_{jj})} = \sqrt{[s^2(X^\top X)^{-1}]_{jj}}$$

```
p <- ncol(X)
hat_s2 <- sum(residuals(lmFit)^2)/(nrow(Credit) - p)
hat_Vb <- hat_s2*solve(t(X)%*%X)
sqrt(diag(hat_Vb))
```

```
 (Intercept)        Limit   StudentYes 
23.069301674  0.004330826 33.279679039 
```

Linear Models: Residuals summary

Residuals: $e = y - \hat{y}$

- A summary of the residuals
- Residual standard error (the square root of the unbiased estimate of the variance of the error term, σ^2)

```
head(residuals(lmFit))
```

```
        1          2          3          4          5          6 
 47.66441 -309.30693 -301.84343 -335.51929 -176.32798  102.01744 
```

```
head(Credit$Balance - predict(lmFit))
```

```
        1          2          3          4          5          6 
 47.66441 -309.30693 -301.84343 -335.51929 -176.32798  102.01744 
```

```
summary(residuals(lmFit))
```

```
    Min.  1st Qu.   Median     Mean  3rd Qu.     Max. 
-637.771 -116.900    6.045    0.000  130.916  434.236 
```

```
summary(lmFit)$sigma
```

```
[1] 199.6745
```

```
sqrt(sum(residuals(lmFit)^2)/(nrow(Credit)-ncol(X)))
```

```
[1] 199.6745
```

Linear Models: Residuals summary

R^2 **coefficient and the adjusted R-squared** (R_c^2)

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$R_c^2 = R^2 - \frac{p-1}{n-p}(1 - R^2)$$

```
summary(lmFit)$r.squared
```

```
[1] 0.8123267
```

```
Rsq <- 1 - sum(residuals(lmFit)^2)/(var(Credit$Balance)*(nrow(Credit) - 1))
Rsq
```

```
[1] 0.8123267
```

```
summary(lmFit)$adj.r.squared
```

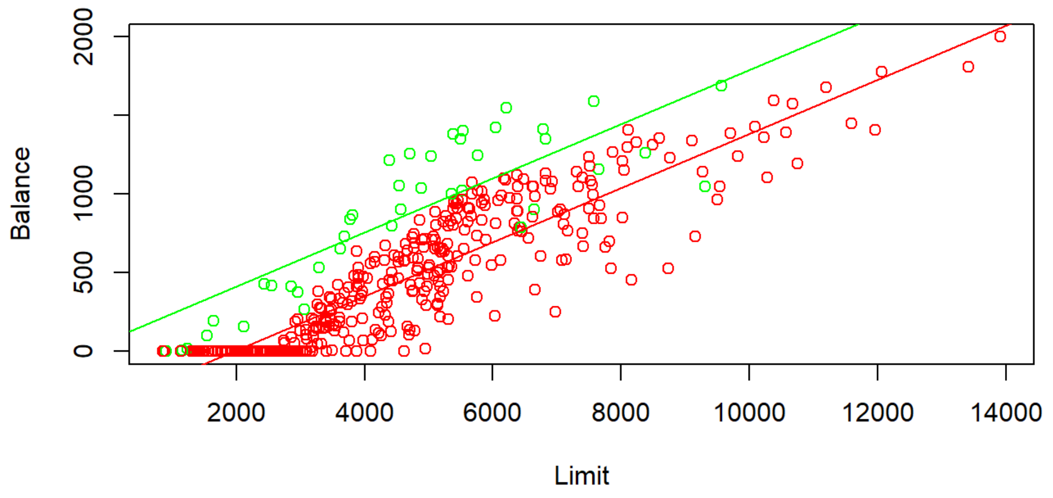
```
[1] 0.8113813
```

```
Rsq - (1-Rsq)*(ncol(X)-1)/(nrow(Credit)-ncol(X))
```

```
[1] 0.8113813
```

Linear Models: Regression lines

```r
with(Credit, plot(Limit, Balance, col = ifelse(Student == "Yes", "green", "red")))
abline(lmFit$coefficients[1:2], col = "red")
abline(c(lmFit$coefficients[1] + lmFit$coefficients[3],
         lmFit$coefficients[2]), col = "green")
```



Remember: This is just a toy example

- We are just using this data and this model setting to show how some quantities are obtained

Linear Models: Further steps

Obtaining the remaining quantitities of the summary

- t value
- $\mathrm{Pr}(> |\mathrm{t}|)$
- F-statistic

Explore violations of the linear model assumptions

- Residual analysis
- $\ldots$

What we need?

- A further assumption

Linear Models: Statistical Inference

Problems of Statistical Inference

- Estimation: **point** and interval estimation
- Hypothesis test
- Prediction (**point** or interval)

By deriving the OLS estimator we have just obtained a point estimator and derived its variance (which provide information on how the estimator is far from the unknown parameter β)

To carry out the remaining inferential results (interval estimation, hypothesis test), we need to

1. Use the asymptotic theory of the least squares or resampling techinque
2. **Introduce a further assumption: the errors are distributed according to a $\mathcal{N}(0, \sigma^2)$ and they are independent**

Normal Linear Model

Assumptions

1. Linearity

$$Y_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \varepsilon_i$$

2. Errors having mean zero, homoschedastic, normally distributed and independent

$$\varepsilon_i \sim \mathcal{N}(0, \sigma^2), \quad \text{independent}, \quad i = 1, \ldots, n$$

$$\varepsilon \sim \mathcal{N}_n(0, \Sigma), \quad \text{where} \quad \Sigma = \sigma^2 I$$

3. Linear independence between explanatory variables

Note

From 1) and 2), we get that the Y_i are independent with

$$Y_i \sim \mathcal{N}(\mu_i, \sigma^2), \quad \text{where} \quad \mu_i = \beta_1 + \beta_2 x_{i2} + \ldots + \beta_p x_{ip}, \quad i = 1, \ldots, n$$

Normal Linear Model: What we need?

By introducing the normality assumption for the errors we can derive

- Distribution of the estimators ($\hat{\beta}$ and $\hat{\sigma}^2$)
- Joint distribution of ($\hat{\beta}, \hat{\sigma}^2$)
- Pivotal quantity to make inference on a single coefficient (confidence interval, hypothesis test)
- Procedure for testing hypothesis on a group of coefficient and make predictions

Note

Friday, we will introduce the likelihood function and deriving the remaining quantities