Session 1

Oleg Deev & Štefan Lyócsa

Masaryk University

In general, regression analysis is concerned about estimating (conditional) expected value (or values) of **variable of interest** given known or **pre-determined** values of one or more independent variables.

- Similar ideas are behind most machine-learning techniques: LASSO, Ridge, Elastic net, Bayesian model averaging, Regression trees, Random forest, Neural Networks, Vector Support Machines,... .

The principle

Are we better off if we ignore information from other variables?

- What is the probability of institution K to default?

Institution	A	B	C	D	E	F	G	H	I	J	K*
Default (1 - yes, 0 - no)	0	0	0	1	1	0	0	0	1	0	?

[redacted]

The principle

Are we better off if we ignore information from other variables?

Institution	A	B	C	D	E	F	G	H	I	J	K*
Default (1 - yes, 0 - no)	0	0	0	1	1	0	0	0	1	0	?
Leverage ratio	5	6	6	4	5	8	10	7	2	8	6

The average Leverage ratio (Tier I/Total consolidated assets) for defaulted corporations is just 3.67 for non-defaulted 8.43.

- Data on leverage ratio seems to be helpful in predicting defaults.
- What is the probability of a default given some level of leverage ratio?
- We could **link** leverage ratio to the probability of default. Statistical methods help to find such 'links'.

Standard (non-matrix form) notation

Let's define:
Y_i - variable of interest (e.g. return on a loan - $RR2_i$).
X_i - explanatory (pre-determined) variable (e.g. verification of the income - $ver2_i$).
u_i - Stochastic (random) residual term.
$i = 1, 2, ..., N$ - index that labels observations.

We assume that Y_i can be calculated given an **expected** value of Y_i given realizations of X_i.
$Y_i = E(Y|X_i) + u_i$

Many possibilities for $E(.)$, linear regression assumes, that:
$Y_i = \beta_0 + \beta_1 X_i + u_i$

β_0 (intercept) a β_1 (slope) are unknown parameters, so called **regression coefficients**.

Residual term

$Y_i = \beta_0 + \beta_1 X_i + u_i$
Re-arranging:
$u_i = Y_i - (\beta_0 + \beta_1 X_i)$

This difference (u_i) is called the **stochastic residual term** or just **residual** or **error term**.
Error term shows that there are also other factors that influence variable of interest Y_i, not just X_i. Properties of u_i are key in regression analysis.

Sample regression curve

In reality, we only **assume** that $Y_i = \beta_0 + \beta_1 X_i + u_i$, and we never have **all** the data from the whole population (or we do not know the data-generating process). What we have is a **sample of data** (presumably a random sample of data that is representative).

In practice, using data and some models, we estimate this regression. The estimated (sample) model is:
$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$

Example

Let $RR2_i$ be the return on the loan and $ver2_i$ a variable with only two values, 1 if income was not verified and 0 otherwise.

$$RR2_i = \beta_0 + \beta_1 ver2_i + u_i$$

Given the data, we estimate the β parameters:

$$RR2_i = 8.58 - 3.52 ver2_i + \hat{u}_i$$

The intercept is 8.58 and the slope is -3.52. It is negative, meaning, that loans with not verified income $(ver2_i = 1)$ have a lower return than loans that have a verified income $(ver2 = 0)$.

Parameter estimation - OLS

The goal is that the sample regression line $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{u}_i$ fits the true data Y_i as 'well as possible'. The difference between the true value and the sample regression line is:
$\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$

What does it mean 'as well as possible'? There are many possibilities:

- $\min\limits_{\hat{\beta}_0,\hat{\beta}_1} \rightarrow \sum_{i=1}^{n} \hat{u}_i = \sum_{i=1}^{n} Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i$
- $\min\limits_{\hat{\beta}_0,\hat{\beta}_1} \rightarrow \sum_{i=1}^{n} |\hat{u}_i| = \sum_{i=1}^{n} |Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i|$
- $\min\limits_{\hat{\beta}_0,\hat{\beta}_1} \rightarrow \sum_{i=1}^{n} \hat{u}_i{}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2$

Ordinary **L**east **S**quares searchers for parameters $\hat{\beta}_0, \hat{\beta}_1$ for which the sum of squared residuals is minimized:

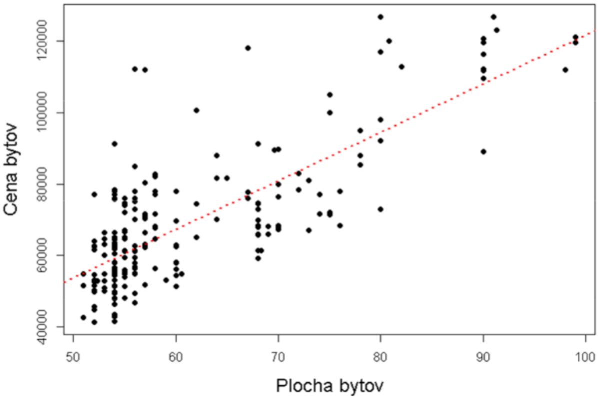
$$\underset{\hat{\beta}_0,\hat{\beta}_1}{min} \rightarrow \sum_{i=1}^{n} \hat{u_i}^2 = \sum_{i=1}^{n} (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 = f(\hat{\beta}_0, \hat{\beta}_1)$$

Linear regression and OLS estimator is far from perfect. It is almost surely a faulty model. But, it can nevertheless be useful. Some assumptions:

1. Model is linear in parameters $Y_i = \beta_0 + \beta_1 X_i$
2. Independent variables are not-stochastic.
3. As $E(u_i|X_i) = 0$ therefore $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$

4. Residuals are homoscedastic $var(u_i|X_i) = \sigma^2$.

Intuitively, if the salary depends on the gender, the error terms should be similar for man and woman.



Výrost, T., Baumohl, E., Lyócsa, Š., (2013). Kvantitatívne metódy v ekonómii 3, s. 218

5. Residuals u_i, u_j, where $i \neq j$ are not correlated.

Beware of the serial dependence in time-series, where error terms might be related in time, e.g. $cor(u_t, u_{t-1}) \neq 0$.

Often, time-series data are subject to seasonality, e.g. tourism arrivals in monthly data, $cor(u_t, u_{t-12}) \neq 0$.

What about spatial dependence?

6. Co-variance between u_i and X_i is zero, $E(u_i, X_i) = 0$

Assume that u_i and X_i are positively correlated. If X_i increases, so does u_i. Therefore coefficient β_2 for larger values of X_i underestimates the effect of X on Y, as the error term increases. Therefore β_2 does not have a meaningful interpretation.

7. Number of observations n should be more than the number of estimated coefficients.

How many parameters are estimated in a linear regression model?

8. The variance of the independent variable X should be finite and positive.
9. Regression is correctly specified.

10. In case of multiple independent variables, there is no perfect co-linearity between them.

Co-linearity between variables arises, if a variables is a property where the given variable can be expressed as a linear combination of all other variables.

Should we require P2P markets to verify the income of the borrower?

The return on the loan is $RR2_i$ and the variable that codes verification of the income is $ver2_i$. One (not the only one) approach is to estimate of the following linear model:

$$RR2_i = \beta_0 + \beta_1 ver2_i + u_i$$

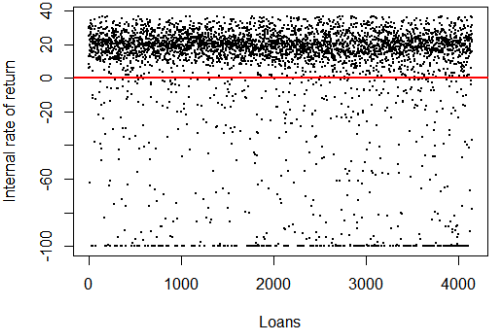
Another one could be a linear regression model:

$$int_i = \beta_0 + \beta_1 ver2_i + u_i$$

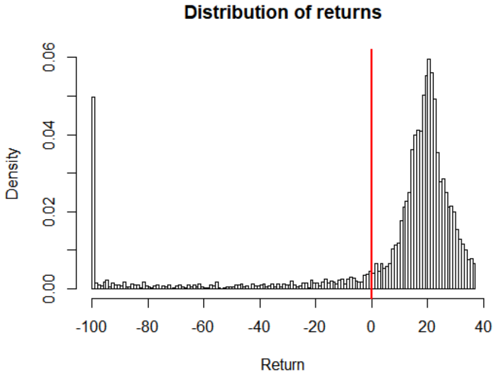
where, int_i is the interest rate on the loan contract on a p.a. basis.

Let's start the R session and open the script FinTech.R

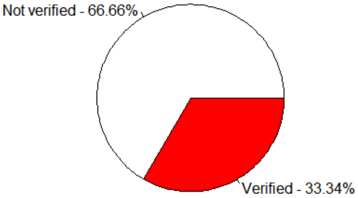
- `names(DT)`
- `plot(DT$RR2,type='p',pch=19,cex=0.25,xlab='Loans', ylab='Internal rate of return')`
- `abline(h=0,lwd=2,col='red')`



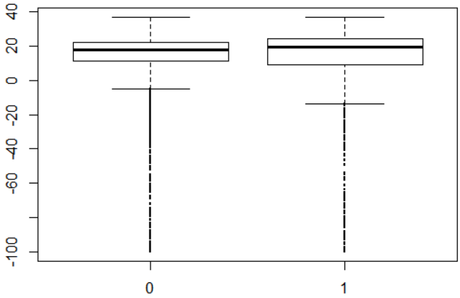
- `hist(DT$RR2,breaks=100,xlab='Return',prob=T, main='Distribution of returns')`
- `abline(v=0,lwd=2,col='red')`



- prct = round(100*table(DT$ver2)/sum(table(DT$ver2)),2)
- prct
- pie(table(DT$ver2),labels=paste(c("Not verified", "Verified"), " ",prct,"%",sep=),col=c("white","red"))



- `boxplot(DT$RR2 ~ DT$ver2,pch=19,cex=0.25)`



- Descriptive statistics

```
• y = DT$RR2
• y = na.omit(y)
• install.packages('lawstat')
• library(lawstat)
• round(c(mean(y),sd(y),min(y),median(y),max(y),
skewness(y),kurtosis(y)),2)
• round(100*sum(y==-100)/length(y),2)
• round(100*sum(y>-100 & y<0)/length(y),2)
• table(DT$ver2)
• prct
```

- OLS model estimation

```
m1 = lm(RR2 ~ ver2,data=DT)
m1
summary(m1)
install.packages("moments")
library(moments)
bptest(m1)
install.packages("sandwich")
library(sandwich)
coeftest(m1, vcov=vcovHC(m1,type='HC0'))
```

Is higher required return (interest rate) associated with lower returns?

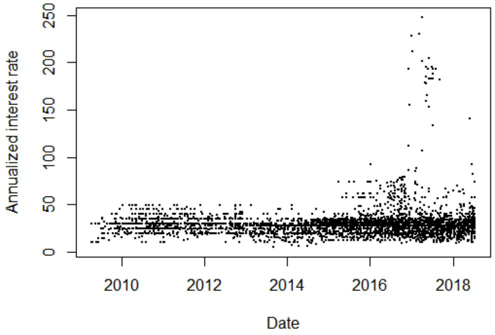
The return on the loan is $RR2_i$ and the interest rate on the loan (annualized) is int_i. We want to estimate:

$$RR2_i = \beta_0 + \beta_1 int_i + u_i$$

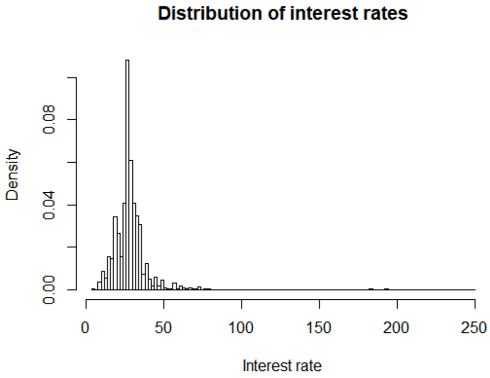
Another one could be a linear regression model:

We already saw data on $RR2_i$, let's continue with int_i.

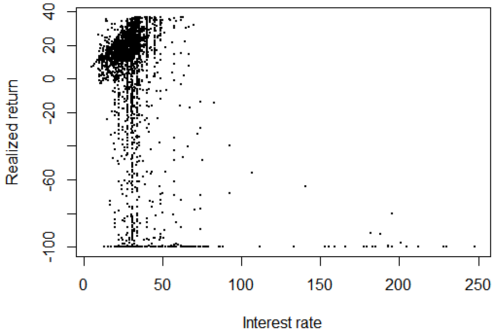
- `plot(y=DT$int,x=DT$date,type='p',pch=19,cex=0.25, xlab='Date',ylab='Annualized interest rate'))`



- `hist(DT$int,breaks=100,xlab='Interest rate',prob=T, main='Distribution of interest rates')`



- Not a textbook example of a nice relationship, but a real one...
- `plot(y=DT$RR2,x=DT$int,pch=19,cex=0.25,xlab='Interest rate',ylab='Realized return')`



- Descriptive statistics

```
• y = DT$int
• y = na.omit(y)
• round(c(mean(y),sd(y),min(y),median(y),max(y),
skewness(y),kurtosis(y)),2)
```

- OLS model estimation

```
m3 = lm(RR2 ~ int,data=DT)
m3
summary(m3)
bptest(m3)
coeftest(m3, vcov=vcovHC(m3,type='HC0'))
```

Multivariate model:

$$Y_i = \beta_0 + \beta_1 X_{i,1} + ... + \beta_p X_{i,p} + u_i$$

Interpretation of β_0 has not changed - an average value of Y given that X_1, X_2 are equal 0. Coefficients $\beta_0, \beta_1, ..., \beta_p$ are referred to as **parcial regression coefficients**, or simply regression coefficients.

Assume model:
$RR2_i = \beta_0 + \beta_1 int_i + \beta_2 ver2_i + u_i$

- Coefficient β_1 gives the change in Y given a unit change in int_i, for otherwise fixed values of $ver2_i$.

There is another (more complicated) way how to arrive to the β_1 coefficient.

The multivariate model:

$RR2_i = \beta_0 + \beta_1 int_i + \beta_2 ver2_i + u_i$

Instead, estimate:

$RR2_i = \alpha_0 + \alpha_1 ver2_i + u_{1,i}$

- If you subtract the effect of $ver2_i$ on $RR2_i$ you are left with $u_{1,i}$, i.e. the unexplained part of $RR2_i$.

$int_i = \gamma_0 + \gamma_1 ver2_i + u_{2,i}$

- If you subtract the effect of $ver2_i$ on int_i you are left with $u_{2,i}$, i.e. the unexplained part of int_i.

$u_{1,i} = \beta_1 u_{2,i} + u_{3,i}$

- Now β_1 is the **net effect** of int_i on $RR2_i$.

Multivariate model:
$Y_i = \beta_0 + \beta_1 X_{1,i} + ... + \beta_p X_{p,i} + u_i$

Two new issues are of concern:

- What if independent variables are linearly interconnected.
- How to evaluate the model fit.

Co-linearity

Assume to have two variables X_1 and X_2, non-existence of co-linearity means that there are no two real numbers λ_1 and λ_2 such, that:

$\lambda_1 X_{1,i} + \lambda_2 X_{2,i} = 0$

If such numbers do exists, we say that the variables X_1 and X_2 are **co-linear**.

For example, if $X_{1,i} = -4X_{2,i}$, we can arrange that so that $1X_{1,i} + 4X_{2,i} = 0$. It follows that the two variables are exactly co-linear.

We rather encounter examples of **near co-linearity** as exact co-linearity. For example, theory suggests that consumption is linearly driven by income and wealth. At the same time, income and wealth are related, but they are not exactly co-linear, e.g. both income and wealth influence consumption, but at least to some extent, they effect the consumption independently.

Model selection (LASSO, Ridge, Elastic net, Bayesian model averaging & model selection, ...) and machine learning techniques (Regression tree, random forest, artificial neural networks) to some extent **alleviate** the problem of near co-linearity.

Recall that coefficient of determination is calculated as:
$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\hat{u}_i^2}{\sum(Y_i - \bar{Y})^2}$

If we start adding independent variables into the model, R^2 is not going to decrease. This is a serious drawback of R^2.

How should we compare (more fairly) models with different number of independent variables?

There are many alternatives that access the fit of a model, while **penalizing increasing number of independent variables**. A standard one is the adjusted coefficient of determination.

$$adjR^2 = 1 - \frac{\frac{\hat{u}_i^2}{n-k}}{\frac{\sum(Y_i-\bar{Y})^2}{n-1}}$$

Where k is the number of parameters of the regression model (including the constant). While $adj.R^2$ can be used to compare multiple models (more fairly), the number itself cannot be interpreted in a same way as R^2.

What factors drive the rate of return on a loan?

- An investor might be interested in higher return.
- A consumer might be interested in interest rate.
- A policy maker might be interested in comparing returns a customers receives on a P2P market with returns on a similar loan of a standard commercial bank or a non-banking institution.

We use all data that we have available and start searching.... The OLS model is often a benchmark model (to beat).

$$RR2_i = \beta_0 + \beta_1 new_i + \beta_2 ver3_i + ... + \beta_p nrodep_i + u_i$$

We split the sample into two parts.

- First sample, **testing**, is used to estimate the model.
- Second sample, **validation**, is used to test the model's accuracy.

```
• NF = 100
• N = dim(DT)[1]
• Sample1 = DT[1:(N-NF),]
• Sample2 = DT[(N-NF+1):N,]
```

Now model estimation:

```
• m7 = lm(RR2 ∼ new+ver3+ver4+lfi+lee+luk+lrs+lsk+age+und
female+lamt+int+durm+educprim+educbasic+ educvocat+educse
espem+esfue+essem+esent+esret+dures+exper+ linctot+noliab
lamntplr+lamteprl+nopearlyrep,data=Sample1)
```

- `summary(m7)`
- `bptest(m7)`
- `coeftest(m7, df = Inf, vcov = vcovHC(m7, type = "HC0"))`

New library installation:

- `install.packages('car')`
- `library(car)`
- `which(vif(m7)>10)`

We now use model $m7$ to predict the return on the next 500 loans.

- `yhat = predict(m7,new=Sample2)`
- `ytrue = Sample2$RR2`
- `plot(y=ytrue,x=yhat,pch=19,cex=0.25,`
`ylim=c(min(yhat,ytrue),max(yhat,ytrue)),`
`xlim=c(min(yhat,ytrue),max(yhat,ytrue)),`
`xlab='Predicted returns',ylab='Realized returns')`
- `cbind(yhat,ytrue)`
- `hist(abs(yhat-ytrue),main='Forecast errors')`
- `mean(abs(yhat-ytrue))`
- `mean((yhat-ytrue)^2))`

Session 1

Oleg Deev & Štefan Lyócsa

Masaryk University