Pricing in Insurance

Basics of Risk Analysis

Dr. Rebecca Lodin, 2023

About If

- Property and casualty insurance
- Largest insurance company in the Nordics
- 3.9 million customers
- Handle 1.4 million claims yearly
- Almost 8000 employees

Who are we?

Rebecca Lodin

– analyst within private insurances, work with stable pricing and preparing data for quality assurance

– Theoretical physics at University of Edinburgh, PhD at Uppsala University

Ingrid Torstensson

– analyst within Commercial insurances, work with stable pricing and reserving

– Engineering Physics at LTH

Lovisa Julin

– analyst within Commercial insurances, work with property insurances – pricing and external data

– Engineering Physics at KTH

Today's Lecture

Price differentiation
Why is it necessary?

Risk
How to differentiate the price customers pay

Premium
Group customers according to factors

GLM
The theory needed to describe risk

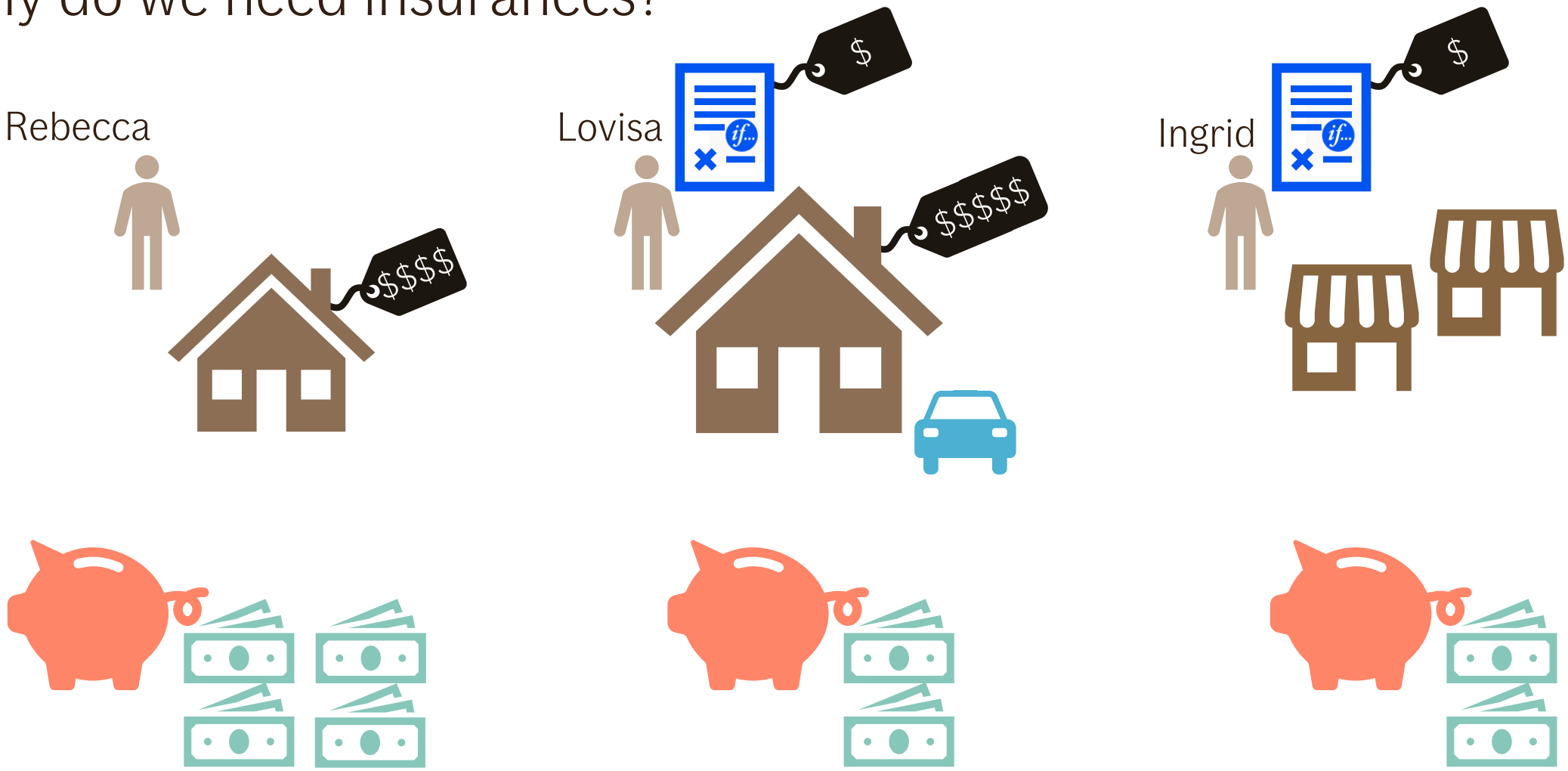
Project
Introduction

Why do we need insurances?

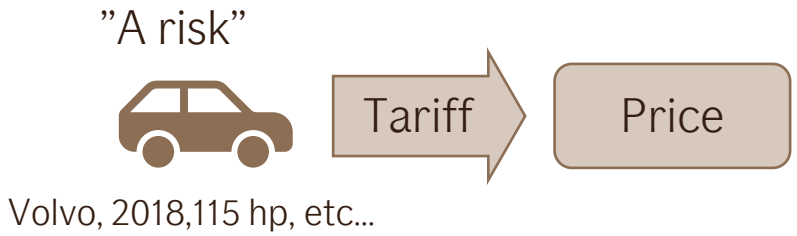
- Examples in Sweden from year 1200
- Insurance that covered fire
- Every homeowner in a region should give compensation to the person affected by the fire

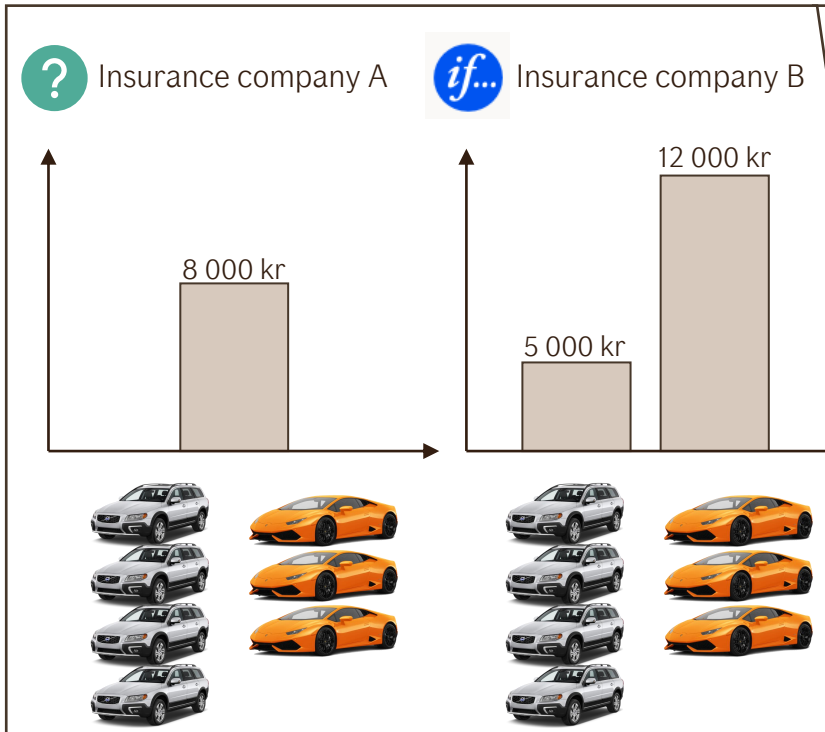
[www.svenskforsakring.se, 2023]

Why do we need insurances?

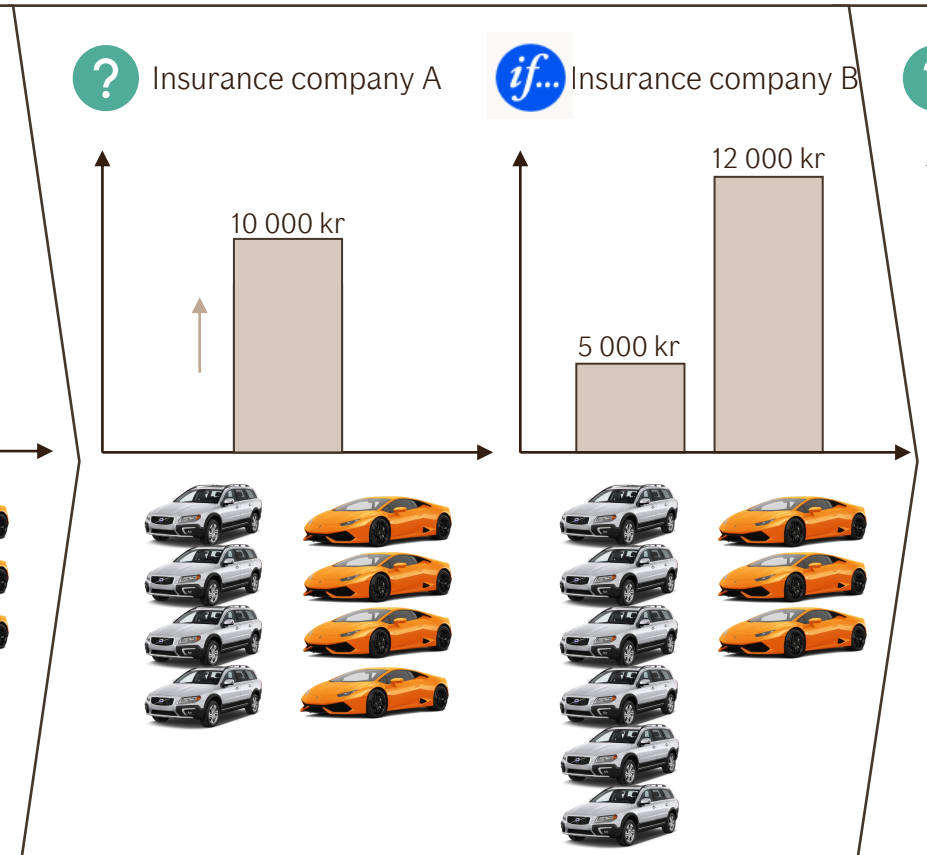


Price differentiation

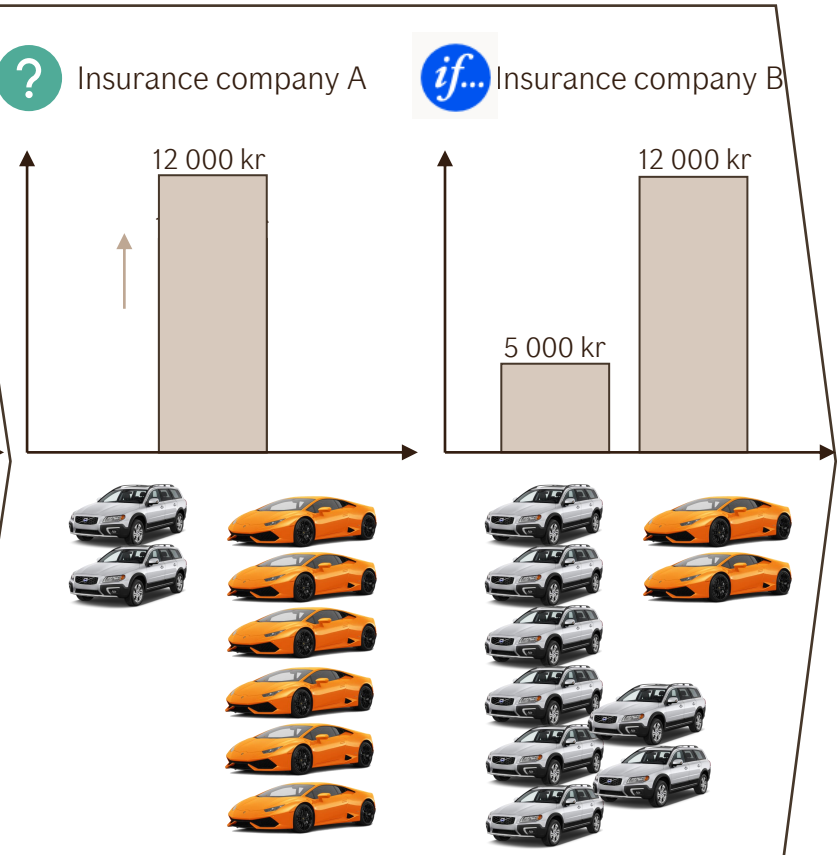




Total income and claims cost is the same for both companies.



Company A has to increase prices to cover increased claims costs.



And so it continues...

Risk

- The **premium** (or price of an insurance) a customer pays is based upon how much risk that customer poses.
- A lower risk customer should pay less premium than a higher risk customer.

Lower risk

Higher risk

Risk

The risk itself is measured in two parts:

- how often a claim occurs and
- how expensive the claims are.

The **claim frequency**: $$\text{Frequency} = \frac{\text{Number of claims}}{\text{Policy years}}$$
measures the average number of claims per policy year.

The **claim severity**: $$\text{Severity} = \frac{\text{Total claim cost}}{\text{Number of claims}}$$
measures the average claim cost.

Risk

The risk is defined as the product of the **claim frequency** and the **claim severity**:

$$\begin{aligned}\text{Risk} &= \text{Frequency} \times \text{Severity} \\ &= \frac{\cancel{\text{Number of claims}}}{\text{Policy years}} \times \frac{\text{Total claim cost}}{\cancel{\text{Number of claims}}} \\ &= \frac{\text{Total claim cost}}{\text{Policy years}}\end{aligned}$$

This is a measure of the average claim cost per policy year.

The higher the **expected annual claim cost** a customer has, the higher the risk.

Premium

Insurance premium = Risk premium (+ Some extra to pay our salary etc.)

The risk premium is not calculated on an individual level, but in customer groups based on what we know about them: the **pricing variables**.

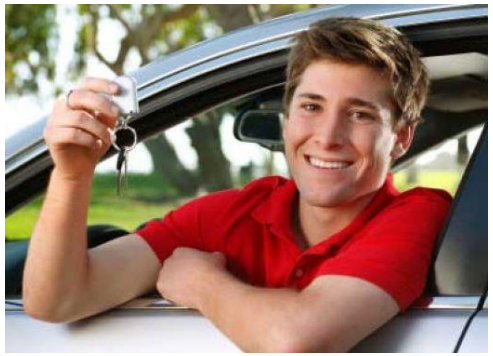
Example:

We will consider an insurance portfolio of professional drivers (e.g. taxi drivers)

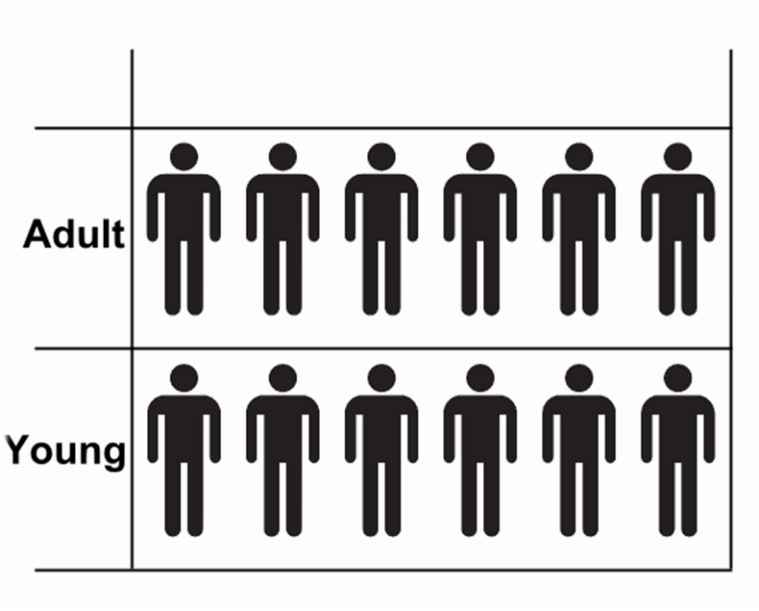
Premium

We know the age of the driver, and can divide them into two age groups:

Adult (31-60 years)



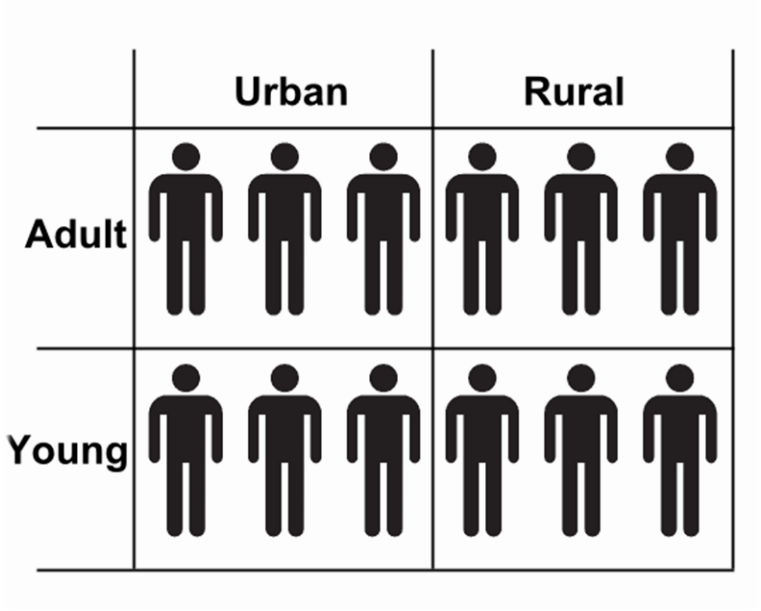
Young (18-30 years)



Premium

We know where the drivers work, and can introduce a geographical category.

Urban

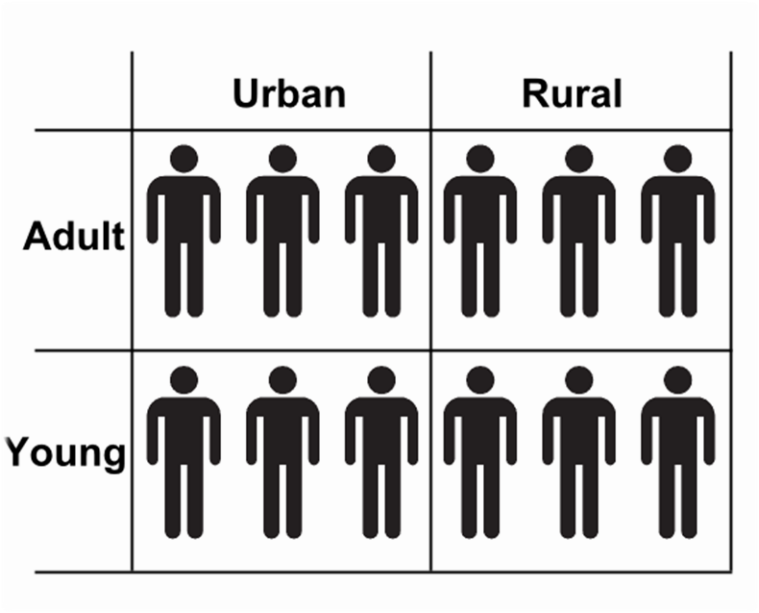


Rural

Premium

We have divided the customers into four groups, which are known as tariff cells.

All the customers within one tariff cell are treated as identical, and in the end they will get the same risk factor. We use GLM-analysis to determine how risky these cells are relative to each other.



Why GLM?

Motivation why GLM instead of LM:

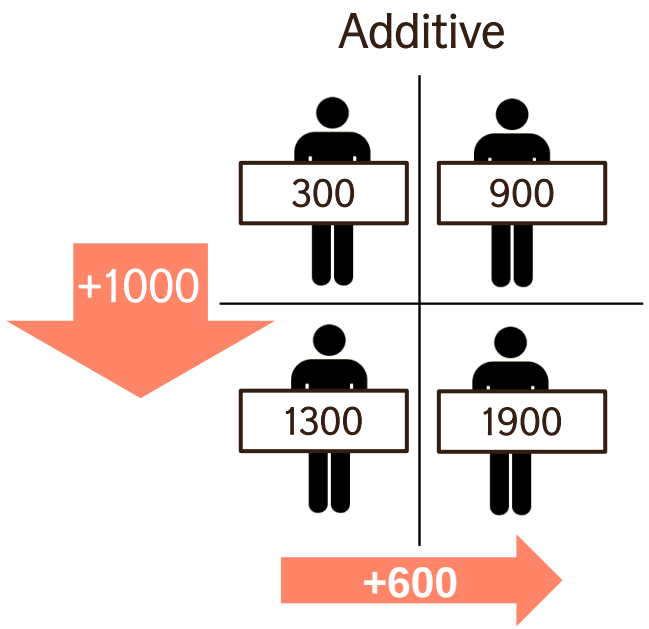
- All response variables are not normally distributed, some responses are always positive like insurance premium.
- Because we would like multiplicative model, easier to handle factors.

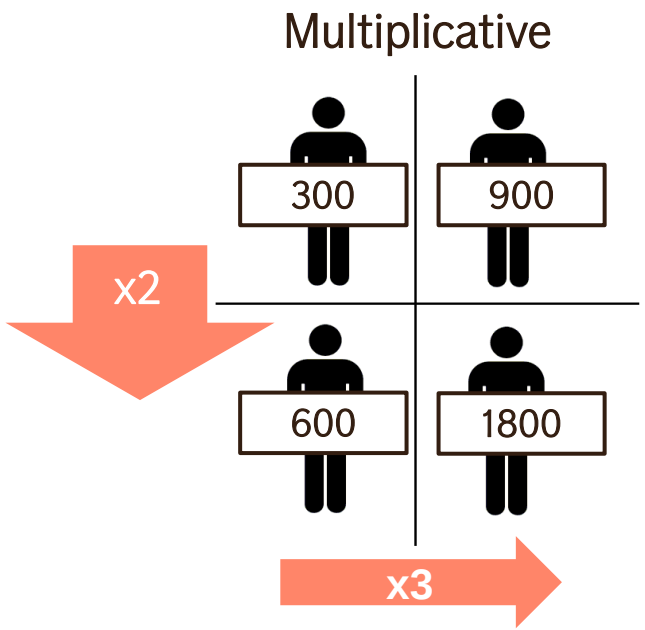
Why GLM?

Let us consider the models we use for pricing.

Pricing models

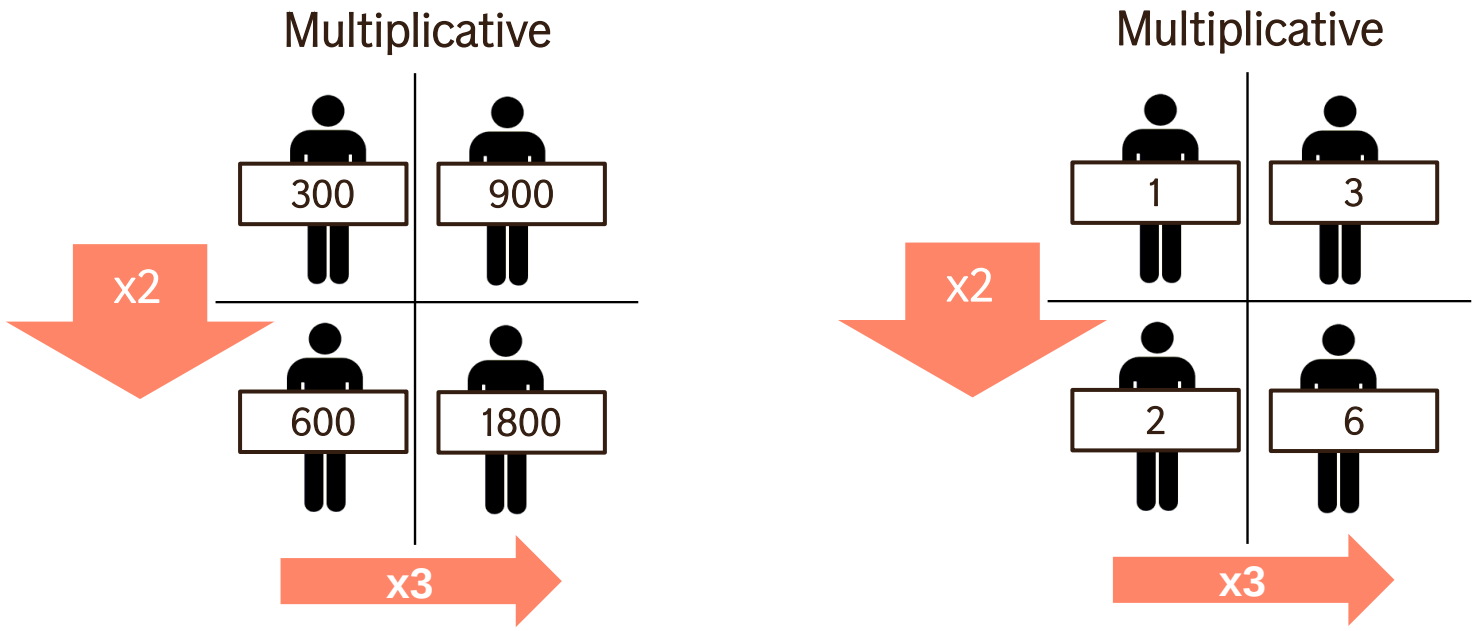
- There should be a relation between the premium that different customers pay.
- This is a simplification, but without it the pricing structure becomes too complicated.
- There are two choices: **additive** or **multiplicative** models.





Multiplicative model

- In the **multiplicative** model we get the relative factors between the customers.
- This is useful for various analysis purposes and is the main reason for choosing the multiplicative model over the additive model.



GLM-theory

- **GLM** (Generalized Linear Models) is a class of statistical methods that generalizes the Linear Models like e.g. simple linear regression.
- **GLM** is a way of expressing the relationship between an observed response variable Y, based on a number of covariates X.
- In insurance jargon: based on historical data Y (observed frequency, severity or risk) and the pricing variables X, we can calculate the expected value of Y.
- E[Y] is the best guess for future values and is how we determine future risk.
- The calculations are done by applying the method of **MLE** (Maximum Likelihood Estimation).

GLM-theory

The generalization from the Linear Models to GLM is done in two ways.

LM	**GLM**
$y_i \sim N(\mu_i, \sigma^2)$	
$\mathbb{E}[y_i] = \eta_i$	

$$\eta_i = \sum_{j=1}^{n} \beta_j x_{ij}$$

In the Linear models, the observed values are assumed to be Normally distributed.

The expectation of Y is η, where η is a **linear predictor**.

The β are the estimates.

GLM-theory

The generalization from the Linear Models to GLM is done in two ways.

LM	**GLM**
$y_i \sim N(\mu_i, \sigma^2)$	$y_i \sim \{\mathbf{EDM}\}$
$\mathbb{E}[y_i] = \eta_i$	$\mathbb{E}[y_i] = g^{-1}(\eta_i)$

$$\eta_i = \sum_{j=1}^{n} \beta_j x_{ij}$$

In the GLM-framework, we allow the response to follow any distribution in the class of **EDM** (Exponential Dispersion Models).

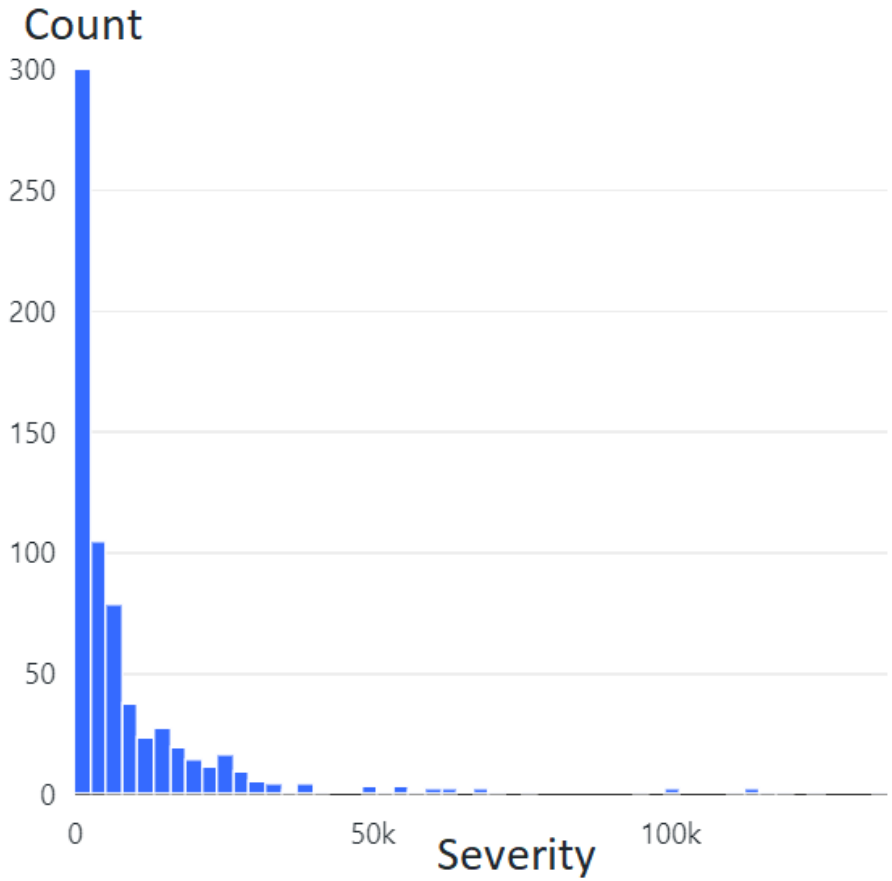
The expectation of Y is related to η through the inverse of a **link function** $g(\cdot)$.

η is the same as for linear models, the β are the estimates.

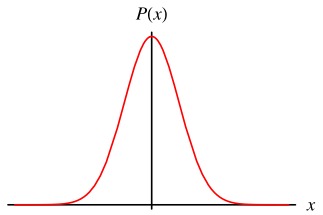
GLM-theory

Let us come back to **claim severity** (average claim cost).

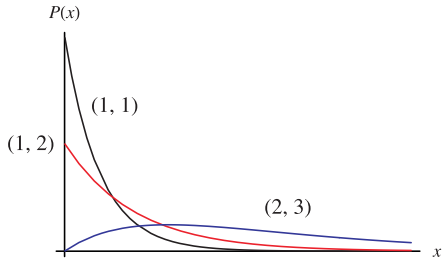
Which distribution does claim severity remind you of?



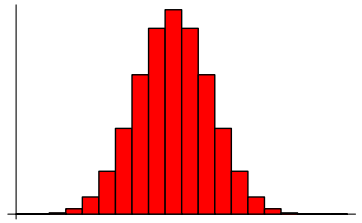
- Normal distribution:



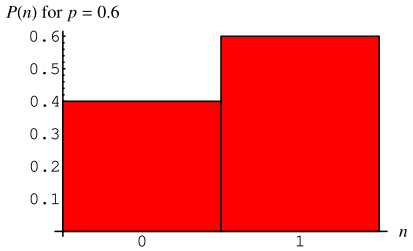
- Gamma distribution:



- Binomial distribution:



- Bernoulli distribution:

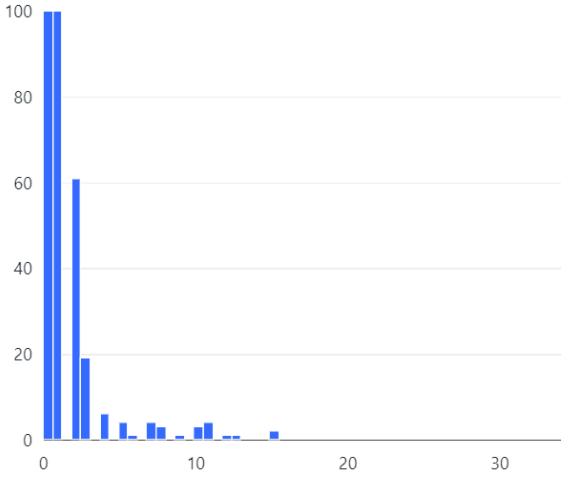


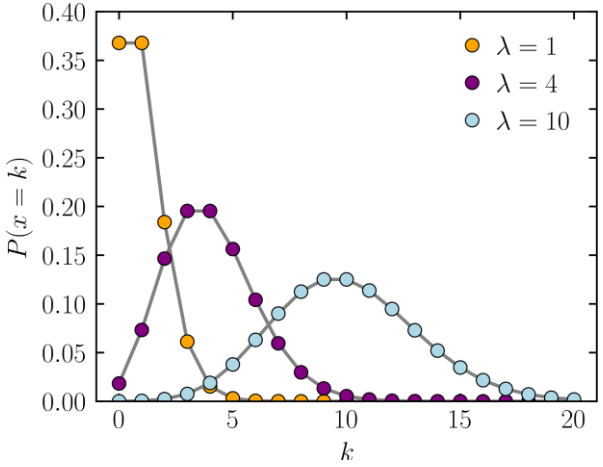
GLM-theory

The **EDM** (Exponential Dispersion Models) is a class of families of statistical distributions. It contains the normal, Poisson and gamma families of distributions (among others).

For our purposes:

The **frequency** response variables are assumed to follow a **Poisson** distribution.





GLM-theory

The **EDM** (Exponential Dispersion Models) is a class of families of statistical distributions. It contains the normal, Poisson and gamma families of distributions (among others).

For our purposes:

The **frequency** response variables are assumed to follow a **Poisson** distribution.

The **severity** response variables are assumed to follow a **gamma** distribution.

What about the **link function**?

GLM-theory: Link function $g(\cdot)$

Idea is to generalise: $\mu_i = \sum_{j=1}^{n} x_{ij}\beta_j$ to $g(\mu_i) = \sum_{j=1}^{n} x_{ij}\,\beta_j$.

Choice of $g(\cdot)$ is strongly related to the distributions of the response variables.

For EDM (exponential dispersion models) the distribution is given by

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta)}{a(\phi)} + c(y_i,\phi)\right).$$

θ_i is a parameter governing the shape of the density of Y (μ_i depends on θ_i).

ϕ is a dispersion parameter of the density of Y (think of it as known).

The **canonical link** is the link function that sets

$$g(\mu_i) = \theta_i = \sum_{j=1}^{n} x_{ij}\beta_j.$$

GLM-theory: Link function $g(\cdot)$

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

Example: Normal distribution

$$f_{Y_i}(y_i;\mu_i\,) = \frac{1}{\sqrt{2\,\pi}\sigma_i}\exp\left(-\frac{(y_i-\mu_i)^2}{2{\sigma_i}^2}\right) = \exp\left(\frac{y_i\mu_i - {\mu_i}^2/2}{{\sigma_i}^2} - \frac{{y_i}^2}{2{\sigma_i}^2} - \ln(\sigma_i) - \ln(\sqrt{2\,\pi})\right)$$

Thus,

$\theta_i = \mu_i$

$b(\theta_i) = {\theta_i}^2/2$

$a(\phi) = {\sigma_i}^2$ with the dispersion parameter $\phi = \sigma_i$

$c(y_i,\phi) = -{y_i}^2/(2{\sigma_i}^2) - \ln(\sigma_i) - \ln(\sqrt{2\,\pi})$

The link function is $g(\mu_i) = \theta_i = \mu_i$.

GLM-theory: Link function $g(\cdot)$

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

Example: Poisson distribution

$$f_{Y_i}(y_i;\mu_i\,) = \frac{1}{y_i!}e^{-\mu_i}\mu_i{}^{y_i} = \exp\big(y_i\ln(\mu_i) - \mu_i - \ln(y_i!)\big)$$

Thus,

$\theta_i = \ln(\mu_i)$

$b(\theta_i) = \mu_i$

$a(\phi) = 1$ since there is no dispersion

$c(y_i,\phi) = -\ln(y_i!)$.

The link function is $g(\mu_i) = \theta_i = \ln(\mu_i)$.

GLM-theory: Link function $g(\cdot)$

$$f_{Y_i}(y_i;\theta_i,\phi) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{a(\phi)} + c(y_i,\phi)\right)$$

Example: Gamma distribution with mean $\mu_i = \alpha_i\beta_i$

$$f_{Y_i}(y_i;\alpha_i\,,\beta_i) = \frac{{y_i}^{\alpha_i-1}e^{-\frac{y_i}{\beta_i}}}{{\beta_i}^{\alpha_i}\Gamma(\alpha_i)} = \exp\left(-\alpha_i\left(\frac{y_i}{\mu_i} - \ln(\mu_i)\right) + (\alpha_i - 1)\ln(y_i) - \alpha_i\ln(\alpha_i) - \ln\Gamma(\alpha_i)\right)$$

Thus,

$\theta_i = 1/\mu_i$

$b(\theta_i) = -\ln(\mu_i)$

$a(\phi) = -1/\alpha_i$ with dispersion parameter $\phi = \alpha_i$

$c(y_i,\phi) = (\alpha_i - 1)\ln(y_i) - \alpha_i\ln(\alpha_i) - \ln\Gamma(\alpha_i)$.

The link function is $g(\mu_i) = \theta_i = 1/\mu_i$.

GLM-theory: Link function $g(\cdot)$

Link function	$g(\mu_i) =$	$\mu_i =$	Normal	Poisson	Gamma
Identity	μ_i	$\sum x_{ij}\beta_j$	*	✓	✓
Log	$\ln(\mu_i)$	$e^{\sum x_{ij}\beta_j}$	✓	*	✓
Inverse	$1/\mu_i$	$(\sum x_{ij}\beta_j)^{-1}$	✓		*
Sqrt	$\sqrt{\mu_i}$	$(\sum x_{ij}\beta_j)^2$		✓	

* Indicates the canonical link

✓ Indicates that it is compatible

GLM-theory

The **frequency** response variables ~ **Poisson** distribution

The **severity** response variables ~ **gamma** distribution

We choose the **link function** to be: $g(\cdot) = \ln(\cdot) \Rightarrow g^{-1}(\cdot) = \exp(\cdot)$

This ensures we get a multiplicative model.

Summary of today's lecture

Risk
Why price differentiation is important
How to differentiate the price customers pay

Premium
Group customers according to pricing variables into tariff cells

GLM
How to generalise from LM
Frequency ~ Poisson distribution
Severity ~ Gamma distribution
Link function: $g^{-1}(\cdot) = \exp(\cdot)$

Introduction to Project II

GLM analysis of insurance for **Business Travel Cancellation**

Introduction

- Travel insurance is normally covered in home insurance
- Does not cover travel done as a part of your occupation
- Companies that require some travelling buy business travel insurance for their employees
- In this project, you will build a price model for the part of this insurance that gives compensation in case of **cancellations**

Price model

You should create a price model on the form

$$price = \gamma_0 \prod_{k=1}^{M} \gamma_{k,i}$$

where γ_0 is the base level and $\gamma_{k,i}$ is the risk factors corresponding to variable *k*, and variable group number *i*.

For example, let *k=1* correspond to number of persons insured. For one particular company, the number of persons insured are 27.

Then, according to the table to the right, $\gamma_1 = 10$.

Number of persons group *i*	Risk factor $\gamma_{1,i}$
1. $\# \leq 1$	0.7
2. $\# \leq 5$	2.2
3. $\# \leq 30$	10
4. $\# \leq 100$	25
5. $\# > 100$	35

RiskYear	Duration	NumberOfClaims	ClaimCost	Num_of_ft_Employee	CompanyAge	Travelling Area	Dangerous areas	Number of Persons
2019	0.166667	0	0.0	30	13.526027	Abroad (whole world)	Excluded	16.0
2020	0.147541	0	0.0	0	0.000000	Abroad (whole world)	Excluded	16.0
2020	0.415301	0	0.0	0	0.000000	Abroad (whole world)	Excluded	16.0
2020	0.161644	0	0.0	0	0.000000	Abroad (whole world)	Excluded	16.0
2016	0.504110	0	0.0	10	8.304110	Abroad (whole world)	Excluded	9.0
...	...	...	...	...	...	...	...	...
2017	0.534247	0	0.0	2	1.638356	In one's home country	Excluded	3.0
2017	0.465753	0	0.0	2	1.638356	In one's home country	Excluded	3.0
2018	0.534247	0	0.0	2	2.638356	In one's home country	Excluded	3.0
2018	0.465753	0	0.0	2	2.638356	In one's home country	Excluded	3.0
2019	0.534247	0	0.0	2	3.638356	In one's home country	Excluded	3.0

Data

Material

SF2930_Regression_analysis_Project_II.pdf

- Project description, instructions and appendix with explanation of variables in data

Cancellation.csv

- Policy and claims data between 2016-2020

GLM.ipynb

- Code example for you to expand
- In notebook style – use Jupyter notebook, Google Colab or similar

Your tasks

- Make one frequency model and one severity model
  - In the frequency model, we try to predict the **yearly number of claims** for a policy with a certain combination of variables
  - In the severity model, we try to predict the **average claim cost**

- Group variables
  - Make groups **risk homogeneous**
  - Have enough data in each group – and at least **one claim** in each group
- Handle missing values
  - One strategy is to put all these values in its own group
- Find a base level

LAB DEMO

Q&A sessions and demo lecture

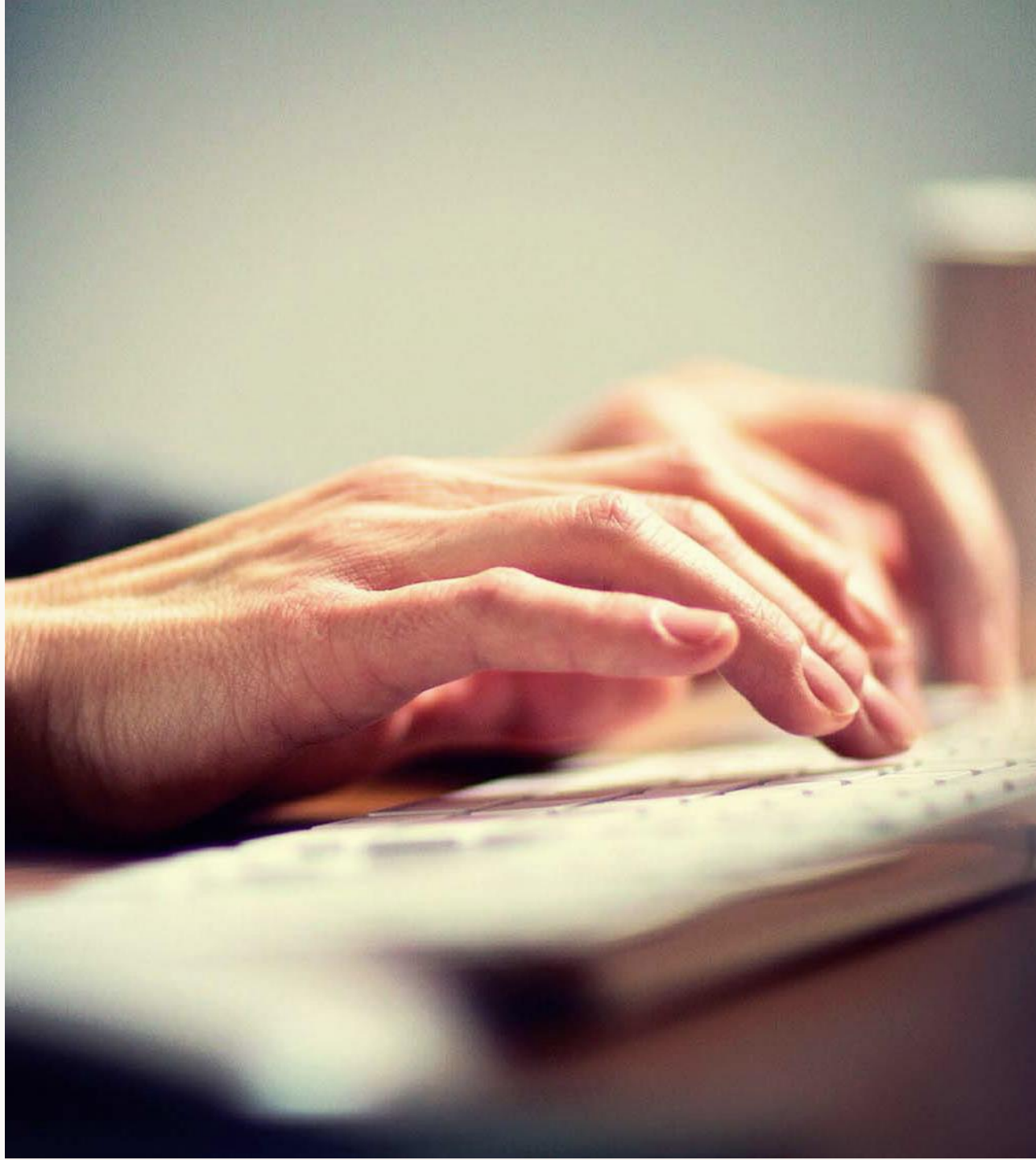
Monday, February 27, 13-15

Tuesday, February 28, 15-17

- You have time to work on the project
- We are available to answer questions
- We might start Tuesday's lecture with a common session depending on questions that have come up during Monday

Thursday, March 2, 8-10

- Insurance game!



Plan for next lecture

Tariff analysis
Mathematics behind a tariff

Test
AIC, BIC, Wald, portfolio profitability, Gini