

Pricing in Insurance

GLM modelling and tests

Ingrid Torstensson, 2023



Today's Lecture

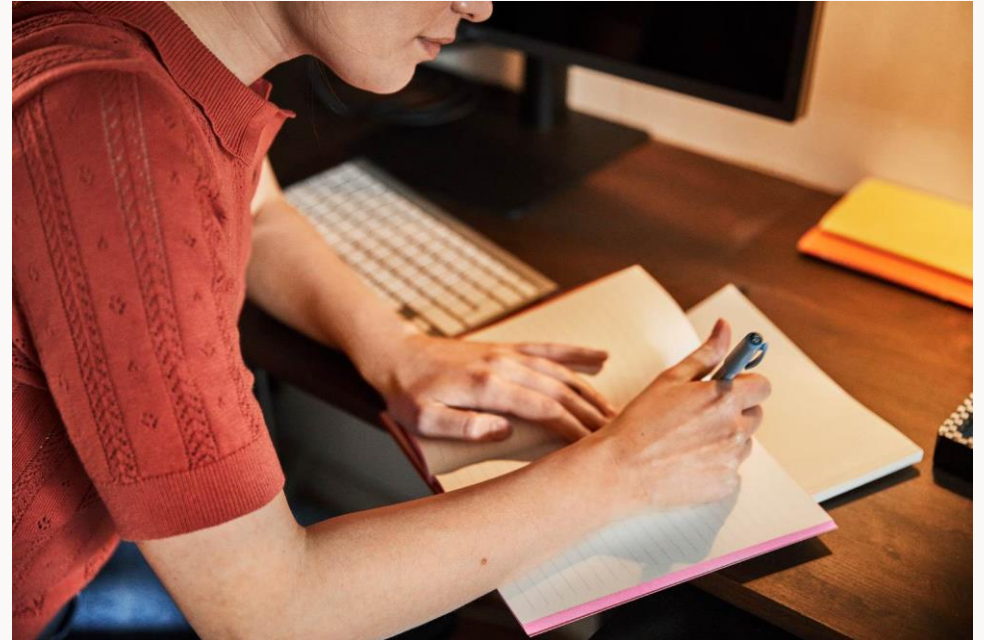
Repetition from last lecture

Tariff analysis

Walk-through of mathematics behind a tariff analysis in an example

Tests

Wald test, Likelihood test, AIC, BIC, Risk Ratio, Gini



Repetition

From last lecture

Risk

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

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

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**.

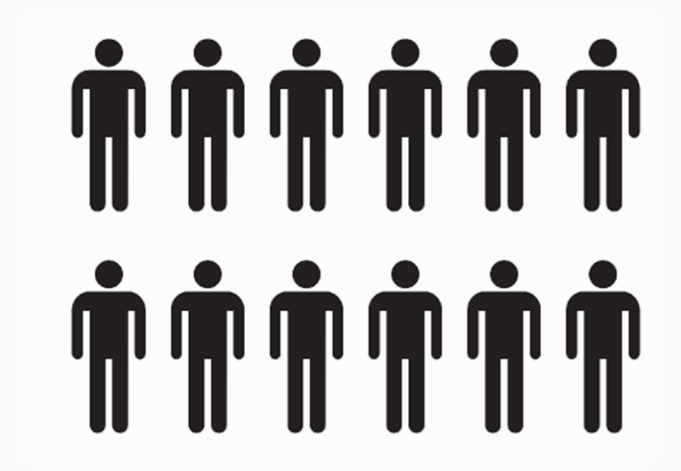
Premium

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 **rating factors**.

Example:

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



Premium

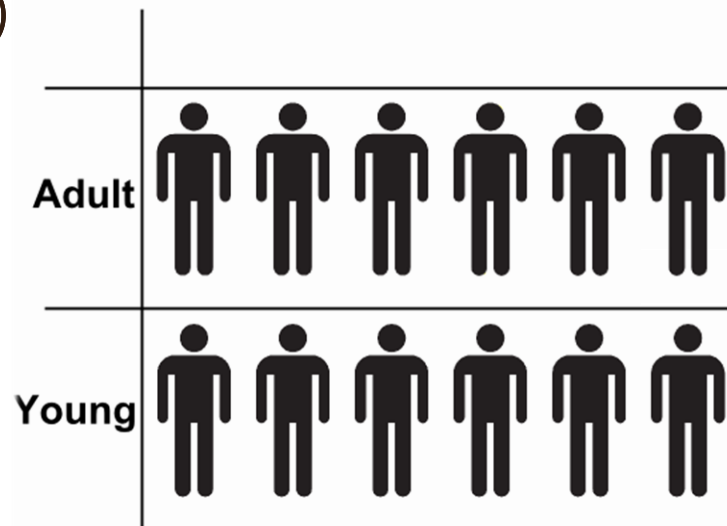
We know the age of the driver, and can divide them into two **age groups**:



Adult (31-60 years)



Young (18-30 years)



Premium













We know where the drivers work, and can introduce a [geographical category](#).



Urban















Rural

	Urban	Rural
Adult	  	  
Young	  	  

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.

	Urban	Rural
Adult	  	  
Young	  	  

GLM – Generalized Linear Models

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).





Example

Tariff analysis with GLM

Example - Claim cost modelling

Returning to our example, we will do the calculations to determine the **frequency** **relativities**. Here are the four **tariff cells** in a **tariff analysis data mart**.

Risk year	Age	Area
2010	adult	rural
2010	adult	urban
2010	young	rural
2010	young	urban

	Urban	Rural
Adult		
Young		

Mentimeter

Question 1

Example - Claim cost modelling

Returning to our example, we will do the calculations to determine the **frequency** **relativities**.

Risk year	Age	Area	Duration
2010	adult	rural	6812
2010	adult	urban	5923
2010	young	rural	5815
2010	young	urban	4923

We add **duration**, which is the time span when the insurance is active. Here we count it in **years**.

Example - Claim cost modelling

Returning to our example, we will do the calculations to determine the **frequency** **relativities**.

Risk year	Age	Area	Duration	Claim cost
2010	adult	rural	6812	1.645.000
2010	adult	urban	5923	289.000
2010	young	rural	5815	3.145.000
2010	young	urban	4923	1.523.000

We add **claim cost** for each tariff cell.

Example - Claim cost modelling

Returning to our example, we will do the calculations to determine the **frequency** **relativities**.

Risk year	Age	Area	Duration	Claim cost	NOC
2010	adult	rural	6812	1.645.000	2103
2010	adult	urban	5923	289.000	586
2010	young	rural	5815	3.145.000	3914
2010	young	urban	4923	1.523.000	1523

We add **NOC**, which means **Number of claims**.

Now we have all we need to calculate **claim frequency**.

$$\text{Frequency} = \frac{\text{Number of claims}}{\text{Policy years}}$$

Example - Claim cost modelling

Returning to our example, we will do the calculations to determine the **frequency relativities**.

Risk year	Age	Area	Duration	Claim cost	NOC	Frequency
2010	adult	rural	6812	1.645.000	2103	0,30872
2010	adult	urban	5923	289.000	586	0,09894
2010	young	rural	5815	3.145.000	3914	0,67309
2010	young	urban	4923	1.523.000	1523	0,30936

We add the calculated **Claim Frequency**.

Example - Claim cost modelling

Returning to our example, we will do the calculations to determine the **frequency relativities**.

Risk year	Age	Area	Duration	Claim cost	NOC	Frequency
2010	adult	rural	6812	1.645.000	2103	0,30872
2010	adult	urban	5923	289.000	586	0,09894
2010	young	rural	5815	3.145.000	3914	0,67309
2010	young	urban	4923	1.523.000	1523	0,30936

The tariff cell for rural adults contains the most data in terms of policy years and therefore we choose it as **base cell**. We calculate **frequency relativities** relative to this cell. To calculate the frequency relativities, we need to decide a **link function**.

Link function

Linear regression is a great choice for modelling linear phenomenon. E.g. how the cost vary depending on how many apples you buy.

When it is not a linear relationship, e.g. temperature and number of persons on a beach, linear regression is not suitable.

Linear regression: $g(\mu_i) = \mu_i$

Example of a GLM (logit): $g(\mu_i) = \ln\left(\frac{\mu_i}{1-\mu_i}\right)$

Link function g is defined as

$$g(\mu_i) = \sum_j x_{ij} \beta_j$$

Where μ_i is the expected value of response variable y_i , x_{ij} are the tariff cells (originates from the data set we use for predicting y_i) and β_j are parameters connecting x_{ij} and μ_i .

From last lecture: Choosing link function

The **frequency** response variables \sim **Poisson** distribution

The **severity** response variables \sim **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.

Connecting this to the previous slide, we get the link function and inverse function:

$$g(\mu_i) = \ln(\mu_i) \quad g^{-1} = \mu_i = e^{\sum x_{ij}\beta_j}$$

Example - Frequency modelling

Age and Area corresponds to x_{ij} and will be used to set up our tariff cells.

Risk year	Age	Area	Duration	Claim cost	NOC	Frequency y_i
2010	adult	rural	6812	1.645.000	2103	0,30872
2010	adult	urban	5923	289.000	586	0,09894
2010	young	rural	5815	3.145.000	3914	0,67309
2010	young	urban	4923	1.523.000	1523	0,30936

As discussed previous lecture:
Using response variables **frequency** y_i , the Poisson-distribution and log-link function, we can find the expected **frequency relativities** for these tariff cells.

We start with modelling the frequency part, and return to severity later.

Example - Frequency modelling

To connect the **variables** young/adult and rural/urban to **risk**, we need to set up a **matrix of tariff cells**.

The matrix has **as many rows as combinations of variables**.

i	Cell	x_{ij}		
1	(A,R)	x_{11}	x_{12}	x_{13}
2	(A,U)	x_{21}	x_{22}	x_{23}
3	(Y,R)	x_{31}	x_{32}	x_{33}
4	(Y,U)	x_{41}	x_{42}	x_{43}

Example - Frequency modelling

The matrix is rewritten with 0 and 1, making each row unique with regards to the set of variables. It is called the **design matrix**.

In this example, we need three columns.

The **first column** of covariates corresponds to the intercept level. This includes **adult A** and **rural R**.

The **second column** corresponds to age group. This corresponds to **young Y** or **not**.

The **third column** to area. This corresponds to **urban U** or **not**.

i	Cell	x_{ij}		
1	(A,R)	1	0	0
2	(A,U)	1	0	1
3	(Y,R)	1	1	0
4	(Y,U)	1	1	1

Basically, we add a factor 1 when the **rating factors** differ from the **base cell** on the first row.

Mentimeter

Question 2

Example - Frequency modelling

We keep the **design matrix** for the variable combinations.

We multiply with the vector of **beta estimates**: intercept, young and urban estimates.

From that we get the **linear predictor** for each tariff cell.

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \\ 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} \beta_1 \\ \beta_1 + \beta_3 \\ \beta_1 + \beta_2 \\ \beta_1 + \beta_2 + \beta_3 \end{bmatrix}$$

i	Cell	$\log(\mu_i)$				
1	(A,R)	β_1				
2	(A,U)	β_1		+	β_3	
3	(Y,R)	β_1	+	β_2		
4	(Y,U)	β_1	+	β_2	+	β_3

Then we can instead of the 0 and 1 describe each variable combination with linear predictors.

μ_i is the expected value of y_i , which means $\mu_i = E[y_i]$ and $y_i \sim \text{Pois}(\mu_i)$

Example - Frequency modelling

With the linear predictors we can calculate the **β estimates**: $\beta_1, \beta_2, \beta_3$

From that, we can find the **μ relativities**: μ_1, μ_2, μ_3

For the claim **frequency** y_i , we assume the **Poisson** distribution:

$$f(y_i; \mu_i) = \Pr(Y_i = y_i) = \frac{\mu_i^{y_i}}{y_i!} e^{-\mu_i}$$

For the four independent responses, we get the log-likelihood function:

$$\ell = \sum_{i=1}^4 y_i \cdot \log(\mu_i) - \log(y_i!) - \mu_i$$

Example - Frequency modelling

Writing out the sum:

$$\begin{aligned}\ell &= \sum_{i=1}^4 -\mu_i + y_i \cdot \log(\mu_i) - \log(y_i!) \\ &= -e^{\beta_1} + y_1 \cdot \log(e^{\beta_1}) - \log(y_1!) \\ &\quad - e^{\beta_1+\beta_3} + y_2 \cdot \log(e^{\beta_1+\beta_3}) - \log(y_2!) \\ &\quad - e^{\beta_1+\beta_2} + y_3 \cdot \log(e^{\beta_1+\beta_2}) - \log(y_3!) \\ &\quad - e^{\beta_1+\beta_2+\beta_3} + y_4 \cdot \log(e^{\beta_1+\beta_2+\beta_3}) - \log(y_4!)\end{aligned}$$

$\mu_1 = e^{\beta_1}$
$\mu_2 = e^{\beta_1+\beta_3}$
$\mu_3 = e^{\beta_1+\beta_2}$
$\mu_4 = e^{\beta_1+\beta_2+\beta_3}$

Example - Frequency modelling

Writing out the sum:

$$\begin{aligned}\ell &= \sum_{i=1}^4 -\mu_i + y_i \cdot \log(\mu_i) - \log(y_i!) \\ &= -e^{\beta_1} + y_1 \cdot \log(e^{\beta_1}) - \log(y_1!) \\ &\quad - e^{\beta_1+\beta_3} + y_2 \cdot \log(e^{\beta_1+\beta_3}) - \log(y_2!) \\ &\quad - e^{\beta_1+\beta_2} + y_3 \cdot \log(e^{\beta_1+\beta_2}) - \log(y_3!) \\ &\quad - e^{\beta_1+\beta_2+\beta_3} + y_4 \cdot \log(e^{\beta_1+\beta_2+\beta_3}) - \log(y_4!)\end{aligned}$$

Example - Frequency modelling

Writing out the sum:

$$\begin{aligned}\ell &= \sum_{i=1}^4 -\mu_i + y_i \cdot \log(\mu_i) - \log(y_i!) \\ &= -e^{\beta_1} + y_1 \cdot \log(e^{\beta_1}) \\ &\quad - e^{\beta_1+\beta_3} + y_2 \cdot \log(e^{\beta_1+\beta_3}) \\ &\quad - e^{\beta_1+\beta_2} + y_3 \cdot \log(e^{\beta_1+\beta_2}) \\ &\quad - e^{\beta_1+\beta_2+\beta_3} + y_4 \cdot \log(e^{\beta_1+\beta_2+\beta_3}) + C\end{aligned}$$

Example - Frequency modelling

Writing out the sum:

$$\begin{aligned}\ell &= \sum_{i=1}^4 -\mu_i + y_i \cdot \log(\mu_i) - \log(y_i!) \\ &= -e^{\beta_1} + y_1 \cdot \beta_1 \\ &\quad - e^{\beta_1+\beta_3} + y_2(\beta_1 + \beta_3) \\ &\quad - e^{\beta_1+\beta_2} + y_3(\beta_1 + \beta_2) \\ &\quad - e^{\beta_1+\beta_2+\beta_3} + y_4(\beta_1 + \beta_2 + \beta_3) + C\end{aligned}$$

Differentiating with regards to the beta-parameters and setting equal to 0.

$$\frac{\partial \ell}{\partial \beta_1} = -e^{\beta_1} + y_1 - e^{\beta_1+\beta_3} + y_2 - e^{\beta_1+\beta_2} + y_3 - e^{\beta_1+\beta_2+\beta_3} + y_4 = 0$$

$$\frac{\partial \ell}{\partial \beta_2} = -e^{\beta_1+\beta_2} + y_3 - e^{\beta_1+\beta_2+\beta_3} + y_4 = 0$$

$$\frac{\partial \ell}{\partial \beta_3} = -e^{\beta_1+\beta_3} + y_2 - e^{\beta_1+\beta_2+\beta_3} + y_4 = 0$$

Example - Frequency modelling

After some algebra, we are left with 3 equations and 3 unknowns; the MLE equations.

$$e^{\beta_1} + e^{\beta_1+\beta_2} + e^{\beta_1+\beta_3} + e^{\beta_1+\beta_2+\beta_3} = y_1 + y_2 + y_3 + y_4$$

$$e^{\beta_1+\beta_2} + e^{\beta_1+\beta_2+\beta_3} = y_3 + y_4$$

$$e^{\beta_1+\beta_3} + e^{\beta_1+\beta_2+\beta_3} = y_2 + y_4$$

Example - Frequency modelling

After some algebra, we are left with 3 equations and 3 unknowns; the MLE equations.

$$e^{\beta_1} + e^{\beta_1+\beta_2} + e^{\beta_1+\beta_3} + e^{\beta_1+\beta_2+\beta_3} = 1.390107$$

$$e^{\beta_1+\beta_2} + e^{\beta_1+\beta_2+\beta_3} = 0.982451$$

$$e^{\beta_1+\beta_3} + e^{\beta_1+\beta_2+\beta_3} = 0.408300$$

This is not trivial to solve analytically, so we solve it numerically.

$$\begin{array}{l} \beta_1 \approx -1.24517 \\ \beta_2 \approx 0.87961 \\ \beta_3 \approx -0.877529 \end{array} \quad \longrightarrow \quad \begin{array}{l} \mu_1 = e^{\beta_1} \approx 0.287921 \\ \mu_2 = e^{\beta_2} \approx 2.409959 \\ \mu_3 = e^{\beta_3} \approx 0.415866 \end{array}$$





The μ parameters are the relativities we are interested in.

Finding frequency factors

The first parameter, μ_1 , is the intercept. Since we are only interested in the **relativities**, we set $\mu_1 = 1$.

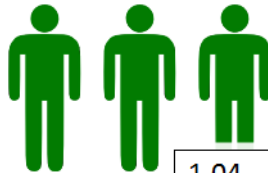



We have finally managed to find the **expected frequency factors** for the tariff cells.

(A,R)	$\mathbb{E}[y_1] = \mu_1$	$= 1$	$= 1.00$
(A,U)	$\mathbb{E}[y_2] = \mu_1\mu_3$	$\approx (1)(0.41)$	$= 0.41$
(Y,R)	$\mathbb{E}[y_3] = \mu_1\mu_2$	$\approx (1)(2.41)$	$= 2.41$
(Y,U)	$\mathbb{E}[y_4] = \mu_1\mu_2\mu_3$	$\approx (1)(0.41)(2.41)$	$= 1.00$

	Urban	Rural
Adult	 0.41	 1.00
Young	 1.00	 2.41

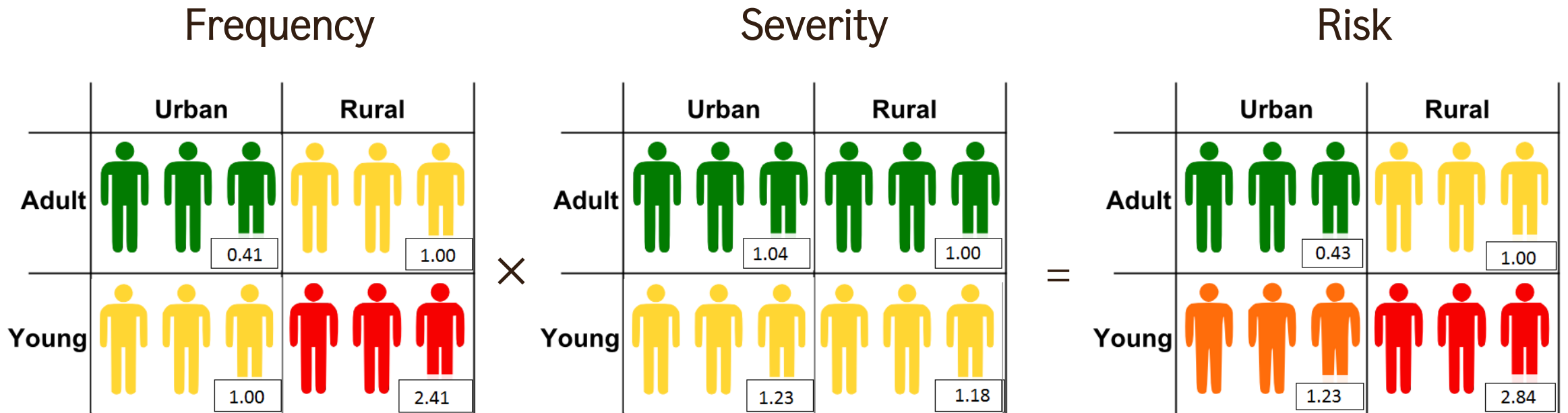
Finding severity factors

By doing the similar calculations for the **severity**, we find the following **expected severity factors**:

	Urban	Rural
Adult	 1.04	 1.00
Young	 1.23	 1.18

Finding risk factors

We get the final **risk factors** by multiplying the frequency with the severity, and have completed the GLM-analysis.



Tests

Tests

Modelling tests

- Hypothesis testing - Wald test
- How good is our model? – Likelihood-ratio test
- Overfitting - AIC and BIC

Business tests

- Portfolio profitability
- Gini

Wald test

Hypothesis test to check if all parameters are relevant.

$$H_0: \beta_j = 0$$

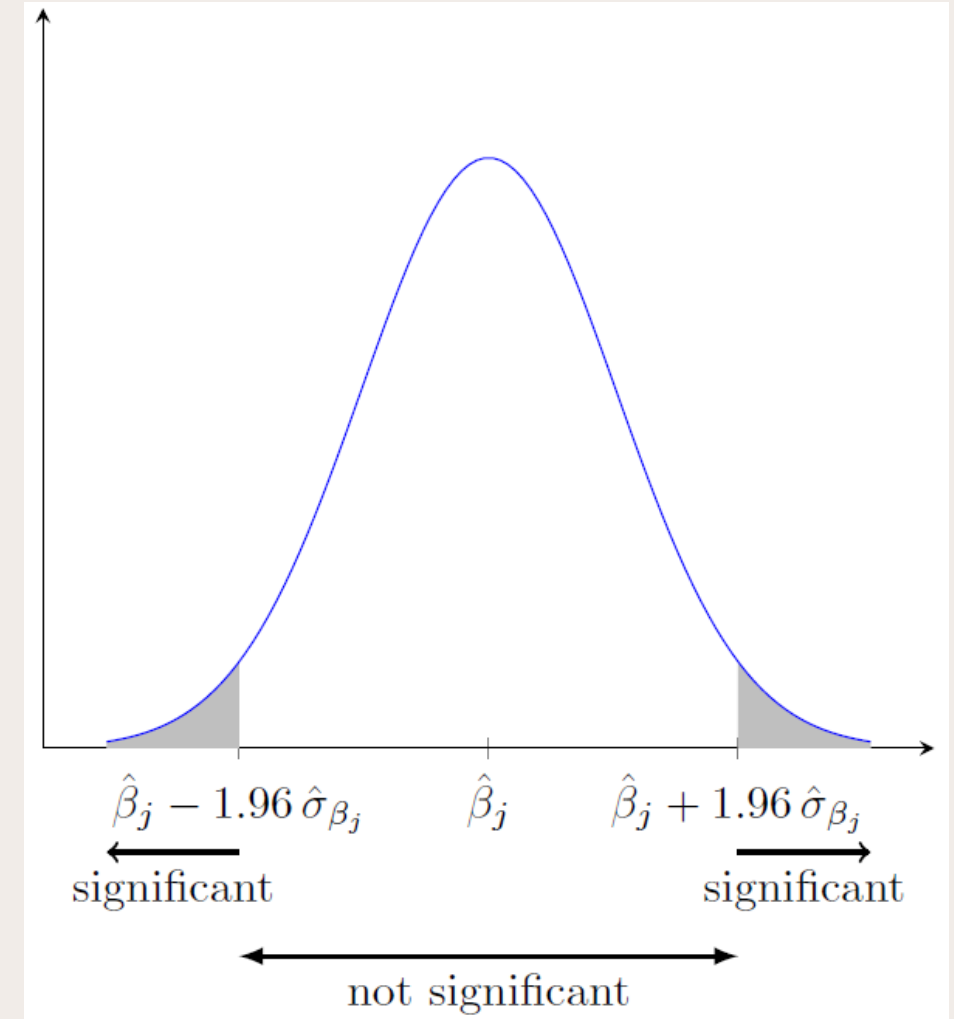
$$H_1: \beta_j \neq 0$$

$$\hat{\beta}_j \sim N(\beta_j, \sigma_{\beta_j}^2)$$

$$I_{\beta_j}: (\hat{\beta}_j \pm Z_{\frac{\alpha}{2}} \cdot \sigma_{\beta_j})$$

where $Z_{\frac{\alpha}{2}} = 1.96$ for $\alpha = 0.05$

If 0 is not within the confidence interval, then H_0 is false and β_j is significant.



Estimate standard error σ_{β_j}

1. Create the Hessian matrix

$$G = \begin{bmatrix} \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_1 \partial \beta_n} \\ \frac{\partial^2 \ell}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_2 \partial \beta_2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_2 \partial \beta_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 \ell}{\partial \beta_n \partial \beta_1} & \frac{\partial^2 \ell}{\partial \beta_n \partial \beta_2} & \cdots & \frac{\partial^2 \ell}{\partial \beta_n \partial \beta_n} \end{bmatrix}$$

2. Insert the maximum likelihood estimates $\beta_1 \dots \dots \beta_n$.

This gives us numbers in the matrix, which is called the evaluated matrix \hat{G} .

3. Calculate the negative inverse of the evaluated matrix, $-\hat{G}^{-1}$, in which the diagonal element with index (j,j) is $\widehat{Var}(\hat{\beta}_j)$

4. The standard error is then $\hat{\sigma}_{\beta_j} = \sqrt{\widehat{Var}(\hat{\beta}_j)}$

Likelihood-ratio test

Full model, FM: $\log \mu_i = \beta_0 + x_{i1}\beta_{i1} + x_{i2} + \cdots + x_{in}\beta_{in}$

Reduced model, RM: $\log \mu_i = \beta_0$

$$\begin{aligned} LR &= 2 \cdot \ln \frac{\mathcal{L}(FM)}{\mathcal{L}(RM)} = 2(\ln \mathcal{L}(FM) - \ln \mathcal{L}(RM)) & \mathcal{L}(RM) \text{ is the likelihood function} \\ &= 2 (\ln(FM) - \ln(RM)) \\ &= \chi^2(\# \text{ parameters in FM} - \# \text{ parameters in RM}) \end{aligned}$$

$H_0 =$ *The reduced model RM gives the same results as the full model FM*

$H_1 =$ *RM does not give the same result as FM*

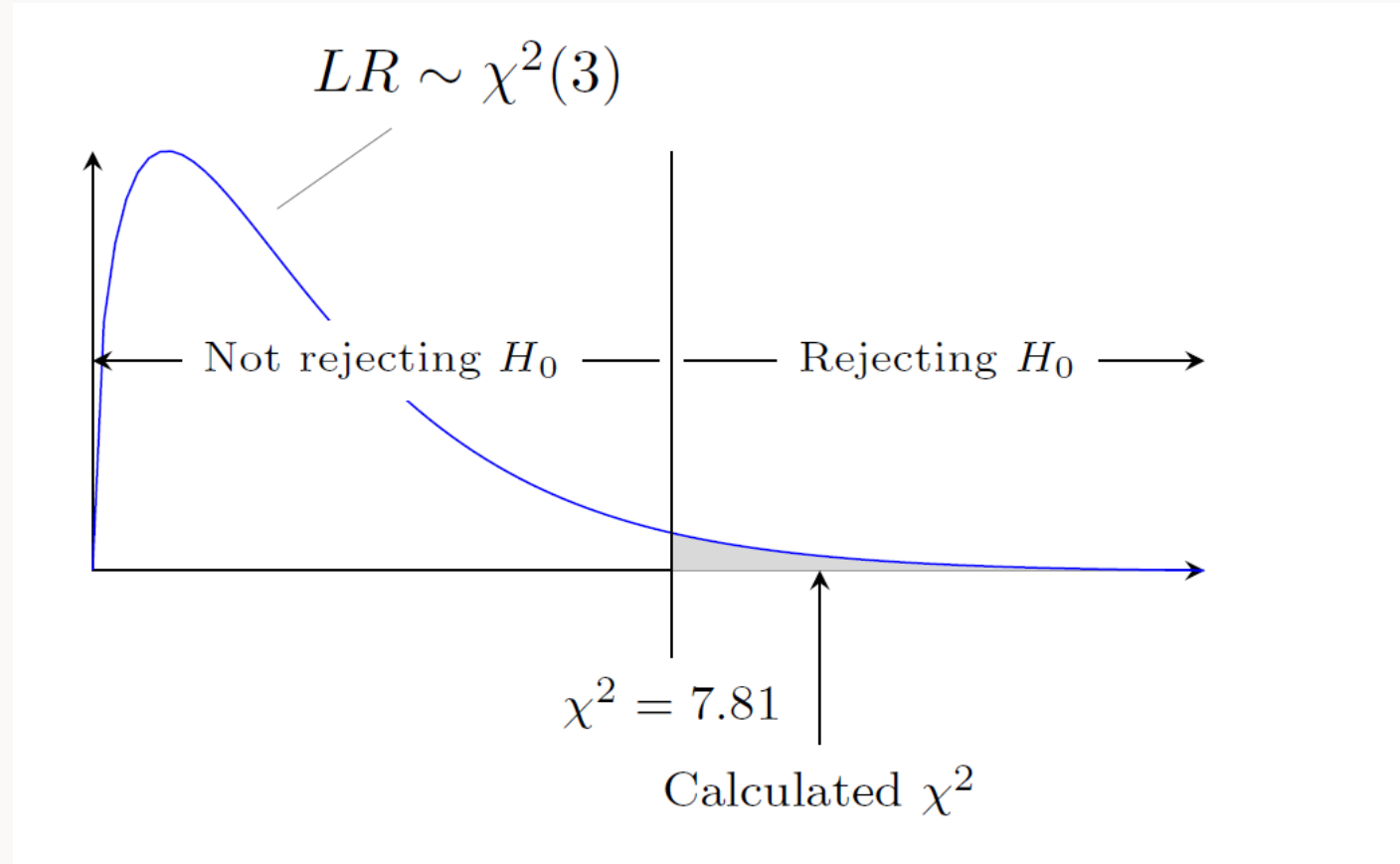
Likelihood-ratio test

To evaluate whether the null hypothesis H_0 can be rejected or not with a confidence of 95 %, we use the $\chi^2(n)$ distribution.

n is the degree of freedom, which corresponds to the difference in number of parameters, which is $n = 4 - 1 = 3$

For a p-value of 0.05, $\chi^2(3)$ corresponds to 7.81.

If the calculated LR for our models is larger than that we can reject H_0 .



Overfitting – AIC and BIC

$$AIC = 2k - 2 \log \hat{\mathcal{L}}$$

k is the number of parameters in the model (including intercept β_0)

$\hat{\mathcal{L}}$ is the maximum likelihood ML estimate of the GLM.

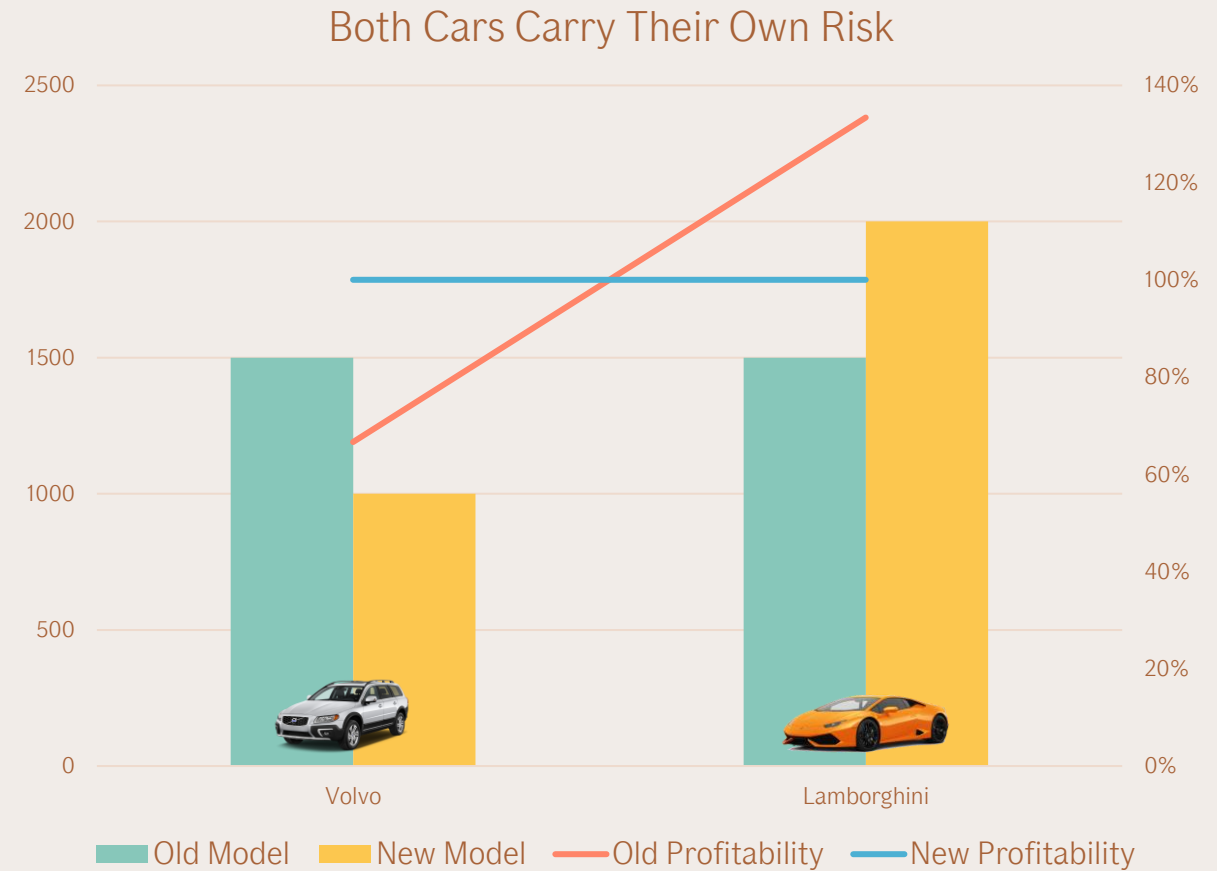
$$BIC = \log n \cdot k - 2 \log \hat{\mathcal{L}}$$

n is the number of observations

Business tests

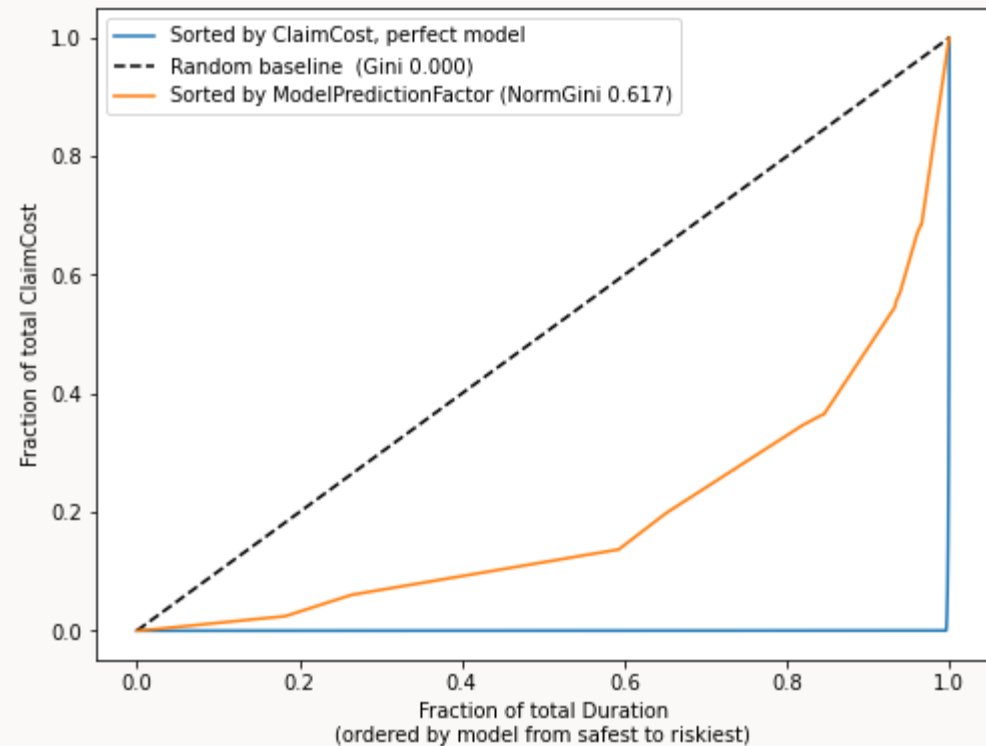
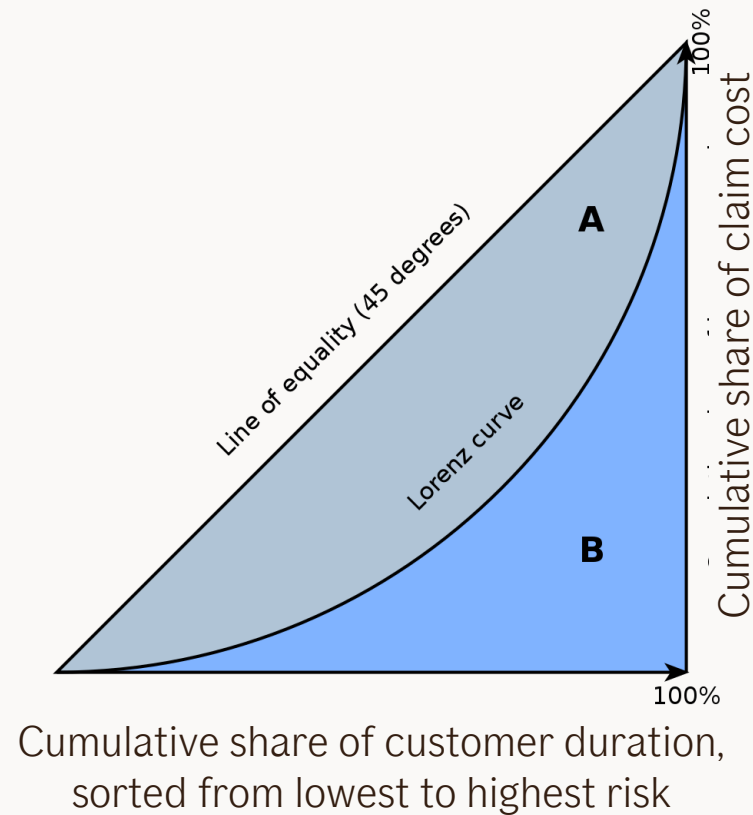
Portfolio Profitability – Is the portfolio correctly priced?

Gini Score – How well do we rank each relative risk



Gini

Gini score is a measure of how well we rank each risk.



$$gini\ score = \frac{\int FlatRate - \int Lorenz}{\int FlatRate - \int LorenzPerfectPrediction} = \frac{A}{A + B}$$

Risk ratio and levelling

$$\text{Risk Ratio} = \frac{\text{Total claim cost}}{\text{Total premium}}$$

$$\text{price} = \gamma_0 \cdot \text{tariff factors}$$

γ_0 is the base level

Risk ratio is a measure that is used on product or portfolio level to find whether the portfolio is priced correctly.

There are often a RR target for a product, and in the project the target is 90 %.

When claim cost modelling for a product is done we need to make sure that we get a good risk ratio, to at least receive enough premium to cover the total claim costs.

To achieve the desired risk ratio, we perform levelling by adjusting the base level γ_0 and apply that on top of the claim cost model.

Summary

Tariff Analysis

Prepare data

Set up design matrix

Calculate β

Risk = Frequency · Severity

Tests

Likelihood-ratio test

Wald test

AIC

BIC

Risk ratio

Gini



