ST4050
Statistical consulting

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

School of Mathematical Sciences
Eolaíochtaí Matamaiticiúla

WE ARE...
THE COLLEGE OF SCIENCE, ENGINEERING & FOOD SCIENCE

WE ARE... STEM @UCC

Module introduction

- **Lectures/practicals:** Tuesdays, 09:00-12:00, WGB 34

Assessment:

- **Individual written report:** 6000 words + images + tables + TOC + cover etc. (70/200 marks, **5pm Sunday 29/03/2026**)
- **Individual interview:** approx. 20-30 minutes (30/200 marks, **date TBD, ≃ last week of course**)

A reminder of the background …

- How are insurance prices made?

- French frequency dataset
  - Why model severity and frequency? Do we have to?
  - What is pseudo/simulated data?

- What is TPL?
  - what factors are useful, and not so useful? *Value* or *power* of car?
  - What other perils are there?

- Risk model vs Pricing model
  - Best estimate, what is that used for?
  - Risk vs. price what's the difference?
  - Restriction process

Overview of the task …

1. Analyse data
2. Review literature
3. fit models (GLM and MLs)
4. Describe methods
5. Analyze results

Act as a consultant for a client. Produce modeling solutions, describe them, compare them.

Note: take note of GI and industry specific points discussed in class. These can be useful and can gain you marks if discussed in reports and presentations in a relevant way!

Let’s start at the end: the report

Sketch of report content

- Cover page
- Declaration
- Abstract
- TOC
- List of tables
- List of figures
- Summary
  - Overview of project, what was achieved. What does the client need to know?

- Introduction
  - Of the problem
  - Motivation and background
  - Point at what will follow in the report

- Goals of the project
  - Replicate work of Noll, Salzmann, Wuthrich 2020, in their tutorial paper
  - Review literature for GLM and ML performance on same dataset
  - Test further developments in GLM methodology vs ML in performance
  - Can GLM beat ML, what further methods/enhancements are required for this?

- Description of data –
  - Data is already well described in paper, so produce a brief high-level analysis and point the reader to the paper for more detail
  - Highlight any new aspects of the data not already shown in the tutorial paper.

- Describe literature –
  - Important step, what GLMs and GLM techniques have been fit on this data, in comparison against ML.
  - What are the performance/accuracy results for GLM vs ML reported in the literature for this dataset, and generally on insurance pricing problems?
  - How consistent is the literature in terms of applying the same data cleaning, and same data partitioning

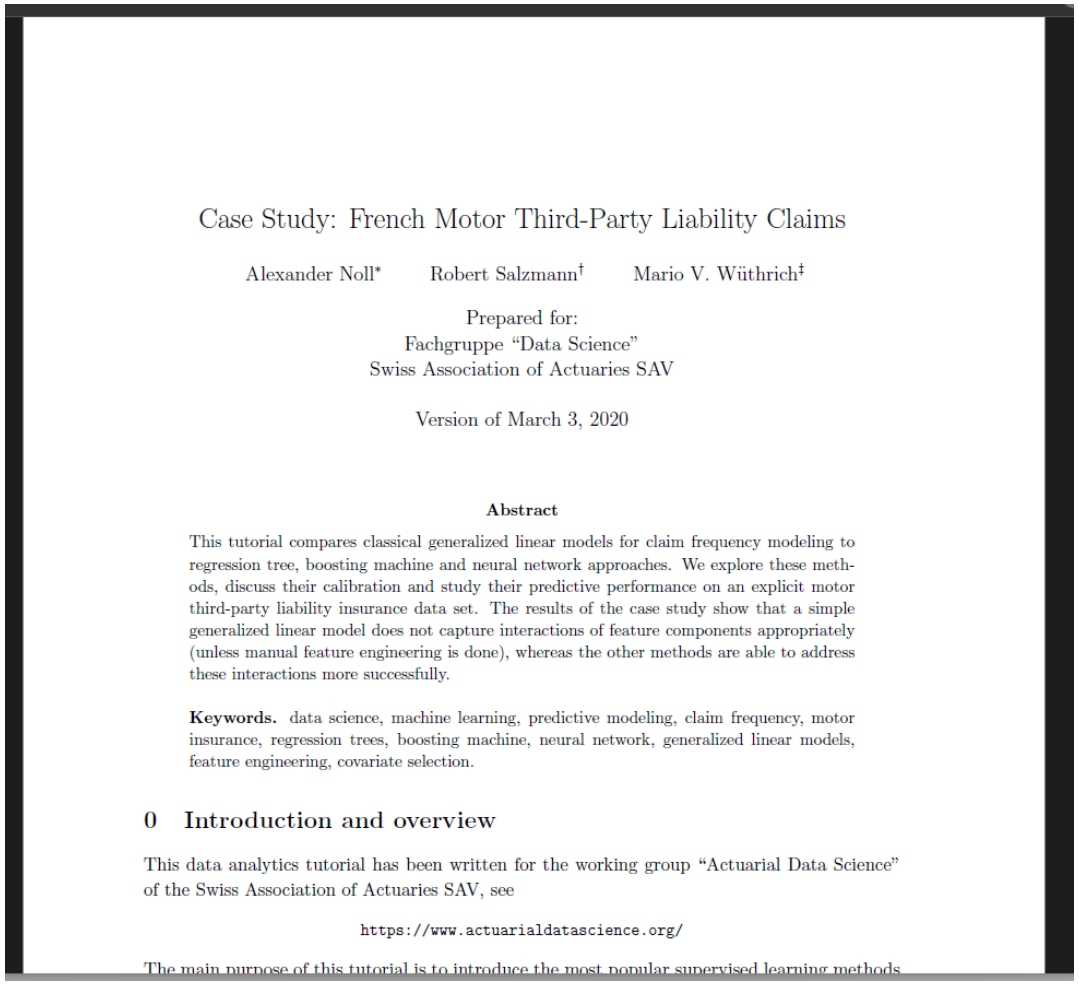
- Modelling:
  - Describe data cleaning and data preparation
  - Describe model architectures and techniques used
  - Scrutinise approach in the tutorial paper

- Describe the results
  - Did you verify published results in the papers?
  - What were the in-sample and out-of-sample performance metrics for each of the models in your study?
  - Describe the breakdown of the benefits of each new step/technique/modelling choice you made

- Discussion
  - Has the comparison of GLM vs ML on this dataset in the literature been fair? Can the GLM compete with ML if more care, more sophistication, is used in the modelling of a GLM?
  - Can the use of an ML model further improve a GLM model?
  - What were the limitations of the work? What are the next steps?

- Conclude
  - What did you get done? Were the goals achieved?

Paper we are focused on …



Case Study: French Motor Third-Party Liability Claims

Alexander Noll* Robert Salzmann† Mario V. Wüthrich‡

Prepared for:
Fachgruppe "Data Science"
Swiss Association of Actuaries SAV

Version of March 3, 2020

Abstract

This tutorial compares classical generalized linear models for claim frequency modeling to regression tree, boosting machine and neural network approaches. We explore these methods, discuss their calibration and study their predictive performance on an explicit motor third-party liability insurance data set. The results of the case study show that a simple generalized linear model does not capture interactions of feature components appropriately (unless manual feature engineering is done), whereas the other methods are able to address these interactions more successfully.

Keywords. data science, machine learning, predictive modeling, claim frequency, motor insurance, regression trees, boosting machine, neural network, generalized linear models, feature engineering, covariate selection.

0 Introduction and overview

This data analytics tutorial has been written for the working group "Actuarial Data Science" of the Swiss Association of Actuaries SAV, see

https://www.actuarialdatascience.org/

The main purpose of this tutorial is to introduce the most popular supervised learning methods

In canvas:

Case Study French Motor Third-Party Liability Claims.pdf

High level view

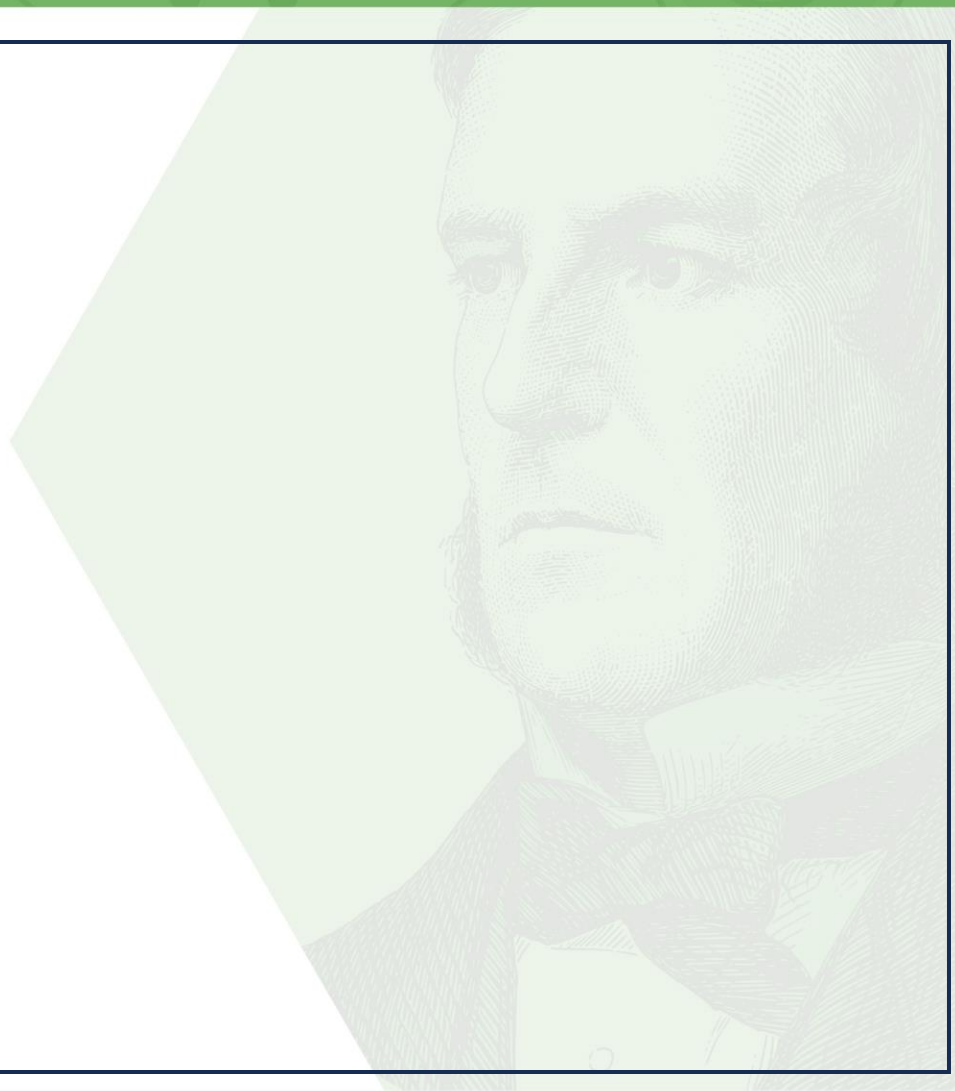
- Is the comparison between GLM and ML in insurance motor pricing literature fair? In particular, on the (standard) French cars dataset, is enough care and skill applied to GLM methods in order to make the comparison against ML a fair one? Using the Tutorial Paper as base, what changes and enhancements can be applied to their GLM to make it more competitive with the ML models?

1. Read the tutorial paper
2. Replicate results
3. Review literature – how have GLMs compared against ML on this dataset? What methods were applied to GLMs in comparison against ML? Were ML models also optimally fitted in literature?
4. Enhance the GLM model in tutorial paper and compare against ML results. Further check if ML model itself has been optimally fit.
5. Report on findings in performance of GLM vs ML.

Consulting: report writing

Report tidy up: have you checked these …

- Labelled all charts and axes, and all tables
- Included a caption, and figure number, for all visuals
- Included page numbers
- Table of contents, list of figures
- Abstract
- Declaration
- Appendix
- References
- Conclusions
- Description of software used, e.g., RStudio version x.xx
- List of abbreviations, GLM, ML, GBM etc.
- Your name, module code, student number, date



Assessment

REPORT 70%

Guideline breakdown of marks:

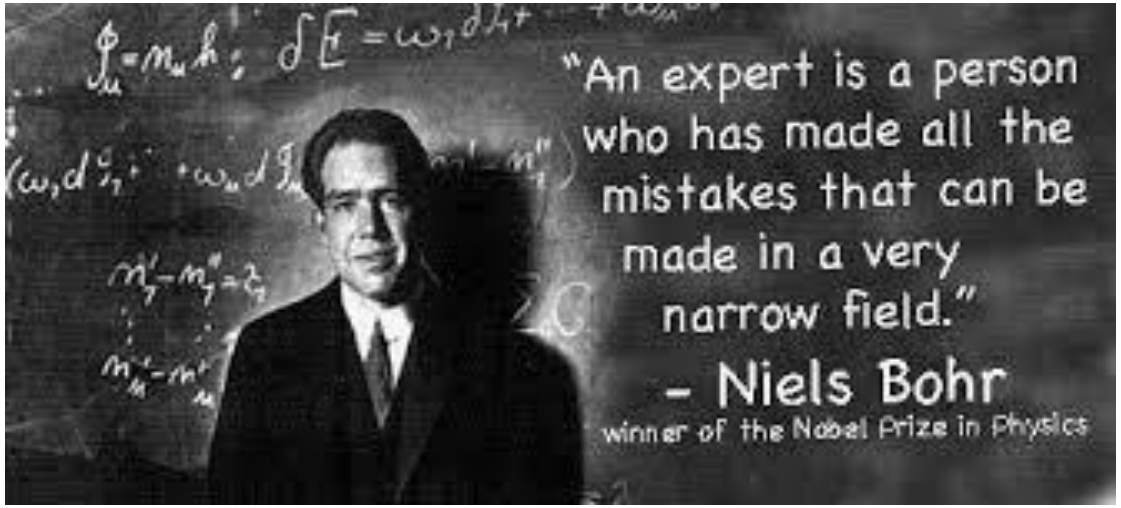
- Introduction and Motivation 10%
- Data description 10 %
- Methodologies 30%
- Results 30%
- Discussion and Conclusion 20%

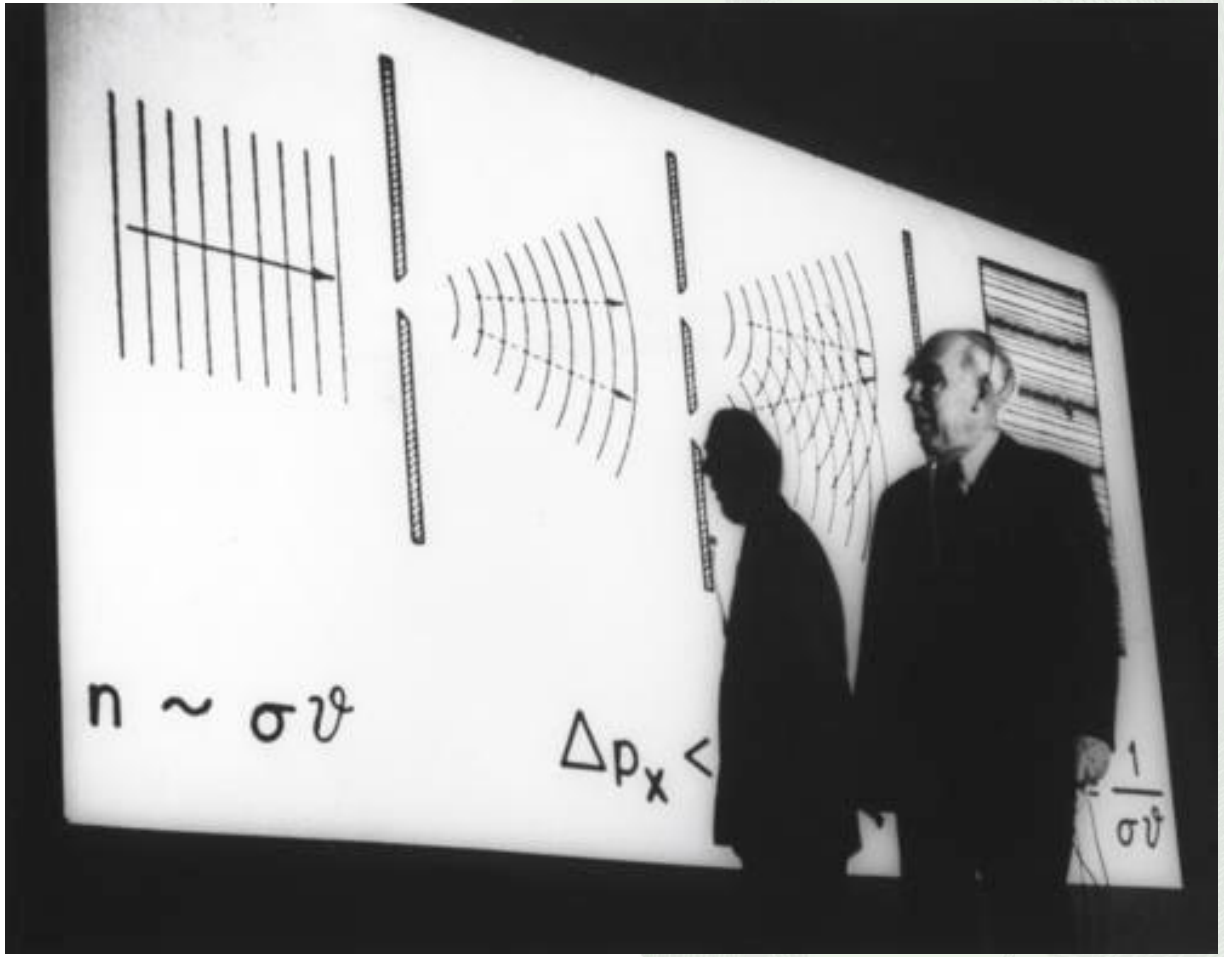
INTERVIEW 30%

- Answering questions on your project work, going through the report, code, methods and results

Time will be tight.

Get started on the report.


"An expert is a person who has made all the mistakes that can be made in a very narrow field."
- Niels Bohr
winner of the Nobel Prize in Physics


$n \sim \sigma v$
$\Delta p_x <$
$\frac{1}{\sigma v}$

A little about frequency

- How is it different from severity?
- Frequency and making the full price: severity, frequency, for each peril. How are propensity models used?
- What would be the most important frequency factors?
- Inflation and time factor – fitting the time control factor.
- Development of claim amounts by reserving – not using very undeveloped claims
- Frequency vs severity, which drives risk more?

Data EDA and preparation

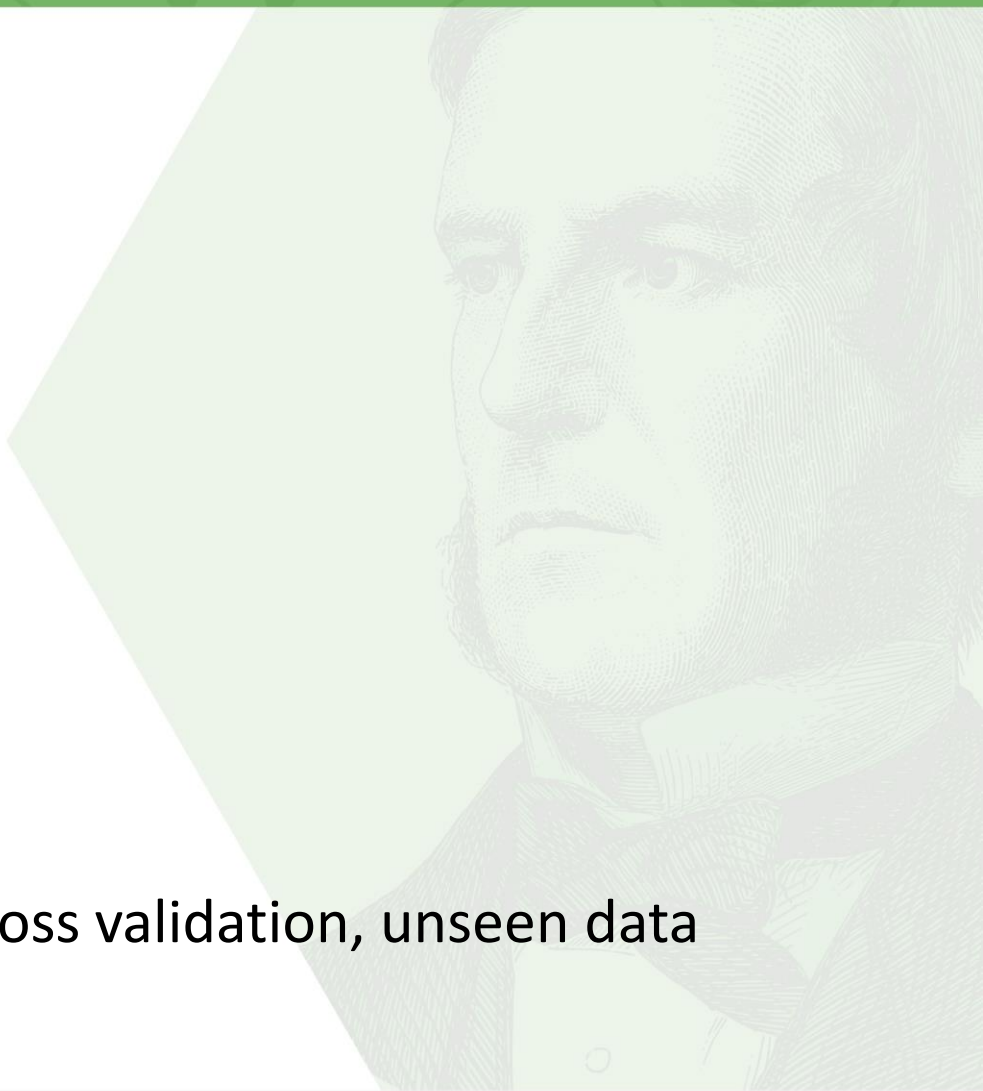
Typically, these questions arise:

- How was the final data made
- Some policies in severity, but not in frequency – what do we notice about these?
- Some policies in frequency, but not in severity – is there a difference?
- Remove large claims, outliers
- Check for duplicates
- Missings? How could you deal with these?

- Remove exposure >1. Should we remove low exposure policies?
- Expand on data description in Tutorial Paper – what extra analysis can you add?
- Frequency tables, scatter plots, 2-way tables between variables to show associations
- Correlation – show the scatter plot!

The language of modelling

- Claim severity, claim count
- Peril
- Loss ratio
- Fitting a model
- Scoring a model
- Relativities
- Parameters
- Factors
- levels
- Train, test, validation, modelling data, holdout data, cross validation, unseen data **(we will use these terms: train / validation / test)**



GLM factor types

Simple factors

- Ungrouped categorical factors
- Easy to model
- In most cases though, grouping is needed
- Example of a factor that would be left as a simple factor?

Custom factors

- A grouped simple factor
- Go through a robust grouping process in making these
- Re-examine groups throughout the fitting process

Variates

- Will likely need pre-processing: capping at low/high levels; testing curves with higher orders of polynomials
- When would you use these?

Interactions

- There are different types of interactions, which we will cover
- Carefully examine the evidence before including an interaction. You're complicating the model, so make sure the interaction is worth it!
- Best test: partition data; fit interactions; check for consistent relativities.
- Consider creating a **compound factor** in your data first.

GLM factor types

Compound factor – handy for interactions!

Example: interact car colour and occupation

… navy_plumber, silver_nurse, etc.

… dark_professional, bright_trades

Interactions

- There are different types of interactions, which we will cover
- Carefully examine the evidence before including an interaction. You're complicating the model, so make sure the interaction is worth it!
- Consider creating a **compound factor** in your data first.

Unrefined GLM – severity model

These factors were found to be predictive…

R 4.3.2 · C:/Users/JohnCondon/OneDrive/John Condon/modules/ST4050/teaching/

```
                                          Estimate Std. Error t value Pr(>|t|)
(Intercept)                              6.913e+00  1.613e-01  42.866  < 2e-16 ***
DrivAgeGLM_grouped_1_2                  -1.303e-01  5.055e-02  -2.579  0.00993 **
DrivAgeGLM_grouped_1_3                  -1.294e-01  5.207e-02  -2.485  0.01298 *
DrivAgeGLM_grouped_1_4                  -8.638e-02  5.295e-02  -1.631  0.10286
DrivAgeGLM_grouped_1_5                  -8.606e-02  5.425e-02  -1.586  0.11267
DrivAgeGLM_grouped_1_6                  -8.585e-02  5.822e-02  -1.474  0.14038
DrivAgeGLM_grouped_1_7                  -2.341e-02  6.376e-02  -0.367  0.71348
DrivAgeGLM_grouped_1_8                  -1.846e-01  9.918e-02  -1.861  0.06271 .
RegionGLMAquitaine                       2.334e-01  1.474e-01   1.583  0.11344
RegionGLMAuvergne                        2.209e-01  1.787e-01   1.236  0.21654
RegionGLMBasse-Normandie                 8.815e-02  1.560e-01   0.565  0.57216
RegionGLMBourgogne                       2.768e-01  1.591e-01   1.740  0.08186 .
RegionGLMBretagne                        1.981e-01  1.453e-01   1.364  0.17272
RegionGLMCentre                          2.121e-01  1.431e-01   1.482  0.13831
RegionGLMChampagne-Ardenne               2.579e-01  2.070e-01   1.246  0.21293
RegionGLMCorse                           4.973e-01  1.835e-01   2.711  0.00672 **
RegionGLMFranche-Comte                   2.884e-01  2.631e-01   1.096  0.27303
RegionGLMHaute-Normandie                 1.845e-01  1.688e-01   1.093  0.27442
RegionGLMIle-de-France                   2.195e-01  1.455e-01   1.508  0.13155
RegionGLMLanguedoc-Roussillon            2.549e-01  1.482e-01   1.720  0.08550 .
RegionGLMLimousin                        1.987e-01  1.727e-01   1.150  0.25011
RegionGLMMidi-Pyrenees                   2.598e-01  1.595e-01   1.629  0.10324
RegionGLMNord-Pas-de-Calais              2.488e-01  1.465e-01   1.699  0.08936 .
RegionGLMPays-de-la-Loire                1.818e-01  1.458e-01   1.247  0.21228
RegionGLMPicardie                        2.608e-01  1.604e-01   1.626  0.10403
RegionGLMPoitou-Charentes                1.922e-01  1.495e-01   1.285  0.19867
RegionGLMProvence-Alpes-Cotes-D'Azur     3.363e-01  1.439e-01   2.338  0.01943 *
RegionGLMRhone-Alpes                     2.240e-01  1.433e-01   1.563  0.11798
Density                                  7.824e-06  5.727e-06   1.366  0.17190
BonusMalus                               2.153e-03  5.117e-04   4.209 2.59e-05 ***
VehPower                                 4.130e-03  4.442e-03   0.930  0.35247
VehBrandGLM_B10                          1.257e-01  5.500e-02   2.286  0.02226 *
VehBrandGLM_B11                          5.761e-03  5.675e-02   0.102  0.91913
VehBrandGLM_B12                          3.199e-02  2.970e-02   1.077  0.28145
VehBrandGLM_B13                          8.466e-02  6.294e-02   1.345  0.17860
VehBrandGLM_B14                         -8.125e-03  1.197e-01  -0.068  0.94588
VehBrandGLM_B2                           5.489e-02  2.321e-02   2.364  0.01807 *
VehBrandGLM_B3                           3.264e-02  3.271e-02   0.998  0.31836
VehBrandGLM_B4                           7.231e-04  4.454e-02   0.016  0.98705
VehBrandGLM_B5                          -2.572e-03  3.716e-02  -0.069  0.94481
VehBrandGLM_B6                          -5.237e-02  4.174e-02  -1.255  0.20968
VehAge                                  -8.495e-03  1.813e-03  -4.685 2.83e-06 ***
AreaGLM_B                                1.063e-02  3.561e-02   0.298  0.76536
AreaGLM_C                               -1.949e-02  2.953e-02  -0.660  0.50926
AreaGLM_D                                3.451e-02  3.131e-02   1.102  0.27046
AreaGLM_E                                1.432e-02  4.047e-02   0.354  0.72337
AreaGLM_F                               -2.709e-01  1.399e-01  -1.936  0.05288 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.8813564)

    Null deviance: 9214.7  on 12658  degrees of freedom
Residual deviance: 9089.7  on 12612  degrees of freedom
AIC: 206278
```

```
> print(coefficients_table)
                                          Variable  Estimate LowerValue UpperValue p_values
(Intercept)                              (Intercept) 1005.6421   728.3790  1388.4476     0.000
DrivAgeGLM_grouped_1_2                DrivAgeGLM_grouped_1_2    0.8778     0.7934     0.9712     0.010
DrivAgeGLM_grouped_1_3                DrivAgeGLM_grouped_1_3    0.8786     0.7917     0.9751     0.013
DrivAgeGLM_grouped_1_4                DrivAgeGLM_grouped_1_4    0.9173     0.8251     1.0197     0.103
DrivAgeGLM_grouped_1_5                DrivAgeGLM_grouped_1_5    0.9175     0.8232     1.0227     0.113
DrivAgeGLM_grouped_1_6                DrivAgeGLM_grouped_1_6    0.9177     0.8169     1.0311     0.140
DrivAgeGLM_grouped_1_7                DrivAgeGLM_grouped_1_7    0.9769     0.8599     1.1097     0.713
DrivAgeGLM_grouped_1_8                DrivAgeGLM_grouped_1_8    0.8314     0.6818     1.0138     0.063
RegionGLMAquitaine                        RegionGLMAquitaine    1.2629     0.9404     1.6960     0.113
RegionGLMAuvergne                          RegionGLMAuvergne    1.2472     0.8723     1.7831     0.217
RegionGLMBasse-Normandie                RegionGLMBasse-Normandie    1.0921     0.7994     1.4922     0.572
RegionGLMBourgogne                        RegionGLMBourgogne    1.3189     0.9595     1.8129     0.082
RegionGLMBretagne                          RegionGLMBretagne    1.2190     0.9117     1.6300     0.173
RegionGLMCentre                              RegionGLMCentre    1.2363     0.9286     1.6460     0.138
RegionGLMChampagne-Ardenne          RegionGLMChampagne-Ardenne    1.2942     0.8554     1.9581     0.213
RegionGLMCorse                              RegionGLMCorse    1.6443     1.1393     2.3732     0.007
RegionGLMFranche-Comte                  RegionGLMFranche-Comte    1.3343     0.7883     2.2585     0.273
RegionGLMHaute-Normandie                RegionGLMHaute-Normandie    1.2026     0.8581     1.6855     0.274
RegionGLMIle-de-France                  RegionGLMIle-de-France    1.2454     0.9309     1.6662     0.132
RegionGLMLanguedoc-Roussillon      RegionGLMLanguedoc-Roussillon    1.2904     0.9593     1.7358     0.086
RegionGLMLimousin                        RegionGLMLimousin    1.2198     0.8635     1.7231     0.250
RegionGLMMidi-Pyrenees                  RegionGLMMidi-Pyrenees    1.2967     0.9426     1.7838     0.103
RegionGLMNord-Pas-de-Calais        RegionGLMNord-Pas-de-Calais    1.2825     0.9569     1.7190     0.089
RegionGLMPays-de-la-Loire          RegionGLMPays-de-la-Loire    1.1994     0.8961     1.6055     0.212
RegionGLMPicardie                        RegionGLMPicardie    1.2980     0.9417     1.7889     0.104
RegionGLMPoitou-Charentes          RegionGLMPoitou-Charentes    1.2119     0.8987     1.6343     0.199
RegionGLMProvence-Alpes-Cotes-D'Azur RegionGLMProvence-Alpes-Cotes-D'Azur    1.3997     1.0498     1.8664     0.019
RegionGLMRhone-Alpes                    RegionGLMRhone-Alpes    1.2511     0.9394     1.6663     0.118
Density                                          Density    1.0000     1.0000     1.0000     0.172
BonusMalus                                    BonusMalus    1.0022     1.0011     1.0032     0.000
VehPower                                      VehPower    1.0041     0.9953     1.0131     0.352
VehBrandGLM_B10                        VehBrandGLM_B10    1.1340     1.0159     1.2658     0.022
VehBrandGLM_B11                        VehBrandGLM_B11    1.0058     0.8979     1.1267     0.919
VehBrandGLM_B12                        VehBrandGLM_B12    1.0325     0.9730     1.0957     0.281
VehBrandGLM_B13                        VehBrandGLM_B13    1.0884     0.9596     1.2343     0.179
VehBrandGLM_B14                        VehBrandGLM_B14    0.9919     0.7808     1.2602     0.946
VehBrandGLM_B2                          VehBrandGLM_B2    1.0564     1.0085     1.1066     0.018
VehBrandGLM_B3                          VehBrandGLM_B3    1.0332     0.9678     1.1030     0.318
VehBrandGLM_B4                          VehBrandGLM_B4    1.0007     0.9154     1.0940     0.987
VehBrandGLM_B5                          VehBrandGLM_B5    0.9974     0.9260     1.0744     0.945
VehBrandGLM_B6                          VehBrandGLM_B6    0.9490     0.8730     1.0316     0.210
VehAge                                          VehAge    0.9915     0.9880     0.9951     0.000
AreaGLM_B                                    AreaGLM_B    1.0107     0.9412     1.0853     0.765
AreaGLM_C                                    AreaGLM_C    0.9807     0.9244     1.0404     0.509
AreaGLM_D                                    AreaGLM_D    1.0351     0.9723     1.1020     0.270
AreaGLM_E                                    AreaGLM_E    1.0144     0.9356     1.0999     0.723
AreaGLM_F                                    AreaGLM_F    0.7627     0.5765     1.0090     0.053
>
```

Presenting a more refined GLM – severity model

```
461
462
463  # ---------------- make the GLM model ---------------
464
465  gamma_glm <- glm(ClaimAmount ~
466
467  CappedDrivAgeGLM_LE21
468  +RegionGLM_coded
469  +CappedDensityGLM
470  +BonusMalus_grouped_
471  +CappedVehPower
472  +VehBrandGLM_coded
473  +CappedVehAgeGLM_v2
474  ,
475  data = severity_data_3_GLM_train,
476  family = Gamma(link = "log"))
477
478
479  # ---------------- AIC comparison ---------------
480  new_AIC  <- AIC(gamma_glm)
481  #AIC_change = new_AIC - previous_AIC
482  #AIC_change
483
484
```

```
Call:
glm(formula = ClaimAmount ~ CappedDrivAgeGLM_LE21 + RegionGLM_coded +
    CappedDensityGLM + BonusMalus_grouped_ + CappedVehPower +
    VehBrandGLM_coded + CappedVehAgeGLM_v2, family = Gamma(link = "log"),
    data = severity_data_3_GLM_train)

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              7.557174   0.092783  81.450  < 2e-16 ***
CappedDrivAgeGLM_LE21   -0.079004   0.027268  -2.897  0.00377 **
RegionGLM_codedHigh      0.127171   0.021396   5.944 2.83e-09 ***
CappedDensityGLM        -0.007403   0.003057  -2.421  0.01547 *
BonusMalus_grouped_high  0.074928   0.015960   4.695 2.69e-06 ***
CappedVehPower           0.009596   0.003689   2.601  0.00930 **
VehBrandGLM_codedlow    -0.062898   0.018869  -3.333  0.00086 ***
CappedVehAgeGLM_v2      -0.019389   0.002345  -8.268  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for Gamma family taken to be 0.8929822)

    Null deviance: 13957  on 19087  degrees of freedom
Residual deviance: 13814  on 19080  degrees of freedom
AIC: 311071

Number of Fisher Scoring iterations: 5

> summary_result <- summary(gamma_glm)
>
```

Presenting a more refined GLM– severity model

```
> # Print the coefficients table
> print(coefficients_table)
                                    Variable  Estimate LowerValue UpperValue p_values
(Intercept)                      (Intercept) 1914.4279  1590.1893  2304.7786    0.000
CappedDrivAgeGLM_LE21    CappedDrivAgeGLM_LE21    0.9240     0.8750     0.9758    0.004
RegionGLM_codedHigh        RegionGLM_codedHigh    1.1356     1.0880     1.1853    0.000
CappedDensityGLM              CappedDensityGLM    0.9926     0.9866     0.9987    0.015
BonusMalus_grouped_high BonusMalus_grouped_high    1.0778     1.0439     1.1128    0.000
CappedVehPower                  CappedVehPower    1.0096     1.0022     1.0171    0.009
VehBrandGLM_codedlow      VehBrandGLM_codedlow    0.9390     0.9043     0.9752    0.001
CappedVehAgeGLM_v2          CappedVehAgeGLM_v2    0.9808     0.9762     0.9854    0.000
>
```

How to present a GLM model – factors and parameters

- Factors, levels and parameters

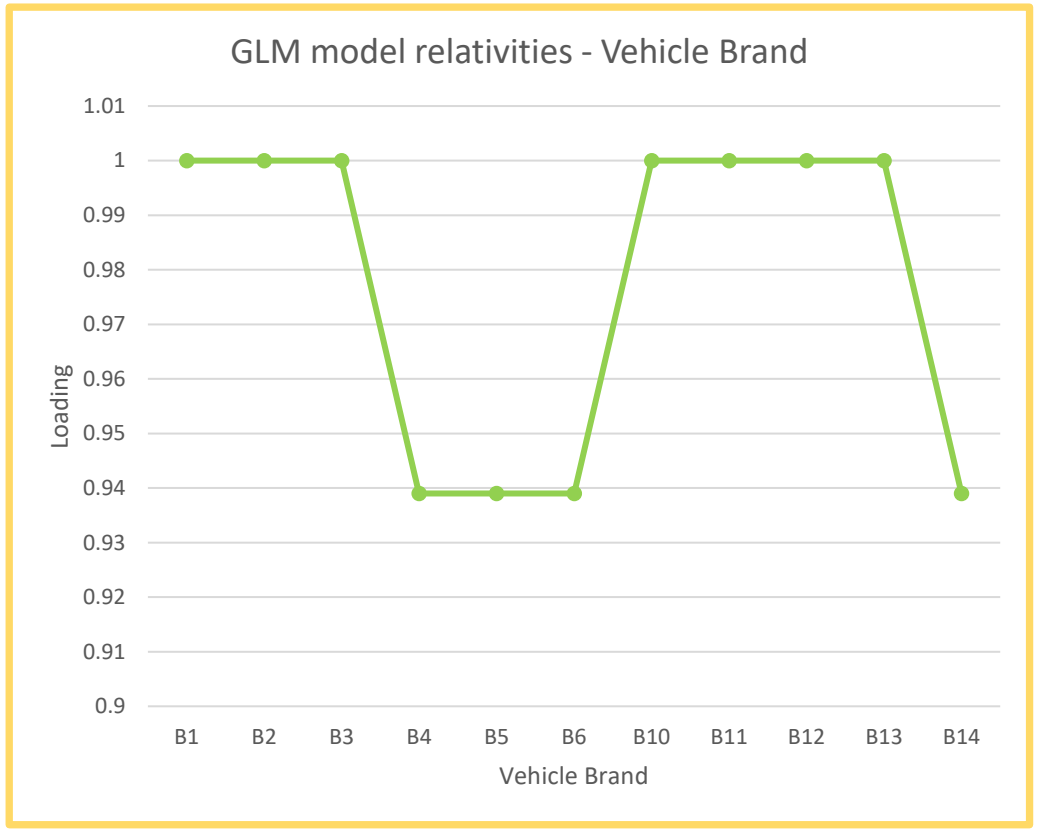
- New factor added – what changes in DOF? What changes in no. of parameters?
  - For categorical factors and for variates

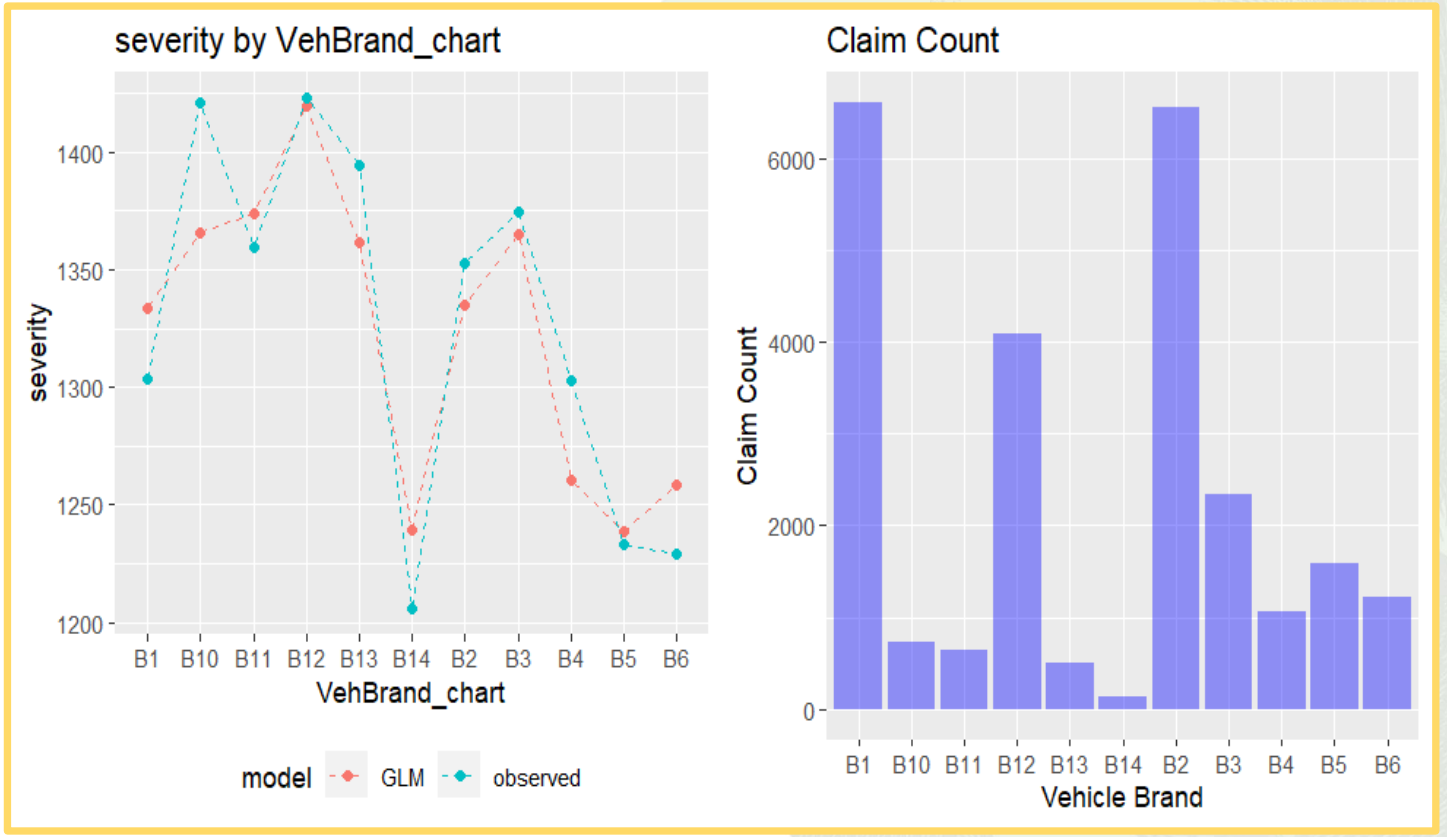
- Grouping parameters - how to group, when to group, and grouping at different stages of the modelling process.
  - group levels that are within the same confidence intervals. What are the issues with this?
  - group levels that have similar relativities and represent similar groups. What are the issues with this?
  - Watch out for small parameters, and very large parameters.
  - Consider the importance of *when* you group levels, not just *what* you group

Look for consistency!

- Interacting a factor with a random number, and checking for consistency in relativities is a common technique applied in industry. Equivalently, partition your train data, and check for consistency in relativities after fitting.

How to present a GLM model – factor by factor





Fitted vs Observed

Fitted vs observed, CA vs OBS

- for factors included
- Not included
- What happens when we add a factor? Adding independent vs related factors?
- The advantage of new factors?

Restrictions and fitted vs OBS

Remove a factor, what happens in the fitted vs observed chart?

Restrictions for protected characteristics, how could you reduce the potential to rate on a protected characteristic? (unfit it? Or other options?)

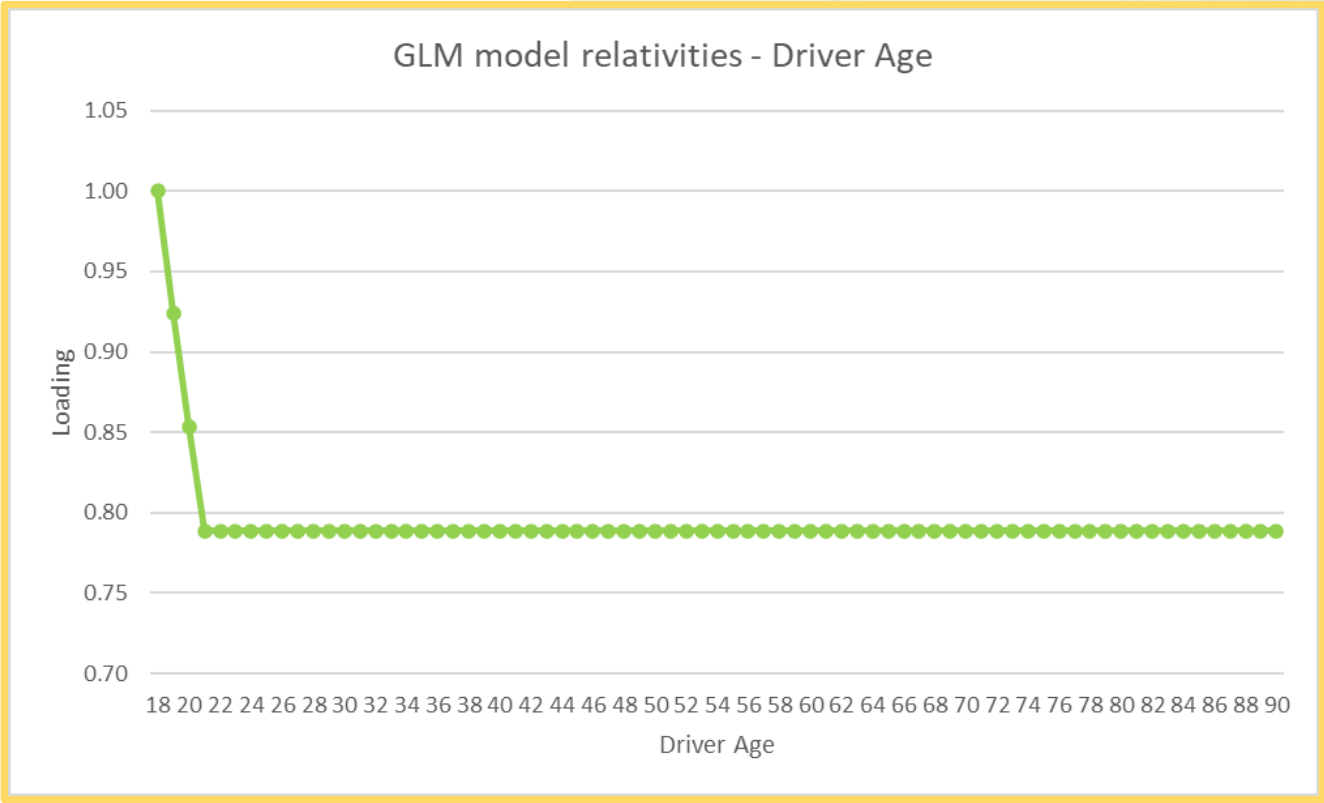
What is happening the obs vs fitted chart during the modelling process? What would you check for at the end of the modelling process?

How to present a GLM model – a variate

```
severity_data_3$CappedDrivAgeGLM_LE40 <- pmin(severity_data_3$DrivAge, 40)  # flat
severity_data_3$CappedDrivAgeGLM_LE21 <- pmin(severity_data_3$DrivAge-18, 3)
severity_data_3$CappedDrivAgeGLM_GE40 <- pmax(severity_data_3$DrivAge, 40)  # flat
```

```
                                Variable  Estimate L
(Intercept)                   (Intercept) 1914.4279
CappedDrivAgeGLM_LE21   CappedDrivAgeGLM_LE21    0.9240
RegionGLM_codedHigh       RegionGLM_codedHigh    1.1356
```

Driver age	Pre-processed	Loading
18	0	0.9240^0 = 1
19	1	0.9240^1 = 0.9240
20	2	0.9240^2 = 0.8538
21	3	0.9240^3 = 0.7889
22	3	0.9240^3 = 0.7889



How to present a GLM model – overall

Vehicle Brand	Loading
B1	1
B2	1
B3	1
B4	0.939
B5	0.939
B6	0.939
B10	1
B11	1
B12	1
B13	1
B14	0.939

Region	Loading
Picardie	1
Ile-de-France	1
Nord-Pas-de-Calais	1
Midi-Pyrenees	1
Languedoc-Roussillon	1
Limousin	1
Poitou-Charentes	1
Aquitaine	1
Bretagne	1
Rhone-Alpes	1
Provence-Alpes-Cotes-D'Azur	1.1356
Auvergne	1
Corse	1.1356
Centre	1
Basse-Normandie	1
Pays-de-la-Loire	1
Bourgogne	1
Franche-Comte	1
Champagne-Ardenne	1
Haute-Normandie	1
Alsace	1

… and so on for all other factors.

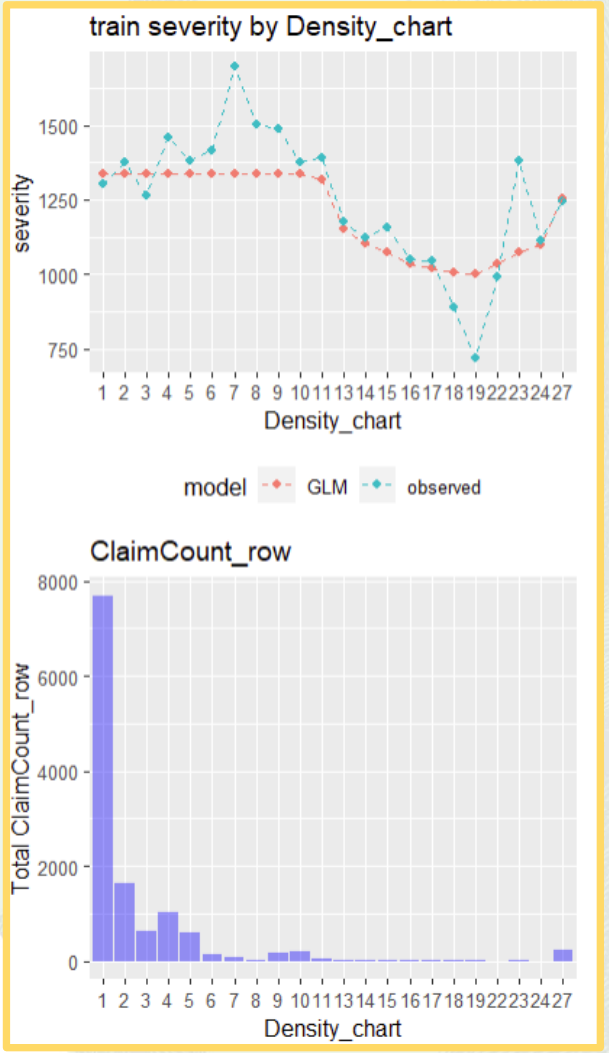
Look at an example calculation…

Explaining variates

```
severity_data_3$CappedDensityGLM = as.numeric(pmax(pmin(0.001*severity_data_3$Density, 26), 10))
```

```
> coefficients_table_1
                                Variable  Estimate LowerValue UpperValue p_values
(Intercept)                   (Intercept) 3950.6557  1669.6133  9348.0812    0.000
CappedDensityGLM         CappedDensityGLM    0.8621     0.7649     0.9717    0.013
CappedDensityGLM_pwr2 CappedDensityGLM_pwr2    1.0040     1.0006     1.0074    0.018
> AIC_model  <- AIC(gamma_glm_train)
> round(AIC_model,0)
[1] 206372
>
```

- For a unit increase in “CappedDensiyGLM”, severity is scaled by 0.8621
- For a unit increase in “CappedDensiyGLM_pwr2”, severity is scaled by 1.0040
- Make a visual description



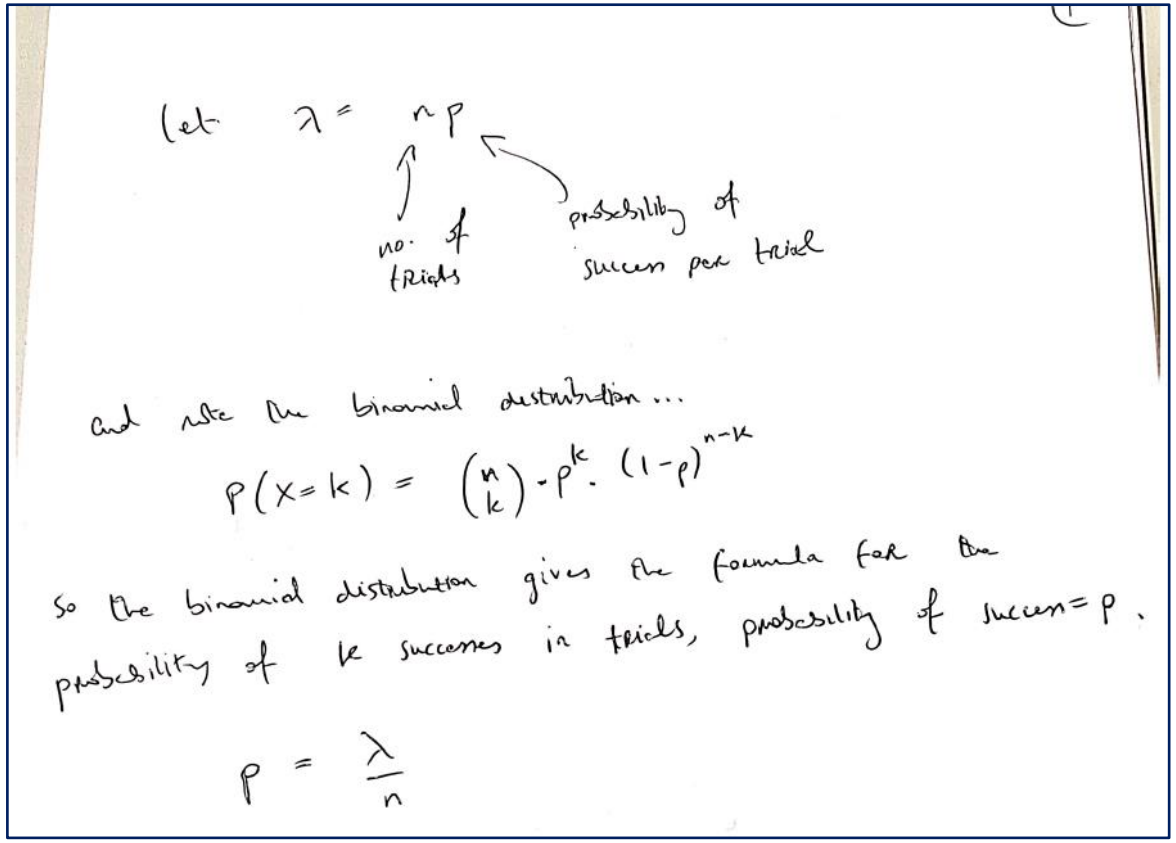
GLM – recap points

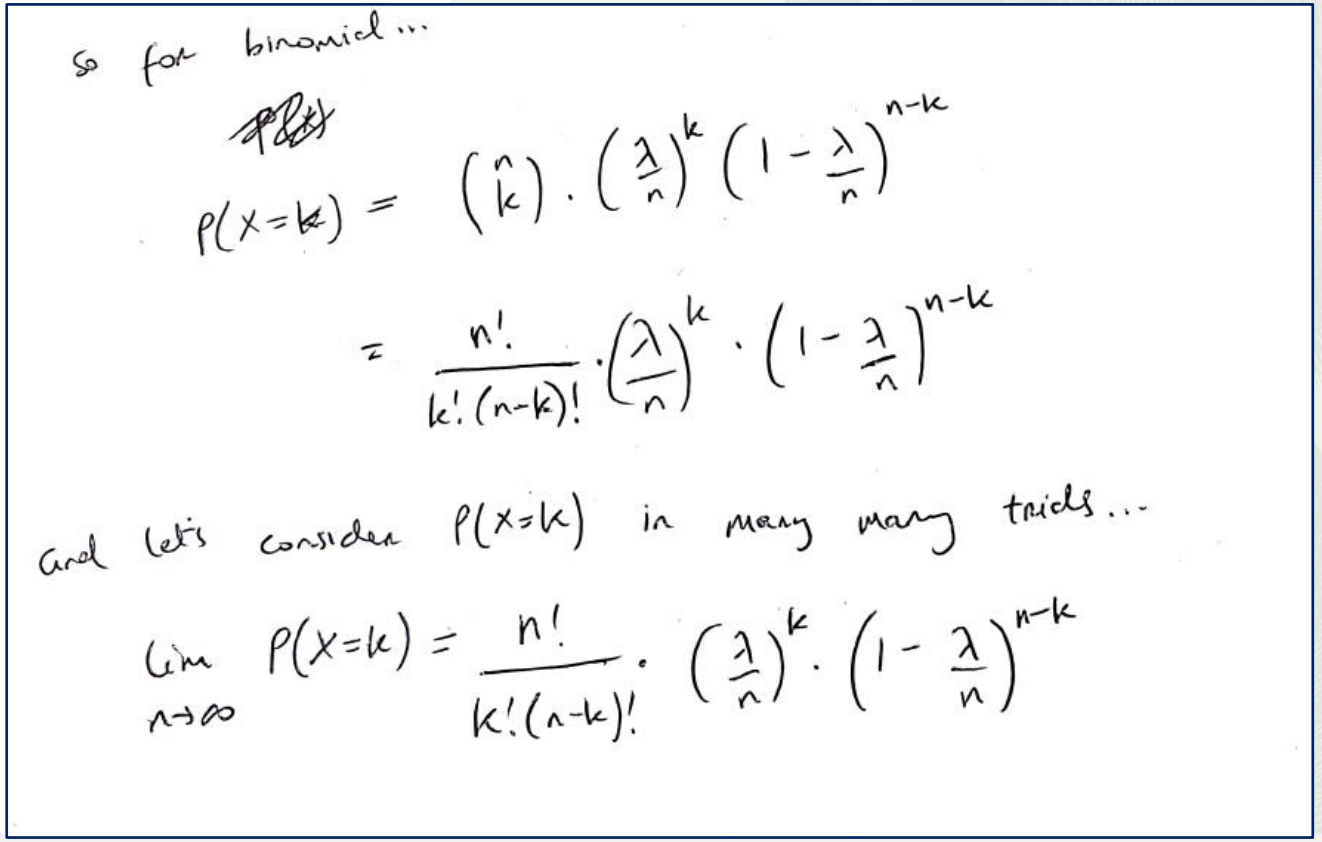
The example calculation with a GLM model:

model factor	IDpol example 1234	model loading
intercept	NA	1900.15
area	A	x 1.20
region	Alsace	x 0.95
driver age	45	x 1.24
vehicle age	5	x 1.04
vehicle power	6	x 0.85
claim severity prediction		**= €2,374.47**

Poisson binomial connection

Consider the count of claims in an insurance policy to follow a binomial distribution. The binomial distribution gives the probability for the number of successes (k) in a number of trials (n). Consider each brief passing moment, microsecond, of the policy as being a “trial”, in a which a claim (or “success”) can happen. The probability of claim in each brief passing moment will be very small, and given by p. As the number of trials, n, in the policy reaches infinity, then the number of claims from the binomial distribution follows a Poisson process.





Poisson binomial connection

(2)

$$\Rightarrow \lim_{n\to\infty} P(X=k) = \lim_{n\to\infty} \frac{\lambda^k}{k!} \cdot \frac{n!}{(n-k)!} \cdot \frac{1}{n^k} \cdot \left(1-\frac{\lambda}{n}\right)^{n-k}$$

$$= \frac{\lambda^k}{k!} \cdot \lim_{n\to\infty} \underbrace{\frac{n!}{(n-k)!} \cdot \frac{1}{n^k}}_{A} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^{n}}_{B} \cdot \underbrace{\left(1-\frac{\lambda}{n}\right)^{-k}}_{C}$$

$$= \frac{\lambda^k}{k!} \cdot \lim_{n\to\infty} A \cdot B \cdot C$$

(A)

$$\lim_{n\to\infty} \frac{n!}{(n-k)!} \cdot \frac{1}{n^k}$$

$$= \lim_{n\to\infty} \frac{(n)(n-1)(n-2)\cdots(n-k)(n-k-1)\cdots(1)}{(n-k)(n-k-1)\cdots(1)} \cdot \frac{1}{n^k}$$

- n terms on top.
- we cancelled out n-k terms
- we are left with k terms

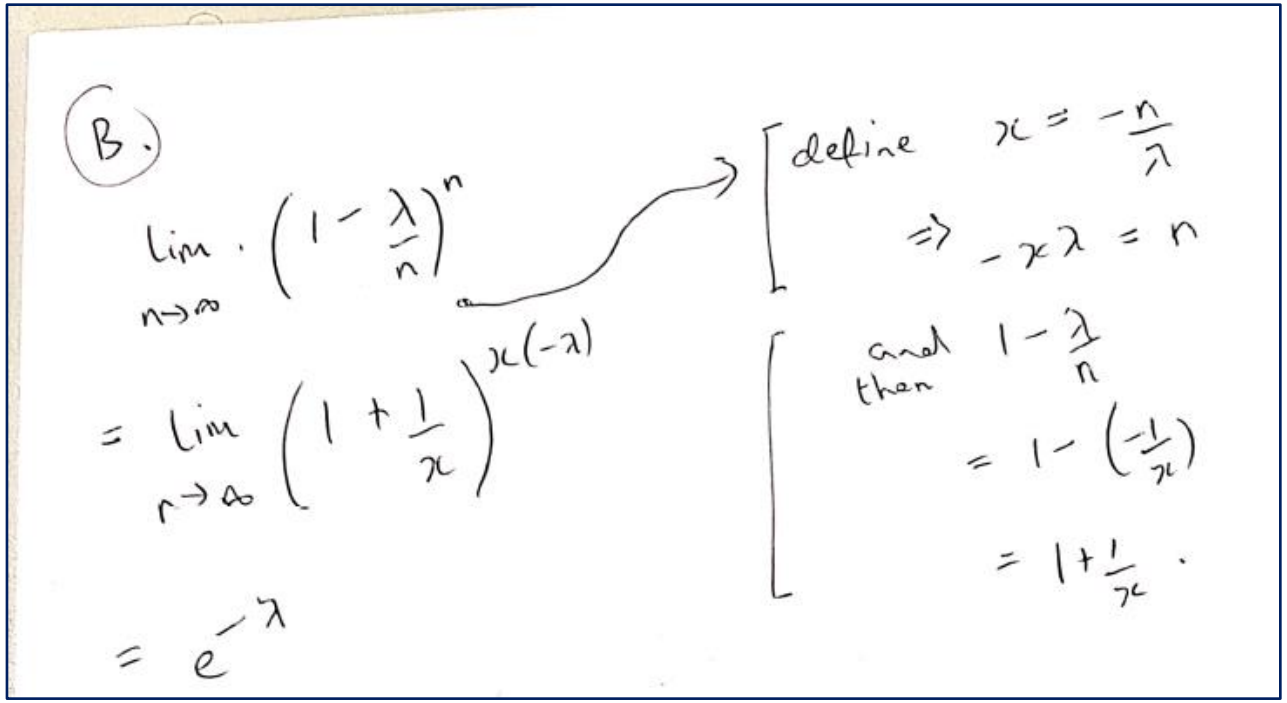
$$= \lim_{n\to\infty} \frac{(n)(n-1)(n-2)\cdots(n-k+1)}{n^k}$$

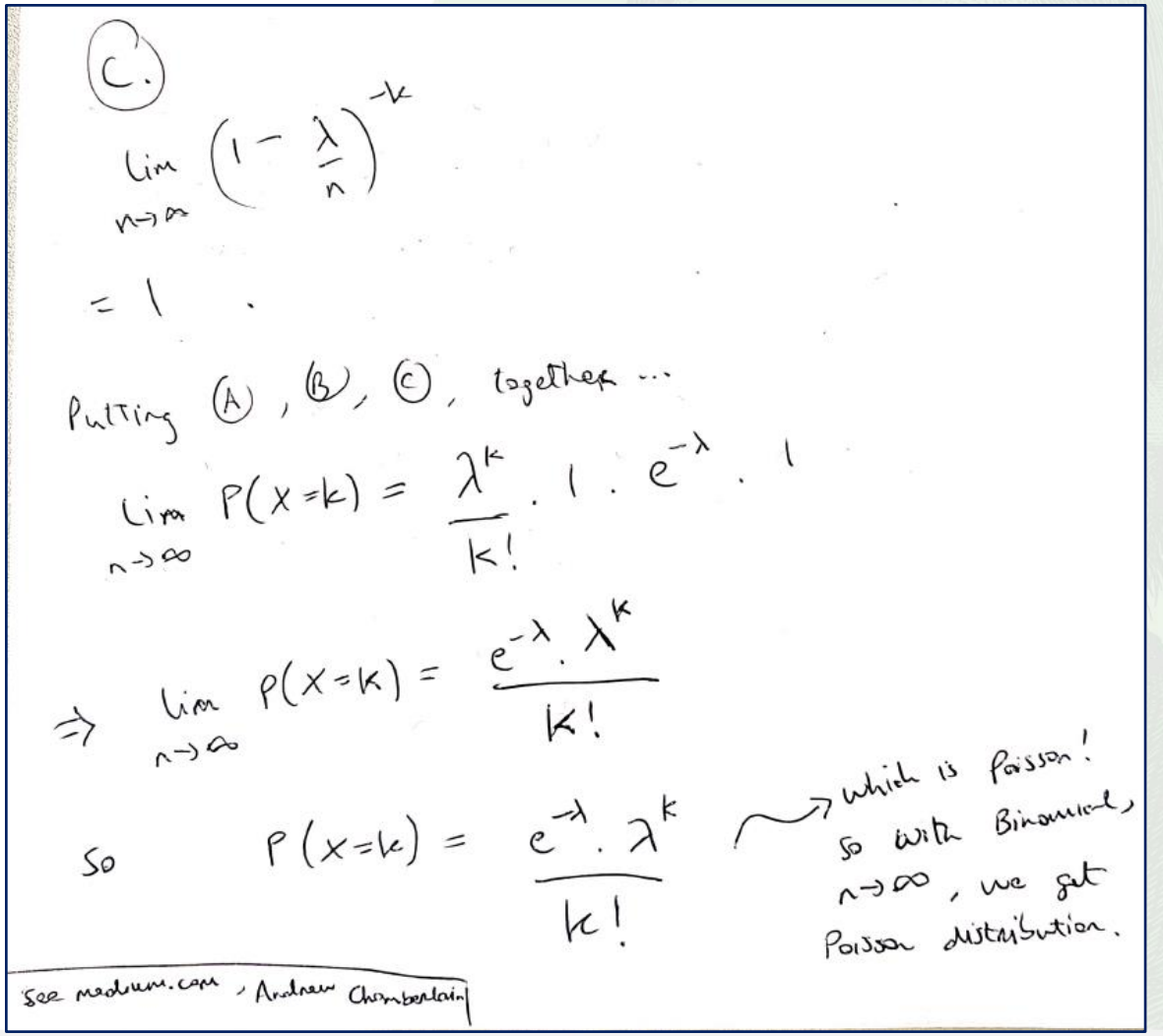
$$= \lim_{n\to\infty} \left(\frac{n}{n}\right)\left(\frac{n-1}{n}\right)\left(\frac{n-2}{n}\right)\cdots\left(\frac{n-k+1}{n}\right)$$

$$= 1 \times 1 \times 1 \times 1 \cdots$$

$$= 1 .$$

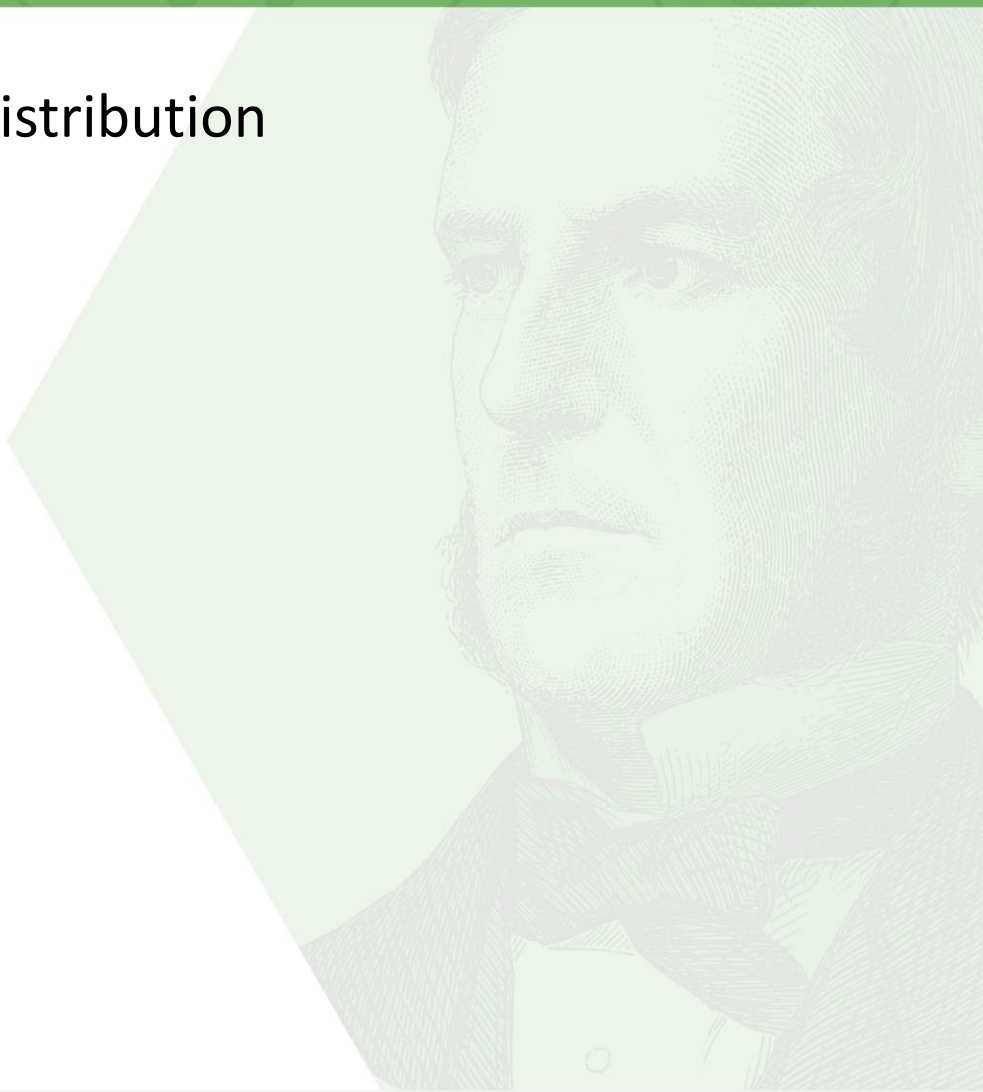
Poisson binomial connection





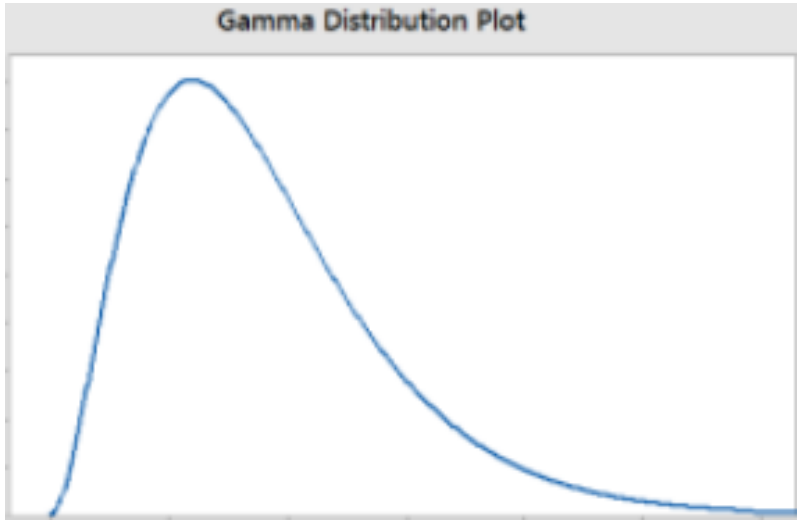
GLM assumption on the distribution of the response

- Imagine we were modelling frequency, with a Poisson distribution
- λ for policy no. 1 is calculated to be 0.1
- So expected count of claims = 0.1
- Actual count of claims could be 0,1,2,3…



GLM assumption on the distribution of the response

- The overall distribution of your data is not a justification to choose a particular distribution for your GLM …
- So even if the distribution of your claim severity data, all plotted on one chart, looks like this (very like Gamma):



- That is not a justification for using Gamma as the distribution of your GLM.
- The assumption in the model, is that the claims in the data are realisations of Gamma (for example) processes, where the μ of that process is a function of the data for the claim and parameters in the model

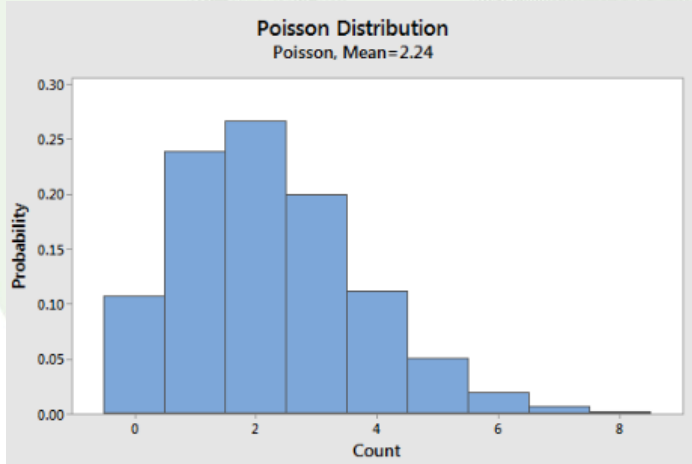
- E.g. so if there is a claim for Mr. X in the data of €10,000, we assume that €10,000 is a realisation of a gamma random variable which has a mean of μ, where μ is a function of the factors and relativities in the model, and is a function of the data for Mr. X. For a log link, μ would be given as:

$$\mu_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$$

- The gamma assumption is that the claims for each policyholder (or to be more precise, each rating cell) are generated from a gamma random variale, where the mean of the R.V is a function of the data and model parameters.

- The same applies for Poisson on frequency side.

- The shape of the overall distribution of claim counts is not justification for picking a Poisson. This distribution could look very non-Poisson in fact, and yet Poisson could still be the perfect assumption.

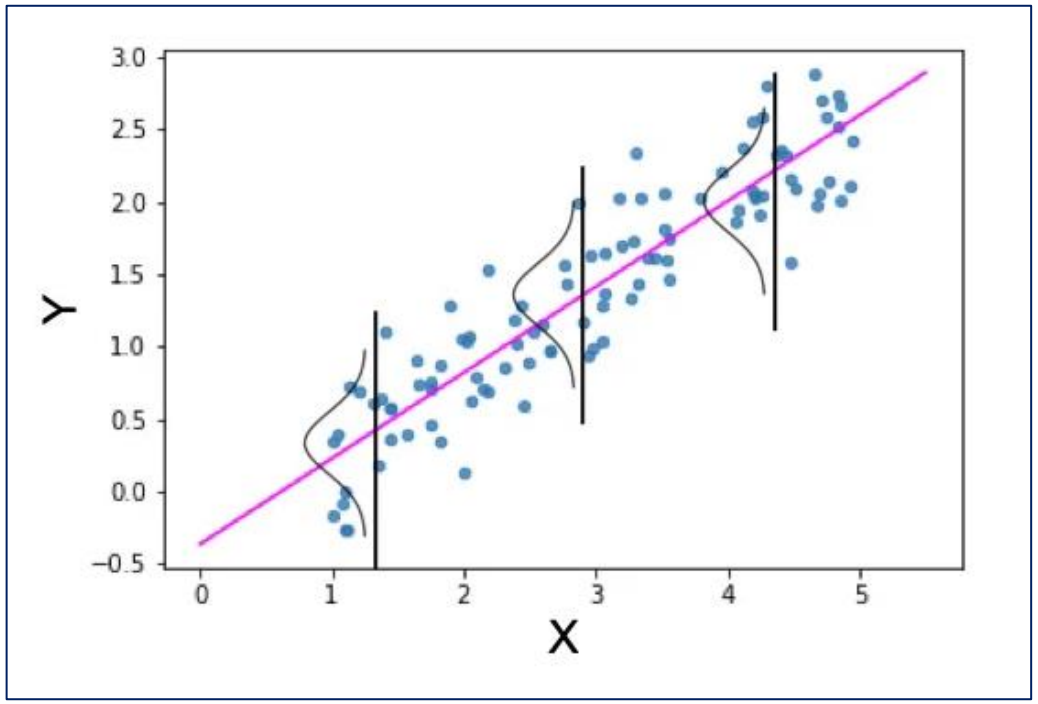


Images link 1, link 2, link 3

GLM assumption on the distribution of the response

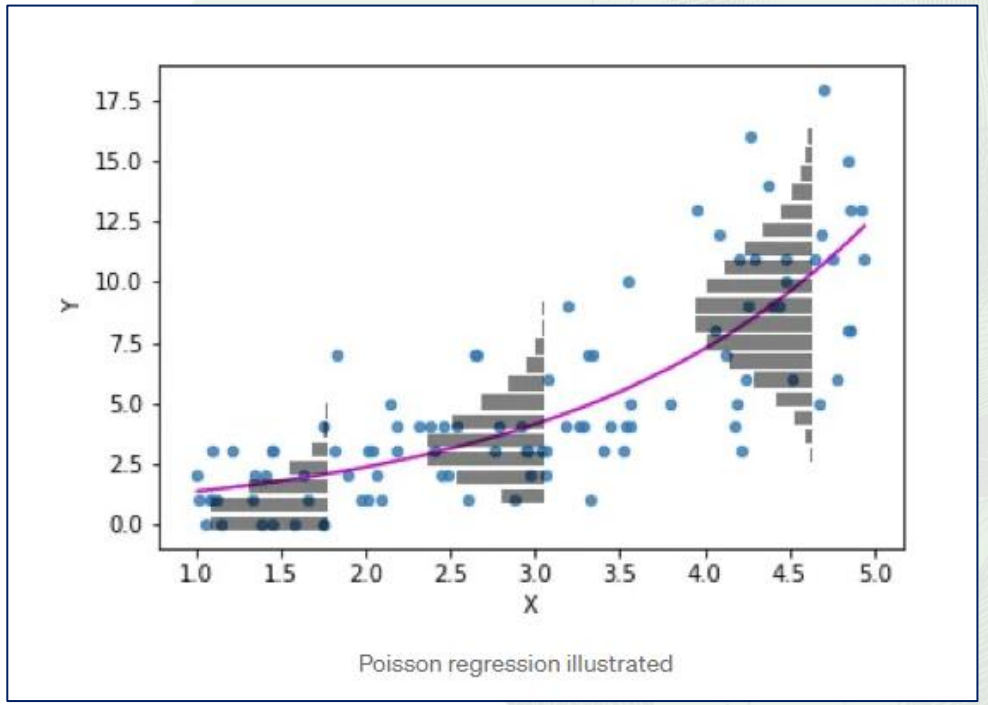
$$\mu_i = b_0 + b_1 x_i$$
$$y_i \sim \mathcal{N}(\mu_i, \varepsilon)$$

Linear regression



$$\ln \lambda_i = b_0 + b_1 x_i$$
$$\Leftrightarrow \lambda_i = \exp(b_0 + b_1 x_i)$$

Inverse of log link function

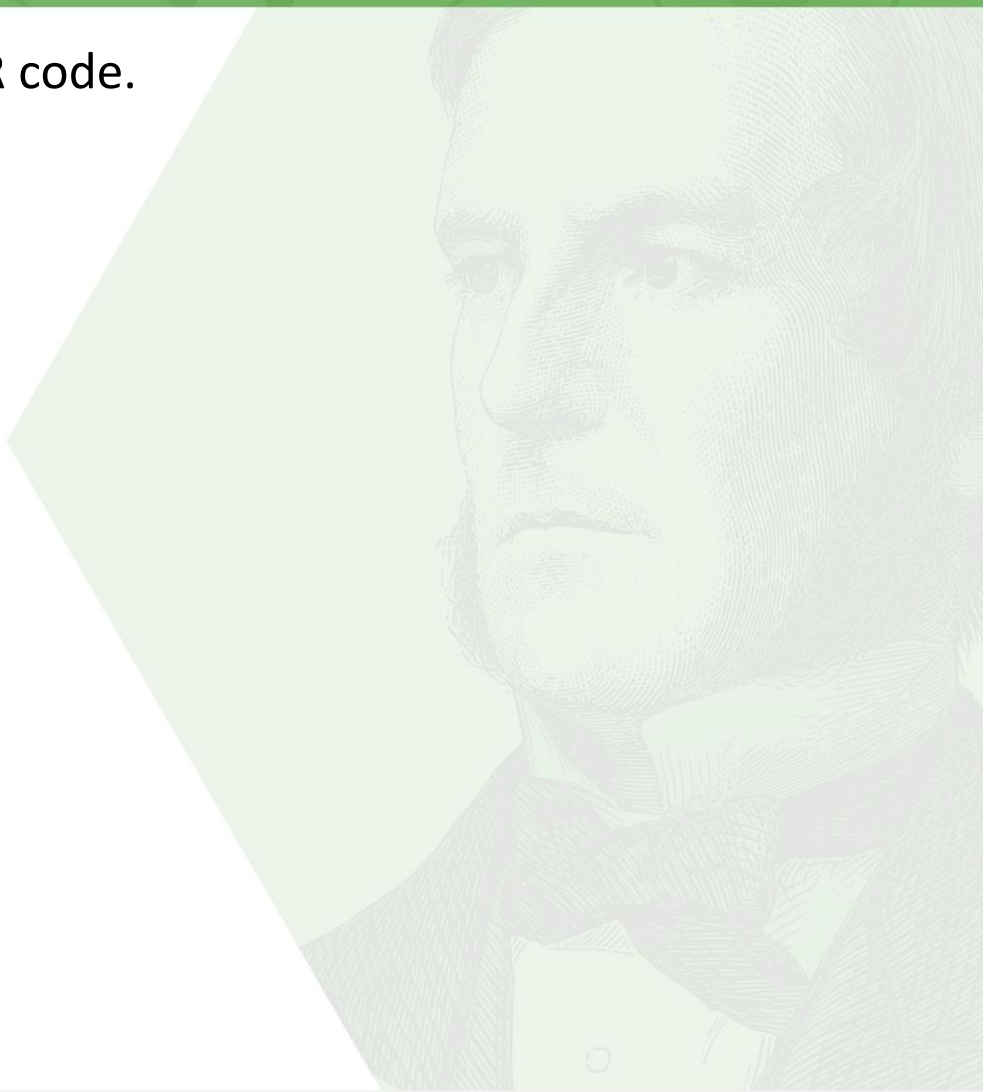


Poisson regression illustrated

Generalized linear models. Introduction to advanced statistical… | by Yuho Kida | Towards Data Science

Poisson deviance excel example

- See Poisson excel uploaded to canvas, and the translation of this to R code.



GLM assumption on the distribution of the response

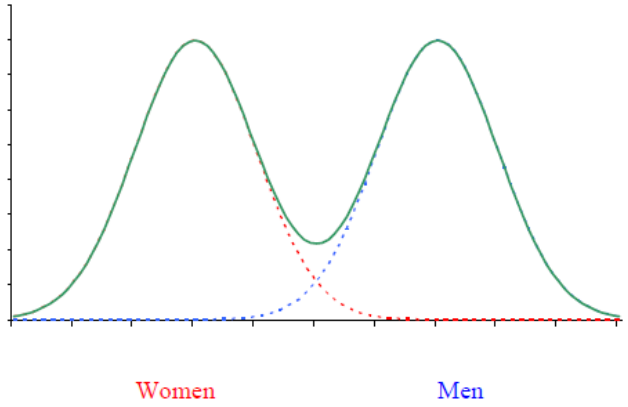
From the Practitioners Guide to GLMs …

Classical linear model assumptions

1.39 Linear models assume all observations are independent and each comes from a Normal distribution.

1.40 This assumption does not relate to the aggregate of the observed item, but to each observation individually. An example may help illustrate this distinction.

Distribution of individual observations

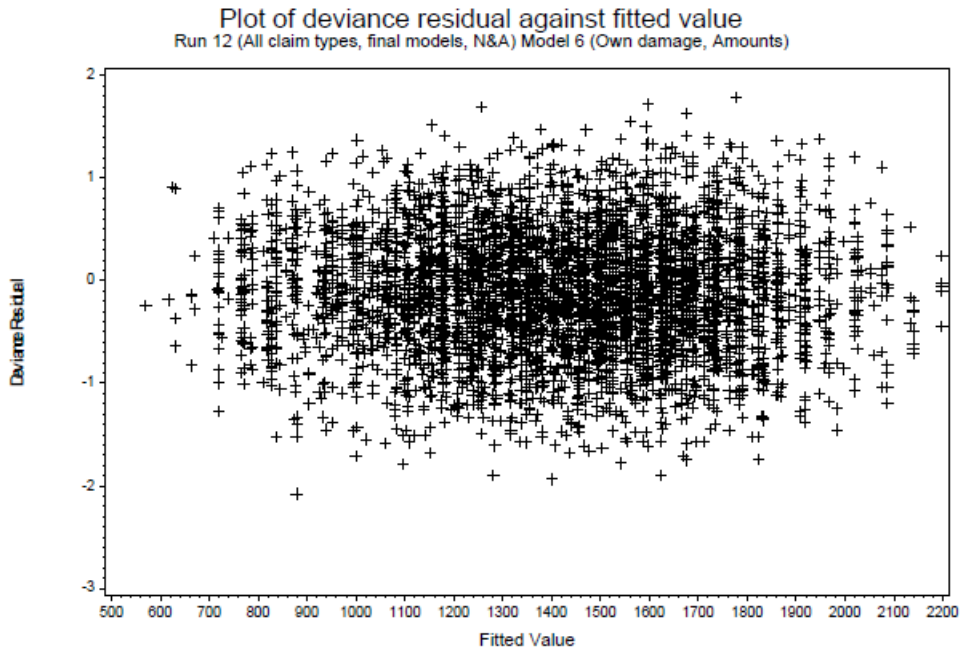


1.41 An examination of average claim amounts by gender may identify that average claim amounts for men are Normally distributed, as are average claim amounts for women, and that the mean of the distribution for men is twice the mean of the distribution for women. The total distribution of average claim amounts across all men and women is not Normally distributed. The only distribution of interest is the distribution of the two separate classes. (In this case there are only two classes being considered, but in a more complicated model there would be one such class for each combination of the rating factors being considered.)

GLM assumption on the distribution of the response

From the Practitioners Guide to GLMs …

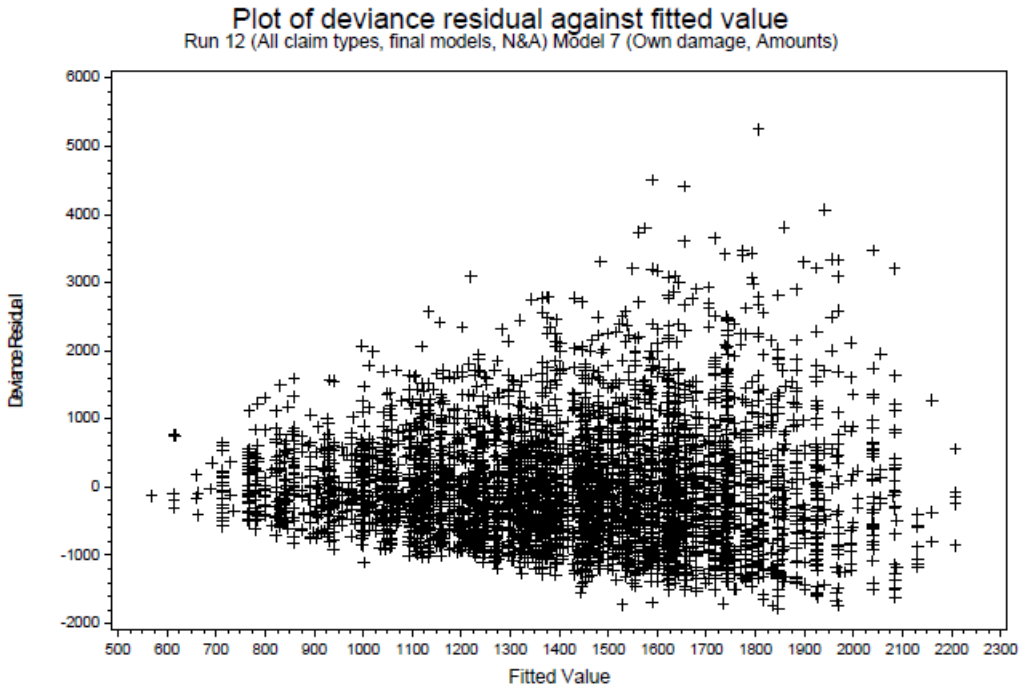
2.60 Observing scatter plots of residuals against fitted values can give an indication of the appropriateness of the error function which has been assumed. For example, if the model form is appropriate then the standardized deviance residuals should be distributed Normal (0,1) regardless of the fitted value. The example scatter plot below shows the result of fitting a GLM with a gamma variance function to data which has been randomly generated on a hypothetical insurance dataset from a gamma distribution (with a mean based on assumed factor effects). It can be seen that moving from the left to the right of the graph the general mean and variability of the deviance residuals is reasonably constant, suggesting (as is known to be the case in this artificial example) that the assumed variance function is appropriate.



GLM assumption on the distribution of the response

From the Practitioners Guide to GLMs …

2.61 Conversely the graph below shows the scaled deviance residuals obtained from fitting a GLM with an assumed Normal error to the same gamma data. In this case the variability increases with fitted value, indicating that an inappropriate error function has been selected and that the variance of the observations increases with the fitted values to a greater extent than has been assumed. This could occur, for example, when a Normal model is fitted to Poisson data, when a Poisson model is fitted to gamma data, or (as is the case here) where a Normal model is fitted to gamma data.

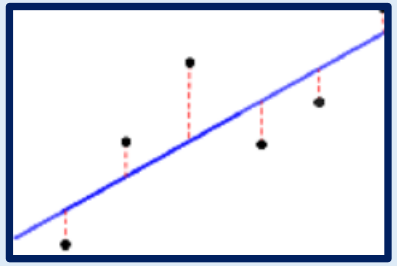


Loss functions

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

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

$$\text{MBE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)$$



$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \cdot 100\%$$

https://www.analyticsvidhya.com/blog/2021/10/evaluation-metric-for-regression-models/

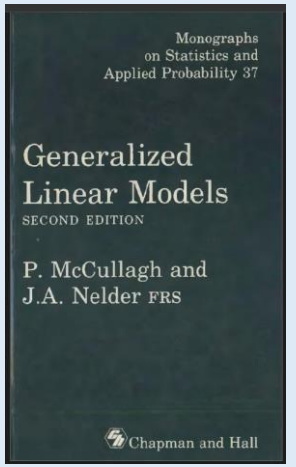
Loss functions

Gamma deviance for a dataset with n records …

$$\text{Poisson deviance, } D = 2 * \sum_{i=1}^{i=n} [actual_i \,.\ln\left(\frac{actual_i}{prediction_i}\right) - (actual_i - prediction_i)]$$

From McCullagh and Nelder …

Normal	$\sum(y-\hat{\mu})^2,$
Poisson	$2\sum\{y\log(y/\hat{\mu})-(y-\hat{\mu})\},$
binomial	$2\sum\{y\log(y/\hat{\mu})+(m-y)\log[(m-y)/(m-\hat{\mu})]\},$
gamma	$2\sum\{-\log(y/\hat{\mu})+(y-\hat{\mu})/\hat{\mu}\},$
inverse Gaussian	$\sum(y-\hat{\mu})^2/(\hat{\mu}^2 y).$



A note on one-ways

- History of rating
  - Underwriters
  - Actuaries before multivariate models
  - GLMs
  - The future?

- Example of factor that appears useless in univariate but is useful in multivariate
  - Age and driver engine size

- Example of factor that undergoes a change in relativities in multivariate view
  - Police officer count, rioter count, riot damage

- Look at the model from a number of one-way splits, sometimes two way splits, optimising a complete multivariate loss function.

A note on one-ways

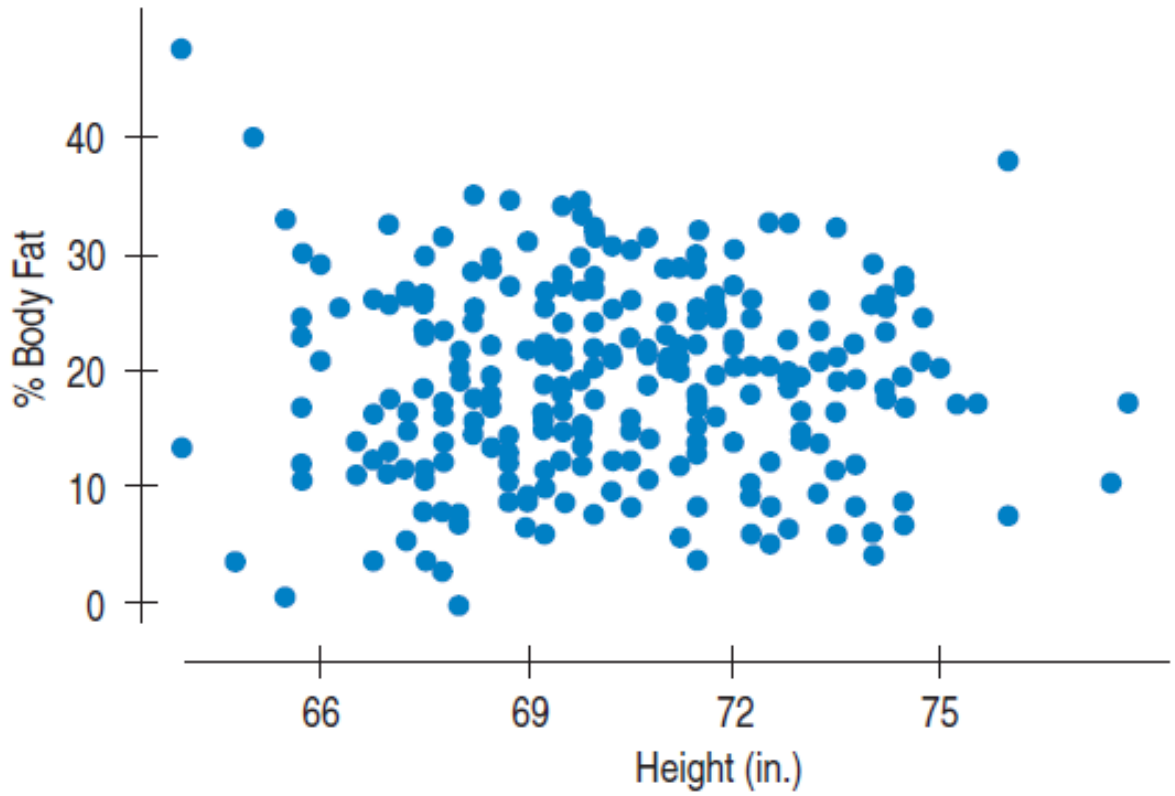


FIGURE 30.1
The scatterplot of *%Body Fat* against *Height* seems to say that there is little relationship between these variables.

Height is not so significant on it its own.

A note on one-ways

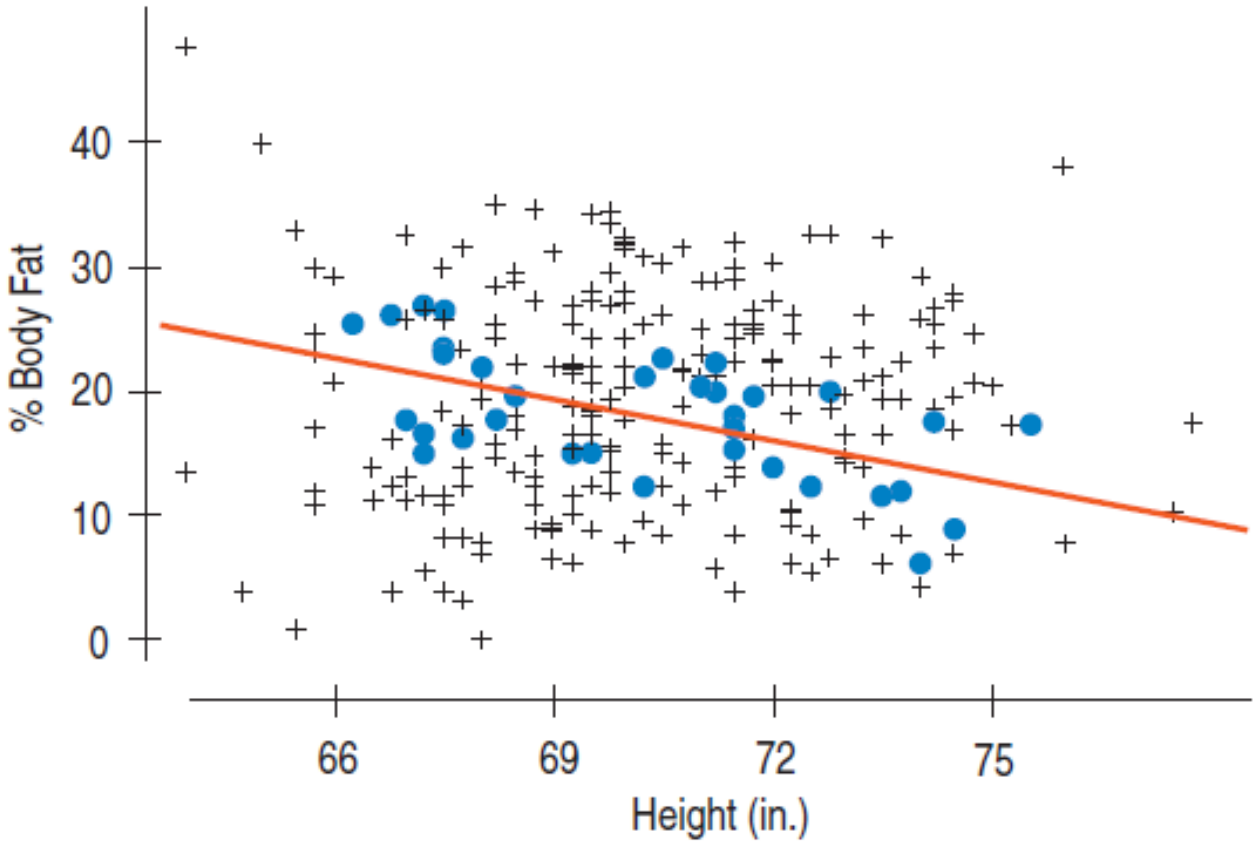


FIGURE 30.2
When we restrict our attention to men with waist sizes between 36 and 38 inches (points in blue), we can see a relationship between *%Body Fat* and *Height*.

But is significant in the presence of waist size

A note on one-ways

- In previous example there is a two-way effect at play
  - The effect of being taller, appears as though it should reduce body fat %
  - But as people get taller, their waist size gets bigger, and that is an effect working in the opposite direction (towards larger body fat%)
  - In this example they balance out to make it appear that height does not relate to body fat %, but it does. The multivariate view shows this.

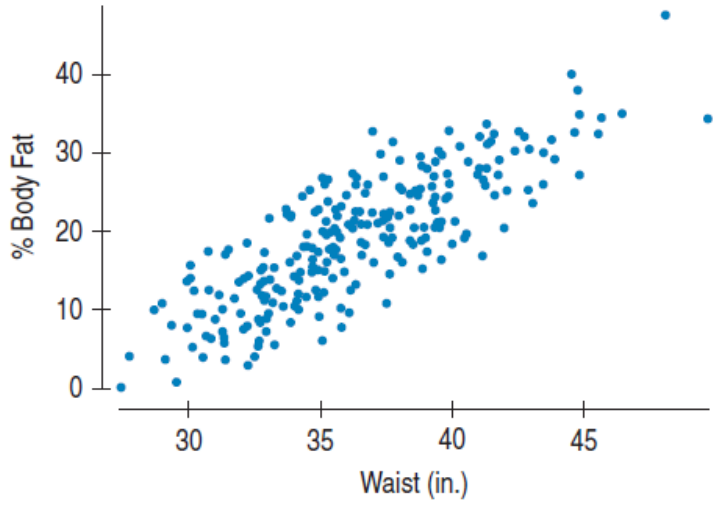
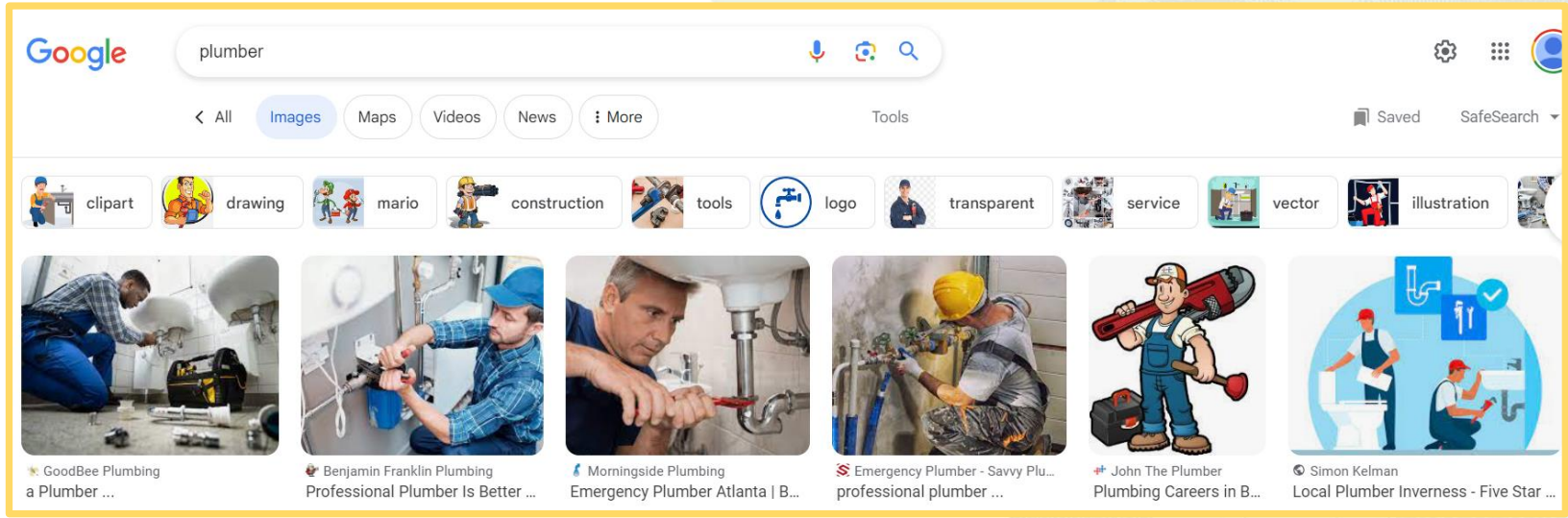


FIGURE 27.1
Percent Body Fat vs. *Waist* size for 250 men of various ages. The scatterplot shows a strong, positive, linear relationship.

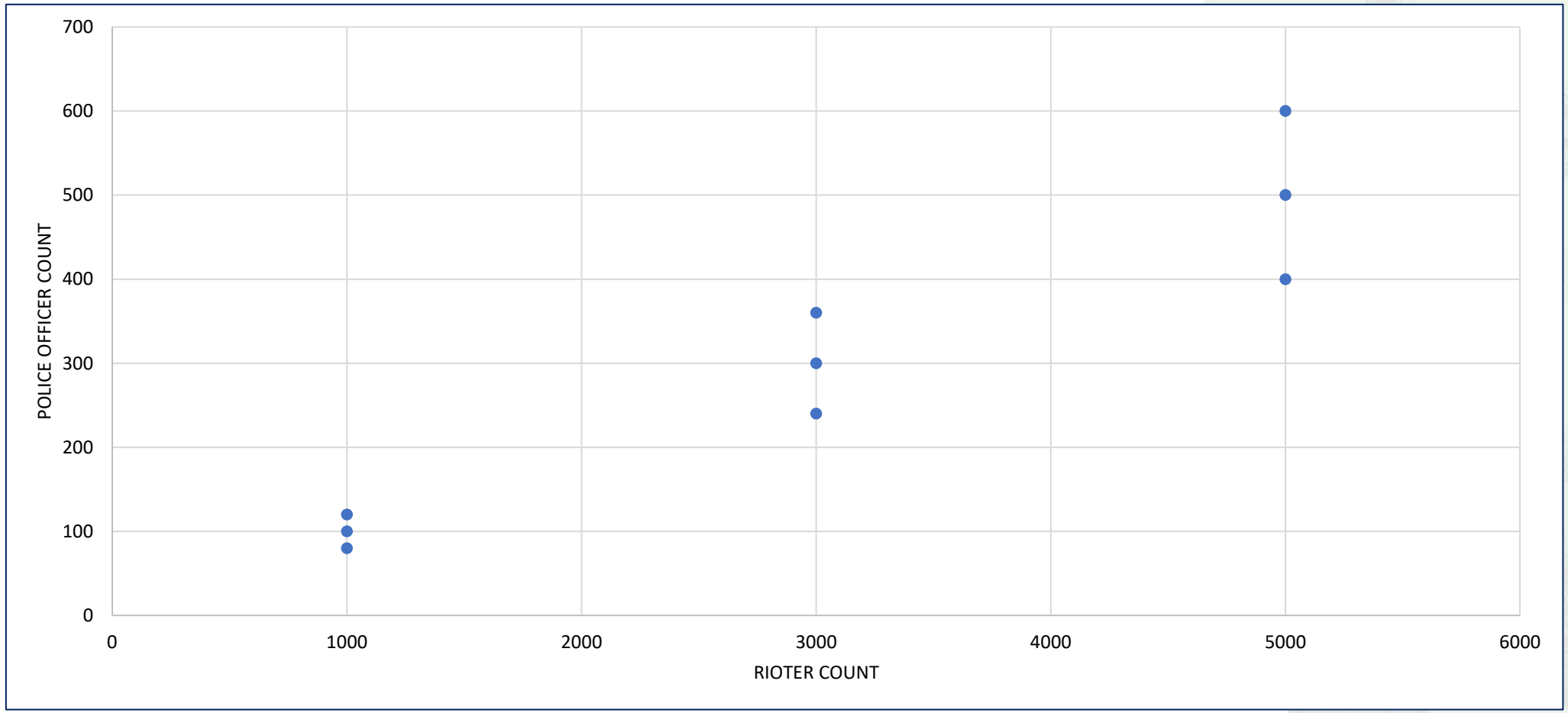
A note on one-ways

- Theoretical example of age and engine size … discuss

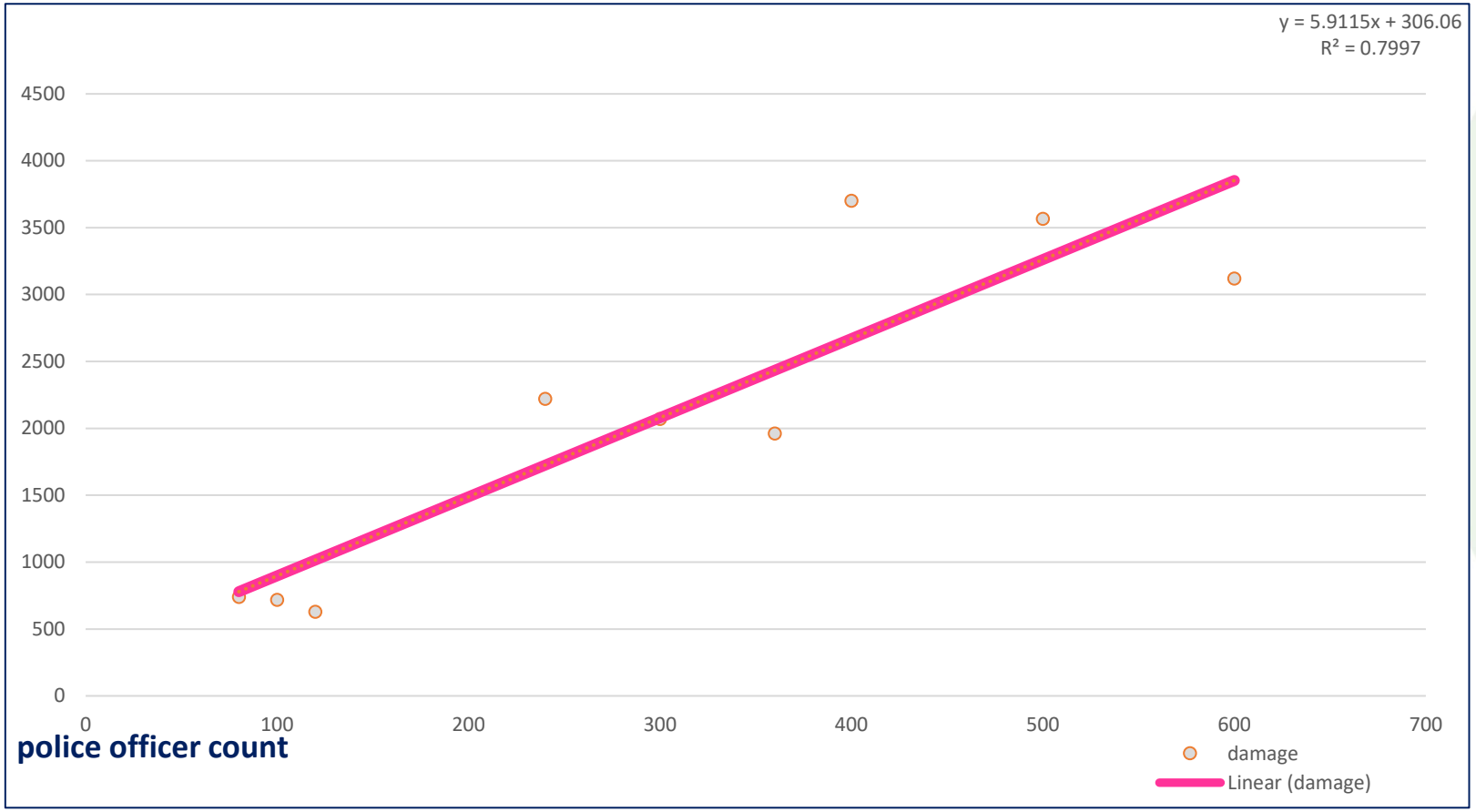
- What good is a univariate test? For example …
  - Only fit gender – what does that tell us?
  - Now fit gender AND occupation – what does that tell us?
  - Discuss…



A note on one-ways



A note on one-ways



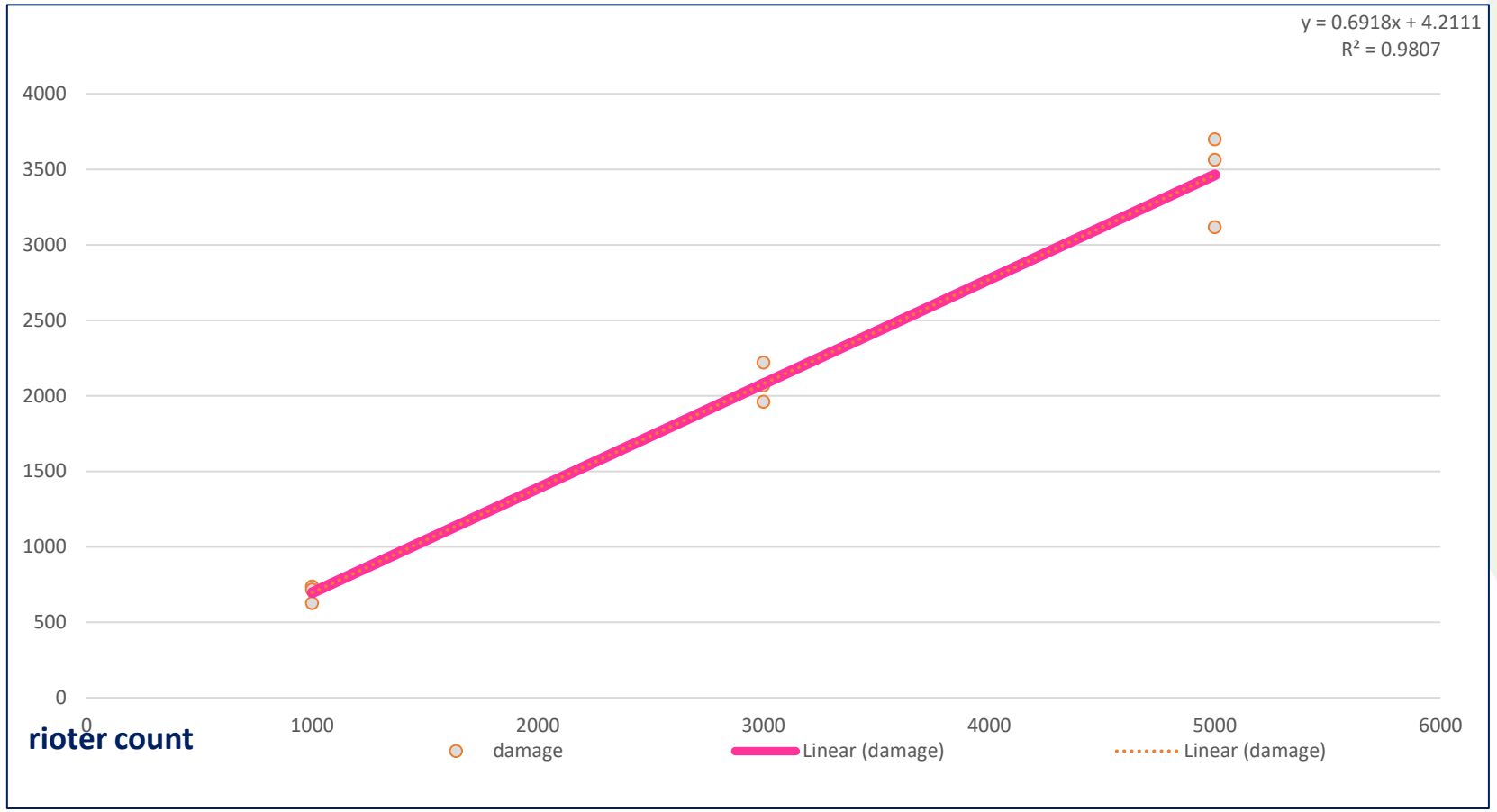
Regression Statistics	
Multiple R	0.882440782
R Square	0.778701733
Adjusted R Square	0.747087695
Standard Error	628.625239
Observations	9

ANOVA

	df	*SS*
Regression	1	9733629.158
Residual	7	2766187.837
Total	8	12499817

	Coefficients	*Standard Error*
Intercept	321.252073	420.2546644
police officer count	6.0	1.214295587

A note on one-ways



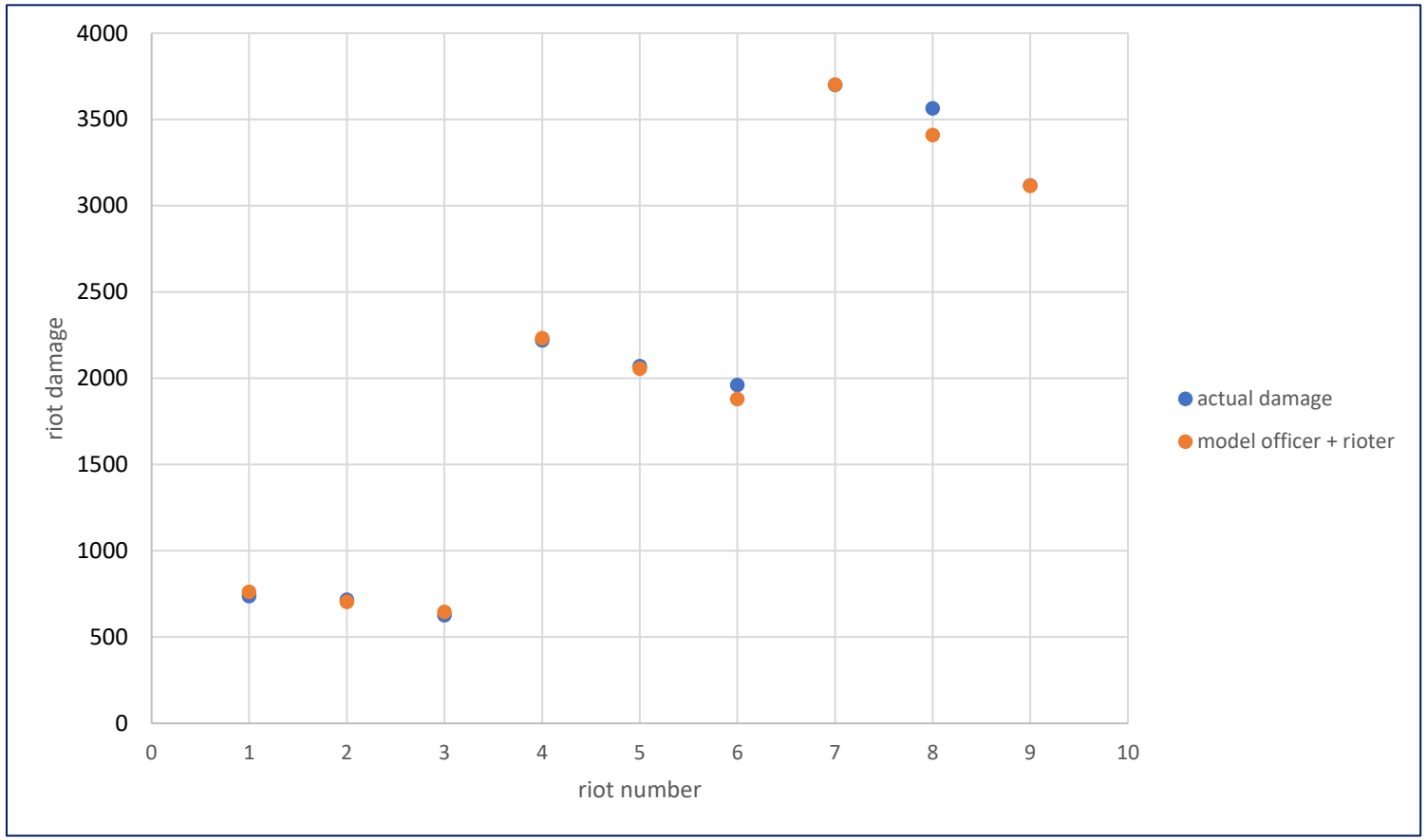
Regression Statistics	
Multiple R	0.98818495
R Square	0.976509496
Adjusted R Square	0.973153709
Standard Error	200.9877111
Observations	9

ANOVA

	df	*SS*
Regression	1	11754960.54
Residual	7	282772.42
Total	8	12037732.96

	Coefficients	*Standard Error*
Intercept	-7.35	140.1319864
rioter count	0.7	0.041026445

A note on one-ways



Regression Statistics	
Multiple R	0.999567162
R Square	0.999134511
Adjusted R Square	0.998846014
Standard Error	40.23636185
Observations	9

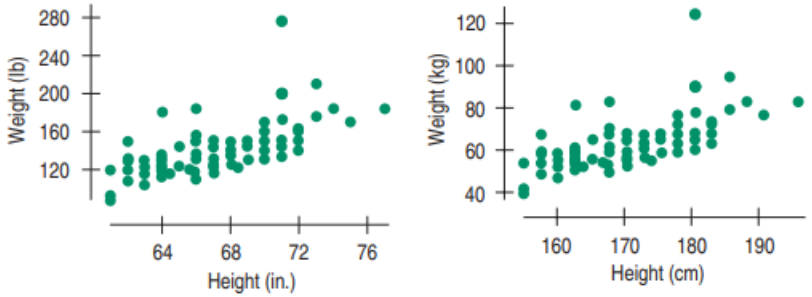
Discuss …

ANOVA

	df	*SS*
Regression	2	11213750.81
Residual	6	9713.788889
Total	8	11223464.6

	Coefficients	*Standard Error*
Intercept	28.47222222	28.05346298
rioter count	**1.0**	0.025409813
police officer count	**-2.9**	0.240458254

Correlation, causation, interactions

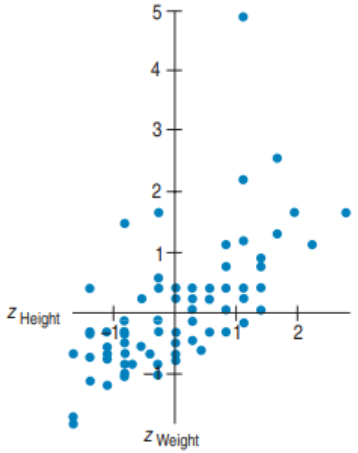


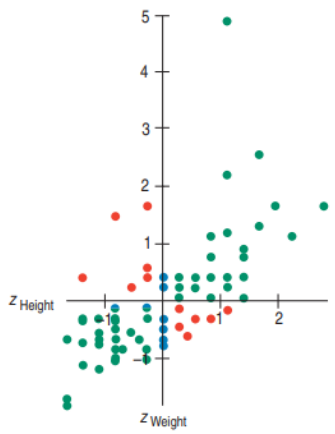
Units don't matter, r works both ways

$$r = \frac{\sum z_x z_y}{n-1}.$$

$$r = \frac{\sum (x-\bar{x})(y-\bar{y})}{\sqrt{\sum (x-\bar{x})^2 (y-\bar{y})^2}} = \frac{\sum (x-\bar{x})(y-\bar{y})}{(n-1)s_x s_y}.$$

$$(z_x, z_y) = \left(\frac{x-\bar{x}}{s_x}, \frac{y-\bar{y}}{s_y}\right).$$





Standardise and plot

3 conditions to satisfy before calculating a correlation coefficient:

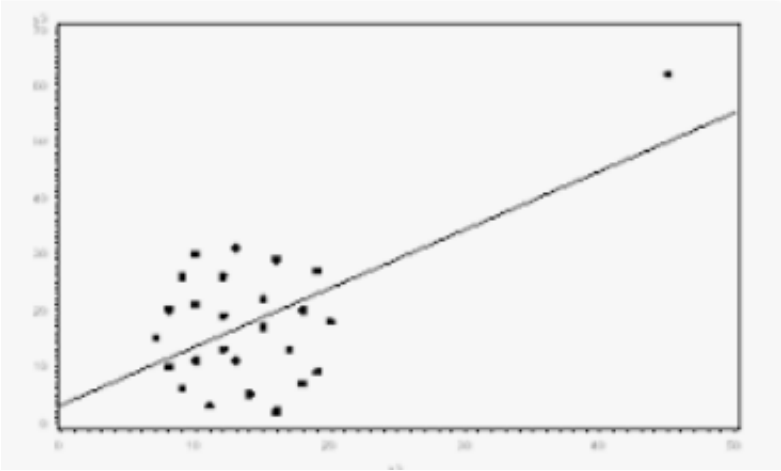
- Quantitative variables only
- Straight-enough condition
- Outlier condition

=> always show a scatter plot with a value of r

correlation measures the strength only of the linear association

Correlation, causation, interactions

Outliers …



Non-linear

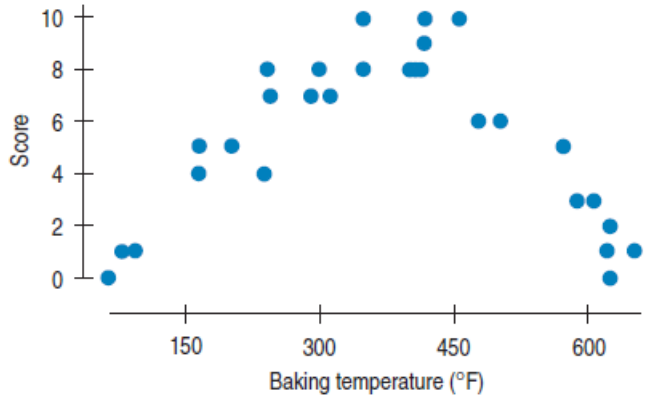


FIGURE 7.10
The relationship between brownie taste *Score* and Baking Temperature *is strong, but not at all linear.*

Quantitative – quantitative

Did you know that there's a strong correlation between playing an instrument and drinking coffee?

Interactions

When the effects of one factor change for different levels of another factor, we say that there's an interaction. To show the interaction, we often plot the treatment means in a single display, called an interaction plot. This plot shows the average of the observations at each level of one factor broken up by the levels of the other factor.

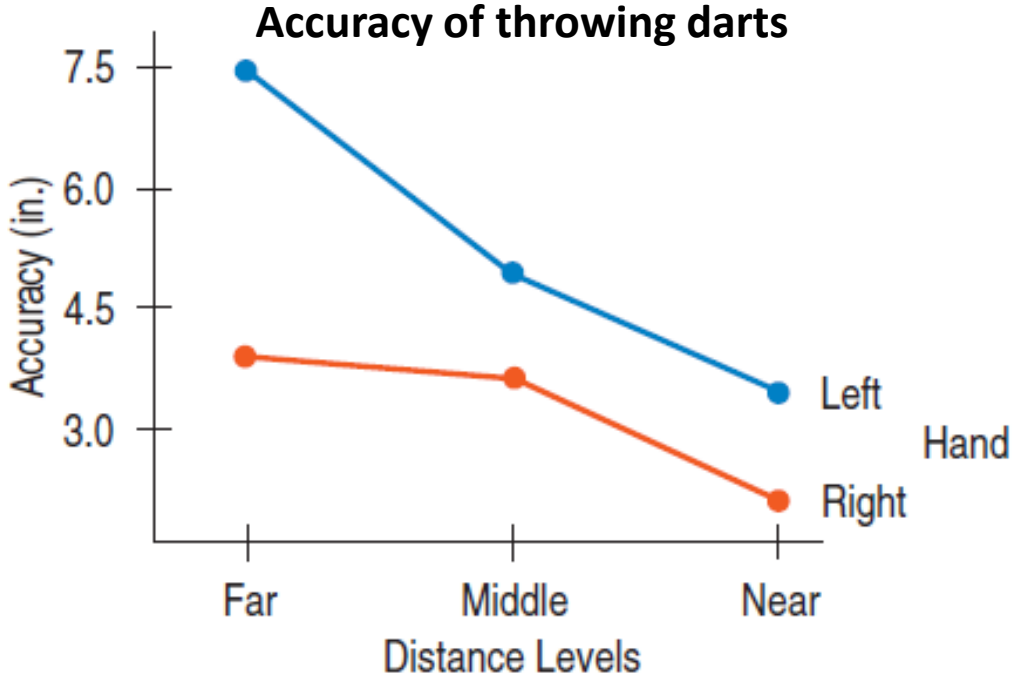
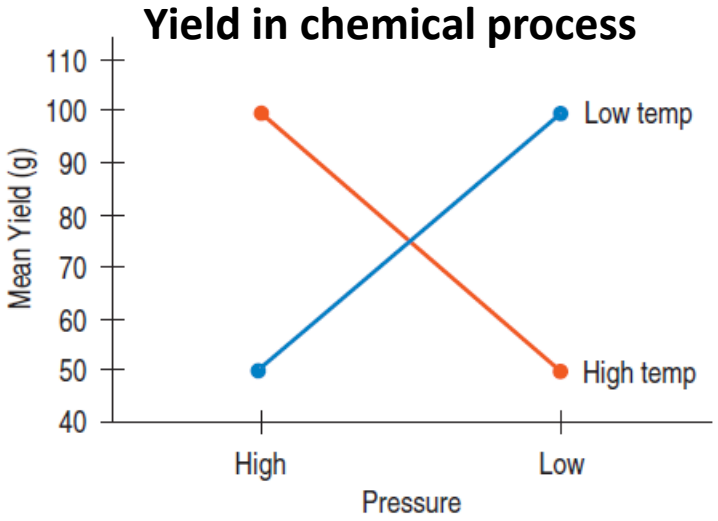


FIGURE 29.5
An interaction plot showing the mean accuracy at the six combinations of *Distance* and *Hand*. Connecting lines for right and left hands shows the greater effect of the far *Distance* on left-hand *Accuracy* than on right-hand *Accuracy*.

If the effect of Distance were constant, the lines in this plot would be parallel. This is the plot of observeds, check model indications after fitting.

Stats: Data and Models. Book by David E. Bock, Paul F. Velleman, and Richard D. De Veaux

Interactions



An interaction plot of *Yield* by *Temperature* and *Pressure*. The main effects are misleading. There is no (main) effect of *Pressure* because the average *Yield* at the two pressures is the same. That doesn't mean that *Pressure* has no effect on the *Yield*. In the presence of an interaction effect, be careful when interpreting the main effects.

Consider some insurance interactions …

Main effects can be very misleading in the presence of interaction terms. Look at this interaction plot. The experiment was run at two temperatures and two pressure levels.
High amounts of material were produced at high pressure with high temperature and at low pressure with low temperature. What's the effect of *Temperature*? Of *Pressure*? Both main effects are 0, but it would be silly (and wrong) to say that neither *Temperature* nor *Pressure* was important. The real story is in the interaction.

Interactions – further example

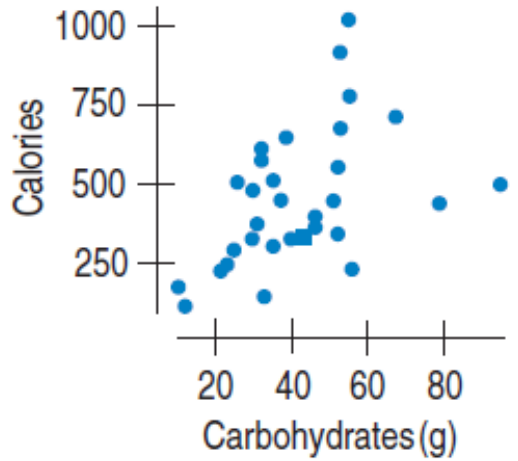
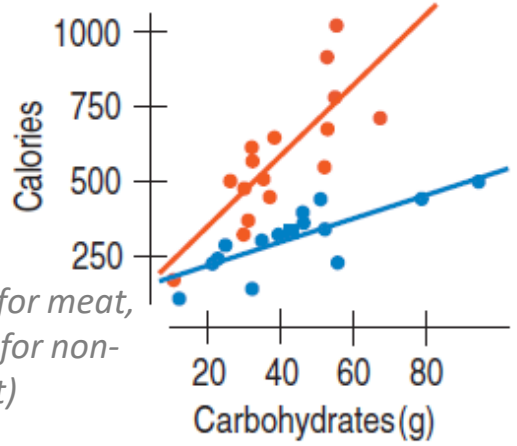


FIGURE 31.3
Calories of *Burger King foods plotted against* Carbohydrates *seems to fan out.*



(red for meat, blue for non-meat)

FIGURE 31.4
Plotting the meat-based and non-meat items separately, we see two distinct linear patterns.

Check for parallel regression lines. When you introduce an indicator variable for a category, check the underlying assumption that the other coefficients in the model are essentially the same for both groups. If not, consider adding an interaction term.

The effect of carb/g differs if it's meat or non-meat

Interactions – further example

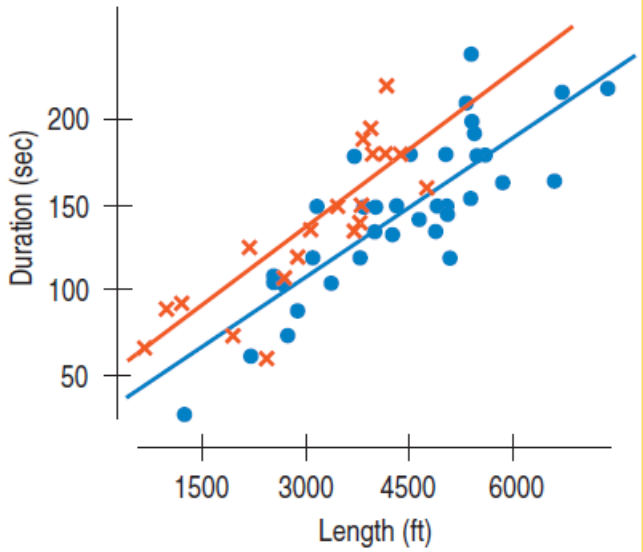


FIGURE 31.2
The two lines fit to coasters with inversions and without are roughly parallel.

One way of checking … split the data and fit twice …

Dependent variable is: Duration
Cases selected according to: No inversions
R-squared = 69.4% R-squared (adjusted) = 68.5%
s = 25.12 with 38 − 2 = 36 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	25.9961	14.10	1.84	0.0734
Length	0.0274	0.003	9.03	≤0.0001

Dependent variable is: Duration
Cases selected according to: Inversions
R-squared = 70.5% R-squared (adjusted) = 69.2%
s = 23.20 with 25 − 2 = 23 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	47.6454	12.50	3.81	0.0009
Length	0.0299	0.004	7.41	≤0.0001

Interactions – further notes

Keep it simple …

Simple factor interacted with simple factor
Probably too many parameters added to the model

Variate interacted with variate
Many different rates possible, unstable, high levels of checks needed

Custom factor with custom factor
Generally the best choice, requires pre-analysis.
Include as a compound factor.
Custom factor with variate
Can work, watch out for large jumps in rates

Check …

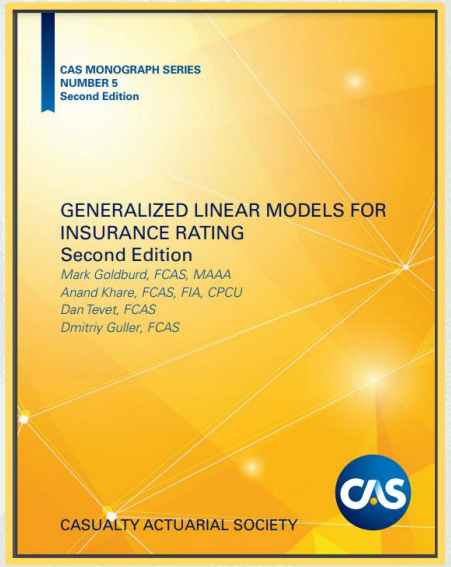
- Check that the fit is improved significantly – improved AIC for example
- Check that the indications are consistent across partitions – the best way to check if the interaction is reliable. You will need to prove it's consistent in your report.
- Check that it makes sense – in a GLM you should have some rationale for why it is needed
- Check it thoroughly!

Ensemble models

- So far, you have built these models:
  - GLM
  - Regression tree model
  - Pruned regression tree model
  - Bagged tree model
  - Random forest model
  - Gradient boosted tree model
- Often, the best model is an **ensemble** of other models. From the CAS GLM text:

> A model that combines information from two or more models is called an **ensemble** model. There are many strategies for combining models, and a full treatment of the subject is beyond the scope of this text. However, a simple, yet still very powerful, means of ensembling is to simply take the straight average of the model predictions.[21] Two well-built models averaged together will almost always perform better than one, and three will perform even better—a phenomenon known as the *ensemble effect*. Generally, the more models the better, though subject to the law of diminishing returns. In fact, ensembling is one notable exception to the parsimony principle in modeling (i.e., the "keep it simple" rule); adding more models to an ensemble—thereby increasing the complexity—will rarely make a model worse.

CAS GLM text



Text uploaded to canvas

Ensemble models

You can average using:

- simple average
- weighted average (use CV to find optimal weights)
- Geometric average (model_**1**_prediction x model_**2**_prediction)$^{1/2}$

From the CAS GLM text:

> Predictive models, like people, each have their strengths and weaknesses. One model may over-predict on one segment of the data while under-predicting on another; a different model is not likely to have the same flaws but may have others. Averaged together, they can balance each other out, and the gain in performance can be significant.

> One caveat though—for the ensemble effect to work properly, the model errors should be as uncorrelated as possible; that is, the models shouldn't all be systematically missing in the same way. (Much as the averaged jelly bean guesses would not work well if everyone guessed similarly.) Thus, if ensembling is to be employed as a model-building strategy, it is best if the models are built by multiple people or teams working *independently*, with little or no sharing of information. Done properly, though, ensembles can be quite powerful; if resources permit, it may be worth it.