Practical Predictive Analytics Seminar

Matthias Kullowatz

Session 3: Predictive Models (with Life example)

September 23, 2020

SOCIETY OF ACTUARIES

Agenda

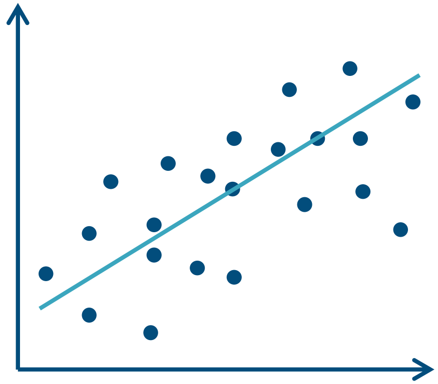
- Questions of interest for actuaries
- Logistic regression theory and application
- Associated theoretical concerns that may arise in the modeling process
- Model validation
- Hands-on time throughout!

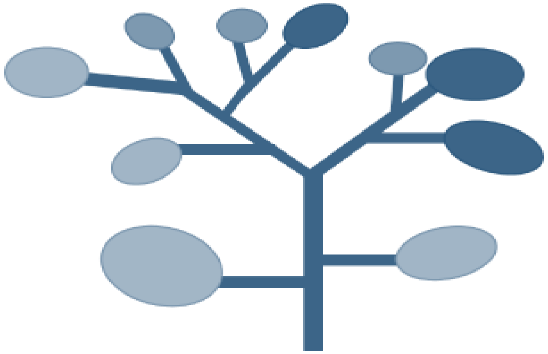
Theory

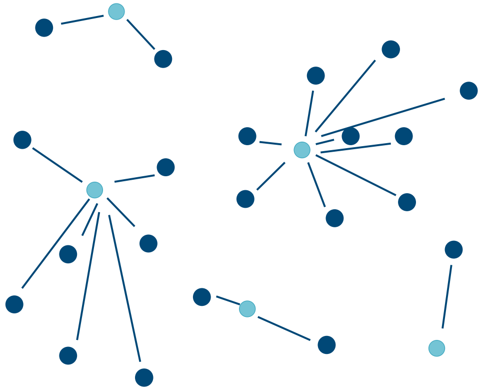
Questions of interest

- When will a policyholder…
  - Lapse?
  - Make a claim or withdrawal?
  - Die?
- How much?
- What drives these “behaviors” and why?
- Are the findings implementable?

Predictive model forms



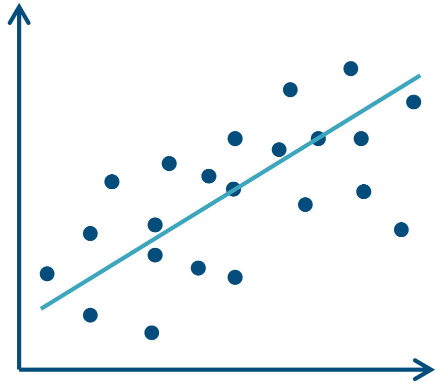




Icon made by Freepik from www.flaticon.com

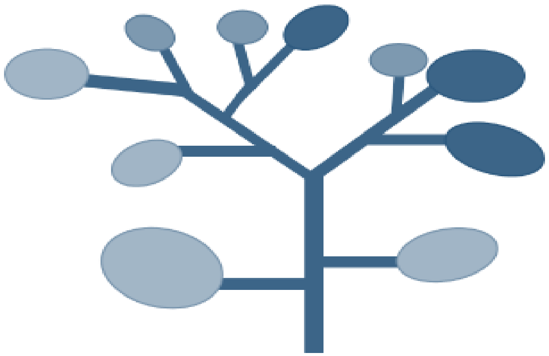
Regression

- OLS, GLM, regularization (ridge, lasso, elastic net)
- Pros
  - Quick fitters
  - Interpretable coefficients and output
  - Harder to overfit
  - Widely used
- Cons
  - Constrained by parametric, functional form
  - Multicollinearity issues



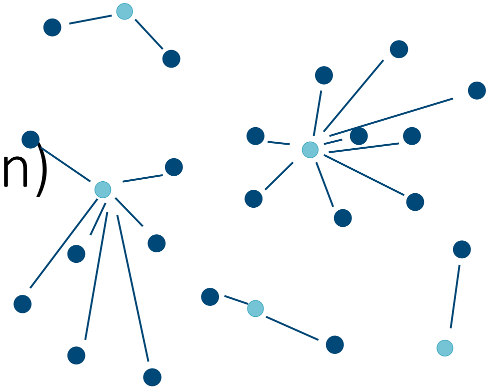
Tree-based models

- Decision trees, random forest, GBM
- Pros
  - Inherently models interactions between drivers
  - Models relationships non-parametrically
- Cons
  - Black-boxy formula (enter: Kshitij)
  - Hard to implement in other software
  - Doesn’t interpolate or extrapolate well



Clustering, et. al.

- Supervised: k-nearest neighbors
- Unsupervised: k-means, hierarchical
- Pros
  - Reduces dimensionality (ease of interpretability)
  - Easy to explain predictions (k-nearest neighbors)
- Cons
  - Sensitive to outliers
  - Reduces dimensionality (loss of information)



Neural networks

- Pros
  - Inherent interaction effects/non-parametric
  - Well-suited for problems with many predictor variables
    - Image recognition and text analysis-type problems
- Cons
  - Black-box formula (even more opaque than GBM/RF)
  - Hard to implement in other software
  - Computationally intensive

Icon made by Freepik from www.flaticon.com

Other modeling methods/techniques

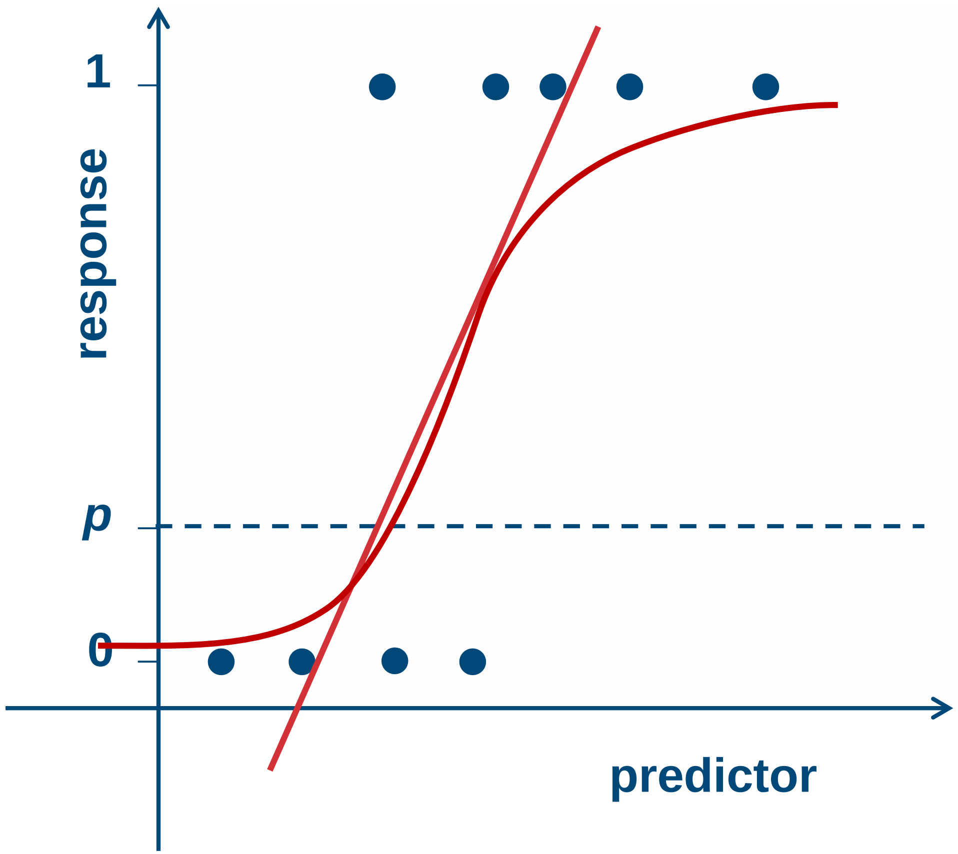
- Survival models
  - Cox proportional hazards
  - Accelerated failure time
- Support vector machines
- Agent-based modeling
- Splines (with regularization)

Icon made by Freepik from www.flaticon.com

Logistic GLM

- For predicting probabilities of binary outcomes
- Link function provides much needed flexibility
- Predictor variables can be quantitative or qualitative

Why a link function?

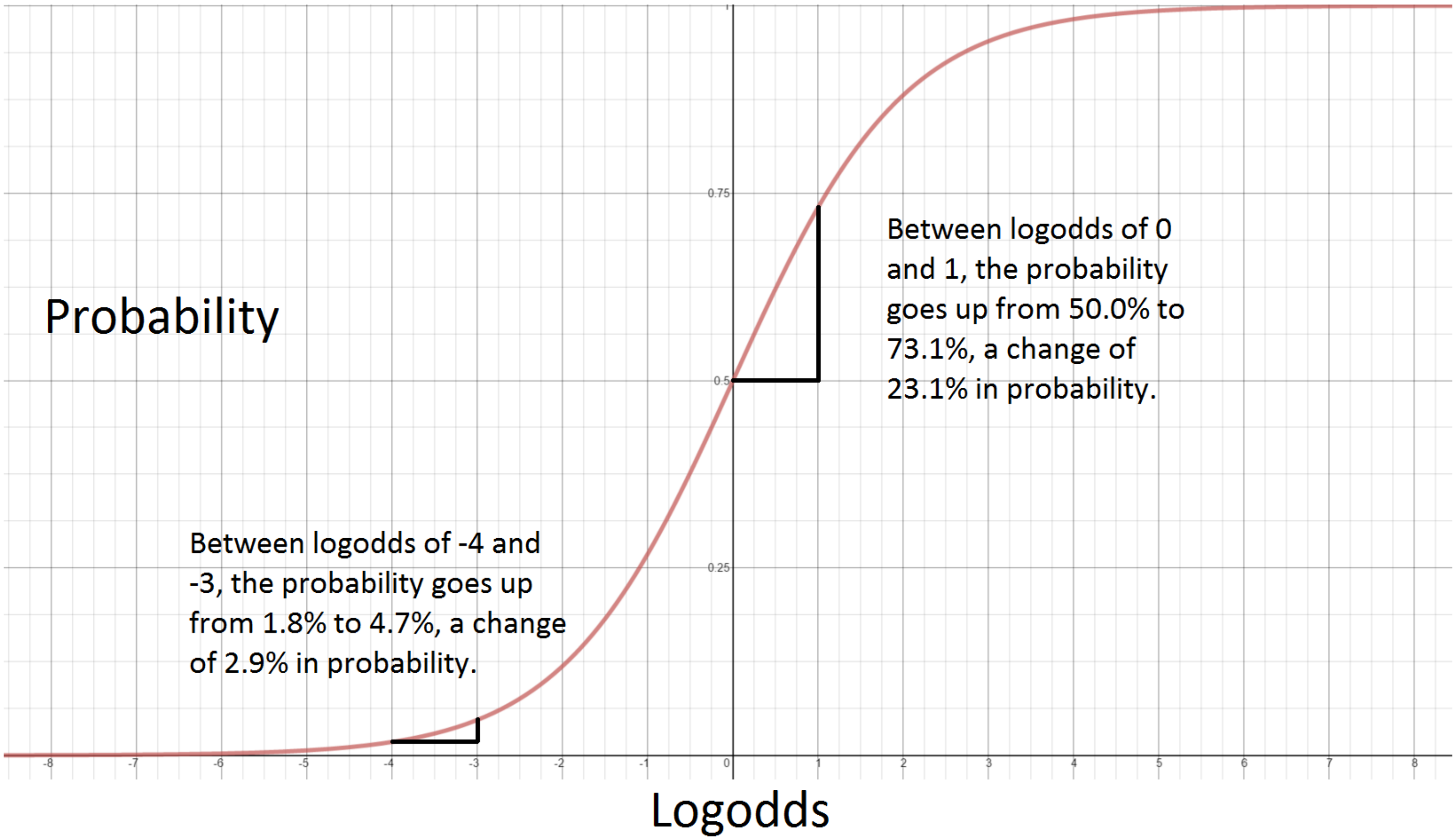


The logistic function

- $\hat{y} = g(L) = \frac{e^L}{1+e^L}$
  - $L = \widehat{\beta_0} + \widehat{\beta_1}x_1 + \widehat{\beta_2}x_2 + \cdots + \widehat{\beta_p}x_p$
  - $\lim_{L\to\infty} g(L) = 1 \; and \; \lim_{L\to-\infty} g(L) = 0$

- $g^{-1}(\hat{y}) = \ln\left(\frac{\hat{y}}{1-\hat{y}}\right) = L$
  - Logit function ("logodds")

Consequences of logit link



Interpretation of coefficients

- $\ln\left(\frac{\hat{y}(x)}{1-\hat{y}(x)}\right) = \widehat{\beta_0} + \widehat{\beta_1}x \Rightarrow \frac{\hat{y}(x)}{1-\hat{y}(x)} = e^{\widehat{\beta_0}+\widehat{\beta_1}x}$

- Continuous x-value:
  - $\frac{\hat{y}(x+1)}{1-\hat{y}(x+1)} \div \frac{\hat{y}(x)}{1-\hat{y}(x)} = \frac{e^{\widehat{\beta_0}+\widehat{\beta_1}(x+1)}}{e^{\widehat{\beta_0}+\widehat{\beta_1}x}}$

    $= e^{\widehat{\beta_1}}$

- Odds ratio

Theoretical extras

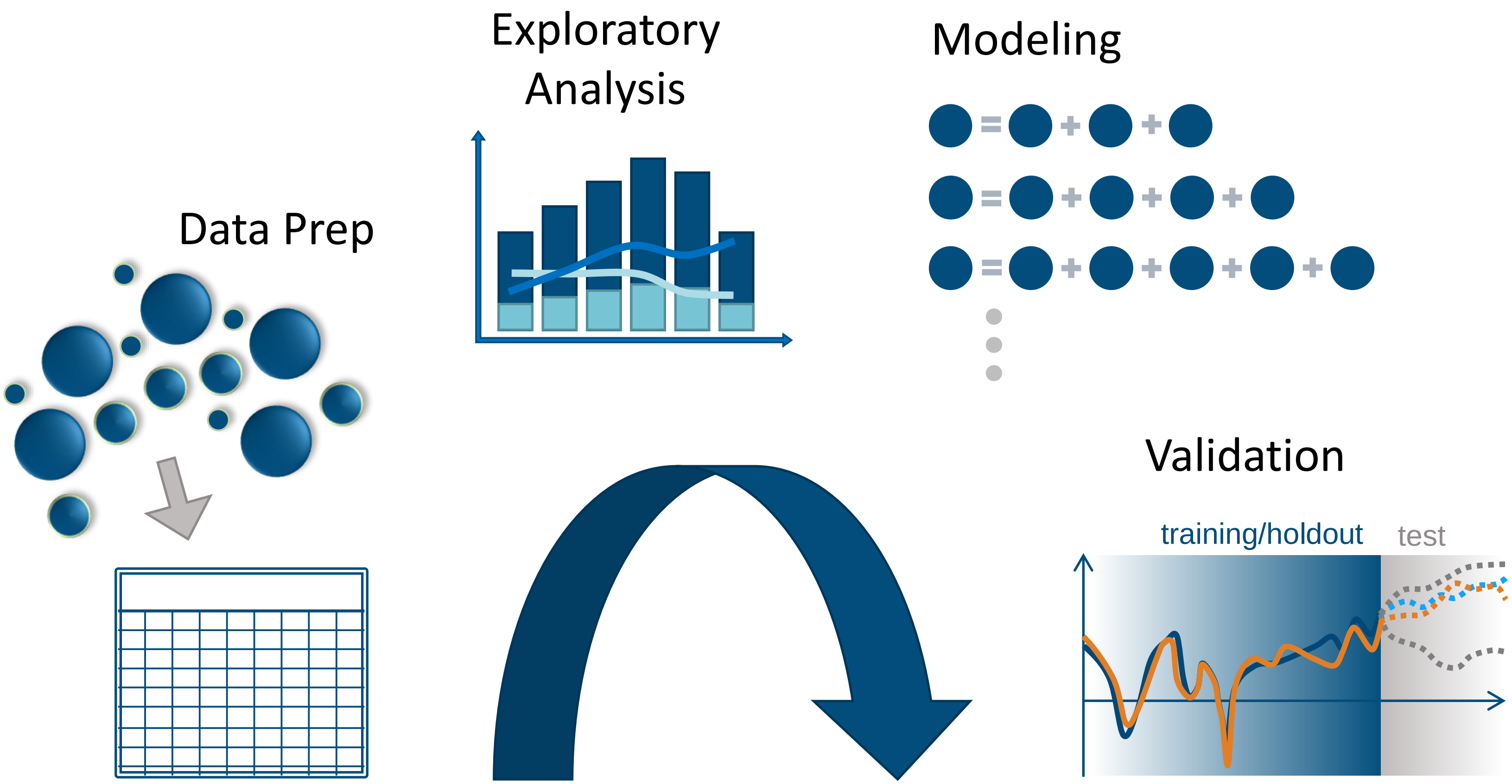
- Independent observations
- The model is fit by maximizing the following:

$$loglikelihood = \sum [Y_i \ln(\hat{y}_i) + (1 - Y_i)\ln(1 - \hat{y}_i)]$$

  - $AIC = -2 \times loglikelihood + 2 \times parameters$
  - $BIC = -2 \times loglikelihood + \ln(N) \times parameters$

Practical concerns

Predictive analytics process



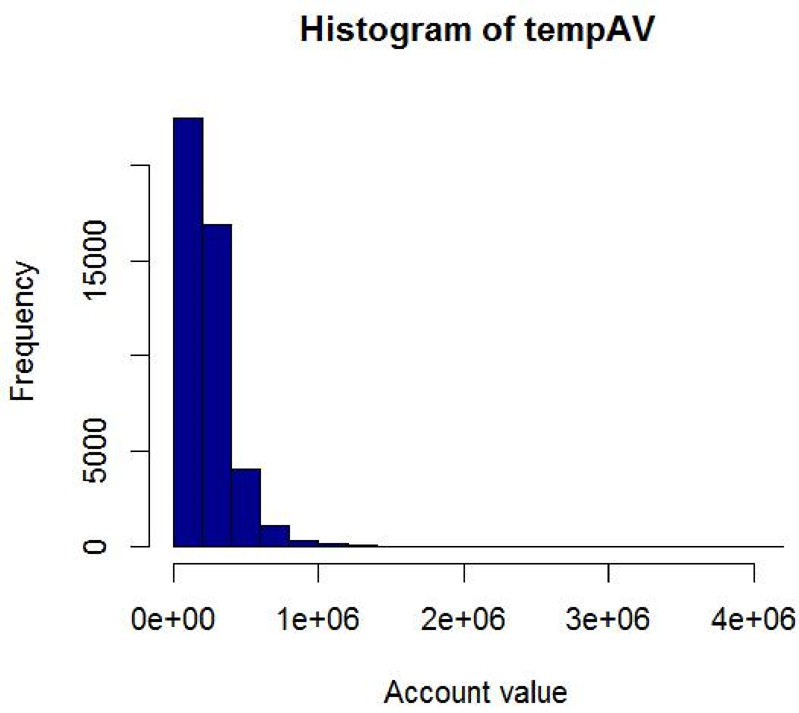
Practical concerns: Data

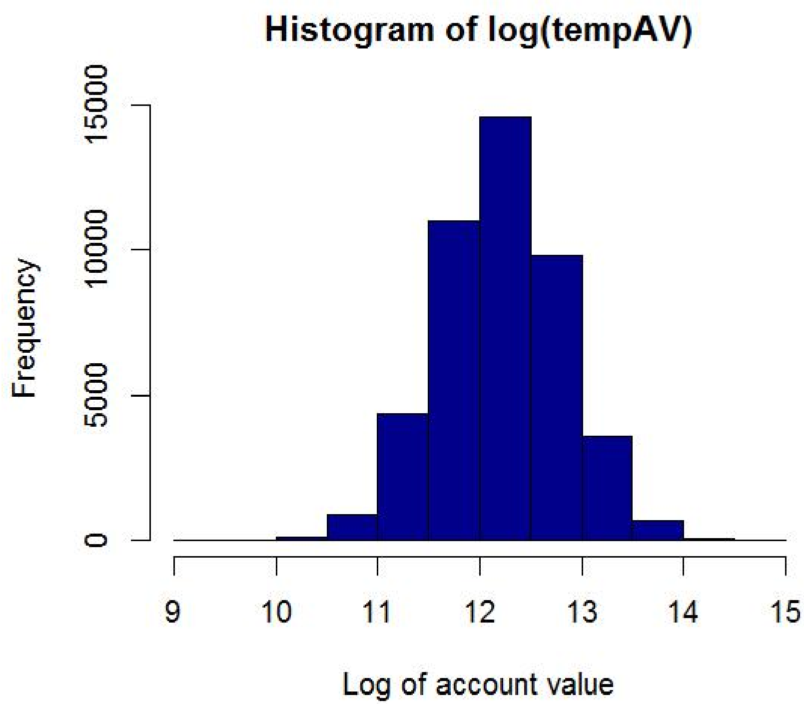
- Formatting variables (1)
- Identifying and dealing with outlier data values (2)
- Accounting for missing data (2)
- Derive new variables for modeling (3)
- Compile dataset into appropriate format (4)

Practical concerns: Modeling

- Holdout dataset (2A)
- Fitting a model (2C)
- Multicollinearity concerns (2E)
- Setting reference levels for factors (DataPrep 2)
- Piecewise terms (2F)
- Undersampling (3)

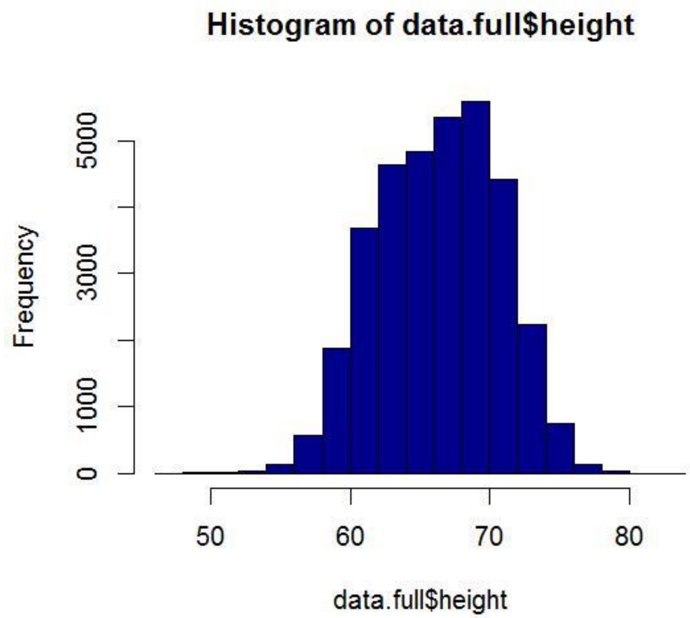
Data outliers

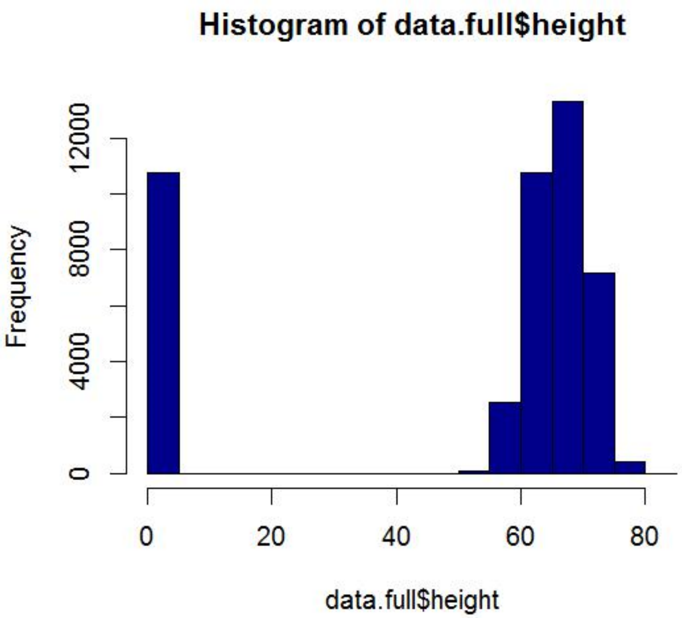




Missing values

```
> summary(data.full$height)
  Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's 
 47.00   64.00   67.00   66.82   70.00   83.00   10739 
```



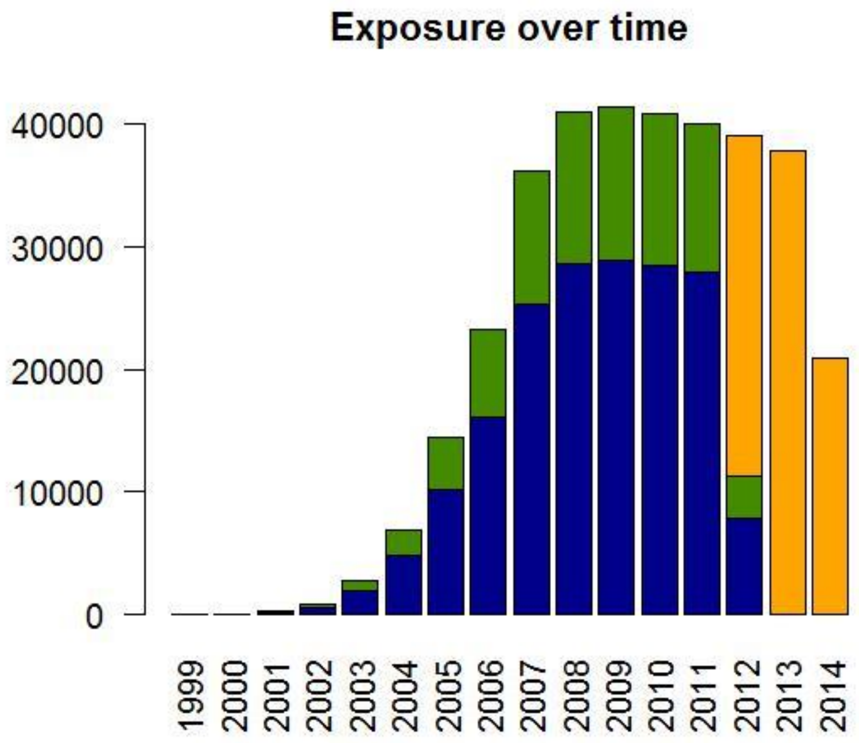


Missing values

Model	NA treatment	Intercept	Height coefficient	Flag coefficient
Death ~ height	Removed	-4.418	0.0100	N/A
Death ~ height + Ind	Set to 0	-3.580	0.0100	-0.838
Death ~ height + Ind	Set to mean	-4.245	0.0100	-0.173
Death ~ height	Set to 0	-3.589	-0.0024	N/A
Death ~ height	Set to mean	-4.343	0.0095	N/A

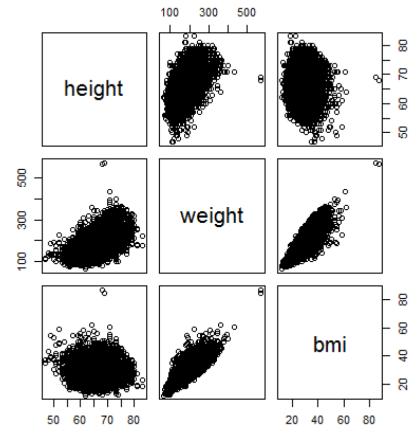
- The first three models are mathematically equivalent
- The second two are biased
- Flag indicates that height was *not* missing

Training versus holdout data



Multicollinearity

- pairs()

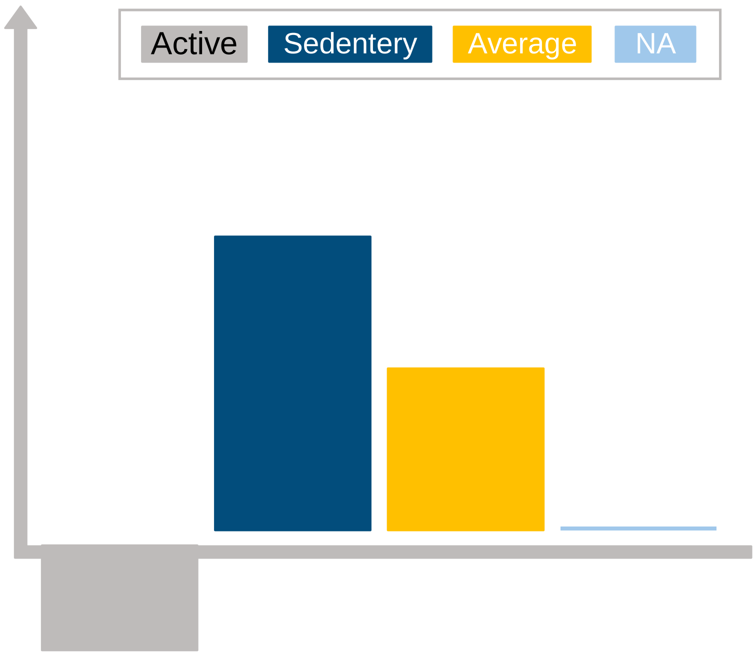


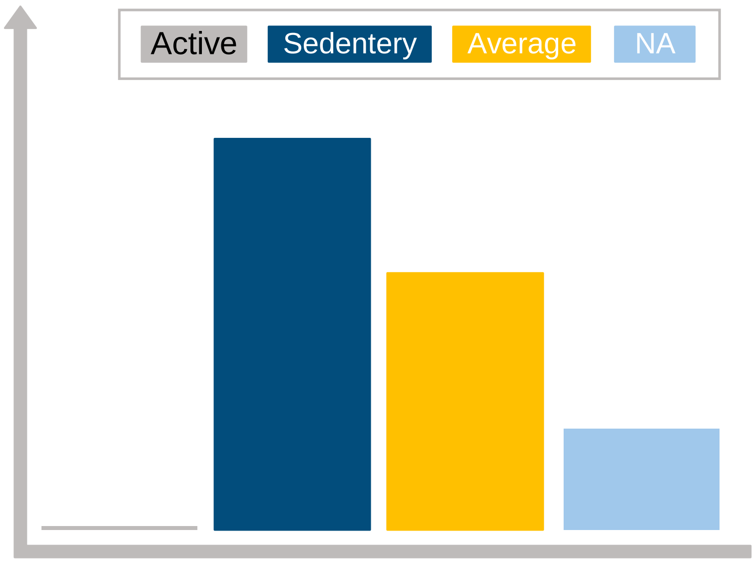
- cor()

	height	weight	bmi
height	1.000000	0.637640	0.052578
weight	0.637640	1.000000	0.795710
bmi	0.052578	0.795710	1.000000

- vif()

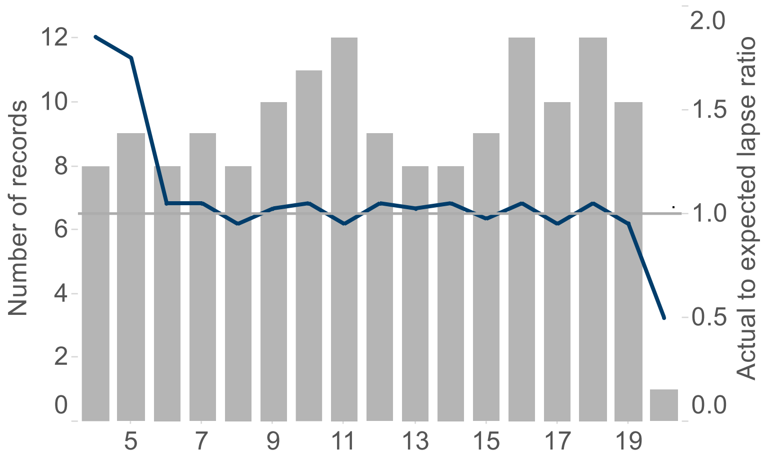
Reference levels



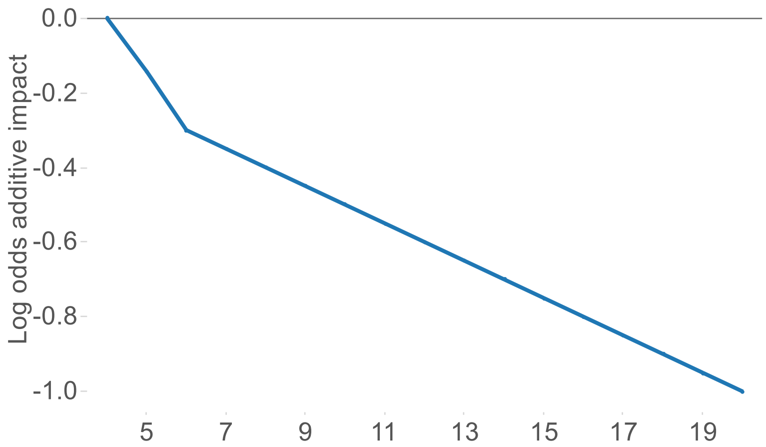


Piecewise linear effects

A/E by predictor before piecewise split



Piecewise impact of example predictor



Undersampling

- For logistic regression, undersampling can help improve runtimes:
  - All deaths (n) +
  - Randomly selected non-deaths (3n)
- Fitting the model Death ~ AttAge

Dataset	Records	Runtime	Intercept	AttAge coefficient
Full	259,284	2.15	-14.13	0.129
Undersampled	25,152	0.12	-10.99	0.123

Hands-on: Fit logistic GLM in R!

SOCIETY OF ACTUARIES

Hands-on: Practical concerns in R!

SOCIETY OF ACTUARIES

Validation

Validation and comparison

- Overall model fit (4A)
  - Bias-variance tradeoff
- Comparison between two candidate models (4B)

Model fit

- R^2
- Log-likelihood/AIC/BIC
- Actual-to-expected plots (4A-i)
- Confusion matrix (4A-ii)
- AUC (4A-iii)

Confusion matrix

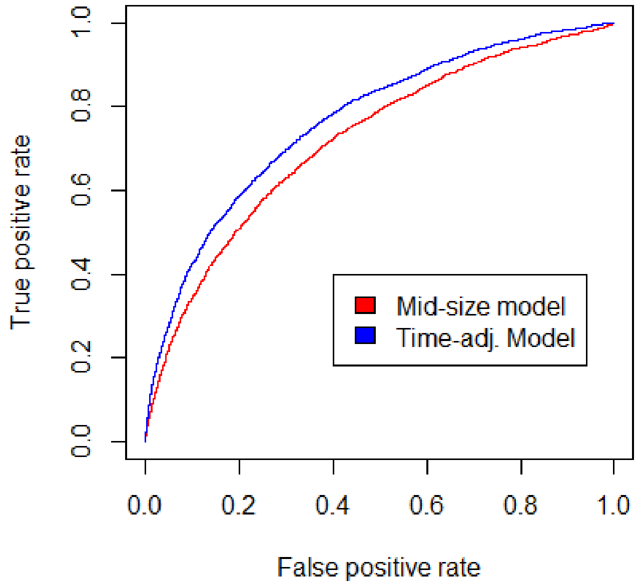
- Select a threshold for predicting the outcome
- Build a 2x2 contingency table

Prediction	Death		
	0	1	Total
0	65,815	835	66,650
1	18,500	1,313	19,813
Total	84,315	2,148	86,463

True positive rate = 1,313/2,148 = 0.658 (1 – Type-II error)
False positive rate = 18,500/84,315 = 0.301 (Type-I error)

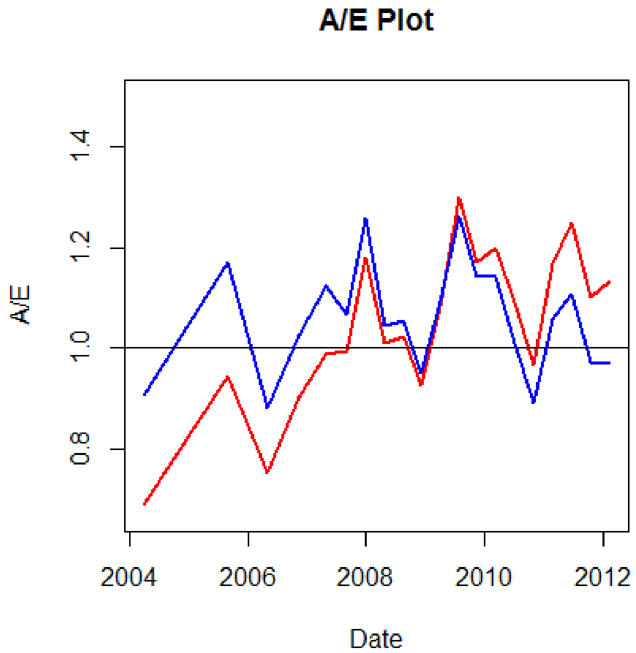
Area under the curve (AUC)

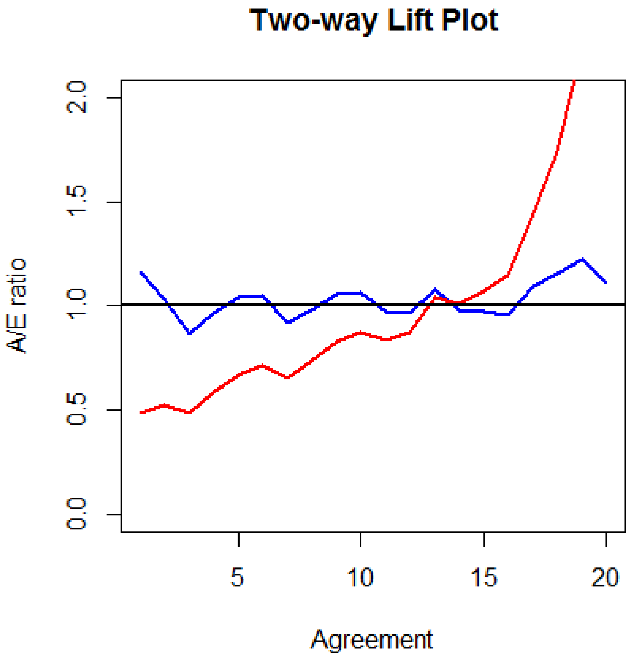
- The curve here is the relationship of the true positive rate and false positive rate as the threshold moves from 0 to 1



Model comparison: Lift charts

- Actual to expected (4B)
- Two-way lift (4B)





Hands-on: Validation in R!

Thank you!

SOCIETY OF ACTUARIES