Research Prototype: Lapse Analysis of Life Insurance Policies in Malaysia with Generalised Linear Models

Nicholas Yeo Chee Lek FIA FASM FSA

Actuarial Society of Malaysia

Founder & Actuary | Nicholas Actuarial Solutions

Chief | learn@AP | Actuarial Partners Consulting

Consulting Actuary | Sunway University Business School

E: nicholas.yeo@n-actuarial.com | T: +6 012 502 3566 | W: www.n-actuarial.com

Today's presentation

Part 1	Part 2
Lapse	Multicollinearity
Predictive Analytics	Over-dispersion
Research Prototype	Model Selection
Objectives	Model Diagnostics
	Actuarial Judgment

Part 1

Lapse

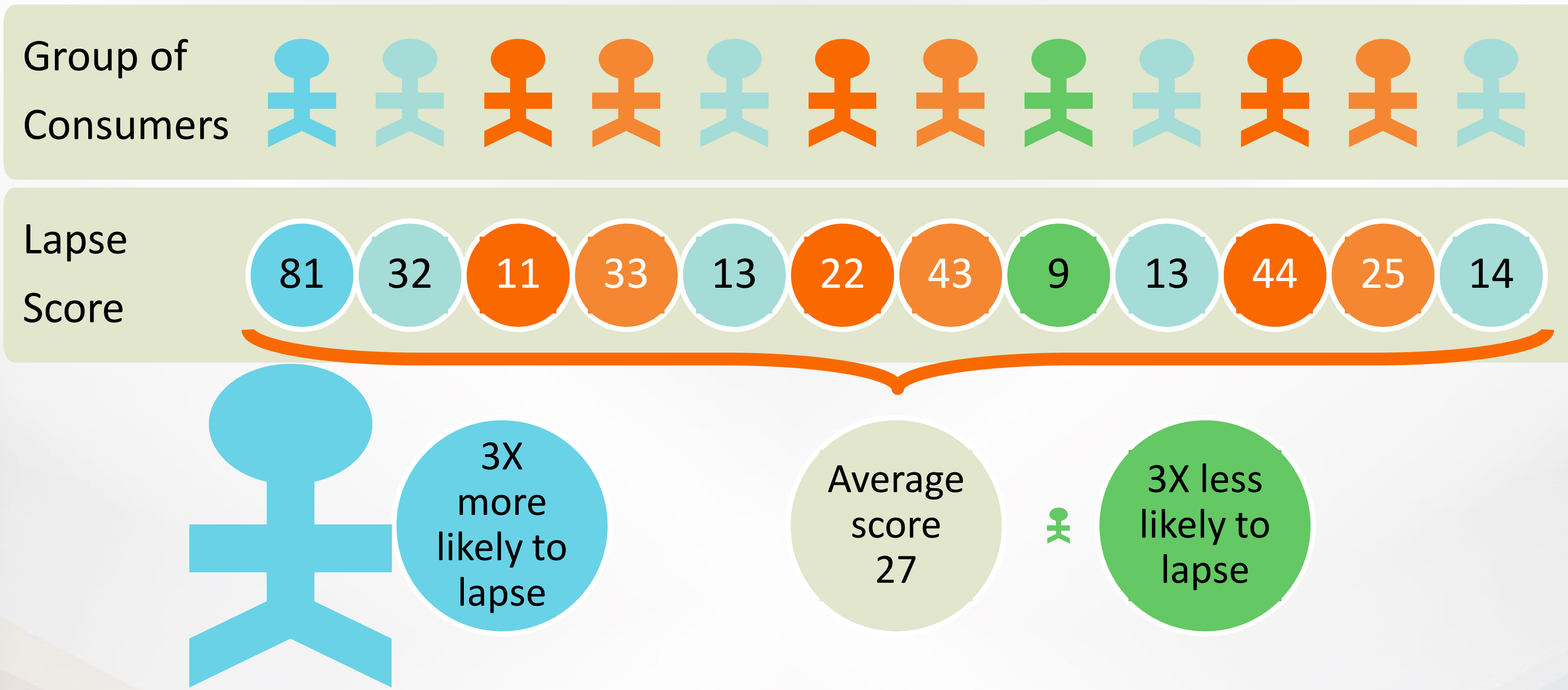
Under-addressed issue in life insurance

In Malaysia, approximately 1 policy lapse for every 2 new cases

Two steps forward, one step back

Similar story everywhere else in the world

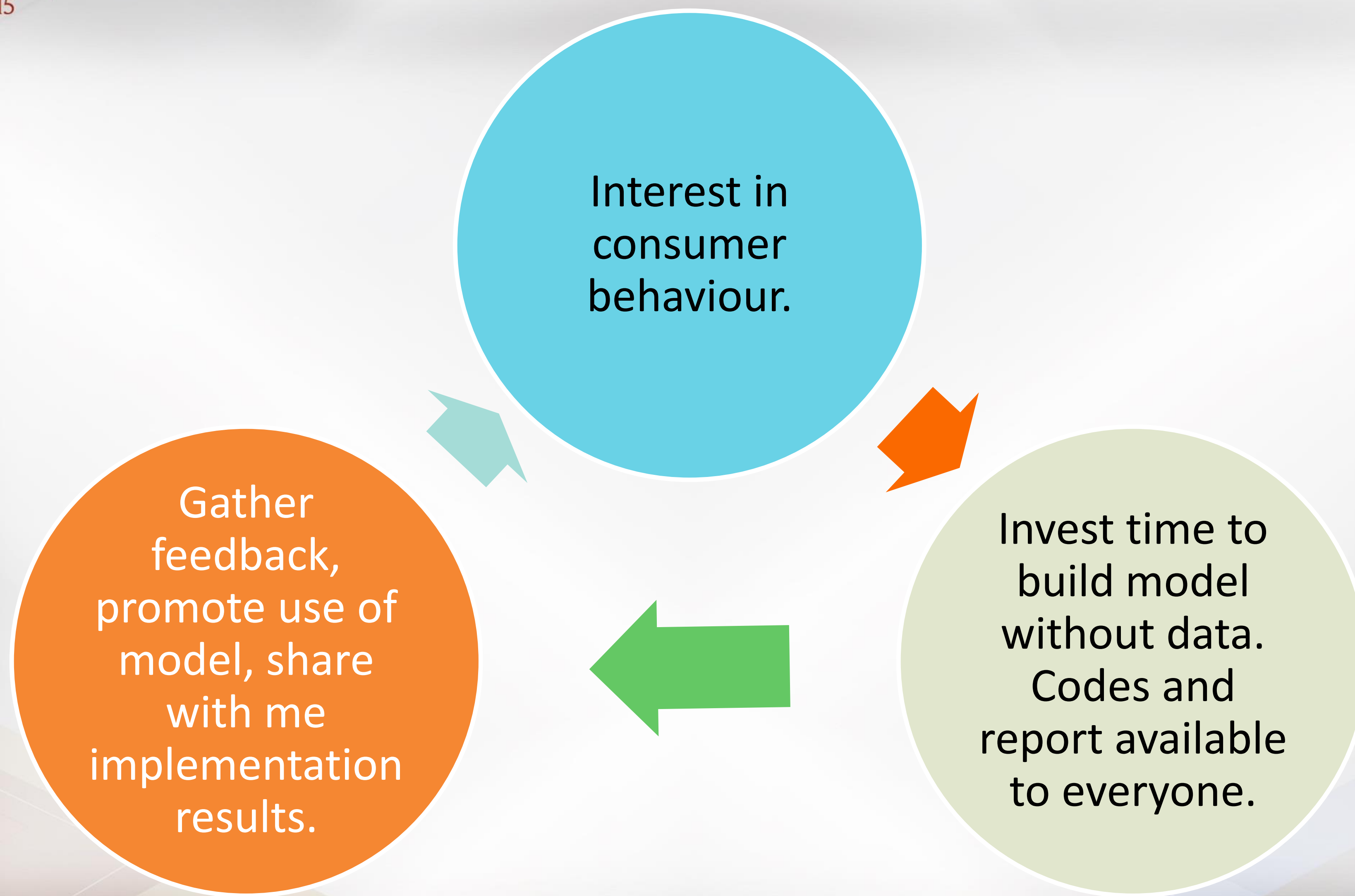
Predictive Analytics



Innovation
Creating Sustainable Value

Predictive Analytics

Objectives


Interest in consumer behaviour.
Invest time to build model without data. Codes and report available to everyone.
Gather feedback, promote use of model, share with me implementation results.

Part 2

GLMs

Generalised Linear Models (GLM) is a family of statistical models

$$Y = g^{-1}(X\beta)$$

$$\text{lapse} = g^{-1}(\text{data} \times \beta)$$

"data can explain lapses"

g is the "link function"

Exploratory Analysis

The following data fields were available:

Lapse	Exposure	Whole Life	Term	Others
Single Premium	Duration 0	Duration 1	Duration 2	Financial Year 2008
Financial Year 2009	Financial Year 2010	Financial Year 2011	Company	

Use histogram, density plots, boxplots and scatterplots etc high school statistics to visualise your data.

Multicollinearity

One weakness of basic GLM is that it cannot easily deal with multicollinearity between the explanatory variables i.e. the “data”.

There is no fixed rule to confirm multicollinearity problem or otherwise.

Examine Pearson’s coefficient of correlation for each pair and Variance Inflation Factor.

Pair Correlation

Examine Pearson's coefficient of coefficient for each pair of explanatory variables.

ot (products group others) and d0 (duration 0) have a correlation of 0.59.

Further consideration during modelling stage.

wl	-0.39	0.07	-0.12	0	-0.09	-0.07
-0.39	tm	-0.14	0.2	-0.25	0.08	0.08
0.07	-0.14	ot	0.11	0.59	-0.13	-0.11
-0.12	0.2	0.11	sp	0.03	0.05	-0.01
0	-0.25	0.59	0.03	d0	0.02	-0.03
-0.09	0.08	-0.13	0.05	0.02	d1	0.33
-0.07	0.08	-0.11	-0.01	-0.03	0.33	d2

Variance Inflation Factors

Explanatory Variable	VIF without company	VIF with company
Wl	1.2275	**93.7007**
Tm	1.3288	**87.0183**
Ot	1.6735	**44.3866**
Sp	1.1019	4.1908
d0	1.7133	**14.3445**
d1	1.2076	1.7100
d2	1.1929	1.9343

$VIF_i \geq 5$ indicates possible problem

$VIF_i \geq 10$ indicates almost certainly a problem

Clear that with explanatory variable company in the data it will create significant multicollinearity issues. We create two models, “company only variable model” and “all other variables model”.

Poisson Model

Lapse can be modelled as a count variable.

Use log link function.

Saturated model: $\log(lapse) = \beta_0 + \beta_2 wl + \beta_3 tm + \beta_4 ot + \beta_5 sp + \beta_6 d0 + \beta_7 d1 + \beta_8 d2 + \sum_i \beta_{9i} year_i + \log(exposure)$

Null model: $\log(lapse) = \beta_0 + \log(exposure)$

Poisson Model

Explanatory Variables	Intercept Value	Intercept P(>|z|)	Coefficient Value	Coefficient P(>|z|)	Residual Deviance	Deg. of Freedom	P(>X)	AIC
Null	-3.1142	<2e-16	NA	NA	370830	74	NA	371710
saturated	-2.7109	<2e-16			245520	63	<2e-16	246422
wl			-0.6079	<2e-16				
tm			-0.8822	<2e-16				
ot			-2.2799	<2e-16				
sp			0.0875	<2e-16				
d0			2.5126	<2e-16				
d1			-0.0353	0.0002				
d2			0.2756	<2e-16				
year1			-0.0487	<2e-16				
year2			-0.1282	<2e-16				
year3			-0.1436	<2e-16				
year4			-0.0915	<2e-16				

Overdispersion

Residual deviance ≈ residual degrees of freedom for a well-fitted model.

Overdispersion arise when residual deviance > residual degrees of freedom i.e. variance of the observations > variance implied by the model. Here overdispersion arise due to:

- the use of summarised data
- potentially more useful and precise explanatory variables e.g. target market, distribution channels, and conservation strategy, are not examined.

Overdispersion

Ways to deal with overdispersion

- refit the model with individual data
- refit model with better explanatory variables
- extend the model to a quasi-Poisson model (variance is a linear function of the mean, “technical fix”)
- use a negative binomial regression model (variance is a quadratic function of the mean, different likelihood function)

Quasi-Poisson Model

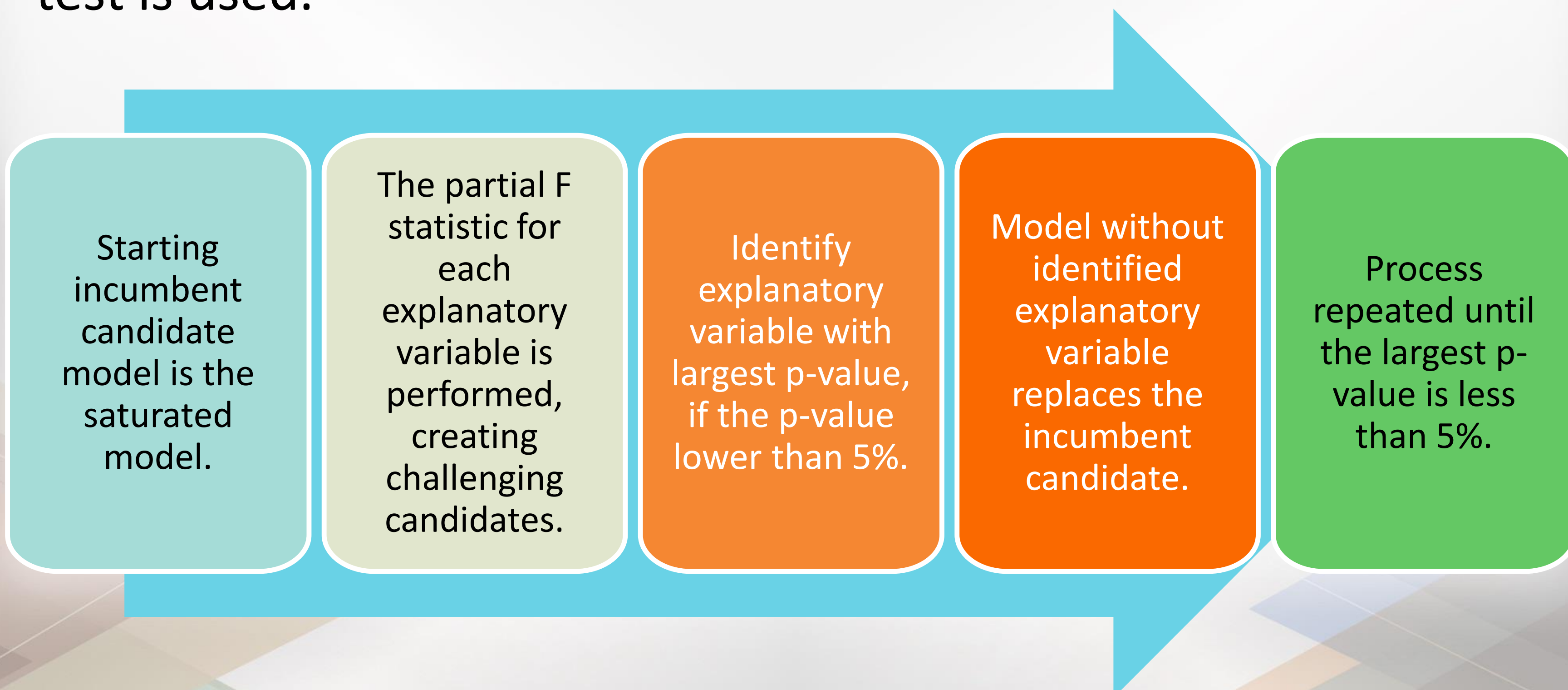
Explanatory Variables	Intercept Value	Intercept P(>|t|)	Coefficient Value	Coefficient P(>|t|)	Residual Deviance	Deg. of Freedom	P(>F)	Dispersion
Null	-3.1142	<2e-16	NA	NA	370830	74	NA	5470
saturated	-2.7109	<2e-16			245520	63	0.0041	3972
wl			-0.6079	0.0463				
tm			-0.8822	0.0007				
ot			-2.2799	0.0067				
sp			0.0875	0.6143				
d0			2.5126	0.0026				
d1			-0.0353	0.9542				
d2			0.2756	0.6289				
year1			-0.0487	0.7356				
year2			-0.1282	0.3793				
year3			-0.1436	0.3225				
year4			-0.0915	0.5159				

Had we used the Poisson model, the predictive power would have been overstated.

Model Selection

Many different approaches to perform model selection.

Here a stepwise backwards elimination algorithm using the F-test is used.



Innovation
Creating Sustainable Value

Stepwise Backwards Partial F-test Algorithm

Iteration	Explanatory Variables	Residual Deviance	Deg. of Freedom	P(>F)	Action
1	none	245520			
	wl	261473	1	0.0047	
	tm	294659	1	0.0007	
	ot	275594	1	0.0072	
	sp	246531	1	0.6124	
	d0	278312	1	0.0005	
	d1	245533	1	0.9537	remove
	d2	246434	1	0.6298	
	year	250843	4	0.8490	
2	none	245533			
	wl	261573	1	0.0450	
	tm	294705	1	0.0007	
	ot	276484	1	0.0060	
	sp	246535	1	0.6111	
	d0	279423	1	0.0042	
	d2	246881	1	0.5554	
	year	250858	4	0.8452	remove

Innovation
Creating Sustainable Value

Stepwise Backwards Partial F-test Algorithm

Iteration	Explanatory Variables	Residual Deviance	Deg. of Freedom	P(>F)	Action
3	none	250858			
	wl	268286	1	0.0332	
	tm	302957	1	0.0004	
	ot	282919	1	0.0044	
	sp	251179	1	0.7688	remove
	d0	284990	1	0.0033	
	d2	253555	1	0.3955	
4	None	251179			
	wl	269525	1	0.0280	
	tm	303196	1	0.0003	
	ot	283359	1	0.0041	
	d0	285887	1	0.0029	
	d2	253959	1	0.3852	remove
5	None	253959			
	wl	271847	1	0.0296	
	tm	304112	1	0.0004	
	ot	290429	1	0.0023	
	d0	294118	1	0.0014	

Stepwise Backwards Partial F-test Algorithm

Explanatory Variables	Intercept Value	Intercept P(>|t|)	Coefficient Value	Coefficient P(>|t|)	Residual Deviance	Deg. of Freedom	P(>F)	Dispersion
Backwards	-2.7469	<2e-16			253959	70	2.6e-5	3700
wl			-0.6250	0.0288				
tm			-0.8666	0.0004				
ot			-2.3221	0.0023				
d0			2.5742	0.0007				

The backwards elimination algorithm yielded:

$$\frac{lapse}{exposure} = e^{-2.7469}e^{-0.6250wl}e^{-0.8666tm}e^{-2.3221ot}e^{2.5742d0}$$

Stepwise Backwards Partial F-test Algorithm

However, the coefficient for d0 is very high, which suggests:

First year policies have $e^{2.5742}$ = 1312% higher lapse rates than other year policies

Lapse rate of $e^{-2.7469}e^{2.5742} = 84.1\%$ for first year endowment policies

Recall ot and d0 have a high Pearson correlation coefficient, and this has manifested into an unsatisfactory model. Consider dropping ot and/or d0.

Applying Actuarial Judgment

Explanatory Variables	Intercept Value	Intercept P(>|t|)	Coefficient Value	Coefficient P(>|t|)	Residual Deviance	Deg. of Freedom	P(>F)	Dispersion
Drop d0	-2.4144	<2e-16			294118	71	0.0015	4472
wl			-0.9900	0.0010				
tm			-0.9357	0.0004				
ot			-0.7734	0.3187				
Drop ot	-2.5850	<2e-16			290429	71	0.0007	4212
wl			-0.9707	0.0008				
tm			-0.8702	0.0009				
d0			0.8259	0.1261				
Drop both	-2.4418	<2e-16			299293	72	0.0008	4499
wl			-1.0693	0.0002				
tm			-0.9240	0.0005				

Drop d0 is a weak candidate as coefficient ot has a high p-value.

Judgment made to select drop ot instead of drop both as drop ot has higher functionality with an extra coefficient.

Final Quasi-Poisson Model

$$\frac{lapse}{exposure} = e^{-2.5850} e^{-0.9707wl} e^{-0.8702tm} e^{0.8259d0}$$

Multiplicative table :

Base Lapse Rate	
Base Lapse Rate	7.54%

X

Product Type	
Whole Life	0.38
Endowment and Others	1.00
Term	0.42

X

Policy Duration	
First Year	2.28
Subsequent Years	1.00

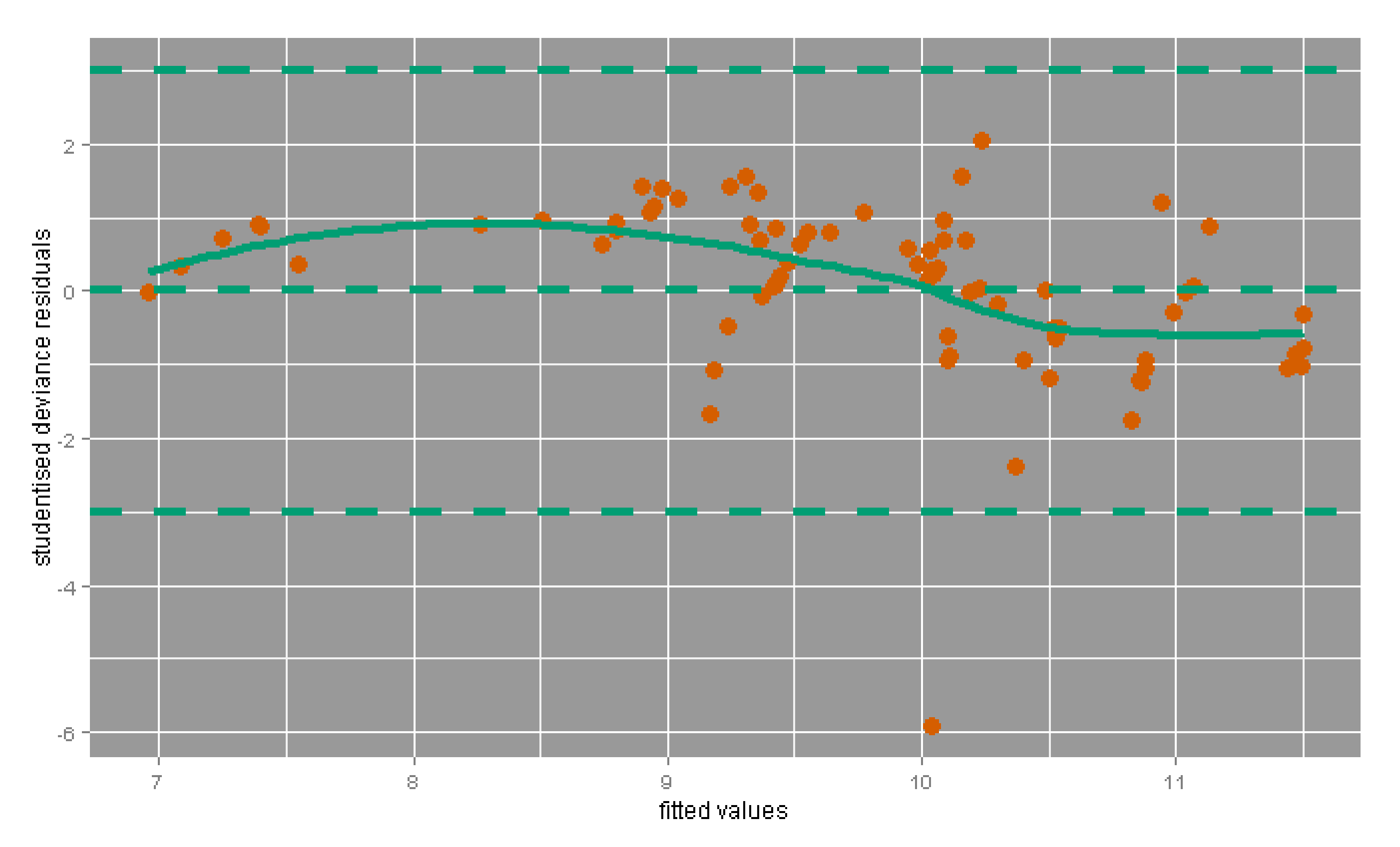
Innovation Creating Sustainable Value
19th AAC BANGKOK • THAILAND 2015

Model Diagnostics

Diagnostic tests, accompanied by generally accepted rule of thumbs, indicate where further investigations are required.

- Studentised deviance residuals – model assumptions
- Hat diagonals – observed response value to fitted value
- Cook's distance – observation on fitted values & coefficients
- COVRATIO – observation on variance & covariance of coefficients
- DFFITS – observation on fitted values
- DFBETA – observation on each coefficients & intercept

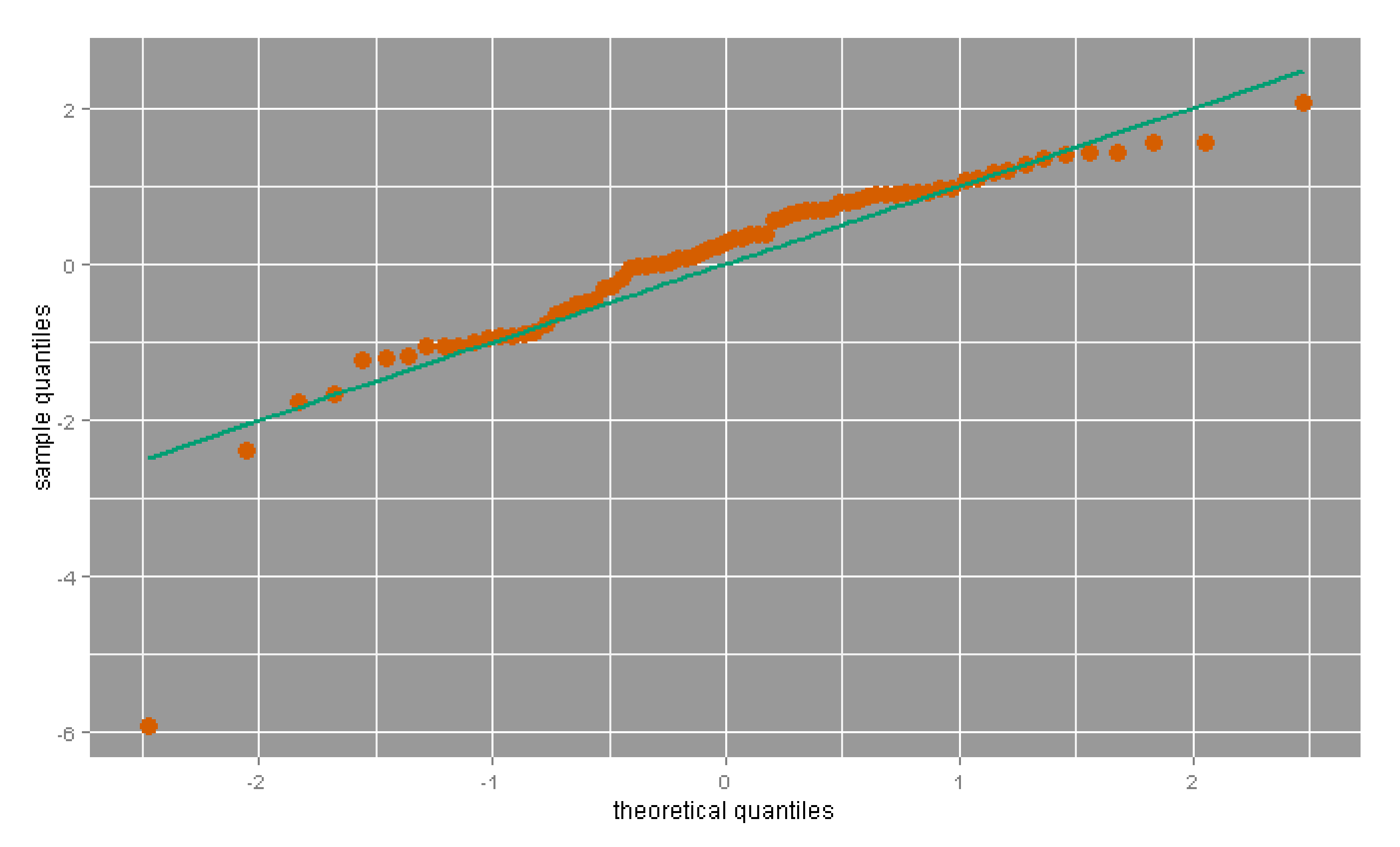
Studentised Deviance Residuals Scatterplot



Roughly evenly distributed around zero, no specific patterns

Values more than 3 are generally considered as outliers

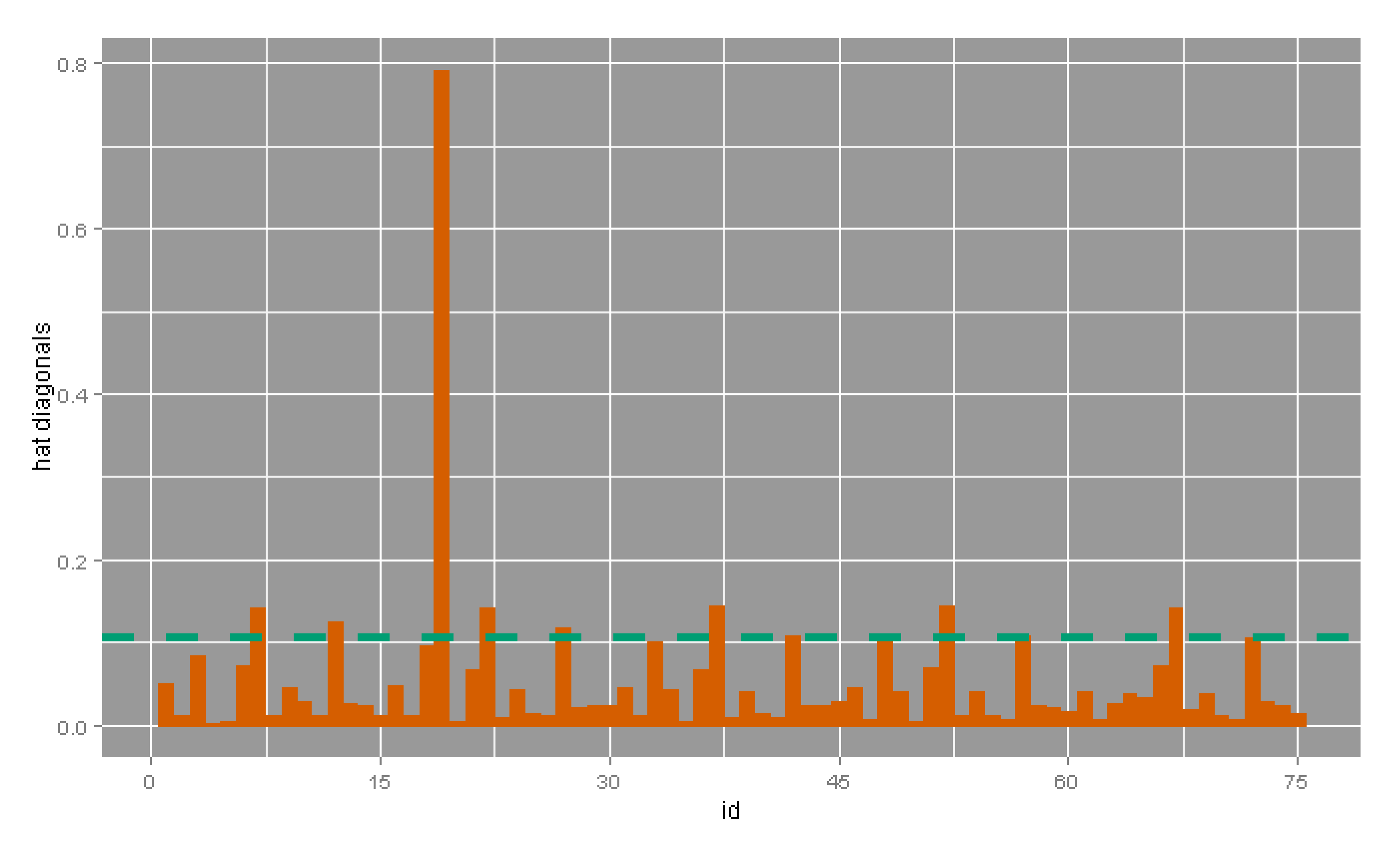
Studentised Deviance Residuals QQ-plot



Approximately normally distributed

Deviation for tail values are common

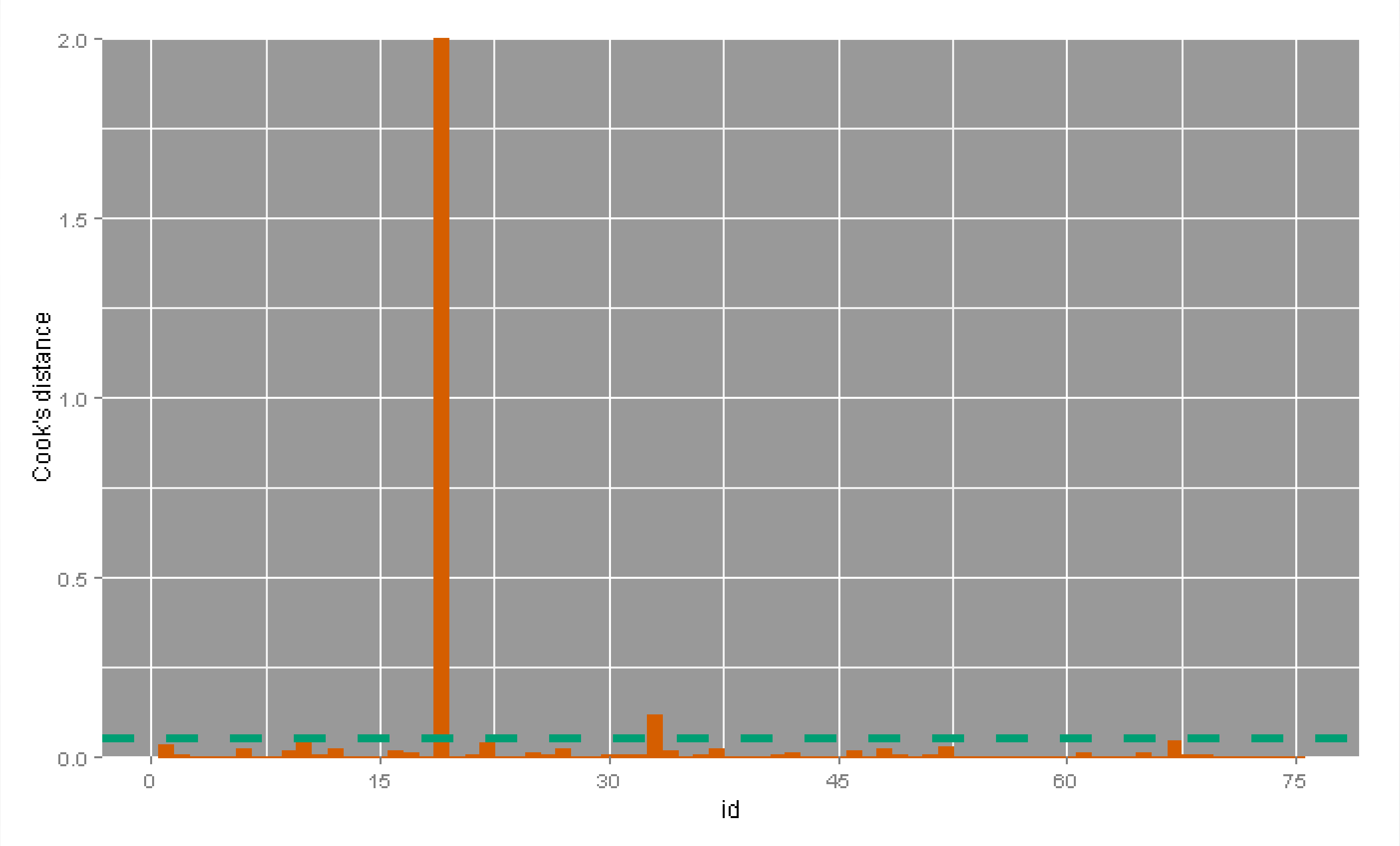
Hat Diagonals



Highly influential observations have hat diagonals larger than

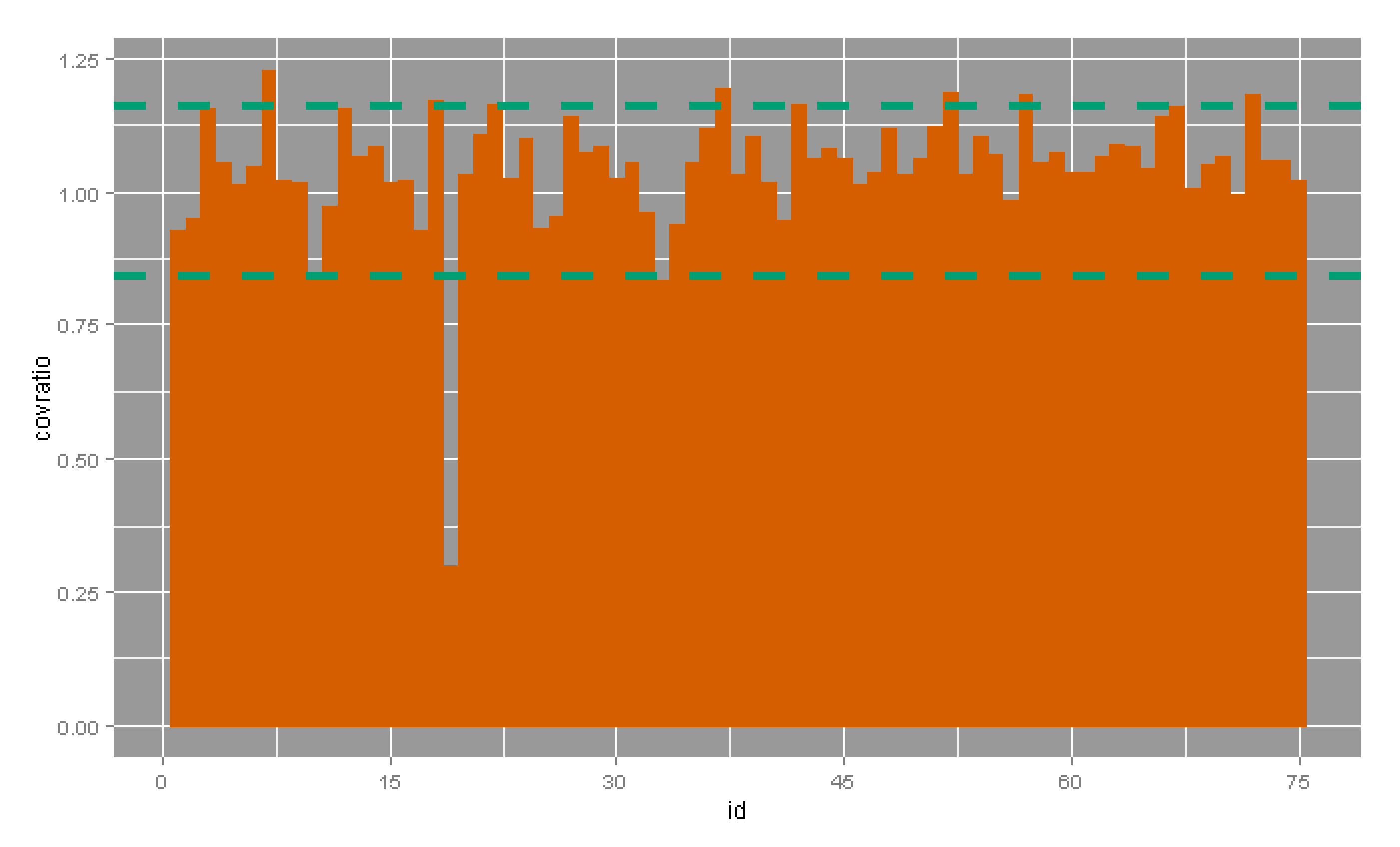
$$\frac{2 \times (\text{number of observations} - \text{ residual degrees of freedom})}{\text{number of observations}}$$

Cook's Distance



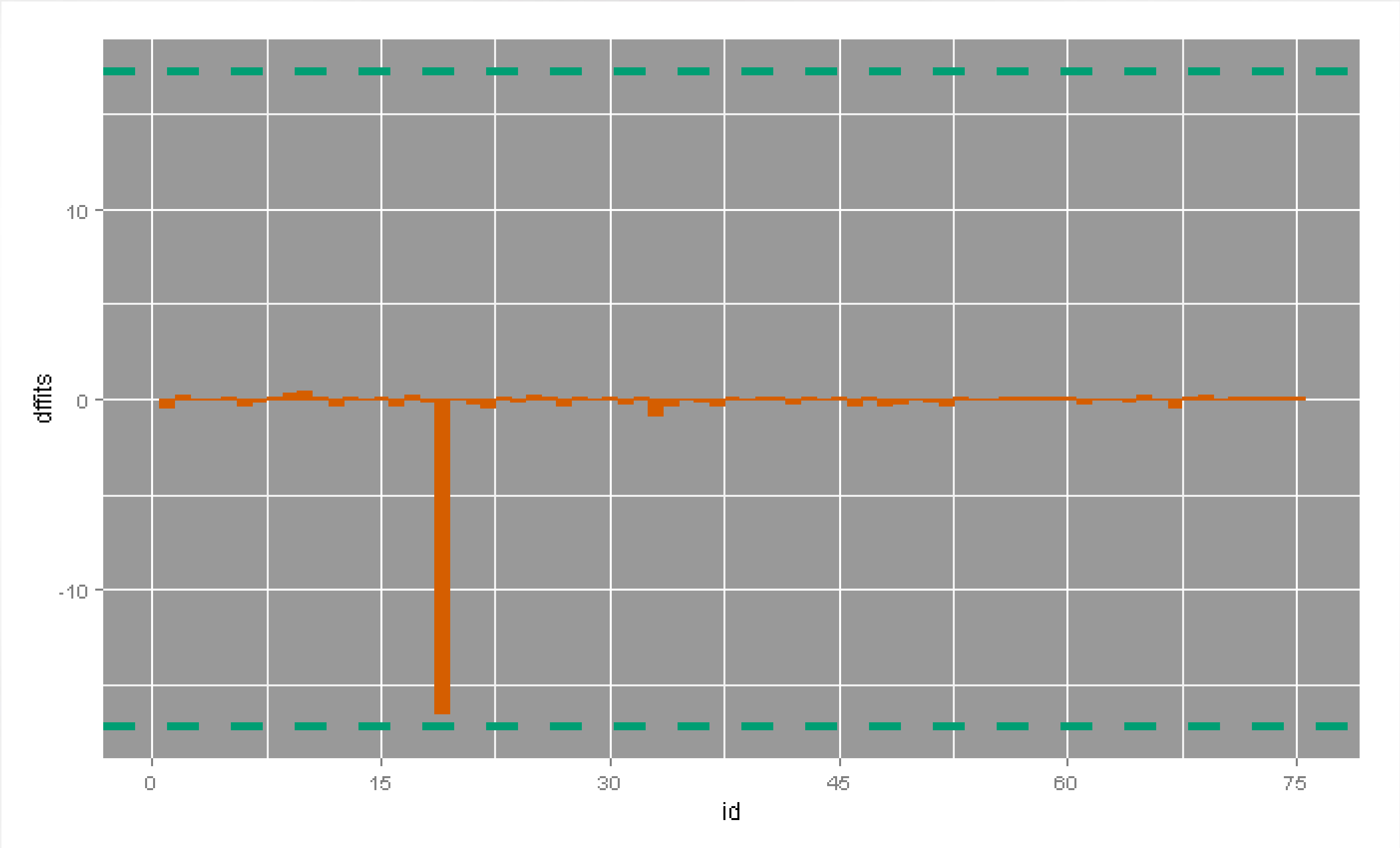
Highly influential observations have Cook's distance value higher than $\frac{4}{\text{number of observations}}$

COVRATIO



Highly influential observations are outside $1 \pm 3 \times \frac{(number\ of\ observations - residual\ degrees\ of\ freedom)}{number\ of\ observations}$

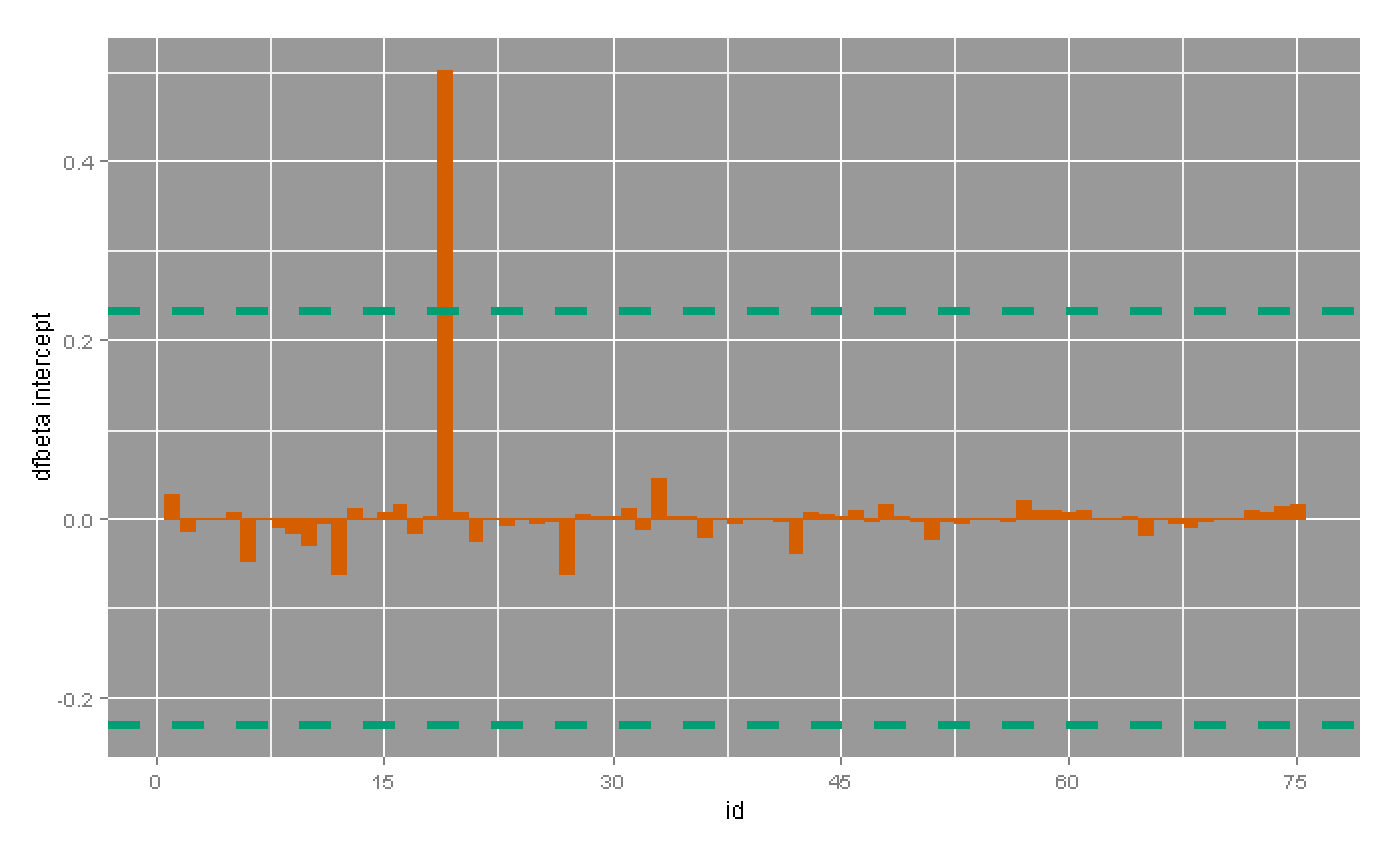
DFFITS



Highly influential observations have DFFITS outside $\pm 2 \times$

$$\sqrt{\frac{(number\ of\ observations - residual\ degrees\ of\ freedom)}{number\ of\ observations}}$$

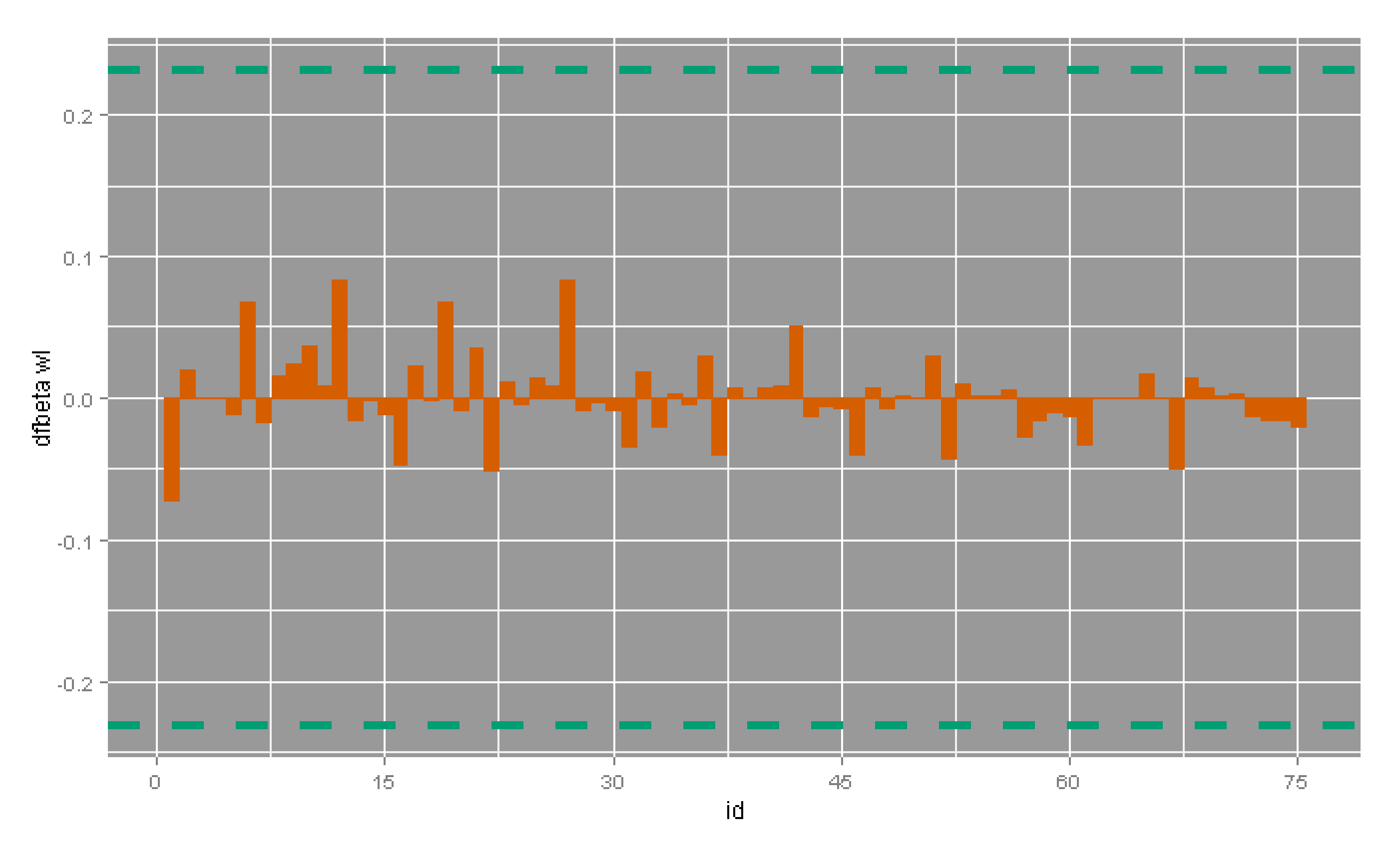
DFBETA Intercept



Highly influential observations have DFBETA outside

$$\pm \frac{2}{\text{number of observations}}$$

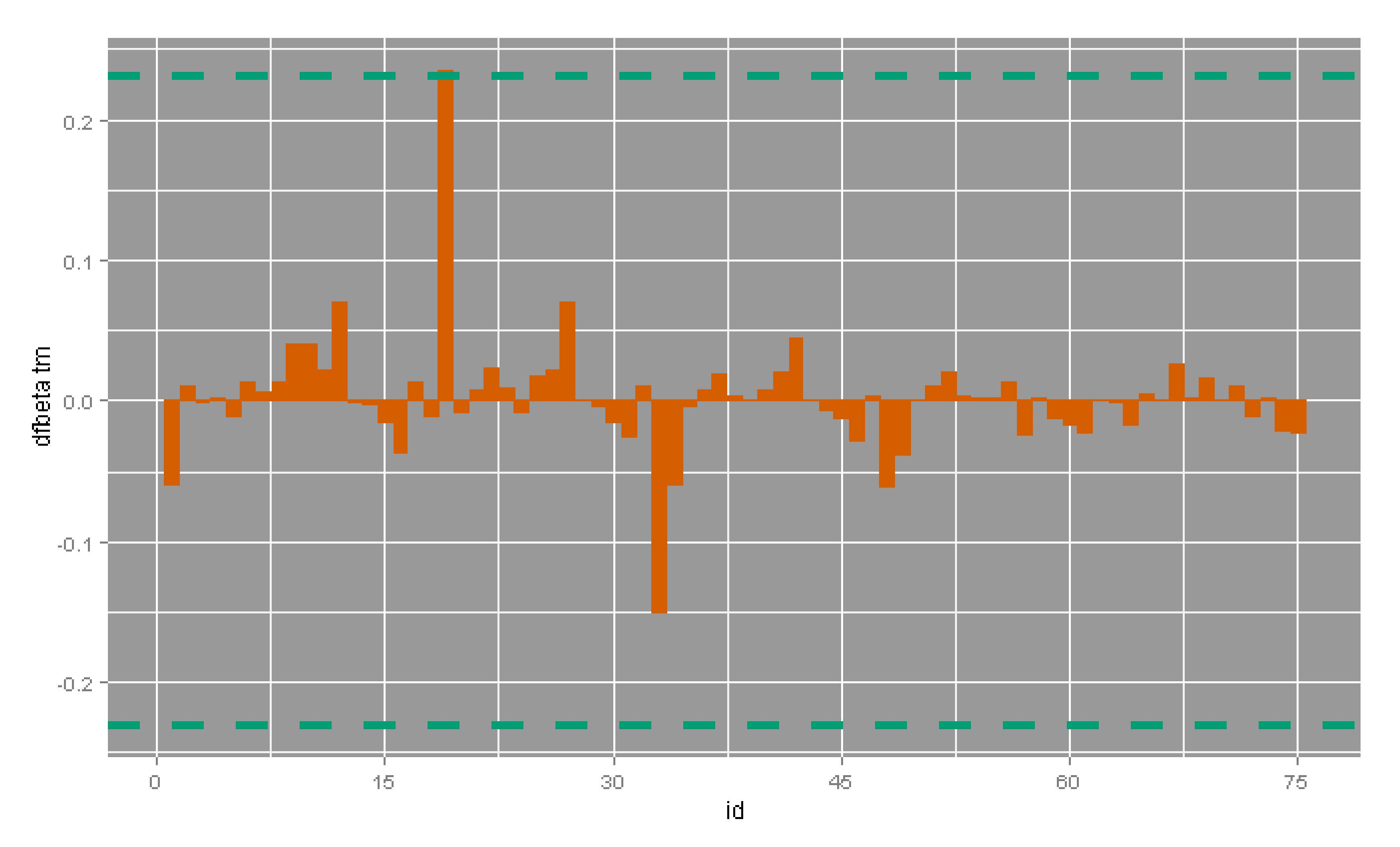
DFBETA wl



Highly influential observations have DFBETA outside

$$\pm \frac{2}{\text{number of observations}}$$

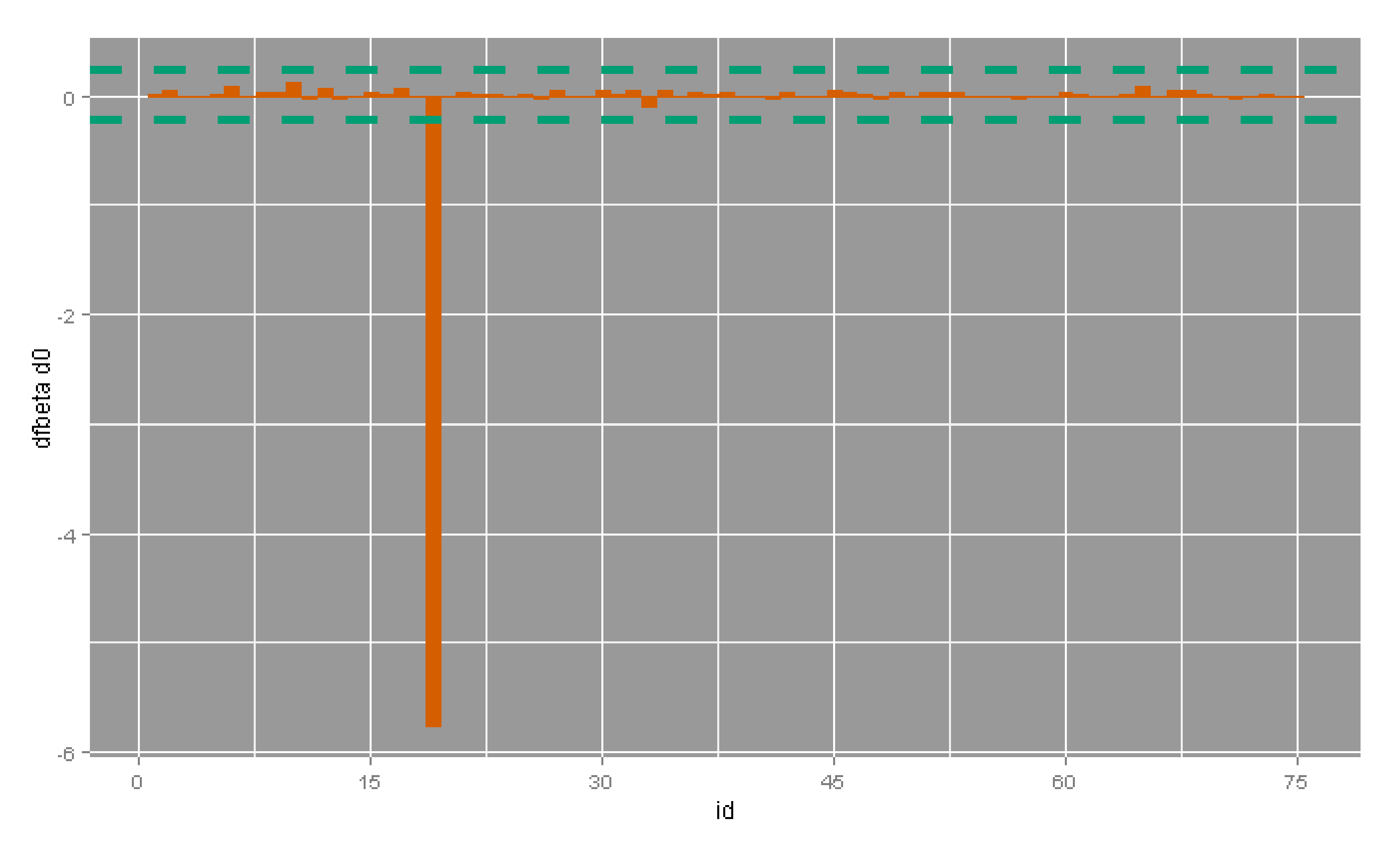
DFBETA tm



Highly influential observations have DFBETA outside

$$\pm \frac{2}{\text{number of observations}}$$

DFBETA d0



Highly influential observations have DFBETA outside

$$\pm \frac{2}{\text{number of observations}}$$

Negative Binomial Model

Similar to Poisson model i.e. lapse is modelled as a count variable, same log link function.

Main difference Poisson requires variance = mean but negative binomial only requires variance as a quadratic function of the mean.

Hence different likelihood function.

Innovation
Creating Sustainable Value

19th AAC BANGKOK • THAILAND 2015

Negative Binomial Model

Explanatory Variables	Intercept Value	Intercept P(>|z|)	Coefficient Value	Coefficient P(>|z|)	Residual Deviance	Deg. of Freedom	P(>X)	AIC	Dispersion
null	-2.7453	<2e-16	NA	NA	79.094	74	NA	1621.8	3.0393
saturated	-2.4589	<2e-16			77.152	63	1.9e-7	1590.9	5.8395
wl			-0.0919	0.7599					
tm			-1.2438	2.10e-8					
ot			-3.4574	1.31e-8					
sp			0.1118	0.5171					
d0			1.7782	0.0007					
d1			-0.4115	0.3343					
d2			0.1450	0.7169					
year1			0.0763	0.6210					
year2			-0.0071	0.9634					
year3			-0.1309	0.4013					
year4			-0.1025	0.5058					

Model Selection

Another model selection approach is to use the stepwise backwards AIC algorithm.

AIC is used to compare between models, rule of thumb is that, all else being equal, the model with a lower AIC is better.

1. Starting incumbent candidate model is the saturated model.
2. Challenging candidates are models each with one less explanatory variable than the incumbent candidate.
3. Identify challenging candidate with lowest AIC, if AIC lower that of incumbent candidate.
4. Model without lowest AIC replaces the incumbent candidate.
5. Process repeated until incumbent candidate has the lowest AIC.

Innovation
Creating Sustainable Value

Stepwise Backwards AIC Algorithm

Iteration	Explanatory Variables	AIC	Action
1	none	1588.9	
	wl	1587.0	
	tm	1609.3	
	ot	1606.2	
	sp	1587.4	
	d0	1597.7	
	d1	1587.3	
	d2	1587.0	
	year	1583.3	remove
2	none	1583.3	
	wl	1581.3	
	tm	1602.3	
	ot	1601.1	
	sp	1581.7	
	d0	1593.3	
	d1	1581.5	
	d2	1581.3	remove

Stepwise Backwards AIC Algorithm

Iteration	Explanatory Variables	AIC	action
3	none	1581.3	
	wl	**1579.3**	remove
	tm	**1600.3**	
	ot	**1599.1**	
	sp	**1579.7**	
	d0	**1591.3**	
	d1	**1579.5**	
4	none	1579.3	
	tm	**1598.8**	
	ot	**1598.1**	
	sp	**1577.8**	
	d0	**1589.1**	
	d1	**1577.6**	remove
5	none	1577.6	
	tm	**1596.9**	
	ot	**1596.1**	
	sp	**1576.0**	remove
	d0	**1587.5**	

Stepwise Backwards AIC Algorithm

Iteration	Explanatory Variables	AIC	action
6	none	1576.0	
	tm	1595.0	
	ot	1594.2	
	d0	1585.7	

The backwards elimination algorithm yielded:

$$\frac{lapse}{exposure} = e^{-2.5630} e^{-1.1534tm} e^{-3.4043ot} e^{1.8449d0}$$

However, the coefficient for d0 is slightly high, implying first policy year lapse rate is $e^{2.5742}$ = 633% higher than other policy years. Again judgment is required.

Innovation
Creating Sustainable Value

Recap

Give individual consumers a lapse score.

This gives insights for more effective conservation actions.

Model lapse as a count variable. Start with a Poisson model.

Use quasi-Poisson or negative binomial due to overdispersion.

Use partial F-test for quasi-Poisson, AIC for negative binomial.

Apply judgment. Analyse diagnostics.

Also Available in the Paper

Assessment of Model Lift

Lapse modelled as a binary variable with binomial model

Manipulation of summarised industry data

Company only model – biproduct of multicollinearity

Accompanying R-codes for generating results and graphs

Innovation
Creating Sustainable Value

Thank You

Nicholas Yeo Chee Lek FIA FASM FSA
Actuarial Society of Malaysia

Founder & Actuary | Nicholas Actuarial Solutions
Chief | learn@AP | Actuarial Partners Consulting
Consulting Actuary | Sunway University Business School

E: nicholas.yeo@n-actuarial.com | T: +6 012 502 3566 | W: www.n-actuarial.com