Ch 10: Linear mixed and generalized linear mixed models

Outline

1. Two industrial problems
2. Simplified analysis of Problem 1 using regression
3. Simplified solution of Problem 2 using generalised linear models
4. Generalized linear mixed models
5. Proper solution of Problem 1 using linear mixed models
6. Proper solution of Problem 2 using generalized linear mixed models
7. Reference

1.1 Problem 1: Big Data

- In regional monopolies such as the electricity and water industries customer satisfaction scores form the basis of regulatory fines (Brint and Fry, 2021)
- The whole system is potentially unfair if there is a detectable regional bias i.e. a regional effect over and above the actual level of service received
- Want to try and measure regional bias using data from `TripAdvisor`
- The data recorded are the hotel, the rating (out of 5, the higher better)
- In this lecture the data are treated as if they are a continuous (numerical) variable

  - **Subtle point. This may not actually be strictly correct but makes it easier to use and interpret the model (shows important aspects of the subject are not just purely mathematical though are probably inherently quantitative)**

1.2 Problem 2: Big Data

- Scores take the values 1:5
  - High 4-5
  - Medium 3
  - Low 1-2
- **This lecture**
- Want to investigate how the probability a survey respondent gives a high/medium/low score depends on the region of the survey respondent

2.1 An (overly) simple model for regional bias

- Is there evidence for a regional bias in this data set?
- Regression model

$$\text{score}_i = \alpha + \beta \text{Region}_i + \text{error}$$

- **Subtleties**

  - for the moment we ignore the effect of different hotels (in practice this would be a simple mistake to make that would give dangerously plausible suggested answers to the question at hand)

  - If the β terms in the regression are significant then we have statistical evidence of a regional effct

  - The β terms help us to quantify the practical effect of any regional bias

2.2 Regression model fitting in R

- **Think R=robot!**
- Usually what you have is separate commands
  1. Run the analysis (command=`lm`)
  2. Display the results (command=`summary`)
- The default in R is to fit a constant term though this is not explicit in the R syntax
- This can be surpressed using a -1 term in the sequence of X-variables.
- Usually you include a constant term but occasionally it is easier to present results without this (this is one of the mistakes we perhaps made in our paper Brint and Fry, 2019 so this can be quite subtle)

2.3 Simplified regression treatment of Problem 1

- Data in the file `ScoreData.xlsx` gives the ratings of 24,365 reviewers for a range of different product ("hotels")
- Could argue over whether this really is "Big Data" but my view would be that this is big enough to be non-trivial and informative
- Some of the data regarding the regional affiliation of the survey respondent is missing and data corresponding to these entries needs to be deleted
- The easiest way of processing this data in R is
  1. Copy the data into notepad
  2. Read the data into R using the command `read.table`
  3. Use the command line in R to organise the data as required
     - Note that the size of the dataset means MS excel is impractical even though the structure of the data is so straightforward

2.4 Reading data into R

- Easiest to save the data file directly onto a USB to make the file path shorter

```
scores<-read.table(‘‘G:ScoreData.txt")
```

- If you were to type `scores` and press enter this would print the data you have just entered in (but probably best not to do this!)
- Next you need to define the variables one by one

```
rating<-scores[,1]
product<-scores[,2]
region<-scores[,3]
regionmissing<-scores[,4]
```

2.5 Processing the data in R

• Next need to delete entries for unassigned regions. This is easiest to do in R using its vectorised nature and the subsetting commands as follows:

```
 rating<-rating[regionmissing<1]
product<-product[regionmissing<1]
region<-region[regionmissing<1]
```

• Need to tell R to put the data in the right form. The data for product and region are labels rather than numerical measurements and R needs to be told this. In R the command to do this is `factor`

```
product<-factor(product)
region<-factor(region)
```

2.6 Analysis of data in R

- **As this is a non-trivial problem worthwhile to look at the regional data to see if we might have made a mistake**
- **In R the command to do this is summary**

```
summary(region)
EasMid LonEsx NE Yor NorWes Scotlan South SouWes
Wales WesMid
2781 3012 3250 3446 1984 3375 1945
1387 2301
```

2.7 Regression analysis in R

- **Think R=robot**
- Two steps
  1. Run the analysis (command=`lm`)
  2. Display the results (command=`summary`)

1. Run the analysis

```
a.lm< -lm(rating~region)
```

2. Display the results

```
summary(a.lm)
```

2.8 Regression results in R I

```
 Coefficients:
Estimate Std.  Error t value Pr(>|t|)
(Intercept) 3.989932 0.020858 191.288 <2e-16 ***
regionLonEsx -0.064965 0.028927 -2.246 0.0247 *
regionNEYor -0.026855 0.028414 -0.945 0.3446
regionNorWes 0.003394 0.028039 0.121 0.9037
regionScotlan -0.013117 0.032325 -0.406 0.6849
regionSouth 0.031994 0.028170 1.136 0.2561
regionSouWes 0.028577 0.032514 0.879 0.3794
regionWales 0.018720 0.036158 0.518 0.6047
regionWesMid 0.024845 0.030998 0.801 0.4229
```

2.9 Interpreting the regression results I

- The convention would be it is the second row of this table downwards that is interesting
- This is because we usually fit a constant or underlying average term to the model. This means that even in the absence of a regional difference we do not expect the average score to be equal to zero
- Taking these results at face value there is evidence $p = 0.0247$ that the results for those from the London region are lower than for the other regions
- **The suggestion would be that electricity and water companies that are based in the London region would be systematically disadvantaged by the current system of regulation and the way that customer satisfaction scores determine fines.**

2.10 Re-parameterising the model

• In this case the estimated effects of the regional bias can be better presented by re-parameterising the model.

• In order to do this we can suppress the constant term in the fitting of the regression model. In R this is achieved by including a -1 term into the list of X-variables

1. Run the analysis

```
b.lm< −lm(rating∼region-1)
```

2. Display the results

```
summary(b.lm)
```

2.11 Regression results in R II

```
 Coefficients:  Estimate Std.  Error t value Pr(>|t|)
regionEasMid 3.98993 0.02086 191.3 <2e-16 ***
regionLonEsx 3.92497 0.02004 195.8 <2e-16 ***
regionNEYor 3.96308 0.01929 205.4 <2e-16 ***
regionNorWes 3.99333 0.01874 213.1 <2e-16 ***
regionScotlan 3.97681 0.02469 161.0 <2e-16 ***
regionSouth 4.02193 0.01893 212.4 <2e-16 ***
regionSouWes 4.01851 0.02494 161.1 <2e-16 ***
regionWales 4.00865 0.02954 135.7 <2e-16 ***
regionWesMid 4.01478 0.02293 175.1 <2e-16 ***
```

2.12 Interpreting the regression results II

• These numerical values should be the same under either parameterisation

• **You have to use the first parameterisation to test for regional differences**

- In this case the significant values indicate that the London scores are lower than the benchmark score for the first East Midlands category

• **The second parameterisation is the easiest way to present the estimated average score**

- In this case the significant values indicate that in none of the regions is the average rating equal to zero

- The numerical estimates obtain suggestion only minor differences in the average scores given by customers from different regions though this isn't formally tested for

• **The estimated customer satisfaction scores should be the same under either parameterisation**

2.13 Different parameterisations give the same answer

- Concentrate for sake of argument on the London region
- The second parameterisation gives that in this case the average customer satisfaction score is 3.92497
- Under the first parameterisation the average customer satisfaction score for the London region would be given by

$$\begin{aligned} \text{Average score} &= \text{Intercept} + \text{London Adjustment} \\ &= 3.989932 - 0.064965 = 3.924967 \end{aligned}$$

3.1 Commands in R – reading in the data

- From last time

```
scores<-read.table("E:ScoreData.txt")
ratings<-scores[,1]
rating<-scores[,1]
product<-scores[,2]
region<-scores[,3]
regionmissing<-scores[,4]
rating<-rating[regionmissing<1]
product<-product[regionmissing<1]
region<-region[regionmissing<1]
```

3.2 R subsetting commands

- In R generate the sequence of high values using
  `high< −1*(rating>3)`
- In R generate the sequence of medium values using
`medium< −1*(rating==3)`
- In R generate the sequence of low values using
`low< −1*(rating<2)`
- **In each case this generates a sequence of values that take the value 1 if the observation belongs to that category (e.g. high) and 0 otherwise**

3.3 Further considerations for probability models in R

• **To fit a binomial glm in R you need the data organised in columns of successes and failures**
• Strictly speaking binomial models are formulated for occasions when you have only two categories (e.g. yes/no, successful/unsuccessful, male/female etc.)
• Our Big Data example with three categories can be modelled using 3-1=2 equations
• A third equation can be fitted but should be redundant
• In practice it might be still worth fitting this third equation just to check that the rest of our conclusions are still in order

3.4 Data structure for probability models in R

• **To fit a binomial glm in R you need the data organised in columns of successes and failures**

• The R command needed to do this is `cbind` which has the effect of binding the required counts of successes and failures together

• For our data example in R use

```
yhigh<-cbind(high, 1-high)
ylow<-cbind(low, 1-low)
ymedium<-cbind(medium, 1-medium)
```

3.5 Fitting binomial generalised linear models in R

- The basic set of commands works as follows
- **Compute the model**

```
high1.glm< −glm(yhigh∼region, family=binomial)
high2.glm< −glm(yhigh∼region,
family=binomial(link=probit))
```

- **Summarise the results**

```
summary(high1.glm)
summary(high2.glm)
```

- **These models can serve as a cross-check of each other in applications. Should expect to have similar models giving you similar interpretations and numerically similar estimates**

3.6 Probability somebody gives a high score – logit model

• Probability is calculated using the R commands on the previous slides

```
 Estimate Std.  Error z value Pr(>|z|)
(Intercept) 1.046081 0.043232 24.197 <2e-16 ***
regionLonEsx -0.114166 0.059219 -1.928 0.0539 .
regionNEYor -0.053799 0.058553 -0.919 0.3582
regionNorWes 0.002846 0.058133 0.049 0.9610
regionScotlan -0.039457 0.066637 -0.592 0.5538
regionSouth 0.075978 0.058890 1.290 0.1970
regionSouWes 0.075283 0.068138 1.105 0.2692
regionWales 0.067004 0.075778 0.884 0.3766
regionWesMid 0.018630 0.064408 0.289 0.7724
```

• **Weak evidence ($p = 0.0539$) that those in the London region may be less likely to give a high score as the coefficient is statistically significant and negative**

3.7 Probability somebody gives a high score – probit model

- Probability is calculated using the R commands on the previous slides

```
 Estimate Std.  Error z value Pr(>|z|)
(Intercept) 0.643412 0.025643 25.091 <2e-16 ***
regionLonEsx -0.068089 0.035304 -1.929 0.0538 .
regionNEYor -0.031994 0.034813 -0.919 0.3581
regionNorWes 0.001688 0.034477 0.049 0.9610
regionScotlan -0.023449 0.039614 -0.592 0.5539
regionSouth 0.044895 0.034808 1.290 0.1971
regionSouWes 0.044487 0.040233 1.106 0.2689
regionWales 0.039611 0.044744 0.885 0.3760
regionWesMid 0.011040 0.038165 0.289 0.7724
```

- **Probit model leads to the same interpretation as the logit model**
- **Weak evidence ($p = 0.0538$) that those in the London region may be less likely to give a high score as the coefficient is statistically significant and negative**

3.8 Numerical calculation – proof that similar models should give similar numbers

• **Example. Using the logit model calculate the probability that somebody from the London region gives a high score**

$$\begin{aligned}
\log\left(\frac{p}{1-p}\right) &= 1.046081 - 0.114166 = 0.931915 \\
\frac{p}{1-p} &= \exp(0.931915) = 2.539367413 \\
p &= (1-p)2.539367413 \\
&= \frac{2.539367413}{3.539367413} \\
&= 0.717463635 = 0.717 \text{ (3 d.p.)}
\end{aligned}$$

3.9 Numerical calculation – proof that similar models should give similar numbers

• **Example. Using the probit model calculate the probability that somebody from the London region gives a high score**

$$\begin{aligned} Z^{-1}(p) &= 0.643412 - 0.068089 \\ &= 0.575323 = 0.58 \text{ (2 d.p.)} \end{aligned}$$

• Keep this calculation to 2dp because of the limited resolution of the tables

$$p = Z(0.58) = 0.71904$$

• **Results thus give very small numerical differences between logit and probit models**

3.10 What is the probability that somebody gives a low score?

- The basic set of commands works as follows
- **Compute the model**

```
low1.glm< −glm(ylow∼region, family=binomial)
low2.glm< −glm(ylow∼region,
family=binomial(link=probit))
```

- **Summarise the results**

```
summary(low1.glm)
summary(low2.glm)
```

- **These models can serve as a cross-check of each other in applications. Should expect to have similar models giving you similar interpretations and numerically similar estimates**

3.11 Probability somebody gives a low score – logit model

• Probability is calculated using the R commands on the previous slides

```
Estimate Std.  Error z value Pr(>|z|)
(Intercept) -3.14340 0.09524 -33.005 <2e-16 ***
regionLonEsx 0.18064 0.12719 1.420 0.156
regionNEYor 0.02004 0.12919 0.155 0.877
regionNorWes -0.01845 0.12851 -0.144 0.886
regionScotlan -0.01334 0.14813 -0.090 0.928
regionSouth 0.02539 0.12796 0.198 0.843
regionSouWes 0.11757 0.14395 0.817 0.414
regionWales 0.16271 0.15733 1.034 0.301
regionWesMid -0.11741 0.14590 -0.805 0.421
```

• **If we take this at face value no evidence ($p > 0.05$) for regional differences in the extent to which people award low scores**

3.12 Probability somebody gives a low score – probit model

• Probability is calculated using the R commands on the previous slides

```
Estimate Std.  Error z value Pr(>|z|)
(Intercept) -1.735207 0.042645 -40.689 <2e-16 ***
regionLonEsx 0.081926 0.057602 1.422 0.155
regionNEYor 0.008987 0.057924 0.155 0.877
regionNorWes -0.008251 0.057477 -0.144 0.886
regionScotlan -0.005966 0.066254 -0.090 0.928
regionSouth 0.011390 0.057387 0.198 0.843
regionSouWes 0.053084 0.065063 0.816 0.415
regionWales 0.073699 0.071491 1.031 0.303
regionWesMid -0.052143 0.064720 -0.806 0.420
```

• **If we take this at face value no evidence ($p > 0.05$) for regional differences in the extent to which people award low scores**

3.13 What is the probability that somebody will sit on the fence?!

- Classical binomial models are constructed for the special case of two categories (yes/no, male/female, successful/unsuccessful etc.)
- In the case of three categories here (high scores, low scores, medium scores) we have used two regression equations to separately estimate the probability of obtaining a high score and the probability of obtaining a low score
- Though, mathematically, this should be a redundant step it is instructive to check this and fit the third regression equation to explain the probability that somebody will award a medium score
- Mathematically, if you have n categories would need $n - 1$ probability calculations
- Mathematically, the technically correct analysis would use multinomial regression models corresponding to multiple (> 2) categories (not discussed here)

3.14 Computing the probability of a medium score

- The basic set of commands works as follows
- **Compute the model**

```
medium1.glm< −glm(ylow∼region, family=binomial)
medium2.glm< −glm(ylow∼region,
family=binomial(link=probit))
```

- **Summarise the results**

```
summary(medium1.glm)
summary(medium2.glm)
```

- **These models can serve as a cross-check of each other in applications. Should expect to have similar models giving you similar interpretations and numerically similar estimates**

3.15 Probability of a medium score – logit model

- **These R commands give**

```
 Estimate Std.  Error z value Pr(>|z|)
(Intercept) -1.655e+00 5.168e-02 -32.035 <2e-16 ***
regionLonEsx 3.642e-02 7.125e-02 0.511 0.609
regionNEYor -9.664e-03 7.050e-02 -0.137 0.891
regionNorWes -9.264e-02 7.048e-02 -1.314 0.189
regionScotlan 1.793e-02 7.980e-02 0.225 0.822
regionSouth -7.973e-02 7.068e-02 -1.128 0.259
regionSouWes -1.273e-01 8.272e-02 -1.539 0.124
regionWales -6.257e-02 9.089e-02 -0.688 0.491
regionWesMid -6.615e-05 7.680e-02 -0.001 0.999
```

- **No evidence ($p > 0.05$) that the probability of awarding a medium score depends on the region**

3.16 Probability of a medium score – probit model

- **These R commands give**

```
 Estimate Std.  Error z value Pr(>|z|)
(Intercept) -9.929e-01 2.856e-02 -34.772 <2e-16 ***
regionLonEsx 2.017e-02 3.946e-02 0.511 0.609
regionNEYor − 5.337e − 033.894e − 02 − 0.1370.891
regionNorWes -5.088e-02 3.873e-02 -1.314 0.189
regionScotlan 9.917e-03 4.416e-02 0.225 0.822
regionSouth_4.383e − 023.886e − 02 − 1.1280.259
regionSouWes -6.978e-02 4.524e-02 -1.542 0.123
regionWales_3.443e − 024.994e − 02 − 0.6890.491
regionWesMid -3.655e-05 4.244e-02 -0.001 0.999
```

- **Probit model leads to the same interpretation as the logit model**
- **No evidence ($p > 0.05$) that the probability of awarding a medium score depends on the region**

4. Generalized linear mixed models

- Linear mixed and generalized linear mixed models
  - "Mixed" term indicates correlation problems caused by the sampling structure
- Same basic interpretation of previous models but ...
  - account for correlations caused by repeat observations and the nature of the data collection
  - should lead to more precise estimates of the regional effect
- Interpretation of the model and calculations involving the fixed effects terms are essentially the same as before
- Section 5 below presents a professional-standard solution to Problem 1
- Section 6 below presents a professional-standard solution to Problem 2

5.1 Analysis and optional packages in R

- Want to fit a linear mixed effects model in R
- To do this you have to download the R package `lme4` – which stands for linear mixed effects
- In R to upload packages use

`Packages` $\longrightarrow$ `load packages` $\longrightarrow$ `lme4` $\longrightarrow$ `OK`

5.2 Installing non-standard packeages from CRAN

- In R to see what packages are available for loading use

`Packages`$\longrightarrow$`Load packages`

- If the package is not there you might have to download the required package from CRAN
- To do this use

`Packages`$\longrightarrow$`Install package(s) ...`$\longrightarrow$`Choose a CRAN mirror`

- Better to choose the UK (London or Bristol) or wherever you are in the world
- You should then be able to see a long list of packages that are available for download
- **This online community and long list of written packages is really the best thing about R**

5.3 Potential problems with the university computer network

- On my work computer standard packages load fine
- On your personal computer should be able to update the list of packages via the R repository CRAN
- This maintenance of packages and an active online R community is probably the best thing about R and the best reason to use it
- However, it may not be possible to update packages directly on a PC on the university network
- **The work-around this is to save the package code to a USB stick and then use the option**

`Packages`$\longrightarrow$`Install package(s) from local files ...`

- **Best thing might be to bring your laptop into class and see if we can get R working ...**
- **On my laptop I found loading R packages fine but fiddlier than it should have been ...**

5.4 Analysis of problem 1

- Analysis in R proceeds in 4 steps
  1. Download the `lme4` package
  2. Fit a model with no regional effect but correlation caused by repeated measures
  3. Fit a model with regional effect but correlation caused by repeated measures
  4. Use a chi-squared test to test for a regional effect by distinguishing between the models fitted in Steps 2-3

5.5 Problem solution in R

Step 2

```
ab1<-lmer(rating~1+(1|product), REML=F)
```

Step 3

```
ab2<-lmer(rating~region+(1|product), REML=F)
```

Explanation of R code

- R command `lmer` stands for Linear Mixed Effects Regression
- As before `rating` is the y-variable we are trying to model.
- Model divides into a **fixed effect** (constant term or a term adjusting for different regional effects) ...
- and a separate **mixed effect** `+(1|product)`
- Need to suppress the REML estimation method in order to run a standard maximum likelihood ratio (χ^2) test

5.6 The problem with mixed effects models

- Mixed effects problems are hard
- Problem structures may be intricate dependant upon the complexity of the dataset and how the data has been collected
- In my simple example

`+(1|product)`

- This means there is an average rating associated with each product (hotel) irrespective of who reviews it
- This is a simple illustrative example (albeit one that does stem from a practical industrial problem)

5.7 Properly testing for a regional effect

- **R command to run a chi-squared test is** `anova`

```
anova(ab1, ab2)
Data:  NULL
Models:
ab1:  rating ~ 1 + (1 | product)
ab2:  rating ~ region + (1 | product)
Df AIC BIC logLik deviance Chisq Chi Df Pr(>Chisq)
ab1 3 68975 69000 -34485 68969
ab2 11 68948 69036 -34463 68926 43.825 8 6.141e-07
***
```

- **Results give significant evidence ($p = 6.141 \times 10^{-7}$) of a regional effect**

5.8 Presenting results in R

- Basic procedure in R works the same way
  1. Run a computational calculation
  2. Show the results but only if explicitly directed by the user
- Remember that R is like a robot (it will do exactly what you tell it to do!)
- In R use the command `summary` to present the results

5.9 Presenting the results in R

```
summary(ab2)
Fixed effects:
             Estimate Std.  Error t value
(Intercept) 3.965336 0.048813 81.235
regionLonEsx -0.119061 0.027666 -4.304
regionNEYor -0.003653 0.027090 -0.135
regionNorWes 0.005216 0.026742 0.195
regionScotlan 0.057297 0.030884 1.855
regionSouth -0.014028 0.026868 -0.522
regionSouWes 0.008482 0.031007 0.274
regionWales 0.014557 0.034456 0.422
regionWesMid 0.016191 0.029519 0.548
```

5.10 Interpretation

• Interpretation and calculations work the same way as before though this time the results should be more robust
• Evidence of a London effect

$$t = |-4.304| = 4.304 > 2;\ p < 0.05$$

• No evidence of an effect for other regions. E.g. for Scotland

$$t = 1.855 < 2;\ p > 0.05$$

• London term is negative and statistically significant. Reaffirms previous suggestion that London companies may be at a disadvantage.

5.11 Question: how to adjust survey scores out of 5 for regional bias I

$$\text{East Midlands} = \text{Intercept} = 3.965336$$

$$\begin{aligned} \text{London} &= \text{Intercept} - 0.119061 \\ &= 3.965336 - 0.119061 = 3.846275 \end{aligned}$$

$$\begin{aligned} \text{North East/Yorkshire} &= \text{Intercept} - 0.003653 \\ &= 3.965336 - 0.003653 = 3.961683 \end{aligned}$$

$$\begin{aligned} \text{North West} &= \text{Intercept} + 0.005216 \\ &= 3.965336 + 0.005216 = 3.970552 \end{aligned}$$

$$\begin{aligned} \text{Scotland} &= \text{Intercept} + 0.057297 \\ &= 3.965336 + 0.057297 = 4.022633 \end{aligned}$$

5.12 Question: how to adjust survey scores out of 5 for regional bias II

$$\begin{aligned} \text{South} &= \text{Intercept} - 0.014028 \\ &= 3.965336 - 0.014028 = 3.951308 \end{aligned}$$

$$\begin{aligned} \text{South West} &= \text{Intercept} + 0.008482 \\ &= 3.965336 + 0.008482 = 3.973818 \end{aligned}$$

$$\begin{aligned} \text{Wales} &= \text{Intercept} + 0.14557 \\ &= 3.965336 + 0.14557 = 4.110906 \end{aligned}$$

$$\begin{aligned} \text{West Midlands} &= \text{Intercept} + 0.016191 \\ &= 3.965336 + 0.016191 = 3.981527 \end{aligned}$$

6.1 Analysis of problem 2 in R

- The basic R command needed is `glmmPQL` in the R package `MASS`
- Essentially what you end up with is a repetition of the logistic regression models of the previous lecture
- There is a further adjustment to reflect the repeated measurements from individual hotels (exactly as we did in the first part of the lecture)

Explanation of R commands

- `glmmPQL` stands for Generalised Linear Mixed Models fitted via Penalized Quasi Likelihood

6.2 Three basic problems

1. Regional effect in the probability of giving a high score?
2. Regional effect in the probability of giving a low score?
3. Regional effect in the probability of giving a medium score?

6.3 Testing for regional effects in the probability of giving a high score

1. Compute the model
   `high2< −glmmPQL(high ∼ region, random = ∼ 1 | product,family=binomial)`
2. Show the results using the command `summary`
   `suumary(high2)`

6.4 R commands and the structure of the model

- Generalized linear mixed models are the harder version of generalized linear models
- As such there are similarities in the statistical interpretation and R commands of the model

`random = ~ 1 | product` reflects the same correlation structure as the linear mixed models discussed previously

`family=binomial` need to specify the distributional family as was the case with logit and probit models

- Numerical examples work in the same way as logistic regression (see below)

6.5 Fixed effects in the probability of giving a high score

```
(Intercept) 1.0760804 0.09293285 23427 11.579118
0.0000
regionLonEsx -0.2214280 0.06144537 23427 -3.603656
0.0003
regionNEYor -0.0109988 0.06057182 23427 -0.181582
0.8559
regionNorWes 0.0102060 0.06014998 23427 0.169675
0.8653
regionScotlan 0.0811167 0.06892505 23427 1.176883
0.2393
regionSouth -0.0064580 0.06077871 23427 -0.106255
0.9154
regionSouWes 0.0426359 0.07024479 23427 0.606962
0.5439
regionWales 0.0589384 0.07794753 23427 0.756129
0.4496
regionWesMid 0.0020951 0.06645689 23427 0.031526
```

6.6 Interpretation

- Retain previous evidence of a London effect ($p = 0.003 < 0.05$)
- Coefficient of London is negative and statistically significant
- Suggests London respondents are less likely to award high marks
- No evidence for any other regional effects $t < 2,\ p > 0.05$)
- It is noteworthy that in this example both linear mixed and generalized linear mixed models point to the same London effect
- In this case similar models should serve as a cross-check of each other

6.7 Numerical example I: Calculate the probability a London respondent gives a high score

Set

$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \text{Regression equation} \\
&= 1.0760804 - 0.2214280 = 0.8546524 \\
\frac{p}{1-p} &= \exp(0.8546524) = 2.350557185 \\
p &= \frac{2.350557185}{3.350557185} = 0.701542177 = 0.702 \text{ (3 d.p.)}
\end{aligned}$$

6.8 Numerical example II: Calculate the probability a Scotland respondent gives a high score

Set

$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \text{Regression equation} \\
&= 1.0760804 + 0.0811167 = 1.1571971 \\
\frac{p}{1-p} &= \exp(1.1571971) = 3.181004731 \\
p &= \frac{3.181004731}{4.181004731} = 0.76082304 = 0.761 \text{ (3 d.p.)}
\end{aligned}$$

• Sanity check. Calculated probability is higher for Scotland than for London

6.9 Testing for regional effects in the probability of giving a low score

1. Compute the model
   `low2<-glmmPQL(low ~ region, random = ~ 1 | product,family=binomial)`
2. Show the results using the command `summary`
   `summary(low2)`

6.10 Fixed effects in the probability of giving a low score

```
Fixed effects:  low ~ region
                Value Std.Error DF t-value p-value
(Intercept) -3.361186 0.1479072 23427 -22.724961
0.0000
regionLonEsx 0.278707 0.1263615 23427 2.205636 0.0274
regionNEYor -0.046161 0.1276307 23427 -0.361674
0.7176
regionNorWes -0.045858 0.1270027 23427 -0.361076
0.7180
regionScotlan -0.184987 0.1463020 23427 -1.264419
0.2061
regionSouth 0.145506 0.1264916 23427 1.150325 0.2500
regionSouWes 0.175125 0.1423506 23427 1.230239 0.2186
regionWales 0.196901 0.1554111 23427 1.266972 0.2052
regionWesMid -0.089078 0.1438397 23427 -0.619284
0.5357
```

6.11 Interpretation

- Retain previous evidence of a London effect ($p = 0.0274$)
- Coefficient of London is positive and statistically significant
- Suggests London respondents are more likely to award low marks
- No evidence for any other regional effects ($t < 2,\ p > 0.05$)
- Look for similar models serving as a cross-check of each other
- Three linear mixed and generalized linear mixed models now all point to the same London effect

6.12 Numerical example I: Calculate the probability a London respondent gives a low score

Set

$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \text{Regression equation} \\
&= -3.361186 + 0.278707 = -3.082479 \\
\frac{p}{1-p} &= \exp(-3.082479) = 0.045845464 \\
p &= \frac{0.045845464}{1.045845464} = 0.043835792 = 0.044 \text{ (3 d.p.)}
\end{aligned}$$

6.13 Numerical example II: Calculate the probability a Scotland respondent gives a low score

Set

$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \text{Regression equation} \\
&= -3.361186 - 0.184987 = -3.546173 \\
\frac{p}{1-p} &= \exp(-3.546173) = 0.028834779 \\
p &= \frac{0.028834779}{1.028834779} = 0.028026637 = 0.028 \text{ (3 d.p.)}
\end{aligned}$$

• Sanity check. Calculated probability is lower for Scotland than for London

6.14 Final cross-check

- Recall we have 3 categories so only need 2 regression equations to describe the system
- As a cross-check fir a generalized linear mixed model to estimate the probability of awarding a medium score
- Technically, this should be a redundant step so no new regional effects should be identified at this stage

6.15 Testing for regional effects in the probability of giving a medium score

1. Compute the model
   ```
   medium2< −glmmPQL(medium ∼ region, random = ∼ 1
   | product,family=binomial)
   ```
2. Show the results using the command `summary`
   ```
   summary(medium2)
   ```

6.16 Fixed effects in the probability of giving a medium score

```
(Intercept) -1.6659027 0.07813808 23427 -21.319984
0.0000
regionLonEsx 0.1104438 0.07210366 23427 1.531737
0.1256
regionNEYor -0.0389842 0.07111155 23427 -0.548212
0.5836
regionNorWes -0.0954841 0.07111438 23427 -1.342684
0.1794
regionScotlan -0.0474800 0.08058379 23427 -0.589200
0.5557
regionSouth -0.0297179 0.07128690 23427 -0.416877
0.6768
regionSouWes -0.1083780 0.08335193 23427 -1.300246
0.1935
regionWales -0.0585106 0.09148767 23427 -0.639546
0.5225
```

6.17 Interpretation

- No evidence ($t < 2,\ p > 0.05$) of a regional effect in this case
- Consistent with the previous slides we find no evidence of any further regional effects
- Though technically a redundant step this serves as a cross-check of our earlier results and suggests we can be more confident that our earlier analysis is correct

6.18 Numerical example I: Calculate the probability a London respondent gives a medium score

Set

$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \text{Regression equation} \\
&= -1.6659027 + 0.1104438 = -1.5554589 \\
\frac{p}{1-p} &= \exp(-1.5554589) = 0.21109249 \\
p &= \frac{0.21109249}{1.21109249} = 0.174299231 = 0.174 \text{ (3 d.p.)}
\end{aligned}$$

6.19 Numerical example II: Calculate the probability a Scotland respondent gives a medium score

Set

$$\begin{aligned}
\ln\left(\frac{p}{1-p}\right) &= \text{Regression equation} \\
&= -1.6659027 - 0.0474800 = -1.7133827 \\
\frac{p}{1-p} &= \exp(-1.7133827) = 0.180255011 \\
p &= \frac{0.180255011}{1.180255011} = 0.152725478 = 0.153 \text{ (3 d.p.)}
\end{aligned}$$

• Sanity check. Both answers are similar reflecting no statistically significant evidence for differences between London and Scotland here

7. Reference

Brint, A. and Fry, J. (2021) Regional bias when benchmarking services using customer satisfaction scores. *Total Quality Management and Business Excellence* **32** 344-358