

# Week 7: Mixed Models

## MATH-516 Applied Statistics

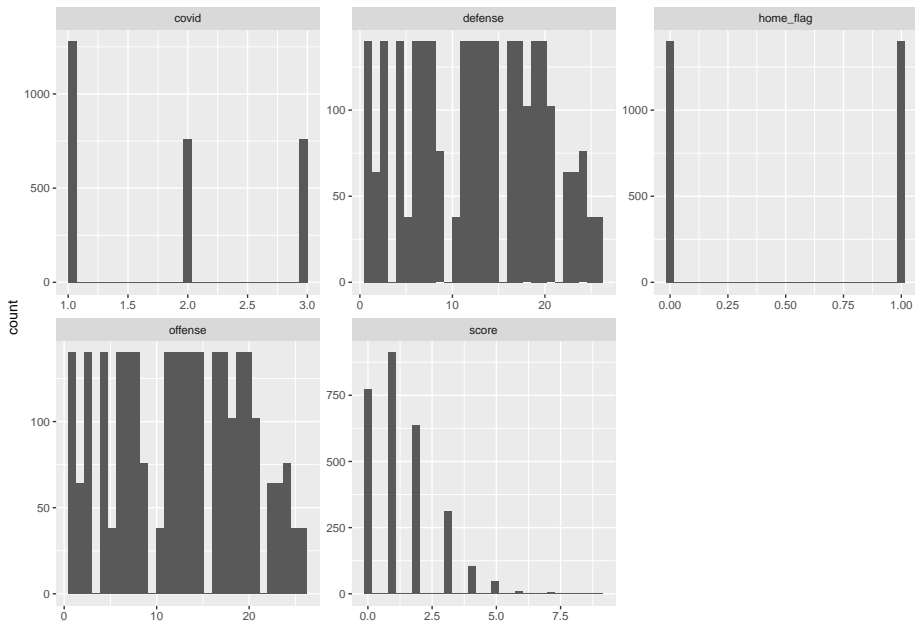
Tomas Masak

Feb 20th 2023

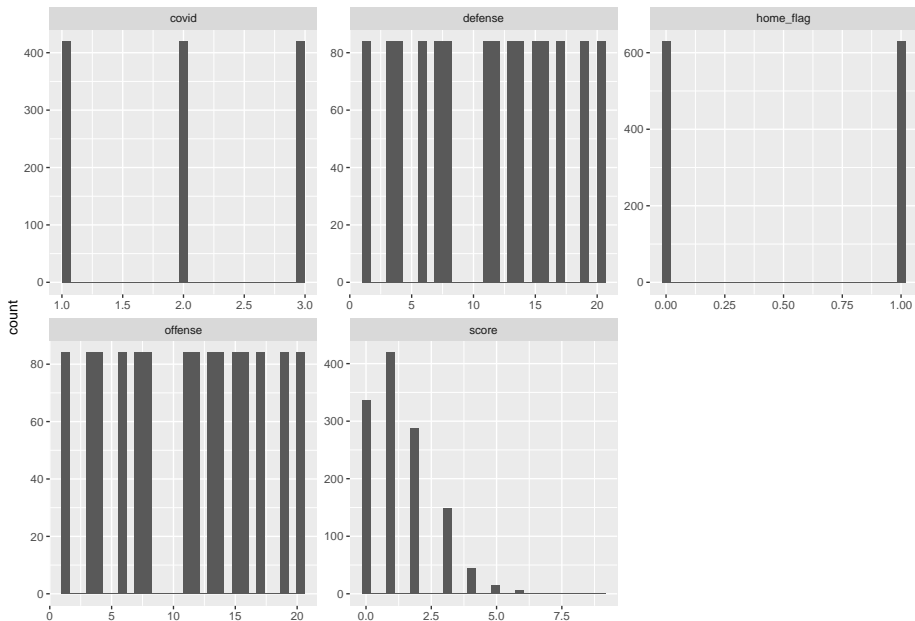
## Section 1

### Project 3: Premier League

# Data Visualized



# Balanced Data



# Models (Balanced Data)

```
m <- glm(score~covid*home_flag+defense+offense, data=subDat, family="poisson")
msub <- glm(score~home_flag+defense+offense, data=subDat, family="poisson")
anova(m,msub,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: score ~ covid * home_flag + defense + offense
## Model 2: score ~ home_flag + defense + offense
##   Resid. Df Resid. Dev Df Deviance Pr(>Chi)
## 1      1226      1352.8
## 2      1230      1357.5 -4   -4.7254  0.3167
```

# Models (Full Data)

```
m <- glm(score~covid*(home_flag+defense+offense), data=Data, family="poisson")
msub <- glm(score~home_flag+defense+offense, data=Data, family="poisson")
anova(m,msub,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: score ~ covid * (home_flag + defense + offense)
## Model 2: score ~ home_flag + defense + offense
##   Resid. Df Resid. Dev   Df Deviance Pr(>Chi)
## 1         2674      2952.6
## 2         2748      3057.7 -74   -105.14   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
library(car)
Anova(m,type=2)
```

```
## Analysis of Deviance Table (Type II tests)
##
## Response: score
##           LR Chisq Df Pr(>Chisq)
## covid           0.91  2  0.633917
## home_flag       16.99  1 3.761e-05 ***
## defense        184.35 25 < 2.2e-16 ***
## offense        345.44 25 < 2.2e-16 ***
## covid:home_flag   5.77  2  0.055895 .
## covid:defense     60.04 35  0.005292 **
## covid:offense     36.35 35  0.405718
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Models (Full Data)

```
minter <- glm(score~covid*(home_flag+defense)+offense, data=Data, family="poisson")
msub <- glm(score~home_flag+covid*defense+offense, data=Data, family="poisson")
anova(m,msub,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: score ~ covid * (home_flag + defense + offense)
## Model 2: score ~ home_flag + covid * defense + offense
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1      2674      2952.6
## 2      2711      2994.7 -37   -42.102   0.2596
anova(minter,msub,test="LRT")
```

```
## Analysis of Deviance Table
##
## Model 1: score ~ covid * (home_flag + defense) + offense
## Model 2: score ~ home_flag + covid * defense + offense
##   Resid. Df Resid. Dev  Df Deviance Pr(>Chi)
## 1      2709      2988.9
## 2      2711      2994.7 -2    -5.7544  0.05629 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
sum(is.na(coefficients(m)))
```

```
## [1] 30
```

- we are on the edge of significance with a model that has too few observations to rely on asymptotics and to estimate all the parameters

# A GLMM

```
library(lme4)
m <- glmer(score ~ covid*home_flag+(1|defense/covid)+(1|offense))
msub <- glmer(score ~ covid*home_flag+(1|defense)+(1|offense))
```

- is `m` a good model?
- can it be simplified to `msub`?
- in both `m` and `msub`, the `covid:home_flag` interaction *looks* significant
  - home advantage reduced during covid and bounced back after covid (not to the original level, but the difference not significant)



# Common Feedback to Reports

- describe data you use, not data you were given
  - (nobody cares whether you got 4 csv files or a single one, and which variables were available but never used because they have absolutely nothing to do with anything, like the betting odds)
- there are some reserved words in statistics such as “significant” or “robust” that are better paraphrased when not used in the reserved meaning (statistical testing, robustness against outliers)
- multiple models vs. a single model
- not taking into account which teams are playing leads to dependence between data
- not including the intercept (i.e. manually discarding the intercept) is problematic since more parameters become inconsistent
  - Poisson vs. multinomial likelihoods
  - it is never a good idea to discard the intercept!
- doesn't vs. does not; let's vs. let us
- description of pre-processing your data (e.g. every match coded twice)

## Section 2

### Common Feedback to Code by Charles

# General comments

Overall good code, just a few remarks:

Major:

- variable names should be descriptive data, data2 (use `raw_data`, `clean_data`)
- Code style (see next slides)
- Colors (see next slides)

Minor:

- commit messages: should describe work done by the commit (ahhhhh might reflect how you feel at the time of commit, but it's not very informative)

# Code style

Code spacing and linting are important<sup>1</sup>

*# Good*

```
do_something_very_complicated(  
  something = "that",  
  requires = many,  
  arguments = "some of which may be long"  
)
```

*# Bad*

```
do_something_very_complicated("that", requires, many, argument  
                               "some of which may be long"  
                               )
```

---

<sup>1</sup>check [<https://style.tidyverse.org/>](<https://style.tidyverse.org/>) for a reference

# Color palette

- understand color scheme selection<sup>2</sup>
- use preset palettes:

```
library(RColorBrewer)
par(mfrow=c(1,2))
display.brewer.pal(name = "Blues", n = 4)
display.brewer.pal(name = "Set1", n = 4)
```



Blues (sequential)



Set1 (qualitative)

---

<sup>2</sup>see [<https://www.gastonsanchez.com/>](<https://www.gastonsanchez.com/>) for more details