Introduction to Heteroscedastic Linear Model and Generalized Linear Model

Application: Testing for Biotech Traits

Jean-Louis Laffont
Pioneer Hi-Bred International, Inc.
A Dupont Company

With: Deanne Wright (Pioneer), Clara Alarcon (Pioneer), David Grothaus (Pioneer), Kirk Remund (Monsanto), Sylvain Grégoire (GEVES)
• When the structure of the data is complex (many factors), the classical Linear Model is a wonderful technique for getting conclusions on the significance of the main effects and their interactions, estimates of the effects, …

• In some situations, there is a departure from the assumptions underlying the use of the classical Linear Model (e.g. experimental errors are either not normally distributed or they have heterogeneous variances).

• In this presentation, some alternative modeling techniques will be introduced.
Dataset 1: estimating source of errors in qualitative ELISA assay for detecting low levels of GMOs
Pioneer Experiment

• Biotech Levels: 0%, 0.11%, 0.51%, 1%, 2%, and 100%.

• 5 separate seed pools of size 3500 spiked with the appropriate number of positive seeds and ground into flour

• 5 flour sub-samples taken

• Qualitative ELISA; response: optical density (OD)

• 25 total observations per biotech level

• Question: quantify flour and sub-sample variation
Dataset 1: estimating source of errors in qualitative ELISA assay for detecting low levels of GMOs

ELISA assay - Side-by-side boxplots - Conditionning on GMO_level x pool #

Strong Gmo_level effect
Dataset 1: estimating source of errors in qualitative ELISA assay for detecting low levels of GMOs

ELISA assay - Normal QQ plots - Conditionning on GMO_level

- Normality of the distributions
- Different variances (heteroscedasticity)
Dataset 1: estimating source of errors in qualitative ELISA assay for detecting low levels of GMOs

Heteroscedasticity confirmed
Dataset 2: proficiency test to assess false negative rates ($Fn$) in qualitative tests for detecting low levels of GMOs
Artificially generated data

- 25 laboratories
- 2 phases
- Around 15 samples that should test positive sent to each laboratory in the 1st phase
- Around 30 samples that should test positive sent to each laboratory in the 2nd phase
- Response: $Fn$ rate
- Questions: Are there lab, phase and lab x phase effects? What are the $Fn$ rate estimates for each Lab?
Dataset 2: proficiency test to assess false negative rates \((Fn)\) in qualitative tests for detecting low levels of GMOs

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>Phase 1</th>
<th>Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab A</td>
<td>0 / 15 (0%)</td>
<td>0 / 31 (0%)</td>
</tr>
<tr>
<td>Lab B</td>
<td>0 / 16 (0%)</td>
<td>1 / 31 (3.2%)</td>
</tr>
<tr>
<td>Lab C</td>
<td>0 / 14 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab D</td>
<td>1 / 15 (6.7%)</td>
<td>1 / 31 (3.2%)</td>
</tr>
<tr>
<td>Lab E</td>
<td>0 / 16 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab F</td>
<td>0 / 15 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab G</td>
<td>0 / 14 (0%)</td>
<td>1 / 31 (3.2%)</td>
</tr>
<tr>
<td>Lab H</td>
<td>0 / 15 (0%)</td>
<td>0 / 31 (0%)</td>
</tr>
<tr>
<td>Lab I</td>
<td>0 / 16 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab J</td>
<td>0 / 15 (0%)</td>
<td>0 / 28 (0%)</td>
</tr>
<tr>
<td>Lab K</td>
<td>1 / 14 (7.1%)</td>
<td>1 / 28 (3.6%)</td>
</tr>
<tr>
<td>Lab L</td>
<td>1 / 14 (7.1%)</td>
<td>2 / 29 (6.9%)</td>
</tr>
<tr>
<td>Lab M</td>
<td>1 / 15 (6.7%)</td>
<td>3 / 29 (10.3%)</td>
</tr>
<tr>
<td>Lab N</td>
<td>0 / 14 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab O</td>
<td>1 / 15 (6.7%)</td>
<td>2 / 31 (6.5%)</td>
</tr>
<tr>
<td>Lab P</td>
<td>2 / 14 (14.3%)</td>
<td>3 / 32 (9.4%)</td>
</tr>
<tr>
<td>Lab Q</td>
<td>1 / 16 (6.3%)</td>
<td>1 / 28 (3.6%)</td>
</tr>
<tr>
<td>Lab R</td>
<td>1 / 16 (6.3%)</td>
<td>3 / 30 (10%)</td>
</tr>
<tr>
<td>Lab S</td>
<td>1 / 15 (6.7%)</td>
<td>5 / 29 (17.2%)</td>
</tr>
<tr>
<td>Lab T</td>
<td>0 / 14 (0%)</td>
<td>1 / 29 (3.4%)</td>
</tr>
<tr>
<td>Lab U</td>
<td>6 / 16 (37.5%)</td>
<td>4 / 30 (13.3%)</td>
</tr>
<tr>
<td>Lab V</td>
<td>0 / 15 (0%)</td>
<td>6 / 31 (19.4%)</td>
</tr>
<tr>
<td>Lab W</td>
<td>4 / 15 (26.7%)</td>
<td>4 / 29 (13.8%)</td>
</tr>
<tr>
<td>Lab X</td>
<td>1 / 15 (6.7%)</td>
<td>4 / 28 (14.3%)</td>
</tr>
<tr>
<td>Lab Y</td>
<td>3 / 15 (20%)</td>
<td>5 / 28 (17.9%)</td>
</tr>
</tbody>
</table>

1 negative result out of 15

Sample from \(B(n_i,0.01)\)

Sample from \(B(n_i,0.05)\)

Sample from \(B(n_i,0.10)\)

Sample from \(B(n_i,0.15)\)
Dataset 2: proficiency test to assess false negative rates ($Fn$) in qualitative tests for detecting low levels of GMOs.
Dataset 1: classical linear model

Model formulation:

7 GMO levels \((i = 1, \ldots, 7)\)
5 pools of seeds from each GMO level \((j = 1, \ldots, 5)\)
5 samples from each pool \((k = 1, \ldots, 5)\)

\(y_{ijk}\): Optical Density

Mixed effects model:

\[ y_{ijk} = \alpha_i + b_{j(i)} + e_{ijk} \]

where \(b_{j(i)} \sim N(0, \sigma_b^2)\)
\(e_{ijk} \sim N(0, \sigma^2)\)

and where

\(\alpha_i\) is the fixed effect part of the model
\(b_{j(i)}\) is the effect of the \(j^{th}\) randomly selected pool from GMO level \(i\)
\(e_{ijk}\) is the effect of the \(k^{th}\) randomly selected sample from the \(j^{th}\) pool of GMO level \(i\)
Dataset 1: classical linear model

Diagnostic plot: normal plot of the residuals by GMO_level:

Normality assumption verified
Dataset 1: classical linear model

Diagnostic plot: scatter plots of residuals vs fitted values, by GMO_level:

Residuals are centered at zero but variability changes with GMO_level.
How to account for heteroscedasticity?

- Variance stabilizing transformation: log, square root, Box-Cox, …

- Weighted Least Squares

  → Linear Heteroscedastic models

  - R.C Littell, G.A. Milliken, W.WW. Stroup, R.D. Wolfinger
    *SAS System for Mixed Models* – Chapter 8

  → Use of variance functions to model the variance structure:
    Stat software offer many capabilities for this.

  Examples:
  - The variance increases linearly with the fitted values
  - The variance is an exponential function of the variance of a covariate
  - Different variances for each level of a stratification variable
    → this later case will be considered in the following
Linear Heteroscedastic model: formulation

7 GMO levels \((i = 1, \ldots, 7)\)
5 pools of seeds from each GMO level \((j = 1, \ldots, 5)\)
5 samples from each pool \((k = 1, \ldots, 5)\)

\(y_{ijk}\): Optical Density

\[ y_{ijk} = \alpha_i + b_{j(i)} + e_{ijk} \quad \text{(mixed effects model)} \quad (1) \]

where \(\alpha_i\) is the fixed effect of GMO level \(i\)

\(b_{j(i)}\) is the random effect of the \(j^{th}\) randomly selected pool from GMO level \(i\)

\(- b_{j(i)} \sim N(0, \sigma^2_b)\)

\(e_{1jk} \sim N(0, \gamma^2_1)\)
\(e_{2jk} \sim N(0, \gamma^2_2)\)
\(\ldots\)
\(e_{ijk} \sim N(0, \gamma^2_i)\)
\(\ldots\)
\(e_{7jk} \sim N(0, \gamma^2_7)\)

Different error variances across GMO levels
Linear Heteroscedastic model: results

The pool variation was not significant after fitting model (1) (Likelihood Ratio test: p-value = 0.9986), thus the following fixed effects model has been fitted:

\[ y_{ijk} = \alpha_i + e_{ijk} \quad (2) \]

with

\[ e_{1jk} \sim N (0, \gamma^2_1) \]
\[ e_{2jk} \sim N (0, \gamma^2_2) \]
\[ \ldots \]
\[ e_{ijk} \sim N (0, \gamma^2_i) \]
\[ \ldots \]
\[ e_{7jk} \sim N (0, \gamma^2_7) \]
Linear Heteroscedastic model: results

The GMO level main effect is highly significant (p-value <.0001)

<table>
<thead>
<tr>
<th>General mean</th>
<th>1.765</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means</td>
<td></td>
</tr>
<tr>
<td>0.00% GMO</td>
<td>0.114 (0.112 ; 0.117)</td>
</tr>
<tr>
<td>0.11% GMO</td>
<td>0.327 (0.315 ; 0.339)</td>
</tr>
<tr>
<td>0.51% GMO</td>
<td>0.976 (0.925 ; 1.026)</td>
</tr>
<tr>
<td>1.00% GMO</td>
<td>1.553 (1.506 ; 1.6)</td>
</tr>
<tr>
<td>2.00% GMO</td>
<td>2.289 (2.22 ; 2.359)</td>
</tr>
<tr>
<td>5.00% GMO</td>
<td>3.316 (3.267 ; 3.365)</td>
</tr>
<tr>
<td>100% GMO</td>
<td>3.783 (3.747 ; 3.819)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Variance Components (x10000)</th>
<th></th>
<th>CV(GMO levels Std Devn. in% of GMO level mean)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Means</td>
<td></td>
<td></td>
</tr>
<tr>
<td>0.00% GMO</td>
<td>0.4 (0.262 ; 0.831)</td>
<td>5.74%</td>
</tr>
<tr>
<td>0.11% GMO</td>
<td>8.3 (5.086 ; 16.144)</td>
<td>8.83%</td>
</tr>
<tr>
<td>0.51% GMO</td>
<td>160.2 (97.697 ; 310.11)</td>
<td>12.97%</td>
</tr>
<tr>
<td>1.00% GMO</td>
<td>138.8 (84.606 ; 268.56)</td>
<td>7.58%</td>
</tr>
<tr>
<td>2.00% GMO</td>
<td>301.5 (183.79 ; 583.39)</td>
<td>7.58%</td>
</tr>
<tr>
<td>5.00% GMO</td>
<td>147.6 (90.013 ; 285.72)</td>
<td>3.66%</td>
</tr>
<tr>
<td>100% GMO</td>
<td>80.7 (49.231 ; 156.27)</td>
<td>2.38%</td>
</tr>
</tbody>
</table>

* figures between parentheses provide an approximate 95% confidence interval of the true values
Linear Heteroscedastic model: benefits for analysis of dataset 1

→ Size of 95% CIs of the true values of the OD means for each GMO_level takes into account variability changes with GMO_level

→ 95% CIs of the true values of the OD variances for each GMO_level are obtained. These variance estimates are very close to those obtained from simple variance calculations.
Linear Heteroscedastic model: fitting with S-PLUS®

. **Model (1):**

```r
> options(contrasts=c("contr.treatment","contr.poly"))
> model1.fit<-lme( od ~ gmo.lev, data = dataset1, random = ~ 1|pool.in,
>                  weights = varIdent(form = ~ 1 | gmo.lev))
```

S-PLUS does not deliver directly the $\gamma_i^2$ s. Instead, the ratio between the standard deviation of the $i^{th}$ level ($i = 2, \ldots, 7$) and the 1st level is delivered.

To get them, we use:

```r
> model1.fit$sigma^2 * c(1, coef(model1.fit$modelStruct$varStruct, unc = F))^2
```

. **Model (2):**

```r
> options(contrasts=c("contr.treatment","contr.poly"))
> model2.fit <- gls(od ~ gmo.lev, data = dataset1,
>                  weights = varIdent(form = ~ 1 | gmo.lev))
```
Linear Heteroscedastic model: fitting with SAS®

. Model (1):

```
proc mixed data = dataset1 MAXITER = 200 CL=WALD COVTEST;
   class gmo_lev pool;
   model od= gmo_lev;
   random pool(gmo_lev);
   repeated /type = un group = gmo_lev;
   lsmeans gmo_lev;
run;
```

. Model (2):

```
proc mixed data = dataset1 MAXITER = 200 CL=WALD COVTEST;
   class gmo_lev;
   model od= gmo_lev;
   repeated /type = un group = gmo_lev;
   lsmeans gmo_lev;
run;
```
Generalized Linear Models (GLM): Introduction

• Generalization of the classical linear model: can also accommodate non normal response distributions

• Focus of this presentation will be on GLMs for binomial data

• Proportions: \( y_i \) (\( i = 1, 2, \ldots, I \))
  → appropriate distribution for the \( i^{th} \) observation:
    binomial distribution \( B(n_i, p_i) \).

• Problems in using a classical linear model for proportions:
  . Constant variance of the response → not verified here
    \( \text{var}(y/n_i) = p_i(1-p_i)/n_i \)
  . Response variable not normally distributed
  . Fitted values that could lie anywhere in the interval \([−∞; +∞]\)
    → for proportions, should lie in the interval \([0; 1]\)
Generalized Linear Models (GLM): Introduction

• Solution:

  - Transformation of the $p_i$ from the range $]0,1[$ to $]-\infty, +\infty[$

Common transformation for proportions: **logistic transformation**:

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right)$$ (the quantity $p_i/(1-p_i)$ is called **odds**)

![Graph showing the logistic transformation of $p_i$](image-url)
Generalized Linear Models (GLM): Introduction

• Solution (cont.):
  
  - Adoption of a linear model for the transformed proportions

  - Non use of Least-squares method to fit the linear model on the transformed proportions

    instead

    use of a maximum-likelihood approach

    → process of model fitting: **iterative weighted least squares**

      (weights of the deviations of the data points from those predicted by the model depend on the $n_i$ and the values fitted by the model)

  - Measure of goodness of fit with the **deviance**

    (in a classical linear model, deviance = residual sum of squares)
Dataset 2: Binomial GLM

Dataset 2: 25 labs ($i = 1, \ldots, 25$)
2 phases ($j = 1, 2$)

$$\text{logit}(p_{ij}) = \log \left( \frac{p_{ij}}{1-p_{ij}} \right) = \beta_0 + \text{lab}_i + \text{phase}_j$$

where:
- $p_{ij}$ is the proportion of $Fn$ in laboratory $i$ and for phase $j$
- $\beta_0$ is a constant
- $\text{lab}_i$ is the effect of the $i$th laboratory ($i = 1, 2, \ldots, 25$)
- $\text{phase}_j$ is the effect of the $j$th phase ($j = 1, 2$)

Vocabulary:
- The function that relates $p$ to the linear component of the model is called the link function (here we have a logistic link function)
- The model for dataset 2 is also referred to as a Binomial GLM
As there are problems in fitting a binomial GLM when all the data for some levels of a factor = 0, we will limit the analysis to a subset of labs:

<table>
<thead>
<tr>
<th>Laboratory</th>
<th>Phase 1</th>
<th>Phase 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab A</td>
<td>0 / 15 (0%)</td>
<td>0 / 31 (0%)</td>
</tr>
<tr>
<td>Lab B</td>
<td>0 / 16 (0%)</td>
<td>1 / 31 (3.2%)</td>
</tr>
<tr>
<td>Lab C</td>
<td>0 / 14 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab D</td>
<td>1 / 15 (6.7%)</td>
<td>1 / 31 (3.2%)</td>
</tr>
<tr>
<td>Lab E</td>
<td>0 / 16 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab F</td>
<td>0 / 15 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab G</td>
<td>0 / 14 (0%)</td>
<td>1 / 31 (3.2%)</td>
</tr>
<tr>
<td>Lab H</td>
<td>0 / 15 (0%)</td>
<td>0 / 31 (0%)</td>
</tr>
<tr>
<td>Lab I</td>
<td>0 / 16 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab J</td>
<td>0 / 15 (0%)</td>
<td>0 / 28 (0%)</td>
</tr>
<tr>
<td>Lab K</td>
<td>1 / 14 (7.1%)</td>
<td>1 / 28 (3.6%)</td>
</tr>
<tr>
<td>Lab L</td>
<td>1 / 14 (7.1%)</td>
<td>2 / 29 (6.9%)</td>
</tr>
<tr>
<td>Lab M</td>
<td>1 / 15 (6.7%)</td>
<td>3 / 29 (10.3%)</td>
</tr>
<tr>
<td>Lab N</td>
<td>0 / 14 (0%)</td>
<td>0 / 29 (0%)</td>
</tr>
<tr>
<td>Lab O</td>
<td>1 / 15 (6.7%)</td>
<td>2 / 31 (6.5%)</td>
</tr>
<tr>
<td>Lab P</td>
<td>2 / 14 (14.3%)</td>
<td>3 / 32 (9.4%)</td>
</tr>
<tr>
<td>Lab Q</td>
<td>1 / 16 (6.3%)</td>
<td>1 / 28 (3.6%)</td>
</tr>
<tr>
<td>Lab R</td>
<td>1 / 16 (6.3%)</td>
<td>3 / 30 (10%)</td>
</tr>
<tr>
<td>Lab S</td>
<td>1 / 15 (6.7%)</td>
<td>5 / 29 (17.2%)</td>
</tr>
<tr>
<td>Lab T</td>
<td>0 / 14 (0%)</td>
<td>1 / 29 (3.4%)</td>
</tr>
<tr>
<td>Lab U</td>
<td>6 / 16 (37.5%)</td>
<td>4 / 30 (13.3%)</td>
</tr>
<tr>
<td>Lab V</td>
<td>0 / 15 (0%)</td>
<td>6 / 31 (19.4%)</td>
</tr>
<tr>
<td>Lab W</td>
<td>4 / 15 (26.7%)</td>
<td>4 / 29 (13.8%)</td>
</tr>
<tr>
<td>Lab X</td>
<td>1 / 15 (6.7%)</td>
<td>4 / 28 (14.3%)</td>
</tr>
<tr>
<td>Lab Y</td>
<td>3 / 15 (20%)</td>
<td>5 / 28 (17.9%)</td>
</tr>
</tbody>
</table>
Dataset 2 – 17 labs: Binomial GLM

- Deviance = 15 on 16 df
  - The deviance is asymptotically distributed as $\chi^2$
  - Lack of fit: p.value = 0.524392 → not significant
    This also indicates that the interaction lab x phase is not significant

- Diagnostic plots:
Dataset 2 – 17 labs: Binomial GLM

- Analysis of deviance → which terms are significant
  → Lab: $p$-value = 0.011760
  → Phase: $p$-value = 0.956128

- Our final model will be simpler:
  $$\text{logit}(p_{ij}) = \log \left( \frac{p_{ij}}{1-p_{ij}} \right) = \beta_0 + \text{lab}_i$$

- The model will provide the predicted values of false negative proportions and their standard errors on the logistic scale. We can get model-based probability estimates on the original scale with approximate 95% confidence for the true probabilities:
  $$\hat{p}_i = \frac{\exp(\hat{\eta}_i)}{1 + \exp(\hat{\eta}_i)}$$
Dataset 2 – 17 labs: Binomial GLM

Model-based probability estimates on the original scale with approximate 95% confidence for the true probabilities:
Dataset 2 – 17 Labs: classical linear model

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum of Sq</th>
<th>Mean Sq</th>
<th>F Value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab</td>
<td>16</td>
<td>1461.844</td>
<td>91.365</td>
<td>2.066</td>
<td>0.078706</td>
</tr>
<tr>
<td>Phase</td>
<td>1</td>
<td>0.012</td>
<td>0.012</td>
<td>0.000</td>
<td>0.987225</td>
</tr>
<tr>
<td>Residuals</td>
<td>16</td>
<td>707.547</td>
<td>44.222</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Departure from Normality**

- Fitted values vs Residuals
- Quantiles of Standard Normal vs Residuals
Dataset 2: classical methods used to analyze proportions

95% confidence intervals for the true \( Lab \ i \ Fn \) rate:

\[
\left[ \frac{y_i}{y_i + (n_i - y_i + 1)F_{0.025,\nu_1,\nu_2}}, \frac{(y_i + 1)F_{0.025,\nu'_1,\nu'_2}}{n_i - y_i + (y_i + 1)F_{0.025,\nu_1,\nu_2}} \right]
\]

where: \( p_i = \frac{y_i}{n_i} \) : observed proportion of \( Fn \) in \( lab \ i \)

\[\nu_1 = 2(n_i - y_i + 1) \quad \nu_2 = 2y_i\]

\[\nu'_1 = 2(y_i + 1) \quad \nu'_2 = 2(n_j - y_j)\]

\( F_{0.025,\nu_1,\nu_2} \) is the upper 0.025 point of the \( F \)-distribution with \( \nu_1 \) and \( \nu_2 \) df

Dataset 2: classical methods vs Binomial GLM vs LM

Probability estimates

Lab B
Lab D
Lab G
Lab K
Lab L
Lab M
Lab O
Lab P
Lab Q
Lab R
Lab S
Lab T
Lab U
Lab V
Lab W
Lab X
Lab Y

Binomial GLM
Simple Linear Model
Classical Method

Probability estimates
GLM : fitting with S-PLUS®

> glm(FN ~ lab + phase, data = dataset2, family = binomial(link=logit))

GLM : fitting with SAS®

```
proc genmod data=dataset2;
    class LAB PHASE;
    model FN/N = LAB PHASE /p ;
run;
```
Binomial GLM: some conclusions

• Classical methods to analyze proportions and binomial GLM provided very similar estimates

• Problem with GLM are zeros; other potential difficulties when interactions are significant…

• Fitting data from dataset2 with a classical LM is not adapted

• Binomial GLM provided answers to significance of main effects and interaction
References


