Chapter 8 & 9: GLM Requirements

A look at model residuals will be paramount in determining whether or not our GLM requirements are met. Our F-statistics are based on reliable estimates of the error variance (MSE). To have a reliable estimate we need to make sure that our errors (residuals) are independent; that it is sensible to consider a single value to summarize the error variance (constant variance across the model), and that there are no unusual patterns in the residuals suggesting that our model is not representing our data well. Furthermore, since we’ll be computing a p-value to make our decisions, we must also have an appropriate reference distribution for our test statistic (the F-distribution). To ensure this, our errors must be normally distributed with common variance (chapter 9).

Independence

Caterpillar growth (from text)

The text was trying to have a nice example of a lack of independence. They describe a dataset that indeed has this problem, but unfortunately the problem is not transparent from the plots. In fact, the evidence manifests more as a problem of a lack of orthogonality (lack of independence of the predictor variables) rather than a lack of independence between the individual units (caterpillars) in the analysis.

The researcher sampled 10 caterpillars from each of three habitats. 5-day weight gain and population density were measured. The gray line is the regression line for the word equation Weight gain = denisty.

So there are two problems here (one problem that manifests in two ways). First, Habitat and PopDen are not orthogonal. Knowing that the density is high, for example, indicates that we’re most likely dealing with observations from habitat 1; similarly, low density implies habitat 3. As a result a second problem manifests with this inappropriate model that does not account for habitat differences: within
each habitat group, the residuals are no longer independent. Note that *within each habitat*, when density is high, the gain tends to be above the modeled average and when density is low we see below average observations. This feature can be exaggerated using a plot of the residuals versus fitted values and then using different plotting symbols for the habitat groups.

Residual: \( e = y - \hat{y} \) the difference between the data value \( y \) and the model predicted value \( \hat{y} \). By looking closely at the residuals, we can see their lack of independence. Pick a residual. Is it above or below the regression line? What about its neighbors? If you can guess what’s happing for its neighbors, the residuals are not independent.

\[
\begin{align*}
> \text{plot(m1, col = mycol[bug$Habitat], pch = 16, which = 1)} \\
> \text{m1 = lm(WeightGain ~ Popden, data = bug)} \\
> \text{plot(resid(m1) ~ fitted(m1), pch = 16, col = mycol[bug$Habitat])} \\
> \text{abline(h=0, lty = 2)} \\
> \text{plot(m1, col = mycol[bug$Habitat], pch = 16, which = 1)}
\end{align*}
\]
Two analyses...which is appropriate?

```r
> m1 = lm(WeightGain ~ Popden, data = bug)
> drop1(m1, test= 'F')

Single term deletions

```

Df | Sum of Sq | RSS | AIC | F value | Pr( F )
---|-----------|-----|-----|---------|--------
<none> | 269.32 | 69.841
Popden | 1 | 354.12 | 623.44 | 93.022 | 36.816 | 1.526e-06 | ***

```r
---
> m1 = lm(WeightGain ~ Popden + Habitat, data = bug)
> drop1(m1, test = 'F')

Single term deletions

```

Df | Sum of Sq | RSS | AIC | F value | Pr( F )
---|-----------|-----|-----|---------|--------
<none> | 150.47 | 56.377
Popden | 1 | 14.618 | 165.09 | 57.158 | 2.5259 | 0.124080 | 1
Habitat | 2 | 118.853 | 269.32 | 69.841 | 10.2686 | 0.000516 | ***

**Pig Growth: Repeated measures**

This textbook doesn’t have extra plots for this. Let’s look more at what’s going on and why a special analysis is necessary. Here there are 10 pigs on two diets (5 on each diet). Each pig is weighed on four occasions (3, 8, 20, and 60 weeks).

First we’ll plot the data and jitter weeks to get a better view. There is a curve pattern so maybe we’ll transform Wt using a base 10 log. We still have a curve, but it’s different. Note what happened to the variability within each week.
Next we consider logging week too. The pattern is still nonlinear, but this is what the book goes with for now.

- How many independent pieces of data do we have?
- How many “points” should appear in a “proper” plot of the data.
- How many independent piece of information do we have?
Consider looking at each of the 10 pigs growth trajectory. How can we analyze this data?

A solution, consider analyzing 4 separate data sets (one for each time) or perhaps pairs (changes over time). Another solution is Multivariate ANOVA (MANOVA) or repeated measures ANOVA (advanced topic).
Tomato Germination: Non-independence

In this experiment there were 5 different levels of watering. An experimental goal was to compare germination rates for the five different watering regimes. There were 50 seed trays and the percent germination was measured for each tray on three occasions.

- How many independent observations are there?
- How many points or lines should be used to represent this data on a plot?
- Does it matter how many seeds were on the trays? (This is an issue not presented at this time, but it turns out that it is very important).

```
> m1 = lm(PercGerm ~ Water, data = tomato)
> anova(m1)

Analysis of Variance Table

Response: PercGerm
            Df  Sum Sq Mean Sq F value Pr(>F)
Water        4 283.23  70.790  1.332 1 0.2608
Residuals 145 7705.6  53.142
```

Why is the above analysis incorrect?
> m1 = lm(PercGerm ~ Water + Date, data = tomato)
> anova(m1)

Analysis of Variance Table

Response: PercGerm

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>4</td>
<td>283.2</td>
<td>70.79</td>
<td>5.0803</td>
</tr>
<tr>
<td>Date</td>
<td>2</td>
<td>5713.0</td>
<td>2856.51</td>
<td>205.0026</td>
</tr>
<tr>
<td>Residuals</td>
<td>143</td>
<td>1992.6</td>
<td>13.93</td>
<td></td>
</tr>
</tbody>
</table>

Why is the above analysis incorrect?

What is a possible way to analyze this data appropriately?

> m1 = lm(PercGerm ~ Water, data = subset(tomato, subset = Date == '3'))
> anova(m1)

Analysis of Variance Table

Response: PercGerm

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Water</td>
<td>4</td>
<td>140.97</td>
<td>35.242</td>
<td>2.2088</td>
</tr>
<tr>
<td>Residuals</td>
<td>45</td>
<td>717.98</td>
<td>15.955</td>
<td></td>
</tr>
</tbody>
</table>