FROM EXPERIMENTAL DESIGNS TO GLMS

WEEK 7: CHOOSING A MODEL AND EXPLAINING ITS RESULTS

Gavin Simpson
Aarhus University

Wednesday, 21 October 2026

WHAT WE WILL DO TODAY

We will use the experiment from Weeks 5–6 to choose a statistical model, fit it in R, and explain what it says about the two diets.

Time	Work
12:30–13:10	Review the experiment; learn how to describe and fit a GLM
13:10–13:40	Choose models for three short studies
13:40–13:55	Break
13:55–14:55	Compare diet effects in two housing conditions; analyse your group's data
14:55–15:10	Break
15:10–16:00	Apply the steps to a calf study; discuss projects; finish with a short question

THE EXPERIMENT WE ARE ANALYSING

There are **64 pens in four rooms**, with six pigs in each pen. Within each room, eight pens receive Control feed and eight receive High fibre feed. Diet is randomly assigned to whole pens.

We follow every pen for **28 days** and record one outcome per pen. The four groups analyse different outcomes from this same study design.

This is a simulated teaching dataset, not evidence about the effects of a real diet.

OPENING RECALL: DISCUSS THE EXPERIMENT

Use the description on the previous slide. Discuss these questions with your group for two minutes:

1. Is diet assigned separately to each pig, or to the whole pen?
2. If the data contain one row per pen, how many rows should there be?
3. Why is it useful to have both diets represented in every room?

Be ready to explain one answer in a sentence.

WHAT DOES EACH GROUP MEASURE?

The **response** is the outcome we want the model to explain.

Group	One measurement for each pen
1	Average daily weight gain, in grams per day
2	Whether any pig needed treatment: 0 means none, 1 means at least one
3	Number of treatment events during the 28 days; a pig may be treated more than once
4	Feed conversion ratio: kg of feed consumed per kg of weight gained

For Group 2, one row describes the whole pen. It does not tell us how many individual pigs were treated.

FOUR THINGS TO STATE WHEN CHOOSING A GLM

A **generalised linear model (GLM)** describes how an expected outcome changes with explanatory variables, such as diet and room.

State:

1. **Response:** what is measured, and what does one row describe?
2. **Formula:** which explanatory variables should be included?
3. **Family:** which probability distribution describes variation in the response?
4. **Link:** how is the expected response related to the sum of the model terms?

The next slides explain these choices using daily weight gain.

READING THE FORMULA

For our study, the formula is `response ~ diet + room`.

- `response` is the outcome column in the data.
- `~` separates the outcome from the explanatory variables.
- `diet` lets the expected outcome differ between Control and High fibre.
- `room` allows for differences between the four rooms.

Including `room` lets us compare diets while accounting for where the pens were kept.

THE LINEAR PREDICTOR AND THE LINK

The **linear predictor** is the intercept plus the contributions of diet and room. An intercept is the starting value for the reference categories: here, Control diet in Room 1.

For daily gain we use an **identity link**. This means the linear predictor is already the expected gain in grams per day; no conversion is needed.

Other links are useful for other outcomes. With a log link, use `exp()` to convert the linear predictor back to a positive expected value. With a logit link, use `plogis()` to convert it back to a probability.

THE FAMILIES AND LINKS SUPPLIED FOR THIS ACTIVITY

Outcome	Family	Link
Average daily gain, a continuous measurement	Gaussian, also called normal	Identity
Whether any pig was treated, recorded as 0 or 1	Binomial	Logit
Number of treatment events	Poisson	Log
Positive feed conversion ratio	Gamma	Log

These are starting assumptions for the teaching examples. A suitable outcome type does not, by itself, establish that a model fits the data well.

OPEN THE DAILY-GAIN EXAMPLE IN POSIT CLOUD

Open `group-1-student.R` in the supplied Week 7 project. Run the first section.

`source()` reads the supplied file of functions. `read_pen_data(1)` reads Group 1's CSV into a data frame called `d` and sets Control and Room 1 as the reference categories.

```
source("glm-helpers.R")
d <- read_pen_data(1)
head(d[c("pen_id", "diet", "room", "response")], 2)
```

```
  pen_id    diet   room response
1   P001 Control Room 1 695.1326
2   P002 Control Room 1 688.3944
```

Each displayed row describes one pen; `response` is that pen's average daily gain.

FIT THE MODEL FOR DAILY GAIN

`glm()` estimates the model from `d`. We show the first two coefficient rows and only their estimates and standard errors.

```
fit <- glm(response ~ diet + room,
           family = gaussian("identity"), data = d)
# Intercept and diet estimate/SE; remaining rows are room terms.
coef(summary(fit))[1:2, c("Estimate", "Std. Error")]
```

```
              Estimate Std. Error
(Intercept)    689.98514   16.46264
dietHigh fibre  57.56809   14.72463
```

`(Intercept)` is the estimated average gain for Control pens in Room 1. `dietHigh fibre` is the estimated High fibre minus Control difference, in grams per day.

HOW TO READ THE COEFFICIENT TABLE

The **Estimate** column contains the fitted coefficient values. A positive diet estimate means the model predicts higher average gain with High fibre than with Control, within the same room.

The **Std. Error** column describes uncertainty in each coefficient estimate. It is not the spread of individual pen measurements.

Look back at the diet row. Does its estimate describe an increase or a decrease in expected gain? What are its units?

CHOOSE THE CONDITIONS FOR OUR PREDICTIONS

We want to predict the expected gain for both diets **in the same room**.

The supplied `prediction_grid(d)` function makes a table containing both diets in each room. It does not contain new observations: it lists the conditions for which we want predictions.

```
grid <- prediction_grid(d)
head(grid, 2)
```

```
        diet   room
1    Control Room 1
2 High fibre Room 1
```

The first two rows ask for Control and High fibre predictions in Room 1.

READ THE PREDICTIONS FOR ROOM 1

`predict_means(fit, grid)` calculates the expected outcomes and approximate 95% confidence intervals.

```
predictions <- predict_means(fit, grid)
subset(predictions, room == "Room 1",
       select = c(diet, room, expected, lower, upper))
```

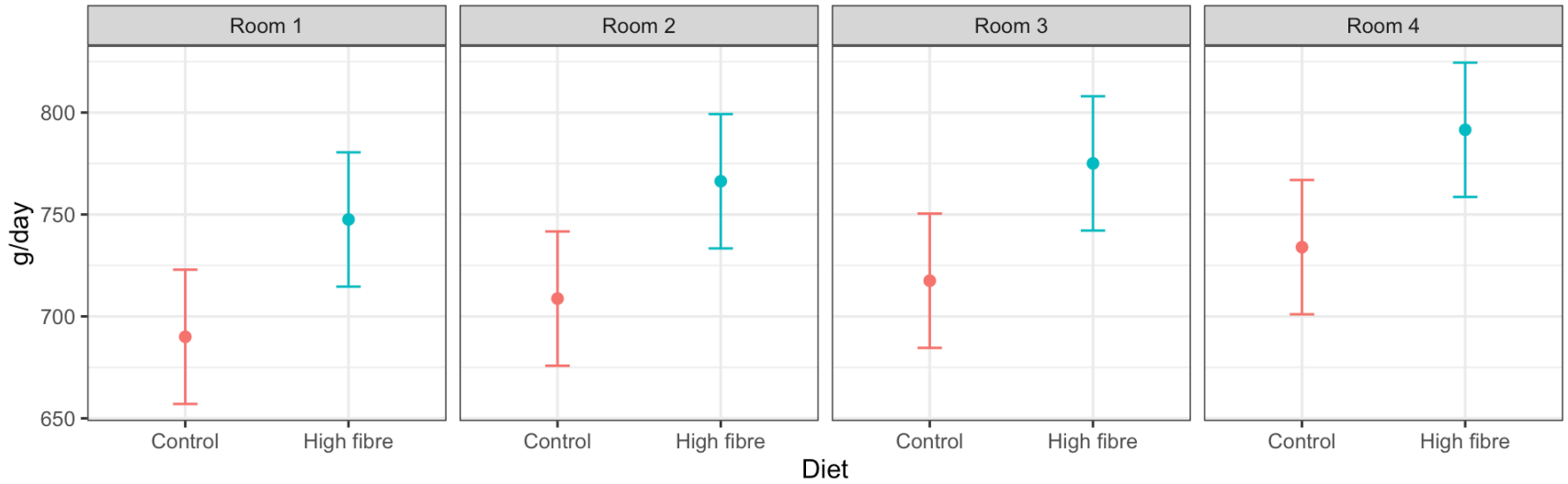
```
        diet   room expected    lower    upper
1    Control Room 1 689.9851 657.0435 722.9268
2 High fibre Room 1 747.5532 714.6116 780.4949
```

`expected` is estimated average gain, in g/day. `lower` and `upper` show uncertainty in that average, assuming the model is appropriate. They do not describe the range containing 95% of individual pens.

COMPARE THE PREDICTIONS IN ALL FOUR ROOMS

`response_spec(1)` supplies Group 1's labels and settings, stored in `spec`. The plotting function uses these to label gain in g/day. Its points show estimated averages; its bars show their confidence intervals.

```
spec <- response_spec(1)
plot_predictions(predictions, spec)
```



Approximate 95% confidence intervals for expected responses in each room.

In Room 1, which diet has the higher estimated average gain? Check that your answer agrees with the diet coefficient.

DESCRIBE THE DIET COMPARISON IN THE CORRECT UNITS

Group	What the model's diet coefficient tells us
1	Difference in expected gain, in g/day
3 and 4	Exponentiating the coefficient gives the High fibre / Control ratio of expected outcomes
2	Exponentiating the coefficient gives a ratio of odds; this needs a separate explanation

For example, an expected-count ratio of 0.60 means the expected count under High fibre is 60% of the Control count in the same room.

FOR THE BINARY OUTCOME: PROBABILITY AND ODDS

A probability of 0.30 means that about 30 out of 100 comparable pens would have at least one pig treated.

Odds are probability divided by one minus probability: here, `0.30 / 0.70`. An **odds ratio** compares those odds between the two diets.

An odds ratio of 0.60 does not mean the probability is multiplied by 0.60. Group 2 should report the model's predicted probability for each diet; the after-class example shows the odds calculation step by step.

CHOOSE MODELS FOR THE SHORT STUDIES

Open the class worksheet. For Studies A and B, write down what is measured, which unit receives diet, the formula, and the supplied family and link.

For Study C, identify the four diet-and-housing combinations and the averages we need to compare. We will explain the interaction formula after the break.

Use the response table and the worked daily-gain example as references while answering.

BEFORE THE BREAK

Keep your written answers. If your group is unsure about a choice, write the question beside it so we can discuss it.

After the break we will look at a study with both diet and housing treatments, then analyse each group's pen data.