# FROM COEFFICIENTS TO EXPECTED OUTCOMES

## WEEK 7: UNDERSTANDING COEFFICIENTS, PREDICTIONS AND INTERACTIONS

Gavin Simpson
Aarhus University

Wednesday, 21 October 2026

# AFTER-CLASS WORK

Allow 60–75 minutes: about 25 minutes using these slides, 25–35 minutes reading, and 10–15 minutes answering the follow-up questions.

These slides explain the examples while you work through the questions. You can use the recap and class results if you do not have the textbook.

Keep your script and answers for 28 October.

# START WITH THE SAME DAILY-GAIN DATA

Open the supplied Week 7 Posit Cloud project. The code below reads Group 1's data and fits the model from class. The outcome is average gain per pen, in g/day.

```
source("glm-helpers.R")
d <- read_pen_data(1)
fit <- glm(response ~ diet + room,
           family = gaussian("identity"), data = d)
coef(summary(fit))[1:2, c("Estimate", "Std. Error")]
```

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

The intercept is expected gain for Control pens in Room 1. The diet row is the estimated High fibre minus Control difference.

# WHAT IS A REFERENCE CATEGORY?

For a categorical variable, R chooses one category as the starting point for comparisons. This is the **reference category**.

Here, Control is the reference diet and Room 1 is the reference room. The intercept describes their combination.

Changing the reference diet changes what the intercept and diet coefficient describe. It should not change the predicted gain for the same diet and room.

# CHANGE THE STARTING POINT FOR THE DIET COMPARISON

We copy `d` and use `relevel()` to make High fibre the reference in that copy. We then refit the same model.

```
d_relevelled <- d
d_relevelled$diet <- relevel(d$diet, ref = "High fibre")
fit_relevelled <- glm(response ~ diet + room,
  family = gaussian("identity"), data = d_relevelled)
c(Control_reference = unname(coef(fit)["(Intercept)"]),
  High_fibre_reference = unname(coef(fit_relevelled)["(Intercept)"]))
```

```
   Control_reference High_fibre_reference
            689.9851             747.5532
```

The two intercepts refer to different diets in Room 1, so their values differ.

# CHECK THE PREDICTIONS AFTER CHANGING THE REFERENCE

`grid` lists both diets in Room 1. We predict gain using each fitted model and put the answers beside one another.

```
grid <- subset(prediction_grid(d), room == "Room 1")
data.frame(grid,
  Control_reference = predict(fit, grid, type = "response"),
  High_fibre_reference = predict(fit_relevelled, grid,
                                 type = "response"))
```

```
        diet   room Control_reference High_fibre_reference
1    Control Room 1          689.9851             689.9851
2 High fibre Room 1          747.5532             747.5532
```

For a given diet, do the two columns give the same predicted gain? Explain why that can happen even though the intercepts differ.

# CONVERTING THE LINEAR PREDICTOR TO AN EXPECTED OUTCOME

The linear predictor adds the intercept and the relevant diet and room contributions. The link tells us how to convert that sum into an expected outcome.

| Link used here | Conversion to the outcome we want to explain |
|---|---|
| Identity | No conversion: already expected gain in g/day |
| Log | Use `exp()` to obtain a positive expected count or feed conversion ratio |
| Logit | Use `plogis()` to obtain a probability between 0 and 1 |

In `predict()`, `type = "response"` asks R to make the conversion for you.

# PROBABILITY AND ODDS DESCRIBE THE SAME CHANCE DIFFERENTLY

For Group 2, the event is **at least one pig in a pen needing treatment**.

If its probability is 0.30, about 30 out of 100 comparable pens would have the event. The odds are `0.30 / 0.70`: three pens with the event for every seven without it.

To turn odds back into probability, divide the odds by one plus the odds. We will use these two conversions in the next example.

# COMPARE TWO DIETS USING AN ODDS RATIO

Use this illustrative example, rather than the fitted classroom numbers: the Control probability is 0.30 and the diet coefficient is −0.50 in a logit model.

Exponentiating the diet coefficient gives the High fibre / Control **odds ratio**. Multiply the Control odds by that ratio, then convert the new odds back to a probability.

Before running the calculation, do you expect the High fibre probability to be above or below 0.30? Explain using the negative coefficient.

# WORK THROUGH THE PROBABILITY CALCULATION

```
control_probability <- 0.30
control_odds <- control_probability / (1 - control_probability)
odds_ratio <- exp(-0.50)
high_odds <- control_odds * odds_ratio
high_probability <- high_odds / (1 + high_odds)
c(probability = high_probability, odds_ratio = odds_ratio,
  probability_change = high_probability - control_probability)
```

```
      probability        odds_ratio probability_change
       0.20631249        0.60653066       -0.09368751
```

Read `probability` as the chance of the pen-level event under High fibre. `probability_change` subtracts the Control probability. The odds ratio and probability ratio are different comparisons.

# RETURN TO THE STUDY WITH DIET AND HOUSING

We now use `data/factorial-pens.csv`, the separate study with 48 pens and four diet-and-housing combinations in every room.

We fit the interaction model because the question is *whether the diet difference changes with housing*. The reference categories are `Control`, `Standard housing` and `Room 1`.

```
f <- read_factorial_data()
interaction <- glm(gain ~ diet * housing + room,
                   family = gaussian("identity"), data = f)
coef(interaction)[
  c("dietHigh fibre", "dietHigh fibre:housingEnriched")]
```

```
             dietHigh fibre dietHigh fibre:housingEnriched
                   43.03714                     35.67392
```

# PREDICT THE FOUR AVERAGE GAINS IN ROOM 1

We list both diets under both housing conditions, keeping room unchanged. The table shows predicted average gains in g/day.

```
grid <- expand.grid(diet = levels(f$diet),
  housing = levels(f$housing), room = "Room 1")
predictions <- predict_means(interaction, grid)
predictions[c("diet", "housing", "expected")]
```

```
        diet  housing expected
1    Control Standard 693.7933
2 High fibre Standard 736.8305
3    Control Enriched 722.7639
4 High fibre Enriched 801.4749
```

Find the Control and High fibre rows for Standard housing. Then find the corresponding two rows for Enriched housing.

# COMPARE THE DIET DIFFERENCES BETWEEN HOUSING CONDITIONS

For each housing condition, subtract the Control prediction from the High fibre prediction. Then compare those two differences.

```
diet_difference <- with(predictions,
  expected[diet == "High fibre"] - expected[diet == "Control"]
)
names(diet_difference) <- levels(f$housing)
diet_difference
```

```
Standard Enriched
43.03714 78.71106
```

```
#c(difference_of_differences = unname(diff(diet_difference)))
diff(diet_difference)
```

```
Enriched
35.67392
```

Complete this sentence: “The estimated High fibre minus Control gain difference is ___ g/day larger in Enriched housing than in Standard housing.”

# BE CAREFUL WHEN COMPARING OTHER KINDS OF OUTCOMES

For gain with an identity link, an interaction asks whether a difference in expected gain changes with housing.

With a log link, it asks whether an expected-outcome ratio changes.

With a logit link, it asks whether an odds ratio changes.

For any model, first say what is measured and which comparison you want to explain.

# PREPARE FOR OBSERVATIONS FROM INDIVIDUAL ANIMALS

Today, each row describes a pen.

Next week, imagine recording a separate outcome for each of its six pigs. The pigs still share their pen's assigned diet, feed and environment.

Why might two pigs sharing a pen have more similar outcomes than pigs in different pens? Does recording more pigs change the number of pens independently assigned a diet?

# READ WITH YOUR CLASS EXAMPLE BESIDE YOU

In Touchon's second edition, use Chapter 7, *Generalized Linear Models*, to revisit a worked example similar to your group's outcome. Look for what is measured, the family and link, and how its fitted coefficients are explained.

Then use Chapter 6, *More Linear Models!*, to revisit interactions and predictions. Connect it to the diet-and-housing comparison we made in class.

If you cannot access the book, use the recap and your group's output for the same practice.