simon-5502-06-slides

Topics to be covered

- What you will learn
  - Review of simple tests of two proportions
  - Concepts behind the logistic regression model
  - Logistic regression with categorical predictors
  - Logistic regression with continuous predictors
  - Logistic regression with interactions
  - Diagnostics

Comparing two binary outcomes

- Is there a difference in the proportion of deaths between male passengers and female passengers on the Titanic?
- Is there difference in the proportion of patients finishing the full three doses of HPV vaccine between Black women and White women?
- Does using a ng tube for feeding in pre-term infants increase the probability of successful breast feeding at six months?

Speaker notes

Most of the statistics, you have seen so far involve a continuous outcome. You can, however, use a binary outcome. Here are three examples comparing a binary outcome between two groups.

Other comparisons involving a binary outcome

- Is there are difference in the proportion of deaths between first class, second class, and third class passengers?
- Does age influence the proportion of women finishing the full three doses of HPV vaccine?
- Controlling for the mother's age, does using a ng tube for feeding in pre-term infants increase the probability of successful breast feeding at six months?

Speaker notes

Here are some more complex comparisons involving a binary outcome. The first example involves a comparison of three proportions, not two. The next example involves a continuous predictor of a binary outcome. The final example involves a comparison of binary outcomes in two groups, but controlling for a third variable.

Hypothesis framework

- $H_0 : \ \pi_1 = \pi_2$
- $H_1 : \ \pi_1 = \pi_2$
- Compute $\hat{p}_1$ and $\hat{p}_2$ from samples
- Accept H_0 if $\hat{p}_1 - \hat{p}_2$ is close to zero.
  - $T = (\hat{p}_1 - \hat{p}_2)/s.e.$
  - 95% CI: $(\hat{p}_1 - \hat{p}_2) \pm Z_{\alpha/2} s.e.$

Speaker notes

The hypothesis to test two proportions uses the symbols π_1 and π_2 to represent the proportions in a population.

The Titanic dataset

```
Rows: 1,313
Columns: 5
$ Name     <chr> "Allen, Miss Elisabeth Walton", "Allison, Miss Helen
Loraine"…
$ PClass   <chr> "1st", "1st", "1st", "1st", "1st", "1st", "1st", "1st",
"1st"…
$ Age      <dbl> 29.00, 2.00, 30.00, 25.00, 0.92, 47.00, 63.00, 39.00, 58.00,
…
$ Sex      <chr> "female", "female", "male", "female", "male", "male",
"female…
$ Survived <dbl> 1, 0, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1, 1, 0, 1, 0, 0, 1, 1,
1…
```

Counts and percentages

```
         Survived
Sex        Yes  No
  female 308 154
  male   142 709
```

```
         Survived
Sex             Yes        No
  female 0.6666667 0.3333333
  male   0.1668625 0.8331375
```

Test for difference in proportions

```
# A tibble: 1 × 9
  estimate1 estimate2 statistic  p.value parameter conf.low conf.high method
      <dbl>     <dbl>     <dbl>    <dbl>     <dbl>    <dbl>     <dbl> <chr>
1     0.667     0.167      332. 3.43e-74         1    0.450     0.550 2-sample
…
# ℹ 1 more variable: alternative <chr>
```

Chi-square test of independence, 1 of 2

- Equivalent to test of two proportions
- Lay out data in two by two table

$$\begin{array}{ccc} & No\ event & Event \\ Treatment & O_{11} & O_{12} \\ Control & O_{21} & O_{22} \end{array}$$

Speaker notes

Chi-square test of independence, 2 of 2

$$\begin{array}{ccc} & No\ event & Event \\ Treatment & E_{11} = n_1(1 - \hat{p}_.) & E_{12} = n_1\hat{p}_. \\ Control & E_{21} = n_2(1 - \hat{p}_.) & E_{22} = n_2\hat{p}_. \end{array}$$

- $X^2 = \Sigma \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$

Speaker notes

Expected counts for Titanic

Observed counts

```
        Survived
Sex      Yes  No
  female 308 154
  male   142 709
```

Expected counts ::: {.cell} ::: {.cell-output .cell-output-stdout}

```
        Survived
Sex           Yes       No
  female 158.3397 303.6603
  male   291.6603 559.3397
```

::: :::

Chisquare test for Titanic

```
# A tibble: 1 × 4
  statistic  p.value parameter method
      <dbl>    <dbl>     <int> <chr>
1      332. 3.43e-74         1 Pearson's Chi-squared test
```

Odds ratio calculation

	No event	Event	Odds
Group1	a	b	
Group2	c	d	

- Odds for group 1 = b/a
- Odds for group 2 = d/c
- Odds for group 1 = $\frac{d/c}{b/a} = \frac{ad}{bc}$
- s.e.(log or) = $\sqrt{\frac{1}{a} + \frac{1}{b} + \frac{1}{c} + \frac{1}{d}}$

Titanic data

	Survived	Died	Total
Female	308	154	462
Male	142	709	851
Total	450	863	1,313

Speaker notes

Titanic data, odds of death

	Survived	Died	Total	Odds
Female	308	154	462	2 to 1 against
Male	142	709	851	4.993 to 1 in favor
Total	450	863	1,313	

Odds ratio = 4.993 / 0.5 = 9.986

Speaker notes

Clearly, a male passenger on the Titanic was more likely to die than a female passenger. But how much more likely? You can compute the odds ratio or the relative risk to answer this question.

The odds ratio compares the relative odds of death in each group. For females, the odds were exactly 2 to 1 against dying (154/308=0.5). For males, the odds were almost 5 to 1 in favor of death (709/142=4.993). The odds ratio is 9.986 (4.993/0.5). There is a ten fold greater odds of death for males than for females.

Odds ratio for survival by sex

```
$data
        Survived
Sex         Yes  No Total
  female 308 154   462
  male   142 709   851
  Total  450 863  1313

$measure
        odds ratio with 95% C.I.
Sex       estimate    lower    upper
  female 1.000000       NA       NA
  male   9.956188 7.662525 13.00928

$p.value
        two-sided
```

Live demo, Review of simple tests of two proportions

Break #1

- What you have learned
  - Review of simple tests of two proportions
- What's coming next
  - Concepts behind the logistic regression model

What is logistic regression?

- Binary outcome
- Categorical or continuous predictors
- Linear on the log odds scale

Speaker notes

The logistic regression model is a model that uses a binary (two possible values) outcome variable. Examples of a binary variable are mortality (live/dead), and morbidity (healthy/diseased). Sometimes you might take a continuous outcome and convert it into a binary outcome. For example, you might be interested in the length of stay in the hospital for mothers during an unremarkable delivery. A binary outcome might compare mothers who were discharged within 48 hours versus mothers discharged more than 48 hours.

The covariates in a logistic regression model represent variables that might be associated with the outcome variable. Covariates can be either continuous or categorical variables.

For binary outcomes, you might find it helpful to code the variable using indicator variables. An indicator variable equals either zero or one. Use the value of one to represent the presence of a condition and zero to represent absence of that condition. As an example, let 1=diseased, 0=healthy.

Why log odds?

- Statistical model of surgery
  - Estimates probability of demise
  - First prediction: probability=1.2
- Log odds prevent out of range predictions

Speaker notes

A logistic regression model examines the relationship between one or more independent variable and the log odds of your binary outcome variable. Log odds seem like a complex way to describe your data, but when you are dealing with probabilities, this approach leads to the simplest description of your data that is consistent with the rules of probability.

Let’s consider an artificial data example where we collect data on the gestational age of infants (GA), which is a continuous variable, and the probability that these infants will be breast feeding at discharge from the hospital (BF), which is a binary variable. We expect an increasing trend in the probability of BF as GA increases. Premature infants are usually sicker and they have to stay in the hospital longer. Both of these present obstacles to BF.

A linear model for probability, 1

GA	prob BF
28	60%
29	62%
30	64%
31	66%
32	68%
33	70%
34	72%

Speaker notes

A linear model would presume that the probability of BF increases as a linear function of GA. You can represent a linear function algebraically as

prob BF = a + b*GA

This means that each unit increase in GA would add b percentage points to the probability of BF. The table shown below gives an example of a linear function.

Figure 1. Hypothetical probabilities from an additive model

This table represents the linear function

prob BF = 4 + 2*GA

which means that you can get the probability of BF by doubling GA and adding 4. So an infant with a gestational age of 30 would have a probability of **4+2*30 = 64**.

A simple interpretation of this model is that each additional week of GA adds an extra 2% to the probability of BF. We could call this an additive probability model.

A linear model of probability, 2

GA	prob BF
28	88 %
29	91 %
30	94 %
31	97 %
32	100%
33	103%
34	106%

Speaker notes

I’m not an expert on BF; what little experience I’ve had with the topic occurred over 67 years ago. But I do know that an additive probability model tends to have problems when you get probabilities close to 0% or 100%**. Let’s change the linear model slightly to the following:

prob BF = 4 + 3*GA

This model would produce the following table of probabilities.

Figure 2. Hypothetical probabilities from an alternative additive model

You may find it difficult to explain what a probability of 106% means. This is a reason to avoid using a additive model for estimating probabilities. In particular, try to avoid using an additive model unless you have good reason to expect that all of your estimated probabilities will be between 20% and 80%.

A multiplicative model for probability

GA	prob BF
28	0.01 %
29	0.03 %
30	0.09 %
31	0.27 %
32	0.81 %
33	2.43 %
34	7.29 %

Speaker notes

It’s worthwhile to consider a different model here, a multiplicative model for probability, even though it suffers from the same problems as the additive model.

In a multiplicative model, you change the probabilities by multiplying rather than adding. Here’s a simple example.

Figure 3. Hypothetical probabilities from a multiplicative model

In this example, each extra week of GA produces a tripling in the probability of BF. Contrast this to the linear models shown above, where each extra week of GA adds 2% or 3% to the probability of BF.

A multiplicative model can’t produce any probabilities less than 0%, but it’s pretty easy to get a probability bigger than 100%. A multiplicative model for probability is actually quite attractive, as long as you have good reason to expect that all of the probabilities are small, say less than 20%.

The relationship between odds and probability

- odds = prob / (1-prob)
- prob = odds / (1+odds)
  - $0 \leq$ prob ≤ 1
  - $0 \leq$ odds $\leq \infty$
    - $0 \leq$ odds against ≤ 1
    - $1 \leq$ odds in favor $\leq \infty$

Speaker notes

Another approach is to try to model the odds rather than the probability of BF. You see odds mentioned quite frequently in gambling contexts. If the odds are three to one in favor of your favorite football team, that means you would expect a win to occur about three times as often as a loss. If the odds are four to one against your team, you would expect a loss to occur about four times as often as a win.

You need to be careful with odds. Sometimes the odds represent the odds in favor of winning and sometimes they represent the odds against winning. Usually it is pretty clear from the context. When you are told that your odds of winning the lottery are a million to one, you know that this means that you would expect to having a losing ticket about a million times more often than you would expect to hit the jackpot.

It’s easy to convert odds into probabilities and vice versa. With odds of three to one in favor, you would expect to see roughly three wins and only one loss out of every four attempts. In other words, your probability for winning is 0.75.

If you expect the probability of winning to be 20%, you would expect to see roughly one win and four losses out of every five attempts. In other words, your odds are 4 to 1 against.

The formulas for conversion are

odds = prob / (1-prob)

and

prob = odds / (1+odds).

In medicine and epidemiology, when an event is less likely to happen and more likely not to happen, we represent the odds as a value less than one. So odds of four to one against an event would be represented by the fraction 1/5 or 0.2. When an event is more likely to happen than not, we represent the odds as a value greater than one. So odds of three to one in favor of an event would be represented simply as an odds of 3. With this convention, odds are bounded below by zero, but have no upper bound.

A log odds model for probability, 1

GA	odds BF
28	27 to 1 against (.037)
29	9 to 1 against (.111)
30	3 to 1 against (.333)
31	1 to 1 (1)
32	3 to 1 in favor (3)
33	9 to 1 in favor (9)
34	27 to 1 in favor (27)

Speaker notes

Let’s consider a multiplicative model for the odds (not the probability) of BF.

Figure 4. Hypothetical odds from a multiplicative model

This model implies that each additional week of GA triples the odds of BF. A multiplicative model for odds is nice because it can’t produce any meaningless estimates.

A log odds model for probability, 2

GA	odds BF	log odds
28	27 to 1 against (.037)	-3.30
29	9 to 1 against (.111)	-2.20
30	3 to 1 against (.333)	-1.10
31	1 to 1 (1)	0.00
32	3 to 1 in favor (3)	1.10
33	9 to 1 in favor (9)	2.20
34	27 to 1 in favor (27)	3.30

Speaker notes

It’s interesting to look at how the logarithm of the odds behave.

Notice that an extra week of GA adds 1.1 units to the log odds. So you can describe this model as linear (additive) in the log odds. When you run a logistic regression model in SPSS or other statistical software, it uses a model just like this, a model that is linear on the log odds scale. This may not seem too important now, but when you look at the output, you need to remember that SPSS presents all of the results in terms of log odds. If you want to see results in terms of probabilities instead of logs, you have to transform your results.

A log odds model for probability, 3

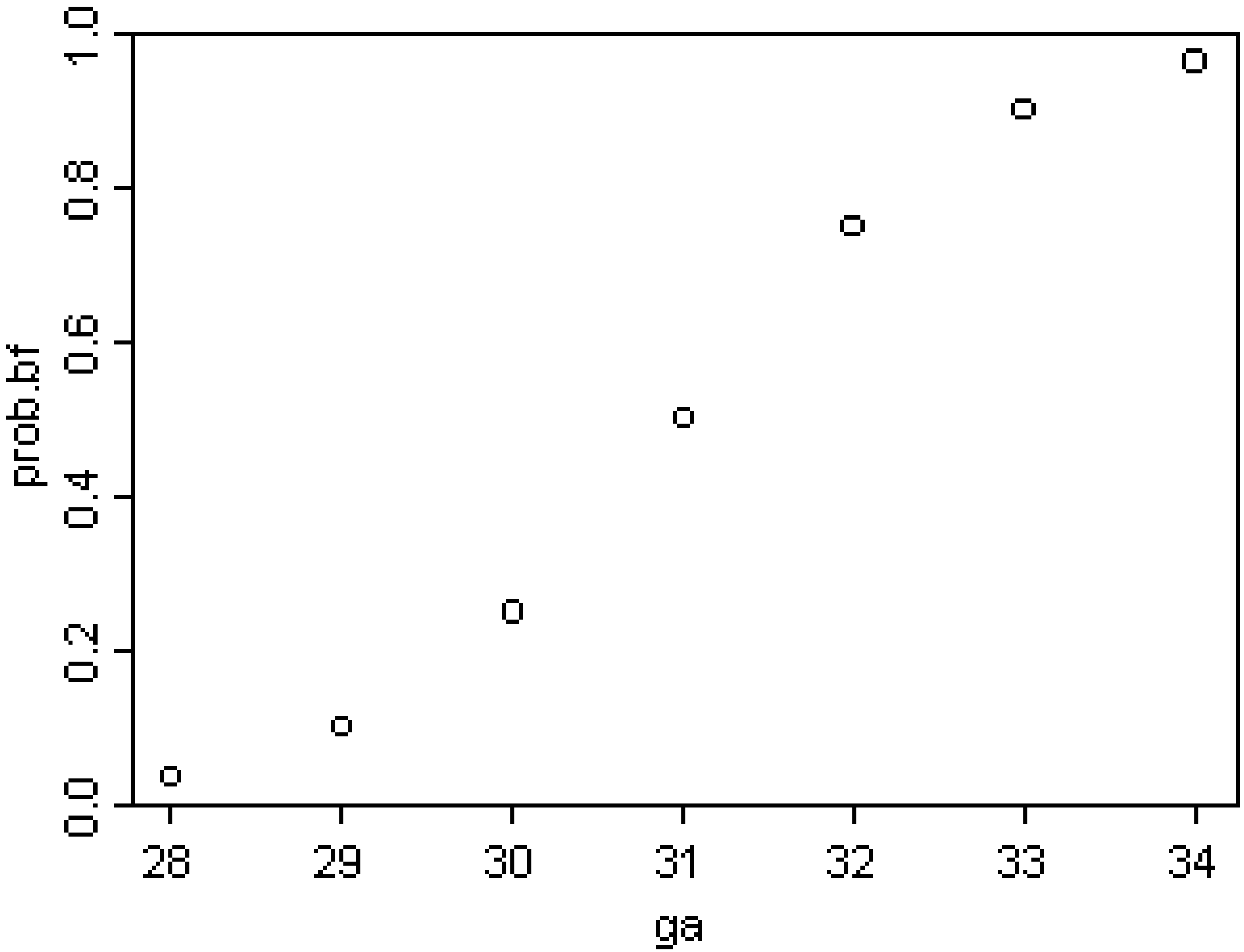
GA	odds BF	prob BF
28	27 to 1 against (.037)	3.6%
29	9 to 1 against (.111)	10.0%
30	3 to 1 against (.333)	25.0%
31	1 to 1 (1)	50.0%
32	3 to 1 in favor (3)	75.0%
33	9 to 1 in favor (9)	90.0%
34	27 to 1 in favor (27)	96.4%

Speaker notes

Let’s look at how the probabilities behave in this model.

Notice that even when the odds get as large as 27 to 1, the probability still stays below 100%. Also notice that the probabilities change in neither an additive nor a multiplicative fashion.

A log odds model for probability, 4



Speaker notes

A graph shows what is happening.

The probabilities follow an S-shaped curve that is characteristic of all logistic regression models. The curve levels off at zero on one side and at one on the other side. This curve ensures that the estimated probabilities are always between 0% and 100%.

An example of a log odds model with real data, 1

GA	Actual prob BF
28	2/6 = 33.3%
29	2/5 = 40.0%
30	7/9 = 77.8%
31	7/9 = 77.8%
32	16/20 = 80.0%
33	14/15 = 93.3%

Speaker notes

There are other approaches that also work well for this type of data, such as a probit model, that I won’t discuss here. But I did want to show you what the data relating GA and BF really looks like.

An example of a log odds model with real data, 2

Speaker notes

I’ve simplified this data set by removing some of the extreme gestational ages.

The table below shows the predicted log odds, and the calculations needed to transform this estimate back into predicted probabilities.

An example of a log odds model with real data, 3

- log odds = -16.72 + 0.577 $\times$ 30 = 0.59
- odds = exp(log odds) = 1.8
- prob = odds / (1+odds) = 0.64

Speaker notes

Let's examine these calculations for GA = 30. The predicted log odds would be the intercept plus the slope times 30.

Convert from log odds to odds by exponentiating.

And finally, convert from odds back into probability.

prob = 1.80 / (1+1.80) = 0.643

The predicted probability of 64.3% is reasonably close to the true probability (77.8%).

You might also want to take note of the predicted odds. Notice that the ratio of any odds to the odds in the next row is 1.78. For example,

3.20/1.80 = 1.78

5.70/3.20 = 1.78

It's not a coincidence that you get the same value when you exponentiate the slope term in the log odds equation.

exp(0.59) = 1.78

This is a general property of the logistic model. The slope term in a logistic regression model represents the log of the odds ratio. This represents the increase (decrease) in risk as the independent variable increases by one unit.

Live demo, Concepts behind the logistic regression model

Break #2

- What you have learned
  - Concepts behind the logistic regression model
- What's coming next
  - Logistic regression with categorical predictors

Always start with descriptive statistics

```
# A tibble: 2 × 2
  Sex    pct
  <chr>  <glue>
1 female 35% (462/1313)
2 male   65% (851/1313)
```

```
# A tibble: 3 × 2
  PClass pct
  <fct>  <glue>
1 3rd    54% (711/1313)
2 2nd    21% (280/1313)
3 1st    25% (322/1313)
```

```
# A tibble: 2 × 2
  Survived pct
  <fct>    <glue>
1 Yes      34% (450/1313)
2 No       66% (863/1313)
```

Compute survival probabilites by sex

```
# A tibble: 2 × 4
# Groups:   Sex [2]
  Sex    Survived pct                odds
  <chr>  <fct>    <glue>            <dbl>
1 female No       33% (154/462)  0.5 
2 male   No       83% (709/851)  4.99
```

```
# A tibble: 2 × 4
# Groups:   Sex [2]
  Sex    Survived pct                odds
  <chr>  <fct>    <glue>            <dbl>
1 female Yes      67% (308/462)   2  
2 male   Yes      17% (142/851)   0.2
```

Logistic model estimates

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl> <glue>
1 (Intercept)   -0.693    0.0987     -7.02 p < 0.001
2 Sexmale        2.30     0.135      17.1  p < 0.001
```

- exp(2.3) = 9.99
- exp(2.3-1.96*0.135) = 7.67
- exp(2.3+1.96*0.135) = 13.0

```
# A tibble: 1 × 4
  term    odds_ratio lower upper
  <chr>        <dbl> <dbl> <dbl>
1 Sexmale       9.99  7.67  13.0
```

Break #3

- What you have learned
  - Logistic regression with categorical predictors
- What's coming next
  - Logistic regression with continuous predictors

Always start with descriptive statistics

```
# A tibble: 1 × 4
  age_mean age_sd age_min age_max
     <dbl>  <dbl>   <dbl>   <dbl>
1     30.4   14.3    0.17      71
```

Logistic regression using Age to predict Survived

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl> <glue>
1 (Intercept)  0.0814    0.174       0.468 p = 0.64
2 Age          0.00879   0.00523     1.68  p = 0.093
```

Back transform

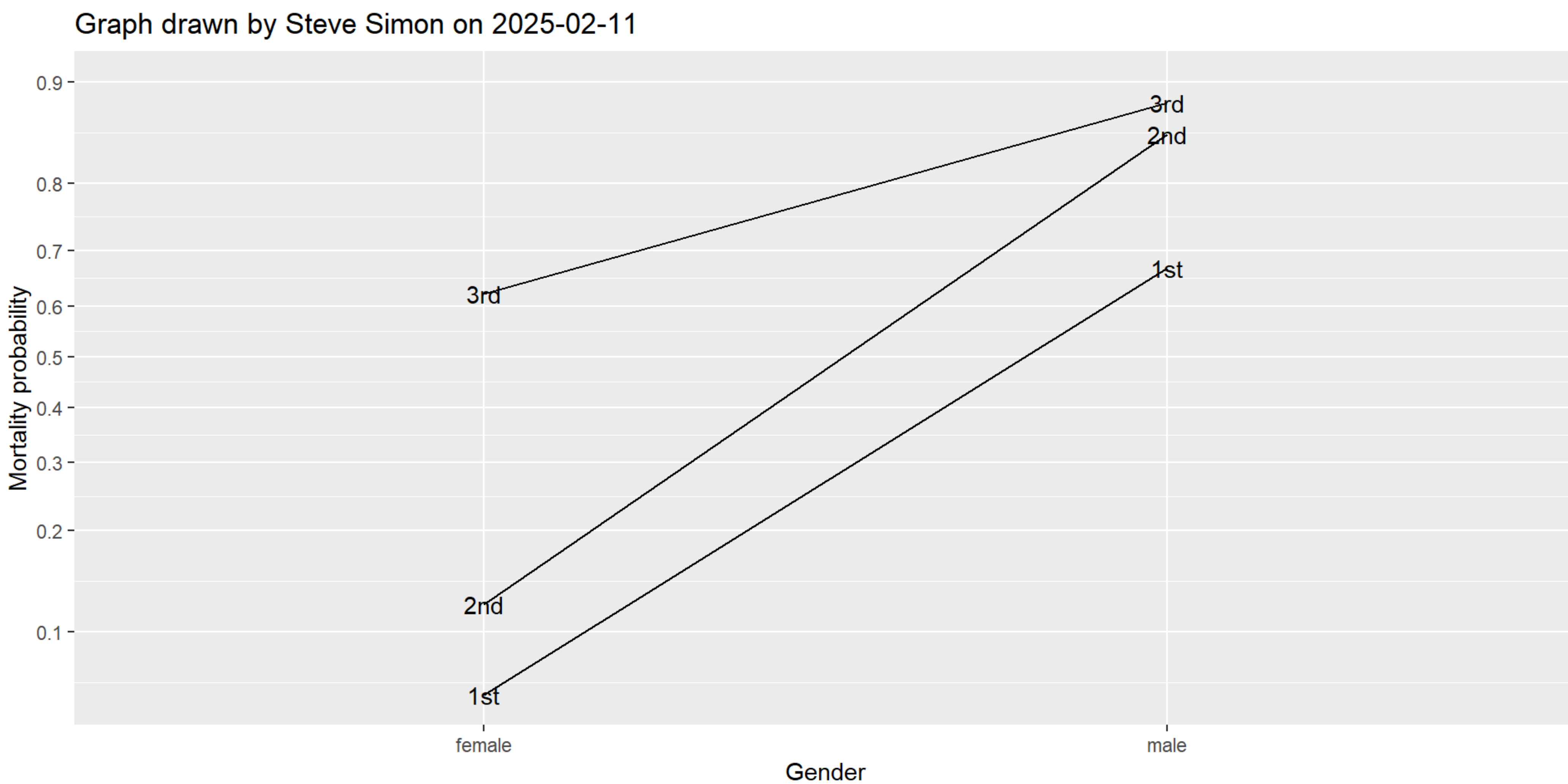
```
# A tibble: 1 × 4
  term  odds_ratio lower upper
  <chr>      <dbl> <dbl> <dbl>
1 Age         1.01 0.999  1.02
```

Live demo, Logistic regression with continuous predictors

Break #4

- What you have learned
  - Logistic regression with continuous predictors
- What's coming next
  - Logistic regression with interactions

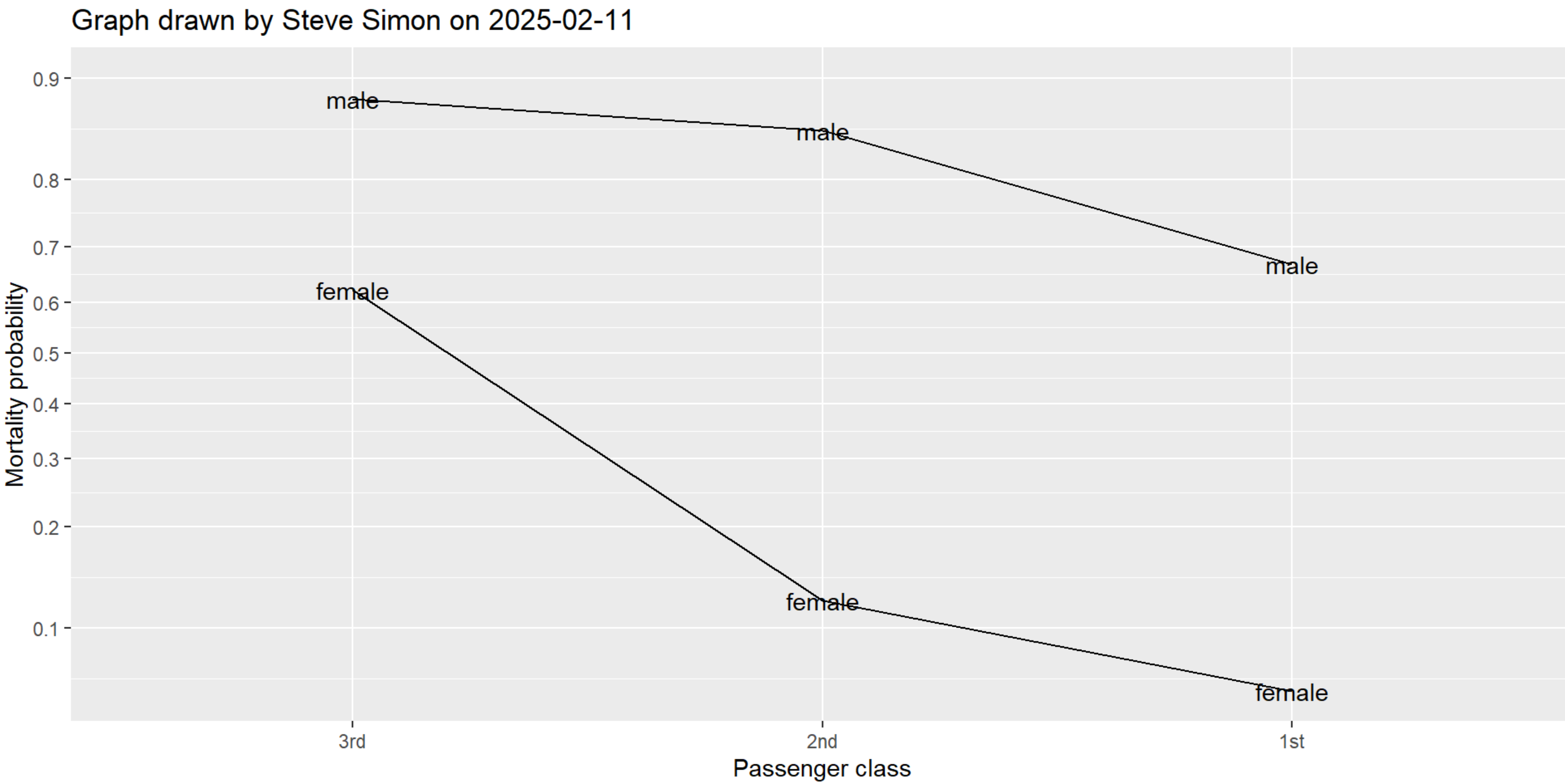
Line plot, 1


Graph drawn by Steve Simon on 2025-02-11
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
Mortality probability
3rd
2nd
1st
3rd
2nd
1st
female
male
Gender

Speaker notes

Notice how the spacing on the y-axis is not linear. This graph used a logit transformation. Logit is short for the log odds transformation. The demo program for this module shows how to graph the interaction on the probability scale or on the log odds scale. If most of the probabilities are between 20% and 80%, you won’t see any major differences between these two plots. If many of the probabilities are small (less than 20%) or large (greater than 80%), then you should only interpret the interaction plot on the log odds scale.

Line plot, 2



Graph drawn by Steve Simon on 2025-02-11
Mortality probability
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
male
male
male
female
female
female
3rd
2nd
1st
Passenger class

Speaker notes

This plot switches what goes on the x-axis and what goes on the individual lines. You should look at both plots, but please only use one in your reports.

Logistic regression model with interaction

```
# A tibble: 2 × 5
  term        estimate std.error statistic p.value
  <chr>          <dbl>     <dbl>     <dbl> <glue>
1 (Intercept)  0.0814    0.174       0.468 p = 0.64
2 Age          0.00879   0.00523     1.68  p = 0.093
```

Live demo, Logistic regression with interactions

Break #5

- What you have learned
  - Logistic regression with interactions
- What's coming next
  - Diagnostics

Diagnostics

- Comparison to null model
- Linearity
  - Only for continuous predictors
- Independence
  - Assessed qualitatively

Live demo, Diagnostics

Summary

- What you have learned
  - Review of simple tests of two proportions
  - Concepts behind the logistic regression model
  - Logistic regression with categorical predictors
  - Logistic regression with continuous predictors
  - Logistic regression with interactions
  - Diagnostics