Regression models for epidemiological research

Advanced Epidemiology

Brooke Anderson

: brooke.anderson@colostate.edu

: www.github.com/geanders

Department of Environmental & Radiological Health Sciences
Environmental Epidemiology Section
Colorado State University

Further reading and sources

- Woodward (2014) *Epidemiology: Study Design and Data Analysis.* Chapman & Hall.
- Vittinghoff et al. (2005) *Regression Methods in Biostatistics: Linear, Logistic, Survival, and Repeated Measures Models.* Springer.
- Harrell. (2001) *Regression Modelling Strategies: With Applications to Linear Models, Logistic and Ordinal Regression, and Survival Analysis.* Springer.

Basics of logarithms

There are a few basics of logartithms you should keep in mind for this lecture (these all use *log* for a natural logarithm):

1. If $log(a) = b$, then $a = e^b$.
2. $log(e^a) = a$. Similarly, $e^{log(a)} = a$.
3. $log(1) = 0$ and $e^0 = 1$.
4. $log(a * b) = log(a) + log(b)$.
5. $log(a/b) = log(a) - log(b)$. As a result, $log(1/a) = -log(a)$.

General model equation for generalized linear regression

Systematic part of generalized linear models (GLMs):

$$g(E[Y_i]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...$$

where:

- $E[Y_i]$ is the expected value of the outcome (Y_i) for observation i
- $g(.)$ is a function linking the expected value of the outcomes to the predictors (identity function for linear regression, logit function for logistic regression, etc.)
- β_0 is the model intercept
- X_1, X_2, X_3, etc., are predictor variables
- β_1, β_2, β_3, etc., are parameters describing the relationship between the predictor variables and the outcome

The right-hand side of this model equation is the **linear predictor**.

General model equation for generalized linear regression

Random part of GLMs:

- What is the distribution of the outcome (Y_i) conditional on the observed values of the predictors (X_1, X_2, etc.)?

Fitting GLMs

GLMs are fit through **maximum-likelihood estimation**. For each candidate model, the **likelihood** of the data (joint probability of the data) under that model is measured. The **maximum-likelihood estimates** for the parameters are the values that maximize the likelihood of the observed data.

General model equation for generalized linear regression

Simple regression

If only one explanatory variable is included, it's a "simple" regression:

$$g(E[Y_i]) = \beta_0 + \beta_1 X_1$$

Multiple regression

If two or more explanatory variables are included, it's a "multiple" regression:

$$g(E[Y_i]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ...$$

An epidemiologist walks into a movie theater...

The mortality outcomes, by sex, of passengers on the Titanic were (note: some passengers with missing data have been excluded):

	Outcome	
Sex	Survived	Died
Female	292	96
Male	135	523

Data obtained from `http://biostat.mc.vanderbilt.edu/wiki/Main/DataSets`. Original data source: Hind (1999) *Encyclopedia Titanica.* `http://atschool.eduweb.co.uk/phind`. Figure source: `http://www.imdb.com`

Example– Surviving the Titanic

Based on this table (i.e., using contingency table methods), determine and discuss:

- Does sex affect the odds of dying during the Titanic sinking?
- Are the odds that someone like Jack will die on the Titanic versus the odds that someone like Rose will?
- How do these two questions differ?
- Are the odds of dying on the Titanic significantly higher (statistically) for males versus females? What about for someone like Jack versus someone like Rose?
- What other information would you like to have to better answer these questions? How would the information help to answer the previous questions?

Example– Surviving the Titanic

	Outcome	
Sex	Survived	Died
Female	292	96
Male	135	523

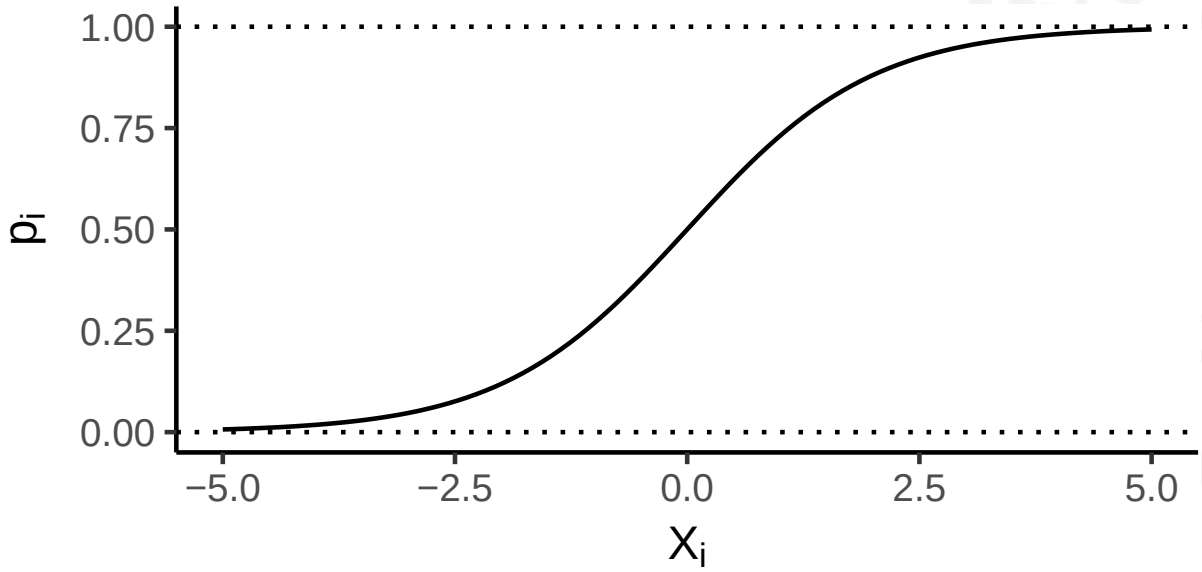
- The odds for dying on the Titanic for females is $\frac{96}{292} = 0.33$
- The odds for dying on the Titanic for males is $\frac{523}{135} = 3.87$
- The odds ratio for dying on the Titanic for males compared to females is $\frac{523*292}{135*96} = 11.78$
- The log odds ratio is $log(11.78) = 2.47$
- The estimated standard error for the log odds ratio is $\sqrt{\frac{1}{292} + \frac{1}{96} + \frac{1}{135} + \frac{1}{523}} = 0.15$
- The 95% confidence interval for the odds ratio is (8.74, 15.88)

Logistic function

Logistic function, with $p_i = Pr(Y_i = 1|X_i)$:

$$p_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_i}}$$

Plot of logistic function with $\beta_0 = 0$, $\beta_1 = 1$ (note that y is always between 0 and 1):



Logit function

Notice what happens if you change the equation so the right-hand side matches the typical model equation format for a GLM.

Before:

$$p_i = \frac{1}{1 + e^{-\beta_0 - \beta_1 X_i}}$$

After:

$$log(\frac{p_i}{1 - p_i}) = \beta_0 + \beta_1 X_i$$

The left-hand side of this equation is the **logit** of p_i.

Logistic regression example

Let's fit a logistic regression for the previous example on sex and risk of dying in the Titanic sinking. The systematic part of this model is:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1,i}$$

where:

- p_i: Probability person i died during the sinking, $Pr(Y_i = 1|X_i)$
- Y_i: Whether person i died during the sinking

$$X_i = \begin{cases} 0 \text{ if person } i \text{ is female} \\ 1 \text{ if person } i \text{ is male} \end{cases}$$

The random part of the model is: Outcomes (Y_i) follow a binomial distribution.

Interpreting coefficients– logistic regression

For males, this equation evaluates to:

$$log(\frac{p_m}{1-p_m}) = \beta_0 + \beta_1 * 1 = \beta_0 + \beta_1$$

For females, this equation evaluates to:

$$log(\frac{p_f}{1-p_f}) = \beta_0 + \beta_1 * 0 = \beta_0$$

Therefore:

- The log odds for males is estimated by $\beta_0 + \beta_1$
- The log odds for females is estimated by β_0.

Interpreting coefficients– logistic regression

Now look at what happens when you subtract the log odds for males from the log odds for females:

$$log(\frac{p_m}{1-p_m}) - log(\frac{p_f}{1-p_f}) = (\beta_0 + \beta_1) - (\beta_0)$$

You can rearrange this to:

$$log\left(\frac{\frac{p_m}{1-p_m}}{\frac{p_f}{1-p_f}}\right) = \beta_1$$

- The log(odds ratio) (which equals the difference in the log(odds)) for males compared to females is estimated by $\hat{\beta}_1$.

Example– Surviving the Titanic

To fit this model, you'll want to have your data in a form where there is one row per observation (person in this case), with columns for the outcome (died [1] / survived [0]) and predictor variable (sex: male [1] / female [0]). The first few rows might look like:

Outcome	Sex
0	0
0	1
1	0
1	1
1	0
0	1

Note that if you take the mean of the `Outcome` column, it gives you the probability of death across the observations ($p_i = E(Y_i)$).

Example– Surviving the Titanic

Here are results from fitting a logistic regression to the data on sex and odds of death on the Titanic:

(Intercept)	-1.112^{***}
	(0.118)
Sex: Male/Female	2.467^{***}
	(0.152)
Log-likelihood	-551.0
Deviance	1102.0
AIC	1106.0
BIC	1115.9

Values for each row of the top of the table are estimated model coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$). Values in parentheses are estimated standard errors for each coefficient. Stars indicate the range of the p-value for the estimated coefficient.

Example– Surviving the Titanic

From the previous table, the p-value for each parameter is for a hypothesis test of:

$$H_0 : \beta_j = 0$$

$$H_a : \beta_j \neq 0$$

We already saw that $\hat{\beta}_1$ in the previous model is estimating the log odds ratio of dying during the Titanic sinking for males versus females. If the odds are identical for males and females, the odds ratio would be 1, so the log odds ratio would be $log(1) = 0$.

Logistic regression example

The systematic model form we fit was:

$$log(\frac{p_i}{1-p_i}) = \hat{\beta}_0 + \hat{\beta}_1 X_{1,i}$$

The model we estimated based on the data is:

$$log(\frac{p_i}{1-p_i}) = -1.112 + 2.467 X_{1,i}$$

where:

$$X_i = \begin{cases} 0 \text{ if person } i \text{ is female} \\ 1 \text{ if person } i \text{ is male} \end{cases}$$

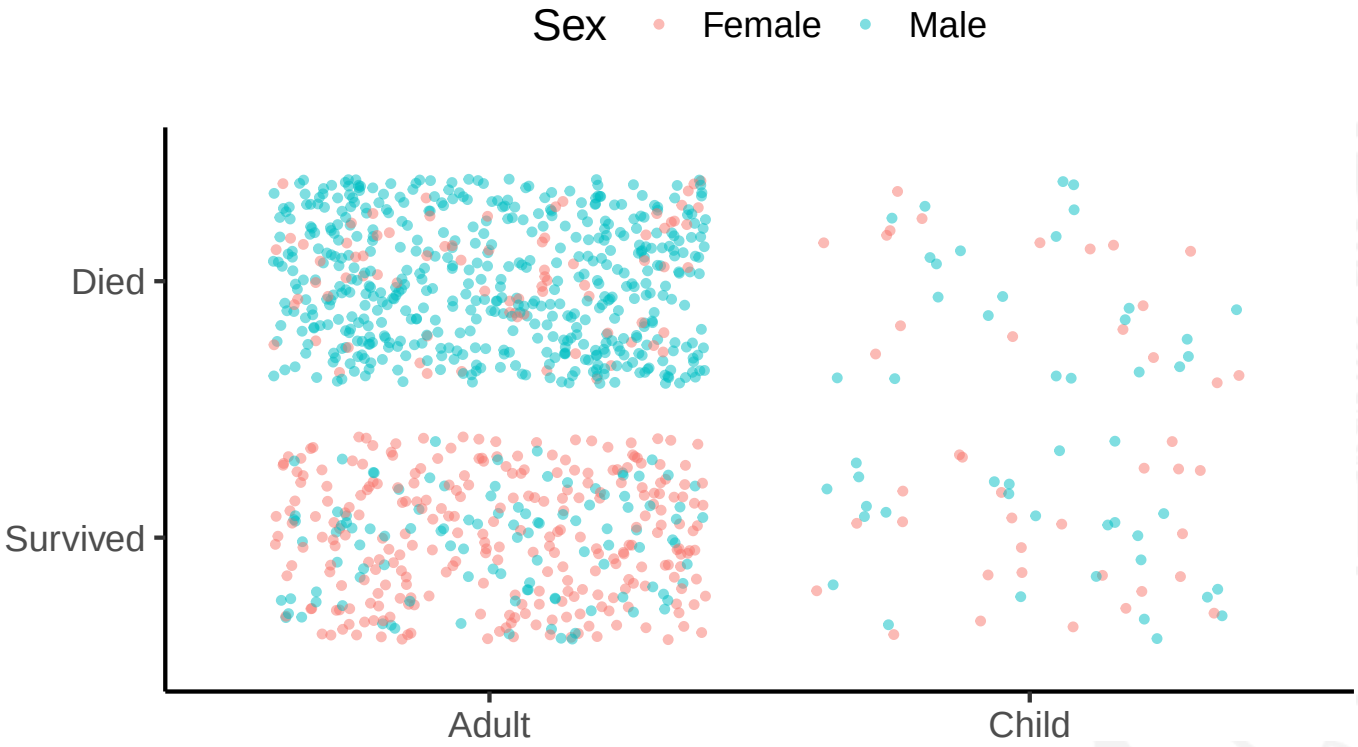
Logistic regression example

$$log(\frac{p_i}{1-p_i}) = -1.112 + 2.467X_{1,i}$$

- The log odds for females is estimated by -1.112. The odds for females of dying during the sinking of the Titanic is estimated as $e^{-1.112} = 0.33$. This corresponds with about a 25% risk of dying.
- The log odds for males is estimated as $-1.112 + 2.467 = 1.355$. The odds for males of dying during the sinking of the Titanic is estimated as $e^{1.355} = 3.88$. This corresponds with about an 80% risk of dying.
- The log odds ratio for males versus females is estimated as 2.467. The odds ratio for males versus females is estimated as $e^{2.467} = 11.78$.
- The standard error of the log odds ratio is estimated as 0.1522. The estimated 95% confidence interval for the log odds ratio is $2.467 \pm 1.96(0.1522) = (2.169, 2.765)$. The 95% confidence interval for the odds ratio is $(e^{2.169}, e^{2.765}) = (8.75, 15.88)$.

Example– Surviving the Titanic

Our analysis so far has not made any consideration for the age of each person in the data. Here is the data divided by age group:



Each point represents a person in the data. Color shows whether the person was male or female. The quadrant in which the point is plotted shows whether the person was an adult or a child (x-axis) and whether the person survived or died (y-axis).

Example– Surviving the Titanic

Two things we might want to consider are:

- Would the odds ratio for males versus females be different if we **adjusted** for age?
- Is there an **interaction** between sex and age in the odds of death during the sinking of the Titanic?

How would you assess these two questions based on a table of the data?

		Outcome	
Sex	Age	Survived	Died
Female	Adult	267	79
	Child	25	17
Male	Adult	109	500
	Child	26	23

Adjusting for a variable

- "Adjusting" estimates the association for a predictor and outcome while ensuring the comparison is within the same strata of another variable (e.g., comparing odds for males versus females while ensuring that adults are compared to adults and children to children).

Idea of adjusting for a variable in Titanic example

We assume there is a constant odds ratio of dying for males versus females, regardless of age. However, in our simple logistic regression, we may not be estimating this odds ratio well because:

1. The proportion of males versus females may differ by age category.
2. The odds of death may differ by age category.

Adjusting for a variable

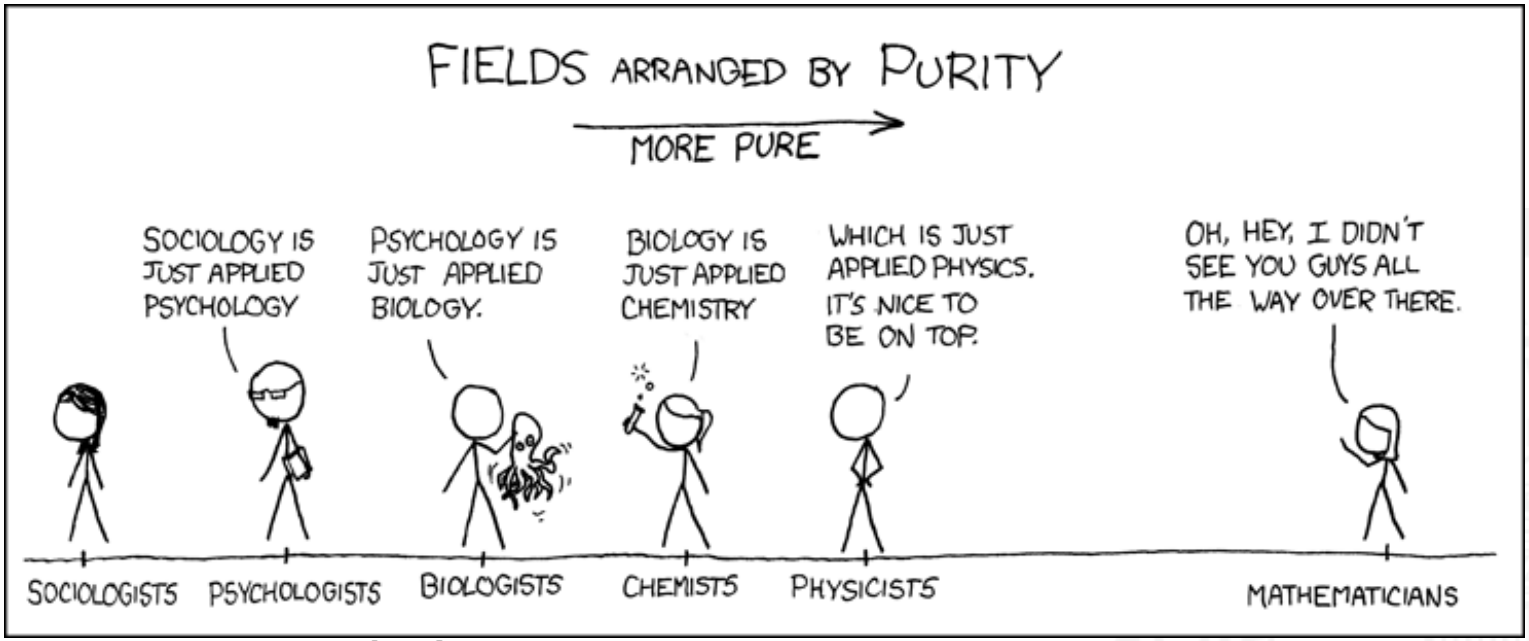


Figure source: `www.xkcd.com`.

- Adjusting for a variable can help correct for bias from a confounder.
- For continuous outcomes, adjusting for a variable that helps explain residual variance in the outcome can improve efficiency

Shisterman et al. (2009) Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology* 20(4):488–295.

Investigating interaction with a variable

Idea of interaction in Titanic example

The odds ratio of dying during the Titanic sinking for males versus females is **different** within each age category. We will estimate separate odds ratios for adults and children.

- If we find that there is an interaction between sex and age in odds of dying, we would say that age **modifies** the association between sex and odds of dying. We would call age an **effect modifier** for this association. (Note that this is *not* mediation!)

Example– Surviving the Titanic

We can fit a model to estimate the log odds ratio of dying during the sinking of the Titanic for males versus females, adjusted for age, with the following regression model:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i}$$

where:

$$X_{1,i} = \begin{cases} 0 \text{ if person } i \text{ is female} \\ 1 \text{ if person } i \text{ is male} \end{cases}$$

$$X_{2,i} = \begin{cases} 0 \text{ if person } i \text{ is an adult} \\ 1 \text{ if person } i \text{ is a child} \end{cases}$$

Example– Surviving the Titanic

From this logistic regression model, here are the log odds that will be estimated for each group:

	Adult	Child
Female	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_2$
Male	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2$

Notice that these estimates always estimate the same difference in log odds between males and females within an age category ($\hat{\beta}_1$), regardless of which age category you consider. This is the estimated log odds ratio for males versus females, *adjusted for* or *controlling for* age category.

Example– Surviving the Titanic

Here are the results from fitting the regression model ("Sex and Age" column):

	Sex only	Sex and Age
(Intercept)	-1.112^{***}	-1.053^{***}
	(0.118)	(0.120)
Sex: Male/Female	2.467^{***}	2.463^{***}
	(0.152)	(0.153)
Age: Child/Adult		-0.641^{*}
		(0.261)
Log-likelihood	−551.0	−548.0
Deviance	1102.0	1096.0
AIC	1106.0	1102.0
BIC	1115.9	1116.9

Example– Surviving the Titanic

When we fit our data to the model, we get the following regression model:

$$log(\frac{p_i}{1-p_i}) = -1.0532 + 2.4634X_{1,i} + -0.6413X_{2,i}$$

where:

$$X_{1,i} = \begin{cases} 0 \text{ if person } i \text{ is female} \\ 1 \text{ if person } i \text{ is male} \end{cases}$$

$$X_{2,i} = \begin{cases} 0 \text{ if person } i \text{ is an adult} \\ 1 \text{ if person } i \text{ is a child} \end{cases}$$

Example– Surviving the Titanic

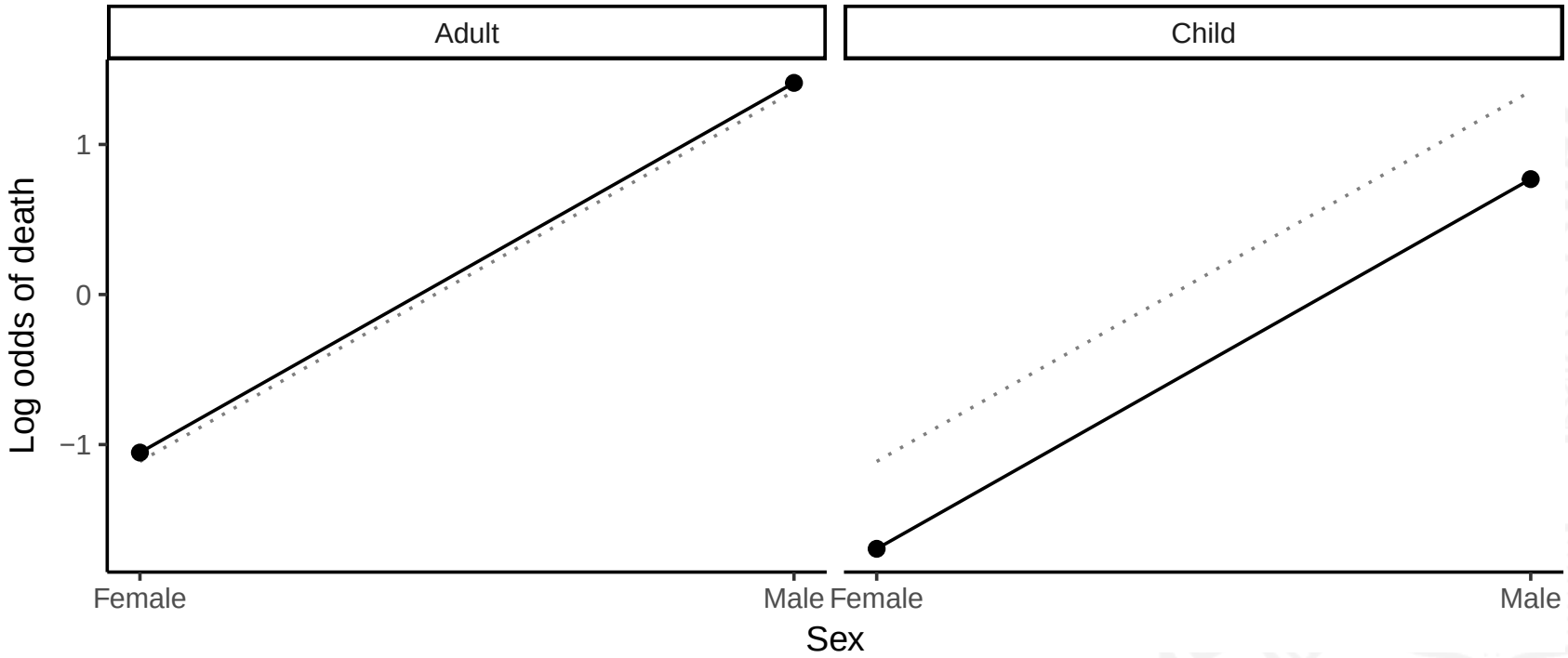
Here are the log odds estimated for each group:

	Adult	Child
Female	-1.0532	-1.0532 - 0.6413
Male	-1.0532 + 2.4634	-1.0532 + 2.4634 - 0.6413

- The estimated log odds ratio for males versus females, *adjusted for age group*, is 2.4634.
- The estimated odds ratio for males versus females, *adjusted for age group*, is $e^{2.4634} = 11.74$.
- The 95% confidence interval for the log odds ratio is $2.4634 \pm 1.96(0.1527) = (2.164, 2.763)$.
- The 95% confidence interval for the odds ratio is $(e^{2.164}, e^{2.763}) = (8.71, 15.84)$.

Example– Surviving the Titanic

Here are graphs of the estimated log odds within each group (the dotted line shows the estimate from the simple logistic regression):



In this analysis, we made the assumption that the odds ratio for males versus females is the same within each age category. However, we have allowed the age categories to have different baseline log odds.

Deviance / Likelihood test statistic

The deviance and log-likeihood values can be used to calculate test statistics for hypothesis tests of **nested** models. Models are nested if:

- The predictors for one model are a subset of the predictors for the other model.
- The models were fit using the same observations.

Information criteria (e.g., AIC, BIC) can be used to compare models whether they are tested or not. However, they can not be used for specific hypothesis tests.

Example– Surviving the Titanic

We can fit a model to estimate the log odds ratio of dying during the sinking of the Titanic for males versus females, with an interaction for age, with the following regression model:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \beta_3 X_{1,i} X_{2,i}$$

where:

$$X_{1,i} = \begin{cases} 0 \text{ if person } i \text{ is female} \\ 1 \text{ if person } i \text{ is male} \end{cases}$$

$$X_{2,i} = \begin{cases} 0 \text{ if person } i \text{ is an adult} \\ 1 \text{ if person } i \text{ is a child} \end{cases}$$

Example– Surviving the Titanic

From this logistic regression model, here are the odds that will be estimated for each group:

	Adult	Child
Female	$\hat{\beta}_0$	$\hat{\beta}_0 + \hat{\beta}_2$
Male	$\hat{\beta}_0 + \hat{\beta}_1$	$\hat{\beta}_0 + \hat{\beta}_1 + \hat{\beta}_2 + \hat{\beta}_3$

Notice that now the difference in log odds for males versus females is different depending whether the person is an adult (difference in log odds of $\hat{\beta}_1$ between males and females) or a child (difference in log odds of $\hat{\beta}_1 + \hat{\beta}_3$ between males and females). We are now estimating different odds ratios for males versus females for the two age categories.

Example– Surviving the Titanic

Here are the results from fitting the regression model ("Sex:Age" column):

	Sex only	Sex+Age	Sex:Age
(Intercept)	-1.112^{***}	-1.053^{***}	-1.218^{***}
	(0.118)	(0.120)	(0.128)
Sex: Male/Female	2.467^{***}	2.463^{***}	2.741^{***}
	(0.152)	(0.153)	(0.166)
Age: Child/Adult		-0.641^{*}	0.832^{*}
		(0.261)	(0.339)
Sex: Male/Female $\times$ Age: Child/Adult			-2.478^{***}
			(0.456)
Log-likelihood	−551.0	−548.0	−534.2
Deviance	1102.0	1096.0	1068.5
AIC	1106.0	1102.0	1076.5
BIC	1115.9	1116.9	1096.3

Example– Surviving the Titanic

When we fit our data to the model, we get the following regression model:

$$log(\frac{p_i}{1-p_i}) = -1.2178 + 2.7411X_{1,i} + 0.8321X_{2,i} + -2.4780X_{1,i}X_{2,i}$$

where:

$$X_{1,i} = \begin{cases} 0 \text{ if person } i \text{ is female} \\ 1 \text{ if person } i \text{ is male} \end{cases}$$

$$X_{2,i} = \begin{cases} 0 \text{ if person } i \text{ is an adult} \\ 1 \text{ if person } i \text{ is a child} \end{cases}$$

Example– Surviving the Titanic

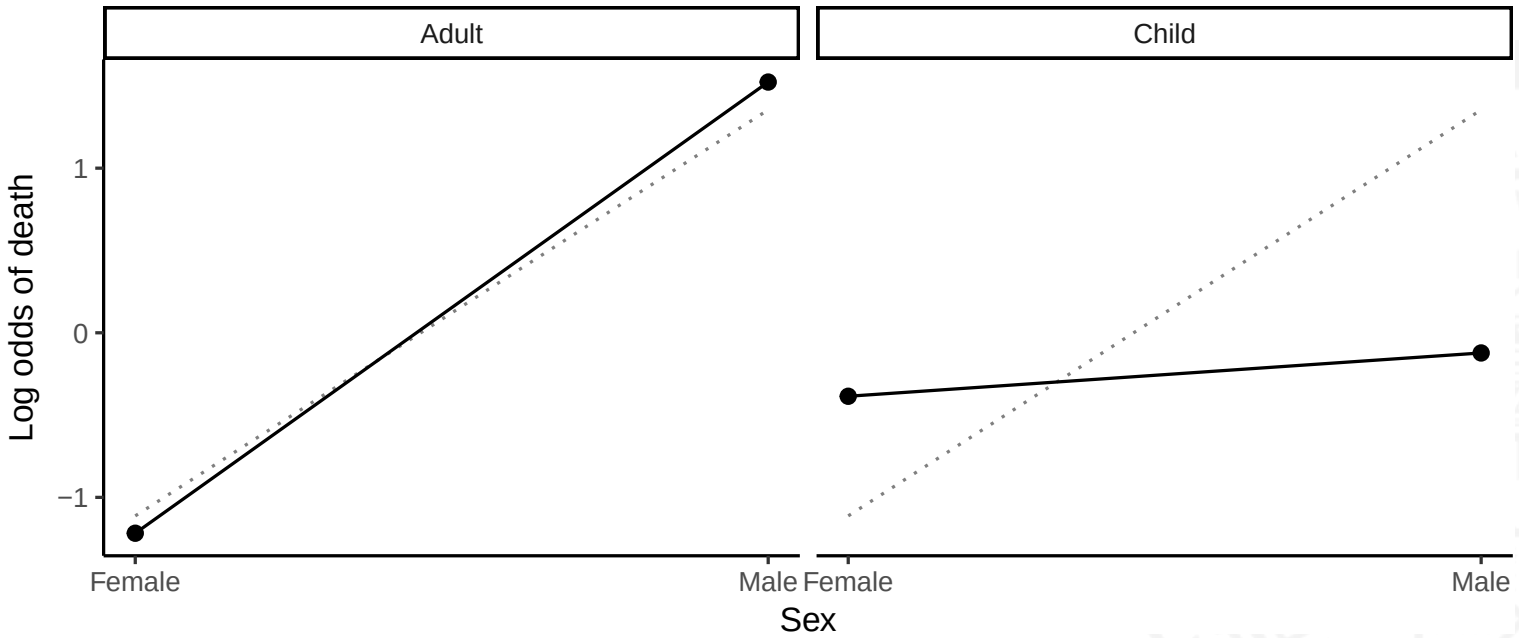
Here are the log odds estimated for each group:

	Adult	Child
Female	-1.2178	-1.2178 + 0.8321
Male	-1.2178 + 2.7411	-1.2178 + 2.7411 + 0.8321 - 2.4780

- The estimated log odds ratio for males versus females, *among adults*, is 2.7411.
- The estimated log odds ratio for males versus females, *among children*, is $2.7411 - 2.4780 = 0.2631$.
- The estimated odds ratio for males versus females is $e^{2.7411} = 15.50$ among adults and $e^{0.2631} = 1.30$ among children.
- The 95% confidence interval for the log odds ratio among adults is $2.7411 \pm 1.96(0.1661) = (2.416, 3.067)$.
- The 95% confidence interval for the odds ratio is $(e^{2.416}, e^{3.067}) = (11.20, 21.47)$.

Example– Surviving the Titanic

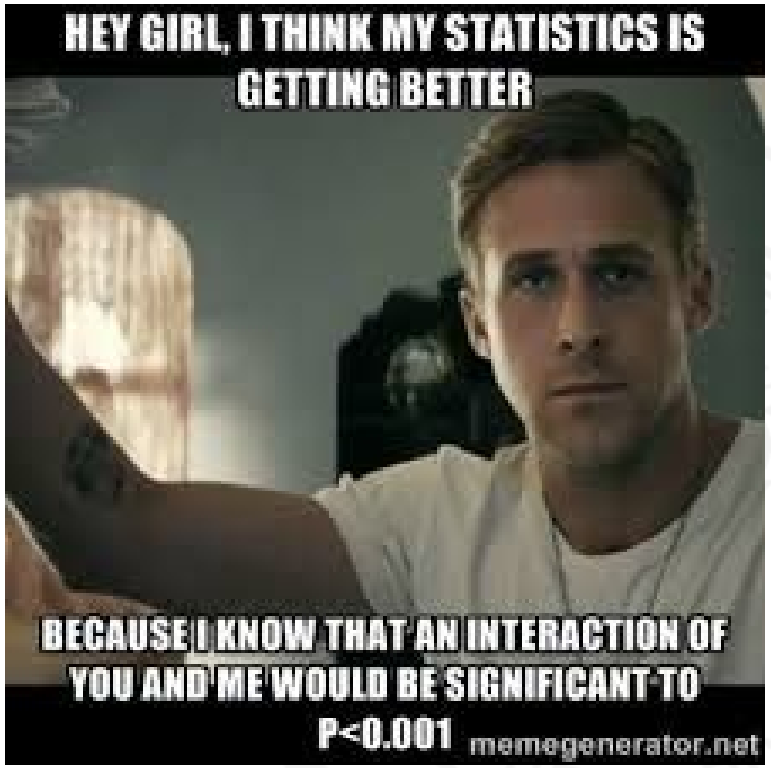
Here are graphs of the estimated log odds within each group (the dotted line shows the estimate from the simple logistic regression):



In this analysis, we have estimated **different** odds ratios within each age category. These two odds ratios are very different, indicating evidence of an interation between age and sex on the odds of dying during the sinking of the Titanic.

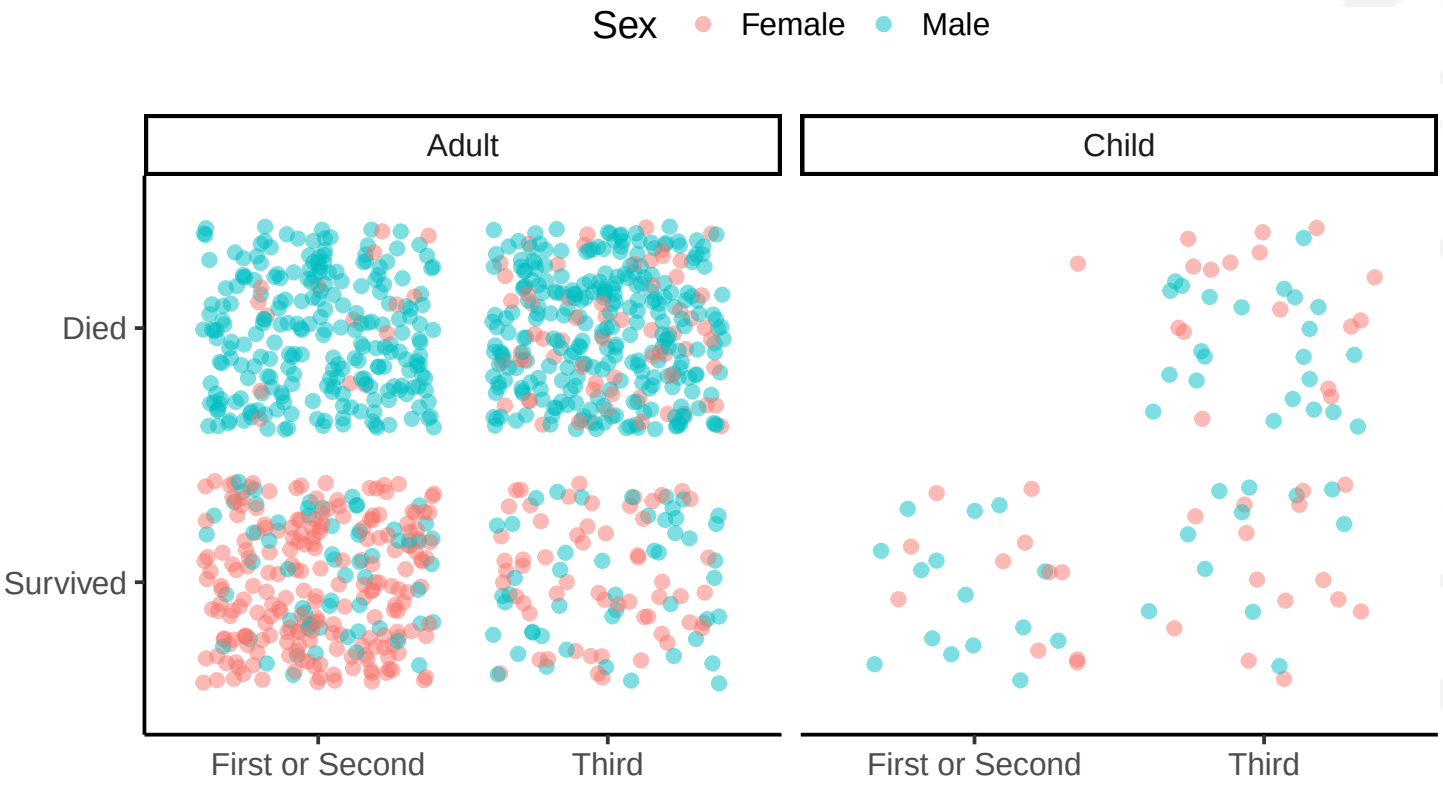
Example– Surviving the Titanic

- **Adjusting for age:** The odds ratio for males versus females of dying during the sinking of the Titanic is very similar with or without adjustment for age.
- **Interaction with age:** There is strong evidence of an interaction between age and sex in the odds of dying during the sinking of the Titanic. While the odds of dying are much higher for males than females among adults, the odds do not vary much by sex among children.



Example– Surviving the Titanic

You could continue expanding the regression model. For example, do you think that the odds ratio for males versus females should be adjusted for ticket class? Do you think there might be an interaction between sex and ticket class? What about age and ticket class?



Generalized linear models

Different forms of GLMs are distinguished by (1) the link function and (2) the distribution of the outcome.

Model	Example outcome	Link	Outcome distribution
Linear	Continuous	Identity: $E(Y)$	Normal
Logistic	Binary	Logit: $log(\frac{E(Y)}{1-E(Y)})$	Binomial
Poisson	Count	Log: $log(E(Y))$	Poisson
Log-binomial	Binary	Log: $log(E(Y))$	Binomial
Additive risk	Binary	Identity: $E(Y)$	Binomial

Generalized linear models– link functions

Here are how the systematic part of a GLM looks for different link functions:

Identity link:

$$E([Y_i]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

Log link:

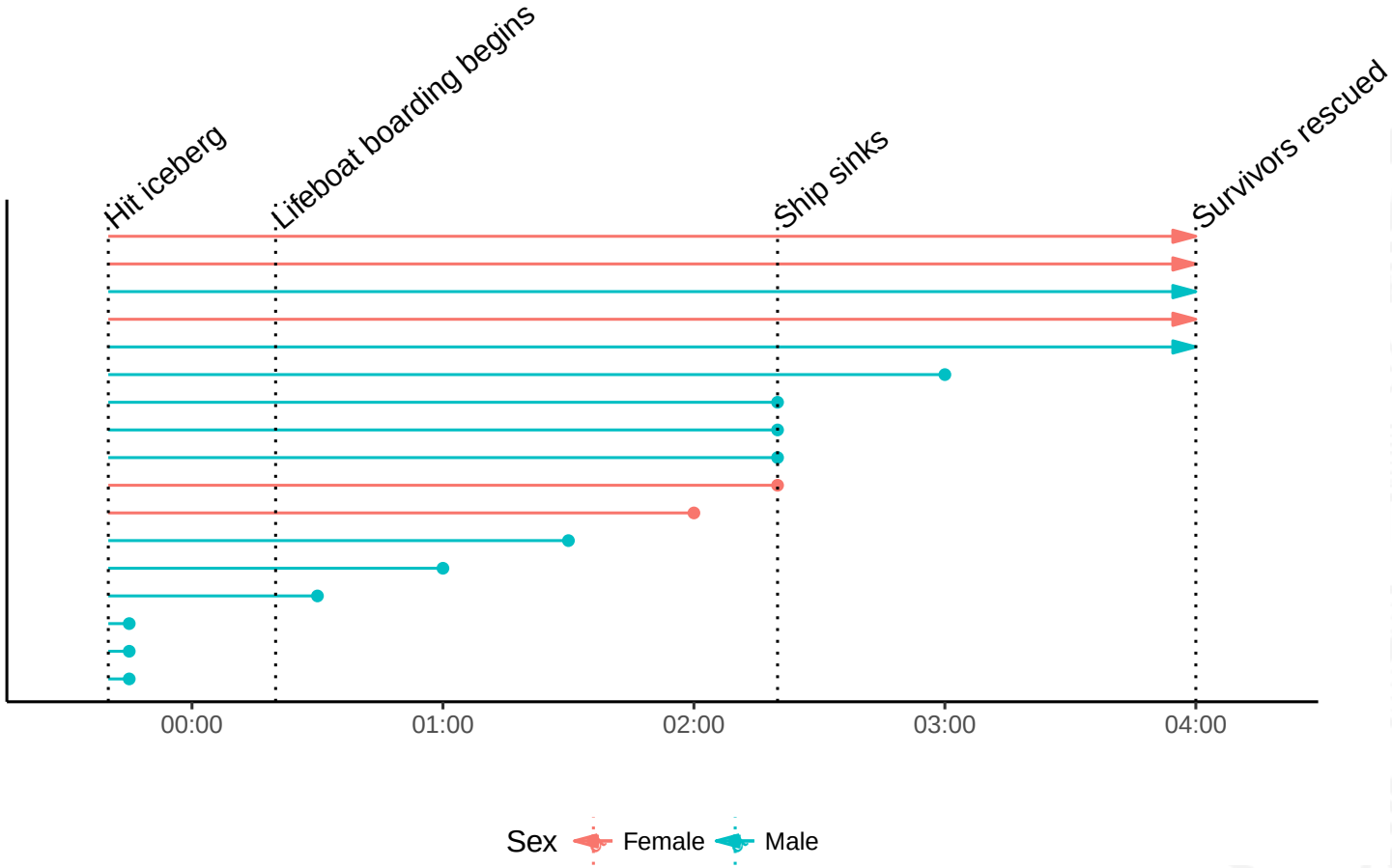
$$log(E[Y_i]) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

Logit link:

$$log\left(\frac{E[Y_i]}{1 - E[Y_i]}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ...$$

Survival analysis

What if we were interested in how long people survived instead of whether they survived?

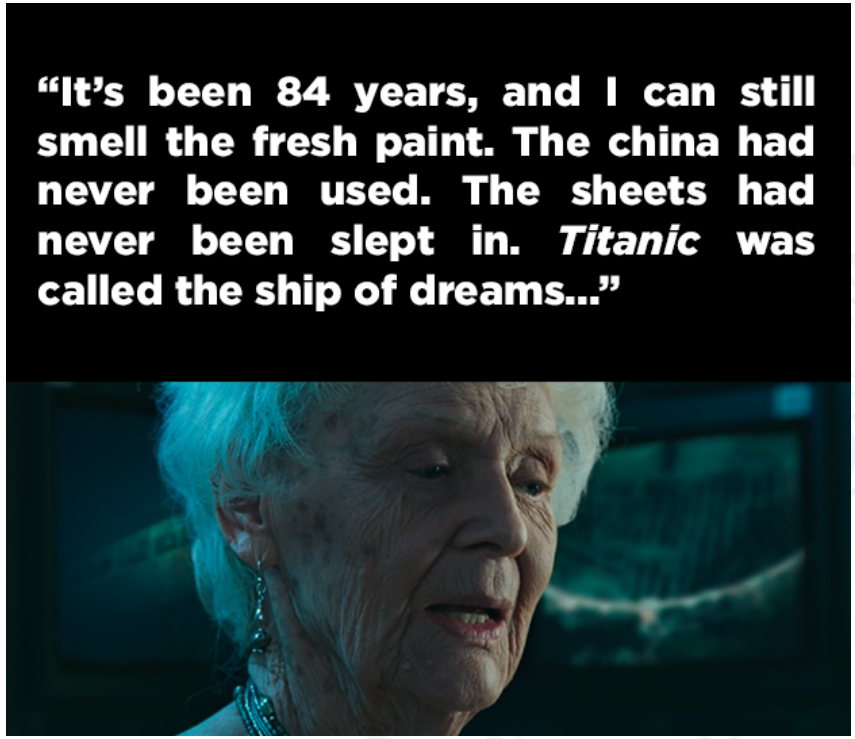


Each line shows a person on the Titanic. The lines with arrows are people who survived. The lines with points show people who died, with the point at the time they died.

Survival analysis

Characteristics of survival analysis:

- Outcome is no longer binary (survived / died)
- Outcome is survival time (time to event)
- Right-censored data– time to event is longer than follow-up time and so unobserved



Survival analysis

For a survival analysis, you can model the log hazard ratio as a linear function of predictors using a proportional hazard model:

$$log[HR(\mathbf{x_i})] = log\frac{h(t|\mathbf{x_i})}{h_0(t)} = \beta_1 X_{1,i} + \beta_2 X_{2,i} + ...$$

where:

- $h_0(t)$: The baseline hazard at time t
- $h(t|\mathbf{x_i})$: The hazard at time t given characteristics $\mathbf{x_i}$
- $HR(\mathbf{x_i})$: The hazard ratio for person i

This model assumes *proportional hazards*. Cox proportional hazard model is a popular semi-parameteric model to fit.

Survival analysis

In the Titanic example, if we were interested in how survival time is associated with sex, we could fit the following Cox proportional hazards model:

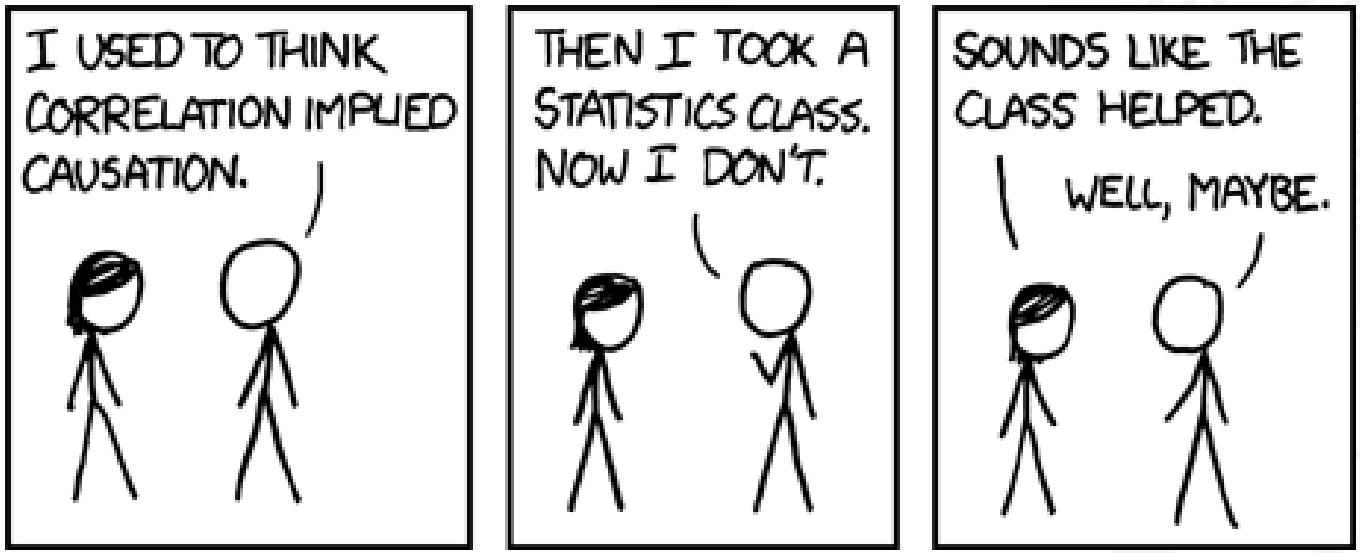
$$log[HR(\mathbf{x_i})] = log\frac{h(t|X_{1,i})}{h_0(t)} = \beta_1 X_i$$

where:

$$X_i = \begin{cases} 0 \text{ if person i is female} \\ 1 \text{ if person i is male} \end{cases}$$

Association versus causation

An important caveat of all these models is that all they guarantee to estimate is association, not causation. There are ways to use regression modeling as a tool in causal inference, but using a regression model does not guarantee a causal interpretation.



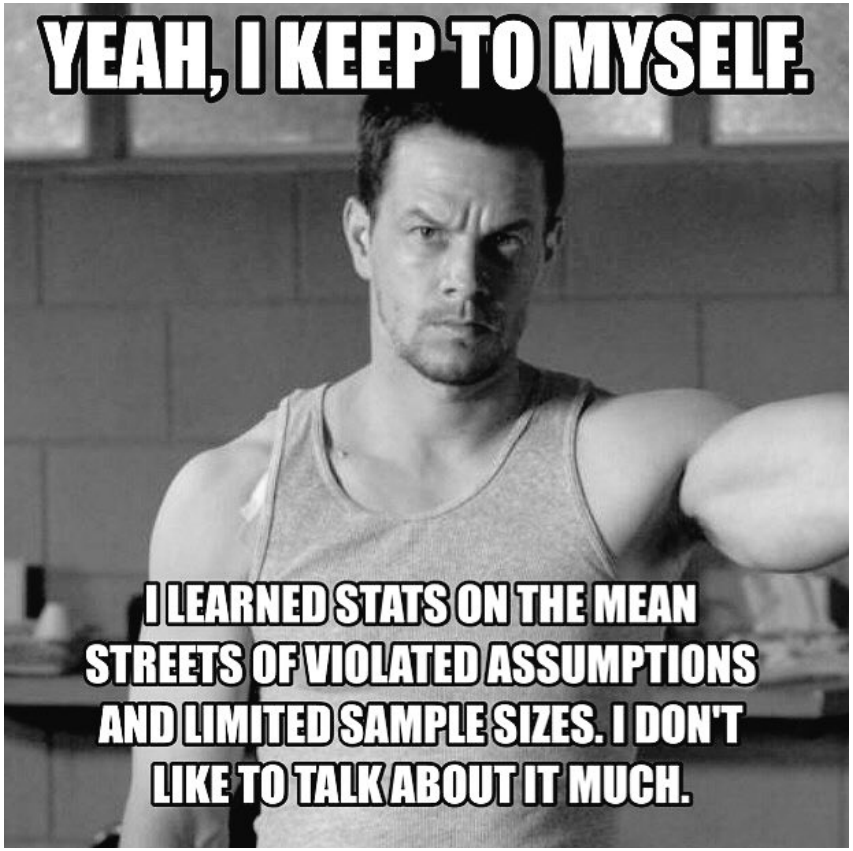
Source: `www.xkcd.com`

Multicollinearity

A second caveat is that you need to careful of including multiple predictors that are strongly correlated. If not, you will run into problems from **multicollinearity**– the regression model will struggle to separate estimated coefficients between the correlated variables.

- For the Titanic example, a model that included ticket class and ticket price might suffer from multicollinearity.
- Implications: (1) instability of coefficient estimates and (2) large standard errors for coefficient estimates.

Model assumptions



Assumptions of GLMs:

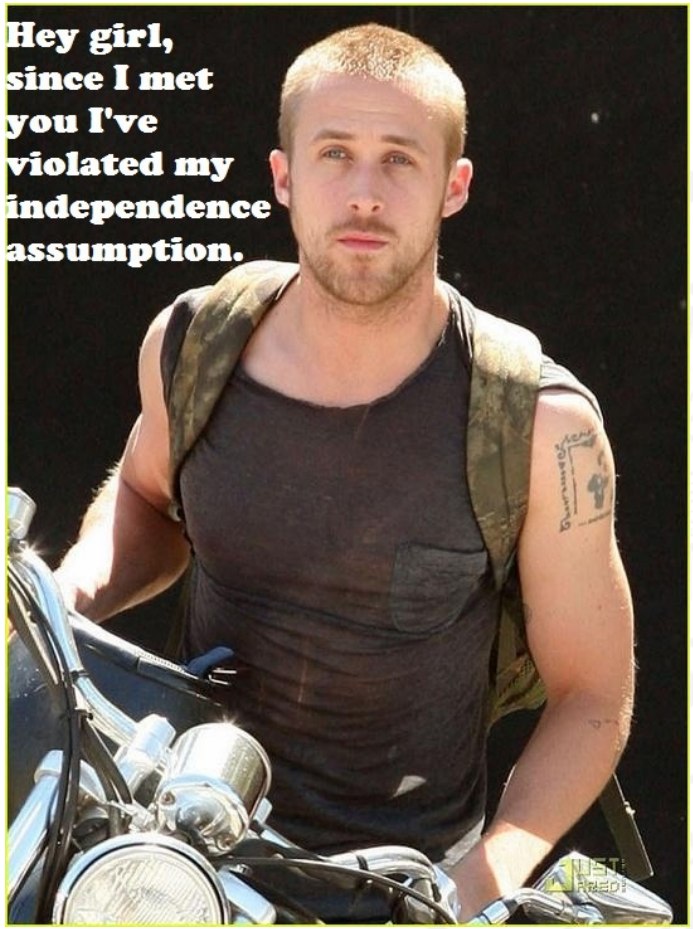
- Independence assumption (observations are independently distributed)
- Outcome follows the specified distribution for a fixed set of covariates
- For continuous predictors, relationship with $g(E[Y_i])$ is linear

Non-independent outcomes

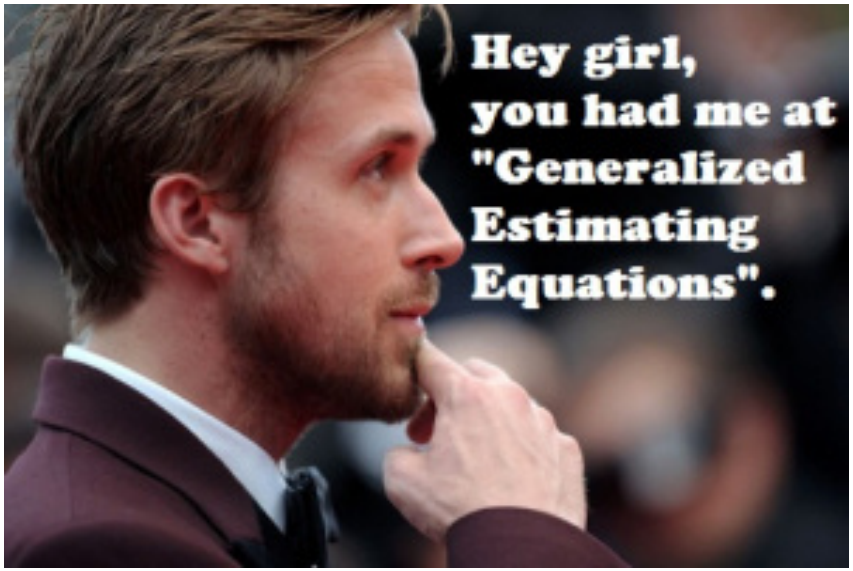
In epidemiological studies, the independence assumption is often violated. Examples of when the independence assumption might be violated include:

- Repeated measures / longitudinal data
- Hierarchical / clustered data (families, schools, multi-site clinical trials)

In the Titanic example, the independence assumption might be violated because many of the passengers were traveling as families, and survival outcomes might be more similar within families than between families.



Non-independent outcomes



There are a number of ways to model data in which the independence assumption is violated. A popular one in epidemiological studies are **Generalized Estimating Equations** (GEEs).

- GEEs accomodate correlated observations.
- You must specify which variables indicate clustering (e.g., family in the Titanic example).
- You must also specify a "working correlation structure" (e.g., exchangeable correlation structure; autoregressive correlation structure).

Non-independent outcomes

For more on GEEs in epidemiologic studies:

- Hanley et al. (2003) Statistical analysis of correlated data using generalized estimating equations: an orientation. *American Journal of Epidemiology.*
- Hubbard et al. (2009) To GEE or not to GEE: comparing estimating function and likelihood-based methods for estimating the associations between neighborhoods and health. *Epidemiology*.