Introduction to Statistical Concepts

William Astle
with thanks to Angela Goncalves

Useful Statistical Concepts

- This course contains a lot of material so we are going to
  - assume some background knowledge and
  - cover a lot of topics but rather superficially
- We will cover:
  - statistical terminology
  - tests for independence in contingency tables
  - linear regression
  - logistic regression
  - Poisson regression
- Examples and exercises in R

Statistics and Statistical Terminology and Modelling

What is Statistics?

- ***Stat[e]****istics* - originally conceived as *the science of the state* - the collection and analysis of facts about a country
- A modern definition: *statistics* is a set of methods for reasoning when there is uncertainty
- It can be thought of loosely as a generalisation of *logic*
- *Logic* is the study of methods for reasoning from statements which are definitely known to be true or false

What is Statistics?

- ***Stat[e]****istics* - originally conceived as *the science of the state* - the collection and analysis of facts about a country
- A modern definition: *statistics* is a set of methods for reasoning when there is uncertainty
- It can be thought of loosely as a generalisation of *logic*
- *Logic* is the study of methods for reasoning from statements which are definitely known to be true or false
- *to reason* **defn**. to think, understand, and form judgements logically.

What is Statistics?

An example of *logical* reasoning:

- Bananas are not spherical
- Apples are coloured red
- I take a fruit from a bowl containing apples, oranges and bananas
- The fruit is 1) spherical and 2) not coloured orange
- Therefore the fruit must be an apple

What is Statistics?

An example of *logical* reasoning:

- Bananas are not spherical
- Apples are coloured red
- I take a fruit from a bowl containing apples, oranges and bananas
- The fruit is 1) spherical and 2) not coloured orange
- Therefore the fruit must be an apple

- This is a logical inference
- No uncertainty to worry about

What is Statistics?

An example of *statistical* reasoning:

- I take a fruit from a bowl containing 3 bananas, 4 apples and 5 oranges.
- The fruit is spherical
- What is the probability that the fruit is an apple?
- $4/(4+5) = 4/9$

What is Statistics?

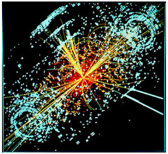
An example of *statistical* reasoning:

- I take a fruit from a bowl containing 3 bananas, 4 apples and 5 oranges.
- The fruit is spherical
- What is the probability that the fruit is an apple?
- $4/(4+5) = 4/9$

- We have observed some data (knowledge that the fruit is spherical) and have drawn a *statistical inference*
- Statistical inferences summarise uncertainty

Who uses statistics?

- Health services, corporations, governments, scientists all need to reason with uncertainty
- e.g. plan health services: *How many new cases of breast cancer will occur in Malta in the next 5 years?*
- e.g. advertisers: *During which TV show is it most profitable to advertise for a new car?*
- e.g. science: (to give a non-Bioinformatics example!) *What is the probability the observed particle decays imply the existence of the Higgs boson*



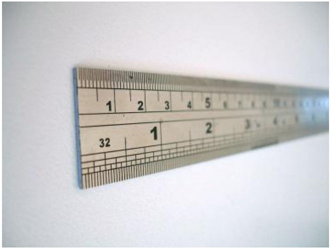
Sources of uncertainty

- Random sampling
- Want to know a fact about a population, but too expensive to ask the question about every member
- e.g. *What is mean age children learn to swim in Malta?*

- Sample 500 children in at random and use the sample average as an uncertain estimate of the population average

Sources of uncertainty

- Measurement error



- e.g. measurement resolution: continuous variables usually **measured** on a discrete scale with a fractional resolution.
- We may measure a person's weight in kg but people do not weigh whole numbers of kg.

Sources of uncertainty

- Complexity - real world phenomena often random
- e.g. time between bus arrivals

Statistical Language

- Scientists frequently study collections of *objects* or *individuals* usually called *study subjects* by statisticians
- Typical aims:
  1. identify qualitative or quantitative *relationships* between measured properties of the study subjects
  2. *summarise uncertainty* about these relationships
  3. make *predictions* about the properties for unobserved individuals

Study Subjects and Variables

- Some examples

Study Subjects	**Properties to be related**
British doctors cohort	smoking, death with lung cancer
mice	genotypes, coat colour
UK Biobank cohort	age, blood haemoglobin level
cancer drugs	molecular structure, mean 5-year survival rate
schools	examination results

Subjects and Variables

- A *variable* is a property of a study subject
- Variables can be
  - observed (i.e. measured, possibly with an associated error)
  - or unobserved, latent or random
- Variables can be categorical or numerical
- Numerical variables can be continuous ‘real numbers’ (e.g. 10.71, 8.23), or discrete counts (e.g. 0, 1, 120)

Subjects and Variables

- Some examples

Definition	**Type**
the sex of person i	Categorical (Female/Male)
the weight of mouse m	Continuous (e.g. measured in kg)
the number brain cells of person i	Count

Random Variables

- A *random variable* is a variable with an uncertain value
- e.g. Define Y by, $Y = 1$ *if the Queen of England dies with lung cancer,* $Y = 0$ *otherwise*
- The value of Y maybe 0 or it maybe 1. Until the Queen dies we will not know which
- We can however estimate Y from measured data
- e.g. The Queen has never smoked but her father died of cancer

Data

- Measurements of variables generate *data*.
- A dataset is usually composed of measurements of multiple variables on many study subjects
- Convention:
  - p denotes the number of variables in a dataset
  - n denotes the number of study subjects
- Data are used to draw inferences about the relationships between variables using statistical models

Models

- A model is a rule for describing a relationship between variables

- Deterministic models are very common in physics e.g.

  $$E = (\Delta m)c^2$$

  describes the relationship between E, the energy emitted when an atomic nucleus transmutes and Δm the change in the mass of the nucleus. The relationship depends on the constant c the speed of light

- This is a deterministic model. If we know Δm precisely we can calculate the energy released E exactly

Statistical Models

- *Statistical* or probabilistic models are alternatives to deterministic models which are used to describe relationships between variables when one or more of the variables is *random*.

- Statistical models are more common in biology and medicine than physics because biological mechanisms are often complex and uncertain and measurements often noisy.

Statistical Model: Example

- This statistical model describes the relationship between:
    - 2 deterministic variables: sex and smoking status
    - a random variable: $Y = 1$ if individual dies with lung cancer, $Y = 0$ if individual dies without lung cancer.
    - The uncertainty in Y is presented as a percentage (a *probability*).

Smoking Status	**Sex**	$\mathbf{Y = 1}$
Smoker	Male	20%
Smoker	Female	10%
Non-smoker	Male	1%
Non-smoker	Female	1%

Events, their Probabilities and Dependencies

Events

- Statisticians often need to model the occurrence of events
- Example of an event: Mrs. Smith had a myocardial infarction between 1/1/2000 and 31/12/2009.
- The occurrence of an event is a binary (dichotomous) variable. There are two possibilities: the event occurs or it does not occur.
- Event occurrence variables can always be coded with $0, 1$ e.g.

  $Y_i = 1 \iff$ person i became pregnant in 2011.
  $Y_i = 0 \iff$ person i did not become pregnant in 2011.

Probability, Odds and log-Odds

- There are many equivalent ways of measuring the plausability of an event.
- We will use three:
  1. probability of the event
  2. odds in favour of the event
  3. log-odds in favour of the event
- These are equivalent in the sense that if you know the value of one measure for an event you can compute the value of the other two measures for the same event

  cf. measuring a distance in kilometres, statute miles or nautical miles

The Probability of an Event

- This is a number π between 0 and 1. We write

  $$\pi = \mathbb{P}(Y = 1)$$

  to mean π is the probability that $Y = 1$.

- $\pi = 1$ means we know the event is certain to occur.

- $\pi = 0$ means we know the event is certain **not** to occur.

- Values between 0 and 1 represent intermediate states of certainty, ordered monotonically.

- Because we are certain one of $Y = 1$ and $Y = 0$ is true and because they cannot be true simultaneously:

  $$\mathbb{P}(Y = 0) = 1 - \mathbb{P}(Y = 1) = 1 - \pi.$$

Odds in Favour of an Event

- The odds in favour of an event is defined as the probability the event occurs divided by the probability the event does not occur.

- The odds in favour of $Y = 1$ is defined as:

$$\mathrm{ODDS}(Y = 1) = \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y \neq 1)} = \frac{\mathbb{P}(Y = 1)}{\mathbb{P}(Y = 0)} = \frac{\pi}{1 - \pi}.$$

- Note:

$$\mathrm{ODDS}(Y = 0) = \frac{1}{\mathrm{ODDS}(Y = 1)} = \frac{1 - \pi}{\pi}.$$

so

$$\mathrm{ODDS}(Y = 1) \times \mathrm{ODDS}(Y = 0) = 1.$$

Interpreting the Odds in Favour of an Event

- An odds is a number between 0 and ∞.
- An odds of 0 means we are certain the event does **not** occur.
- An increased odds corresponds to increased belief in the occurrence of the event.
- An odds of 1 corresponds to a probability of 1/2.
- An odds of ∞ corresponds to certainty the event occurs.

Log-odds in Favour of an Event

- The log odds in favour of an event is defined as the log of the odds in favour of the event:

$$\log \mathrm{ODDS}(Y=1) = \log \frac{\mathbb{P}(Y=1)}{\mathbb{P}(Y=0)} = \log \frac{\pi}{1-\pi}.$$

- Note

$$\log \mathrm{ODDS}(Y=1) = -\log \mathrm{ODDS}(Y=0) = \log \frac{1-\pi}{\pi}$$

Interpreting the Log-odds in Favour of an Event

- A log-odds is a number between $-\infty$ and ∞.
- A log odds of $-\infty$ means we are certain the event does **not** occur.
- An increased log-odds corresponds to increased belief in the occurrence of the event.
- A log-odds of 0 corresponds to a probability of $1/2$.
- A log-odds of ∞ corresponds to certainty the event occurs.

Moving between Probability, Odds and Log-odds

- You can use the following table to compute one measure of probability from another:

	$\mathbb{P}$	ODDS	$\log$ ODDS
$\mathbb{P}(Y = 1) = \pi$	■	$\frac{\pi}{1-\pi}$	$\log \frac{\pi}{1-\pi}$
$\text{ODDS}(Y = 1) = o$	$\frac{o}{1+o}$	■	$\log o$
$\log \text{ODDS}(Y = 1) = x$	$\frac{e^x}{1+e^x}$	e^x	■

- Choose the row corresponding to the quantity you start with and the column corresponding to the quantity you want to compute.
- $\log \frac{\pi}{1-\pi}$ is often written $\text{logit}(\pi)$.
- $\frac{\exp(x)}{1+\exp(x)}$ is often written $\text{inv.logit}(x)$ (sometimes $\text{expit}(x)$).

Dependency Between Events

- Sometimes we are interested in understanding the dependency between events
- e.g.
  - Event A = Study subject no. 12 has measured genotype GG at rs2383206
  - Event B = Study subject no. 12 had a heart attack between 2000 and 2009
- Cross-classify into four probabilities:

	Event A occured	Event A did not occur
Event B occured	π_{AB}	$\pi_{\bar{A}B}$
Event B did not occur	$\pi_{A\bar{B}}$	$\pi_{\bar{A}\bar{B}}$

Measuring Dependency Between Events

- Dependency between events is measured using *odds ratios*

	A occured	A did not occur	
B occured	π_{AB}	$\pi_{\bar{A}B}$	$\pi_B (\equiv \pi_{AB} + \pi_{\bar{A}B})$
B did not occur	$\pi_{A\bar{B}}$	$\pi_{\bar{A}\bar{B}}$	$1 - \pi_B$
	$\pi_A (\equiv \pi_{AB} + \pi_{A\bar{B}})$	$1 - \pi_A$	

- Two possible odds ratios:
  - odds in favour of A when B is true divided by the odds in favour of A when B is false
  - odds in favour of B when A is true divided by the odds in favour of B when A is false
- It happens both are equal:

$$\mathrm{OR}(A,B) \equiv \frac{\mathrm{ODDS}(A|B)}{\mathrm{ODDS}(A|\bar{B})} = \frac{\mathrm{ODDS}(B|A)}{\mathrm{ODDS}(B|\bar{A})} = \frac{\pi_{AB}\pi_{\bar{A}\bar{B}}}{\pi_{\bar{A}B}\pi_{A\bar{B}}}$$

Properties of Odds Ratios and log Odds Ratios

- The log odds ratio is defined as the log of the odds ratio $\mathrm{LOR}(A, B) \equiv \log \mathrm{OR}(A, B)$
- $\mathrm{OR}(A, B) > 1$, (equivilantly $\mathrm{LOR}(A, B) > 0$) $\implies$ A and B are positively correlated; when A is true B is more likely to be true than when A is false (and vice versa)
- $\mathrm{OR}(A, B) < 1$, (equivilantly $\mathrm{LOR}(A, B) < 0$) $\implies$ A and B are negatively correlated; when A is true B is less likely to be true than when A is false (and vice versa)
- $\mathrm{OR}(A, B) = 1$, (equivilantly $\mathrm{LOR}(A, B) = 0$) $\implies$ A and B are uncorrelated; when A is true B has the same likelihood of being true as when A is false (and vice versa)

Independence of Events

	A occured	A did not occur	
B occured	π_{AB}	$\pi_{\bar{A}B}$	π_B
B did not occur	$\pi_{A\bar{B}}$	$\pi_{\bar{A}\bar{B}}$	$1 - \pi_B$
	π_A	$1 - \pi_A$	

- Events A and B are *independent* if and only if $\pi_{AB} = \pi_A \pi_B$
- When A and B are independent:
  - knowledge of whether or not A has occurred gives you no information about whether or not B be has occurred
  - A and B are uncorrelated, i.e.

$$\mathrm{OR}(A,B) \equiv \frac{\pi_{AB}\pi_{\bar{A}\bar{B}}}{\pi_{\bar{A}B}\pi_{A\bar{B}}} = 1$$

$$\mathrm{LOR}(A,B) \equiv \log(\mathrm{OR}(A,B)) = 0$$

Testing for Departures from Independence

- We can perform statistical tests to determine the strength of evidence against pairs of events being independent
- To perform such a test we need to collect data on multiple instances of the events
- e.g. Suppose we are interested in the relationship between genotypes at the SNP rs2383206 and risk of heart attack
- We collect 1000 cases (heart attack in ten year window) and 1000 controls (no heart attack in ten year window)

Testing for Departures from Independence

	AA rs2383206 genotype	AG rs2383206 genotype	GG rs2383206 genotype
Heart attack 2000-2009	248	436	244
No heart attack 2000-2009	185	436	379

- Two statistical tests i) Pearson's test ii) Fisher's test
- Pearson's test is computationally efficient but is only accurate when each cell has a count of at least 5
- Fisher's test is computationally intensive for tables with very large counts

Testing for Departures from Independence

```
> genotype_counts
     [,1] [,2] [,3]
[1,]  248  436  244
[2,]  185  436  379
> fisher.test(genotype_counts)

      Fisher’s Exact Test for Count Data

data:  genotype_counts
p-value = 1.539e-08
alternative hypothesis: two.sided

> chisq.test(genotype_counts)

      Pearson’s Chi-squared test

data:  genotype_counts
X-squared = 35.781, df = 2, p-value = 1.699e-08
```

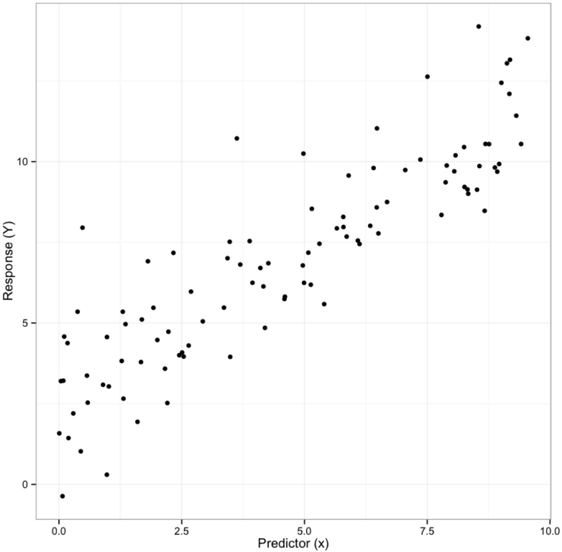
Regression Modelling

Regression Models

- A regression model describes the relationship between the *average* value of a random *response* variable and the value of values of one or more *predictor* variables

- A regression is defined by
  1. the random *response* variable
  2. a list of *predictor* variables
  3. a regression *equation*
  4. a *distribution* for the value of the random response variable

Response and Univariate Predictor



The Response Variable

- The *response* (sometimes *outcome* or *dependent*) is random
- The notation Y_i is usually used to indicate the response value of study subject i
- $\mathbf{E}Y_i$ is used to denote the *average* or *expected* value of the response variable for study subject i
- Responses variables can be continuous or categorical (binary or count)
- The word *response* is used by analogy with a treatment-response experiment
- Such an experiment allocates subjects to treatment classes and seeks to identify differences in the distribution of the responses between the treatments

The Response Variable Distribution

- The response is usually modelled as a random variable with a particular parametric form (shape):
- e.g. a Normal distribution:

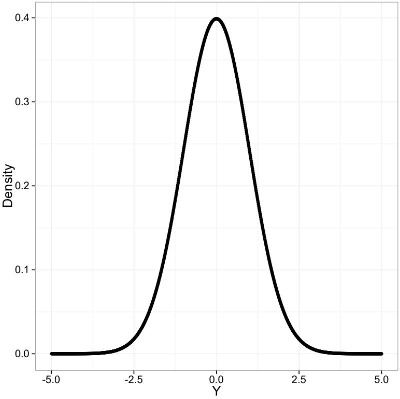
$$Y_i \sim N(\mu_i, \sigma^2)$$

- e.g. a Bernouilli distribution (i.e. a 0/1 distribution):

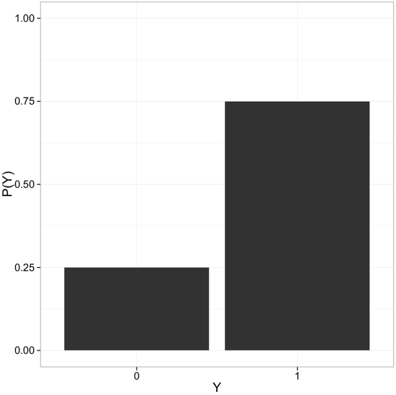
$$\begin{aligned} \mathbb{P}(Y_i = 1) &= \mu_i \\ \mathbb{P}(Y_i = 0) &= 1 - \mu_i \end{aligned}$$

- Note in both these cases we have written $\mu_i = \mathbb{E}Y_i$

Response Distributions



- Normal distribution



- Bernoulli distribution (0/1 distribution)

Predictors

- Predictors are deterministic (non-random) variables
- The aim of regression modelling is to associate a predictor with a response or to associate multiple predictors with the average value of a response variable
- Predictor variables are usually numerical
- Categorical variables can be used as predictors but the categories must be coded numerically

Predictor Notation

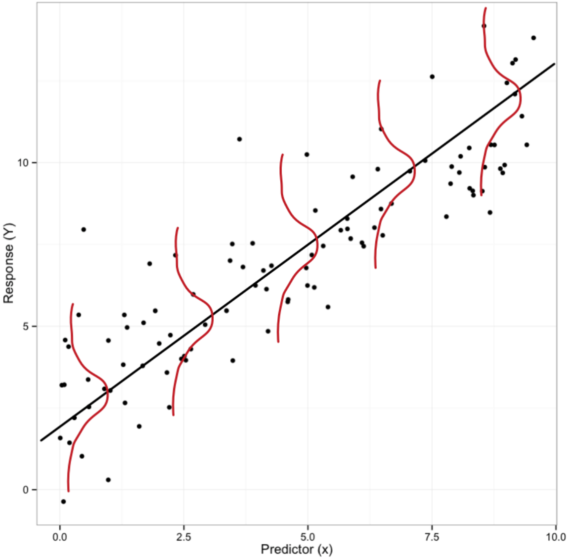
- The letter x is generally used to denote data from predictor variables (although other letters are used, e.g. z is common)
- If a regression model has a single predictor x then x_i is used to denote the value of the predictor measured on study subject i
- If a regression model has multiple predictors, double subscripts are used. x_{ij} denotes the data measured on study subject i for predictor variable j

Regression Equation

- This is the deterministic bit of the regression model
- It describes how the average value of the response varies with the predictor variables
- e.g. a univariate (one predictor) linear regression equation has the form

$$\mathbb{E}Y_i = \mu_i = \alpha + x_i\beta$$

Response and Univariate Predictor



General Multiple Regression Equation

$$g(\mathbb{E}Y_i) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$$

- This general equation can be applied with:
  - a range of probability distributions for the response variable
  - multiple predictor variables

Expected Value of Response

$$g(\mathbb{E}Y_i) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + \ldots + x_{ip}\beta_p$$

- $\mathbb{E}Y_i$ is the *expected value* of the response for subject i
- $\mathbb{E}Y_i$ can be thought of as the mean value of Y in an infinitely large group of hypothetical study subjects who have the same predictor variable measurements as study subject i

Linear Predictor

$$g(\mathbb{E}Y_i) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$$

- The right hand side of the equation is called the *linear predictor*
- α is the intercept
- β_1, ... β_p are the regression coefficients
- The intercept and the regression coefficients are numbers
- α, β_1,... β_p are usually unknown
- The purpose of statistical analysis is to estimate α, β_1,... β_p from data

Link Function

$$g(\mathbb{E}Y_i) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$$

- g is called the *link* function
- g is always a monotonic, strictly increasing function
- This means that an increase in the linear predictor corresponds to an increase in $\mathbb{E}Y_i$

Purpose of the Link Function

$$g(\mathbb{E}Y_i) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$$

- In principle α, β_1,... β_p can each take any value between $-\infty$ and ∞
- Consequently, the linear predictor can be any value between $-\infty$ and ∞
- Sometimes the distribution of Y_i is such that $\mathbb{E}Y_i$ can only take a certain set of values
- e.g. If Y_i is binomial taking values $0/1$ then $0 \leqslant \mathbb{E}Y_i \leqslant 1$
- The link function allows us to map the set of possible values of $g(\mathbb{E}Y_i)$ to the whole number line

Intercept

$$g(\mathbb{E}Y_i) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$$

- α is the intercept term
- α represents the value of $g(\mathbf{E}Y_i)$ taken by a hypothetical study subject i which has predictor variables all equal to zero.
- i.e. $g(\mathbb{E}Y_i) = \alpha$, $x_{ij} = 0$ for all j.

Regression Coefficients

$$g(\mathbb{E}Y_i) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$$

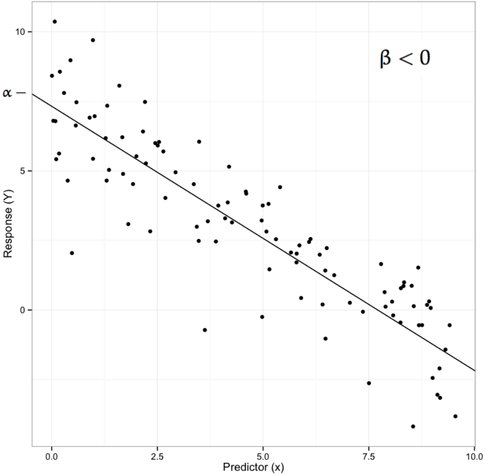
- The regression coefficient β_j defines the relationship between the jth predictor variable and the response variable

- If β_j is equal to zero then a change in the value x_{ij} has no effect on the distribution of the response

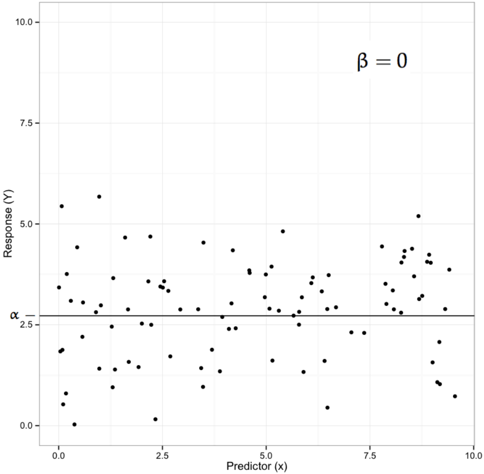
- If $\beta_j > 0$ then an increase in the value of the x_{ij} increases the average value of the response $\mathbb{E}Y_i$

- If $\beta_j < 0$ than in increase in the value of x_{ij} decreases the average value of the response $\mathbb{E}Y_i$

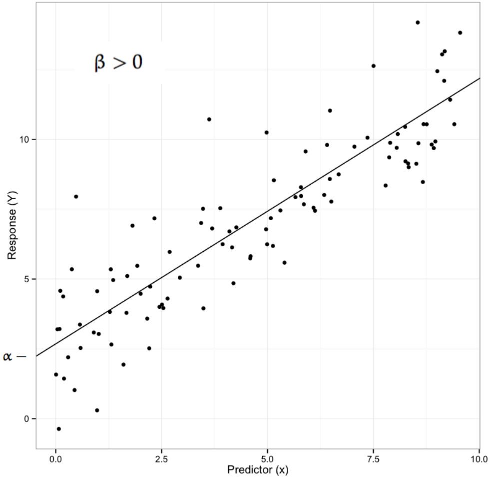
Regression Coefficients



Regression Coefficients



Regression Coefficients



Why Use Regression?

- There are two reasons for doing regression modelling:
  1. Inference of relationships between variables
  2. Prediction of response values in new subjects with given predictor values

- We address both these questions by *fitting the model* to data

- When we fit the model we can draw inferences about relationships by:
  1. Obtaining point estimates of the regression coefficients
  2. Quantifying our uncertainty about the regression coefficients

- Point estimates of the regression coefficients can then be used to predict the response in new study subjects.

Estimating Regression Coefficients

- There are a number of methods for estimating regression coefficients from a dataset of measured values for the response and predictor variables.
- We will touch briefly on the most widely used method, *maximum likelihood estimation* although the details are not terribly important in practice
- Maximum likelihood estimation is the standard method implemented in most widely used statistical software
- Other methods include *methods of moments estimation* and *Bayesian estimation* neither of which we will consider

Likelihood

- Given a regression model and a dataset we can write down the *likelihood function*

- The likelihood function is a multivariate function which assigns a a number to each possible value of the regression coefficients

$$L(\alpha, \beta_1, \beta_2, ...\beta_p) = \prod_i \mathbb{P}(Y_i|\alpha, \beta_1, \beta_2, ...\beta_p)$$

- It is calculated by multiplying the probability of each observed response value at the desired values of the parameters

Maximum Likelihood Estimation

- The maximum likelihood estimate (MLE) of the regression coefficients is the set of values for the regression coefficients for which the likelihood is largest

- The MLE is usually denoted using a hat symbol. $\hat{\beta}_1$ is the maximum likelihood estimate of the regression coefficient corresponding to predictor variable 1.

$$L(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2, ...\hat{\beta}_p) = \max_{\alpha, \beta_1, \beta_2, ... \beta_p} \prod_i \mathbb{P}(Y_i | \alpha, \beta_1, \beta_2, ...\beta_p)$$

- Intuitively the MLE of the regression coefficients is the value of the regression coefficients which makes the observed data most probable

Uncertainty

- The MLE gives us a *point estimate* for regression coefficients
- However, estimates are almost never correct
- To draw an inference about the relationship between a predictor and the response, we usually want to say something about our *uncertainty* about the corresponding regression coefficient
- One method of summarising uncertainty is to quote a *confidence interval*

Confidence Intervals

- A confidence interval is a pair of numbers L (the lower limit) and U (the upper limit) together with an associated *confidence level*.
- The confidence level is quoted as a percentage (normally 95% is used)
- Given a confidence level of $\gamma\%$ a lower $L(\gamma\%)$ and an upper $U(\gamma\%)$ limit can calculated from the observed data
- There are many methods for calculating confidence intervals. We will not go into the details
- However, most methods of calculating a confidence interval rely on the likelihood function

Interpretation of Confidence Intervals

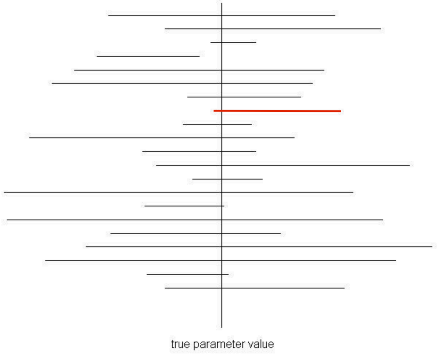
- The interpretation of confidence intervals can be counter-intuitive at first
- Uncertainty is quantified using the idea of imaginary replicate experiments
- Suppose, in an imaginary world in which time and money are no object, we:
  1. repeat our experiment very many times
  2. generate a new dataset on each occasion
  3. calculate a new $\gamma\%$ level confidence interval for the coefficient β using each dataset

  then approximatelp $\gamma\%$ of those confidence intervals should contain the true value of the parameter. i.e.

  $$L(\gamma\%) < \beta < U(\gamma\%)$$

  in $\gamma\%$ of the imaginary replicates

Interpretation of Confidence Intervals



- Red line is the interval calculated from the actual dataset
- Black lines are the imaginary intervals from repeat experiments
- 95% of lines cross true parameter value

Response Prediction with Regression Models

- Once we have obtained estimates of regression coefficients we can predict the value of the response for a new study subject i with known predictor values, using the regression equation

$$g(\mathbb{E}Y_i) = \alpha + x_{i1}\beta_1 + x_{i2}\beta_2 + ... + x_{ip}\beta_p$$

1. Denote the predicted value by $\hat{Y}_i$. Plug in the maximum likelihood estimates of the coefficients:

$$g(\hat{Y}_i) = \hat{\alpha} + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + ... + x_{ip}\hat{\beta}_p$$

2. Invert the link function:

$$\hat{Y}_i = g^{-1}(\hat{\alpha} + x_{i1}\hat{\beta}_1 + x_{i2}\hat{\beta}_2 + ... + x_{ip}\hat{\beta}_p)$$

Linear Regression

Simple Linear Regression

- The regression equation for simple linear regression is:

$$\mathbb{E}Y_i = \mu_i = \alpha + \beta \times x_i$$

- Note that the link function g is the identity function for linear regression.
- The assumption here is that the relationship between x and $\mathbb{E}Y_i$ is a straight line
- The *slope* of the line is β

Interpretation of α

- To interpret α put $x_i = 0$ into the regression equation:

  $$\mathbb{E}Y_i = \alpha + \beta \times x_i$$

  then

  $$\mathbb{E}Y_i = \alpha$$

- α is the average value of the response variable amongst study subjects for which the predictor variable is zero.

Interpretation of β

- To interpret β put $x = z$ and $x = z + 1$ for study subjects i and i' into the regression equation to obtain:

$$\mathbb{E}Y_i = \alpha + \beta \times z \tag{1}$$

$$\mathbb{E}Y_{i'} = \alpha + \beta \times (z + 1) \tag{2}$$

  then take equation (1) from equation (2)

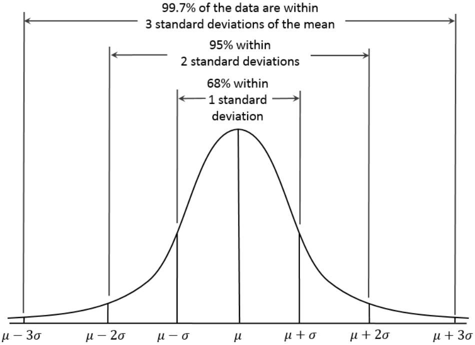
$$\mathbb{E}Y_{i'} - \mathbb{E}Y_i = \beta \tag{3}$$

- β is the difference in the average value of the response variable between groups of study subjects for which the predictor variable differs by one unit.

Linear Regression Response

- For linear regression the response distribution is assumed to be *normal* (sometimes called Gaussian).

$$Y_i \sim N(\mu_i, \sigma^2) \quad \text{equivalently} \quad Y_i \sim N(\alpha + \beta \times x_i, \sigma^2)$$



Linear Regression Errors

- The quantity

  $$\begin{aligned} \epsilon_i &= Y_i - \mu_i \\ &= Y_i - (\alpha + \beta \times x_i) \end{aligned}$$

  is the *error* corresponding to study subject i

- The distributional assumption of linear regression is equivalent to the assumption that the errors are normally distributed, with mean zero:

  $$\epsilon_i \sim N(0, \sigma^2)$$

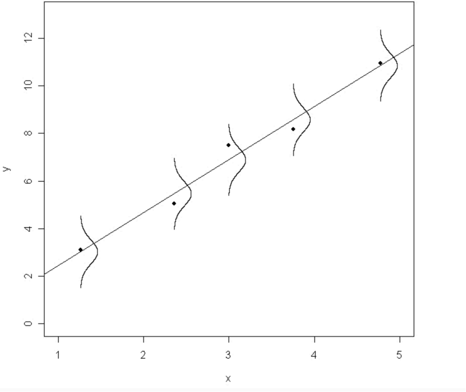
- The error variance σ^2 is the same for each study sample

- σ^2 can be estimated from the data using maximum likelihood

Linear Regression Error Assumption

- We can put the regression equation and distribution assumption into a single statement:

$$Y_i = \alpha + \beta \times x_i + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2)$$



Linear Regression Residuals

- We define residual for study subject *i* by:

  $$r_i = Y_i - (\hat{\alpha} + \hat{\beta} \times x_i)$$

  Recall that the error for study subject *i* is defined by

  $$\epsilon_i = Y_i - (\alpha + \beta \times x_i)$$

- Note that residuals and errors are **not** the same.
- Errors are unknown because we don't know α and β
- Residuals can be computed from the data
- Residuals can be thought of as estimates of errors

Properties of Linear Regression Residuals

- Although residuals and errors are not the same, residuals have similar properties to errors:
  1. The mean (and sum) of the residuals for a study sample is equal to zero
  2. Residuals are normally distributed
  3. The variance of the residuals should not depend on the value of the predictors
- The first property holds regardless of the validity of the modelling assumptions
- The second and third properties only holds if the model assumptions are valid. Specifically only if
  1. The relationship between x and Y is linear
  2. The *errors* are normally distributed
  3. The variance of the errors is constant

Checking Modelling Assumptions

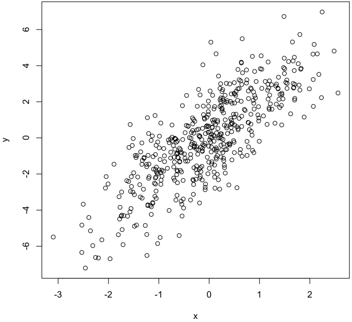
- Before we rely on an inference made from a linear regression model, we should always verify that the modelling assumptions hold
- Specifically we should check
  1. $\mathbb{E}Y$ is a linear function of x
  2. The properties of the residuals are consistent with the assumption about the distribution of Y

Check Linearity

- Suppose the R variable `y` is a vector containing data from a response variable Y and the R variable `x` is a vector containing data from a predictor variable x.

- We can generate a plot of y against x with the command

```
>  plot(x,y)
```



Fitting a Linear Model in R

- We can fit a linear regression in R using the `lm` function.

  ```
  >  fit.obj = lm(y~x)
  ```

- Fits the regression equation

  $$\mathbb{E}Y = \alpha + \beta \times x$$

- The result of the model fit is stored in the R object `fit.obj`

Extracting the Residuals

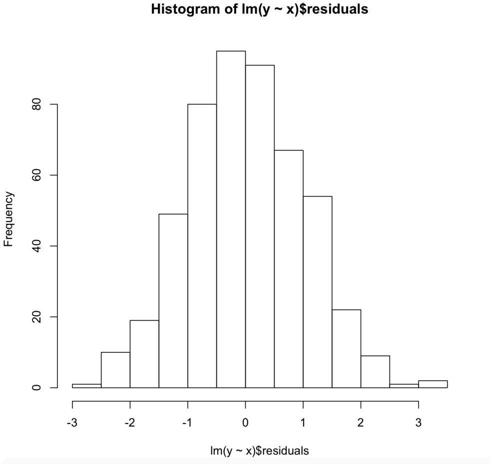
- The residuals can be extracted from the linear regression object using `fit.obj$residuals`
- For example to draw a histogram of the residuals you can type:

  ```
  >  hist(fit.obj$residuals)
  ```

- Alternatively you can do the fitting and plotting in one statement, without storing a model object:

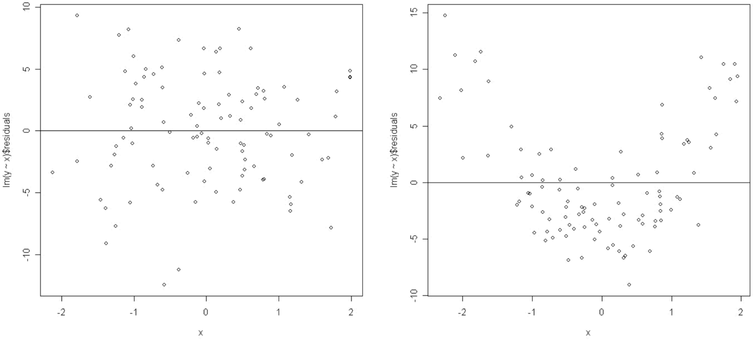
  ```
  >  hist(lm(y~x)$residuals)
  ```

Histogram of the Residuals



- By examining a histogram of the residuals we can check the normality assumption holds

Plot the Residuals vs. the Predictor



- Plot the residuals against the predictor variable to verify that the distribution of the residuals is independent of x

```
>  plot(x, lm(y~x)$residuals)
```

Maximum Likelihood Estimation

- For linear regression there are formulae for the maximum likelihood estimates of the regression coefficients:

$$\hat{\beta} = \frac{\sum_i^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_i^n (x_i - \bar{x})^2}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

- However we do not need to worry about these too much as R will do the calculations for us

Viewing Model Fit Information in R

- The simplest way to view model fit information in R is to type the name of a fitted model object and hit return:

```
> fit.obj = lm(y~x)
> fit.obj

Call:
lm(formula = y ~ x)

Coefficients:
(Intercept)            x
     0.9803       1.9794
```

- This prints the MLEs of the coefficients

Printing Confidence Intervals Using R

- 95% confidence limits can be computed with the `confint` function

```
> confint(fit.obj)
                2.5 %   97.5 %
(Intercept) 0.9349681 1.025687
x           1.9343508 2.024393
```

- A different confidence level can be specified if desired, e.g. 99%:

```
> confint(fit.obj, level=0.99)
                0.5 %   99.5 %
(Intercept) 0.9206313 1.040023
x           1.9201208 2.038623
```

The `display` command

```
> display(fit.obj)
lm(formula = y ~ x)
            coef.est coef.se
(Intercept) 0.98     0.023
x           1.98     0.013
---
n = 500, k = 2
residual sd = 0.51, R-Squared = 0.94
```

- Prints: the MLE, the standard errors of the coefficient MLEs, the standard deviations of the regression coefficients, the residual standard deviation and R^2

Standard Errors of the MLEs

- Back to the idea of imaginary repeated experiments
- Suppose, in an imaginary world we:
  1. repeat our experiment very many times
  2. generate a new dataset on each occasion
  3. estimate a new MLE $\hat{\beta}$ using each dataset
- The MLE is a random variable under this replication process
- The standard error of $\hat{\beta}$ denoted $SE(\hat{\beta})$ is defined as the standard deviation of the MLE.

Proportion of Variance Explained

- R^2 is the proportion of the variance in the response which is explained by the predictor.
- R^2 is a number between 0 and 1
- R^2 is a measure of the correlation between `x` and `y`.
- When $R^2 = 1$, `x` is perfectly correlated with `y` and the residuals are all equal to 0
- When $R^2 = 0$, `x` contains no information about `y`.

Residual Standard Deviation

- The residual standard deviation is what it says on the tin:

$$sd(\hat{\epsilon}) = \sqrt{\frac{1}{n}\sum_{i}^{n}(\epsilon_i - \bar{\epsilon})^2}$$

The R `summary` Command

```
> summary(fit.obj)

Call:
lm(formula = y ~ x)

Residuals:
     Min       1Q   Median       3Q      Max
-1.39718 -0.35082 -0.00092  0.31271  1.60025

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98033    0.02309   42.46   <2e-16 ***
x            1.97937    0.01291   86.38   <2e-16 ***
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1

Residual standard error: 0.5149 on 498 degrees of freedom
Multiple R-squared:  0.9374,^^IAdjusted R-squared:  0.9373
F-statistic:  7462 on 1 and 498 DF,  p-value: < 2.2e-16
```

p-values

- The *p*-value in the `Pr(>|t|)` column of the `summary` command is a measure of the weight of evidence against the *null* hypothesis that the regression coefficient in that row is equal to zero.
- The null hypothesis is so called because it refers to the assumed position that there is no association between the predictor and the response.
- Usually the evidence must be strong before a null hypothesis is rejected
- A *p*-value is a number between 0 and 1. The smaller the number the greater the evidence against the null hypothesis. Typically a *p*-value at least as small as 0.05 is required to reject a null hypothesis.

Interpretation of *p*-values

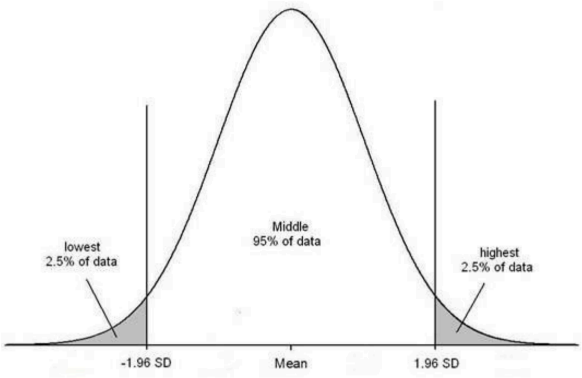
- The interpretation of *p*-values, is based on the idea of imaginary repeated experiments.
- Suppose, in an imaginary world we:
  1. repeat our experiment very many times
  2. generate a new dataset on each occasion
  3. calculate a new *p*-value level using each dataset

  then **assuming the null hypothesis is true** $\alpha \times 100\%$ of the calculated *p*-values should be less than α
- Small *p*-values are rare when the null hypothesis is true

Computing Confidence Intervals Manually

- Although *R* provides the `confint` function, confidence intervals can also be computed manually from standard errors
- Not all statistical software provides functions to compute confidence intervals so this is a useful skill
- Standard errors are listed in the second column of the `summary` output. (They are also printed by the `display` command)
- Manual calculation of confidence intervals is based the assumption that the MLE of the regression coefficient follows a normal distribution.

Computing Confidence Intervals Manually



- We can compute a 95% confidence interval for a regression coefficient using a normal approximation:

$$\hat{\beta} - 1.96 \times SE(\hat{\beta}) < \beta < \hat{\beta} + 1.96 \times SE(\hat{\beta})$$

Multiple Linear Regression

- Multiple linear regression is very similar to simple linear regression
- More than one predictor is now allowed on the right handside of the equation

$$\mathbb{E}Y_i = \mu_i = \alpha + \beta_1 \times x_{i1} + \beta_2 \times x_{i2} + ...\beta_1 \times x_{ip}$$

- The assumptions about the distribution of Y_i (normal, homogeneous variance) are the same as those for simple linear regression.

Fitting a Multiple Linear Regression

- A multiple linear regression can be fitted with the `lm` command.

```
> fit.obj=lm(y~x1+x2)
```

- Information can be extracted from the model object using the functions already seen: `confint`, `display` and `summary`.

When to Use Multiple Linear Regression

- Multiple linear regression is useful when more than one predictor is thought to associate with the response simultaneously

- By fitting both predictors in the same model we can get more precise estimates of the regression coefficients

Fitting a Multiple Linear Regression

```
> summary(fit.obj)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-2.0953 -0.7377 -0.1590  0.7445  3.0638

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 0.001302   0.154968   0.008   0.9933
x1          0.987808   0.165837   5.957 3.13e-07 ***
x2          0.424832   0.158882   2.674   0.0103 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.089 on 47 degrees of freedom
Multiple R-squared:  0.4533,^^IAdjusted R-squared:  0.4301
F-statistic: 19.49 on 2 and 47 DF,  p-value: 6.864e-07
```

Multiple Linear Regression: Interpretation of α

- To interpret α put $x_{ij} = 0$ into the regression equation for each predictor:

  $$\mathbb{E}Y_i = \mu_i = \alpha + \beta_1 \times x_{i1} + \beta_1 \times x_{i1} + ...\beta_1 \times x_{ip}$$

  then

  $$\mathbb{E}Y_i = \alpha$$

- α is the average value of the response variable amongst study subjects for which every predictor variable is zero.

Interpretation of β_j

- To interpret β_j, the regression coefficient for the jth predictor variable, put $x = z$ for study subjects i and i' and $x = z + 1$ into the regression equation to obtain:

$$\mathbb{E}Y_i = \alpha + \beta_1 \times x_{i1} + ...\beta_j \times z + ...\beta_p \times x_{ip} \tag{4}$$

$$\mathbb{E}Y_{i'} = \alpha + \beta_1 \times x_{i'1} + ...\beta_j \times (z + 1) + ...\beta_p \times x_{i'p} \tag{5}$$

then take equation (1) from equation (2)

$$\mathbb{E}Y_{i'} - \mathbb{E}Y_i = \beta$$

- β is the difference in the average value of the response variable between groups of study subjects for which the predictor variable differs by one unit.

Logistic Regression

Motivation for (Multiple) Logistic Regression

- We want to model $\mathbb{P}(Y = 1)$ in terms of a set of predictor variables X_1, X_2,... X_p (for univariate regression $p = 1$).
- In linear regression we use the regression equation

$$\mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p \tag{6}$$

- However, for a binary Y (0 or 1), $\mathbb{E}(Y) = \mathbb{P}(Y = 1)$.
- We cannot now use equation (6), because the left hand side is a number between 0 and 1 while the right hand side is potentially a number between $-\infty$ and ∞.
- Solution: replace the LHS with $\operatorname{logit} \mathbb{E} Y$:

$$\operatorname{logit} \mathbb{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

Logistic Regression Equation Written on Three Scales

- We defined the regression equation on the logit or $\log$ ODDS scale:

$$\log \text{ODDS}(Y = 1) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

- On the ODDS scale the same equation may be written:

$$\text{ODDS}(Y = 1) = \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)$$

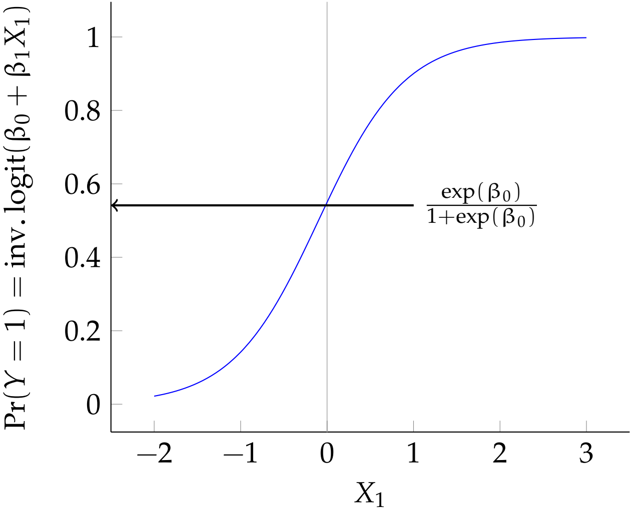
- On the probability scale the equation may be written:

$$\mathbb{P}(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}$$

Interpreting the Intercept

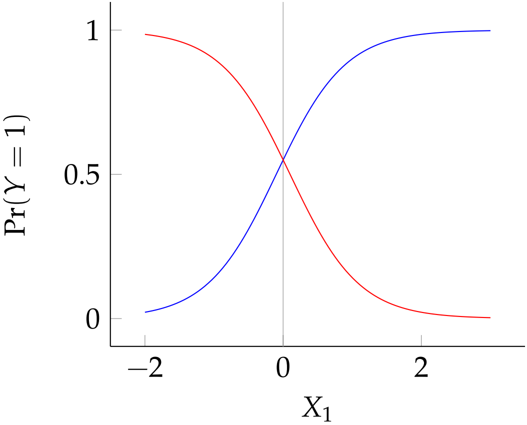
- In order to obtain a simple interpretation of the intercept we need to find a situation in which the other parameters $(\beta_1, ..., \beta_p)$ vanish.
- This happens when $X_1, X_2..., X_p$ are all equal to 0.
- Consequently we can interpret β_0 in 3 equivalent ways:
  1. β_0 is the log-odds in favour of $Y = 1$ when $X_1 = X_2... = X_p = 0$.
  2. β_0 is such that $\exp(\beta_0)$ is the odds in favour of $Y = 1$ when $X_1 = X_2... = X_p = 0$.
  3. β_0 is such that $\frac{\exp(\beta_0)}{1+\exp(\beta_0)}$ is the probability that $Y = 1$ when $X_1 = X_2... = X_p = 0$.
- You can choose any one of these three interpretations when you make a report.

Univariate Picture: Intercept



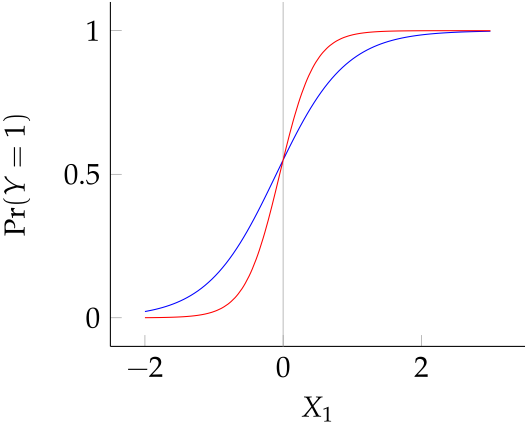
- $\mathbb{P}(Y = 1)$ *vs.* X_1 when $p = 1$ (univariate regression).

Univariate Picture: Sign of β_1



- When $\beta_1 > 0$, $\mathbb{P}(Y = 1)$ increases with X_1 (blue curve).
- When $\beta_1 < 0$, $\mathbb{P}(Y = 1)$ decreases with X_1 (red curve).

Univariate Picture: Magnitude of β_1



- $\beta_1 = 2$ (blue curve), $\beta_1 = 4$ (red curve).
- When $|\beta_1|$ is greater, changes in X_1 more strongly influence the probability that the event occurs.

Interpreting β_1: Univariate Logistic Regression

- To obtain a simple interpretation of β_1 we need to find a way to remove β_0 from the regression equation.
- On the log-odds scale we have the regression equation:

$$\log \text{ODDS}(Y = 1) = \beta_0 + \beta_1 X_1$$

- This suggests we could consider looking at the difference in the log odds at different values of X_1, say $t + z$ and t.

$$\log \text{ODDS}(Y = 1 | X_1 = t + z) - \log \text{ODDS}(Y = 1 | X_1 = t)$$

which is equal to

$$\beta_0 + \beta_1(t + z) - (\beta_0 + \beta_1 t) = z\beta_1.$$

Interpreting β_1: Univariate Logistic Regression

- By putting $z = 1$ we arrive at the following interpretation of β_1:

  β_1 *is the additive change in the log-odds in favour of* $Y = 1$ *when* X_1 *increases by* 1 *unit.*

- We can write an equivalent second interpretation on the odds scale:

  $\exp(\beta_1)$ *is the multiplicative change in the odds in favour of* $Y = 1$ *when* X_1 *increases by* 1 *unit.*

β_1 as a Log-odds Ratio

- The first interpretation of β_1 expresses the equation:

  $$\log \frac{\text{ODDS}(Y = 1|X_1 = t + z)}{\text{ODDS}(Y = 1|X_1 = t)} = z\beta_1$$

  whilst the second interpretation expresses the equation:

  $$\frac{\text{ODDS}(Y = 1|X_1 = t + z)}{\text{ODDS}(Y = 1|X_1 = t)} = \exp(z\beta_1).$$

- The quantity $\frac{\text{ODDS}(Y=1|X_1=t+z)}{\text{ODDS}(Y=1|X_1=t)}$ is the odds-ratio in favour of $Y = 1$ for $X_1 = t + z$ *vs.* $X_1 = t$.

Interpreting Coefficients in Multiple Logistic Regression

- The interpretation of regression coefficients in multiple logistic regression is similar to the interpretation in univariate regression.
- We dealt with β_0 previously.
- In general the coefficient β_k (corresponding to the variable X_k) can be interpreted as follows:

  β_k *is the additive change in the log-odds in favour of* $Y = 1$ *when* X_k *increases by* 1 *unit, while the other predictor variables remain unchanged.*
- As in the univariate case, an equivalent interpretation can be made on the odds scale.

Fitting a Logistic Regression in R

- We fit a logistic regression in R using the `glm` function:

```
> output <- glm(sta ~ sex, data=icu1.dat, family=binomial)
```

- This fits the regression equation

$$\operatorname{logit} \mathbb{P}(\text{sta} = 1) = \beta_0 + \beta_1 \times \text{sex}.$$

- `data=icu1.dat` tells `glm` the data are stored in the data frame `icu1.dat`.
- `family=binomial` tells `glm` to fit a logistic model.
- As an aside, we can use `glm` as an alternative to `lm` to fit a linear model, by specifying `family=gaussian`.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- Summary of the distribution of the deviance residuals.
- Deviance residuals measure how well the observations fit the model. The closer a residual to 0 the better the fit of the observation.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- $\hat{\beta}_0$, the maximum likelihood estimate of the intercept coefficient β_0.
- $\frac{\exp(\hat{\beta}_0)}{1+\exp(\hat{\beta}_0)}$ is an estimate of $\mathbb{P}(\text{sta} = 1)$ when $\text{sex} = 0$

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- $SE(\hat{\beta}_0)$, the standard error of the maximum likelihood estimate of β_0.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- z-value for a Wald-statistic, $z = \hat{\beta}_0 / SE(\hat{\beta}_0)$
- p-value for test of null hypothesis $\beta_0 = 0$ via the Wald-test.
- $p = 2\Phi(z)$, where Φ is the cdf of the normal distribution.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- Significance codes for *p*-values.
- List of *p*-value thresholds (the critical values) corresponding to significance codes.

Logistic Regression: `glm` Output in R

```
Call:
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.6876  -0.6876  -0.6559  -0.6559   1.8123

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- All entries are as for intercept row but apply to β_1 rather than to β_0.

Computing a 95% Confidence Interval from `glm`

```
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
 sex1         0.1054     0.3617   0.291    0.771
---
Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

- We can compute a 95% confidence interval for a regression coefficient using a normal approximation:

$$\hat{\beta}_k - 1.96 \times SE(\hat{\beta}_k) < \beta_k < \hat{\beta}_k + 1.96 \times SE(\hat{\beta}_k)$$

- Plugging in the numbers for β_1:

$$0.105 - 1.96 \times 0.362 < \beta_1 < 0.105 + 1.96 \times 0.362$$
$$-0.603 < \beta_1 < 0.814$$

Computing a 95% Confidence Interval on Odds Scale

- We can compute a 95% confidence interval for the odds-ratio parameter $\exp(\beta_1)$ by transforming the limits to the new scale (see table above).
- Start with the log-odds scale interval:

$$-0.603 < \beta_1 < 0.814$$

- Transform the limits:

$$\begin{aligned} \exp(-0.603) < \exp(\beta_1) < \exp(0.814) \\ 0.547 < \exp(\beta_1) < 2.257 \end{aligned}$$

Logistic Regression with Dummy Variables

- A dummy variable is a 0/1 representation of a dichotomous catagorical variable.
- Such a numeric representation allows us to use categorical variables as predictors in a regression model.
- For example the dichotomous variable sex can be coded

$$
\begin{aligned}
\mathsf{s}ex_i = 0 &\;\; \text{means individual } i \text{ is male} \\
\mathsf{s}ex_i = 1 &\;\; \text{means individual } i \text{ is female}
\end{aligned}
$$

Logistic Regression with Dummy Variables

- Suppose we fit the regression specified by the equation

$$\operatorname{logit} \mathbb{P}(Y_i = 1) = \beta_0 + \beta_1 \mathsf{sex}_i.$$

- Recall one interpretation of β_1:

  $\exp(\beta_1)$ *is the multiplicative change in the odds in favour of* $Y = 1$ *as* sex *increases by* 1 *unit.*

- The only unit increase possible is from 0 to 1, so we can write an interpretation in terms of male/female:

  $\exp(\beta_1)$ *is multiplicative change of the odds in favour of* $Y = 1$ *as a male becomes a female.*

- A bit ridiculous, so better to say:

  $\exp(\beta_1)$ *is the odds-ratio (in favour of* $Y = 1$*) for females vs. males.*

Multiple Logistic Regression Example

- Data on admisssions to an intensive care unit (ICU).
- sta - outcome variable, status on leaving: dead=1, alive=0.
- loc - level of consciousness: no coma/stupor=0, deep stupor=1, coma=2.
- sex - male=0, female=1.
- ser - service at ICU: medical=0, surgical=1.
- ser and sex are dummy variables
- loc is a categorical/factor variable with 3 levels.

Multiple Logistic Regression ICU Example

- Summarise the data:

```
> summary(icu1.dat)
      sta          loc        sex        ser
 Min.   :0.0   0:185   0:124   0: 93
 1st Qu.:0.0   1:  5   1: 76   1:107
 Median :0.0   2: 10
 Mean   :0.2
 3rd Qu.:0.0
 Max.   :1.0
```

- 20% leave ICU dead.

- Categories 1 and 2 of loc are rare, not many people arrive in a stupor/deep coma. This variable may not be very informative.

- sex and ser are reasonably well balanced.

Multiple Logistic Regression ICU Example

- Take an initial look at the 2-way tables cross classifying the outcome with each predictor variable in turn.
- vital status (rows) *vs.* sex (columns):

```
> table(icu1.dat$sta, icu1.dat$sex)
      0   1
  0 100  60
  1  24  16
```

- Observed death rate in males: $24/124 = 0.19$
- Observed death rate in females: $16/76 = 0.21$
- Without doing a formal test, looks significantly different.

Multiple Logistic Regression ICU Example

- vital status (rows) *vs.* service type at ICU (columns):

```
> table(icu1.dat$sta, icu1.dat$ser)

     0  1
  0 67 93
  1 26 14
```

- Observed death rate at medical unit (ser=0): $26/93 = 0.28$
- Observed death rate at surgical unit (ser=1): $14/107 = 0.13$

Multiple Logistic Regression ICU Example

- vital status (rows) *vs.* level of consciousness (columns):

```
> table(icu1.dat$sta, icu1.dat$loc)

      0   1   2
  0 158   0   2
  1  27   5   8
```

- Few observations but higher death rate amongst those in a stupor or coma.

Multiple Logistic Regression ICU Example

- Take an initial look at the 2-way tables cross classifying each pair of predictors.
- sex (rows) *vs.* service type (columns):

```
> table(icu1.dat$sex, icu1.dat$ser)

     0  1
  0 54 70
  1 39 37
```

- Rate of admission to SU in males: $70/124 = 0.56$
- Rate of admission to SU in females: $37/76 = 0.48$
- Some correlation to be aware of but confounding of ser by sex seems unlikely given weak effect of sex.

Multiple Logistic Regression ICU Example

- sex (rows) *vs.* level of consciousness (columns):

```
> table(icu1.dat$sex, icu1.dat$loc)

      0   1   2
  0 116   3   5
  1  69   2   5
```

- Hard to say much, maybe females have higher levels of loc.

Multiple Logistic Regression ICU Example

- Service type (rows) *vs.* level of consciousness (columns):

```
> table(icu1.dat$ser, icu1.dat$loc)

      0   1   2
  0  84   2   7
  1 101   3   3
```

- Hard to say much.
- loc may not to be a useful variable due to low variability.

Multiple Logistic Regression ICU Example

- Now look at univariate regressions.

```
glm(formula = sta ~ sex, family = binomial, data = icu1.dat)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -1.4271     0.2273  -6.278 3.42e-10 ***
sex1          0.1054     0.3617   0.291    0.771
---
$intercept.ci
[1] -1.8726220 -0.9816107

$slopes.ci
[1] -0.6035757  0.8142967

$OR
    sex1
1.111111

$OR.ci
[1] 0.5468528 2.2575874
```

- Wide confidence interval for sex including $OR = 1$.

Multiple Logistic Regression ICU Example

```
glm(formula = sta ~ ser, family = binomial, data = icu1.dat)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -0.9466     0.2311  -4.097 4.19e-05 ***
ser1         -0.9469     0.3682  -2.572   0.0101 *
---
$intercept.ci
[1] -1.3994574 -0.4937348

$slopes.ci
[1] -1.6685958 -0.2252964

$OR
     ser1
0.3879239

$OR.ci
[1] 0.1885116 0.7982796
```

- $OR < 1$ so being in surgical unit may lower risk of death.
- CI implies at least 20% effect.

Multiple Logistic Regression ICU Example

```
Call:
glm(formula = sta ~ loc, family = binomial, data = icu1.dat)
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -1.7668     0.2082  -8.484  < 2e-16 ***
loc1          18.3328  1073.1090   0.017 0.986370
loc2           3.1531     0.8175   3.857 0.000115 ***
---
$intercept.ci
[1] -2.174912 -1.358605

$slopes.ci
             [,1]        [,2]
[1,] -2084.922247 2121.587900
[2,]     1.550710    4.755395
```

- Huge *SE*, should be wary of using this variable.

Multiple Logistic Regression ICU Example

Summary of univariate analyses:

- Vital status not significantly associated with sex.
- Vital status associated with service type at 5% level.
- Admission to surgical unit associated with reduced death rate.
- loc variable not very useful, will now drop.

Multivariate Logistic Regression ICU Example

- Multivariate analysis:

```
Call:
glm(formula = sta ~ sex+ser, family = binomial, data = icu1.dat)
Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.96129    0.27885  -3.447 0.000566 ***
sex1         0.03488    0.36896   0.095 0.924688
ser1        -0.94442    0.36915  -2.558 0.010516 *
---
$intercept.ci
[1] -1.5078281 -0.4147469

$slopes.ci
           [,1]      [,2]
[1,] -0.6882692  0.758025
[2,] -1.6679299 -0.220904

$OR
     sex1      ser1
1.0354933 0.3889063
```

Multivariate Logistic Regression ICU Example

Main Conclusions:

- Univariate and multivariate parameter models show same pattern of significance.
- Direction of association of service variable the same.
- Admission to surgical unit associated with reduced death rate $(OR = 0.39,\ 95\%\ CI = (0.19, 0.80)$.

Prediction In Logistic Regression

- Suppose we fit a logistic regression model and obtain coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, ...\hat{\beta}_p$.
- Suppose we observe a set of predictor variables $X_{i1}, X_{i2}, ...X_{ip}$ for a new individual i.
- If Y_i is unobserved, we can estimate the log-odds in favour of $Y_i = 1$ using the following formula:

$$\text{logit} \frac{\hat{\pi}_i}{1 - \hat{\pi}_i} = \hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip}$$

- Equivilently an estimate of the probability that $Y_i = 1$:

$$\hat{\pi}_i = \frac{\exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip})}{1 + \exp(\hat{\beta}_0 + \hat{\beta}_1 X_{i1} + \hat{\beta}_2 X_{i2} + ... + \hat{\beta}_p X_{ip})}$$

- $\hat{\pi}_i$ can be thought of as a prediction of Y_i.

Prediction In Logistic Regression Using R

- We can use the `predict` function to calculate $\hat{\pi}_i$

```
> output <- glm(sta ~ sex, data=icu1.dat, family=binomial)
> newdata <- data.frame(sex=as.factor(c(0,0,1,1)),
             ser=as.factor(c(0,1,0,1)))

> newdata
  sex ser
1   0   0
2   0   1
3   1   0
4   1   1
```

- Predict on the log-odds scale (i.e. $\log \frac{\hat{\pi}_i}{1-\hat{\pi}_i}$) :

```
> predict(output, newdata=newdata)
        1          2          3          4
-0.9612875 -1.9057045 -0.9264096 -1.8708266
```

- Predict on the probability scale (i.e. $\hat{\pi}_i$) :

```
> predict(output, newdata=newdata, type="response")
        1         2         3         4
0.2766205 0.1294642 0.2836537 0.1334461
```

Multivariate Logistic Regression Example

- Return to ICU example and consider additional variables age and typ.
- sta - outcome variable, status on leaving: dead=1, alive=0.
- sex - male=0, female=1.
- ser - service at ICU: medical=0, surgical=1.
- age - in years
- typ - type of admission: elective=0, emergency=1.

Multivariate Logistic Regression ICU Example

- Look at the joint distribution of the new predictors and the outcome:
- vital status (rows) *vs.* admission type (columns):

```
> table(icu2.dat$sta, icu2.dat$typ)

      0   1
  0  51 109
  1   2  38
```

- Observed death rate for elective admissions: $2/53 = 0.04$
- Observed death rate for emergencies: $38/147 = 0.25$
- Much higher risk of death for admission as an emergency.

Multivariate Logistic Regression ICU Example

- Look at the joint distribution of ser and typ:
- service at ICU (rows) *vs.* admission type (columns):

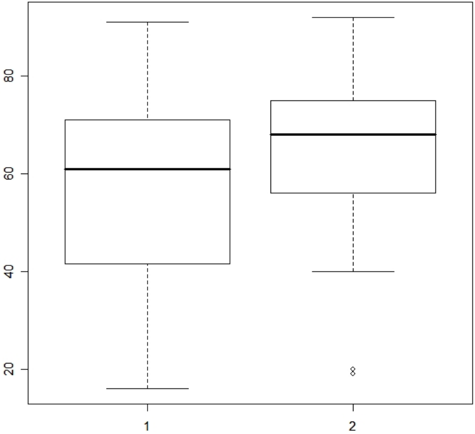
```
> table(icu2.dat$ser, icu2.dat$typ)

     0  1
  0  1 92
  1 52 55
```

- ser and typ are highly correlated.
- We know both variables are associated with outcome
- One might be a confounder for the other

Multivariate Logistic Regression ICU Example

- Box showing distribution of age stratified by vital status

```
>    boxplot(list(icu2.dat$age[icu2.dat$sta==0],
           icu2.dat$age[icu2.dat$sta==1]))
```



Multivariate Logistic Regression ICU Example

- Multivariate analysis:

```
Call:
glm(formula = sta ~ sex + ser + age + typ, family = binomial,
    data = icu2.dat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.2753  -0.7844  -0.3920  -0.2281   2.5072

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.26359    1.11678  -4.713 2.44e-06 ***
sex1        -0.20092    0.39228  -0.512  0.60851
ser1        -0.23891    0.41697  -0.573  0.56667
age          0.03473    0.01098   3.162  0.00156 **
typ1         2.33065    0.80238   2.905  0.00368 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
65050
```

- There is now no significant difference between medical and surgical service types: (ser) has lost its significance.

Multivariate Logistic Regression ICU Example

- Multivariate analysis on odds scale:

```
$OR
     sex1       ser1        age       typ1 
 0.8179766  0.7874880  1.0353364 10.2846123 

$OR.ci
          [,1]      [,2]
[1,] 0.3791710  1.764602
[2,] 0.3477894  1.783083
[3,] 1.0132920  1.057860
[4,] 2.1340289 49.565050
```

- age has a strong effect odds ratio of 1.035 for a 1 year change in age.

- Corresponds to an odds ratio of $1.035^{10} = 1.41$ for a 10 year change in age.

Multivariate Logistic Regression ICU Example

- Multivariate analysis on odds scale:

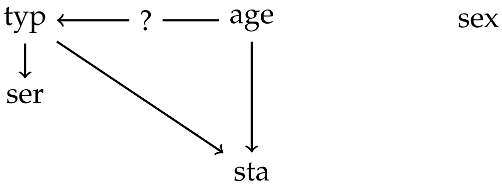
```
$OR
      sex1        ser1         age       typ1
 0.8179766  0.7874880  1.0353364 10.2846123

$OR.ci
          [,1]      [,2]
[1,] 0.3791710  1.764602
[2,] 0.3477894  1.783083
[3,] 1.0132920  1.057860
[4,] 2.1340289 49.565050
```

- age has a strong effect: odds ratio of 1.035 for a 1 year change in age.

- Corresponds to an odds ratio of $1.035^{10} = 1.41$ for a 10 year change in age.

Multivariate Logistic Regression ICU Example

- Draw a causal diagram (DAG)



- Arrow illustrates the direction of causality
- Causality (and so arrows) must obey temporal ordering
- Admission type (emergency/elective) determined before service type (medical/surgical)
- Further evidence that `typ` is the confounder: `ser` is not significant in the multivariate model

Poisson Regression

Estimating Rates

- Context: Suppose we are interested in how the *rate* at which a particular kind of event occurs depends on a set of predictor variables.
- e.g. factors affecting the rates at which people visit their general practitioners.
- We might be interested in estimating the joint effects of a number of predictors simultaneously: e.g. sex, age, employment status, smoking status.

Rates Per Unit of What?

- Because events usually accumulate over time, the denominator of an event rate often includes time.
- Nevertheless other dimensions can be included in the denominator together with time. e.g. if the unit of observation is a population we might look at deaths from liver cancer per unit time *per person*.
- Further, time may not appear at all in the denominator if we are not modelling event rates. e.g we might wish to model the number of skin lesions per unit area of skin.
- From now on we will assume we are modelling events in time so rates are measured with denominator time.
- It is easy enough to adjust the methods of the lecture to other denominators if you wish.

Estimating Rates Using Counts

- Poisson regression links event rate parameters to count data.
- This allows us to estimate the effect of a change in a predictor variable on event rates from counts of the number of times the event occurs to various units of observation.
- e.g. We might have counts of the number of times 10 000 people visit their GP over the period of a year.
- The name *Poisson regression* because it uses the Poisson distribution for event counts.

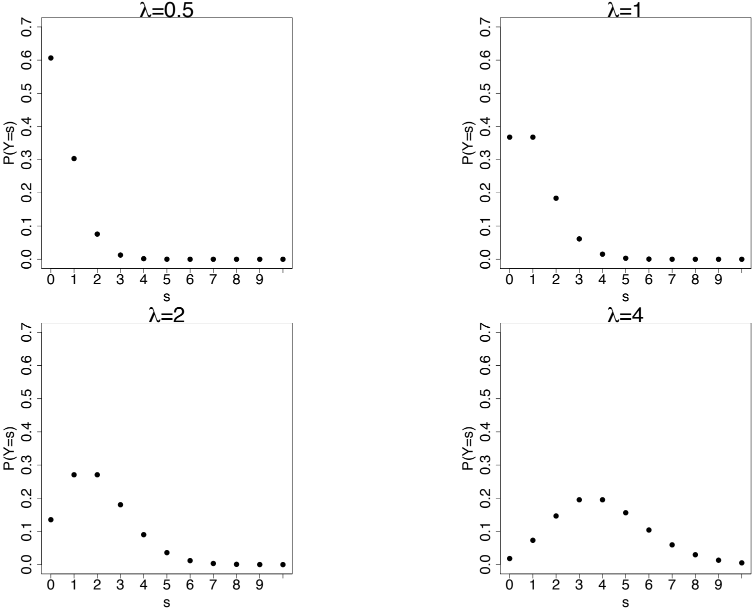
Poisson Distribution

- The Poisson distribution is used for modelling counts.
- A random outcome variable Y has a Poisson distribution with mean λ, if it has probability mass function:

$$\mathbb{P}(Y = s) = \frac{\lambda^s}{s!} \exp(-\lambda)$$

- Y is a count, so s can be any whole number $0, 1, 2, 3$....
- Counts are positive so the mean parameter must be greater than zero. i.e. $\lambda > 0$.
- Some maths shows that if Y has a Poisson distribution with mean λ then $\text{var}(Y) = \lambda$.

Poisson Distribution: Some Examples



Poisson Regression: The Regression Equation

- Poisson regression models Y_i, the event count for observation unit i, as a Poisson distribution with mean λ_i.
- The Poisson regression equation is:

$$\log \mathbb{E} Y_i = \log \lambda_i = \log T_i + \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$$

- λ_i is linked to the predictor variables via $\log \lambda_i$.
- This is because λ_i is a positive number whereas $\log \lambda_i$ can be any number between $-\infty$ and ∞.
- Y_i counts the number of events that occurred in the length of time T_i for which observation unit i was 'at risk.'
- The X_{ij} are the usual predictor variables measured for individual i.

Poisson Regression: Comments on T_i

$$\log \mathbb{E}Y_i = \log \lambda_i = \log T_i + \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$$

- Exponentiate both sides of the regression equation:

$$\mathbb{E}Y_i = T_i \exp(\alpha) \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip})$$

- We see that the expected event count for observation unit i is proportional to the time T_i that the unit was at risk.
- Divide both sides by T_i

$$\frac{\mathbb{E}Y_i}{T_i} = \exp(\alpha) \exp(\beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip})$$

We can interpret this as the event rate for an observation unit with predictor values $X_{i1}, X_{i2}, ...X_{ip}$.

Poisson Regression: Comments on T_i

$$\log \mathbb{E}Y_i = \log T_i + \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$$

- We need to be careful about the unit of measurement of T_i.
- Strictly, it doesn't make sense to take the logarithm of a physical measurement with a unit.
- We can get away with this notation because, as we will see, the units of $\exp(\alpha)$ contain the reciprocal of the unit of T_i so that $T_i \exp(\alpha)$ is unitless.
- e.g. if T_i has units of years then $\exp(\alpha)$ has units years^{-1} i.e. units of '*per year*.'

Poisson Regression: Interpretation of the Intercept

$$\log \mathbb{E}Y_i = \log T_i + \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip}$$

- To interpret the intercept, as usual, we set the X_{ij}s to zero:

  $$\log \mathbb{E}Y_i = \log T_i + \alpha$$

  rearranging:

  $$\exp(\alpha) = \mathbb{E}Y_i / T_i$$

- $\exp(\alpha)$ is the event rate for an observation unit for which all the X_js are equal to zero.

Poisson Regression: Interpretation of the β_j

- To interpret β_j, as usual, we set the X_j to two different values differing by 1. Say $X_{i_1 j} = z + 1$ for observation i_1:

  $$\log \mathbb{E} Y_{i_1} = \log T_{i_1} + \alpha + \beta_1 X_{i_1 1} ... + \beta_j (z+1) ... + \beta_p X_{i_1 p}$$

  and $X_{i_2 j} = z$ in observation i_2:

  $$\log \mathbb{E} Y_{i_2} = \log T_{i_2} + \alpha + \beta_1 X_{i_2 1} ... + \beta_j z ... + \beta_p X_{i_2 p}$$

- Assuming $X_{i_1 j} = X_{i_2 j}$ for all the other j, looking at the difference of these equations we see:

  $$\log \mathbb{E} Y_{i_1} - \log \mathbb{E} Y_{i_2} = \beta_j + \log T_{i_1} - \log T_{i_2}$$

  so rearranging:

  $$\exp(\beta_j) = \frac{\mathbb{E} Y_{i_1} / T_{i_1}}{\mathbb{E} Y_{i_2} / T_{i_2}}$$

Poisson Regression: Interpretation of the β_j

- So

  $$\exp(\beta_j) = \frac{\mathbb{E}Y_{i_1}/T_{i_1}}{\mathbb{E}Y_{i_2}/T_{i_2}}$$

  is a rate ratio.

- $\exp(\beta_j)$ is the ratio of the event rates for a pair of observations with values of X_j which differ by 1 but for which the other predictor variables are the same.

- $\exp(\beta_j)$ is the multiplicative change in the event rate for an observational unit when the predictor X_j increases by one unit while the other predictor variables remain fixed.

Lung Cancer Example

- You can fit a Poisson regression in R using the `glm` command.
- We will look at a lung cancer example.
- 56 122 thousand individuals were followed over one year.
- Number of lung-cancer deaths were counted and grouped by smoking status and age.

Poisson Regression

Lung Cancer Example

```
> lung.cancer
     age          smoke  pop dead
1  40-44             no  656   18
2  45-59             no  359   22
3  50-54             no  249   19
4  55-59             no  632   55
5  60-64             no 1067  117
6  65-69             no  897  170
7  70-74             no  668  179
8  75-79             no  361  120
9    80+             no  274  120
10 40-44  cigarPipeOnly  145    2
11 45-59  cigarPipeOnly  104    4
12 50-54  cigarPipeOnly   98    3
13 55-59  cigarPipeOnly  372   38
14 60-64  cigarPipeOnly  846  113
15 65-69  cigarPipeOnly  949  173
16 70-74  cigarPipeOnly  824  212
17 75-79  cigarPipeOnly  667  243
18   80+  cigarPipeOnly  537  253
19 40-44 cigarrettePlus 4531  149
20 45-59 cigarrettePlus 3030  169
21 50-54 cigarrettePlus 2267  193
22 55-59 cigarrettePlus 4682  576
23 60-64 cigarrettePlus 6052 1001
24 65-69 cigarrettePlus 3880  901
25 70-74 cigarrettePlus 2033  613
26 75-79 cigarrettePlus  871  337
27   80+ cigarrettePlus  345  189
28 40-44 cigarretteOnly 3410  124
29 45-59 cigarretteOnly 2239  140
30 50-54 cigarretteOnly 1851  187
31 55-59 cigarretteOnly 3270  514
32 60-64 cigarretteOnly 3791  778
33 65-69 cigarretteOnly 2421  689
34 70-74 cigarretteOnly 1195  432
35 75-79 cigarretteOnly  436  214
36   80+ cigarretteOnly  113   63
```

```
## pop is the population size (in 1000s) followed in each age/smoking category
## e.g. 656 000 non-smokers aged 40-44 were followed over the year
## dead is the death count from lung cancer in each age/smoking category
## e.g. 18 non-smokers aged 40-44 died over the year.

## Note that numbers of people followed in each category do not reflect the population
## level differences in the category sizes. Otherwise we would be dealing with a
## population where 90% of people smoke.

## This doesn’t matter: Poisson regression is concerned with rates and
## comparisons of rates between exposure categories. The proportions of people in
## each exposure category do not need to reflect population proportions.

## The pop variable is analogous to the "time at risk" variable T in the slides.
## In this case, because each unit of observation is a group of people, we are
## dealing with events occurring in person-time rather than events occurring
## just in time. So T is person-time at risk for each unit of exposure.

## First lets fit a Poisson regression regressing dead on age, assuming smoking
## does not cause lung cancer.

## In order to specify the log(T) term of the slides, (log(pop) here) we use the
## offset command inside the regression formula:

> age.mod.obj<-glm(dead˜offset(log(pop))+age, data=lung.cancer,family=poisson)
> age.mod.obj

Call:  glm(formula = dead ˜ offset(log(pop)) + age, family = poisson,
    data = lung.cancer)

Coefficients:
(Intercept)     age45-59     age50-54     age55-59     age60-64     age65-69
    -3.3957       0.5560       0.9881       1.3715       1.6290       1.9571
   age70-74     age75-79       age80+
     2.2058       2.4578       2.6875

Degrees of Freedom: 35 Total (i.e. Null);  27 Residual
Null Deviance:     4056
```

```
## We need to think about what the unit of the event rate is for this model.
## We are following the individuals for one year, so the denominator of the
## unit includes time in units of 1 year. Each unit of observation is
## a cohort of people in a particular exposure group (defined by age interval
## and smoking category).  Since we are measuring the size of these units in 1000s
## of people, the denominator should also include a unit of 1000 persons.

## rate unit = per 1000 persons * per 1 year = per 1000 person years

## What is an estimate of the lung cancer death rate in the age 40-44 category?

## The regression coefficients are stored in age.mod$coef:

> age.mod.obj$coef
(Intercept)    age45-59    age50-54    age55-59    age60-64    age65-69
 -3.3957217   0.5560324   0.9881493   1.3714516   1.6289950   1.9571451
   age70-74    age75-79      age80+
  2.2057743   2.4577851   2.6874888

## age40-44  is the reference category of the factor variable, so we can
## compute the rate as:

> exp(age.mod.obj$coef[1])
(Intercept)
 0.03351636

## There are 0.034 lung cancer deaths per 1000 person years in 40-44 year olds.

## What is the estimated rate ratio for lung-cancer deaths comparing
## 60-64 year olds with 40-44 year olds?

## We can compute this rate-ratio as:

> exp(age.mod.obj$coef[5])
age60-64
5.098748

## The lung cancer rate ratio is 5.1 (note as a ratio of two quantities
## with the same units this is a unitless quantity). The lung cancer rate is about
## 5 times as high in 60-64 year olds as in 40-44 year olds

## What is the lung cancer death rate in the age 60-64 category?

## This is the product of the two things we have just calculated: the rate in the
## 40-44 year old category and the rate ratio comparing the 60-64 year olds
```

```
## with the 40-44 year olds.

## We can calculate it directly like this:

> exp(age.mod.obj$coef[1]+age.mod.obj$coef[5])
(Intercept)
  0.1708915

## The lung cancer death rate in 60-64 year olds is 0.17 deaths
## per 1000 person years.

## Now lets fit a model including the various smoking categories as dummy variables
## by adding smoke to the model:

> age.smoke.mod.obj<-glm(dead~offset(log(pop))+age+smoke,
     data=lung.cancer,family=poisson)
> summary(age.smoke.mod.obj)

Call:
glm(formula = dead ~ offset(log(pop)) + age + smoke, family = poisson,
    data = lung.cancer)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.06055  -0.54773   0.06431   0.29963   1.48348

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)         -3.68002    0.06824 -53.929  < 2e-16 ***
age45-59             0.55388    0.07999   6.924 4.38e-12 ***
age50-54             0.98039    0.07682  12.762  < 2e-16 ***
age55-59             1.37946    0.06526  21.138  < 2e-16 ***
age60-64             1.65423    0.06257  26.439  < 2e-16 ***
age65-69             1.99817    0.06279  31.824  < 2e-16 ***
age70-74             2.27141    0.06435  35.296  < 2e-16 ***
age75-79             2.55858    0.06778  37.746  < 2e-16 ***
age80+               2.84692    0.07242  39.310  < 2e-16 ***
smokecigarPipeOnly   0.04781    0.04699   1.017    0.309
smokecigarretteOnly  0.41696    0.03991  10.447  < 2e-16 ***
smokecigarrettePlus  0.21796    0.03869   5.633 1.77e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)
```

```
    Null deviance: 4055.984  on 35  degrees of freedom
Residual deviance:   21.487  on 24  degrees of freedom
AIC: 285.51

Number of Fisher Scoring iterations: 4

## The estimated rate ratio comparing those who smoke cigarettes only,
## with those who do not smoke, within the same age band is:
> exp(age.smoke.mod.obj$coeff[11])
smokecigarretteOnly
           1.517341

## Compute a 95% confidence interval for this rate ratio.

## We can use confint to get a 95% confidence interval for the
## regression coefficients (on the log-rate scale):

> intervals<-confint(age.smoke.mod.obj)
Waiting for profiling to be done...
> intervals
                          2.5 %     97.5 %
(Intercept)         -3.81561558 -3.5480606
age45-59             0.39729564  0.7110067
age50-54             0.83034745  1.1316137
age55-59             1.25314788  1.5090627
age60-64             1.53341845  1.7787749
age65-69             1.87691591  2.1231296
age70-74             2.14695537  2.3993120
age75-79             2.42712423  2.6929365
age80+               2.70602687  2.9900292
smokecigarPipeOnly  -0.04414904  0.1400901
smokecigarretteOnly  0.33928488  0.4957578
smokecigarrettePlus  0.14271999  0.2944183

## Then exponentiate to get the intervals on the rate scale:

> exp(intervals)
                          2.5 %     97.5 %
(Intercept)          0.02202415  0.0287804
age45-59             1.48779572  2.0360399
age50-54             2.29411568  3.1006561
age55-59             3.50134746  4.5224900
age60-64             4.63399086  5.9225961
age65-69             6.53332439  8.3572514
age70-74             8.55876044 11.0155956
```

```
age75-79             11.32626342 14.7749991
age80+               14.96968079 19.8862627
smokecigarPipeOnly    0.95681134  1.1503774
smokecigarretteOnly   1.40394324  1.6417420
smokecigarrettePlus   1.15340679  1.3423453

## The relevant interval is (1.40394324,1.6417420)
## If we repeated the experiment many times, each time estimating a
## confidence interval, the estimated intervals will include the true
## rate ratio parameter 95% of the time.

## We could also have done this using a normal approximation, using the
## standard formula:
## estimate - 1.96*se < true value < estimate+1.96*se
## to calculate the interval on the log-rate scale. We can then exponentiate
## to get the interval on the rate-scale. In a single step for each limit:

## lower limit of interval:
> exp(age.smoke.mod.obj$coeff[11]+1.96*0.03991)
smokecigarretteOnly
           1.640799

## upper limit of interval:
> exp(age.smoke.mod.obj$coeff[11]-1.96*0.03991)
smokecigarretteOnly
           1.403173

## It seems likely smoking does cause lung cancer.
```

Negative Binomial Regression

- When Y has a Poisson distribution with mean λ, $\mathrm{var}(Y) = \lambda$.
- In practice however, count data often have $\mathbb{E}Y < \mathrm{var}(Y)$. This is *over dispersion*
- It is common with count data from sequencing reads
- Over dispersion can be consequence of "mixing" of multiple Poisson distributions each with a slightly different mean (e.g. due to batch effects)
- There are count regression models which account for over dispersion
- Negative binomial regression is one such examples (`glm.nb` in R MASS package)

The End!