Introduction to Logistic Regression

Session 4

presented by:
Imen Hammami

OxPal Research Fellowship Series 2021

February, 2020

What will this session tell you?

- Understand when to use logistic regression (LR).
- Interpret a LR model with a binary exposure variable.
- Assess model fit.

Your new friend

Jane Superbrain 2.0

- ☺ She steals the brains of top statisticians.
- ☺ She appears in red boxes to tell you really important things.

Content

Layout

Recap

Regression models are used to:

- Describe relationship between 2 or more variables where one of these is 'dependent' ('response', 'outcome').
- Predict the value of the dependent variable for a given value of the independent variable.

We can describe the linear relationship between y and x as:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 \ x_i$$

We can use the linear model to investigate the association between number of hours studying and exam scores as:

$$\text{exam\^{}score} = \hat{\beta}_0 + \hat{\beta}_1 \text{ hours studying}$$

Exam score

We could describe the linear relationship between hours studying and exam scores using the linear regression model:

$$\hat{\text{score}} = \hat{\beta}_0 + \hat{\beta}_1 \text{ hours}$$

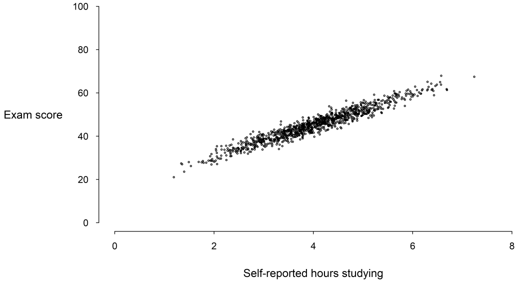


Figure 1: Exam scores versus hours studying.

Exam score (Cont.)

Using the equation of the regression line, calculate the estimated exam score for hours studying$= 6$.

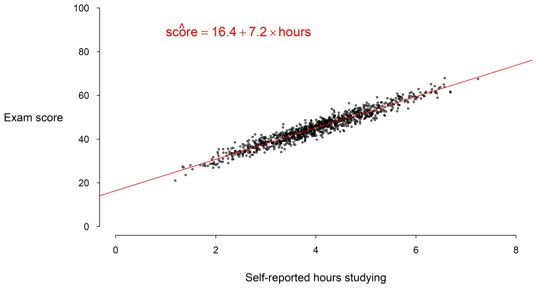


Figure 1 (Cont.): Exam scores versus hours studying.

Exam score (Cont.)

Using the equation of the regression line, calculate the estimated exam score for hours studying$= 6$.

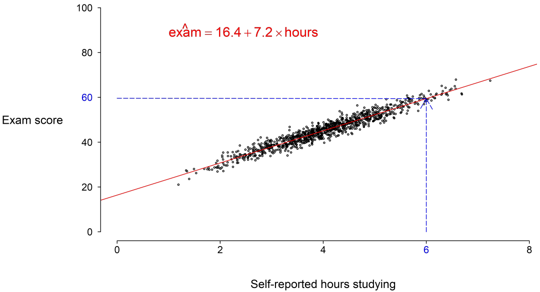


Figure 1 (Cont.): Exam scores versus hours studying.

Exam score (Cont.)

Let's assume a student passes the exam if score ≥ 50.

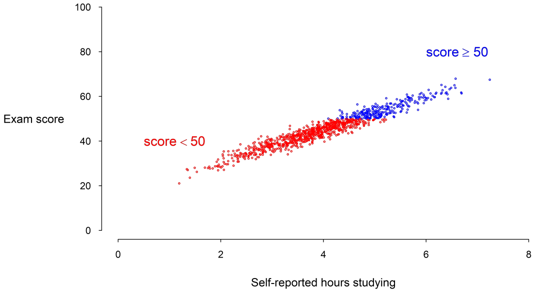


Figure 1 (Cont.): Exam scores versus hours studying.

Passing exam (0 vs 1)

How could we model this relationship between passing exam (0 vs 1) and hours studying?

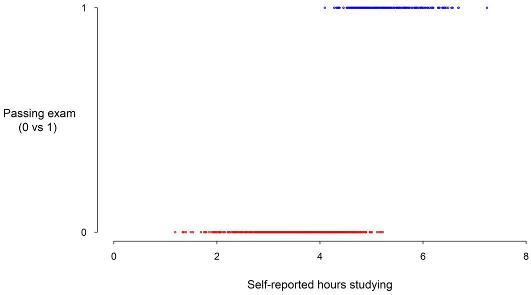


Figure 2: Passing exam versus hours studying.

Passing exam (0 vs 1) (Cont.)

The line does not fit the data very well. It goes below 0 and above 1.

If we take values of Y between 0 and 1 to be probabilities, this does not make sense.

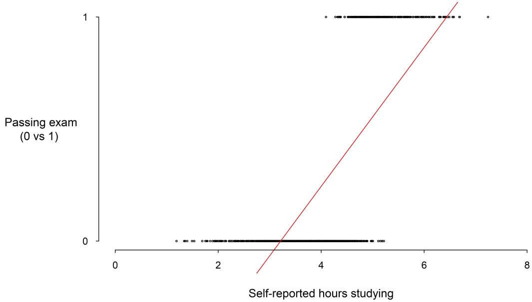


Figure 2 (Cont.): Passing exam versus hours studying.

Probability of passing exam [0,1]

How could we link the probability of passing exam to the continuous predictor ‘hours studying’? The risk is constrained to fall in the interval [0,1].

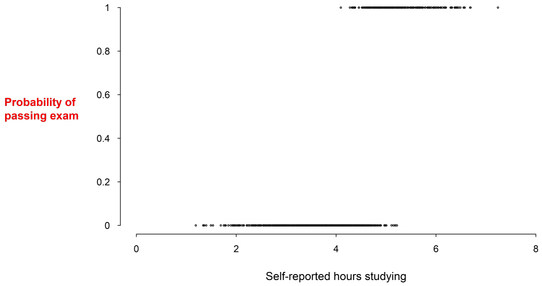


Figure 3: Probability of passing exam versus hours studying.

S-shaped curve

We can use a S-shaped curve to model the relationship between probability of passing and hours studying.

What's the probability of passing for hours studying$= 5$?

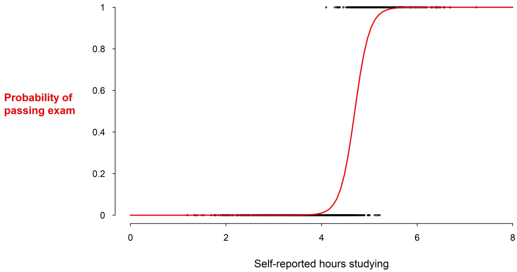


Figure 3 (Cont.): Probability of passing exam versus hours studying.

S-shaped curve

We can use an S-shaped curve to model the relationship between probability of passing and hours studying.

What's the probability of passing for hours studying$= 5$?

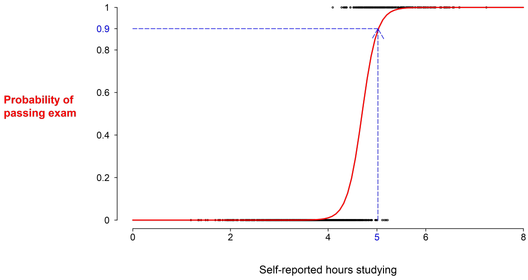


Figure 3 (Cont.): Probability of passing exam versus hours studying.

S-shaped curve (Cont.)

What's the minimum number of hours studying required for $\geq 50\%$ chance to pass exam?

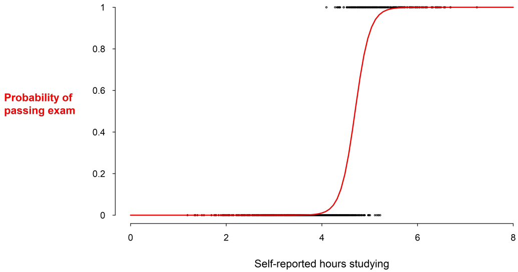


Figure 3 (Cont.): Probability of passing exam versus hours studying.

S-shaped curve (Cont.)

What's the minimum number of hours studying required for $\geq$ 50% chance to pass exam?

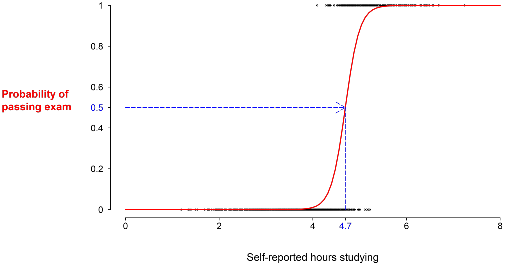
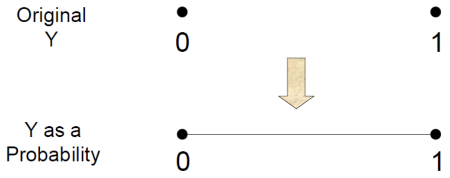


Figure 3 (Cont.): Probability of passing exam versus hours studying.

Original Y (disease 0 vs 1)

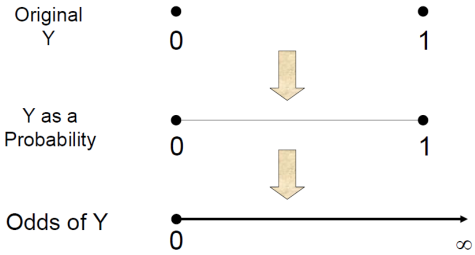
We need to transform the dichotomous Y into a continuous variable Y'. We need a (link) function that takes a dichotomous Y and gives us a continuous Y'.

Y as probability [0,1]



If we work with Y as a probability, what function F(Y) goes from [0,1] interval to the real line? We know at least one function that goes the other way round (but we won't use that one!).

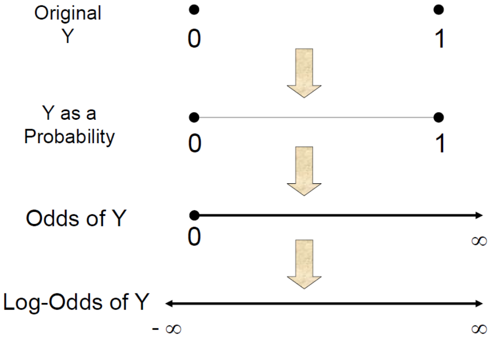
Odds of Y



Let's look at an alternative approach based on odds.

Taking the odds of Y occurring moves us from the [0,1] interval to the half-line $[0, +\infty[$ (odds are always non-negative).

Log-odds of Y



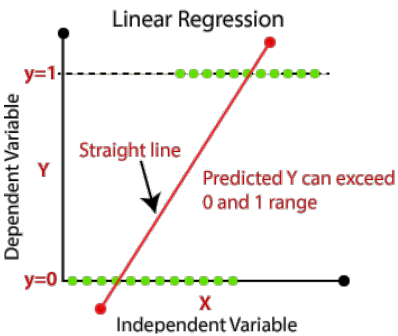
As a final step, let's take the log of the odds.
This is called the logit function:

$$Y' = \text{logit}(Y) = \log(\text{odds } Y) = \log(Y/(1 - Y))$$

(Y as probability of disease).

The logistic model

The smooth S-shaped curve is known as the logistic (or logit) model.



Assuming a linear relationship between log(odds Y) and a predictor X, we can fit a linear regression model with log(odds Y) as the dependent variable and X as the independent variable.

Important points

Jane Superbrain 2.0

Properties of the logistic model:

- Allows for a smooth change in risk throughout the range of X.
- Has the property that risk increases slowly up to a threshold range of X, followed by a more rapid increase and a subsequent leveling off of risk.
- This shape is consistent with many dose response relationships (e.g. likelihood of toxicity response to varying levels of treatment).

Layout

Heart data

Consider a prospective cohort study conducted for the purpose of studying the determinants of ischaemic heart disease (IHD) among 844 men without prior cardiovascular disease.

The men were subsequently followed-up for 10 years, at which point the investigators wanted to assess whether baseline levels of serum cholesterol were associated with IHD mortality.

The investigators conducted a nested case-control study of 68 IHD cases and 138 controls from the main cohort and measured cholesterol levels in these participants.

2x2 Contingency Table

The following Stata output shows the cross-tabulation of the number of participants with a diagnosis of IHD by baseline serum cholesterol (high versus low).

```
. tab ihd hichol1, m
```

Ischaemic heart disease	High serum cholesterol 0	1	Total
0	73	65	138
1	24	44	68
Total	97	109	206

1. Calculate proportions with IHD in those with and without high cholesterol.
2. Calculate odds of IHD in those with and without high cholesterol.

2x2 Contingency Table (Cont.)

```
. tab ihd hichol1, m
```

Ischaemic heart disease	High serum cholesterol 0	1	Total
0	73	65	138
1	24	44	68
Total	97	109	206

Proportions

$$P(\text{ihd} = 1|\text{hichol} = 1) = 44/(44 + 65) = 0.40$$
$$P(\text{ihd} = 1|\text{hichol} = 0) = 24/(24 + 73) = 0.25$$

2x2 Contingency Table (Cont.)

```
. tab ihd hichol1, m
```

Ischaemic heart disease	High serum cholesterol 0	1	Total
0	73	65	138
1	24	44	68
Total	97	109	206

Odds

$$\text{odds}(\text{ihd} = 1|\text{hichol} = 1) = 44/65 = 0.68$$
$$\text{odds}(\text{ihd} = 1|\text{hichol} = 0) = 24/73 = 0.33$$

Odds ratio

$$\begin{aligned}\text{odds ratio} &= \text{odds}(\text{ihd} = 1|\text{hichol} = 1)/\text{odds}(\text{ihd} = 1|\text{hichol} = 0)\\ &= 0.68/0.33 = 2.06\end{aligned}$$

The simple logistic regression model

Suppose a logistic regression of ischemic heart disease (ihd) on high baseline serum cholesterol (hichol1) is performed:

$$\log(\text{odds of ihd}) = \beta_0 + \beta_1 \text{hichol1}$$

Based on the equation above, what are β_0 and β_1?

For an unexposed person (i.e. with low cholesterol), substitute hichol1 $=$ 0 into the model:

$$\log(\text{odds of ihd}) = \beta_0 + \beta_1 \times 0 = \beta_0$$

β_0 is log odds in unexposed.

The simple logistic regression model (Cont.)

Suppose a logistic regression of ischemic heart disease (ihd) on high baseline serum cholesterol (hichol1) is performed:

$$\log(\text{odds of ihd}) = \beta_0 + \beta_1 \text{hichol1}$$

For an exposed person (i.e. with low cholesterol), substitute hichol1 $= 1$ into the model:

$$\log(\text{odds of ihd}) = \beta_0 + \beta_1 \times 1 = \beta_0 + \beta_1$$

The calculation above could be rewritten as follows

$$\begin{aligned}
\beta_1 &= \log(\text{odds in exposed}) - \beta_0 \\
&= \log(\text{odds in exposed}) - \log(\text{odds in unexposed}) \\
&= \log(\text{odds in exposed}/\text{odds in unexposed}) \\
&= \log(\text{odds ratio})
\end{aligned}$$

β_1 is log odds ratio.

The simple logistic regression model (Cont.)

Based on our previous calculations (2x2 table):
$OR = 2.06$
odds in unexposed $= 0.33$

The logistic regression of ischemic heart disease (ihd) on high baseline serum cholesterol (hichol1) is given by:

$$\begin{aligned} \log(\text{odds of ihd}) &= \beta_0 + \beta_1 \text{hichol1} \\ &= \log(0.33) + \log(2.06) \times \text{hichol1} \\ &= -1.11 + 0.72 \times \text{hichol1} \end{aligned}$$

Fitting a logistic regression model

```
. *-- log scale;
. logit ihd hichol1, noheader nolog
```

ihd	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hichol1	0.722	0.306	2.36	0.018	0.123	1.321
_cons	-1.112	0.235	-4.73	0.000	-1.574	-0.651

$$\log(\text{odds of ihd}) = -1.11 + 0.72 \times \text{hichol1}$$

Interpretation

```
. *-- get ORs;
. logit ihd hichol1, or noheader nolog
```

ihd	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hichol1	2.059	0.630	2.36	0.018	1.131	3.749
_cons	0.329	0.077	-4.73	0.000	0.207	0.521

```
Note: _cons estimates baseline odds.
```

A person with high cholesterol have a 2-fold higher odds of IHD as compared to a person with low cholesterol (OR=2.06 [1.13, 3.75], $p = 0.018$).

Important points

Jane Superbrain 2.0

- In logistic regression, we model the log (odds of disease) as the outcome:

$$\log(\text{odds of disease}) = \beta_0 + \beta_1 x$$

  where

  $\beta_0 =$ log odds in the unexposed

  $\beta_1 =$ log OR

- We use this model to estimate log OR and hence OR (with 95% CI, p-value).
- We use a statistics package to fit logistic regression models.
- Estimation is done using the method of maximum likelihood*.

Important points (Cont.)

Jane Superbrain 2.0

Study designs in which logistic regression may be used:

- A cross-sectional study

  Model parameters are interpreted as above.

  If outcome is not rare, then OR will overestimate the risk ratio.

- An unmatched case-control study

  Parameter β_1 is log OR, but β_0 (log odds in unexposed) is not interpretable.

Confounding by age

Looking at the regression results below, describe the impact of age on the association between IHD and baseline serum cholesterol.

```
. logit ihd hichol1, or noheader nolog
```

ihd	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hichol1	2.059	0.630	2.36	0.018	1.131	3.749
_cons	0.329	0.077	-4.73	0.000	0.207	0.521

```
. logit ihd hichol1 age, or noheader nolog
```

ihd	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hichol1	1.884	0.596	2.00	0.045	1.014	3.502
age	1.071	0.021	3.56	0.000	1.031	1.112
_cons	0.008	0.009	-4.41	0.000	0.001	0.069

Confounding by age (Cont.)

```
. logit ihd hichol1, noheader nolog
```

ihd	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hichol1	0.722	0.306	2.36	0.018	0.123	1.321
_cons	-1.112	0.235	-4.73	0.000	-1.574	-0.651

```
. logit ihd hichol1 age, noheader nolog
```

ihd	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
hichol1	0.633	0.316	2.00	0.045	0.014	1.253
age	0.068	0.019	3.56	0.000	0.031	0.106
_cons	-4.808	1.090	-4.41	0.000	-6.945	-2.672

The confounding effect of age can be quantified by computing the percentage difference between the crude and adjusted coefficients $(\beta_{\text{unadjusted}}\text{-}\beta_{\text{adjusted}})/\beta_{\text{unadjusted}}$ ($\sim 12\%$)

Layout

1. From linear to logistic regression
   - The linear model
   - Redefining the dependent variable
2. The logistic regression model
   - The unadjusted model
   - Adjusting for confounders
3. Assessing model fit
   - Classification threshold
   - Performance of a classification model

Hits and misses

What percent of the observation the model correctly predicts?

```
. qui logit ihd hichol1 age, noheader nolog

. lstat

Logistic model for ihd

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        14            10  |         24
     -     |        54           128  |        182
-----------+--------------------------+-----------
   Total   |        68           138  |        206

Classified + if predicted Pr(D) >= .5
True D defined as ihd != 0
```

Hits and misses (Cont.)

1. Use model to generate the probability p that each observation will have the disease.
2. Use a cutoff $\pi = 0.5$. If $p \geq \pi$ predict ihd$= 1$, if $p < \pi$ predict ihd$= 0$.
3. Check predictions against the actual outcomes in the data.

```
. qui logit ihd hichol1 age, noheader nolog

. lstat

Logistic model for ihd

              -------- True --------
Classified |         D            ~D  |      Total
-----------+--------------------------+-----------
     +     |        14            10  |         24
     -     |        54           128  |        182
-----------+--------------------------+-----------
   Total   |        68           138  |        206

Classified + if predicted Pr(D) >= .5
True D defined as ihd != 0
```

Hits and misses (Cont.)

Output shows summary of correct and incorrect predictions.

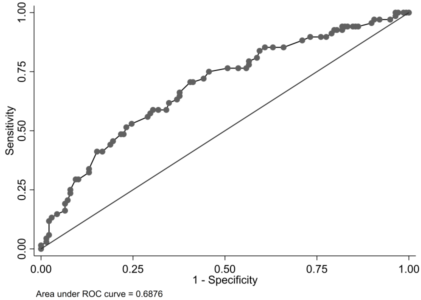
```
Classified + if predicted Pr(D) >= .5
True D defined as ihd != 0
--------------------------------------------------
Sensitivity                     Pr( +| D)   20.59%
Specificity                     Pr( -|~D)   92.75%
Positive predictive value       Pr( D| +)   58.33%
Negative predictive value       Pr(~D| -)   70.33%
--------------------------------------------------
False + rate for true ~D        Pr( +|~D)    7.25%
False - rate for true D         Pr( -| D)   79.41%
False + rate for classified +   Pr(~D| +)   41.67%
False - rate for classified -   Pr( D| -)   29.67%
--------------------------------------------------
Correctly classified                        68.93%
--------------------------------------------------
```

Overall success rate $= (14 + 128)/206 = 68.93\%$

ROC curve

We can imagine changing the cutoff point π continuously from 0 to 1. The ROC curve plots the sensitivity ($S_e = P(+|D)$) and 1−specificity ($S_p = P(-|\bar{D})$) as π goes from 0 to 1.

Area under the curve is 0.6876.



Important points

Jane Superbrain 2.0

Using logistic regression to make a classifier

The goal here is to model and predict if a given observation (row in dataset) has disease or not based on other variables/features in the dataset.

1. Split dataset into training, validation and test sets.
2. Build logistic (logit) model on the training set.
3. Tune the parameters of the classifier on the validation set.
4. Assess model performance using the test set.

What did this session tell you?

- Understand when to use logistic regression (LR).
- Interpret a LR model with a binary exposure variable.
- Assess model fit.

Matching cases and controls based on age

The relationship between IHD and baseline serum cholesterol (low versus high) was further investigated by identifying each individual who developed IHD and matching that person, based on age, to an individual who had not developed IHD.

The count data on the resulting matches are tabulated below.

	Not developing IHD (controls)	
Developing IHD (cases)	High cholesterol	Low cholesterol
High cholesterol	20	21
Low cholesterol	11	16

Based on these data, calculate the odds ratio for the association between IHD and serum cholesterol level (high versus low) among individuals of the same age.

Matching cases and controls based on age (Cont.)

This is a matched case-control study and the estimated odds ratio is based on the discordant pairs b & c.

b is the number of pairs in which emp developing IHD have high baseline serum cholesterol and their matched emp not developing IHD have low baseline serum cholesterol.

c is number of pairs in which emp developing IHD have low baseline serum cholesterol and their matched emp not developing IHD have high baseline serum cholesterol.

Matching cases and controls based on age (Cont.)

	Not developing IHD (controls)	
Developing IHD (cases)	High cholesterol	Low cholesterol
High cholesterol	20	21
Low cholesterol	11	16

OR $= b/c = 21/11 = 1.9$