Probabilities & Data
Week 11: Logistic Regression

DR NICK FEWSTER-YOUNG

Topics to be covered

- Linear Regression – the basics to get us into logistic regression
- Logistic Regression
  - Predicting binary categorical outcome variables using Categorical and Continuous explanatory variables.
- Model fit and diagnostics.
- Odds ratios.

Motivation

We introduce logistic regression as a tool for building models when there is a categorical response variable with two levels (outcomes).

Logistic regression is a type of generalized linear model (GLM) for response variables where regular regression does not work very well.

In particular, the response variable in these settings often takes a form where residuals look completely different from the Normal distribution.

Very useful and has many applications in determining whether a customer defaults on a loan or payment, if a email is spam or not, if you win or loss, if you buy a used version or new version of Mario Cart **64** ☺ **.**

Regression – the basics

- Predictors (explanatory variables)
- A Response variable
- An Error Distribution (Residuals)
- Link Function
- Coefficients – Estimation
- Variance Function

Predictors & The Response

- Predictor variables
  - The inputs to the model
  - The *Cause* variable
  - The variable which is independent

- The Response
  - The output in the model
  - The *Effect* variable
  - The variable which we want to predict
  - The variable which is dependent

Linear Regression

Let y be the response variable.

Let x be the explanatory variable or predicator variable.

Historically there is data where x_i explains y_i.

If we assume the **link function** in this case is a linear equation then

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where the coefficients β_0, β_1 and ϵ_i (errors) are the quantities such that the above is true.

By the **method of least squares**, we determine estimates on the coefficients β_0, β_1 which estimate the following equation:

$$y \sim \widehat{\beta_0} + \widehat{\beta_1} x.$$

Coefficients of a model

- The coefficients of a model is a critical part of determining an accurate model. There are different methods which can be used and different link functions or functions to model different scenarios.
- The coefficients fit the model best they can and are estimated to minimize the error in the model.
- They are sample estimates for the population parameters.
- How do we calculate them?

Error terms and its Distribution

The error terms ϵ_i are the result of the least squares method to find the best fitting coefficients β_0, β_1 and thus are assumed to be independent random variables being Normally distributed with mean zero and a constant variance σ^2. Therefore, the response variable should be

$$y \sim N(\mu, \sigma^2)$$

where $\mu = \beta_0 + \beta_1 x$.

- However, this simple model and generalisation of it with multiple predicators breaks down when the response variable, y is binary. That means we want to use a Logistic Regression model.
- Moreover, it means the family of the errors terms becomes a **Binomial Distribution**.

Logistic Regression

- Used to predict a binary (dichotomous) categorical response variable from one or more categorical and/or continuous explanatory variables.
- The response variable y is a dummy variable coded 0 if a condition is not present and 1 if it is.
- Instead of predicting the value of y from variable x we are interested to predict the probability of y occurring given known values of x.
- Thus we investigate how the probability that a successful outcome occurs depends upon each value of explanatory variable x.

Problems with using linear regression

A simple linear regression model would have a hard time fitting a straight line!

You would typically get the correct answers in terms of the sign and significance of coefficients.

There are four problems:

- The error terms do not have constant variance.
- The error terms are not Normally distributed.
- Probabilities are bounded between 0 and 1.
- If the response is coded 1 = Yes and 0 = No and your regression equation predicts 1.1 or -0.4, what does that mean?

Example

Consider the game, Mario Cart 64 (I'm sure you have heard of it). You are planning on buying a copy on Ebay and analyzing the condition (second hand or new) to price. If you were plot the two variables, well one is categorical and the other is continuous. Thus, it would look like this and with a linear regression line:

That is ugly! How well is that going to work as a prediction?

We need a Logistic Regression Model!

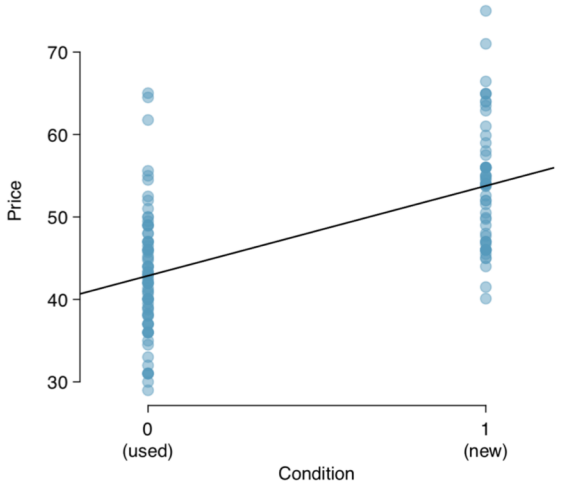


Figure 8.4: Scatterplot of the total auction price against the game's condition. The least squares line is also shown.

Logistic : y is binary

For instance, if we were to model the expected value of this type of binary response, then the probability of it taking the value one, say π, directly as a linear equation of the explanatory variables then the fitted values for the response could be greater than 1. This is absurd!

Also, the assumption of Normality of the errors, ϵ_i would not make sense and if we write the equation as

$$y = \pi(x_1) + \epsilon$$

then if $y = 1$ then $\epsilon = 1 - \pi(x_1)$ with probability $\pi(x)$ and if $y = 0$ then $\epsilon = \pi(x)$ with probability $1 - \pi(x)$. So ϵ has a distribution with mean 0 and variance equal to $\pi(x)(1 - \pi(x))$.

This is what we call the conditional distribution of a binary response variable and follows a Binomial distribution with probability given by the conditional mean, $\pi(x)$.

Link Function for Logistic Regression

Instead of modelling the expected value of the response variable directly as a linear function of the explanatory variables, we use a transformation instead called the logistic or logit function.

The logit function (a type of link function) is defined as follows and leads to the model

$$\text{Logit}(\pi) = \log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x.$$

The Logit of a probability is just the natural log of the odds of the response taking the value 1. Therefore, we can rearrange and write

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}.$$

See that $0 \leq \pi(x) \leq 1$, so it remains a probability and useful.

The parameters β_0, β_1 are estimated by using maximum likelihood. ☺

Examples

Here we create a spam filter with a single predictor: **multiple.** This variable indicates whether more than one email address was listed in the ***To field*** of the email. The following logistic regression model was fit using **R**:

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.12 - 1.81 \times \boldsymbol{multiple}$$

If an email is randomly selected and it has just one address in the ***To field***, what is the probability it is spam?

What if more than one address is listed in the ***To field***?

Spam

If an email is randomly selected and it has just one address in the ***To field***, what is the probability it is spam?

If there is only one email in the ***To field,*** then multiple takes the value 0 and the right side of the model equation equals -2.12. Thus, solving for p_i:

$$p_i = \frac{e^{-2.12-1.81\times multiple}}{1+e^{-2.12-1.81\times multiple}} = \frac{e^{-2.12}}{1+e^{-2.12}} = 0.11.$$

Just as we labeled a fitted value of y_i with a hat in a single variable case, we will do the same since it is a estimate for the probability: $\widehat{p_i} = 0.11$.

For each extra multiple address added on ***To field*** then the right side of the model equation is -3.93, which corresponds to a probability $\hat{p}_i = 0.02$. Thus, we can see from the behavior that the probability will decrease as we increase the ***To field*** variable.

Downside – What's next?

While the information about whether the email is addressed to multiple people is a helpful start in classifying email as spam or not, the probabilities of 11% and 2% are not dramatically different, and neither provides very strong evidence about which particular email messages are spam. To get more precise estimates, we'll need to include many more variables in the model.

Summary: Odds and Probability

- The Logit model can be used for binary logistic regression:

$$\text{Log}\left(\frac{p}{1-p}\right) = b_0 + b_1 x$$

- Logit is the natural Log of the odds ratio, that is $\frac{p}{1-p}$ with p being the probability y takes the value 1 ad 1-p is the probability y takes the value 0.

$$P(Y = 1\ ) = p$$

- We use the natural log function for binary responses and so we can fit a S-shaped curve to what appears to a be linear model to interpret probabilities.

- By taking the exponent (exponential function), we obtain the Odd ratio

$$Odds\ Ratio = \frac{p}{1-p} = e^{b_0 + b_1 x}$$

Summary (cont.)

For one predictor variable (explanatory variable), we have P(y),, the probability of y occurs:

$$P(Y = 1) = \frac{e^{b_0 + b_1 x}}{1 + e^{b_0 + b_1 x}}$$

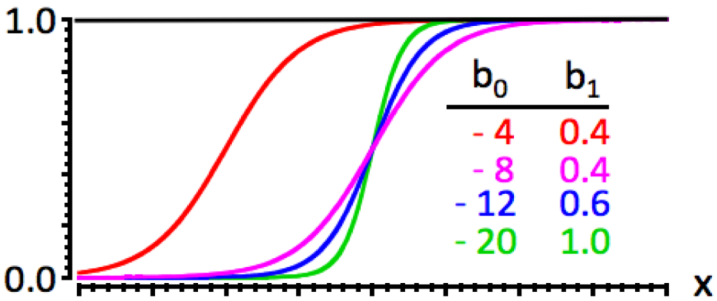
which is $p = \frac{odds}{1+odds}$ and where

$b_1 = 0$ implies $P(Y = 1)$ is the same at each level of x

$b_1 > 0$ implies $P(Y = 1)$ increases as x increases

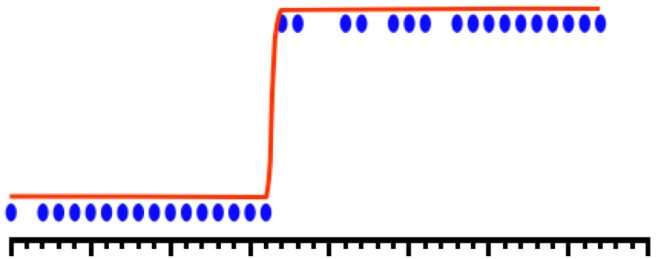
$b_1 < 0$ implies $P(Y = 1)$ decreases as x increases

Different b's

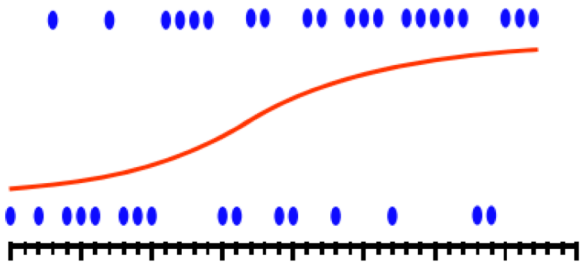


$$P(y=1)=\frac{e^{(b_0+b_1x_1)}}{1+e^{(b_0+b_1x_1)}}$$

Data with a sharp cut-off point should have a large value of b_1



Data with a lengthy transition should have a small value of b_1



Case Study 1 - Health

- ESR is the rate at which red blood cells (erythrocytes) settle out of suspension in blood plasma, when measured under standard conditions. If ESR increases when the level of certain proteins in the blood plasma rise in association with conditions such as in chronic infections and malignant diseases. Now the actual value of ESR is not of great importance, rather if it is less than 20 units which indicates a healthy individual.

➢ The question of interest is whether there is any association between the probability of an ESR greater than 20 units and the levels of a plasma proteins, fibrinogen.

Download the data file: **plasma** from course website.

Logistic Analysis

Firstly, let's look at the conditional probability density plots of the response variable given fibrinogen.

In **R,** construct this plot and comment on a feature. The code is below:

```
> layout(matrix(1:2, ncol = 2))
> cdplot(ESR ~ fibrinogen, data = plasma)
```

- It appears that higher levels of protein leads to ESR being above 20.

- We can now fit a logistic model to the data using the **glm** function!

Case Study 1: Fit the model

The code to fit the model is given by:

```
> plasma_glm_1 <- glm(ESR ~ fibrinogen, data = plasma, + family = binomial())
```

The formula implicitly defines a parameter for the global mean.

A description and summary of the model can be produced as well now by using the code:

```
> summary(plasma_glm_1)
```

and we can construct a 95% confidence interval for the parameter (coefficient) for *fibrinogen* by using the code

```
> confint(plasma_glm_1, parm = "fibrinogen")
```

Summary

> summary(plasma_glm_1)

```
R> summary(plasma_glm_1)

Call:
glm(formula = ESR ~ fibrinogen, family = binomial(), data = plasma)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-0.930  -0.540  -0.438  -0.336   2.479

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   -6.845      2.770   -2.47    0.013
fibrinogen     1.827      0.901    2.03    0.043

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 30.885  on 31  degrees of freedom
Residual deviance: 24.840  on 30  degrees of freedom
AIC: 28.84

Number of Fisher Scoring iterations: 5
```

> confint(plasma_glm_1, parm = "fibrinogen")

```
R> confint(plasma_glm_1, parm = "fibrinogen")
 2.5 % 97.5 %
 0.339  3.998
```

Interpretations

We see that the regression coefficient for *fibrinogen* is significant at the 5% level.

An increase of one unit in this variable increases the log-odds in favor of an ESR value greater than 20 by an estimated 1.83 with 95% confidence interval between 0.339 and 3.998.

These values are more useful converted to **Odds** or **Probabilities,** if we convert the coeffienent of *fibrinogen* to an odd by exponentiating the estimate

R> exp(coef(plasma_glm_1)["fibrinogen"])

Odds for 1 increase in fibrinogen = 6.22

and a confidence interval for this estimate of

R> exp(confint(plasma_glm_1, parm = "fibrinogen"))

CI = (1.4, 54.5)

The confidence interval is very wide because there are few observations overall and very few where the ESR value is greater than 20.

Conclusions & Predictions

Thus it seems likely that increased values of fibrinogen lead to a greater probability of an ESR value greater than 20.

- **Predictions**

We commonly want to make predictions based on certain values of *fibrinogen* this can be done by using the logit function with the coefficients:

$$b_0 = -6.845 \ \ and \ \ b_1 = 1.827.$$

Otherwise, we can use the **R** code to predict probabilities by

```
prob <- predict(plasma_glm_1, type = "response")
```

This obtains all the probabilities for the values imputed for *fibrinogen*.

Prediction

- If the average value of *fibrinogen* is 3 units, what is the probability of ESR being above 20 units.

$$P(Y=1)=\frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}}=\frac{e^{-6.845+1.827\times3}}{1+e^{-6.845+1.827\times3}}=0.2036$$

- Therefore there is a 20.36% chance that ESR is above 20 units when on average fibrinogen level is 3 units.

- Calculate some for yourself ☺

Case Study 2 – Agree or Disagree

In a survey carried out in 1970's each respondent was asked if he or she agreed or disagreed with a political statement.

The question of interest is whether the responses of men and women differ, and how if the number of years of education affected the response.

We will focus on the responses from women today and in turn, does the number of years of education affect their response?

The data (**womensrole**) is clumped together and needs to be pulled apart to fit a model.

```
> data("womensrole", package = "HSAUR2")
```

The individual observations have been grouped into counts of numbers of agreements and disagreements for the two explanatory variables, gender and education. To fit a logistic regression model to such grouped data using the glm function we need to specify the number of agreements and disagreements as a two-column matrix on the lefthand side of the model formula.

We first fit a model that includes the two explanatory variables using the code

```
> fm1 <- cbind(agree, disagree) ~ education
```

```
> womensrole_glm_1 <- glm(fm1, data = womensrole, family = binomial())
```

We can look at the summary:

> summary(womensrole_glm_1)

```
Call:
glm(formula = fm1, family = binomial(), data = womensrole)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-2.74737  -0.88358  -0.07487   0.86240   3.10504

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.50334    0.17843   14.03   <2e-16 ***
education   -0.27065    0.01541  -17.56   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 451.722  on 40  degrees of freedom
Residual deviance:  64.025  on 39  degrees of freedom
AIC: 206.09

Number of Fisher Scoring iterations: 4
```

Interpretation

- From the summary output, it appears that education has a highly significant part to play in predicting whether a respondent will agree with the statement read to them.

- As years of education increase the probability of agreeing with the statement declines.

Advanced (Workshop only): Try and construct a model for the males and determine if the gender is important.

Conclusion: The respondent's gender is apparently unimportant.

Conclusions & Predictions

Thus it seems likely that increased values of *years of education* lead to a decrease in probability of an agreeing with the statement.

- **Predictions**

We commonly want to make predictions based on certain values of *the number of education years* this can be done by using the logit function with the coefficients:

$$b_0 = 2.50334 \ \ and \ \ b_1 = -0.27065.$$

Otherwise, we can use the **R** code to predict probabilities by

```
prob <- predict(womensrole_glm_1, type = "response")
```

This obtains all the probabilities for the values imputed for the *number years of education*.

Prediction

- If the average value of *Level (number of years) of Education* is 5 units, what is the probability of *agree?*

$$P(Y=1) = \frac{e^{b_0+b_1x}}{1+e^{b_0+b_1x}} = \frac{e^{2.50334-0.27065\times5}}{1+e^{2.50334-0.27065\times5}} = 0.7603$$

- Therefore there is a % chance of agreeing with the statement when on average the number of years of education (level) is 5 years.

- Calculate some for yourself ☺

Week 11 is over!

Next week is our final week and we will look at:

Monte Carlo Methods!!!

Workshop Question!

Note: This question is not recorded and is a hands on question which replaces the tutorial!

Challenger disaster, Part I. On January 28, 1986, a routine launch was anticipated for the Challenger space shuttle. Seventy-three seconds into the flight, disaster happened: the shuttle broke apart, killing all seven crew members on board. An investigation into the cause of the disaster focused on a critical seal called an O-ring, and it is believed that damage to these O-rings during a shuttle launch may be related to the ambient temperature during the launch. The table below summarizes observational data on O-rings for 23 shuttle missions, where the mission order is based on the temperature at the time of the launch. Temp gives the temperature in Fahrenheit, Damaged represents the number of damaged O-rings, and Undamaged represents the number of O-rings that were not damaged.

The Data

Shuttle Mission	1	2	3	4	5	6	7	8	9	10	11	12
Temperature	53	57	58	63	66	67	67	67	68	69	70	70
Damaged	5	1	1	1	0	0	0	0	0	0	1	0
Undamaged	1	5	5	5	6	6	6	6	6	6	5	6

Shuttle Mission	13	14	15	16	17	18	19	20	21	22	23
Temperature	70	70	72	73	75	75	76	76	78	79	81
Damaged	1	0	0	0	0	1	0	0	0	0	0
Undamaged	5	6	6	6	6	5	6	6	6	6	6

Questions

(a) Each column of the table above represents a different shuttle mission. Examine these data and describe what you observe with respect to the relationship between temperatures and damaged O-rings.

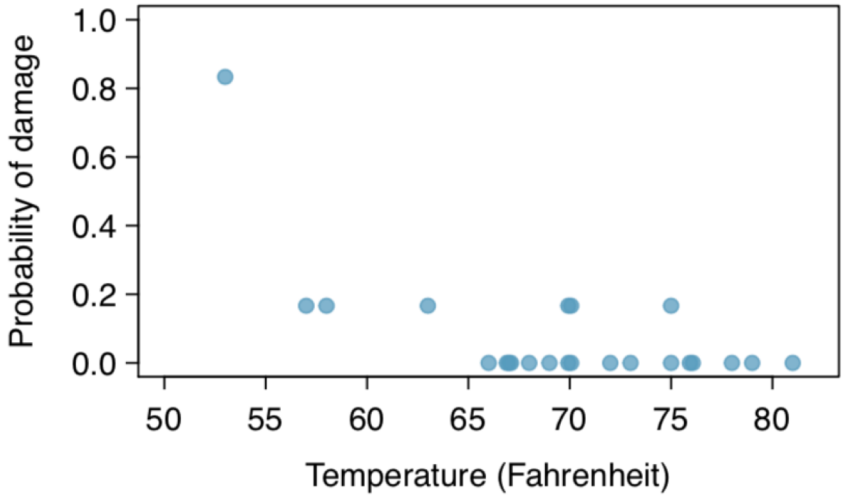
(b) Failures have been coded as 1 for a damaged O-ring and 0 for an undamaged O-ring, and a logistic regression model was fit to these data. A summary of this model is given below. Describe the key components of this summary table in words.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	11.6630	3.2963	3.54	0.0004
Temperature	-0.2162	0.0532	-4.07	0.0000

(c) Write out the logistic model using the point estimates of the model parameters. Based on the model, do you think concerns regarding O-rings are justified? Explain.

Workshop Question

Challenger disaster, Part II. O-rings that were identified as a plausible explanation for the breakup of the Challenger space shuttle 73 seconds into take off in 1986. The investigation found that the ambient temperature at the time of the shuttle launch was closely related to the damage of O-rings, which are a critical component of the shuttle. See this earlier exercise if you would like to browse the original data.



Workshop (Question)

The data provided in the previous exercise are shown in the plot. The logistic model fit to these data may be written as

$$\log\left(\frac{\hat{p}}{1-\hat{p}}\right) = 11.6630 - 0.2162 \times Temperature$$

where $\hat{p}$ is the model-estimated probability that an O-ring will become damaged.

(a) Use the model to calculate the probability that an O-ring will become damaged at each of the following ambient temperatures: 51, 53, and 55 degrees Fahrenheit. The model-estimated probabilities for several additional ambient temperatures are provided below, where subscripts indicate the temperature:

$$\hat{p}_{57} = 0.341 \qquad \hat{p}_{59} = 0.251 \qquad \hat{p}_{61} = 0.179 \qquad \hat{p}_{63} = 0.124$$
$$\hat{p}_{65} = 0.084 \qquad \hat{p}_{67} = 0.056 \qquad \hat{p}_{69} = 0.037 \qquad \hat{p}_{71} = 0.024$$

Last Slide

(b) Describe any concerns you may have regarding applying logistic regression in this application, and note any assumptions that are required to accept the model's validity.