Ch 9: Qualitative response regression models

Outline

1. Background and overview
2. Regressing probabilities
3. The linear probability model (don't use this!)
4. The logit and probit models
5. Worked numerical example

1.1 Summary of the last lecture

- Discussed dummy variable regression models
- In this case the Y-variable remains a continuous measurement but you may have qualitative or categorical X-variables
- In this lecture we discuss regression models where the Y-variable is categorical (often best interpreted as a probability of being in one of two categories)
- Examples might include (yes/no, present/absent, Islamic bank/non-Islamic bank, high score/not a high score etc.)
- These probability models are inherently more complicated

1.2 Statistical overview

• Classical regression models are great but run into problems very quickly

  - e.g. simple surveys can generate surprisingly complex data

• This complicated data necessitates more complicated statistical models – generalized linear models

• I studied generalized linear models during my BSc and MSc. However, it is possible to do an entire statistics degree without covering generalized linear models.

1.3 Where the subject goes ...

• Have written a textbook chapter on Generalized Linear Models (Bingham and Fry, 2010)
• I have used Generalized Linear Models to ...
  - Model complicated survey data
  - Model customer satisfaction data in an industrial problem (see Chapter 10)
  - To model the effect of the academic journal on the perceived quality of published research papers (serious practical problem in academia)
• **Past dissertation students of mine of also successfully used related models**

2.1 Regressing or estimating probabilities

- **Basic problem is to calculate the probability of being in certain categories and how this depends on the X-variables**

2.2 Qualitative dependent variable models

• Qualitative dependent variable models are regression models in which the Y-variable can be categorised as e.g. yes/no, present/absent etc.
• Some of the terminology used can be confusing. In my experience the terms **successes** and **failures** are often used in R and R coding
• In economics and finance these models are often known as dichotomous/dummy variable regression models. I think the term probability regression model gives a clearer idea of what you are trying to do
• Illustrative industrial example: How does the probability that an airplane component fails depend on the applied load?

2.3 How does the probability of component failure depend on the applied load?

Load	Tested	Failures
2500	50	10
2700	70	17
2900	100	30
3100	60	21
3300	40	18
3500	85	43
3700	90	54
3900	50	33
4100	80	60
4300	65	51

2.4 R code to read in the data

```
load<-c(2500, 2700, 2900, 3100, 3300, 3500, 3700,
3900, 4100, 4300)
tested<-c(50, 70, 100, 60, 40, 85, 90, 50, 80, 65)
failures<-c(10, 17, 30, 21, 18, 43, 54, 33, 60, 51)
```

- **Calculate the proportion of successful components**

```
probsuccess<-1-failures/tested
```

- **Extra code needed for fitting logit and probit models**
- **Need a column of successes and failures listed side by side.**

```
successes<-tested-failures
fasteners<-cbind(successes, failures)
```

3.1 Linear probability model

- Consider the following simple model

$$\text{Probability of success} = \beta_1 + \beta_2 \text{Load} + u_i, \tag{1}$$

where the u_i in equation (1) is a normally distributed error term

- This model is known as the linear probability model for two reasons

1. It is understood that the left hand side of the equation refers to a probability
2. The model is just a linear regression model that we have seen in previous lectures

3.2 Interpreting the model

$$\text{Probability of success} = \beta_1 + \beta_2 \text{Load} + u_i$$

- It is natural to expect that the component is more likely to fail as the load applies increases
- Equivalently we might anticipate that the success probability is likely to decrease as the load applied increases
- This means it is natural to anticipate finding $\beta_2 < 0$ in the above

3.3 The linear probability model

- The fitted value of the regression model in (1) is to be interpreted as the probability that a component succeeds given the load that is applied
- Since this is a probability we must have that

$$0 \leq \beta_1 + \beta_2 \text{Load} + u_i \leq 1. \tag{2}$$

- However, there is no guarantee that equation (2) holds unless additional constraints are imposed
- In particular we might anticipate that very low values of the load and (respectively very high values of the load) may result in estimated probabilities being greater than 1 (respectively less than zero)

3.4 The linear probability model

- The fitted value of the regression model in (1) is to be interpreted as the probability that a component succeeds given the load that is applied
- Since this is a probability we must have that

$$0 \leq \beta_1 + \beta_2 \text{Load} + u_i \leq 1. \tag{3}$$

- However, there is no guarantee that equation (3) holds unless additional constraints are imposed
- We can also show that the linear probability model violates two of the standard regression modelling assumptions
  1. Non-normality of the disturbances
  2. Heteroscedasticity
- **Whilst true these points are a little artificial as anybody sensible would feel uneasy about estimated probabilities potentially being either less than 0 or greater than 1!**

3.5 Non-normality of u_i

- Since probabilities ≥ 0 we must have

$$\beta_1 + \beta_2 \text{Load} + u_i \geq 0; \;\; u_i \geq -\beta_1 - \beta_2 \text{Load}$$

- Since probabilities ≤ 1 we must have

$$\beta_1 + \beta_2 \text{Load} + u_i \leq 1; \;\; u_i \leq 1 - \beta_1 - \beta_2 \text{Load}$$

- The conclusion

$$-\beta_1 - \beta_2 \text{Load} \leq u_i \leq 1 - \beta_1 - \beta_2 \text{Load},$$

shows that u_i can only take certain bounded values and so cannot be normally distributed

3.6 Non-normality of u_i

- Since probabilities ≥ 0 we must have

$$\beta_1 + \beta_2 \text{Load} + u_i \geq 0; \;\; u_i \geq -\beta_1 - \beta_2 \text{Load}$$

- Since probabilities ≤ 1 we must have

$$\beta_1 + \beta_2 \text{Load} + u_i \leq 1; \;\; u_i \leq 1 - \beta_1 - \beta_2 \text{Load}$$

- The conclusion

$$-\beta_1 - \beta_2 \text{Load} \leq u_i \leq 1 - \beta_1 - \beta_2 \text{Load},$$

shows that u_i can only take certain bounded values
- The bounds depend on the value of the load
- The distribution of the u_i and hence the variance of the u_i therefore depend on the value of the load
- Since the variance of the u_i depends on the load the model is heteroscedastic

3.7 Summary of the linear probability model

• The model is intuitively appealing combining probability estimates with simple linear regression
• The model has been misguidedly popularised by classic econometric texts like Gujarati and Porter (2009)
• Whilst the linear probability model may give sensible answers in the middle of the sample extreme X-values are liable to lead to probability estimates either less than zero are greater than 1
• To ensure sensible probability estimates that lie between 0 and 1 the logit and probit models should be used

4.1 Mathematical motivation

- Linear regression

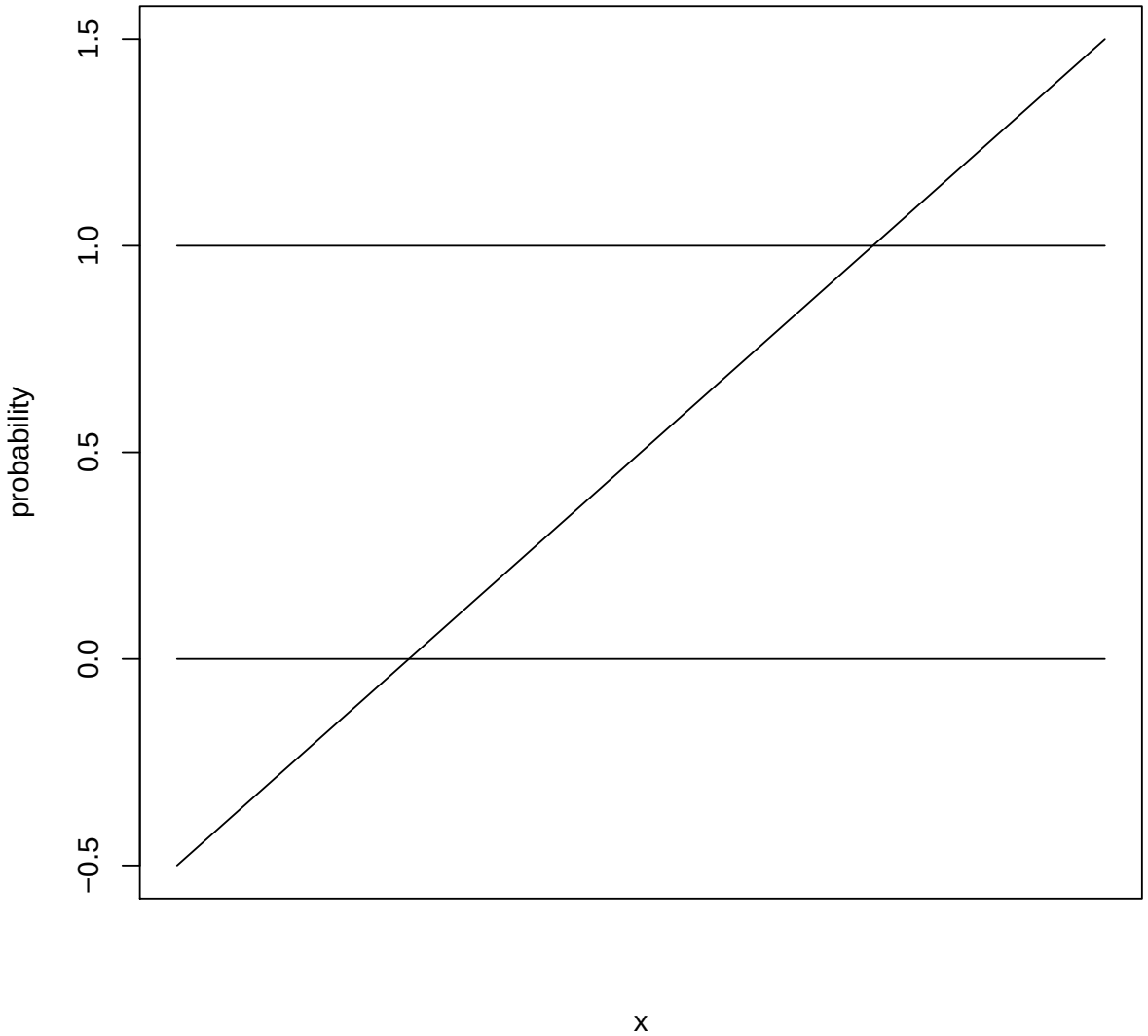
$$Y = \beta_0 + \beta_1 X_1 + ... + \beta_p X_P. \tag{4}$$

- Probability regression

$$f(\text{Probability of } Y) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_P. \tag{5}$$

- Both these equations can take either positive or negative values depending on the values of x and β
- **However, genuine probabilities must lie between 0 and 1**
- The function $f(\cdot)$ in equation (5) is known as the **link function**
- $f(\cdot)$ literally **links** the regression part of the model to the probability calculation

4.2 Problems using regression to estimate probabilities



4.3 Fixing regression ...

• Keep the regression bit of the model -– don't throw the baby out with the bathwater!

$$\beta_0 + \beta_1 X_1 + ... + \beta_p X_P$$

• Since the above equation takes values in $(-\infty, \infty)$ we need a link function $f(\cdot)$ to "squash" equation the equation so that we get sensible estimates of probabilities

$$f(\pi) = \beta_0 + \beta_1 X_1 + ... + \beta_p X_P$$

• If done in the right way much of the interpretation stays quite similar to standard regression
• e.g. if $\beta_i > 0$ as X_i increases, the probability increases
• e.g. if $\beta_i < 0$ as X_i increases, the probability decreases

4.4 Logistic regression

- "Regressing or explaining probabilities"
- How do you squash probabilities to lie between 0 and 1?
- Suppose you have a logistic regression model

$$\ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_i$$

- The probability can be calculated as

$$\begin{aligned}
\frac{p_i}{1-p_i} &= \exp(\beta_0 + \beta_1 x_i) \\
p_i &= (1-p_i)\exp(\beta_0 + \beta_1 x_i) \\
p_i &= \frac{\exp(\beta_0 + \beta_1 x_i)}{1+\exp(\beta_0 + \beta_1 x_i)}
\end{aligned}$$

4.5 Probit regression

- "Regressing or explaining probabilities"
- How do you squash probabilities to lie between 0 and 1?
- Suppose you have a probit regression model

$$Z^{-1}(p_i) = \beta_0 + \beta_1 x_i,$$

where Z^{-1} denotes the inverse CDF of a normal distribution

- Using Tables the probability can be calculated as

$$\pi_i = Z(\beta_0 + \beta_1 x_i),$$

where $Z(\cdot)$ denotes the normal distribution value from tables

4.6 R code for LMs and GLMS

• To fit probability regressions in R you need to be aware of some background mathematics and the data structure

Background mathematics.

• Technically, speaking logistic and probit regression are examples of **Generalised Linear Models** (Bingham and Fry, 2010).

• This means that the R command used is `glm` for **Generalised Linear Model**

• By contrast regression models are known as a **Linear Models** (Bingham and Fry, 2010).

• This means that the R command used for regression is `lm` for **Linear Model**

4.7 R code for GLMs

- In addition to the above you also have to specify the family of distributions used (binomial) to tell R that you are regressing probabilities
- You also have to tell R what link function you are using. The default for binomial generalised linear models is the logit. If you don't specify the link function R will fit a logistic regression model
- All these things are essentially bits of R syntax but it is important to be aware that these do have a mathematical and statistical origin and underpinning

4.8 Data structure for probability models in R

• **To fit a binomial glm in R you need the data organised in columns of successes and failures**

• The R command needed to do this is `cbind` which has the effect of binding the required counts of successes and failures together

• For our data example in R use

```
tested<-c(50, 70, 100, 60, 40, 85, 90, 50, 80, 65)
failures<-c(10, 17, 30, 21, 18, 43, 54, 33, 60, 51)
successes<-tested-failures
fasteners<-cbind(successes, failures)
```

4.9 Fitting binomial generalised linear models in R

- The basic set of commands works as follows
- **Compute the model**

```
a.glm< −glm(fasteners∼load, family=binomial)
b.glm< −glm(fasteners∼load,
family=binomial(link=probit))
```

- **Summarise the results**

```
summary(a.glm)
summary(b.glm)
```

- **These models can serve as a cross-check of each other in applications. Should expect to have similar models giving you similar interpretations and numerically similar estimates**

5.1 Worked numerical example: Aircraft fasteners

Load	Tested	Failures
2500	50	10
2700	70	17
2900	100	30
3100	60	21
3300	40	18
3500	85	43
3700	90	54
3900	50	33
4100	80	60
4300	65	51

1. Fit linear probability, logistic and probit models to this data and interpret the results
2. What load would cause 50% of the components to fail?
3. What is the success probability if loads of 1999kg and 4891kg are applied?

5.2 Linear probability model in R

- **With the data in the right format you can now fit a linear probability model using the `lm` command as follows:**

```
a.lm<-lm(probsuccess~load)
summary(a.lm)
```

- **Tells the computer to do the work and then summarise the results in a separate step**

5.3 Binomial GLM in R

- **With the data in the right format you can now fit a binomial glm as follows**

```
a.glm<-glm(fasteners~load, family=binomial)
summary(a.glm)
b.glm<-glm(fasteners~load,
family=binomial(link=probit))
summary(b.glm)
```

- **Tells the computer to do the work and then summarise the results in a separate step**

5.4 Interpretation of results

- All statistical software packages produce redundant information
- **For the purposes of this course I could expect you to calculate a t-statistic**

1. Is the variable in question significant?
2. If it is significant if the variable increases does the probability increase or decrease (is the sign of the random variable positive or negative?)

- **Liable to actually be a lot easier than it might seem at first**

5.5 Linear probability model

```
 Estimate Std.  Error t value Pr(>|t|)
(Intercept) 1.692e+00 3.543e-02 47.76 4.08e-11 ***
load -3.460e-04 1.027e-05 -33.68 6.60e-10 ***
```

- Interpret these results as

1. Load has a significant impact upon the probability of success $p < 6.60 \times 10^{-10}$
2. As load increases the probability of success decreases. A 1kg increase in the load means the success probability decreases by around 3.46×10^{-4}

- The hand calculation of the z or t-statistic (similar to an exam-type question) would be

$$t = \frac{|\text{Estimate}|}{\text{Standard Error}} = \frac{|-3.460 \times 10^{-4}|}{1.027 \times 10^{-5}} = 33.69036 > 2.0$$

Therefore, $p < 0.05$

- **Note here that the two inequality signs point the different way if you have done this correctly**

5.6 Logistic regression model

```
Estimate Std.  Error z value Pr(>|z|)
(Intercept) 5.3397115 0.5456932 9.785 <2e-16 ***
load -0.0015484 0.0001575 -9.829 <2e-16 ***
```

- Interpret these results as

1. Load has a significant impact upon the probability of success $p < 2e - 16$
2. As load increases the probability of success decreases

- The hand calculation of the z or t-statistic (similar to an exam-type question) would be

$$t = \frac{|\text{Estimate}|}{\text{Standard Error}} = \frac{|0.0015484|}{0.0001575} = 9.831111 > 2.0$$

Therefore, $p < 0.05$

- **Note here that the two inequality signs point the different way if you have done this correctly**

5.7 Extra interpretation

• Since the coefficient of load is negative and statistically significant, as load increases the probability of success decreases

5.8 Probit regression model

```
Coefficients:  Estimate Std.  Error z value Pr(>|z|)
(Intercept) 3.271e+00 3.213e-01 10.18 <2e-16 ***
load -9.488e-04 9.281e-05 -10.22 <2e-16 ***
```

- Interpret these results as

1. Load has a significant impact upon the probability of success $p < 2e - 16$
2. As load increases the probability of success decreases

- The hand calculation of the t-statistic (similar to an exam-type question) would be

$$t = \frac{|\text{Estimate}|}{\text{Standard Error}} = \frac{|-9.488e - 04|}{9.281e - 05} = 10.22303631 > 2.0$$

Therefore, $p < 0.05$

- **Note here that the two inequality signs point the different way if you have done this correctly**

5.9 Extra interpretation

Exam

• Since the coefficient of load is negative and statistically significant, as load increases the probability of success decreases

Outside of the exam

• Look for two different sources of information telling you the same thing ("cross-checks")

• Always worth cross-checking numerical information like this with a graph

• Note that the logistic and probit models give you similar answers in this example and this should always usually be the case

5.10 Plotting the data as a sanity check

```
proportion=failed/tested
plot(load, proportion)
```

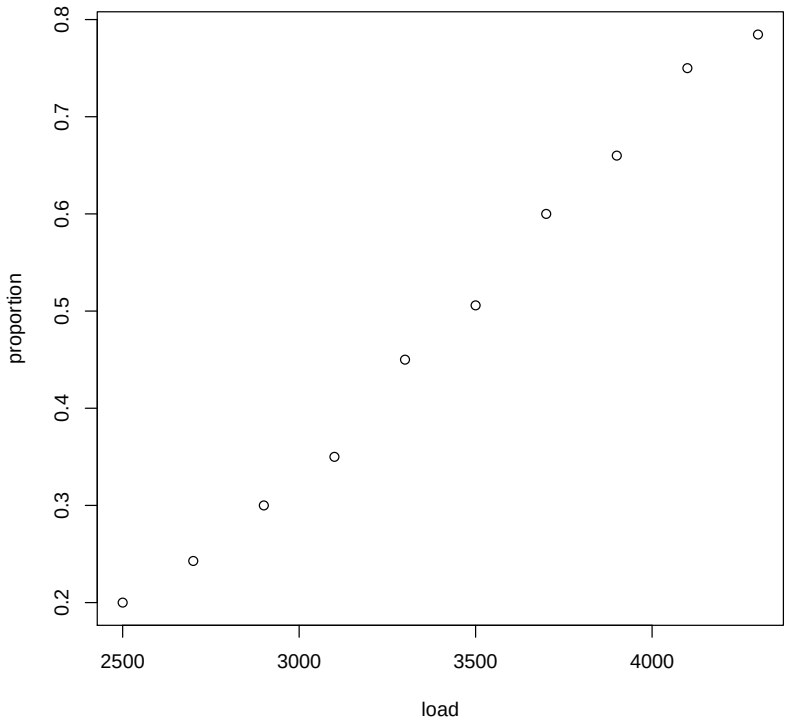


Figure: Suggests the failure rate increases as load increases

5.11 Extra calculation example I: linear probability model

- For what value of the load is there a 50% failure rate (implies a 50% success rate)?
- From the above the fitted model is

$$\text{Success probability} = 1.692 - 3.460 \times 10^{-4} \text{load}$$

- Set

$$\begin{aligned} 0.5 &= 1.692 - 3.46 \times 10^{-4} \text{load} \\ \text{load} &= \frac{0.5 - 1.692}{-3.46 \times 10^{-4}} = 3445.087 \end{aligned}$$

5.12 Extra calculation example I: Logistic model

- For what value of the load is there a 50% failure rate (implies a 50% success rate)?

```
 Estimate Std.  Error z value Pr(>|z|)
(Intercept) 5.3397115 0.5456932 9.785 <2e-16 ***
load -0.0015484 0.0001575 -9.829 <2e-16 ***
```

- From the output the fitted equation is

$$\ln\left(\frac{p}{1-p}\right) = 5.3397115 - 0.0015484 \times \text{load}$$

- Putting in $p = 0.5$ gives

$$0 = 5.3397115 - 0.0015484\text{load};\ \text{load} = \frac{5.3397115}{0.0015484} = 3448.535kg$$

5.13 Extra calculation example I: Probit model

- For what value of the load is there a 50% failure rate (implies a 50% success rate)?

```
 Estimate Std.  Error z value Pr(>|z|)
(Intercept) 3.271e+00 3.213e-01 10.18 <2e-16 ***
load -9.488e-04 9.281e-05 -10.22 <2e-16 ***
```

- From the output the fitted equation is

$$Z^{-1}(p) = 3.271 - 0.0009488 \times \text{load}$$

- From tables since $Z(0) = 0.5$, $Z^{-1}(0.5) = 0$ and so

$$0 = 3.271 - 0.0009488\text{load}; \ \text{load} = \frac{3.271}{0.0009488} = 3447.513kg$$

5.14 Extra calculation example II: linear probability model

- **What is the success probability if loads of 1999kg and 4891kg are applied?**
- **Solution**

$$\text{Probability} = 1.692 - 3.46 \times 10^{-4}(1999) = 1.000346$$

$$\text{Probability} = 1.692 - 3.46 \times 10^{-4}(4891) = -0.000286$$

- **Do you think these are sensible probability estimates?!**

5.15 Extra calculation example II: logit model

- **What is the success probability if loads of 1999kg and 4891kg are applied?**

$$\text{Probability} = \frac{\exp(5.3397115 - (0.0015484)(1999))}{1 + \exp(5.3397115 - (0.0015484)(1999))} = 0.9041716$$

$$\text{Probability} = \frac{\exp(5.3397115 - (0.0015484)(4891))}{1 + \exp(5.3397115 - (0.0015484)(4891))} = 0.09678113$$

- **These are more sensible probability estimates!**

5.16 Extra calculation example II: probit model

• **What is the success probability if loads of 1999kg and 4891kg are applied?**

• In R the function $Z^{-1}(\cdot)$ is calculated using the command `pnorm(·)`

$$\begin{aligned} \text{Probability} &= Z^{-1}(3.271 - 9.488 \times 10^{-4} 1999) \\ &= Z^{-1}(1.374349) = 0.9153333 \end{aligned}$$

$$\begin{aligned} \text{Probability} &= Z^{-1}(3.271 - 9.488 \times 10^{-4} 4891) \\ &= Z^{-1}(-1.369581) = 0.08540887 \end{aligned}$$

• **These are again more sensible probability estimates and should roughly match the values obtained from the logit model**

5.17 Summary

- The linear probability model is surprisingly popular but not really a viable model (so don't use this!)
- The logit and probit model are more viable ways of regressing and estimating probabilities (so do use these!)
- In our numerical example the linear probability model gives reasonable probability estimates in the middle of the sample
- The linear probability model is likely to provide poor estimates of probabilities either close to zero or close to one
- In practical examples would usually expect to see similar numerical estimates from viable models (e.g. the logit and probit models here). This gives a sense of robustness to the results obtained and their interpretation.