GLM I - Variations on Logistic Regression

October 28th, 2020

Latent Variables

Set up

- Suppose that students answer questions on a test and that a specific student has an aptitude T . A particular question might have difficulty d and the student will get the answer correct only if $T > d$. Now if we consider d fixed and T as a random variable then the probability that the student gets the answer wrong is:

$$p = P(T \leq d) = F(d)$$

- This is the distribution function
- How do we define the density function?

Function Definitions

- Cumulative distribution function as a logistic distribution
- The logistic distribution is actually a logistic function on the cumulative distribution, not on the density function
- Normally when we say this is a normal distribution, we're talking about the density function.
- This formula below is actually called the logistic because of what it looks like in the cumulative distribution.

$$F(y) = \frac{exp(y-\mu)/\sigma}{1 + exp(y-\mu)/\sigma}$$

- What is the relationship between the density function and the cumulative distribution function?
  - One is the right derivative of the other. The density function is the derivate of the cumulative distribution function or the integral from negative infinity to a value is a the cumulative distribution function.
  - The right derivative is the derivate taken involving sequences only from the right
  - Cumulative distribution function can be discontinues, so they aren't necessarily differentiable everywhere but they are differentiable from one side
- You can rewrite this if you combine this and the previous equation, you can rewrite it so the log(p) is equal to the formula below.
- You can rewrite it where beta0 = -mu/sigma and beta1 = 1/sigma

$$logit(p) = -\mu/\sigma + d/\sigma logit(p) = \beta_0 + d\beta_1$$

- This is now just a regular logistic regression model!

Graph

- Assume difficulty (d) = 1
- Latent Variable T has mean = -1 and sigma = 1

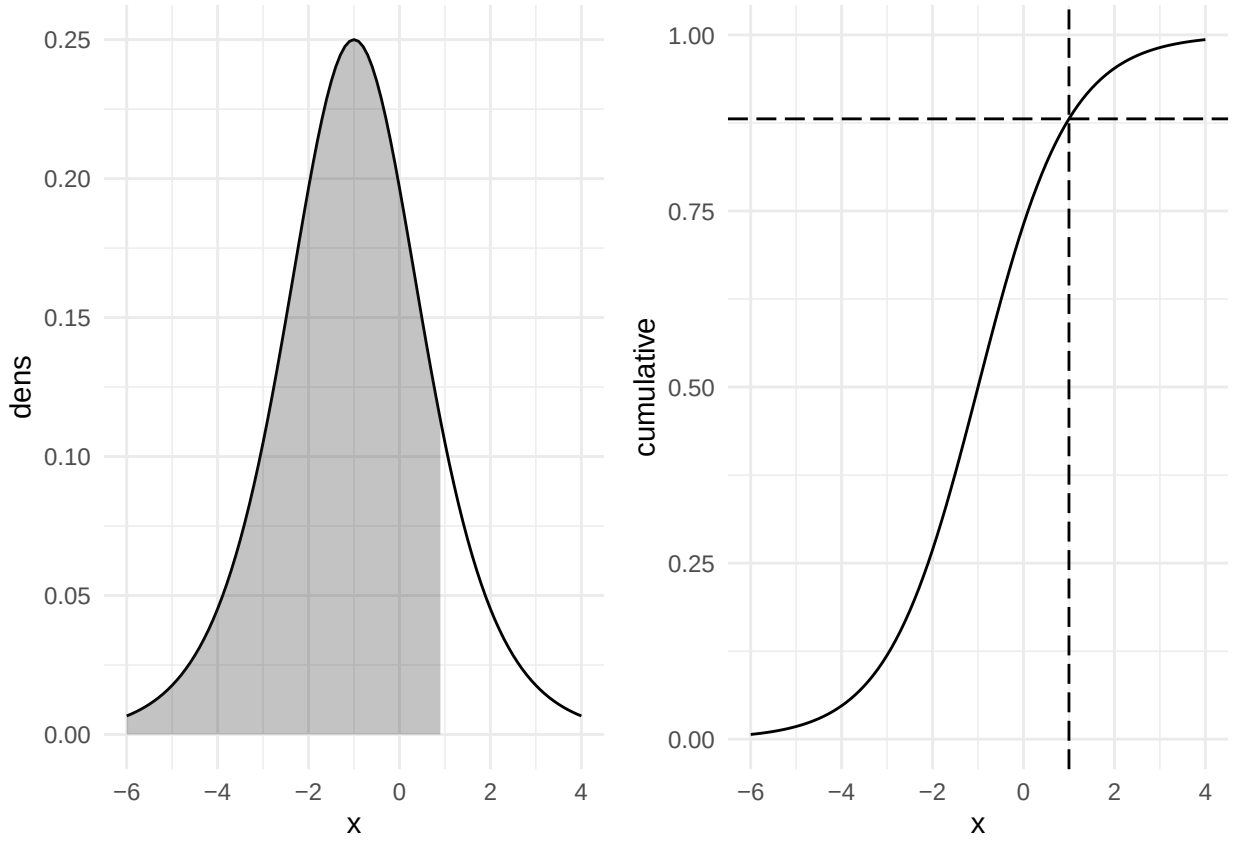
- Note that the graph on the left appears normal

```
test_data <- tibble(
  x = seq(-6, 4, 0.1),
  dens = dlogis(x, location = -1, scale = 1),
  cumulative = plogis(x, location = -1, scale = 1)
)

p1 <- ggplot(test_data, aes( x = x, y = dens)) +
  geom_line() +
  geom_area(data = filter(test_data, x < 1), alpha = .3)

p2 <- ggplot(test_data, aes(x = x, y = cumulative)) +
  geom_line() +
  geom_hline(yintercept = plogis(1, location = -1, scale = 1), linetype = "longdash") +
  geom_vline(xintercept = 1, linetype = "longdash")

grid.arrange(p1, p2, ncol = 2)
```



- The left graph is the density function. the shaded area is everything that is to the left of d=1.
  - It's basically the probability the person gets its wrong, that is the area underneath the curve
- An easier way of doing that is looking at the right graph, which is the cumulative distribution function. This gives you the area underneath the curve as the function itself. So the area under the curve of the left graph is whatever the horizonal dotted line value is, let's say .83. The vertical line is the d = 1. The area underneath the curve is the intersection point.
- This is showing the relationship between the density function and the cumulative distribution function.

Link Functions

Link Function Requirements

- In a lot of cases, like before, what we're doing is we're using the logistic link (logit link) when we're doing binomial regression
- But there are other link functions that are available

Only have two requirements, - Bounds the probability between 0 and 1 - Monotonic, which means that they don't have a curve up and down in them. Once it starts raising it keeps raising

Other link functions in `glm`

- Probit - where variable is normally distributed: $\eta = \Phi^{-1}(p)$
- Complementary Log-log: $\eta = log(-log(1-p))$
- Cauchit: $\eta = tan^{-1}(\pi(p-1/2))$

Show example with `bliss` data

- When we talking about it earlier, we said that there's not much you can do to change them unless you have a good reason to. But it might not always matter
- Now we're going to talk about a situation where it might matter
- The bliss data looks at the effect of poison on bugs. Trying to say as the poison concentration goes up from 0-4 whatever those unites are, how many bugs are alive and how many are dead. Looks like 30 bugs per trail. It follows a nice pattern whereas the concentration of the poison goes up its more effective of killing the bugs
- Look at different concentration of a drug

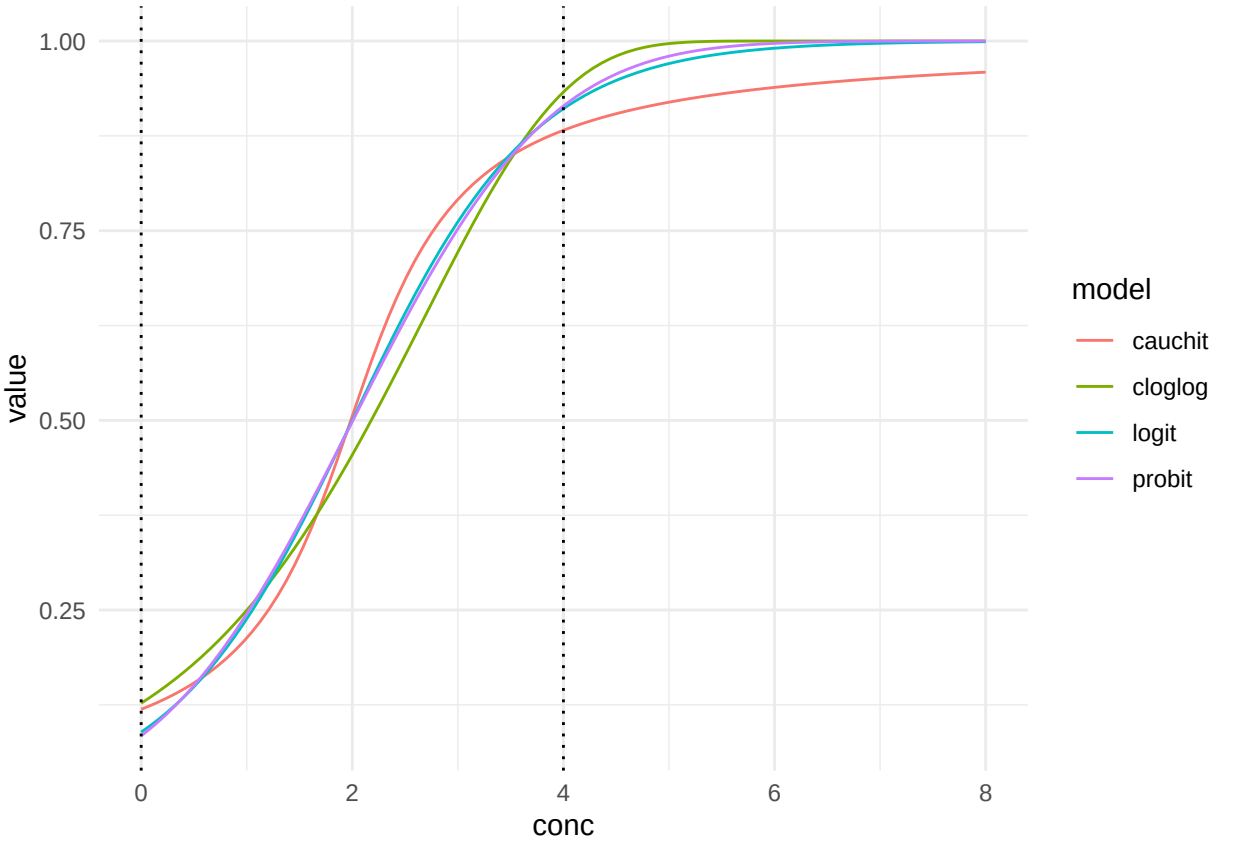
```
bliss
```

dead	alive	conc
2	28	0
8	22	1
15	15	2
23	7	3
27	3	4

```
# fit a bunch of models, with the 4 different link fns that are here
mlogit <- glm(cbind(dead,alive) ~ conc, family=binomial, data=bliss)
mprobit <- glm(cbind(dead,alive) ~ conc, family=binomial(link=probit), data=bliss)
mcloglog <- glm(cbind(dead,alive) ~ conc, family=binomial(link=cloglog), data=bliss)
mcauchit <- glm(cbind(dead,alive) ~ conc, family=binomial(link=cauchit), data=bliss)
```

Look at Predictions

- Just going to look at the predicted response for each value of the concentration



- Here are the 4 different models, and there response over the different values of concentration. From 0-4 is what we have for the data. What we see here is that the values here in the central on the left of 4, the chance of that bug dies is pretty similar for all of the 4 logits curves.
- However once you start to look at extreme values, you see a little bit more separation. You have the Cauchit is a good deal lower than the other 3, which kind of quickly converge at 100%.
- The issues that they're showing that it seems to separate at the tails only becomes stronger when the variables don't actually give you responses in between 0 and 100.
- One of the nice things here is that if death is the outcome we're looking for, the first row is close to 1% where the last row is close to 99% death. Here were have a very good range of the data, around 90%. That actually makes the link functions move a lot closer together. But the tails of the graphs often diverge more when you look at these different link functions.

Takeaways

- The tails are different between these link functions making them important to look at when discussing poisons
- There are substances whose harmful effects only become apparent at large dosages where the observed probabilities are sufficiently larger than zero to become estimable without immense sample sizes.
- Asbestos is a good example of this, most studies use workers that are exposed to high levels of asbestos - but what about low levels over a long period of time?
- Link function is typically logistic by default - it's easier to interpret
- The takeaway is that when you're talking about poisons or things that are harmful but it really works with a lot of different things. Where you might have something that doesn't have very often at normal concentrations but definitely happen at high concentrations.
- The example the book talks about asbestos, a harm substance. but a lot of the studies that look at it

look at high levels of asbestos exposures in folks who are working in construction. And so that data can be used to fit one part of the curve, maybe the curve close to the vertical dotted line. But it's difficult to learn about what happens at lower concentrations over a long period of time.

- There's not real way of choosing a link function without having some sort of prior knowledge as to why that relationship might exist. That's just one of the challenges that's good to think about when you are fitting these regressions. If you are fitting a regression for either end of an extreme low or high prob then predict outside of that, it's a very hard thing to justify to pick a default link function and then look at it. You might have to look at a number of them over time.
- Obviously the link function that we typically us it logit, it's much easier to interpret.
- The tails are different so it's important to discuss, but they also all converge to one pretty much as the same time. So if we're discussing large affects and small affects shouldn't the choice of link function not matter too much ?
- No because if we're talking about out here at 99% that's one thing and if we're talking about the very edge of 1% that's another thing. But the difference might be for example what if the difference between 5% risk vs. 1% risk. Relatively speaking that's a pretty big difference, but on an absolute scale it's not that big of a difference. So even though we see that these curves are, on an absolute scale, very close to each other. On a relative scale they can be very different.
- The book looks at these curves on a negative side as well which is weird. But I guess they're saying that the concentration isn't an actual physical concentration. Just thought that was weird so didn't include it.

Prospective and Retrospective Sampling

- Another way that logistics and logit work nicely in studies is what happens when we talk about prospective and retrospective sampling.

Data Set up

- Infant respiratory disease in their first year by feeding (bottle, breast milk, or supplement) and sex (boy or girl).

```
xtabs(disease/(disease + nondisease) ~ sex + food, babyfood)
```

sex/food	Bottle	Breast	Suppl
Boy	0.1681223	0.0951417	0.1292517
Girl	0.1250000	0.0668103	0.1259843

Prospective vs Retrospective

- Prospective - predictors are fixed and then the outcome is observed
  - Also called a cohort study
  - That would mean for example that you recruit a bunch of baby boys and then you can sort of write down how they will be feeding and then you can start the study and observe the outcome
- Retrospective - outcome is fixed and predictors are observed
  - Get a infants that have the outcome
  - Obtain a sample of infants without outcome
  - The way you could do that is go to a doctor's office and say find me everyone who has this respiratory disease and take a look at them, what gender are they and how did they feed.
  - You would also find another sample of infants that didn't have the outcome and write the observations of them

- The prospective and retrospective refers to when did the outcome happen, has the outcome already happened when you are looking at the study.
- This can be a little confusing because you can have a retrospective study where the outcome isn't necessarily fixed. It feels retrospective because the data has already been recorded.
- But in this case we're going to be talking about it when an outcome is fixed and what are these predictors.
- The only advice there is when someone says it's a retrospective study make sure you understand what is retrospective and what is not. Or basically it's important to know how patients are selected.

Set up prospective study

- Just boys that are either breast or bottle fed

```
babyfood[c(1,3),]
```

	disease	nondisease	sex	food
1	77	381	Boy	Bottle
3	47	447	Boy	Breast

- Given that the sex is boy and infant is breast fed - what are the log odds of having a respiratory disease? So if we're looking at breastfeed row, the second row, what are odds of having the disease? The odds is $47/447 = .10$ which is 10%. the prob would be $47/447+47 = .095$, which is 9.5%.
- Given that the sex is boy and infant is bottle fed what are the log odds of resp disease? The odds would be 77/381.
- How do we get the odds ratio? Divide the two odds.

Answer and difference

- If we look at the log odds and subtracting them is the same as taking the odds, dividing them and then taking the log of it.
- That's sort of the rule of logarithms.

```
# risk of bottle feeding
# this is on the log scale
log(77/381) - log(47/447)
```

```
## [1] 0.653417
```

```
log((77/381)/(47/447))
```

```
## [1] 0.653417
```

- What does the difference represent? the difference in .65.
- Is there an increase in risk from bottle feeding? Yes, because when you turn the log odds ratio into an odds ratio by exponentiating it. And so it'll be 1.91. If the odds ratio is greater than 1, that means the risk increased.
- If the log odds ratio was negative, then when you exponentiate it you would get a number less than 1 which mean that risk would decreased.
- For example if you reverse it, saying what's the risk of breast feeding vs bottle feeding you'd get -.65 and you would exponentiate it and get something less than 1.
- This was setting it up as a prospective study. We can do the same thing and sort of set it up backwards for retrospective.

What if this was a retrospective study?

```
babyfood[c(1,3),]
```

	disease	nondisease	sex	food
1	77	381	Boy	Bottle
3	47	447	Boy	Breast

```
log(77/47) - log(381/447)
```

```
## [1] 0.653417
```

- 77/47 is looking at everyone who has the disease already
- Because remember if we're selecting for people who have disease, that's sort of our fixed variable. Before food was the fixed variable, now disease is the fixed variable.
- So the odds, 77/47, represent given that you have the disease what are the odds that you were bottle feed.
- When we talk about the retrospective study that makes sense with how searched for the data. Because we said give me everyone who has the disease, and looked at how many were bottle feed and breast feed.
- The odds, 381/447, is the odds that given that you don't have the disease that you were bottle feed.
- So then we can do the same thing and take the difference and you'll find that it's the same exact value. This is convenient result that is not possible with the other function because you're not taking the log odds of things. .65 is the odds ratio, or the increasing likelihood that if you have the disease you were bottle feed, as opposed to the increase in odds of having the disease given that you were bottle feed. It's weird because we're looking at the same exact data, but we have to frame it different depending on how we actually calculated it.
- This is one of those nice things about the logit works is that there's symmetry between the two.
- Really it's the property of both logarithm and odds ratio, that basically you can sort of rewrite things like the equation below.

$$log(\frac{a/b}{c/d}) = log(\frac{ad}{bc}) = log(\frac{a/c}{b/d})$$

- This is not the case with other link functions

Pros/Cons of each Study Type

- Retrospective
  - Easier to find rare/long term outcomes
    - If the outcome is very rare, you might have to sample a big part of the population to find that outcome.
    - For example, it's very hard to analyze the entire population to find artificial heart use because it's a tiny % of the populational. On the other hand there's a registry of people who have artificial hearts
  - May not find all predictors
    - In the example they might not be available, in the example all of what happened before the artificial heart has been collected before.
  - Because if you have a very rare medical condition, how do you handle imbalance in the dataset? What you can do in these retrospective studies is that you might be able to assess the relative importance of different predictors. But you can't really give an absolute prediction.
  - So for example, if you have 150 people with an artificial heart and 150 without an artificial heart, the base rate of art heart in your population is 50%. But that is not the rate that it is in the general population. However if you have a balanced cohort (a properly balanced design) among other confounding factors, you can make statements about the relative importance of factors.

For example the odds ratio you can say something, but you can't say anything about the actual prediction on the data

- Prospective
  - Less susceptible to bias (in retrospective you must make sure that the populations for each outcome are similar/balanced)
    - For example if you recruit 1000 people you don't know what the outcome is, you're not biasing for that outcome. I don't know what it's going to be so there's no way for you to bias.
    - If you do the retrospective you're basically biasing it by design, to have percussions to make sure things are well balanced. you also have a lot more control and often accuracy on how data is collected. For example if you know you want to study feeding, when you enroll patients in a retrospective study you can say as part of the study I'm going to collect this information. That's probably going to be more accurate than if you just asked them what were you're feeding habits.
  - More control/possibly accurate in data collection
    - Both of these things come with the downside that they are often harder logistically to do or that they are more expensive. Clinical trials are generally retrospective studies and they're very expensive. These two pieces of it are reasons of it.
  - Only prospective studies can generate proper predictions
    - Why is that? What information are we missing in a retrospective design?
    - Because you have a better representation of what happens in a random sample of people. Not in a specific group of people for an outcome.

Covariates

- In context of all of this, you often have different covariates that you are studying. These are things that you care about and what you're actually interested in observing
- Let probability that individual is included in study if they do not have the disease = π_0
- Probability that individual is included in study if they have the disease = π_1
- What is the relationship between these two in prospective study?
  - $\pi_0 = \pi_1$
  - In a prospective study, you don't know who has the disease so if you're sampling at random, then the probability will be equal to each other, not compliments
- What is it in retrospective?
  - $\pi_1 > \pi_0$
  - By definition you're finding people with the disease, pi1 is greater than pi0.
- For example, say 2000 people, 2 win the lotto and the rest 1998 don't. Say we sample 1000 random people, we would get 1 person to win the lotto and 999 don't. We have an equal probability of choosing win and lose, 1/2 and 999/1998 which is both 50%. As opposed to if you pick the 2 lotto winners and then 2 people from the other group, you're at 100% for win and 2/1998 for lose.

Covariates

- For a given x, $p^*(x)$ is the probability that an individual has the disease given that they were included in the study, while $p(x)$ is the probability that someone has the disease regardless if they are in the study. We care about $p(x)$, the general prob. the x here is different risk factors.

$$p^*(x) = \frac{\pi_1 p(x)}{\pi_1 p(x) + \pi_0(1 - p(x))}$$

- The equation above is using bayes rule to say how you can take an unconditional probability and turn it into a conditional one. - So the unconditional probability that you have the disease to the conditional probability that you have the disease given that you are included in the study - Through some rearranging, you can actually write it as the equation below

$$logit(p^*(x)) = log(\pi_1/\pi_0) + logit(p(x))$$

- In the equation above, what we're saying is the logit of the probability you have the disease given that you're in the study is equal to the probability that you have the disease, regardless if you're in the study or not, plus a constant (involving pi) - Typically we don't know what this constant is, but that is okay. Because what we can still do is you can still make inferences on the relative effects of the different parameters in this equation. - This shows we can still do this even in a retrospective study - This means that we can estimate effects of different covariates, but not necessary predict outcome.

- In a prospective study, the constant would be log(1) which is zero. If these are equal, then the unconditional probability is the same as the conditional probability. And if they're not equal, then there's this constant.

Prediction and Effective Doses

Set up

- Prediction and confidence intervals

```
data(bliss, package="faraway")
# Create a model, the concentration of the poison is an independent variable
# And we're predicting the probability of death
lmod <- glm(cbind(dead,alive) ~ conc, family=binomial,data=bliss)
(lmodsum <- summary(lmod))
```

```
##
## Call:
## glm(formula = cbind(dead, alive) ~ conc, family = binomial, data = bliss)
##
## Deviance Residuals:
##       1        2        3        4        5
## -0.4510   0.3597   0.0000   0.0643  -0.2045
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -2.3238     0.4179  -5.561 2.69e-08 ***
## conc          1.1619     0.1814   6.405 1.51e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 64.76327  on 4  degrees of freedom
## Residual deviance:  0.37875  on 3  degrees of freedom
## AIC: 20.854
##
## Number of Fisher Scoring iterations: 4
```

Prediction

```
# We can predict for a new variable
# Use the predict function, use the new data of concentration of 2.5 and return the standard error
# You can put T for true
# This is on the linear scale, we get a log odds of .58 and the se is .23
(pred <- predict(lmod,newdata=data.frame(conc=2.5),se=T))
```

```
## $fit
##         1
## 0.5809475
##
## $se.fit
## [1] 0.2262995
##
## $residual.scale
## [1] 1
```

```
# We can create a prediction, the inverse logit of this value, the fit +/-
# Here are our predictions of death of .53 to.73
# Saying there's between a 53-73% chance that this poison will kill a bug at a concentration of 2.5
ilogit(c(pred$fit) + c(-1, 1) * 1.96 * pred$se.fit)
```

```
## [1] 0.5342962 0.7358471
```

ED50

- Effective dose 50 is when there is a 50% chance of outcome (in this case death)
- In this case, you would call it a lethal dose
- What you want is the p to = 1/2. What odds does a prob of 1/2 refer to? 1. And the log of 1 is 0.
- So basically what you're doing is you're setting the logit(p)=0, because the logit(1/2)=0. and you put it into you're equation over here, you do a little algebra and you get -beta0/beta1

$$logit(p) = logit(1/2) = 0 = \beta_0 + \beta_1 * E\hat{D}50 E\hat{D}_{50} = -\hat{\beta}_0/\beta_1$$

```
(ld50 <- -lmod$coef[1]/lmod$coef[2])
```

```
## (Intercept)
##           2
```

- In this case lethal does is 2.

ED50 Standard Error

- What if you want to get the standard error of that value? It's hard to get because we don't really have the estimate -beta0/beta1. And so have to do this transformation.
- Use delta method to estimate variance of transformed parameter
- The variance of the transformed variable is equal to the derivative of the transformation times the variance of the original variable.
- We're transforming two variables and so we have to do this matrix multiplication

$$var(g(\hat{\theta})) = g'(\hat{\theta})^T var(\hat{\theta}) g'(\hat{\theta})$$

```r
# Matrix multiplication in R is %*%
# The top line is a vector, the derivative of the estimate in respect to each of those two variables
dr <- c(-1/lmod$coef[2],lmod$coef[1]/lmod$coef[2]^2)
# This gives the estimate of the variance and so we take the square root of it
# The [,] gives you a vector instead of a matrix, otherwise this will be a 1x1 matrix
# We're taking this derivative and we're doing matrix multiplication times the covariance of the model
sqrt(dr %*% lmodsum$cov.un %*% dr)[,]
```

```
## [1] 0.1784367
```

- The takeaway is this is this way that you can get the standard error for something that doesn't really come out of the model

```r
# 95% CI around the estimate of 50%, which was 2
c(2-1.96*0.178,2+1.96*0.178)
```

```
## [1] 1.65112 2.34888
```

- The result is the confidence interval

Matched Case-Control Studies

Problem

- One of the problems is this whole idea of confounding. A confounder is something that effects both the variable that you're interested in studying. For example a drug or something like that and the outcome that you're studying. If there is a confounding variable, you can't really figure out what the effect of the drug is because it could be due to this confounding variable
- How do we control for confounding?
  - We can model away using regular regression techniques
- Hard to do if confounding variables are not balanced between cases and controls

Matching groups

- One way to deal with this is to match each patient with an outcome with another similar person without the outcome.
- One way that you could deal with that is matching patients on different confounders
  - The idea is if you have 2 groups of people, and you're not interested in age particularly, you can match people on having similar ages. So for every time you have a 56 year old receive the drug, you have a 56 year old that didn't receive the drug. Similarly you add other things, gender race, education, you can add all these different things as confounders.
- This becomes harder to do once you have more variables/continuous variables.
  - One thing that is tough is that once you have a bunch of these confounders it becomes hard to match people.
  - Example, is a 56 year old man similar to a 55 year old man? What about a 56 year old woman?
- One way to think about it is that you might have to have matches that are close enough.
- Downsides
  - You lose the ability to infer the properties of what you matched on. E.g. if you match on just age, then you won't be able to estimate the effect of age.
  - Relative effects of others may be found - but not the absolute prediction of an outcome.

- If you do have these matches and these groups, some cases it's a one-to-one match or one-three match, you can follow some of the things that they talk about in the book
- This can also be done so that the control set is matched more than 1:1 (e.g. 1:3 has 3 controls for each case)
- See chapter 4.5 in book
- Once you have this data matched, what you can do is you can basically use this regression technique at page 77 to basically estimate the parameters. And what we have said before where we can estimate the parameter but not the actual prediction itself.
- There are much better ways of doing this, such as the propensity score analysis is a common method.