CHAPTER 6

Binary Dependent Variables

Example 6.1. *As a terrorism researcher, I notice many things. For instance, on a trip to Washington, DC, I noticed that I had been stopped at the security checkpoint for an extended search for each of the past four times I flew. The Transit Security Adminstration (TSA) officials told me that I was randomly chosen and that there is no other reason the computer kept flagging me.*

I decided to test this statement. To do this, I asked my extended network of associates about their experiences with the TSA. For every flight they took over the past six months, I asked for their destination and their origin city, as well as some demographic information (including their research specialties and school affiliations). From this information, I want to predict who will and who will not be stopped by the TSA.

Thus far, we have examined linear regression where the dependent variable is unbounded and when the dependent variable is bounded. These cases cover a wide variety of cases where the dependent variable is continuous. Examples of dependent variables we can now use include heights, incomes, vote proportions, distances, and so forth.

However, we do not have the abilities yet to handle dependent variables which are discrete. These types of variables include count variables (ages, deaths, numbers of fires), ordinal variables

Individual	Insurance	Age	Income ($000)
1	0	25	20
2	0	30	40
3	0	21	30
4	0	39	25
5	1	55	55
6	1	40	60
7	1	40	45
8	1	44	30

Table 6.1: *Data to accompany Example 6.2 in the text.*

(importance level), nominal variables (different outcomes), and dichotomous variables (presence of a characteristic). These types of variables are all limited in that there is no allowed outcome between two adjacent outcomes: A person is either pregnant or not; You can have 3 or 4 fires, not 3.5 fires; A hurricane is a Category 4 or Category 5, not category 4.25.

This chapter covers the case of dichotomous variables. The next chapter covers the case of nominal and ordinal variables. Chapter **??** covers the case of count variables.

6.1 Binary variables

A dichotomous variable is one that can take one of two values: 1 or 0, True or False, Yes or No. In research, these variables include the incidence of terrorism, the election of a specific party to power, the existence of a fire, and the failure of a plane. In each of these cases, there are only two possible values. This is the hallmark of dichotomous variables. The problem with using the typical ordinary least squares is that predictions will invariably fall outside a logical range.

Example 6.2. *The decision to buy tornado insurance is related to several variables, including age and income. Table 6.1 includes records of several individuals. Fit this data with a simple linear model:*

$$insurance = \beta_0 + \beta_1\, age + \beta_2\, income$$

Next, predict whether Individual X will buy tornado insurance, given that his age is 65, and his income is $125,000. Finally, determine if the assumptions of Ordinary Least Squares are violated with this model.

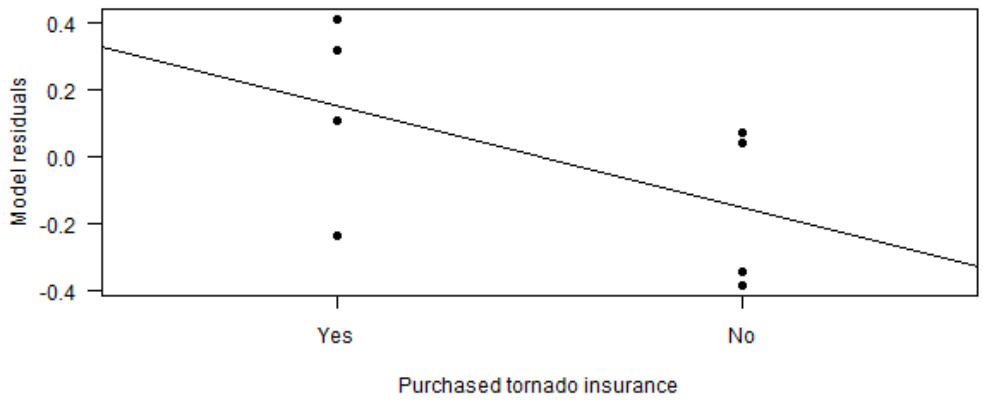

Figure 6.1: *Scatterplot of the residuals against the values of the dependent variable. Note the existence of a relationship between the two (diagonal line). As such, the linear model is not appropriate in this case.*

Solution: Using our statistical program, we get $insurance = -1.03 + 0.028 \times age + 0.014 \times income$ as our regression equation. Using the provided information, we predict Individual X will buy tornado insurance at 2.47.

To check the assumptions of OLS, let us merely check that the variance is constant. To do this, let us plot the residuals against the values of the dependent variable. Figure 6.1 shows that the residuals are not independent of the dependent variable — a violation of our assumptions.

Furthermore, calculations show that the variance for those who bought insurance is about 60% higher than for those who did not (0.817 vs 0.590). This is an example of non-constant variance.

Therefore, we conclude that our model is not appropriate for this data. ◇

To solve the first problem, we could create a decision rule that any predicted value above $\tau = 0.500$ will be treated as a 'Buy' prediction, and any predicted value less than $\tau = 0.500$ will be treated as a 'not buy' prediction. This is actually what we do in practice, although the threshold τ may be changed, depending on the application and the model.

The second problem is more serious and not as easily solved. One may consider performing a transformation on the dependent variable to make it unbounded. A logit transformation would be a natural transformation for this; however, we would also have to subtract a small value, $\varepsilon > 0$, from all '1' values and add it to all '0' values to avoid problems with infinity. Unfortunately, this transformation would not take care of the relationship between the residuals and the (transformed) dependent variables.

Instead of seeing the existence of a relationship between the residuals and the dependent variable as a problem, let us realize that this relationship tells us that there is *more* information in the data that we can model at this point.

6.2 Latent Variable Modeling

Instead of modeling the outcome, let us model the underlying probability that the person will purchase tornado insurance. This has the advantage of being a continuous variable instead of a dichotomous variable. As such, we can model it using previous techniques. The disadvantage is that there is an additional step in predicting the outcome: selecting a threshold value, τ, above which we predict the individual bought insurance; below which, not.

Thus, our research model in the tornado insurance example is

$$\text{logit}\Big(\mathbb{P}\left[\text{gets insurance}\right]\Big) = \tilde{y} = \beta_0 + \beta_1 \text{ age} + \beta_2 \text{ income} \tag{6.1}$$

We use the logit function for the same reason we used it before: to transform the bounded variable into an unbounded variable. The right hand side of Eqn 6.1 is a linear function that can take on all real values. It is actually called the "linear predictor" for this reason. Figure 6.2 shows a schematic of what we are actually modeling. The diagonal line is the line of best fit for the linear predictor. The horizontal line is the threshold value we chose to distinguish between 'Buy' predictions and 'Not Buy' predictions.

If we need to actually calculate the probability that Person X will purchase tornado insurance, we can calculate it from the linear predictor: $\mathbb{P}\left[\text{gets insurance}\right] = \text{logistic}\Big(\tilde{y}\Big)$.

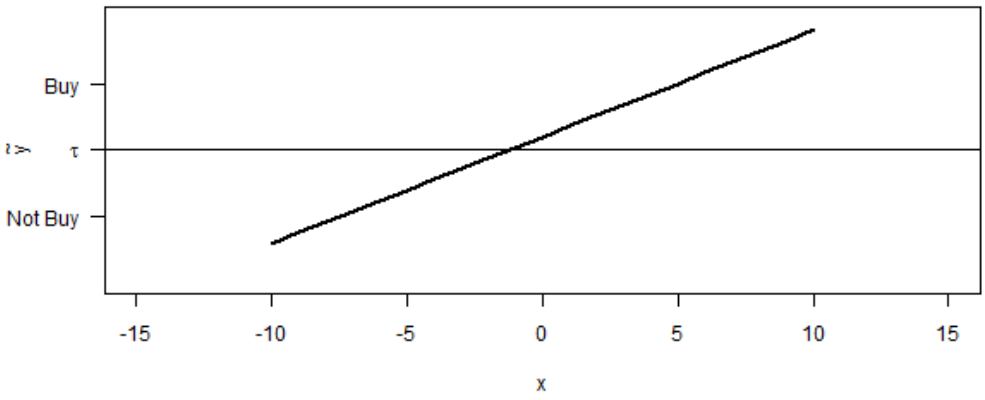

Figure 6.2: *Plot of the linear predictor and a possible threshold (τ) for a typical latent variable model.*

6.3 The mathematics

In an abstract way, this is already how we have been modeling. When we estimated our ordinary linear model, we were not really modelling the outcomes, we were modelling the *expected value* of the those outcomes. From this, we now know what that "line of best fit" actually means — It is the equation of the expected values. The observations are Normally distributed around that expected value. In other words,

$$\hat{y} \sim \mathcal{N}(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p, \sigma^2)$$

We call $\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$ the linear predictor, since it is a linear expression that predicts the expected value of the distribution.

Looking at linear regression in these terms means you are thinking in terms of generalizing the linear model. This is why this paradigm is called generalized linear modeling (GLM). The GLM paradigm requires three items: the linear predictor, the distribution, and the link function — three things we have already seen.[1]

Returning to our discussion of the binary dependent variable, we need to find a distribution to match our outcome and a link function to link our linear predictor with our expected value.

[1]For the linear regression, the link function is called the identity link, since there is no need to modify it to match the limits of the distribution. Other distributions, however, do set certain limits on the linear predictor.

One distribution that has only two outcomes is the Bernoulli distribution. For the Bernoullio, the probability of getting a '1' is p; of a '0', $1 - p$. In mathematical form, this means the probability density function is

$$f_X(x) = \begin{cases} p^x(1-p)^{1-x} & x \in \{0, 1\} \\ 0 & \text{otherwise} \end{cases}$$

Calculating the expected value of the Bernoulli is very straight forward using the definition of expected value:

$$\begin{aligned} \mathbb{E}[X] &:= \sum x_i f(x_i) \\ &= 0 f_X(0) + 1 f_X(1) \\ &= 0 * (1-p) + 1(p) \end{aligned}$$

Thus, the expected value of a Bernoulli is p, the probability of getting a '1'.

As for the link function, note that p is bounded: $p \in (0, 1)$. We have already met a link function that can handle this — the logit function.

With this, we have that the outcomes are distributed as

$$\hat{y} \sim \mathcal{B}]\nabla\backslash(p)$$

Here, $\text{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$.

★ **Notice**: *Here is what you need to take away from this discussion: The distribution must fit the possible outcomes. The link must translate the bounds on the parameter to the linear predictor. Both require you to know some distributions, which is why we had a chapter of them. It may be good to go back and re-read Chapter* **??**, *because you know another reason for studying them.*[2]

This section provided a link between what we have been doing with the Normal distribution and what we will be doing over the next few chapters with other distributions. As the linear model had no bounds on the dependent variable, there was no reason to introduce the GLM paradigm. We introduced the residuals are errors, but here we see those residuals as a natural consequence of the distribution.

[2] Yes, there is no such chapter yet, but there will be.

★ ★ ★

Let us now return to the binary dependent variable. The distribution is the Bernoulli.[3] The link function is the logit. The observed outcomes are 0 or 1. In the next example, the outcomes are actually Head or Tail.

Example 6.3. *Let us imagine an experiment, where we have a series of 100 coins. Were these coins all fair, then the probability of getting a Heads on any throw would be* $p = \frac{1}{2}$. *However, let us assume these coins are not necessarily fair, but that they are weighted in a very specific manner: In order of weighting towards Heads, the probability of getting a Heads on Coin i is* $p_i = p_1 + 0.005i$. *If we are allowed to flip each coin only once, is there a way of estimating* p_1*? If you would like to try, dataset* `chL-coinflips.csv` *contains one such run of flips.*

What the experiment provides us is a good look at a typical binary response variable and what we are actually looking for. In the thought experiment, the response variable was the Head or Tail that was flipped. The independent variable was the coin's position in the line, as the probability of a coin coming up heads is directly related to its position on the line ($p_i = p_1 + 0.005i$). Finally, we do not observe this probability; we only observe the outcome of the flip, which is a random variable based on a function of the observed values.

This model is very similar to all dichotomous dependent variable models: The existence of a group becoming a terrorist group is related to the economy, the group's separation from the state, and the existence of grievances. The outbreak of a hurricane in the Atlantic is related to the mean water temperature, the level of African dust in the air, and the strength of the El Niño in the Pacific. The vote of a person for a candidate depends on age, income, political affiliation, and employment type. In each of these examples, we are actually modeling the *probability* of the event — the expected value of the event — not the event itself. As we cannot observe the probability, we are reduced to using the event as a proxy for that probability.

Thus, our modeling equation is $P = XB + E$ and not $Y = XB + E$ as it is in linear regression (see the chapter on linear regression).

Warning: *In modeling the probability, there is one additional layer of complexity that must enter: the link function. In Chapter* **??** *(the chapter on variable transformations), we introduced the link function as a*

!

[3] The Bernoulli is a special case of the Binomial distribution. This is why most programs will call this distribution the Binomial, not the Bernoulli.

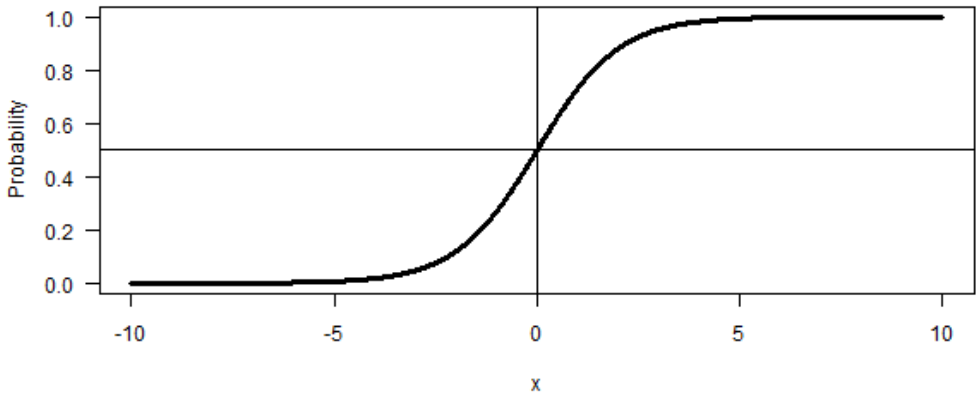

Figure 6.3: *A plot of the logistic function. Note that it is a symmetric function. In this context, 'symmetric' means that you can rotate it 180 degrees around some point and get the same function.*

manner of dealing with continuous variables that needed transforming. Here, since we are modeling the probability, we are still dealing with such transformations. The usual link function for binary dependent variables is the logit function ($\text{logit}(x) = \log(\frac{1}{1-x})$)*. It is the function that transforms the data into the transform space. In addition to the link, we need the inverse link in order to get us out of transform space back into reality.*[4]

This is important to keep in mind, because all of the results given by these procedures are given in the transformed units. Furthermore, it is important to note that these link functions are rarely simple function; they have been created to handle certain types of response variables. Thus, while we are familiar with the logit link function, others we may wish to employ include the cauchit, probit, log-log, and complementary log-log functions.

6.4 The canonical link

For a binary response variable, the canonical link function is the logit link. This link is characterized by being symmetric and having moderately thin tails (see Figure 6.3).

[4]The inverse of the logit is called the logistic function: $\text{logistic}(x) = \frac{1}{1+exp(-x)}$.

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.2929	0.5384	-4.26	0.0000
trial	0.0345	0.0087	3.97	0.0001

Table 6.2: *Results of performing logit regression on the coin flip data. Note that these coefficient estimates are in logit units. As such, any predictions done using them will have to be transformed into level units using the inverse link (logistic).*

The symmetry may be important when you are dealing with events that are balanced — neither rare nor frequent. The tail thickness may be important when you think there is a sharp transition between 0 and 1 in your data. In reality, Political Science theory is not so clear as to give you guidance in which link function you should use. As such, try several and see which one gives the best fit. Of course, if there is a traditional link function used in your field, you should probably use that one as a default. Thus, health science researchers should try the probit in lieu of the logit.

Example 6.4. *Let us revisit our coin-flipping experiment from above. Our goal is to determine the probability of flipping a Head for that first coin. Once we get that probability, we have estimates of all of the probabilities for all the coins.*

Solution: As we have no evidence to the contrary, let us use the canonical link function, the logit. Our steps are quite similar to the steps we performed when we had to transform the dependent variable:

1. Read in the data
2. Model the data using a generalized linear model
3. Predict outcomes from the model
4. Transform our predictions using the inverse link

Note that there is a step missing from when we transformed our dependent variable: We do not have to actually transform the dependent variable. The generalized linear modeling does that for us.

Also note that we are no longer using linear models, we are using generalized linear models. Generalized linear models allow you to specify the distribution of the error terms (of the dependent

variable); linear models required them to be distributed Normally. Generalized linear models are more general than linear models.

In R, the general form of the command is

```
glm(formula, family, link)
```

For binary response variables, the family will be the binomial distribution. Thus, for this example, the command will be

```
glm(head ~ trial, family=binomial(link=logit), data=data)
```

I used the data parameter, since I did not attach the data earlier. I included `link=logit` even though this is the default setting to remind myself what the link function is.

The result from this command is the regression table in Table **??**. Again, note that the parameter estimates (and predictions) will be in logit units. You will have to use the logistic function to get the predictions in level units.

Recall that the original question asked us to determine p_1. There are a couple ways of doing that. The best will depend on the numbers involved. Since we want p_1, we know it is equal to the logistic of the intercept plus one times the coefficient, $logistic(-2.292937 + 0.034492) = 0.09462345$. The other way is to use the predict function and take the logistic of that. You will get the same answer (within rounding error). Thus, our estimate of $p_1 = 0.095$.

We can also plot the probability curve on a graph of the outcomes, as we did in Figure **??**. ◇

6.5 Prediction accuracy

The next natural question concerns prediction accuracy. In linear regression, we used R^2 to help us determine how well the model fit the data — an R^2 value close to 1.00 indicated good fit, while an R^2 value close to 0.00 indicated a poor fit. If we recall, the R^2 value was calculated using variances of the original dataset and the square of the errors in the fitted model. Similar processes can be used in this context.

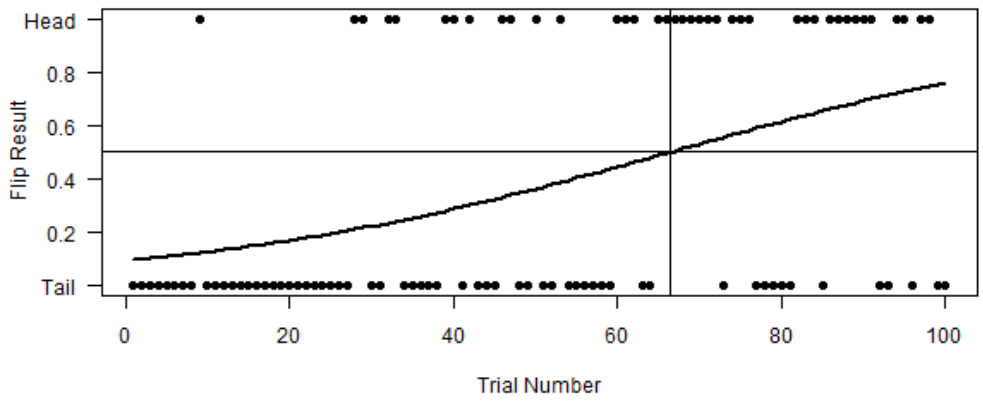

Figure 6.4: *Overlayed plot of the outcome of the experiment with the estimated probabilities superimposed. The horizontal line is the $\tau = 0.500$ threshold. The vertical line corresponds to a trial number corresponding to that threshold ($t = 66.5$). Thus, this model predicts that all coins above number 66.5 have a probability of greater than half of coming up Heads. Because this data is synthetic, I happen to know that Coin 71 is the fair coin.*

6.5.1 Accuracy rate

Let us define the accuracy rate to be the number of correct predictions divided by the total number of predictions. This makes inherent sense, as it reads as the percent of correct predictions. For this model, assuming a threshold value of $\tau = 0.500$, we see that there are 17 misclassified Heads and 12 misclassified Tails, for an accuracy rate of 0.710.

6.5.2 Relative accuracy

Of course, having an accuracy rate of 0.710 does not tell us the entire story. Just as the R^2 was based on a ratio of the model variance to the data variance, a better accuracy number would be the accuracy of the model relative to the accuracy of the data. The accuracy of the data refers to just selecting the modal category as our prediction. In this example, the modal category is Tails. Thus, the accuracy of selecting the modal category is 0.610 (there are 39 heads flipped in this data). So, the relative accuracy is

$$A_R = \frac{0.710}{0.610} = 1.164$$

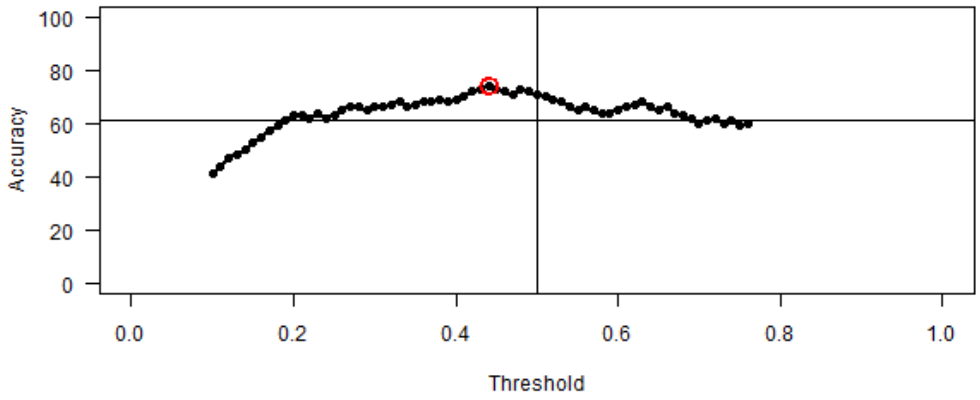

Figure 6.5: *A plot of the accuracy of the model against various thresholds. The horizontal line corresponds to the accuracy of selecting the modal category (the base accuracy). The vertical line corresponds to the threshold* $\tau = 0.500$*. The circled point represents the maximal threshold,* $\tau = 0.440$ *and accuracy* $= 0.740$*.*

Thus, the model does a 16.4% better job of prediction than does just selecting the modal category.

6.5.3 Maximum accuracy

In each of the above measures, we assumed our threshold was $\tau = 0.500$. In some cases, this is a logical threshold. In some cases, it is arbitrary. If we treat it as a parameter, we may be able to get a better model.

The plan is straight forward: Calculate the accuracy for various values of the threshold. That threshold which gives us the best accuracy will be our maximal threshold. Doing this by hand is prohibitive. Using a script to loop through all threshold values is much better. Figure 6.5 is a plot of the calculated accuracy for various thresholds. Note that the maximal threshold is not $\tau = 0.500$, but is $\tau = 0.440$, and the accuracy is 0.740 for that threshold.

6.5.4 ROC curve

There are other types of errors, more-specific types, that are useful in other fields. If we look back to Figure **??**, we see that the threshold line (horizontal) and the corresponding trial line (vertical)

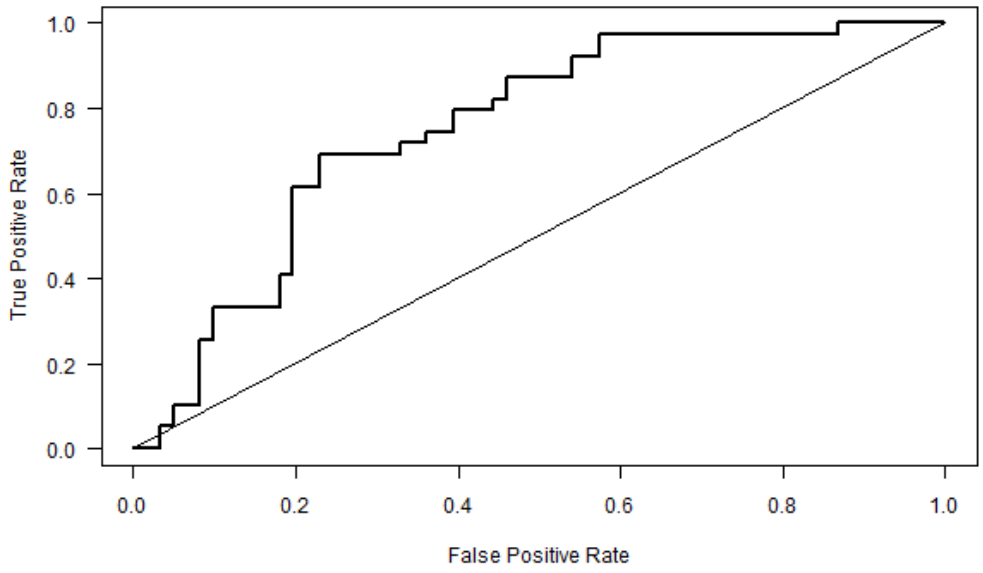

Figure 6.6: *A receiver operating characteristic curve for the coin flipping model. The diagonal line represents a random model. The thicker line represents our model. The farther the ROC curve is above the random line, the better the model is at distinguishing between the two cases (Head and Tail, here). The area under the ROC curve is a measure of the goodness of the model. Here, A' = 0.7516.*

divide the dataset into four parts. The lower-left quadrant are those Tails that are predicted by the model and the threshold value to be Tails. The upper-right quadrant are those Heads that are predicted to be Heads. The lower-right quadrant are Tails predicted to be Heads. The upper-left quadrant are Heads predicted to be Tails. These four are also referred to as True Negatives, True Positives, False Positives, and False Negatives. The error of the model (per above) is just the sum of the False Negatives and the False Positives.

For out coin flipping example (and with $\tau = 0.500$), we can write out a confusion matrix to show all four of these, both in magnitude and in rates:

$$\begin{bmatrix} FN = 17 & TP = 22 \\ TN = 49 & FP = 12 \end{bmatrix} \Longleftrightarrow \begin{bmatrix} FNR = \frac{17}{17+22} = 0.4359 & TPR = \frac{22}{22+17} = 0.5641 \\ TNR = \frac{49}{49+12} = 0.8033 & FPR = \frac{12}{12+49} = 0.1967 \end{bmatrix}$$

The True Negative Rate is also called specificity, and the True Positive Rate is called the specificity. You will come across these two terms in the field of biostatistics, because they mirror what physicians want out of their tests.

The receiver operating characteristic (ROC) curve is a graphical representation of the true positive rate (TPR) against the false positive rate (FPR) as the threshold is changed. Thus, to plot a ROC curve, one would calculate the sensitivity and the false positive rate for various values of the threshold, then plot sensitivity against the FPR. Figure 6.6 shows the ROC curve for our coin model.

In general, the closer the ROC curve approaches the left axis and upper axis, the better the model. As such, we can define a single number that tells us how good our model is — the area under the ROC curve, A'. The area under the ROC curve is a useful number in that it equals the probability that a model will classify a positive instance higher than a randomly chosen negative one. In other words, A' is the probability that the model scores a true Head higher than a true Tail.

Calculating the area is very straight forward, in a Calculus I manner: If we divide up the area under a curve into many vertical strips, then the area of the curve is the sum of the areas of those strips, which is just the height times the base. In this case, the height of the curve is just the function value, FPR. The width of each strip is the difference in consecutive FPRs. Thus, in `R`, the command will be `-sum(tpr*diff(c(1,fpr)))`. What this command does is it multiplies the height of the curve (`tpr`) by the width of each strip (`diff(c(1,fpr))`), then adds them together (`sum`). The negative sign is a result of the ROC curve actually being drawn from the top-right to the bottom-left, which is backwards to how areas are calculated.

The appendix to this Chapter provides the scripts used.

6.6 The complementary log-log link

Beyond the logit link, there are several other available links functions. Actually, for binary response variables, all that is required of the link functions is for it to smoothly map $g\colon (0,1) \to \mathbb{R}$ and to have an inverse that smoothly maps $f\colon \mathbb{R} \to (0,1)$. There are a plethora of functions that fit these requirements. As mentioned earlier, the logit link is symmetric. If you are dealing with rare-events data, you may not want to use a symmetric link function. The complementary log-log link is

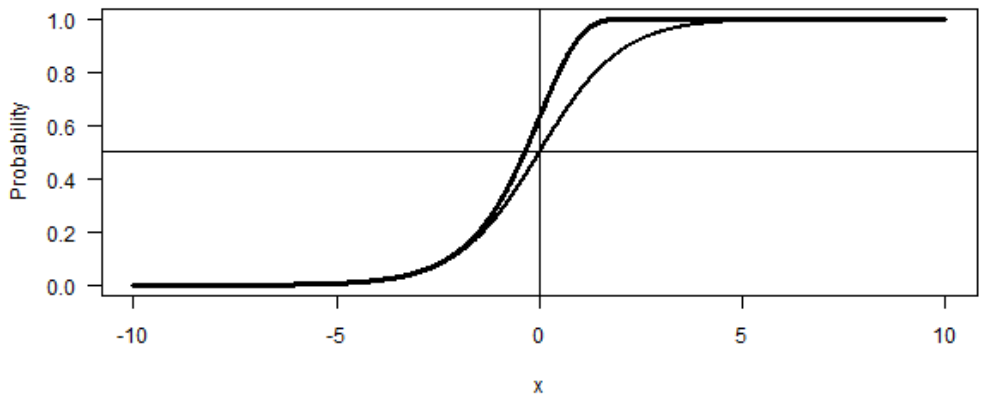

Figure 6.7: *Plot of the complementary log-log function (upper curve) on top of the logit. Note the difference in shapes between the two curves. The complementary log-log function approaches its maximum value much faster than does the logit.*

asymmetrical and often fits the bill. The formula for the complementary log-log is

$$g(\mu_i) := \log\Big(-\log(1-\mu_i)\Big)$$

Its inverse is

$$f(\eta_i) = 1 - \exp\Big(-\exp(x)\Big)$$

The plot of the complementary log-log function is seen in Figure 6.7, overlaid with the same plot for the logit link. Note the difference in shapes. Recall that the logit link is symmetric. The complementary log-log is not. It approaches its maximum value much faster than the logit.

Because of this asymmetry, it will fit models differently. Let us fit the coin data with a complementary log-log link. The command is

```
glm(head   trial, family=binomial(link=cloglog), data=data)
```

Note that the only change is in the link clause. The results of this new model are provided in Table 6.3. Note that the direction of effect is the same in both models. Unfortunately, as the first model is

	Estimate	Std. Error	z value	Pr(> |z|)
(Intercept)	-2.0651	0.4353	-4.74	0.0000
trial	0.0244	0.0063	3.86	0.0001

Table 6.3: *The results of fitting the coin flip data with a complementary log-log link (cf Table 6.2).*

in logit units and the second model is in complementary log-log units, comparing the coefficients tells us nothing. Comparing predictions tells us much more. Using the logit model, the prediction for p_1 was 0.095. Using the complementary log-log model, the prediction was $p_1 = 0.122$, which is closer to the true value of $p_1 = 0.15$.

6.6.1 Comparing models

Which model, in general, is better? Should we go through the trouble of using a complementary log-log link, or is a typical logit link good enough? The answer depends.

You can calculate all of the error (and accuracy) measurements we calculated in Section 6.5 and compare the two. However, if there is a difference is that difference statistically significant? While there are a few (very few) statistical tests on the accuracy measures we have already encountered, the best deals with comparing the two models. There are two tests I would like to introduce. The first is the Akaike comparison test. The second is the likelihood-ratio test. Both can only be used when the underlying data is the same in both models.

Akaike comparison test

There is a statistic that is usually produced whenever regression (of any form) is produced. This statistic is called the Akaike Information Criteria, and is abbreviated AIC. In the general case, $AIC = 2k - 2\ln(L)$, where k is the number of parameters being estimated and L is the likelihood of the model. [5] A lower AIC indicates a better model (from an Occam's Razor standpoint). As such, comparing the two AIC values for the two models will allow us to decide if one model is better than the other. An arbitrary rejection point is 8; that is, if one model is 8 AIC point better than the other, the latter model is eliminated. For our two coin flipping models, $AIC_{logit} = 118.48$, and $AIC_{cloglog} = 119.74$. Thus, while the logit is a better model from the AIC standpoint, it is not

[5] The quantity $-2\ln(L)$ is often called the deviance of the model, which will come in handy in Section **??**.

sufficiently better to completely ignore the complementary log-log model.