# Chapter 16:

## Binary Dependent Variables

Thus far, we have examined linear regression where the dependent variable is continuous — either unbounded or bounded. These cases cover a wide variety of instances — but not all. Examples of dependent variables we can now use include heights, incomes, vote proportions, distances, and so forth.

However, we do not yet have the ability to handle dependent variables which were discrete. Such variables include dichotomous variables (presence of a characteristic), count variables (ages, deaths, numbers of fires), ordinal variables (importance level), and nominal variables (different outcomes). These types of variables are all limited in that there are *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.

❧❧❧

As a terrorism researcher, I notice many things. For instance, on a recent trip to Washington, DC, I noticed that I had been stopped at the security checkpoint for an extended search in each of the past four times I flew. The Transit Security Administration (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 origin cities, as well as some demographic information (including their research specialties, school affiliations, and country). From this information, I want to predict who will and who will not be stopped by the TSA.

If none of the variables I measured were statistically significant, then I would have no reason to doubt the words of the TSA agents. If, however, one or more of these covariates and cofactors are statistically significant, then this would be evidence of profiling taking place.

The previous chapter introduced the Generalized Linear Model paradigm (GLM). Modeling with GLMs requires that we know three things:

1. the distribution of the dependent variable (thus a formula for the expected value of the dependent variable), $\mu$,
2. the linear predictor, $\eta$, and
3. a (bijective) function linking the two, $g(\mu) = \eta$.

In that chapter, we showed that the Classical Linear Model is just a special case of the Generalized Linear Model. Specifically, the CLM is just a GLM using the Gaussian (Normal) distribution and the identity link. In this chapter, we cover the case of dichotomous (binary) dependent variables. In the following pages, we determine the appropriate distribution and the canonical link function.

## 16.1: Binary Dependent 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: success and failure. This is the hallmark of dichotomous variables. Before Nelder and Wedderburn (1972) created the GLM framework, statisticians created special paradigms for binary dependent variable problems.

They did so because the Classical Linear Model invariably makes predictions outside the logical range, demonstrates heteroskedasticity, and has residuals that are not Normally distributed — all violations of the OLS assumptions. To demonstrate this, let us model tornado insurance purchasing using age and income and the Classical Linear Model (fit using OLS).

*Example* 16.1. The decision to buy tornado insurance is related to several variables, including age and income. Table 16.1 includes records of several individuals. Fit this data with a simple linear model using OLS:

$$\texttt{insurance} = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{income}$$

Next, predict whether Bob 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 and data.

| Individual | Insurance | Age | Income ($000) |
|---|---|---|---|
| 1 | 0 | 25 | 20 |
| 2 | 0 | 30 | 30 |
| 3 | 0 | 21 | 30 |
| 4 | 0 | 35 | 25 |
| 5 | 0 | 28 | 27 |
| 6 | 1 | 80 | 90 |
| 7 | 1 | 55 | 25 |
| 8 | 1 | 40 | 60 |
| 9 | 1 | 40 | 65 |
| 10 | 1 | 25 | 125 |

**Table 16.1:** *Insurance pseudo data to accompany Example 16.1 in the text, in which we predict a person purchasing tornado insurance based on the person's age and income.*

Using our statistical program, we get the following as our linear regression equation

$$\texttt{insurance} = -0.4277 + 0.0130 \times \texttt{age} + 0.0088 \times \texttt{income}$$

Using the provided information, we predict Bob will buy tornado insurance at (with?)

$$\begin{aligned} \texttt{insurance} &= -\,0.4277 + 0.0130 \times \texttt{age} + 0.0088 \times \texttt{income} \\ &= -\,0.4277 + 0.0130 \times 65 + 0.0088 \times 125 \\ &= +\,1.5121 \end{aligned}$$

What does this value actually mean? I don't know, either.

Next, to check the assumptions of OLS, let us merely check the assumption of homoskedasticity (constant variance). To do this, we plot the residuals against the values of the dependent variable. Figure 16.1 shows that the variation in the residuals significantly differs across the two groups in this model — a violation of our assumptions. In fact, calculations show that the variance for those who bought insurance is about 24 times higher than for those who did not (0.1325 vs. 0.0055). This is an example of non-constant variance. Performing the usual F-test, we also see that this difference is statistically significant ($F = 0.0416, \nu_n = 4, \nu_d = 4, p = 0.0093$).

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

❧ ❧ ❧

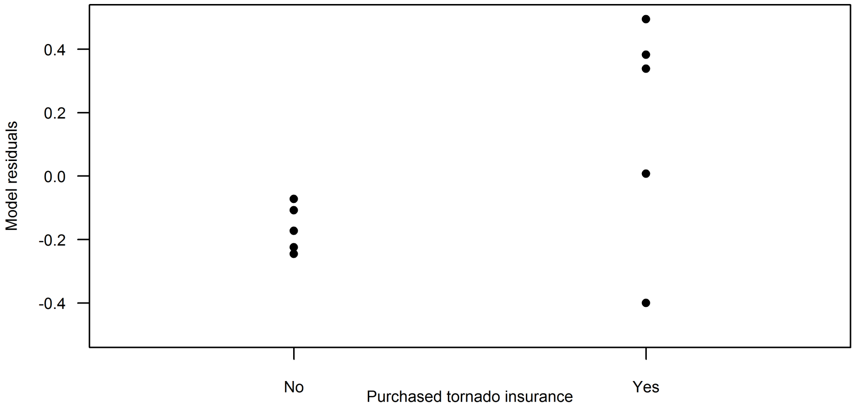

**Figure 16.1:** *Scatterplot of the residuals against the values of the dependent variable. Note the different variances for the two groups. As such, the linear model is not appropriate in this case.*

There were two problems with this analysis. First, the model predicted an outcome that did not make sense. Second, the model violated the assumptions of Ordinary Least Squares. To solve the first problem, we *could* create a decision rule that any predicted value above the threshold $\tau = 0.500$ will be treated as a 'Buy' prediction, and any predicted value less than $\tau = 0.500$ will be treated as a 'not buy' prediction.

The second problem is more serious and not so 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, all of the dependent variables are either 1 or 0, which means the transformed values will be either $+\infty$ or $-\infty$. Furthermore, this transformation would usually not take care of the relationship between the residuals and the (transformed) dependent variables.

***Note***: There is a tendency to feel disappointed when our model violates certain assumptions, such as here. However, *instead* of seeing the existence of a relationship between the residuals and the dependent variable as a problem, let us realize such a relationship tells us that there is *more information* in the data than we are modeling at this point. As an interested researcher, we want to use that information to get more from our data. Thus, assumption violation is not a step backwards; it is a path towards a deeper understanding of the data generating process.

## 16.2: Latent Variable Modeling

In Example 16.1, we discovered that Bob has a *something* of 1.5121 to buy tornado insurance. What is the *something*? Our gut really wants us to say that it is the probability that he buys tornado insurance. Unfortunately, it cannot be as the value is greater than 1. Notice, however, that we have just made an unconscious step in our minds: We are no longer thinking in terms of modeling the actual outcome (1 or 0); we are thinking in terms of modeling the *probability* of a success.

In other words, we are now modeling a variable we cannot measure — a latent variable. Instead of modeling the actual outcome, we now think in terms of modeling the underlying probability that the person will purchase tornado insurance. This has the dual advantage of being a continuous variable and of being bounded by 0 and 1, *exclusive*.

As such, we can model it using previous techniques. Remember that the predicted value will be a *probability*, not an actual outcome. To predict the outcome, there is an additional step: selecting a threshold value, $\tau$, above which we predict the individual bought insurance; below which, not. The traditional threshold value is $\tau = 0.500$; however, there is no reason we cannot alter it to better fit the data (Section 16.5.3).

Thus, our research model in the tornado insurance example becomes

$$\operatorname{logit}\Big(\mathbb{P}\left[\texttt{insurance}\right]\Big) = \beta_0 + \beta_1\texttt{age} + \beta_2\texttt{income} \tag{16.1}$$

We use the logit function for the same reason we used it before (Chapter 14): to transform the bounded variable into an unbounded variable. The right hand side of Eqn 16.1 is $\eta$, a linear function that can take on all real values — the linear predictor. Figure 16.2 shows a schematic of what we are actually modeling. The diagonal line in the top Figure 16.2 is the line of best fit for the linear predictor. The horizontal line is the threshold value we chose to distinguish between 'Success' predictions and 'Failure' predictions, which corresponds to logit($\tau$) in this top graph, $\tau$ in the bottom. The bottom figure is the linear predictor back-transformed into 'probability' units. The horizontal line is the actual $\tau$ chosen, here $\tau = 0.500$.

If we need to actually calculate the probability that Bob will purchase tornado insurance, we can calculate it from the linear predictor:

$$\operatorname{logit}\Big(\mathbb{P}\left[\text{insurance}\right]\Big) = \eta$$

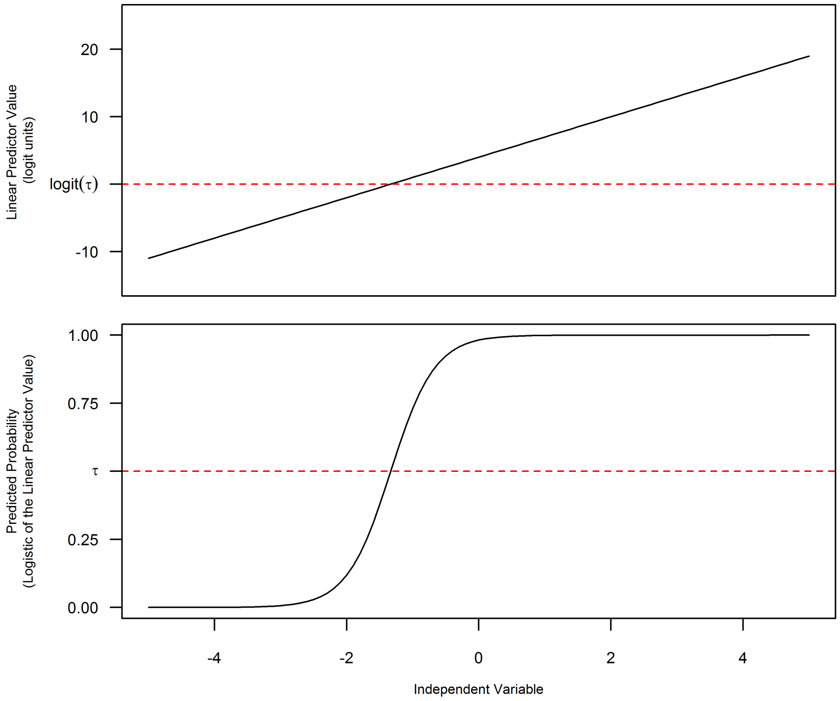

**Figure 16.2:** *Plot of the linear predictor and a possible threshold for a typical latent binary dependent variable model. The logit of the Linear Predictor is in level units (proportion units).*

This is equivalent to

$$\mathbb{P}\left[\text{insurance}\right] = \text{logistic}\left(\eta\right)$$

❧ ❧ ❧

This section examined the relationship between the line of best fit for the linear predictor, $\eta$, and the predicted probability of a success. However, we did not discuss how that line of best fit was determined. The next section does just that.

## 16.3: The Mathematics

When we model using the Classical Linear Model, we actually model/predict the *expected value* of the dependent variable. In the above insurance example, we modeled/predicted the probability of a person purchasing tornado insurance. What is the connection? It is that $\mathbb{E}[X] = p$.

Remember from Chapter 15, performing GLM estimation requires that we know three things about our data and our model: the linear predictor, the distribution of the dependent variable, and the function that links the two domains. The previous section discussed the linear predictor ($\eta = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{income}$) and a link function ($\text{logit}(\mu)$) for our example. That only leaves the distribution of the dependent variable.

What are the possible values of the dependent variable? They are $\left\{0, 1\right\}$. What distribution has only these two outcomes? It is the Bernoulli distribution.[1] For the Bernoulli distribution, the probability of getting a '1' is $p$ and the probability of getting a '0' is $1 - p$. Mathematically, this means the probability mass function (pmf) is

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

Strictly speaking, the probability mass function is not as important as the expected value of this distribution. Why? Remember that the Generalized Linear Model paradigm models the *expected value*, $\mathbb{E}[X]$ of the distribution of the dependent variable.

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

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

Thus, the expected value of a Bernoulli random variable is $p$, the success probability.

[1] The Bernoulli distribution is a special case of the Binomial, where $n = 1$; that is, if $X \sim Bern(p)$, then $X \sim Bin(1, p)$.

This fact makes the results of modeling more apparent: As the GLM paradigm models the expected value, when we use the Bernoulli distribution, we end up modeling the probability of a success, which is what we want.

> ***Note***: Recall that one of the assumptions of Ordinary Least Squares is that the variance is constant with respect to the dependent variable. When the outcomes are Bernoulli random variables, we can actually prove that the variance is *not* constant with respect to the predicted probabilities.
>
> To see this, let $X \sim Bern(p)$. With this, and with the probability mass function above, we use the definition of variance to calculate $\mathbb{V}[X]$:

$$\begin{aligned}
\mathbb{V}[X] &:= \sum_i (x_i - \mu)^2 f_X(x_i) \\
&= (0-\mu)^2 f_X(0) + (1-\mu)^2 f_X(1) \\
&= (0-p)^2 f_X(0) + (1-p)^2 f_X(1) \\
&= p^2(1-p) + (1-p)^2 p \\
&= p(1-p)\Big[p + (1-p)\Big]
\end{aligned}$$

This last line simplifies to $\mathbb{V}[X] = p(1-p)$, as $p + (1-p) = 1$, which means $\mathbb{V}[X]$ is a function of $p$ and is not a constant with respect to the prediction variable ($p$). Thus, binary dependent variables violate the assumption of homoskedasticity — by definition.

***Note***: The variance is a quadratic function of the probability of a success, $\mathbb{V}[X] = p(1-p)$. From this formula, we see that we are most unsure (the variance is highest) when the probability of a success is $p = 0.500$. Check that this makes sense: Which has a more uncertain outcome, a fair coin ($p = 0.500$) or a two-headed coin ($p = 1.000$)?

***Note***: Finally, let us note that *not all forms* of heteroskedasticity can be handled in this manner, even when there are only two outcomes. How do we handle such data? One way is to use a different estimation routine than Maximum Likelihood Estimation (MLE). One option is called Quasi-Likelihood Estimation (QLE). When this becomes important, we will revisit it.

❧ ❧ ❧

Now that we understand our choice of distribution a bit better, and the resulting expected value, let us examine the third facet: the link function. First, note that $p$ is bounded: $p \in (0,1)$. Thus we need a function that takes a doubly-bounded variable and transforms it into an unbounded variable. We have already met a link function that can handle this — the logit function (see Chapter 14).[2]

And so, we have the three necessary components to use Generalized Linear Models in this example:

- the linear predictor,

$$\eta = \beta_0 + \beta_1 \texttt{age} + \beta_2 \texttt{income}$$

- the distribution of the dependent variable,

$$\texttt{insurance} \sim Bin(1,p)$$

  with the formula for the expected value,

$$\mu = p$$

- and the link function,

$$\text{logit}(\mu) = \eta$$

***Note***: Here is what you need to take away from this section: 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 there is an appendix for them.

### 16.3.1 Deriving the *Canonical* Link

In Chapter 15, we mentioned that each distribution has a canonical link. Let us derive the canonical link for

[2] Here, I must mention that the logit is *not* the only appropriate link function. *Any* monotonic function that maps $(0,1) \mapsto \mathbb{R}$ is appropriate. This includes the entire class of quantile functions, of which the probit is a member.

The choice of the link function often reduces to tradition within your field. However, Social Science theory is getting advanced enough to suggest link functions that are more appropriate than others.

the Bernoulli distribution. As a side note, one does not have to understand this section to use Generalized Linear Models.

The steps to determine the canonical link are the same for the Binomial as it was for the Gaussian (Chapter 15):

1. Write the probability mass function (pmf).
2. Write the probability mass function in the required form.
3. Read off the canonical link.

For this distribution, this results in:

$$\begin{aligned}
\text{pmf}: \quad & p^x(1-p)^{1-x} \\
& = \exp\left[\log\left(p^x(1-p)^{1-x}\right)\right] \\
& = \exp\left[\log\left(p^x\right)+\log\left((1-p)^{1-x}\right)\right] \\
& = \exp\left[x\log\left(p\right)+(1-x)\log\left(1-p\right)\right] \\
& = \exp\left[x\log\left(p\right)+\log\left(1-p\right)-x\log\left(1-p\right)\right] \\
& = \exp\left[x\left(\log(p)-\log(1-p)\right)+\log(1-p)\right] \\
& = \exp\left[x\log\left(\frac{p}{1-p}\right)+\log(1-p)\right] \\
& = \exp\left[x\,\text{logit}(p)+\log(1-p)\right] \\
& = \exp\left[\frac{x\,\text{logit}(p)+\log(1-p)}{1}+0\right]
\end{aligned}$$

This is in the required form.

$$\text{Required form}: \quad \exp\left[\frac{x\theta-b(\theta)}{a(\phi)}+c(y,\theta)\right]$$

Thus, reading off the standard form, we have the following:

- $x=x$

| Link | | Inverse Link | |
|---|---|---|---|
| Logit | $\log(\mu/(1-\mu))$ | Logistic | $(1+\exp(-\eta))^{-1}$ |
| Probit | $\Phi^{-1}(\mu)$ | | $\Phi(\eta)$ |
| Cauchit | $\tan\left(\pi\left(\mu-\frac{1}{2}\right)\right)$ | | $\arctan(\eta)/\pi+\frac{1}{2}$ |
| Loglog | $-\log(-\log(\mu))$ | | $\exp(-\exp(-\eta))$ |
| Complementary Loglog | $\log(-\log(1-\mu))$ | | $1-\exp(-\exp(\eta))$ |

**Table 16.2:** *A list of several possible link functions (not all) to use for binary dependent variables. For the case of the Bernoulli distribution, remember that $\mu = p$.*

- $\theta = \text{logit}(p)$
- $a(\phi) = 1$
- $b(\theta) = \log(1-p) = -\log(1+e^{\theta})$
- $c(y,\theta) = 0$

As such, the canonical link is the logit function, $g(\mu) \equiv \text{logit}(p)$.

❧ ❧ ❧

As mentioned in Chapter 15, we do not *have* to use the canonical link. *Any* monotonic, increasing function that maps the restricted domain to the unrestricted domain works. Thus, there are several options for the link function. Table 16.2 gives some options.

The logit link is the canonical link. The probit link is frequently used in biostatistics. Its advantage is that it is based on the Normal distribution, with which we are intimately familiar. There is usually little difference between predictions made with the logit link and those made by the probit link. The coefficient estimates will usually differ by a factor of approximately 3.7, and the levels of significance will usually be close. The cauchit link is a symmetric link with heavy tails, as compared to the logit and the probit links (see Figure 16.3).

The loglog link and the complementary loglog link are asymmetric links. The loglog link has a heavy right tail; the complementary loglog link, a heavy left tail (see Figure 16.4). Social Science literature is only now beginning to be able to state which of the three types of link functions will be most appropriate for the given model (symmetric, heavy left, heavy right).

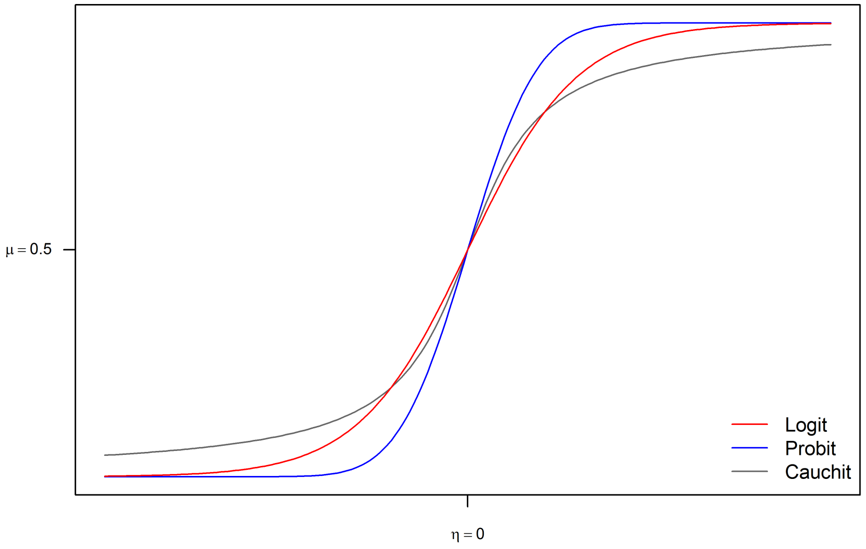

**Figure 16.3:** *A graph of three symmetric links (logit, probit, and cauchit). Note that they all cross when the linear predictor $\eta = 0$ and that they cross at $\mu = 0.5$.*

> ***Note***: R has the built-in ability to model using the following link functions for the Binomial (Bernoulli) distribution: logit, probit, cauchit, log, and cloglog. The RFS package adds the loglog link function. Thus, once that library is loaded, one only has to type the following to perform loglog regression:
>
> ```
> glm( y~x, family=binomial(link=make.link("loglog")) )
> ```

❧ ❧ ❧

Again, the link choice is usually a matter of tradition, rarely of theory. Each of the link functions should report the same variables being statistically significant. So, from a theory-testing standpoint, the link functions are rather interchangeable. With that said, *predictions* will vary depending on the link function chosen. Thus, if prediction is important then you will want to investigate different link functions.

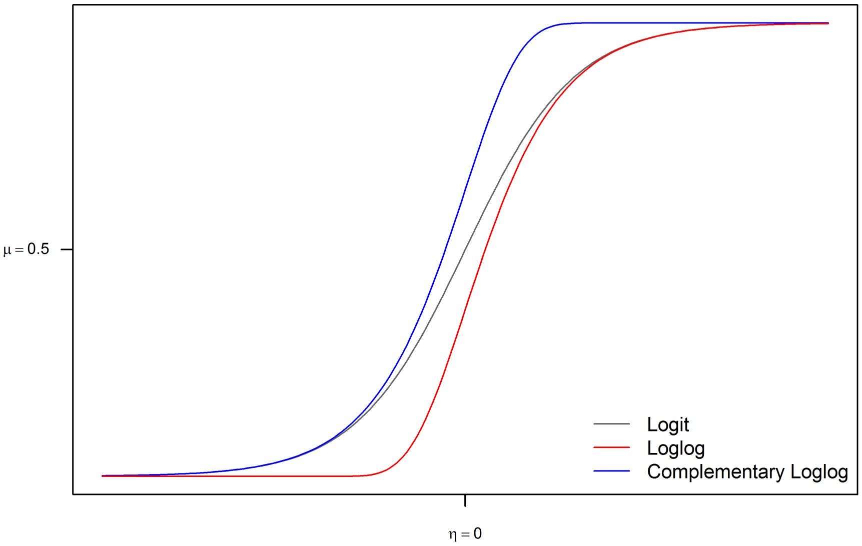

**Figure 16.4:** *A graph of two asymmetric links (complementary loglog and loglog functions) with the symmetric logit link for comparison.*

## 16.4: Modeling with the Logit

For a binary response variable, the canonical link function is the logit link. This link is characterized by being symmetric and having relatively thin tails (see Figure 16.5).

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 success and failure in your data. In reality, current social science theory is rarely 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 use that one as a default. Thus, Political Scientists should start with the logit, while the health science researchers should start with the probit.

*Example* 16.2. Since binary dependent variable regression is important to understand, let us look at it from a different direction: Let us imagine an experiment where we have a series of 100 coins. Were these coins all fair, then the probability of getting a Head on any throw would be $p = \frac{1}{2}$. However, let

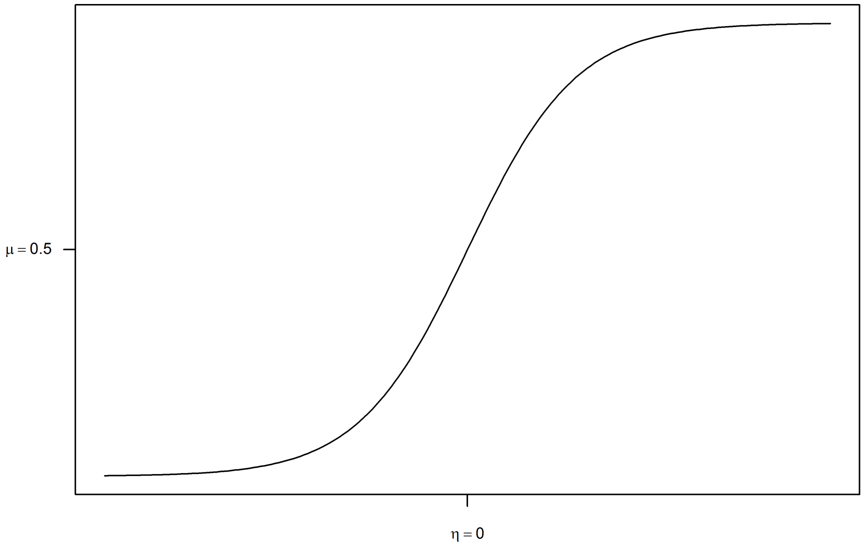

**Figure 16.5:** *A plot of the canonical link function. Note that it is a symmetric function. In this context, 'symmetric' means that you can rotate it 180 degrees around some point and get the same function.*

us assume these coins are not necessarily fair, but that they are weighted in a very specific manner: Coin $i$ has a probability of flipping a Head of $p_i = p_0 + 0.005i$. Now, if we were allowed to flip each coin *only once*, is there a way of estimating $p_0$ from the data?

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 the GLM paradigm (specify the distribution, the linear estimator, and the link function)
3. Predict outcomes from the model
4. Back-transform the predictions using the inverse of the link function

***Note***: There is a step missing from when we previously transformed our dependent variable: We do not have to transform the dependent variable.

|  | Estimate | Std. Error | z-value | p-value |
|---|---|---|---|---|
| Constant Term | -2.2929 | 0.5384 | -4.26 | $\ll 0.0001$ |
| Trial Number | 0.0345 | 0.0087 | 3.97 | 0.0001 |

**Table 16.3:** *Results of performing logit regression on the coin flip data,* `coinflips`*. 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 of the link function (the logistic).*

Generalized Linear Modeling does that for us in R. We do, however, have to back-transform the predictions.

In R, the general form of the command is, showing the most important parameters,

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

Only `formula` is required. If `family` is missing, the Gaussian (or Normal) distribution will be assumed. If `link` is missing, the canonical link for that family will be assumed. If `data` is missing, the current data will be assumed.

For binary response variables, the family will need to be the Binomial distribution. Thus, for the coin example, the command will be

```
m1 <- glm(head~trial, family=binomial(link=logit), data=coin)
```

I used the `data` parameter, as I did not attach the data earlier. I also included `link=logit` even though this is the default setting for the Binomial family in order to remind myself of the link function.

The results from this command are summarized in Table 16.3. 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 units of probability.

Recall that the original question asked us to determine $p_1$, the probability of getting a Head on the first coin. 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: $\text{logistic}(-2.2929 + 1 \times 0.0345) = 0.0946$.

The other way is to use the predict function and take the logistic of that value. You will get the same answer (within rounding error). The function call used is

```
predict(m1, newdata=data.frame(trial=1))
```

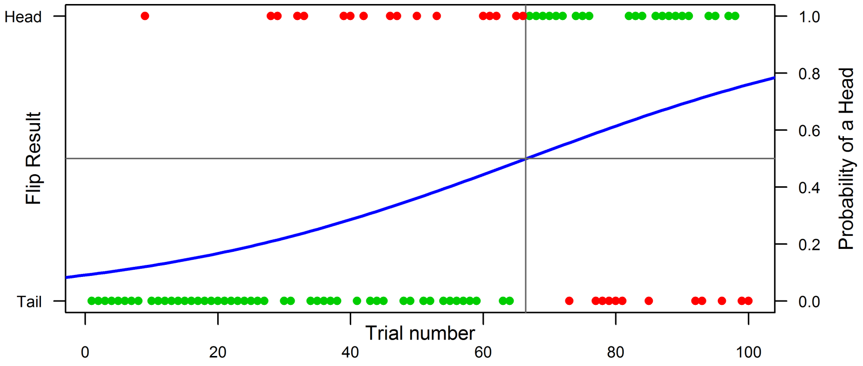

**Figure 16.6:** *Overlayed plot of the outcome of the experiment with the estimated probabilities superimposed. The horizontal line is the $\tau = 0.500$ threshold. The vertical line corresponds to a trial number corresponding to that threshold ($\tau = 66.4$). Thus, this model predicts that all coins above number 66.4 have a probability of greater than half of coming up Heads. Red dots are misclassified, green dots are properly classified.*

This gives an answer of $-2.2584$. The logistic of -2.2584 is our estimate of $p_1$, which is $p_1 = 0.0946$.

If we want to, we can also plot the probability curve on a graph of the outcomes (see Figure 16.6). With such a graph, we could estimate which coin is most fair. With the graph, we could also get a feel for how well the model represents the data.

***Note***: The linear predictor is represented in the curve graphed in Figure 16.6. Note, however, that the curve is *not* linear. This is because the curve in Figure 16.6 is actually the logistic of the linear predictor.

## 16.5: Prediction Accuracy

The next natural questions concern issues of goodness-of-fit: How good is the model? This question can be answered using many related accuracy measures.

Recall that 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 — a PRE measure — was calculated using variances of the original dataset and the square of the errors in the fitted model (Section 12.3). Similar processes can be used in this context to create a *pseudo* $R^2$ measure.

### 16.5.1 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 a measure of goodness of fit since it reads as the proportion of correct predictions.

There is no native accuracy function in R. However, the RFS package provides one. The accuracy() function takes four parameters: data (the data variable), y (the binary dependent variable), model (the model you fit with the data), and t (the threshold). The optional parameter, rate, tells the function to return the accuracy rate (default) or the number of accurate predictions (rate=FALSE). Thus, to determine the accuracy of this model for this data using the usual threshold value of $\tau = 0.500$, we would use

```
accuracy(data=coin, y=coin$head, model=m1, t=0.500)
```

The result of this command is 0.710, which agrees with our by-hand calculations. Thus, we conclude that this model correctly predicts 71% of the time for this data.

### 16.5.2 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 null model. The accuracy of the null model refers to merely selecting the modal category as our prediction. In this example, the modal category is Tails as there were 61 Tails in the data. Thus, the accuracy of merely selecting the modal category is $61 \div 100 = 0.610$. So, the *relative* accuracy is

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

Thus, the model does a 16.4% better job of prediction than does just predicting 'Tail' all of the time.

There actually is a proportional reduction in error (PRE) measurement associated with the relative accuracy. Recall that the $R^2$ value was valuable because it measured the proportion of error explained by the model. For binary dependent variable regression, we can calculate something similar.

$$PRE = 1 - \frac{\text{error with model}}{\text{error without model}}$$

Here, we can see that a pseudo-$R^2$ measure for this data and this model (and this threshold) is

$$1 - \frac{1-0.710}{1-0.610} \approx 0.2564$$

Thus, we can state that this model (and this threshold) reduced the error by 25.64%. Note that there is a bad quality of this measure: while it can never be greater than 1.0, it *can* be less than zero. However, it will only be less than zero when *your* model is worse than no model.

> ***Note***: There are many different ways of calculating a pseudo-$R^2$. Each of the measures are based on different definitions of 'error,' just as the $R^2$ and the adjusted $R^2$ are both based on different definitions of error. Researchers do not agree on much about pseudo-$R^2$ measures except that they are not useful *in vacuo*, and rarely useful in concert with other measures.

This is why I am offering it here, alongside many other measures of fit. Getting to know your model results is just as important as getting to know your data.

### 16.5.3 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 arbitrarily chosen. If we treat it as a parameter, we may be able to get a better prediction model.

The plan is straight forward: Calculate the accuracy for various values of the threshold. The threshold that gives us the best accuracy will be our optimal threshold. Doing this by hand is prohibitive. Using a script to loop through all threshold values is much easier:

```
a <- numeric()
for(i in 1:100) {
  t <- i/100
  a[i] <- accuracy(coin, coin$head, m1, t=t)
}
```

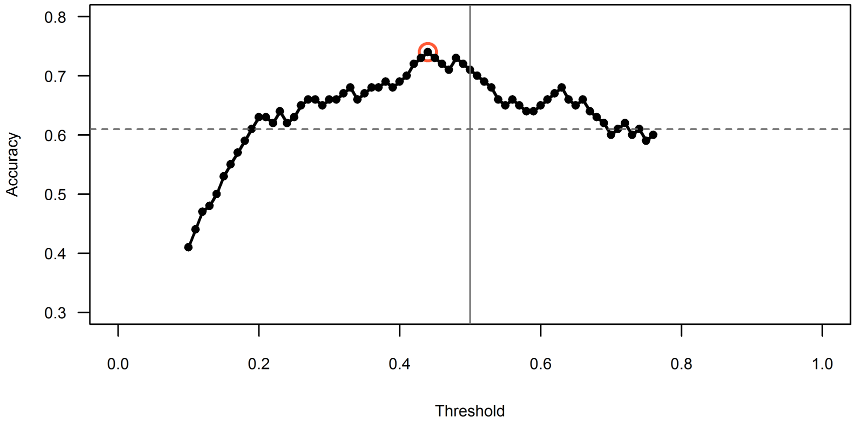

**Figure 16.7:** *A plot of the accuracy of the model against various thresholds. The horizontal line corresponds to the accuracy of selecting the modal category (the base accuracy). The vertical line corresponds to the threshold* $\tau = 0.500$. *The circled point represents the maximal threshold,* $\tau = 0.440$ *and accuracy* $= 0.740$.

Figure 16.7 is a plot of the calculated accuracy for various thresholds. Note that the optimal threshold is not $\tau = 0.500$, but $\tau = 0.440$, and the maximal accuracy is 0.740 for that threshold. Note, however, that there is little difference in accuracies between this optimal threshold ($\tau = 0.440, A = 0.740$) and the traditional threshold ($\tau = 0.500, A = 0.710$).

### 16.5.4 The ROC Curve

There are other types of errors, more-specific types, that are useful in other fields. If we look back to Figure 16.6, we see that the threshold line (horizontal) and the corresponding trial line (vertical) divide the dataset into four parts. The lower-left quadrant are those Tails that are correctly predicted by the model and the threshold value to be Tails. The upper-right quadrant are those Heads that are correctly predicted to be Heads. The lower-right quadrant are Tails incorrectly predicted to be Heads. The upper-left quadrant are Heads incorrectly predicted to be Tails. These four types of errors are also referred to as True Negatives, True Positives, False Positives, and False Negatives, respectively.

For our 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:

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

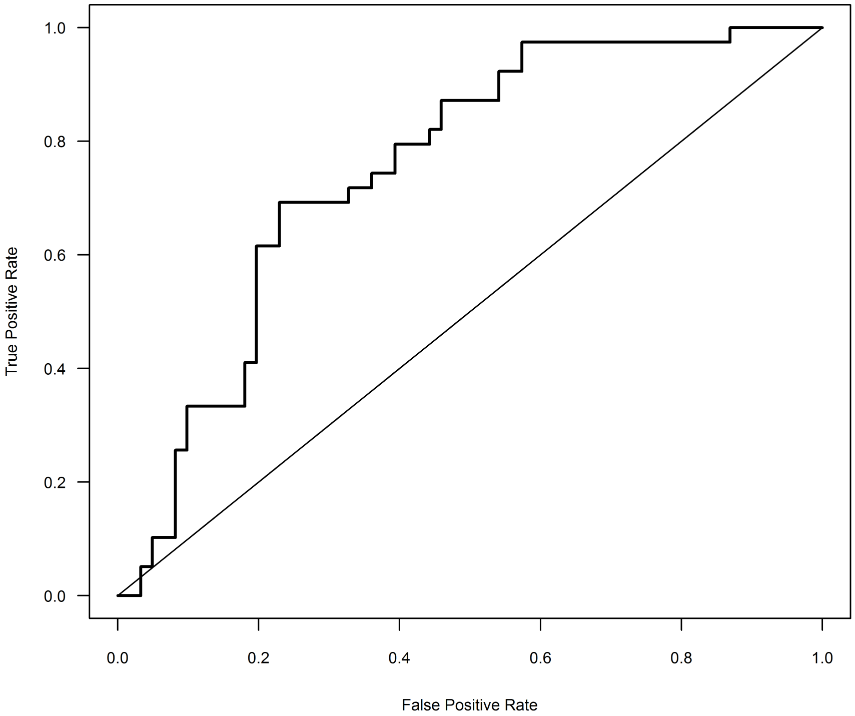

**Figure 16.8:** *A receiver operating characteristic curve for the coin flipping model. The diagonal line represents a random model. The thicker line represents our model. The farther the ROC curve is above the random line, the better the model is at distinguishing between the two cases (Head and Tail, here). The area under the ROC curve is a measure of the goodness of the model. Here,* $A' = 0.7516$.

The true negative rate (TNR) is also called *specificity*, and the true positive rate (TPR) is called the *sensitivity*. You will come across these two terms in the field of biostatistics, because they mirror what physicians and biomedical researchers want out of their diagnostic tests.

The receiver operating characteristic (ROC) curve is a graphical representation of the true positive rate 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 16.8 shows the ROC curve for our coin model.

In general, the closer the ROC curve approaches the left and upper axes, the better the model. As such, we can define a single number that tells

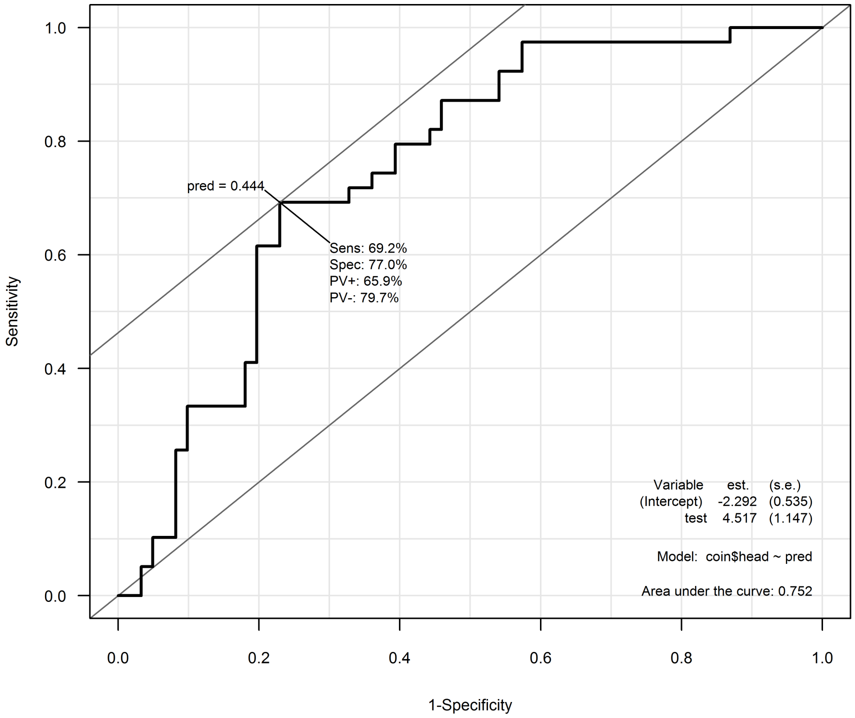

**Figure 16.9:** *The receiver operating characteristic curve for the coin flipping model using the* `ROC` *command from the* `Epi` *package.*

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 negative one. In other words, $A'$ is the probability that the model scores a true Head (success) higher than a true Tail (failure). Calculating the area is very straight forward, in a geometry/Riemann Sum manner.

> ***Note***: There is an entire `R` package dedicated to ROC curves, `Epi`. To create ROC graphs and to calculate the area under the curve in that package, first load it using `library(Epi)`, then use the command
>
> ```
> ROC(test, stat, plot="ROC")
> ```

Here, `test` is the predicted probability of success for each datum from model (a continuous variable bounded by 0 and 1), `stat` is the binary dependent variable, and `plot="ROC"` produces a ROC plot. This graph (Figure 16.9) is a bit more useful than the simple graph in Figure 16.8, as it contains some useful statistics, including the AUC and the optimal threshold, $\tau$, which is the threshold value closest to the upper-left corner.

***Note***: This optimal value is only optimal if the costs of making each type of error is the same.

## 16.6: Modeling with Other Links

The logit regression we did above is quite sufficient if all you want to do is fit the data using logit regression. If, on the other hand, you want to better understand the process that gave you the data, you will want to try different link functions to determine if any of the alternative links do an appreciably better job of fitting your data. The logit link is symmetric. You should also use the probit link as a check on your model: If the results are comparable, then the conclusions are strengthened; if not, there is something wrong with your model.

In addition to using a second symmetric link function, you should use the two main asymmetric link functions: the complementary loglog and the loglog link function.

### 16.6.1 The Complementary Loglog Link

As mentioned earlier, there are several other available links functions beyond the logit link (see Table 16.2). Actually, for binary response variables, all that is required of the link functions is for it to smoothly map $g\colon (0,1) \mapsto \mathbb{R}$ and to have an inverse that smoothly maps $f\colon \mathbb{R} \mapsto (0,1)$. 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 loglog link is asymmetric and is often useful (Figure 16.10).[3]

The formula for the complementary log-log is

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

[3] You may see the complementary loglog link function referred to by its abbreviation — cloglog.

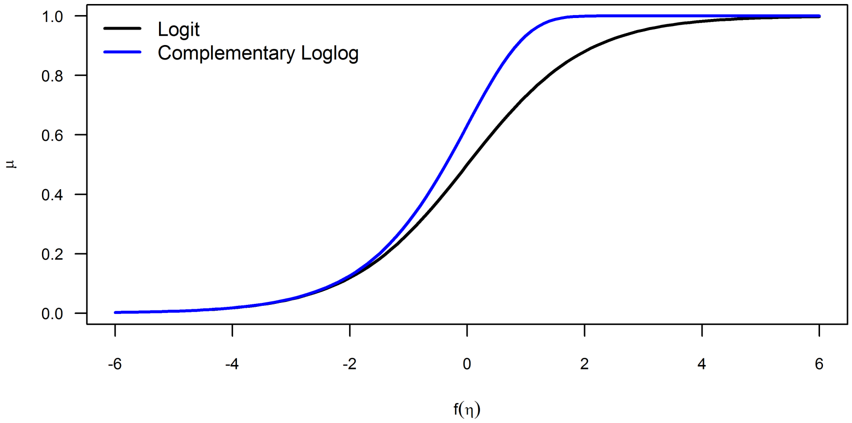

**Figure 16.10:** *Plot of the complementary loglog function (upper curve) on top of the logit. Note the difference in shapes between the two curves. The asymmetric complementary loglog function approaches its maximum value much faster than does the symmetric logit.*

Its inverse is

$$f(\eta) = 1 - \exp\Big(-\exp(\eta)\Big)$$

The plot of the complementary loglog function is seen in Figure 16.10, overlaid with the same plot for the logit link. Note the difference in shapes. Recall that the logit link is symmetric. The complementary loglog 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 loglog link. The command is

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

Note that the only change is in the link clause. The results of this new model are provided in Table 16.4. Note that the direction of effect is the same in both models. Unfortunately, as the first model is in logit units and the second model is in complementary loglog units, comparing the magnitude of 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 loglog model, the prediction is $p_1 = 0.122$, which is closer to the true value of $p_1 = 0.150$.

### 16.6.2 The Loglog Link

A second useful asymmetrical link function is the loglog link (Figure 16.11). Note that the asymmetric loglog link rises to

| | Estimate | Std. Error | z value | Pr(> \| z \|) |
|---|---|---|---|---|
| Constant term | -2.0651 | 0.4353 | -4.74 | $\ll 0.0001$ |
| Trial number | 0.0244 | 0.0063 | 3.86 | 0.0001 |

**Table 16.4:** *The results of fitting the coin flip data with a complementary loglog link (cf. Table 16.3). As before, the magnitudes of the estimates cannot be compared across different link functions; however, the direction of effect can.*

its maximum much slower than either the symmetric logit link or the asymmetric complementary loglog link. Because of this functional shape, it will be better at fitting certain data sets better than the other link functions discussed.

In reality, there is a functional relationship between the complementary loglog and the loglog link functions. They are 180° rotations of each other. Thus, statistical programs either have no support for either or have support only one. `R` has native support for only the complementary loglog link. However, with the `RFS` package, it is straight forward to perform loglog regression.

The command to perform the loglog regression on this data is the same as before, except for the link parameter, which is now

```
link=make.link("loglog")
```

***Note***: The `link` parameter is a bit more cumbersome than has been previously. This is because the loglog link is not native to `R`. Because of this, the `make.link` helper function is needed. Make sure you have already loaded the `RFS` package.

With this, I leave it as an exercise for you to show that the effect of trials in the loglog model is 0.0233 and that the predicted probability of a head for Coin 1 using this model is 0.0498.

## 16.7: Model Selection

Which of the models is best? That is a model selection question. Model selection *procedures* (not tests) attempt to balance competing desires — accuracy and parsimony — to create the 'best' model, by some standard. For linear

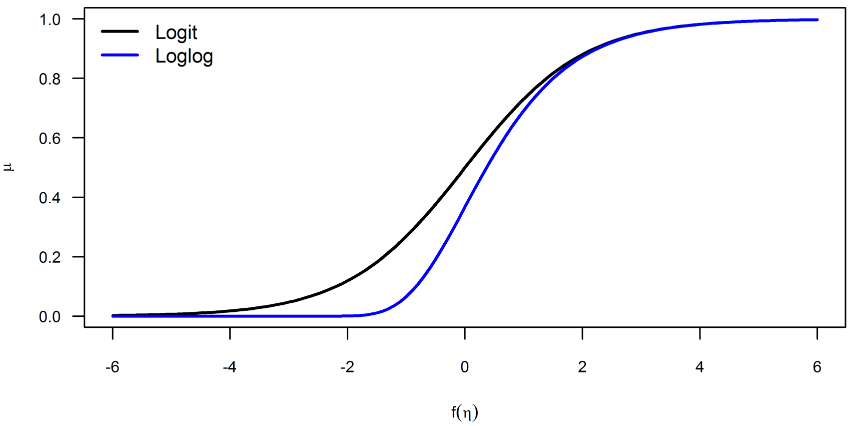

**Figure 16.11:** *Plot of the loglog function (upper curve) on top of the logit. Note the difference in shapes between the two curves. The asymmetric loglog function approaches its maximum value much slower than does the symmetric logit.*

models, we discussed the $R^2$ value as a measure of accuracy. However, we noted that adding variables to the model can never decrease the $R^2$ value, and will usually increase it. Thus, there is a pressure to increase the number of variables. However, science is guided by the philosophy of William of Occam and his Razor: Models should be as simple as possible, but no simpler. In other words, we should only include variables if the theory warrants it.

***Note***: Make no mistake, all models are wrong. As social scientists, we are searching for useful ones.

In linear regression, we corrected for the pressure to keep adding variables by using the adjusted $R^2$ as a guide. This value penalizes the model for the number of variables it has. Thus, unless the variable is statistically significant, there is no benefit to adding it to the model. This is why many scientists use the adjusted $R^2$ measure to help them determine the better model.

There is neither a true $R^2$ nor a true adjusted $R^2$ value for discrete dependent variable models. Thus, there has been much work in creating an appropriate measure to use for model selection. Three different measures are frequently used in the literature: Akaike's Information Criterion (Akaike 1974), Bayesian Information Criterion (Schwarz 1978), and Likelihood Ratio Test (Wilks 1938). Each of these three penalizes additional variables in a different manner and to a different degree. The one you select depends on the one available to you and the relationship between the two models.

### 16.7.1 Akaike Information Criterion

One of the first attempts to explicitly penalize for additional parameters (variables) was done by Hirotugu Akaike (1974). In his paper, he developed (albeit without mathematical rigor) a comparative measure of 'model goodness' that can be used to select the better of two models. The Akaike Information Criterion (AIC) score can be calculated whenever Maximum Likelihood Estimation is used to estimate the model parameters. The formula for the AIC is

$$AIC := -2\ln(\mathcal{L}) + 2k$$

Here, $k$ is the number of parameters being estimated in the model and $\mathcal{L}$ is the likelihood of the data with the model.[4]

The procedure to determine if one model is better than the other is straight-forward:

1. Calculate the AIC for Model A.
2. Calculate the AIC for Model B.
3. The model with the *lower* AIC score is the preferred model.

Its simplicity is its strength. Its weakness is that this measure, called the minimum information theoretical criterion (MAICE) in the paper, has no known probability distribution. As such, there is no way to determine whether the model with the lower AIC is *enough* better to justify eliminating the other from the discussion: If the AIC of Model 1 is 3 less than the AIC of Model 2, do we completely ignore Model 2?

This question actually leads to several "rules of thumb" that determine when that difference is "large enough." The rule I tend to use is that I drop a model if the AIC difference is greater than 8 (others use a threshold of 5).

That there is no statistical distribution to the AIC score only means the test is not optimal. In his paper, Akaike concurs (1974:722):

[4] The quantity $-2\ln(\mathcal{L})$ is often called the deviance of the model, which will be used in Section 16.7.3.

> Although the present author has no proof of optimality of MAICE it is at present the only procedure applicable to every situation where the likelihood can be properly defined and it is actually producing very reasonable results without very much amount of help of subjective judgement.

The R function that calculates the Akaike Information Criterion is AIC(). Using this function, the AIC for each of the three coin models are $AIC_{logit} = 118.48$, $AIC_{cloglog} = 119.74$, and $AIC_{loglog} = 117.05$. Thus, while the loglog is the better model from the AIC standpoint, it is not sufficiently better to completely ignore the other two models (the AIC improvement is not greater than 8). As such, this procedure is inconclusive with respect to the model we should choose.

***Note***: Please keep in mind that for the AIC to be valid in comparing models, the dependent variables must be the same across the models. If not, then this process cannot be used (nor any of these methods).

### 16.7.2 Bayesian Information Criterion

Akaike's paper did not give a mathematically solid reason why there should be a 2 point penalty for each additional estimated parameter (the $2k$ factor). This created an opening for other researchers to improve upon Akaike's proof and to create different penalty factors. Schwarz (1978) took Akaike's idea and put it on a more solid foundation. He humbly called his measure the Bayesian Information Criterion (BIC), others may refer to it as the Schwartz Information Criterion (SIC) or the Schwarz Bayesian Criterion (SBC).

Its formula is quite similar to the AIC:

$$BIC := -2\ln(\mathcal{L}) + k\log(n)$$

Here, $k$ is the number of parameters being estimated, $n$ is the number of data points, and $\mathcal{L}$ is the likelihood of the model. Thus, the difference between the AIC and the BIC is the effect of the additional parameter. In the AIC, each additional parameter penalizes the score by 2 points; in the BIC, $\log(n)$ points — usually a much greater penalty.

The process to select the better of two models is the same as for the AIC: Select the model with the lower BIC score. Furthermore, the requirement that the dependent variables are the same between the models remains.

There is actually no reason to prefer using the BIC over the AIC. As the penalty for the BIC is a function of the size of the data, it will not change between comparable models. Thus, if Model A has a lower AIC than Model B, it will also have a lower BIC.

### 16.7.3 Likelihood Ratio Test

Frequently, we wish to determine if a group of variables are jointly significant in the model. To do this, we compare the two nested models. We say that Model B is *nested* in Model A if Model A contains all the variables as Model B. For instance, let Model A contain the variables `X1`, `X2`, `X3`, `X4`, and `X5`. Let Model B contain variables `X1`, `X2`, and `X3`. Here, Model B is nested within Model A. Now, if we want to determine if variables `X4` and `X5` are jointly significant, then we merely compare Models A and B. To do this, we can use the AIC or the BIC, bu the Likelihood Ratio Test is more appropriate.

The Likelihood Ratio Test is superior to the AIC and BIC, when it can be used, because there is a known probability distribution for the test statistic. As such, we can determine whether Model A is *significantly* better than Model B — whether variables `X4` and `X5` are jointly significant.

The procedure is also straight forward:

1. Calculate the deviance for Model A.
2. Calculate the deviance for Model B.
3. The difference between the two deviances is distributed as a Chi-squared random variable with degrees of freedom equal to the parameter (variable) difference in the two models.

The deviance of a model is defined as

$$D := -2\ln(\mathcal{L})$$

Thus, if Model B is nested in Model A, the test statistic is equal to

$$X2 := D_B - D_A \sim \chi^2_{v_A - v_B}$$

Here, $v_A$ is the number of variables in Model A; $v_B$, in Model B.

*Example* 16.3. Let us assume that Model A uses three variables, $X1$, $X2$, and $X3$, and has a log-likelihood of -20, and Model B uses one variable, $X1$, and has a log-likelihood of -22. Are variables $X2$ and $X3$ jointly significant?

This is an application of the Likelihood Ratio Test. The test statistic is

$$\begin{aligned} X2 &:= D_B - D_A \\ &= \Big(-2\ln(\mathcal{L}_B)\Big) - \Big(-2\ln(\mathcal{L}_A)\Big) \\ &= \Big(-2(-22)\Big) - \Big(-2(-20)\Big) \\ &= 44 - 40 \\ &= 4 \end{aligned}$$

This test statistic is distributed as a Chi-squared random variable with $3-1 = 2$ degrees of freedom; that is,

$$X2 \sim \chi^2_2$$

A Chi-squared table gives us a p-value of approximately $p = 0.15$. This is close to what R gives us: `pchisq(4,df=2,lower.tail=FALSE)` = 0.135. Thus, we conclude at the $\alpha = 0.05$ level that we cannot reject the null hypothesis and we conclude that the restricted model is not significantly different from the full model; that is, we conclude that the two variables are not jointly significant and we can use Model B in lieu of Model A with little loss.

## 16.8: Conclusion

This chapter covered a lot of material. First, we examined how to fit binary dependent variable models. The GLM paradigm allows us to easily fit such models. As in all uses of the GLM paradigm, we need to know three things: the distribution of the dependent variable, the linear predictor, and the link function that connects the two.

For binary dependent variables, we need to realize that the dependent variable is distributed Binomially. The linear predictor is the usual combination of our independent variables. The canonical link is the logit link. Additional link functions include the probit, loglog, and complementary loglog functions.

The chapter proceeded to examine issues of determining how well a model fits the data. Accuracy, relative accuracy, and maximum accuracy measures were examined. Additionally, we examined the ROC curve and how it gives us additional information about our model.

Finally, we examined general techniques to select between two models. Three methods were examined. The first two did not require that the two models be nested. Both the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) only required that the dependent variables be the same. Also in both cases, the model with the lower score was the preferred model, although when that difference was less than 8 there was no reason to jettison the higher-scoring model.

The Likelihood Ratio Test was superior to the two Information Criterion tests as the test statistic has a known distribution. Thus, we could test the statistical significance of multiple variables at once. The drawback to using the Likelihood Ratio test is that the compared models needed to be nested. When determining the statistical significance of several variables at once, this is not an issue; when deciding to include Variable A *or* Variable B, it is an issue.

The next chapter continues our examination of discrete dependent variables. Frequently, our outcome variable is a *count* of events. In such a case, we cannot use the techniques discussed in this chapter as the dependent variable takes on more than just two values. We also cannot apply the techniques of Chapter 15, as the dependent variable is not continuous.

Staying in the realm of GLMs allows us to fit such variables easily. All we need to do is determine the appropriate distribution, the linear predictor, and the link function.

## 16.9: R Functions

In this chapter, we were introduced to several R functions that will be useful in the future. These are listed here.

### 16.9.0 Packages

**RFS**

**Epi**

### 16.9.0 Statistics

**lm(formula)** This function performs linear regression on the data, with the supplied formula. As there is much information contained in this function, you will want to save the results in a variable.

**glm(formula)** This function performs generalized linear model estimation on the given formula. There are three additional parameters that can (and often should) be specified.

The `family` parameter specifies the distributional family of the dependent variable, options include `gaussian`, `binomial`, `poisson`, `gamma`, `quasibinomial`, and `quasipoisson`. If this parameter is not specified, R assumes `gaussian`.

The `link` parameter specifies the link function for the distribution. If none is specified, the canonical link is assumed.

Finally, the `data` parameter specifies the data from which the formula variables come. This is the same parameter as in the `lm()` function.

**predict(model, newdata)** As with almost all statistical packages, R has a predict function. It takes two parameters, the model, and a dataframe of the independent values from which you want to predict. If you omit `newdata`, then it will predict based on the independent variables of the data itself, which can be used to calculate residuals. The dataframe must list all independent variables with their associate new values. You can specify multiple new values for a single independent variable.

**accuracy()** This function in the RFS package determines the predictive accuracy of a provided model. It takes three necessary parameters: data, truth, model, and threshold. It has the optional parameter of returning the *number* of correct classifications (rate=FALSE).

**AIC(model)** This function calculates the Akaike Informations Criterion score for the provided model. The model needs to have been fit using Maximum Likelihood Estimation.

**BIC(model)** This function in the RFS package calculates Schwarz's Bayesian Information Criterion (BIC) for the provided model.

**deviance()** This function returns the deviance in the model. This value is useful in the Likelihood Ratio Test.

**pchisq(x)** This gives the value of the cumulative distribution function (CDF) under the Chi-squared distribution. The necessary parameter is the number of degrees of freedom, df=. By default, it returns the lower-tail probability. Usually, we will want to have the upper-tail probability, thus we will use the lower.tail=FALSE parameter.

**var.test(x,y)** This function performs an F test, which compares the variances of two samples (x and y) from Normal populations. It can only compare two samples. If you need to compare more than two samples for equality of variance, you will need to perform either a Bartlett test or a Fligner-Killeen test.

### 16.9.0 Graphics

**ROC** This function in the Epi package performs ROC analysis on the data. It provides a ROC graph as well as some statistical values.

### 16.9.0 Programming

**for** This command is one of the basic control-constructs in the R language (as in most programming languages). The usual use is for(var in seq) expr, where var is the looping variable (the variable that equals the current loop number). The parameter seq is a vector of values. Usually, seq = something like 1:100, which is a vector of values from 1 to 100. Finally, expr is the expression (or series of expressions) that are performed for each value in the seq vector.

## 16.10: Extensions

This section offers suggestions on things you can practice from this chapter. Save the scripts in your Chapter 16 folder. For each of the following problems, please save the associated R script in the chapter folder as ext0x.R, where x is the problem number.

1. In Example 16.1, we suggested that you fit the provided pseudo data with linear regression and the OLS method. Please do so now and save the script as ext01.R.
2. From Section 16.6.2, please fit the coin data with the formula head~ trial and the loglog link. What is the predicted probability of getting a Head on Coin 15? Save this script as ext02.R.
3. Use the coinflip data (coinflips.csv ) to estimate the coin that is closest to being fair (a probability of producing a head is closest to 0.500). Use multiple link functions and select which you think is the best. Save this script as ext03.R.
4. Let us revisit the ssm data. One of the variables is passed, which is a binary variable indicating whether the ballot measure passed. Your job is to predict the proportion of voters in Maine who will vote in favor of the bill to ban same sex marriage. Do not use the pctFavor variable. Decide which model you are supposed to use. Prove that your model is the best model available. Make your prediction of the vote share. Include graphs if you would like, but only if the graph helps to illustrate your point. Save this script as ext04.R.

## 16.11: Applications

- Judi Bartfeld and Myoung Kim. (2010) "Participation in the School Breakfast Program: New Evidence from the ECLS-K." *Social Service Review* 84(4): 541–62.

- Regina P. Branton. (2009) "The Importance of Race and Ethnicity in Congressional Primary Elections." *Political Research Quarterly* 62(3): 459–73.

- Denise Gammonley, Ning Jackie Zhang, Kathryn Frahm, and Seung Chun Paek. (2009) "Social Service Staffing in U.S. Nursing Homes." *Social Service Review* 83(4): 633–50.

- Michael A. Neblo. (2009) "Meaning and Measurement: Reorienting the Race Politics Debate." *Political Research Quarterly* 62(3): 474–84.

- Lenna Nepomnyaschy and Irwin Garfinkel. (2011) "Fathers' Involvement with Their Nonresident Children and Material Hardship." *Social Service Review* 85(1): 3–38.

- Joseph G. Pickard, Megumi Inoue, Letha A. Chadiha, and Sharon Johnson. (2011) "The Relationship of Social Support to African American Caregivers' Help-Seeking for Emotional Problems." *Social Service Review* 85(2): 247–66.

- Brian Kelleher Richter, Krislert Samphantharak, and Jeffrey F. Timmons. (2009) "Lobbying and Taxes." *American Journal of Political Science* 53(4): 893–909.

- Lori E. Ross, Rachel Epstein, Corrie Goldfinger, and Christina Yager. (2009) "Policy and Practice regarding Adoption by Sexual and Gender Minority People in Ontario." *Canadian Public Policy / Analyse de Politiques* 35(4): 451–67.

## 16.12: References and Further Readings

- Hirotugu Akaike. (1974) “A New Look at Statistical Identification Model.” *IEEE Transactions on Automatic Control* 19(6): 716–23.

- Hirotugu Akaike. (1977) “On Entropy Maximization Principle.” In: P. R. Krishnaiah (Editor). *Applications of Statistics: Proceedings of the Symposium Held at Wright State University, Dayton, Ohio, 14-18 June 1976*. New York: North Holland Publishing, 27–41.

- George Casella and Roger L. Berger. (2002) *Statistical Inference*, Second edition. New York: Duxbury.

- Peter McCullagh and John A. Nelder. (1989) *Generalized Linear Models*. London: Chapman and Hall.

- Gideon E. Schwarz. (1978) “Estimating the dimension of a model.” *Annals of Statistics* 6(2): 461–64.

- Samuel S. Wilks (1938) “The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses.” *The Annals of Mathematical Statistics* 9(1): 60–62.