1
Logit Regression

Logit regression is a part of a larger class of generalized linear models (GLM). In this chapter we first briefly discuss GLMs, and then move on into a more in-depth discussion of logistic regression. Once again, the examples in this chapter use the `tbur` data set.

1.1 Generalized Linear Models

GLMs provide a modeling structure that can relate a linear model to response variables that do not have normal distributions. The distribution of Y is assumed to belong to one of an exponential family of distributions, including the Gaussian, Binomial, and Poisson distributions. GLMs are fit to the data by the method of maximum likelihood.

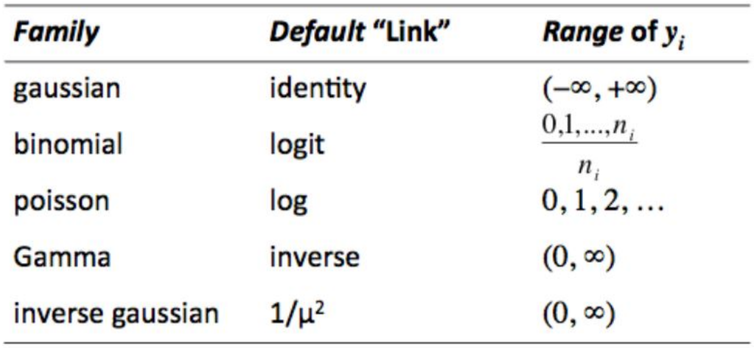
Like OLS, GLMs contain a stochastic component and a systematic component. The systematic component is expressed as:

$$\eta = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} \tag{1.1}$$

However, GLMs also contain a "link function" that relates the response variable, Y_i, to the systematic linear component, η. Table 16.1 shows the major exponential "families" of GLM models, and indicates the kinds of link functions involved in each. Note that OLS models would fall within the Gaussian family. In the next section we focus on the binomial family, and on logit estimation in particular.

1.2 Logit Estimation

Logit is used when predicting limited dependent variables, specifically those in which Y is represented by 0's and 1's. By virtue of the binary dependent variable, these models do not meet the key assumptions of OLS. Logit uses maximum likelihood estimation (MLE), which is a counterpart to minimizing least squares. MLE identifies the probability of obtaining the sample as a function of the model parameters (i.e., the X's). It answers the question, what are the values for B's that make the sample most likely? In other words, the likelihood function expresses the probability of obtaining the observed data as a function of the model



Family	***Default "Link"***	***Range of y_i***
gaussian	identity	$(-\infty, +\infty)$
binomial	logit	$\frac{0,1,...,n_i}{n_i}$
poisson	log	$0, 1, 2, \ldots$
Gamma	inverse	$(0, \infty)$
inverse gaussian	$1/\mu^2$	$(0, \infty)$

Figure 1.1: Exponential "families" of GLM Models

parameters. Estimates of A and B are based on maximizing a likelihood function of the observed Y values. In logit estimation we seek $P(Y=1)$, the probability that $Y=1$. The odds that $Y=1$ is expressed as:

$$O(Y=1) = \frac{P(Y=1)}{1-P(Y=1)}$$

Logits, L, are the natural logarithm of the odds:

$$\begin{aligned} L &= log_e O \\ &= log_e \frac{P}{1-P} \end{aligned}$$

They can range from $-\infty$, when $P=0$, to ∞, when $P=1$. L is the estimated systematic linear component:

$$L = A + B_1 X_{i1} + \ldots + B_k X_{ik}$$

By reversing the logit we can obtain the predicted probability that $Y=1$ for each of the i observations.

$$P_i = \frac{1}{1-e^{-L_i}} \tag{1.2}$$

where $e = 2.71828\ldots$, the base number of natural logarithms. Note that L is a linear function, but P is a non-linear S-shaped function as shown in Figure 1.2. Also note, that Equation 16.2 is the link function that relates the linear component to the non-linear response variable.

In more formal terms, each observation, i, contributes to the likelihood function by P_i if $Y_i = 1$, and by $1-P_i$ if $Y_i = 0$. This is defined as:

$$P_i^{Y_i}(1-P_i)^{1-Y_i}$$

The likelihood function is the product (multiplication) of all these individual contributions:

$$\ell = \prod P_i^{Y_i}(1-P_i)^{1-Y_i}$$

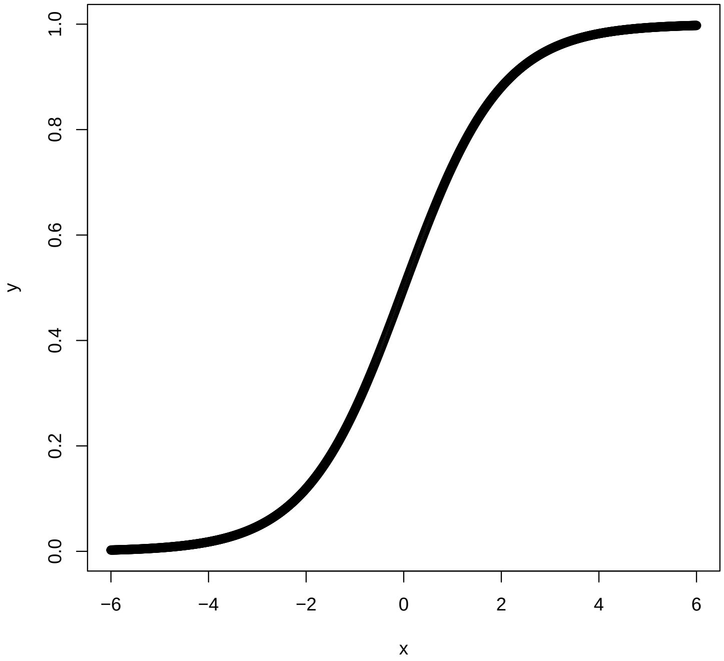


Figure 1.2: Predicted Probability as a Logit Function of X

The likelihood function is the largest for the model that best predicts $Y = 1$ or $Y = 0$, therefore when the predicted value of Y is correct and close to 1 or 0, the likelihood function is maximized.

To estimate the model parameters, we seek to maximize the log of the likelihood function. We use the log because it converts the multiplication into addition, and is therefore easier to calculate. The log likelihood is:

$$\log_e \ell = \sum_{i=1}^{n} [Y_i \log_e P_i + (1 - Y_i) \log_e (1 - P_i)]$$

The solution involves taking the first derivative of the log likelihood with respect to each of the B's, setting them to zero, and solving the simultaneous equation. The solution of the equation isn't linear, so it can't be solved directly. Instead, it's solved through a sequential estimation process that looks for successively better "fits" of the model.

For the most part, the key assumptions required for logit models are analogous to those required for OLS. The key differences are that (a) we do not assume a linear relationship between the Xs and Y, and (b) we do not assume normally distributed, homoscedastistic residuals. The key assumptions that are retained are shown below.

The following example uses demographic information to predict beliefs about anthropogenic climate change.

```
logit1 <- glm(glbcc ~ age + gender + education + income, data = ds.temp, family = binomial())
summary(logit1)
```

Logit Assumptions and Qualifiers

- The model is correctly specified
  - True conditional probabilities are logistic function of the X's
  - No important X's omitted; no extraneous X's included
  - No significant measurement error
- The cases are independent
- No X is a linear function of other X's
  - Increased multicollinearity leads to greater imprecision
- Influential cases can bias estimates
- Sample size: $n - k - 1$ should exceed 100
  - Independent covariation between the Xs and Y is critical

```
##
## Call:
## glm(formula = glbcc ~ age + gender + education + income, family = binomial(),
##     data = ds.temp)
##
## Deviance Residuals:
##    Min      1Q  Median      3Q     Max
## -1.707  -1.250   0.880   1.053   1.578
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.432e-01  2.344e-01   1.891 0.058689 .
## age         -1.079e-02  3.116e-03  -3.462 0.000535 ***
## gender      -3.131e-01  8.804e-02  -3.557 0.000375 ***
## education    1.580e-01  2.513e-02   6.288 3.22e-10 ***
## income      -2.380e-06  8.013e-07  -2.970 0.002977 **
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3114.5  on 2281  degrees of freedom
## Residual deviance: 3047.4  on 2277  degrees of freedom
## AIC: 3057.4
##
```

```
## Number of Fisher Scoring iterations: 4
```

As we can see age and gender are both negative and statistically significant predictors of climate change opinion. Below we discuss logit hypothesis tests, goodness of fit, and how to interpret the logit coefficients.

1.2.1 Logit Hypothesis Tests

In some ways, hypothesis testing with logit is quite similar to that using OLS. The same use of p-values is employed, however they differ in how they are derived. The logit analysis makes use of the Wald z-statistic, which is similar to the t-stat in OLS. The Wald z score compares the estimated coefficient to the asymptotic standard error, (aka the normal distribution). The p-value is derived from the asymptotic standard-normal distribution. Each estimated coefficient has a Wald z-score and a p-value that shows the probability that the null hypothesis is correct, given the data.

$$z = \frac{B_j}{SE(B_j)} \tag{1.3}$$

1.2.2 Goodness of Fit

Given that logit regression is estimated using MLE, the goodness-of-fit statistics differ from those of OLS. Here we examine three measures of fit: log-likelihood, the pseudo R^2, and the Akaike information criteria (AIC).

Log-Likelihood

To test for the overall null hypothesis that all B's are equal to zero, similar to an overall F-test in OLS, we can compare the log-likelihood of the demographic model with 4 IVs to the initial "null model" which includes only the intercept term. In general, a smaller log-likelihood indicates a better fit. Using the deviance statistic G^2 (aka the likelihood-ratio test statistic), we can determine whether the difference is statistically significant. G^2 is expressed as:

$$G^2 = 2(\log_e L_1 - \log_e L_0) \tag{1.4}$$

where L_1 is the demographic model and L_0 is the null model. The G^2 test statistic takes the difference between the log likelihoods of the two models and compares that to a χ^2 distribution with q degrees of freedom, where q is the difference in the number of IVs. We can calculate this in R. First, we run a null model predicting belief that greenhouse gases are causing the climate to change, using only the intercept:

```
logit0 <- glm(glbcc ~ 1, data = ds.temp)
summary(logit0)

##
## Call:
## glm(formula = glbcc ~ 1, data = ds.temp)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.5732  -0.5732   0.4268   0.4268   0.4268
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.57318    0.01036   55.35   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for gaussian family taken to be 0.2447517)
##
##     Null deviance: 558.28  on 2281  degrees of freedom
## Residual deviance: 558.28  on 2281  degrees of freedom
## AIC: 3267.1
##
## Number of Fisher Scoring iterations: 2
```

We then calculate the log likelihood for the null model,

$$\log_e L_0 \tag{1.5}$$

```
logLik(logit0)

## ’log Lik.’ -1631.548 (df=2)
```

Next, we calculate the log likelihood for the demographic model,

$$\log_e L_0 \tag{1.6}$$

Recall that we generated this model (dubbed “logit1”) earlier:

```
logLik(logit1)

## ’log Lik.’ -1523.724 (df=5)
```

Finally, we calculate the G statistic and perform the chi-square test for statistical significance:

```
G <- 2*(-1523 - (-1631))
G

## [1] 216

pchisq(G, df = 3, lower.tail = FALSE)

## [1] 1.470144e-46
```

We can see by the very low p-value that the demographic model offers a significant improvement in fit.

The same approach can be used to compare nested models, similar to nested F-tests in OLS. For example, we can include ideology in the model and use the `anova` function to see if the ideology variable improves model fit. Note that we specify the χ^2 test.

```
logit2 <- glm(glbcc ˜ age + gender + education + income + ideol, family = binomial(),
    data = ds.temp)
summary(logit2)
```

```
##
## Call:
## glm(formula = glbcc ˜ age + gender + education + income + ideol,
##     family = binomial(), data = ds.temp)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6661  -0.8939   0.3427   0.8324   2.0212
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  4.055e+00  3.211e-01  12.629  < 2e-16 ***
## age         -4.287e-03  3.630e-03  -1.181 0.237701
## gender      -2.044e-01  1.023e-01  -1.998 0.045702 *
## education    1.009e-01  2.934e-02   3.440 0.000582 ***
## income      -1.042e-06  8.939e-07  -1.166 0.243485
## ideol       -7.900e-01  3.763e-02 -20.993  < 2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 3114.5  on 2281  degrees of freedom
## Residual deviance: 2404.0  on 2276  degrees of freedom
## AIC: 2416
##
## Number of Fisher Scoring iterations: 4
```

```
anova(logit1, logit2, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model 1: glbcc ˜ age + gender + education + income
## Model 2: glbcc ˜ age + gender + education + income + ideol
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1      2277     3047.4
## 2      2276     2404.0  1   643.45 < 2.2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
```

As we can see, adding ideology significantly improves the model.

Pseudo R^2

A measure that is equivalent to the R^2 in OLS does not exist for logit. Remember that explaining variance in Y is not the goal of MLE. However, a "pseudo" R^2 measure exists that compares the residual deviance of the null model with that of the full model. Like the R^2 measure, pseudo R^2 ranges from 0 to 1 with values closer to 1 indicating improved model fit.

Deviance is analogous to the residual sum of squares for a linear model. It is expressed as:

$$\text{deviance} = -2(\log_e L) \tag{1.7}$$

It is simply the log-likelihood of the model multiplied by a -2. The pseudo R^2 is 1 minus the ratio of the deviance of the full model L_1 to the deviance of the null model L_0:

$$\text{pseudo}R^2 = 1 - \frac{-2(\log_e L_1)}{-2(\log_e L_0)} \tag{1.8}$$

This can be calculated in `R` using the full model with ideology.

```
pseudoR2 <- 1 - (logit2$deviance/logit2$null.deviance)
pseudoR2

## [1] 0.2281165
```

The pseudo R^2 of the model is 0.2281165. Note that the 0.2281165 is only an approximation of explained variance, and should be used in combination with other measures of fit such as AIC.

Akaike Information Criteria

Another way to examine goodness-of-fit is the Akaike information criteria (AIC). Like the adjusted R^2 for OLS, the AIC takes into account the parsimony of the model by penalizing for the number of parameters. But AIC is useful only in a comparative manner – either with the null model or an alternative model. It does not purport to describe the percent of variance in Y accounted for, as does the pseudo R^2.

AIC is defined as -2 times the residual deviance of the model plus two times the number of parameters; k IVs plus the intercept:

$$\text{AIC} = -2(\log_e L) + 2(k+1) \tag{1.9}$$

Note that smaller values are indicative of a better fit. The AIC is most useful when comparing the fit of alternative (not necessarily nested) models. In R, AIC is given as part of the `summary` output for a `glm` object, but we can also calculate it and verify.

```
aic.logit2 <- logit2$deviance + 2*6
aic.logit2

## [1] 2416.002
```

```
logit2$aic

## [1] 2416.002
```

1.2.3 Interpreting Logits

The logits, L, are logged odds, and therefore the coefficients that are produced must be interpreted as logged odds. This means that for each unit change in ideology the predicted logged odds of believing climate change has an anthropogenic cause decrease by -0.7900119. This interpretation, through mathematically straightforward, is not terribly informative. Below we discuss two ways to make the interpretation of logit analysis more intuitive.

Calculate Odds

Logits can be used to directly calculate odds by taking the antilog of any of the coefficients:

$$anti\log = e^B$$

For example, the following retuns odds for all the IVs.

```
logit2 %>% coef() %>% exp()

## (Intercept)          age       gender    education       income        ideol
##   57.6608736    0.9957225    0.8151353    1.1062128    0.9999990    0.4538394
```

Therefore, for each 1-unit increase in the ideology scale (i.e., becoming more conservative) the odds of believing that climate change is human caused decrease by 0.4538394.

Predicted Probabilities

The most straightforward way to interpret logits is to Equation 1.2.3. To calculate the effect of a particular independent variable, X_i, on the probability of $Y = 1$, set all X_j's at their means, then calculate:

$$\hat{P} = \frac{1}{1 + e^{-\hat{L}}}$$

We can then evaluate the change in predicted probabilities that Y=1 across the range of values in X_i.

This procedure can be demonstrated in a two steps. First, create a data frame holding all the variables except ideology at their mean. Second, use the `predict` function to calculate the predicted probabilities for each level of ideology. Indicate `type = "response"`.

```
log.data <- data.frame(age = mean(ds.temp$age), gender = mean(ds.temp$gender),
    education = mean(ds.temp$education), income = mean(ds.temp$income), ideol = 1:7)
log.data$predicted <- predict(logit2, newdata = log.data, type = "response")
log.data
```

```
##        age    gender education   income ideol predicted
## 1 60.10517 0.4119194  5.093777 70627.24     1 0.9665366
## 2 60.10517 0.4119194  5.093777 70627.24     2 0.9291203
## 3 60.10517 0.4119194  5.093777 70627.24     3 0.8560967
## 4 60.10517 0.4119194  5.093777 70627.24     4 0.7297255
## 5 60.10517 0.4119194  5.093777 70627.24     5 0.5506305
## 6 60.10517 0.4119194  5.093777 70627.24     6 0.3573709
## 7 60.10517 0.4119194  5.093777 70627.24     7 0.2015226
```

The output shows, for each case, the ideology measure for the respondent followed by the estimated probability (p) that the individual believes man-made greenhouse gasses are causing climate change. We can also graph the results, with 95% confidence intervals. This is shown in Figure 1.3.

```
preds <- predict(logit2, newdata = log.data, se.fit = T, type = "link")
lower <- plogis(with(preds, fit - 1.96 * se.fit))
upper <- plogis(with(preds, fit + 1.96 * se.fit))
pred <- log.data$predicted
log.df <- data.frame(pred, lower, upper, ideol = 1:7)

ggplot(log.df, aes(ideol, pred)) + geom_point() + geom_errorbar(ymin = lower,
    ymax = upper, width = 0.2)
dev.off()
```

We can see that as respondents become more conservative, the probability of believing that climate change is man-made decreases at what appears to be an increasing rate.

1.3 Summary

As an analysis and research tool, logit modeling expands your capabilities beyond those that can reasonably be estimated with OLS. Now you can accommodate models with binary dependent variables. Logit models are a family of generalized linear models that are useful for predicting the odds or probabilities, of outcomes for binary dependent variables. This chapter has described the manner in which logits are calculated, how model fit can be characterized, and several methods for making the logit results readily interpretable.

Perhaps one of the greatest difficulties in applications of logit models is the clear communication of the meaning of the results. The estimated coefficients show the change in the log of the odds for a one unit increase in the X variable – not the usual way to describe effects. But, as described in this chapter, these estimated coefficients can be readily transformed into changes in the odds, or the logit itself can be "reversed" to provide estimated probabilities. Of particular utility are logit graphics, showing the estimated shift in Y from values of zero to one; the estimated probabilities of Y=1 for cases with specified combinations of values in the X variables; and estimates of the ranges of probabilities for Y=1 across the ranges of values in any X.

In sum, the use of logit models will expand your ability to test hypotheses to include a range of outcomes that are binary in nature. Given that a great deal of the phenomena of interest in the policy and social sciences are of this kind, you will find this capability to be an important part of your research toolkit.

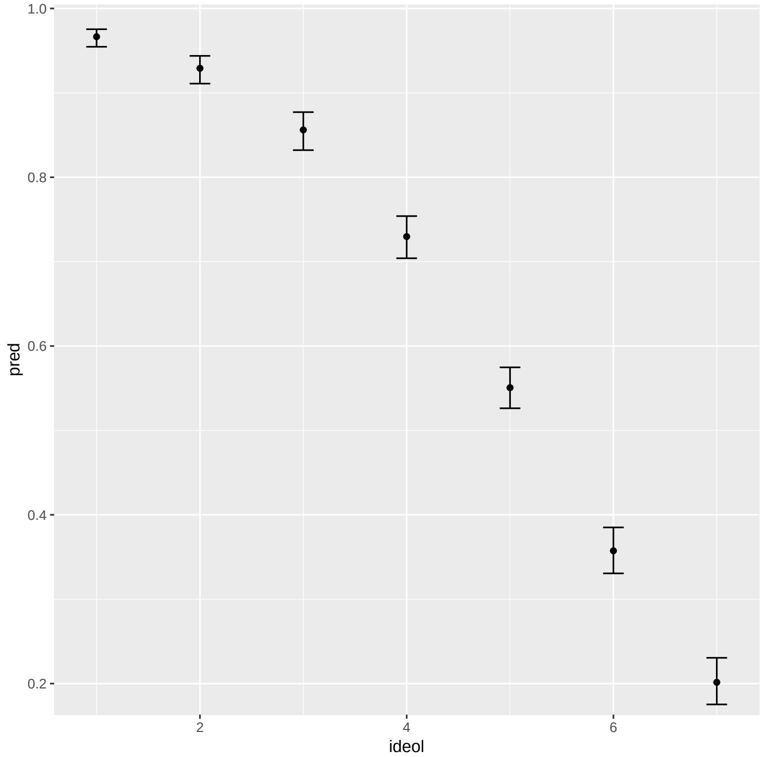


Figure 1.3: Predicted Probability of believing that Greenhouse Gases cause Climate Change by Ideology