CHAPTER 5

Logistic regression

Logistic regression is the standard way to model binary outcomes (that is, data y_i that take on the values 0 or 1). Section 5.1 introduces logistic regression in a simple example with one predictor, then for most of the rest of the chapter we work through an extended example with multiple predictors and interactions.

5.1 Logistic regression with a single predictor

Example: modeling political preference given income

Conservative parties generally receive more support among voters with higher incomes. We illustrate classical logistic regression with a simple analysis of this pattern from the National Election Study in 1992. For each respondent i in this poll, we label $y_i = 1$ if he or she preferred George Bush (the Republican candidate for president) or 0 if he or she preferred Bill Clinton (the Democratic candidate), for now excluding respondents who preferred Ross Perot or other candidates, or had no opinion. We predict preferences given the respondent's income level, which is characterized on a five-point scale.[1]

The data are shown as (jittered) dots in Figure 5.1, along with the fitted *logistic regression* line, a curve that is constrained to lie between 0 and 1. We interpret the line as the probability that $y = 1$ given x—in mathematical notation, $\Pr(y = 1|x)$.

We fit and display the logistic regression using the following R function calls:

R code
```
fit.1 <- glm (vote ~ income, family=binomial(link="logit"))
display (fit.1)
```

to yield

R output
```
             coef.est coef.se
(Intercept)    -1.40    0.19
income          0.33    0.06
  n = 1179, k = 2
  residual deviance = 1556.9, null deviance = 1591.2 (difference = 34.3)
```

The fitted model is $\Pr(y_i = 1) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$. We shall define this model mathematically and then return to discuss its interpretation.

The logistic regression model

It would not make sense to fit the continuous linear regression model, $X\beta + \text{error}$, to data y that take on the values 0 and 1. Instead, we model the probability that $y = 1$,

$$\Pr(y_i = 1) = \text{logit}^{-1}(X_i\beta), \tag{5.1}$$

under the assumption that the outcomes y_i are independent given these probabilities. We refer to $X\beta$ as the *linear predictor*.

[1] See Section 4.7 for details on the income categories and other variables measured in this survey.

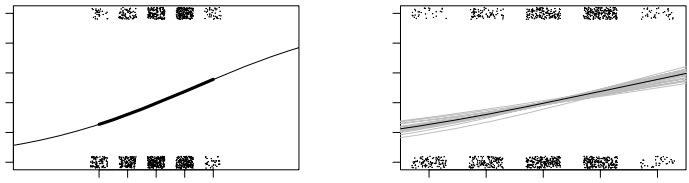

Figure 5.1 *Logistic regression estimating the probability of supporting George Bush in the 1992 presidential election, as a function of discretized income level. Survey data are indicated by jittered dots. In this example little is revealed by these jittered points, but we want to emphasize here that the data and fitted model can be put on a common scale. (a) Fitted logistic regression: the thick line indicates the curve in the range of the data; the thinner lines at the end show how the logistic curve approaches 0 and 1 in the limits. (b) In the range of the data, the solid line shows the best-fit logistic regression, and the light lines show uncertainty in the fit.*

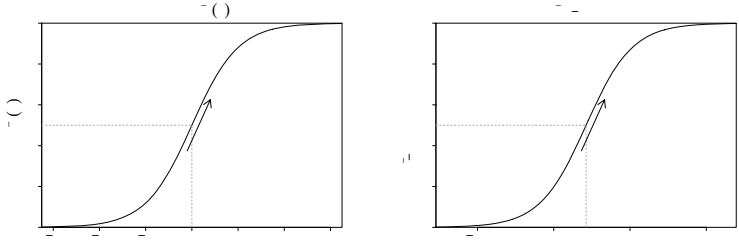

Figure 5.2 *(a) Inverse-logit function $\text{logit}^{-1}(x)$: the transformation from linear predictors to probabilities that is used in logistic regression. (b) An example of the predicted probabilities from a logistic regression model: $y = \text{logit}^{-1}(-1.40 + 0.33x)$. The shape of the curve is the same, but its location and scale have changed; compare the x-axes on the two graphs. For each curve, the dotted line shows where the predicted probability is 0.5: in graph (a), this is at $\text{logit}(0.5) = 0$; in graph (b), the halfway point is where $-1.40 + 0.33x = 0$, which is $x = 1.40/0.33 = 4.2$.*
The slope of the curve at the halfway point is the logistic regression coefficient divided by 4, thus 1/4 for $y = \text{logit}^{-1}(x)$ and 0.33/4 for $y = \text{logit}^{-1}(-1.40 + 0.33x)$. The slope of the logistic regression curve is steepest at this halfway point.

The function $\text{logit}^{-1}(x) = \frac{e^x}{1+e^x}$ transforms continuous values to the range $(0, 1)$, which is necessary, since probabilities must be between 0 and 1. This is illustrated for the election example in Figure 5.1 and more theoretically in Figure 5.2.

Equivalently, model (5.1) can be written

$$\begin{aligned} \Pr(y_i = 1) &= p_i \\ \text{logit}(p_i) &= X_i\beta, \end{aligned} \tag{5.2}$$

where $\text{logit}(x) = \log(x/(1-x))$ is a function mapping the range $(0, 1)$ to the range $(-\infty, \infty)$. We prefer to work with logit^{-1} because it is natural to focus on the mapping from the linear predictor to the probabilities, rather than the reverse. However, you will need to understand formulation (5.2) to follow the literature and also when fitting logistic models in Bugs.

The inverse-logistic function is curved, and so the expected difference in y corresponding to a fixed difference in x is not a constant. As can be seen in Figure 5.2, the steepest change occurs at the middle of the curve. For example:

- $\text{logit}(0.5) = 0$, and $\text{logit}(0.6) = 0.4$. Here, adding 0.4 on the logit scale corresponds to a change from 50% to 60% on the probability scale.
- $\text{logit}(0.9) = 2.2$, and $\text{logit}(0.93) = 2.6$. Here, adding 0.4 on the logit scale corresponds to a change from 90% to 93% on the probability scale.

Similarly, adding 0.4 at the low end of the scale moves a probability from 7% to 10%. In general, any particular change on the logit scale is compressed at the ends of the probability scale, which is needed to keep probabilities bounded between 0 and 1.

5.2 Interpreting the logistic regression coefficients

Coefficients in logistic regression can be challenging to interpret because of the nonlinearity just noted. We shall try to generalize the procedure for understanding coefficients one at a time, as was done for linear regression in Chapter 3. We illustrate with the model, $\Pr(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$. Figure 5.1 shows the story, but we would also like numerical summaries. We present some simple approaches here and return in Section 5.7 to more comprehensive numerical summaries.

Evaluation at and near the mean of the data

The curve of the logistic function requires us to choose where to evaluate changes, if we want to interpret on the probability scale. The mean of the input variables in the data is often a useful starting point.

- As with linear regression, the *intercept* can only be interpreted assuming zero values for the other predictors. When zero is not interesting or not even in the model (as in the voting example, where income is on a 1–5 scale), the intercept must be evaluated at some other point. For example, we can evaluate Pr(Bush support) at the central income category and get $\text{logit}^{-1}(-1.40 + 0.33 \cdot 3) = 0.40$.

  Or we can evaluate Pr(Bush support) at the mean of respondents' incomes: $\text{logit}^{-1}(-1.40 + 0.33 \cdot \bar{x})$; in *R* we code this as[2]

  ```
  invlogit (-1.40 + 0.33*mean(income))
  ```
  R code

  or, more generally,

  ```
  invlogit (coef(fit.1)[1] + coef(fit.1)[2]*mean(income))
  ```
  R code

  For this dataset, $\bar{x} = 3.1$, yielding $\Pr(\text{Bush support}) = 0.40$ at this central point.
- A difference of 1 in income (on this 1–5 scale) corresponds to a positive difference of 0.33 in the logit probability of supporting Bush. There are two convenient ways to summarize this directly in terms of probabilities.
  - We can evaluate how the probability differs with a unit difference in x near the central value. Since $\bar{x} = 3.1$ in this example, we can evaluate the logistic regression function at $x = 3$ and $x = 2$; the difference in $\Pr(y = 1)$ corresponding to adding 1 to x is $\text{logit}^{-1}(-1.40+0.33\cdot3) - \text{logit}^{-1}(-1.40+0.33\cdot2) = 0.08$.

[2] We are using a function we have written, `invlogit <- function (x) {1/(1+exp(-x))}`.

A difference of 1 in income category corresponds to a positive difference of 8% in the probability of supporting Bush.

- Rather than consider a discrete change in x, we can compute the derivative of the logistic curve at the central value, in this case $\bar{x} = 3.1$. Differentiating the function $\text{logit}^{-1}(\alpha + \beta x)$ with respect to x yields $\beta e^{\alpha+\beta x}/(1 + e^{\alpha+\beta x})^2$. The value of the linear predictor at the central value of $\bar{x} = 3.1$ is $-1.40+0.33\cdot 3.1 = -0.39$, and the slope of the curve—the "change" in $\Pr(y=1)$ per small unit of "change" in x—at this point is $0.33e^{-0.39}/(1 + e^{-0.39})^2 = 0.13$.
- For this example, the difference on the probability scale is the same value of 0.13 (to one decimal place); this is typical but in some cases where a unit difference is large, the differencing and the derivative can give slightly different answers. They will always be the same sign, however.

The "divide by 4 rule"

The logistic curve is steepest at its center, at which point $\alpha + \beta x = 0$ so that $\text{logit}^{-1}(\alpha+\beta x) = 0.5$ (see Figure 5.2). The slope of the curve—the derivative of the logistic function—is maximized at this point and attains the value $\beta e^0/(1+e^0)^2 = \beta/4$. Thus, $\beta/4$ is the maximum difference in $\Pr(y = 1)$ corresponding to a unit difference in x.

As a rule of convenience, we can take logistic regression coefficients (other than the constant term) and divide them by 4 to get an upper bound of the predictive difference corresponding to a unit difference in x. This upper bound is a reasonable approximation near the midpoint of the logistic curve, where probabilities are close to 0.5.

For example, in the model $\Pr(\text{Bush support}) = \text{logit}^{-1}(-1.40 + 0.33 \cdot \text{income})$, we can divide 0.33/4 to get 0.08: a difference of 1 in income category corresponds to no more than an 8% positive difference in the probability of supporting Bush. Because the data in this case actually lie near the 50% point (see Figure 5.1), this "divide by 4" approximation turns out to be close to 0.13, the derivative evaluated at the central point of the data.

Interpretation of coefficients as odds ratios

Another way to interpret logistic regression coefficients is in terms of *odds ratios*. If two outcomes have the probabilities $(p, 1-p)$, then $p/(1-p)$ is called the *odds*. An odds of 1 is equivalent to a probability of 0.5—that is, equally likely outcomes. Odds of 0.5 or 2.0 represent probabilities of $(1/3, 2/3)$. The ratio of two odds—thus, $(p_1/(1-p_1))/(p_2/(1-p_2))$—is called an odds ratio. Thus, an odds ratio of 2 corresponds to a change from $p = 0.33$ to $p = 0.5$, or a change from $p = 0.5$ to $p = 0.67$.

An advantage of working with odds ratios (instead of probabilities) is that it is possible to keep scaling up odds ratios indefinitely without running into the boundary points of 0 and 1. For example, going from an odds of 2 to an odds of 4 increases the probability from 2/3 to 4/5; doubling the odds again increases the probability to 8/9, and so forth.

Exponentiated logistic regression coefficients can be interpreted as odds ratios. For simplicity, we illustrate with a model with one predictor, so that

$$\log\left(\frac{\Pr(y=1|x)}{\Pr(y=0|x)}\right) = \alpha + \beta x. \tag{5.3}$$

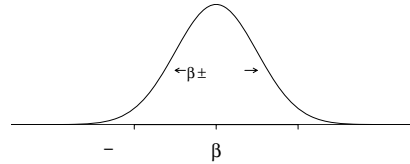

Figure 5.3 *Distribution representing uncertainty in an estimated regression coefficient (repeated from page 40). The range of this distribution corresponds to the possible values of β that are consistent with the data. When using this as an uncertainty distribution, we assign an approximate 68% chance that β will lie within 1 standard error of the point estimate, $\hat{\beta}$, and an approximate 95% chance that β will lie within 2 standard errors. Assuming the regression model is correct, it should happen only about 5% of the time that the estimate, $\hat{\beta}$, falls more than 2 standard errors away from the true β.*

Adding 1 to x (that is, changing x to $x+1$ in (5.3)) has the effect of adding β to both sides of the equation. Exponentiating both sides, the odds are then multiplied by e^{β}. For example, if $\beta = 0.2$, then a unit difference in x corresponds to a multiplicative change of $e^{0.2} = 1.22$ in the odds (for example, changing the odds from 1 to 1.22, or changing p from 0.5 to 0.55).

We find that the concept of odds can be somewhat difficult to understand, and odds ratios are even more obscure. Therefore we prefer to interpret coefficients on the original scale of the data when possible.for example, saying that adding 0.2 on the logit scale corresponds to a change in probability from $\text{logit}^{-1}(0)$ to

Inference

Coefficient estimates and standard errors. The coefficients in classical logistic regression are estimated using maximum likelihood, a procedure that can often work well for models with few predictors fit to reasonably large samples (but see Section 5.8 for a potential problem).

As with the linear model, the standard errors represent estimation uncertainty. We can roughly say that coefficient estimates within 2 standard errors of $\hat{\beta}$ are consistent with the data. Figure 5.3 shows the normal distribution that approximately represents the range of possible values of β. For the voting example, the coefficient of income has an estimate $\hat{\beta}$ of 0.33 and a standard error of 0.06; thus the data are roughly consistent with values of β in the range $[0.33 \pm 2 \cdot 0.06] = [0.21, 0.45]$.

Statistical significance. As with linear regression, a coefficient is considered "statistically significant" if it is at least 2 standard errors away from zero. In the voting example, the coefficient of income is statistically significant and positive, meaning that we can be fairly certain that, in the population represented by this survey, positive differences in income generally correspond to positive (not negative) differences in the probability of supporting Bush for president.

Also as with linear regression, we usually do *not* try to interpret the statistical significance of the intercept. The sign of an intercept is not generally of any interest, and so it is usually meaningless to compare it to zero or worry about whether it is statistically significantly different from zero.

Finally, when considering multiple inputs, we follow the same principles as with linear regression when deciding when and how to include and combine inputs in a model, as discussed in Section 4.6.

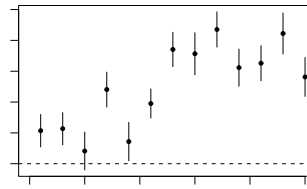

Figure 5.4 *Coefficient of income (on a 1–5 scale) with ±1 standard-error bounds in logistic regressions predicting Republican preference for president, as estimated separately from surveys in the second half of the twentieth century. The pattern of richer voters supporting Republicans has increased since 1970. The data used in the estimate for 1992 appear in Figure 5.1.*

Predictions. Logistic regression predictions are probabilistic, so for each unobserved future data point $\tilde{y}_i$, there is a predictive probability,

$$\tilde{p}_i = \Pr(\tilde{y}_i = 1) = \text{logit}^{-1}(\tilde{X}_i\beta),$$

rather than a point prediction. For example, for a voter not in the survey with income level 5 (recall the 5-point scale in Figure 5.1), the predicted *probability* of supporting Bush is $\Pr(\tilde{y}_i = 1) = \text{logit}^{-1}(-1.40 + 0.33 \cdot 5) = 0.55$. We do not say that our prediction for the *outcome* is 0.55, since the outcome $\tilde{y}_i$—support for Bush or not—itself will be 0 or 1.

Fitting and displaying the model in R

After fitting the logistic regression using the `glm` function (see page 79), we can graph the data and fitted line (see Figure 5.1a) as follows:

R code
```
plot (income, vote)
curve (invlogit (coef(fit.1)[1] + coef(fit.1)[2]*x), add=TRUE)
```

(The R code we actually use to make the figure has more steps so as to display axis labels, jitter the points, adjust line thickness, and so forth.) Figure 5.1b has dotted lines representing uncertainty in the coefficients; we display these by adding the following to the plotting commands:

R code
```
sim.1 <- sim (fit.1)
for (j in 1:10){
  curve (invlogit (sim.1$beta[j,1] + sim.1$beta[j,2]*x),
    col="gray", lwd=.5, add=TRUE)}
```

We demonstrate further use of the `sim` function in Chapter 7.

Displaying the results of several logistic regressions

We can display estimates from a series of logistic regressions in a single graph, just as was done in Section 4.7 for linear regression coefficients. Figure 5.4 illustrates with the estimate ±1 standard error for the coefficient for income on presidential preference, fit to National Election Studies pre-election polls from 1952 through 2000. Higher income has consistently been predictive of Republican support, but the connection has become stronger over the years.

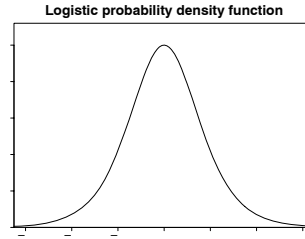

Figure 5.5 *The probability density function of the logistic distribution, which is used for the error term in the latent-data formulation (5.4) of logistic regression. The logistic curve in Figure 5.2a is the cumulative distribution function of this density. The maximum of the density is 0.25, which corresponds to the maximum slope of 0.25 in the inverse-logit function of Figure 5.2a.*

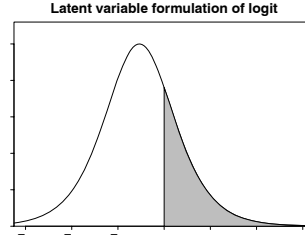

Figure 5.6 *The probability density function of the latent variable z_i in model (5.4) if the linear predictor, $X_i\beta$, has the value -1.07. The shaded area indicates the probability that $z_i > 0$, so that $y_i = 1$ in the logistic regression.*

5.3 Latent-data formulation

We can interpret logistic regression directly—as a nonlinear model for the probability of a "success" or "yes" response given some predictors—and also indirectly, using what are called unobserved or *latent* variables. In this formulation, each discrete outcome y_i is associated with a continuous, unobserved outcome z_i, defined as follows:

$$\begin{aligned} y_i &= \begin{cases} 1 & \text{if } z_i > 0 \\ 0 & \text{if } z_i < 0 \end{cases} \\ z_i &= X_i\beta + \epsilon_i, \end{aligned} \tag{5.4}$$

with independent errors ϵ_i that have the *logistic* probability distribution. The logistic distribution is shown in Figure 5.5 and is defined so that

$$\Pr(\epsilon_i < x) = \text{logit}^{-1}(x) \text{ for all } x.$$

Thus, $\Pr(y_i = 1) = \Pr(z_i > 0) = \Pr(\epsilon_i > -X_i\beta) = \text{logit}^{-1}(X_i\beta)$, and so models (5.1) and (5.4) are equivalent.

Figure 5.6 illustrates for an observation i with income level $x_i = 1$ (that is, a person in the lowest income category), whose linear predictor, $X_i\beta$, thus has the value $-1.40 + 0.33 \cdot 1 = -1.07$. The curve illustrates the distribution of the latent variable z_i, and the shaded area corresponds to the probability that $z_i > 0$, so that $y_i = 1$. In this example, $\Pr(y_i = 1) = \text{logit}^{-1}(-1.07) = 0.26$.

Interpretation of the latent variables

Latent variables are a computational trick but they can also be interpreted substantively. For example, in the pre-election survey, $y_i = 1$ for Bush supporters and 0 for Clinton supporters. The unobserved continuous z_i can be interpreted as the respondent's "utility" or preference for Bush, compared to Clinton: the sign of the utility tells us which candidate is preferred, and its magnitude reveals the strength of the preference.

Only the sign of z_i, not its magnitude, can be determined directly from binary data. However, we can learn more about the z_i's given the logistic regression predictors. In addition, in some settings direct information is available about the z_i's; for example, a survey can ask "feeling thermometer" questions such as, "Rate your feelings about George Bush on a 1–10 scale, with 1 being the most negative and 10 being the most positive."

Nonidentifiability of the latent variance parameter

The logistic probability density function in Figure 5.5 appears bell-shaped, much like the normal density that is used for errors in linear regression. In fact, the logistic distribution is very close to the normal distribution with mean 0 and standard deviation 1.6—an identity that we discuss further on page 118 in the context of "probit regression." For now, we merely note that the logistic model (5.4) for the latent variable z is closely approximated by the normal regression model,

$$z_i = X_i\beta + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, \sigma^2), \tag{5.5}$$

with $\sigma = 1.6$. This then raises the question, why not estimate σ?

We cannot estimate the parameter σ in model (5.5) because it is not identified when considered jointly with the regression parameter β. If all the elements of β are multiplied by a positive constant and σ is also multiplied by that constant, then the model does not change. For example, suppose we fit the model

$$z_i = -1.40 + 0.33x_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, 1.6^2).$$

This is equivalent to the model

$$z_i = -14.0 + 3.3x_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, 16^2),$$

or

$$z_i = -140 + 33x_i + \epsilon_i, \quad \epsilon_i \sim \text{N}(0, 160^2).$$

As we move from each of these models to the next, z is multiplied by 10, but the *sign* of z does not change. Thus all the models have the same implications for the observed data y: for each model, $\Pr(y_i = 1) \approx \text{logit}^{-1}(-1.40 + 0.33x_i)$ (only approximate because the logistic distribution is not exactly normal).

Thus, model (5.5) has an essential indeterminacy when fit to binary data, and it is standard to resolve this by setting the variance parameter σ to a fixed value, for example 1.6, which is essentially equivalent to the unit logistic distribution.

5.4 Building a logistic regression model: wells in Bangladesh

We illustrate the steps of building, understanding, and checking the fit of a logistic regression model using an example from economics (or perhaps it is psychology, or public health): modeling the decisions of households in Bangladesh about whether to change their source of drinking water.

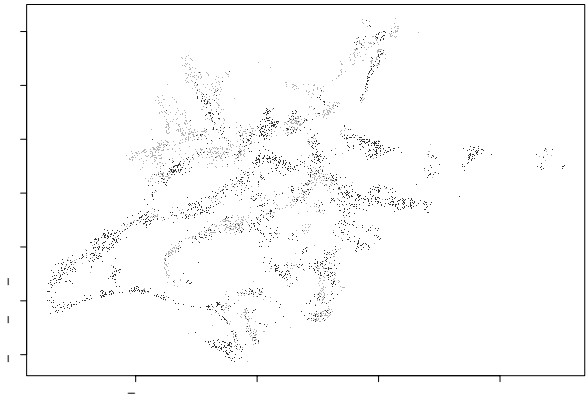

Figure 5.7 *Wells in an area of Araihazar upazila, Bangladesh. Light and dark dots represent wells with arsenic greater than and less than the safety standard of 0.5 (in units of hundreds of micrograms per liter). (The wells are located where people live. The empty areas between the wells are mostly cropland.) Safe and unsafe wells are intermingled in most of the area, which suggests that users of unsafe wells can switch to nearby safe wells.*

Background

Many of the wells used for drinking water in Bangladesh and other South Asian countries are contaminated with natural arsenic, affecting an estimated 100 million people. Arsenic is a cumulative poison, and exposure increases the risk of cancer and other diseases, with risks estimated to be proportional to exposure.

Any locality can include wells with a range of arsenic levels, as can be seen from the map in Figure 5.7 of all the wells in a collection of villages in a small area of Bangladesh. The bad news is that even if your neighbor's well is safe, it does not mean that yours is safe. However, the corresponding good news is that, if your well has a high arsenic level, you can probably find a safe well nearby to get your water from—if you are willing to walk the distance and your neighbor is willing to share. (The amount of water needed for drinking is low enough that adding users to a well would not exhaust its capacity, and the surface water in this area is subject to contamination by microbes, hence the desire to use water from deep wells.)

In the area shown in Figure 5.7, a research team from the United States and Bangladesh measured all the wells and labeled them with their arsenic level as well as a characterization as "safe" (below 0.5 in units of hundreds of micrograms per liter, the Bangladesh standard for arsenic in drinking water) or "unsafe" (above 0.5). People with unsafe wells were encouraged to switch to nearby private or community wells or to new wells of their own construction.

A few years later, the researchers returned to find out who had switched wells. We shall perform a logistic regression analysis to understand the factors predictive of well switching among the users of unsafe wells. In the notation of the previous section, our outcome variable is

$$y_i = \begin{cases} 1 & \text{if household } i \text{ switched to a new well} \\ 0 & \text{if household } i \text{ continued using its own well.} \end{cases}$$

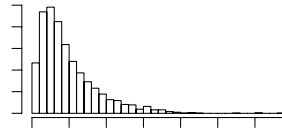

Figure 5.8 *Histogram of distance to the nearest safe well, for each of the unsafe wells in the Araihazar dataset (see Figure 5.7).*

We consider the following inputs:

- A constant term
- The distance (in meters) to the closest known safe well
- The arsenic level of respondent's well
- Whether any members of the household are active in community organizations
- The education level of the head of household.

We shall first fit the model just using distance to nearest well and then put in arsenic concentration, organizational membership, and education.

Logistic regression with just one predictor

We fit the logistic regression in R:

R code
```
fit.1 <- glm (switch ~ dist, family=binomial(link="logit"))
```

Displaying this yields

R output
```
glm(formula = switch ~ dist, family=binomial(link="logit"))
            coef.est coef.se
(Intercept)   0.6060  0.0603
dist         -0.0062  0.0010
  n = 3020, k = 2
  residual deviance = 4076.2, null deviance = 4118.1 (difference = 41.9)
```

The coefficient for `dist` is -0.0062, which seems low, but this is misleading since distance is measured in meters, so this coefficient corresponds to the difference between, say, a house that is 90 meters away from the nearest safe well and a house that is 91 meters away.

Figure 5.8 shows the distribution of `dist` in the data. It seems more reasonable to rescale distance in 100-meter units:

R code
```
dist100 <- dist/100
```

and refitting the logistic regression yields

R output
```
glm(formula = switch ~ dist100, family=binomial(link="logit"))
            coef.est coef.se
(Intercept)     0.61    0.06
dist100        -0.62    0.10
  n = 3020, k = 2
  residual deviance = 4076.2, null deviance = 4118.1 (difference = 41.9)
```

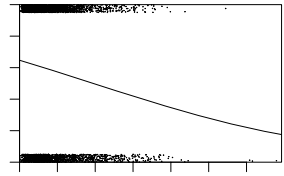

Figure 5.9 *Graphical expression of the fitted logistic regression,* Pr(*switching wells*) = $logit^{-1}(0.61 - 0.62 \cdot$ `dist100`$)$*, with (jittered) data overlain. The predictor* `dist100` *is* `dist`*/100: distance to the nearest safe well in 100-meter units.*

Graphing the fitted model

In preparing to plot the data, we first create a function to jitter the binary outcome while keeping the points between 0 and 1:

R code
```
jitter.binary <- function(a, jitt=.05){
  ifelse (a==0, runif (length(a), 0, jitt), runif (length(a), 1-jitt, 1))
}
```

We can then graph the data and fitted model:[3]

R code
```
switch.jitter <- jitter.binary (switch)
plot (dist, switch.jitter)
curve (invlogit (coef(fit.1)[1] + coef(fit.1)[2]*x), add=TRUE)
```

The result is displayed in Figure 5.9. The probability of switching is about 60% for people who live near a safe well, declining to about 20% for people who live more than 300 meters from any safe well. This makes sense: the probability of switching is higher for people who live closer to a safe well.

Interpreting the logistic regression coefficients

We can interpret the coefficient estimates using evaluations of the inverse-logit function and its derivative, as in the example of Section 5.1. Our model here is

$$\Pr(\text{switch}) = \text{logit}^{-1}(0.61 - 0.62 \cdot \text{dist100}).$$

1. The constant term can be interpreted when `dist100` $= 0$, in which case the probability of switching is $\text{logit}^{-1}(0.61) = 0.65$. Thus, the model estimates a 65% probability of switching if you live right next to an existing safe well.
2. We can evaluate the predictive difference with respect to `dist100` by computing the derivative at the average value of `dist100` in the dataset, which is 0.48 (that is, 48 meters; see Figure 5.8). The value of the linear predictor here is $0.61 - 0.62 \cdot 0.48 = 0.31$, and so the slope of the curve at this point is $-0.62e^{0.31}/(1+e^{0.31})^2 = -0.15$. Thus, adding 1 to `dist100`—that is, adding 100 meters to the distance to the nearest safe well—corresponds to a negative difference in the probability of switching of about 15%.

[3] Another display option, which would more clearly show the differences between households that did and did not switch, would be to overlay separate histograms of `dist` for the switchers and nonswitchers.

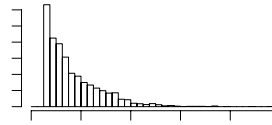

Figure 5.10 *Histogram of arsenic levels in unsafe wells (those exceeding 0.5) in the measured area of Araihazar, Bangladesh (see Figure 5.7).*

3. More quickly, the "divide by 4 rule" gives us $-0.62/4 = -0.15$. This comes out the same, to two decimal places, as was calculated using the derivative because the curve passes through the 50% point right in the middle of the data (see Figure 5.9).

In addition to interpreting its magnitude, we can look at the statistical significance of the coefficient for distance. The slope is estimated well, with a standard error of only 0.10, which is tiny compared to the coefficient estimate of -0.62. The approximate 95% interval is $[-0.82, -0.42]$, which is clearly statistically significantly different from zero.

Adding a second input variable

We now extend the well-switching example by adding the arsenic level of the existing well as a regression input. At the levels present in the Bangladesh drinking water, the health risks from arsenic are roughly proportional to exposure, and so we would expect switching to be more likely from wells with high arsenic levels. Figure 5.10 shows the arsenic levels of the unsafe wells before switching.

R code
```
fit.3 <- glm (switch ~ dist100 + arsenic, family=binomial(link="logit"))
```

which, when displayed, yields

R output
```
             coef.est coef.se
(Intercept)     0.00    0.08
dist100        -0.90    0.10
arsenic         0.46    0.04
  n = 3020, k = 3
  residual deviance = 3930.7, null deviance = 4118.1 (difference = 187.4)
```

Thus, comparing two wells with the same arsenic level, every 100 meters in distance to the nearest safe well corresponds to a *negative* difference of 0.90 in the logit probability of switching. Similarly, a difference of 1 in arsenic concentration corresponds to a 0.46 *positive* difference in the logit probability of switching. Both coefficients are statistically significant, each being more than 2 standard errors away from zero. And both their signs make sense: switching is easier if there is a nearby safe well, and if a household's existing well has a high arsenic level, there should be more motivation to switch.

For a quick interpretation, we divide the coefficients by 4: thus, 100 meters more in distance corresponds to an approximately 22% lower probability of switching, and 1 unit more in arsenic concentration corresponds to an approximately 11% positive difference in switching probability.

Comparing these two coefficients, it would at first seem that distance is a more important factor than arsenic level in determining the probability of switching.

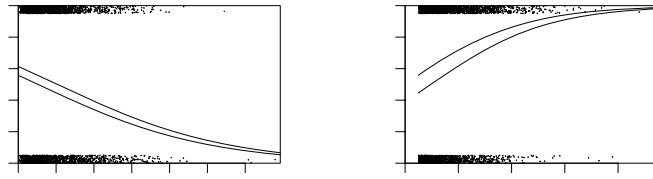

Figure 5.11 *Fitted logistic regression of probability of switching from an unsafe well as a function of two variables, plotted (a) as a function of distance to nearest safe well and (b) as a function of arsenic level of existing well. For each plot, the other input variable is held constant at different representative values.*

Such a statement is misleading, however, because in our data `dist100` shows less variation than `arsenic`: the standard deviation of distances to the nearest well is 0.38 (in units of 100 meters), whereas arsenic levels have a standard deviation of 1.10 on the scale used here. Thus, the logistic regression coefficients corresponding to 1-standard-deviation differences are $-0.90 \cdot 0.38 = -0.34$ for distance and $0.46 \cdot 1.10 = 0.51$ for arsenic level. Dividing by 4 yields the quick summary estimate of a 1-standard-deviation difference in distance or arsenic level corresponding to an 8% negative difference or a 13% positive difference, respectively, in Pr(switch).

Comparing the coefficient estimates when adding a predictor

The coefficient for `dist100` changes from -0.62 in the original model to 0.90 when arsenic level is added to the model. This change occurs because wells that are far from the nearest safe well are also likely to be particularly high in arsenic.

Graphing the fitted model with two predictors

The most natural way to graph the regression of y on two predictors might be as a three-dimensional surface, with the vertical axis showing $\Pr(y = 1)$ as a function of predictors plotted on the two horizontal axes.

However, we find such graphs hard to read, so instead we make separate plots as a function of each of the two variables; see Figure 5.11. As with the lines in Figure 3.4, we can plot the focus input variable on the x-axis and use multiple lines to show the fit for different values of the other input. To produce Figure 5.11a, we first plot the (jittered) data points, forcing zero to be included in the x-range of the plot because it is a natural baseline comparison for distance:

```
plot (dist, switch.jitter, xlim=c(0,max(dist)))
```
R code

We next add the fitted curves:

```
curve (invlogit (cbind (1, x/100,  .5) %*% coef(fit.3)), add=TRUE)
curve (invlogit (cbind (1, x/100, 1.0) %*% coef(fit.3)), add=TRUE)
```
R code

We need to divide x by 100 here because the plot is in the scale of meters but the model is defined in terms of `dist100` = `dist`/100.

The object created by `cbind(1,x/100,.5)` is an $n \times 3$ matrix constructed from a column of 1's, the vector `x` (used internally by the `curve` function), and a vector of `.5`'s. In constructing the matrix, R automatically expands the scalars `1` and `.5` to the length of the vector `x`. For the two lines, we pick arsenic levels of 0.5

and 1.0 because 0.5 is the minimum value of arsenic concentration (since we are only studying users of unsafe wells), and a difference of 0.5 represents a reasonable comparison, given the distribution of arsenic levels in the data (see Figure 5.10).

Similar commands are used to make Figure 5.11b, showing the probability of switching as a function of arsenic concentration with distance held constant:

R code
```
plot (arsenic, switch.jitter, xlim=c(0,max(arsenic)))
curve (invlogit (cbind (1, 0, x) %*% coef(fit.3)), add=TRUE)
curve (invlogit (cbind (1,.5, x) %*% coef(fit.3)), add=TRUE)
```

5.5 Logistic regression with interactions

We continue our modeling by adding the interaction between the two inputs:

R code
```
fit.4 <- glm (switch ~ dist100 + arsenic + dist100:arsenic,
  family=binomial(link="logit"))
display (fit.4)
```

which yields

R output
```
                coef.est coef.se
(Intercept)        -0.15    0.12
dist100            -0.58    0.21
arsenic             0.56    0.07
dist100:arsenic    -0.18    0.10
  n = 3020, k = 4
  residual deviance = 3927.6, null deviance = 4118.1 (difference = 190.5)
```

To understand the numbers in the table, we use the following tricks:

- Evaluating predictions and interactions at the mean of the data, which have average values of 0.48 for `dist100` and 1.66 for `arsenic` (that is, a mean distance of 48 meters to the nearest safe well, and a mean arsenic level of 1.66 among the unsafe wells).
- Dividing by 4 to get approximate predictive differences on the probability scale.

We now interpret each regression coefficient in turn.

- *Constant term:* $\text{logit}^{-1}(-0.15) = 0.47$ is the estimated probability of switching, if the distance to the nearest safe well is 0 and the arsenic level of the current well is 0. This is an impossible condition (since arsenic levels all exceed 0.5 in our set of unsafe wells), so we do not try to interpret the constant term. Instead, we can evaluate the prediction at the average values of `dist100` $= 0.48$ and `arsenic` $= 1.66$, where the probability of switching is $\text{logit}^{-1}(-0.15 - 0.58 \cdot 0.48 + 0.56 \cdot 1.66 - 0.18 \cdot 0.48 \cdot 1.66) = 0.59$.
- *Coefficient for distance*: this corresponds to comparing two wells that differ by 1 in `dist100`, if the arsenic level is 0 for both wells. Once again, we should not try to interpret this.

  Instead, we can look at the average value, `arsenic` $= 1.66$, where distance has a coefficient of $-0.58 - 0.18 \cdot 1.66 = -0.88$ on the logit scale. To quickly interpret this on the probability scale, we divide by 4: $-0.88/4 = -0.22$. Thus, at the mean level of arsenic in the data, each 100 meters of distance corresponds to an approximate 22% *negative* difference in probability of switching.
- *Coefficient for arsenic:* this corresponds to comparing two wells that differ by 1 in `arsenic`, if the distance to the nearest safe well is 0 for both.

  Instead, we evaluate the comparison at the average value for distance, `dist100` $=$

0.48, where arsenic has a coefficient of $0.56 - 0.18 \cdot 0.48 = 0.47$ on the logit scale. To quickly interpret this on the probability scale, we divide by 4: $0.47/4 = 0.12$. Thus, at the mean level of distance in the data, each additional unit of arsenic corresponds to an approximate 12% *positive* difference in probability of switching.

- *Coefficient for the interaction term:* this can be interpreted in two ways. Looking from one direction, for each additional unit of arsenic, the value -0.18 is added to the coefficient for distance. We have already seen that the coefficient for distance is -0.88 at the average level of arsenic, and so we can understand the interaction as saying that the importance of distance as a predictor increases for households with higher existing arsenic levels.

  Looking at it the other way, for each additional 100 meters of distance to the nearest well, the value -0.18 is added to the coefficient for arsenic. We have already seen that the coefficient for distance is 0.47 at the average distance to nearest safe well, and so we can understand the interaction as saying that the importance of arsenic as a predictor decreases for households that are farther from existing safe wells.

Centering the input variables

As discussed earlier in the context of linear regression, before fitting interactions it makes sense to center the input variables so that we can more easily interpret the coefficients. The centered inputs are:

R code
```
c.dist100 <- dist100 - mean(dist100)
c.arsenic <- arsenic - mean(arsenic)
```

We do not fully standardize these—that is, we do not scale by their standard deviations—because it is convenient to be able to consider known differences on the original scales of the data (100-meter distances and arsenic-concentration units).

Refitting the interaction model using the centered inputs

We can refit the model using the centered input variables, which will make the coefficients much easier to interpret:

R code
```
fit.5 <- glm (switch ~ c.dist100 + c.arsenic + c.dist100:c.arsenic,
  family=binomial(link="logit"))
```

We center the *inputs*, not the *predictors*. Hence, we do not center the interaction (`dist100*arsenic`); rather, we include the interaction of the two centered input variables. Displaying `fit.5` yields

R output
```
                    coef.est coef.se
(Intercept)          0.35     0.04
c.dist100           -0.88     0.10
c.arsenic            0.47     0.04
c.dist100:c.arsenic -0.18     0.10
  n = 3020, k = 4
  residual deviance = 3927.6, null deviance = 4118.1 (difference = 190.5)
```

Interpreting the inferences on this new scale:

- *Constant term:* $\text{logit}^{-1}(0.35) = 0.59$ is the estimated probability of switching, if `c.dist100` $=$ `c.arsenic` $= 0$, that is, if distance to nearest safe well and arsenic level are at their averages in the data. (We obtained this same calculation, but with more effort, with our earlier model with uncentered inputs.)

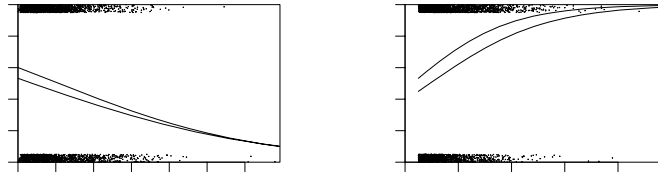

Figure 5.12 *Fitted logistic regression of probability of switching from an unsafe well as a function of distance to nearest safe well and arsenic level of existing well, for the model with interactions. Compare to the no-interaction model in Figure 5.11.*

- *Coefficient for distance*: this is the coefficient for distance (on the logit scale) if arsenic level is at its average value. To quickly interpret this on the probability scale, we divide by 4: $-0.88/4 = -0.22$. Thus, at the mean level of arsenic in the data, each 100 meters of distance corresponds to an approximate 22% *negative* difference in probability of switching.
- *Coefficient for arsenic:* this is the coefficient for arsenic level if distance to nearest safe well is at its average value. To quickly interpret this on the probability scale, we divide by 4: $0.47/4 = 0.12$. Thus, at the mean level of distance in the data, each additional unit of arsenic corresponds to an approximate 12% *positive* difference in probability of switching.
- *Coefficient for the interaction term:* this is unchanged by centering and has the same interpretation as before.

The predictions for new observations are unchanged. The linear centering of the predictors changes the interpretations of the coefficients but does not change the underlying model.

Statistical significance of the interaction

As can be seen from the regression table on the previous page, `c.dist100:c.arsenic` has an estimated coefficient of -0.18 with a standard error of 0.10. The estimate is not quite 2 standard errors away from zero and so is not quite statistically significant. However, the negative sign makes sense—it is plausible that arsenic level becomes a less important predictor for households that are farther from the nearest safe well, and the magnitude of the association is also plausible. So we keep the interaction in, following our general rules for regression coefficients and statistical significance, as given in Section 4.6.

Graphing the model with interactions

The clearest way to visualize the interaction model is to plot the regression curves as a function for each picture. The result is shown in Figure 5.12, the first graph of which we make in R as follows (with similar commands for the other graph):

R code

```
plot (dist, switch.jitter, xlim=c(0,max(dist)))
curve (invlogit (cbind(1,x/100, .5, .5*x/100) %*% coef(fit.4)), add=TRUE)
curve (invlogit (cbind(1,x/100,1.0,1.0*x/100) %*% coef(fit.4)), add=TRUE)
```

As Figure 5.12 makes clear, the interaction is not large in the range of most of the data. The largest pattern that shows up is in Figure 5.12a, where the two lines

intersect at around 300 meters. This graph shows evidence that the differences in switching associated with differences in arsenic level are large if you are close to a safe well, but with a diminishing effect if you are far from any safe well. This interaction makes some sense; however, there is some uncertainty in the size of the interaction (from the earlier regression table, an estimate of −0.18 with a standard error of 0.10), and as Figure 5.12a shows, there are only a few data points in the area where the interaction makes much of a difference.

The interaction also appears in Figure 5.12b, this time in a plot of probability of switching as a function of arsenic concentration, at two different levels of distance.

Adding social predictors

Are well users more likely to switch if they have community connections or more education? To see, we add two inputs:

- `assoc` = 1 if a household member is in any community organization
- `educ` = years of education of the well user.

We actually work with `educ4` = `educ`/4, for the usual reasons of making its regression coefficient more interpretable—it now represents the predictive difference of adding four years of education.[4]

R output

```
glm(formula = switch ~ c.dist100 + c.arsenic + c.dist100:c.arsenic +
    assoc + educ4, family=binomial(link="logit"))
                   coef.est coef.se
(Intercept)         0.20     0.07
c.dist100          -0.88     0.11
c.arsenic           0.48     0.04
c.dist100:c.arsenic -0.16     0.10
assoc              -0.12     0.08
educ4               0.17     0.04
  n = 3020, k = 6
  residual deviance = 3905.4, null deviance = 4118.1 (difference = 212.7)
```

For households with unsafe wells, belonging to a community association surprisingly is *not* predictive of switching, after controlling for the other factors in the model. However, persons with higher education are more likely to switch: the crude estimated difference is $0.17/4 = 0.04$, or a 4% positive difference in switching probability when comparing households that differ by 4 years of education.[5]

The coefficient for education makes sense and is statistically significant, so we keep it in the model. The coefficient for association does not make sense and is not statistically significant, so we remove it. (See Section 4.6 for a fuller discussion of including or excluding regression predictors.) We are left with

R output

```
glm(formula = switch ~ c.dist100 + c.arsenic + c.dist100:c.arsenic +
    educ4, family = binomial(link = "logit"))
                   coef.est coef.se
(Intercept)         0.15     0.06
```

[4] The levels of education among the 3000 respondents varied from 0 to 17 years, with nearly a third having zero. We repeated our analysis with a discrete recoding of the education variable (0 = 0 years, 1 = 1–8 years, 2 = 9–12 years, 3 = 12+ years), and our results were essentially unchanged.

[5] Throughout this example, we have referred to "coefficients" and "differences," rather than to "effects" and "changes," because the observational nature of the data makes it difficult to directly interpret the regression model causally. We continue causal inference more carefully in Chapter 9, briefly discussing the arsenic problem at the end of Section 9.8.

```
c.dist100           -0.87     0.11
c.arsenic            0.48     0.04
c.dist100:c.arsenic -0.16     0.10
educ4                0.17     0.04
  n = 3020, k = 5
  residual deviance = 3907.9, null deviance = 4118.1 (difference = 210.2)
```

Adding further interactions

When inputs have large main effects, it is our general practice to include their interactions as well. We first create a centered education variable:

R code

```
c.educ4 <- educ4 - mean(educ4)
```

and then fit a new model interacting it with distance to nearest safe well and arsenic level of the existing well:

R output

```
glm(formula=switch~c.dist100 + c.arsenic + c.educ4 + c.dist100:c.arsenic +
  c.dist100:c.educ4 + c.arsenic:c.educ4, family=binomial(link="logit"))
                    coef.est coef.se
(Intercept)          0.36     0.04
c.dist100           -0.90     0.11
c.arsenic            0.49     0.04
c.educ4              0.18     0.04
c.dist100:c.arsenic -0.12     0.10
c.dist100:c.educ4    0.32     0.11
c.arsenic:c.educ4    0.07     0.04
  n = 3020, k = 7
  residual deviance = 3891.7, null deviance = 4118.1 (difference = 226.4)
```

We can interpret these new interactions by understanding how education modifies the predictive difference corresponding to distance and arsenic.

- *Interaction of distance and education:* a difference of 4 years of education corresponds to a difference of 0.32 in the coefficient for `dist100`. As we have already seen, `dist100` has a negative coefficient on average; thus positive changes in education *reduce* distance's negative association. This makes sense: people with more education probably have other resources so that walking an extra distance to get water is not such a burden.
- *Interaction of arsenic and education:* a difference of 4 years of education corresponds to a difference of 0.07 in the coefficient for `arsenic`. As we have already seen, `arsenic` has a positive coefficient on average; thus increasing education *increases* arsenic's positive association. This makes sense: people with more education could be more informed about the risks of arsenic and thus more sensitive to increasing arsenic levels (or, conversely, less in a hurry to switch from wells with arsenic levels that are relatively low).

As before, centering allows us to interpret the main effects as coefficients when other inputs are held at their average values in the data.

Standardizing predictors

We should think seriously about standardizing all predictors as a default option when fitting models with interactions. The struggles with `dist100` and `educ4` in this example suggest that standardization—by subtracting the mean from each of the continuous input variables and dividing by 2 standard deviations, as suggested near the end of Section 4.2—might be the simplest approach.

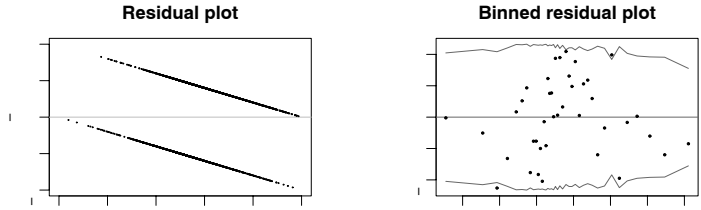

Figure 5.13 *(a) Residual plot and (b) binned residual plot for the well-switching model shown on page 96. The strong patterns in the raw residual plot arise from the discreteness of the data and inspire us to use the binned residual plot instead. The bins are not equally spaced; rather, each bin has an equal number of data points. The light lines in the binned residual plot indicate theoretical 95% error bounds.*

5.6 Evaluating, checking, and comparing fitted logistic regressions

Residuals and binned residuals

We can define residuals for logistic regression, as with linear regression, as observed minus expected values:

$$\text{residual}_i = y_i - \text{E}(y_i|X_i) = y_i - \text{logit}^{-1}(X_i\beta).$$

The data y_i are discrete and so are the residuals. For example, if $\text{logit}^{-1}(X_i\beta) = 0.7$, then $\text{residual}_i = -0.7$ or $+0.3$, depending on whether $y_i = 0$ or 1. As a result, plots of raw residuals from logistic regression are generally not useful. For example, Figure 5.13a plots residuals versus fitted values for the well-switching regression.

Instead, we plot *binned residuals* by dividing the data into categories (bins) based on their fitted values, and then plotting the average residual versus the average fitted value for each bin. The result appears in Figure 5.13b; here we divided the data into 40 bins of equal size.[6] The dotted lines (computed as $2\sqrt{p(1-p)/n}$, where n is the number of points per bin, $3020/40 = 75$ in this case) indicate ± 2 standard-error bounds, within which one would expect about 95% of the binned residuals to fall, if the model were actually true. One of the 40 binned residuals in Figure 5.13b falls outside the bounds, which is not a surprise, and no dramatic pattern appears.

Plotting binned residuals versus inputs of interest

We can also look at residuals in a more structured way by binning and plotting them with respect to individual input variables or combinations of inputs. For example, in the well-switching example, Figure 5.14a displays the average residual in each bin as defined by distance to the nearest safe well, and Figure 5.14b shows average residuals, binned by arsenic levels.

This latter plot shows a disturbing pattern, with an extreme negative residual in the first three bins: people with wells in the lowest bin (which turns out to correspond to arsenic levels between 0.51 and 0.53) are about 20% less likely to

[6] There is typically some arbitrariness in choosing the number of bins: we want each bin to contain enough points so that the averaged residuals are not too noisy, but it helps to have many bins so as to see more local patterns in the residuals. For this example, 40 bins seemed to give sufficient resolution, while still having enough points per bin. Another approach would be to apply a nonparametric smoothing procedure such as lowess (Cleveland, 1979) to the residuals.

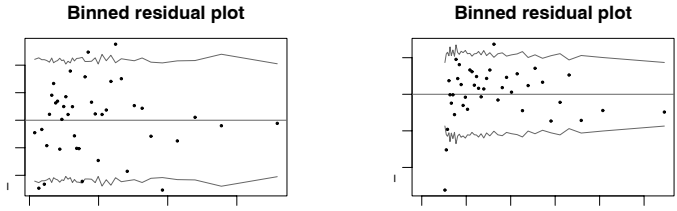

Figure 5.14 *Plots of residuals for the well-switching model, binned and plotted versus (a) distance to nearest well and (b) arsenic level. The dotted lines in the binned residual plot indicate theoretical 95% error bounds that would be appropriate if the model were true. The second plot shows a problem with the model in the lowest bins of arsenic levels.*

switch than is predicted by the model: the average predicted probability of switching for these users is 49%, but actually only 32% of them switched. There is also a slight pattern in the residuals as a whole, with positive residuals (on average) in the middle of the range of arsenic and negative residuals at the high end.

Considering a log transformation

To experienced regression modelers, a rising and then falling pattern of residuals such as in Figure 5.14b is a signal to consider taking the logarithm of the predictor on the x axis—in this case, arsenic level. Another option would be to add a quadratic term to the regression; however, since arsenic is an all-positive variable, it makes sense to consider its logarithm. We do not, however, model distance on the log scale, since the residual plot, as shown in Figure 5.13a, indicates a good fit of the linear model.

We define

R code
```
log.arsenic <- log(arsenic)
c.log.arsenic <- log.arsenic - mean (log.arsenic)
```

and then fit the same model as before, using `log.arsenic` in place of `arsenic`:

R output
```
glm(formula = switch ~ c.dist100 + c.log.arsenic + c.educ4 +
    c.dist100:c.log.arsenic + c.dist100:c.educ4 + c.log.arsenic:c.educ4,
    family = binomial(link = "logit"))
                        coef.est coef.se
(Intercept)              0.35     0.04
c.dist100               -0.98     0.11
c.log.arsenic            0.90     0.07
c.educ4                  0.18     0.04
c.dist100:c.log.arsenic -0.16     0.19
c.dist100:c.educ4        0.34     0.11
c.log.arsenic:c.educ4    0.06     0.07
  n = 3020, k = 7
  residual deviance = 3863.1, null deviance = 4118.1 (difference = 255)
```

This is qualitatively similar to the model on the original scale: the interactions have the same sign as before, and the signs of the main effects are also unchanged.

Figure 5.15a shows the predicted probability of switching as a function of arsenic level. Compared to the model in which arsenic was included as a linear predictor

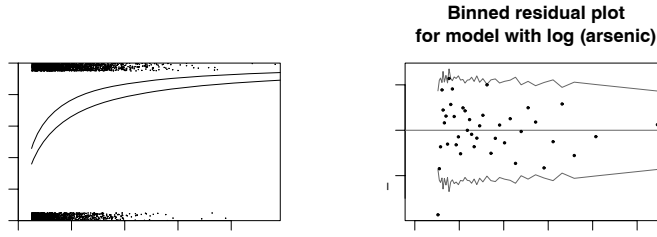

Figure 5.15 *(a) Probability of switching as a function of arsenic level (at two different values of* `dist` *and with education held constant at its average value), for the model that includes arsenic on the logarithmic scale. Compared to Figure 5.11b (the corresponding plot with arsenic level included as a linear predictor), the model looks similar, but with a steeper slope at the low end of the curve and a more gradual slope at the high end. (b) Average residuals for this model, binned by arsenic level. Compared to Figure 5.14b, the residual plot still shows problems at the lowest arsenic levels but otherwise looks cleaner.*

(see Figure 5.11b on page 91), the curves are compressed at the left and stretched out at the right.

Figure 5.15b displays the residuals for the log model, again binned by arsenic level. Compared to the earlier model, the residuals look better but there is still a problem at the very low end. Users of wells with arsenic levels just above 0.50 are less likely to switch than predicted by the model. At this point, we do not know if this can be explained psychologically (measurements just over the threshold do not seem so bad), through measurement error (perhaps some of the wells we have recorded as 0.51 or 0.52 were measured before or after and found to have arsenic levels below 0.5), or for some other reason.

Error rate and comparison to the null model

The *error rate* is defined as the proportion of cases for which the deterministic prediction—guessing $y_i = 1$ if $\text{logit}^{-1}(X_i\beta) > 0.5$ and guessing $y_i = 0$ if $\text{logit}^{-1}(X_i\beta) < 0.5$—is wrong. In R, we could write:

```
error.rate <- mean ((predicted>0.5 & y==0) | (predicted<.5 & y==1))
```
R code

The error rate should always be less than 1/2 (otherwise we could simply set all the β's to 0 and get a better-fitting model), but in many cases we would expect it to be much lower. We can compare it to the error rate of the *null model*, which is simply to assign the same probability to each y_i. This is simply logistic regression with only a constant term, and the estimated probability will simply be the proportion of 1's in the data, or $p = \sum_{i=1}^{n} y_i/n$ (recalling that each $y_i = 0$ or 1). The error rate of the null model is then p or $1-p$, whichever is lower.

For example, in the well-switching example, the null model has an error rate of 42% (58% of the respondents are switchers and 42% are not, thus the model with no predictors gives each person a 58% chance of switching, which corresponds to a point prediction of switching for each person, and that guess will be wrong 42% of the time). Our final logistic regression model (as calculated in R as shown) has an error rate of 36%. The model correctly predicts the behavior of 64% of the respondents.

The error rate is not a perfect summary of model misfit, because it does not distinguish between predictions of 0.6 and 0.9, for example. But, as with R^2 for the linear model, it is easy to interpret and is often helpful in understanding the model fit. An error rate equal to the null rate is terrible, and the best possible error rate is zero. Thus, the well-switching model is not particularly impressive with an error rate of 38%, a mere 4% better than simply guessing that all people will switch.

This low error rate does not mean the model is useless—as the plots showed, the fitted model is highly predictive of the probability of switching. But most of the data are close to the mean level of the inputs (distances of less than 100 meters to the nearest safe well, and arsenic levels between 0.5 and 1.0), and so for most of the data, the simple mean prediction, Pr(switch)=0.58, works well. The model is informative near the extremes, but relatively few data points are out there and so the overall predictive accuracy of the model is not high.

Deviance

For logistic regressions and other discrete-data models, it does not quite make sense to calculate residual standard deviation and R^2, for pretty much the same reason that the models are not simply fit by least squares—the squared error is not the mathematically optimal measure of model error. Instead, it is standard to use *deviance*, a statistical summary of model fit, defined for logistic regression and other generalized linear models to be an analogy to residual standard deviation.

For now, you should know the following properties of deviance:

- Deviance is a measure of error; lower deviance means better fit to data.
- If a predictor that is simply random noise is added to a model, we expect deviance to decrease by 1, on average.
- When an informative predictor is added to a model, we expect deviance to decrease by more than 1. When k predictors are added to a model, we expect deviance to decrease by more than k.

For classical (non-multilevel) models, the deviance is defined as -2 times the logarithm of the likelihood function (up to an arbitrary additive constant, since we are always comparing deviances, never evaluating them on their own).

For example, in the first model fit to the well-switching example, the display on page 88 reports that the "null deviance" is 4118.1 and the "residual deviance" is 4076.2. The null deviance corresponds to the null model, with just the constant term. Thus, by adding `dist` as a predictor in the model, the deviance has decreased by 41.9. This is much more than the expected decrease of 1 if the predictor were noise, so it has clearly improved the fit.

The next fitted model uses `dist100` = `dist/100` as a predictor instead. The deviance stays at 4076.2, because linear transformations have no effect on predictions in classical regression models. (We shall see, however, that linear transformations can make a difference in multilevel models.)

We then add the `arsenic` predictor, and the deviance decreases to 3930.7, a drop of 145.5—once again, much more than the expected decrease of 1 if the new predictor were noise, so it has clearly improved the fit.

The following model including the interaction between `dist` and `arsenic` has a residual deviance of 3927.6, a decrease of 3.1 from the previous model, only a bit more than the expected decrease of 1 if the new predictor were noise. This decrease in deviance is not statistically significant (we can see this because the coefficient for the added predictor is less than 2 standard errors from zero) but, as discussed

in Section 5.5, we keep the interaction in the model because it makes sense in the applied context.

Adding the social predictors `assoc` and `educ` to the regression decreases the deviance to 3905.4, implying better prediction than all the previous models. Removing `assoc` increases the deviance only a small amount, to 3907.9. Adding interactions of education with distance and arsenic level reduces the deviance by quite a bit more, to 3891.7.

Transforming arsenic on to the log scale—that is, removing `arsenic` from the model and replacing it with `log.arsenic`, takes the deviance down to 3863.1, another large improvement.

For multilevel models, deviance is generalized to the deviance information criterion (DIC), as described in Section 24.3.

5.7 Average predictive comparisons on the probability scale

As illustrated, for example, by Figure 5.11 on page 91, logistic regressions are nonlinear on the probability scale—that is, a specified difference in one of the x variables does *not* correspond to a constant difference in $\Pr(y = 1)$. As a result, logistic regression coefficients cannot directly be interpreted on the scale of the data. Logistic regressions are inherently more difficult than linear regressions to interpret.

Graphs such as Figure 5.11 are useful, but for models with many predictors, or where graphing is inconvenient, it is helpful to have a summary, comparable to the linear regression coefficient, which gives the expected, or average, difference in $\Pr(y=1)$ corresponding to a unit difference in each of the input variables.

Example: well switching in Bangladesh

For a model with nonlinearity or interactions, or both, this *average predictive comparison* depends on the values of the input variables, as we shall illustrate with the well-switching example. To keep the presentation clean at this point, we shall work with a simple no-interaction model,

R code
```
fit.10 <- glm (switch ~ dist100 + arsenic + educ4,
  family=binomial(link="logit"))
```

which yields

R output
```
            coef.est coef.se
(Intercept)    -0.21    0.09
dist100        -0.90    0.10
arsenic         0.47    0.04
educ4           0.17    0.04
  n = 3020, k = 4
  residual deviance = 3910.4, null deviance = 4118.1 (difference = 207.7)
```

giving the probability of switching as a function of distance to the nearest well (in 100-meter units), arsenic level, and education (in 4-year units).

Average predictive difference in probability of switching, comparing households that are next to, or 100 meters from, the nearest safe well. Let us compare two households—one with `dist100` = 0 and one with `dist100` = 1—but identical in the other input variables, `arsenic` and `educ4`. The *predictive difference* in probability

of switching between these two households is

$$\delta(\text{arsenic}, \text{educ4}) = \text{logit}^{-1}(-0.21 - 0.90 \cdot 1 + 0.47 \cdot \text{arsenic} + 0.17 \cdot \text{educ4}) - \text{logit}^{-1}(-0.21 - 0.90 \cdot 0 + 0.47 \cdot \text{arsenic} + 0.17 \cdot \text{educ4}). \quad (5.6)$$

We write δ as a function of `arsenic` and `educ4` to emphasize that it depends on the levels of these other variables.

We average the predictive differences over the n households in the data to obtain:

$$\text{average predictive difference: } = \frac{1}{n}\sum_{i=1}^{n} \delta(\text{arsenic}_i, \text{educ4}_i). \quad (5.7)$$

In R:

R code

```
b <- coef (fit.10)
hi <- 1
lo <- 0
delta <- invlogit (b[1] + b[2]*hi + b[3]*arsenic + b[4]*educ4) -
         invlogit (b[1] + b[2]*lo + b[3]*arsenic + b[4]*educ4)
print (mean(delta))
```

The result is -0.20, implying that, on average in the data, households that are 100 meters from the nearest safe well are 20% less likely to switch, compared to houseohulds that are right next to the nearest safe well, at the same arsenic and education levels.

Average predictive difference in probability of switching, comparing households with existing arsenic levels of 0.5 and 1.0. We can similarly compute the predictive difference, and average predictive difference, comparing households at two different arsenic levels, assuming equality in distance to nearest safe well and education levels. We choose `arsenic` $= 0.5$ and 1.0 as comparison points because 0.5 is the lowest unsafe level, 1.0 is twice that, and this comparison captures much of the range of the data (see Figure 5.10 on page 90). Here is the computation:

R code

```
hi <- 1.0
lo <- 0.5
delta <- invlogit (b[1] + b[2]*dist100 + b[3]*hi + b[4]*educ4) -
         invlogit (b[1] + b[2]*dist100 + b[3]*lo + b[4]*educ4)
print (mean(delta))
```

The result is 0.06—so this comparison corresponds to a 6% difference in probability of switching.

Average predictive difference in probability of switching, comparing householders with 0 and 12 years of education. Similarly, we can compute an average predictive difference of the probability of switching for householders with 0 compared to 12 years of education (that is, comparing `educ4` $= 0$ to `educ4` $= 3$):

R code

```
hi <- 3
lo <- 0
delta <- invlogit (b[1]+b[2]*dist100+b[3]*arsenic+b[4]*hi) -
         invlogit (b[1]+b[2]*dist100+b[3]*arsenic+b[4]*lo)
print (mean(delta))
```

which comes to 0.12.

Average predictive comparisons in the presence of interactions

We can perform similar calculations for models with interactions. For example, consider the average predictive difference, comparing `dist` = 0 to `dist` = 100, for the model that includes a distance × arsenic interaction:

R code
```
fit.11 <- glm (switch ~ dist100 + arsenic + educ4 + dist100:arsenic,
  family=binomial(link="logit"))
```

which, when displayed, yields

R output
```
                coef.est coef.se
(Intercept)        -0.35    0.13
dist100            -0.60    0.21
arsenic             0.56    0.07
educ4               0.17    0.04
dist100:arsenic    -0.16    0.10
  n = 3020, k = 5
  residual deviance = 3907.9, null deviance = 4118.1 (difference = 210.2)
```

Here is the R code for computing the average predictive difference comparing `dist1` = 1 to `dist1` = 0:

R code
```
b <- coef (fit.11)
hi <- 1
lo <- 0
delta <- invlogit (b[1] + b[2]*hi + b[3]*arsenic + b[4]*educ4 +
                   b[5]*hi*arsenic) -
         invlogit (b[1] + b[2]*lo + b[3]*arsenic + b[4]*educ4 +
                   b[5]*lo*arsenic)
print (mean(delta))
```

which comes to −0.19.

General notation for predictive comparisons

Considering each input one at a time, we use the notation u for the *input of interest* and v for the vector of all other inputs. Suppose we are considering comparisons of $u = u^{(1)}$ to $u = u^{(0)}$ with all other inputs held constant (for example, we have considered the comparison of households that are 0 meters or 100 meters from the nearest safe well). The *predictive difference* in probabilities between two cases, differing only in u, is

$$\delta(u^{(\text{hi})}, u^{(\text{lo})}, v, \beta) = \Pr(y{=}1|u^{(\text{hi})}, v, \beta) - \Pr(y{=}1|u^{(\text{lo})}, v, \beta), \tag{5.8}$$

where the vertical bar in these expressions is read "conditional on" (for example, the probability that $y = 1$ given $u^{(\text{hi})}$, v, and β).

The average predictive difference then averages over the n points in the dataset used to fit the logistic regression:

$$\Delta(u^{(\text{hi})}, u^{(\text{lo})}) = \frac{1}{n}\sum_{i=1}^{n} \delta(u^{(\text{hi})}, u^{(\text{lo})}, v_i, \beta), \tag{5.9}$$

where v_i represents the vector of other inputs (in our example, arsenic and education levels) for data point i. These expressions generalize formulas (5.6) and (5.7).

For models with interactions, the predictive difference formula (5.8) must be computed carefully, with awareness of where each input enters into the regression model. The distinction between input variables (in this case, distance, arsenic, and

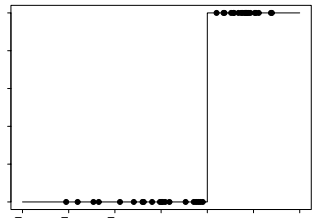

Figure 5.16 *Example of data for which a logistic regression model is nonidentifiable. The outcome y equals 0 for all data below $x = 2$ and 1 for all data above $x = 2$, hence the best-fit logistic regression line is $y = \text{logit}^{-1}(\infty(x - 2))$, which has an infinite slope at $x = 2$.*

education) and predictors (constant term, distance, arsenic, education, and distance $\times$ arsenic) is crucial. We discuss average predictive comparisons further in Section 21.4.

5.8 Identifiability and separation

There are two reasons that a logistic regression can be nonidentified (that is, have parameters that cannot be estimated from the available data and model, as discussed in Section 4.5 in the context of linear regression):

1. As with linear regression, if predictors are collinear, then estimation of the linear predictor, $X\beta$, does not allow separate estimation of the individual parameters β. We can handle this kind of nonidentifiability in the same way that we would proceed for linear regression, as described in Section 4.5.

2. A completely separate identifiability problem, called *separation*, can arise from the discreteness of the data.

   - If a predictor x_j is completely aligned with the outcome, so that $y = 1$ for all the cases where x_j exceeds some threshold T, and $y = 0$ for all cases where $x_j < T$, then the best estimate for the coefficient β_j is ∞. Figure 5.16 shows an example. Exercise 5.11 gives an example with a binary predictor.

   - Conversely, if $y = 1$ for all cases where $x_j < T$, and $y = 0$ for all cases where $x_j > T$, then $\hat{\beta}_j$ will be $-\infty$.

   - More generally, this problem will occur if any linear combination of predictors is perfectly aligned with the outcome. For example, suppose that $7x_1 + x_2 - 3x_3$ is completely positively aligned with the data, with $y = 1$ if and only if this linear combination of predictors exceeds some threshold. Then the linear combination $7\hat{\beta}_1 + \hat{\beta}_2 - 3\hat{\beta}_3$ will be estimated at ∞, which will cause at least one of the three coefficients $\beta_1, \beta_2, \beta_3$ to be estimated at ∞ or $-\infty$.

   One way to handle separation is using a Bayesian or penalized-likelihood approach (implemented for R in the `brlr` package) that provides a small amount of information on all the regression coefficients, including those that are not identified from the data alone. (See Chapter 18 for more on Bayesian inference.)

5.9 Bibliographic note

According to Cramer (2003, chapter 9), logistic regression was introduced for binary data in the mid-twentieth century and has become increasingly popular as computational improvements have allowed it to become a routine data-analytic tool.

For more on income and voting in presidential elections, see Gelman, Shor, et al. (2005). The example of drinking water in Bangladesh is described further by van Geen et al. (2003) and Gelman, Trevisani, et al. (2004).

Binned residual plots and related tools for checking the fit of logistic regressions are discussed by Landwehr, Pregibon, and Shoemaker (1984), Gelman, Goegebeur, et al. (2000), Pardoe and Cook (2002), and Pardoe (2004).

Deviance is discussed by McCullagh and Nelder (1989); related ideas include the Akaike (1973) information criterion (AIC), C_p (Mallows, 1973), and the deviance information criterion (DIC; Spiegelhalter et al., 2002). See also Fox (2002) for an applied overview and Gelman et al. (2003, sections 6.7–6.8) for a Bayesian perspective.

Nonidentifiability of logistic regression and separation in discrete data are discussed by Albert and Anderson (1984), Lesaffre and Albert (1989), Heinze and Schemper (2003), as well as in the book by Agresti (2002). Zorn (2005) proposes a Bayesian resolution, following Firth (1993).

5.10 Exercises

1. The folder `nes` contains the survey data of presidential preference and income for the 1992 election analyzed in Section 5.1, along with other variables including sex, ethnicity, education, party identification, and political ideology.
   (a) Fit a logistic regression predicting support for Bush given all these inputs. Consider how to include these as regression predictors and also consider possible interactions.
   (b) Evaluate and compare the different models you have fit. Consider coefficient estimates and standard errors, residual plots, and deviances.
   (c) For your chosen model, discuss and compare the importance of each input variable in the prediction.

2. Without using a computer, sketch the following logistic regression lines:
   (a) $\Pr(y = 1) = \text{logit}^{-1}(x)$
   (b) $\Pr(y = 1) = \text{logit}^{-1}(2 + x)$
   (c) $\Pr(y = 1) = \text{logit}^{-1}(2x)$
   (d) $\Pr(y = 1) = \text{logit}^{-1}(2 + 2x)$
   (e) $\Pr(y = 1) = \text{logit}^{-1}(-2x)$

3. You are interested in how well the combined earnings of the parents in a child's family predicts high school graduation. You are told that the probability a child graduates from high school is 27% for children whose parents earn no income and is 88% for children whose parents earn $60,000. Determine the logistic regression model that is consistent with this information. (For simplicity you may want to assume that income is measured in units of $10,000).

4. Perform a logistic regression for a problem of interest to you. This can be from a research project, a previous class, or data you download. Choose one variable

of interest to be the outcome, which will take on the values 0 and 1 (since you are doing logistic regression).

(a) Analyze the data in R. Use the `display()` function to summarize the results.

(b) Fit several different versions of your model. Try including different predictors, interactions, and transformations of the inputs.

(c) Choose one particular formulation of the model and do the following:

i. Describe how each input affects $\Pr(y = 1)$ in the fitted model. You must consider the estimated coefficient, the range of the input values, and the nonlinear inverse-logit function.

ii. What is the error rate of the fitted model? What is the error rate of the null model?

iii. Look at the deviance of the fitted and null models. Does the improvement in fit seem to be real?

iv. Use the model to make predictions for some test cases of interest.

5. In a class of 50 students, a logistic regression is performed of course grade (pass or fail) on midterm exam score (continuous values with mean 60 and standard deviation 15). The fitted model is $\Pr(\text{pass}) = \text{logit}^{-1}(-24 + 0.4x)$.

(a) Graph the fitted model. Also on this graph put a scatterplot of hypothetical data consistent with the information given.

(b) Suppose the midterm scores were transformed to have a mean of 0 and standard deviation of 1. What would be the equation of the logistic regression using these transformed scores as a predictor?

(c) Create a new predictor that is pure noise (for example, in R you can create `newpred <- rnorm (n,0,1)`). Add it to your model. How much does the deviance decrease?

6. Latent-data formulation of the logistic model: take the model $\Pr(y = 1) = \text{logit}^{-1}(1 + 2x_1 + 3x_2)$ and consider a person for whom $x_1 = 1$ and $x_2 = 0.5$. Sketch the distribution of the latent data for this person. Figure out the probability that $y=1$ for the person and shade the corresponding area on your graph.

7. Limitations of logistic regression: consider a dataset with $n = 20$ points, a single predictor x that takes on the values $1, \ldots, 20$, and binary data y. Construct data values $y_1, \ldots, y_{20}$ that are inconsistent with any logistic regression on x. Fit a logistic regression to these data, plot the data and fitted curve, and explain why you can say that the model does not fit the data.

8. Building a logistic regression model: the folder `rodents` contains data on rodents in a sample of New York City apartments.

(a) Build a logistic regression model to predict the presence of rodents (the variable `rodent2` in the dataset) given indicators for the ethnic groups (`race`). Combine categories as appropriate. Discuss the estimated coefficients in the model.

(b) Add to your model some other potentially relevant predictors describing the apartment, building, and community district. Build your model using the general principles explained in Section 4.6. Discuss the coefficients for the ethnicity indicators in your model.

9. Graphing logistic regressions: the well-switching data described in Section 5.4 are in the folder `arsenic`.

(a) Fit a logistic regression for the probability of switching using log (distance to nearest safe well) as a predictor.

(b) Make a graph similar to Figure 5.9 displaying Pr(switch) as a function of distance to nearest safe well, along with the data.

(c) Make a residual plot and binned residual plot as in Figure 5.13.

(d) Compute the error rate of the fitted model and compare to the error rate of the null model.

(e) Create indicator variables corresponding to `dist` < 100, $100 \leq$ `dist` < 200, and `dist` > 200. Fit a logistic regression for Pr(switch) using these indicators. With this new model, repeat the computations and graphs for part (a) of this exercise.

10. Model building and comparison: continue with the well-switching data described in the previous exercise.

(a) Fit a logistic regression for the probability of switching using, as predictors, distance, log(arsenic), and their interaction. Interpret the estimated coefficients and their standard errors.

(b) Make graphs as in Figure 5.12 to show the relation between probability of switching, distance, and arsenic level.

(c) Following the procedure described in Section 5.7, compute the average predictive differences corresponding to:

i. A comparison of `dist` $= 0$ to `dist` $= 100$, with `arsenic` held constant.
ii. A comparison of `dist` $= 100$ to `dist` $= 200$, with `arsenic` held constant.
iii. A comparison of `arsenic` $= 0.5$ to `arsenic` $= 1.0$, with `dist` held constant.
iv. A comparison of `arsenic` $= 1.0$ to `arsenic` $= 2.0$, with `dist` held constant.

Discuss these results.

11. Identifiability: the folder `nes` has data from the National Election Studies that were used in Section 5.1 to model vote preferences given income. When we try to fit a similar model using ethnicity as a predictor, we run into a problem. Here are fits from 1960, 1964, 1968, and 1972:

R output

```
glm(formula = vote ~ female + black + income,
  family=binomial(link="logit"), subset=(year==1960))
            coef.est coef.se
(Intercept) -0.14     0.23
female       0.24     0.14
black       -1.03     0.36
income       0.03     0.06

glm(formula = vote ~ female + black + income,
  family=binomial(link="logit"), subset=(year==1964))
            coef.est coef.se
(Intercept)  -1.15     0.22
female       -0.09     0.14
black       -16.83   420.40
income        0.19     0.06

glm(formula = vote ~ female + black + income,
  family=binomial(link="logit"), subset=(year==1968))
            coef.est coef.se
(Intercept)  0.47     0.24
```

```
female      -0.01     0.15
black       -3.64     0.59
income      -0.03     0.07

glm(formula = vote ~ female + black + income,
  family=binomial(link="logit"), subset=(year==1972))
            coef.est coef.se
(Intercept)  0.67     0.18
female      -0.25     0.12
black       -2.63     0.27
income       0.09     0.05
```

What happened with the coefficient of `black` in 1964? Take a look at the data and figure out where this extreme estimate came from. What can be done to fit the model in 1964?