

# Homework Assignment 8: Fair's Affairs

36-402, Advanced Data Analysis, Spring 2011

## SOLUTIONS

```
library(AER)
data(Affairs)
```

### 1. ANSWER:

- (a) When dealing with an counting variable  $Y$  with a known (not estimated) upper limit  $m$ , we can try to model it as having a binomial distribution, with  $m$  trials and some success probability  $p$  which we need to estimate. In binomial logistic regression, the log odds of success is still linear in the predictors, but we use this binomial distribution to build the likelihood, instead of the Bernoulli distribution we used for binary logistic regression. (Actually, binary is just the special case where  $m = 1$ .) In this case, we can roughly think of this as having started with 12 monthly, binary observations for each subject (“did they have an affair in January?”, “did they have an affair in February?”, etc.), and collapsing it by summing the observations. The fitted value from the logistic regression is then a prediction of the probability of having an affair in any given month.

In R, we need to provide the `glm()` function with two columns of responses bound together: the first counts successes and the second one failures. (There are multiple examples of this in the assigned reading from Faraway.)

```
log.r.num = glm(cbind(affairs,12-affairs)~., data=Affairs,family=binomial)
summary(log.r.num)
```

Notice that setting the right-hand side of the formula, after the `~`, to be just `.` means “all the other variables in the data frame”.

(age) is significant with a negative coefficient, suggesting that older people are less likely to have affairs; specifically, every year of age lowers the predicted log odds of an affair in any given month by 0.043. (yearsmarried) is positively associated with affairs; each year married increases the monthly log odds by 0.15. (religiousness) has the second-largest-in-absolute-value coefficient; every step on the 5-point scale lowers the log-odds by 0.44. (occupation) has a small but significant positive effect on affairs. (rating) has the largest-in-absolute-value coefficient, which is negative, i.e., people who give

high ratings to their marriage are predicted to be unlikely to have affairs.

- (b) Here we are just predicting the whether they had any affairs at all, and ignoring how many. We need to make sure now that the response variable *is* binary; here's one way to do it:

```
log.r = glm((affairs>0) ~ ., data = Affairs, family=binomial)
summary(log.r)
```

This logistic regression gives mostly the same conclusions as the last one, but with less statistical significance. (rating) and (religiousness) are still the two most important predictors, while (age) and (yearsmarried) are hardly significant. Additionally, (occupation) is not significant.

2. ANSWER: There is some variation between models both in coefficient sign and in significance levels. For example, (childrenyes) and **education** had negative coefficients in the first model and positive coefficients in the second one, though none of these four coefficients was significant. Perhaps the largest single difference is in the evaluation of whether (occupation) is significant; the first model rates it highly significant while the second model finds it not significant at all.

Differences between these models arise because they are being asked to predict different, though related, things. Variables which are unimportant for *whether* someone has an affair could be important in determining how many affairs they have, if they have any. It is common to be more confident about the importance variables which show up as significant for multiple related models (here, say, the rating of the marriage), as these variables are “robust with respect to model specification”. This seems reasonable, but it is hard to make the argument precise.

3. ANSWER:

- (a) For each individual, the first model estimates the probability of having an affair in a given month. Then the probability of having no affairs at all over 12 months is the probability of no affair in one month raised to the 12th power:

```
p.none.log.r.num = (1-fitted(log.r.num))^12
```

The second logistic regression fits directly the probability of having an affair over the whole, so the probability of having no affair is estimated as 1 minus the probability of an affair:

```
p.none.log.r = 1-fitted(log.r)
```

- (b) 

```
plot(x=p.none.log.r.num, y=p.none.log.r,
     main = "Comparing estimated probabilities of no affairs",
     xlab = "First model; response as number of affairs",
     ylab = "Second model; response as affair/no-affair",
```

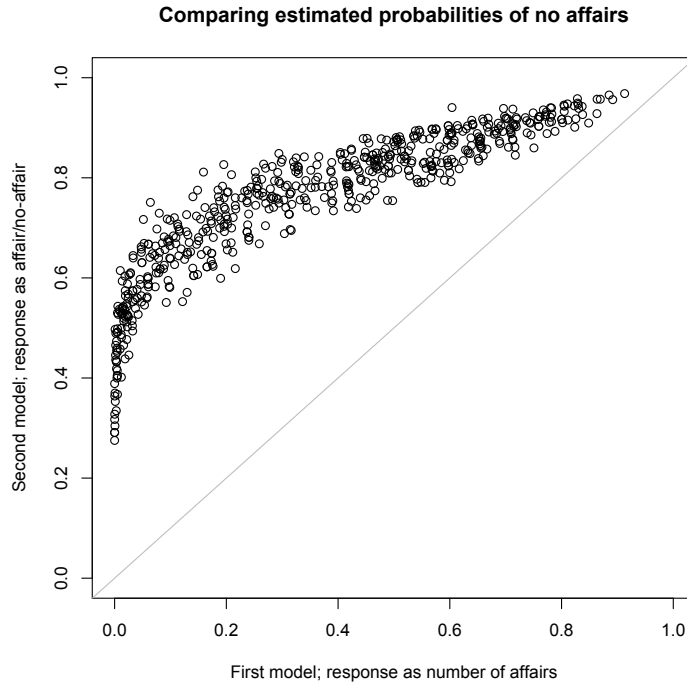


Figure 1: Comparing estimated probabilities of *not* having an affair in the previous year.

```
xlim=c(0,1),ylim=c(0,1))
abline(0,1,col="grey")
```

- (c) Figure 1 shows that the two model's predicted probabilities tend (with exceptions) to go up or down together; they agree, though only roughly, about who is relatively more or less likely to have an affair. But the figure also shows that the second model's estimates for the probability of going for 12 months without an affair are much higher (above the 45-degree line) than the corresponding estimates from the first model. Since they disagree, at least one model must be wrong.
4. ANSWER: The easiest way to do this is to re-cycle the code from Lecture 16. The functions used to check calibration there took a model as an input, on the presumption that the model was built for binary data. We re-write them just a little to take vectors of fitted probabilities, as prepared in the previous problem.

# Based on code from lecture 16

```

# Inputs: lower and upper limits of the probability range to check
# Vector of predicted probabilities
# Vector of binary success/fail responses
# Presumes: fitted.probs and responses have the same length
# fitted.probs and responses are in the same order
# fitted.probs contains probabilities for the type of event in responses
# Outputs: actual frequency of positive responses
# average of predicted probabilities
# rough standard error for predicted probabilities
frequency.vs.probability <- function(p.lower,p.upper=p.lower+0.1,
                                     fitted.probs,responses) {
  # Which rows of the data have fitted probabilities in our range of interest?
  indices <- (fitted.probs >= p.lower) & (fitted.probs < p.upper)
  # For plotting purposes, what's our average predicted probability over this
  # range?
  ave.prob <- mean(fitted.probs[indices])
  # How big a standard error should we see for binomial data with this predicted
  # probability?
  # Rough calculation, could improve by taking account of variation in
  # predicted probabilities, not reliable if difference between p.lower
  # and p.upper was substantial
  se <- sqrt(ave.prob*(1-ave.prob)/sum(indices))
  # How often does the event actually happen?
  # Presumes data is either 0/1 or TRUE/FALSE valued
  frequency <- mean(responses[indices])
  out <- list(frequency=frequency,ave.prob=ave.prob,se=se)
  return(out)
}

```

(a) Use the function.

```

> frequency.vs.probability(p.lower=0,p.upper=0.1,
  fitted.probs=1-p.none.log.r.num,responses=(Affairs$affairs > 0))
$frequency
[1] 0

$ave.prob
[1] 0.08712382

$se
[1] 0.2820164

```

None of the individuals predicted to have less than 10% *annual* chance of an affair did so. On the other hand, one can check that there was only one such person!

(b) Again, take code from lecture 16:

```
f.vs.p <- sapply((0:9)/10,frequency.vs.probability,
```

```

        fitted.probs=1-p.none.log.r.num,
        responses=(Affairs$affairs > 0))
f.vs.p <- data.frame(frequency=unlist(f.vs.p["frequency",]),
                    ave.prob=unlist(f.vs.p["ave.prob",]),
                    se=unlist(f.vs.p["se",]))

```

This applies the function to the probability brackets (0, 0.1], (0.1, 0.2], etc., through (0.9, 1.0], and then stores the results in a data frame. Now plot it:

```

plot(f.vs.p$ave.prob,f.vs.p$frequency,xlim=c(0,1),ylim=c(0,1),
     xlab="Predicted probabilities",ylab="Observed frequencies",
     main="Calibration of first model (response as number of affairs)")
rug(1-p.none.log.r.num,col="grey")
abline(0,1,col="grey")
segments(x0=f.vs.p$ave.prob,y0=f.vs.p$ave.prob-1.96*f.vs.p$se,
        y1=f.vs.p$ave.prob+1.96*f.vs.p$se)

```

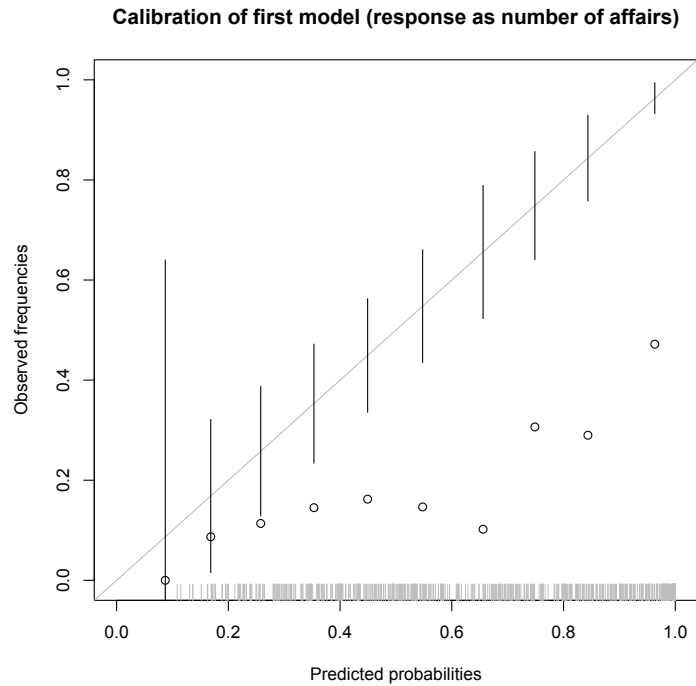


Figure 2: Calibration plot for the first model. Dots show the actual, observed frequency of affairs in each bracket of predicted probabilities; vertical lines are approximate 95% sampling intervals around the predicted probability.

(c) The only thing we have to change is the vector of fitted probabilities.

```
f.vs.p2 <- sapply((0:9)/10,frequency.vs.probability,
                 fitted.probs=1-p.none.log.r,
                 responses=(Affairs$affairs > 0))
f.vs.p2 <- data.frame(frequency=unlist(f.vs.p2["frequency",]),
                    ave.prob=unlist(f.vs.p2["ave.prob",]),
                    se=unlist(f.vs.p2["se",]))

plot(f.vs.p2$ave.prob,f.vs.p2$frequency,xlim=c(0,1),ylim=c(0,1),
     xlab="Predicted probabilities",ylab="Observed frequencies",
     main="Calibration of second model (response as affairs/no affair)")
rug(1-p.none.log.r,col="grey")
abline(0,1,col="grey")
segments(x0=f.vs.p2$ave.prob,y0=f.vs.p2$ave.prob-1.96*f.vs.p2$se,
        y1=f.vs.p2$ave.prob+1.96*f.vs.p2$se)
```

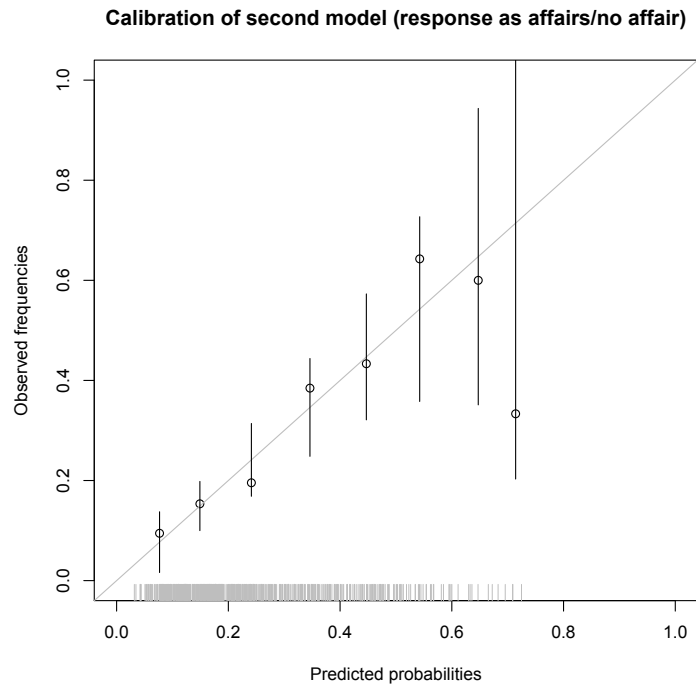


Figure 3: Calibration plot for the second model; see previous figure.

(d) Figure 2 shows the results for the first model and Figure 3 is for the second. Each figure includes a 45-degree line for comparison, and error bars for the predicted probabilities. The points for the first model mostly fall well below the main diagonal, meaning that

the actual frequency is much lower than the predicted probability — it is strongly over-estimating the probability of having at least one affair. The second model does rather better, and seems to be at least roughly calibrated. (Notice that the error bars on the far right in Figure 3 are very wide, because there are very few observations out there.)

It is important to remember that while we want our models to be calibrated, calibration is not enough. For example, a model which predicted a 24.9% probability of anyone having an affair, no matter what, would be calibrated (since that is the actual frequency of having at least one affair), but would clearly miss a lot.

5. ANSWER: Two variables absolutely must be treated as categorical, **sex** and **childrenyes**. However, coding them as 0 or 1 and treating them as numerical values has the same effect as treating them as categorical variables and introducing dummy/indicator variables.

Three variables, religiousness, occupation, and the rating of marriage, while coded with numbers, cannot be regarded as measured on a linear scale. That is, we have no reason to think that the difference between a religiousness of 5 (“very religious”) and of 4 (“somewhat religious”) is as big as the difference between a 2 (“not at all religious”) and a 1 (“anti-religious”). These are *ordinal* variables, because they do fall in a definite order, but treating them the same as age or the number of years married is dubious. Using dummies for these categories seems more appropriate.

Close examination of the table shows that while age, number of years married, years of education, and indeed even the number of affairs are all, in principal, numerical variables, they have each been coded into a few values. So when we see “7” for the number of years someone has been married, that could mean anything from 6 to 8 years, while “15” for that variable could mean 12, 13, or 40 or more. Treating these as ordinary numbers which we can do arithmetic on is somewhat dubious, and it might be better to treat them, too, as ordinal variables.

So, in the end, we can make a case that *all* of the variables in this data set should be treated as categorical.

6. ANSWER:
  - (a) The data is probably better than nothing, but it consists of (i) self-reports about (ii) a sensitive topic which people often lie about, collected from (iii) a self-selected subset (survey-answers) of (iv) a non-random convenience sample (readers of *Psychology Today*). Moreover, it was collected over forty years ago, in a very different social context. In particular, in the late 1960s and early 1970s it became vastly easier to get a legal divorce, which presumably changed the incentives to stay married to one person but have sex with someone else.

- (b) If it comes down to a choice between these two models, the second (binary response) one is at least pretty well calibrated, while the first is *not*.

The fact that the variables can all be seen as categorical (some, more precisely, as ordinal) suggests using conditional density estimation with appropriate kernels (as in Lecture 6). Doing so<sup>1</sup> produces a model more accurate than either logistic regression. However, the only variables *it* says are relevant are religiousness and the rating of the marriage.

---

<sup>1</sup>See Qi Li and Jeff Racine, “Predictor Relevance and Extramarital Affairs”, *Journal of Applied Econometrics* **19** (2004): 533–535; on the class website. You can reproduce their analysis using `npcdens` in the `np` package.