

ESM 206B
Hampton
April 16, 2009

Homework 2 – Regression with binary dependent variables

Environmental economists often want to estimate the demand for “green” goods. The file `appledata.csv` on the ESM 206B website contains data from a telephone survey in which a (fictional) “ecologically friendly” apple was described. Each family was (randomly) presented with a set of prices for regular apples and the eco-labeled prices, and asked how many pounds of each type of apple they would buy. Here we focus on whether the family would buy *any* ecolabeled apples (variable “ecochoice”).

The following table defines the variables.

<u>Variable</u>	<u>Definition</u>
ecochoice	=1 if the consumer said they would buy any ecolabeled apples, otherwise =0
ecopratio	The price per lb for eco-apples divided by the price per lb of regular apples
pci	The per-capita income of the consumer’s family (in \$000s)
educ	The number of years of school attended by the consumer
age	The age of the consumer

You should try to do this homework in R by modifying the script from lab to this setting. You can start by saving the `appledata.csv` file in the Logistic folder we created in lab. Also save a copy of the `LogisticRegression.R` script file and rename the new script file something like ‘`LogisticRegressionHm2.R`’. Your challenge will be to modify `LogisticRegressionHm2.R` in a way that will allow you to answer the questions below. (Note: Make sure to change the directory in the R Console so that it’s set to your Logistic folder).

See attached R script

1. What factors do you think might influence a respondent’s choice to buy the eco-apples?

The quantity of eco-friendly apples an individual demands is probably a function of price ratio, the individual’s preferences for a ‘cleaner’ environment, and their income.

2. Which of these factors are contained within the dataset?

Price and income are directly included. There is no direct information on environmental attitudes. Younger individuals, however, may have stronger

preferences for environmental quality – perhaps because they will be living in this environment longer!

3. Estimate a *linear model* using the data. (For simplicity do not include interaction terms).
- Are all of the factors significant?

Yes, all of the coefficients are significantly different from zero at $\alpha=0.05$.

- Do each of the factors have an effect in the *direction* that you expected (that is, do they have the right sign)? If any do not, speculate on why they don't.

Nothing seems too surprising here. The price ratio is negatively correlated with eco purchases as expected and so is age. Education and income are positive correlated with eco purchases as we might expect.

- Interpret your significant coefficients? (i.e. quantify the meaning of your results).

The coefficient on X3, for example, means that an increase of one year of education increases the probability of an eco purchase by 0.017 holding the other variables constant. The coefficient on X1 means that an increase in the price ratio of 0.5, for example, decreases the probability of an eco purchase by $(0.5)(0.643)= 0.32$.*

- According to this model, what is the probability that a 32 year-old with 18 years of education and in a family with per-capita income of \$75,000 will buy the eco-friendly apple when the price ratio is 1.5?

$$P(Y=1 \mid X1 =1.5, X2=75, X3=18, X4=32) =$$

$$1.29 + -.64(1.5) + .003(75) + .0175(18) - .0023(32) = 0.796$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	1.2900186	0.1568022	8.227	1.04e-15	***
x1	-0.6435072	0.0750386	-8.576	< 2e-16	***
x2	0.0030805	0.0008658	3.558	0.000401	***
x3	0.0175479	0.0075375	2.328	0.020211	*
x4	-0.0023278	0.0010945	-2.127	0.033815	*

 Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

- What are some of the concerns you have about using a linear regression model in this setting?

One problem is that we can construct realistic situations in which our estimated probabilities will not fall within the bounds of 0 and 1. For example, suppose we have the same individual as above but the price ratio is only 1.05 instead of 1.5. We now have:

$$P(Y=1 | X_1 = 1.05, X_2=75, X_3=18, X_4=32) = 1.29 + -.64(1.05) + .003(75) + .0175(18) - .0023(32) = 1.08$$

So our predicted probability is greater than 1, which doesn't make any sense. This is true even though we have selected X values that are well within the range of our observed data.

Another potential problem is that the linear model does not fit the data as well as a logistic regression. The AIC for our linear model is 744.27 and we will see that we can do better with a logistic regression on this criterion.

*> AIC(linear)
[1] 744.2702*

4. Estimate a *logistic regression* using the data

- a. Are all of the factors significant?

They are indeed.

- b. Does each of the factors have an effect in the *direction* that you expected (that is, do they have the right sign)? If any do not, speculate on why they don't.

Again, none of the results seem surprising.

- c. According to this model, what is the probability that a 32 year-old with 18 years of education and in family with per-capita income of \$75,000 will buy the eco-friendly apple when the price ratio is 1.5? (Use the $P = 1 / (1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4)})$ as we discussed in lab). How does this probability compare with the estimate from 3d?

Recall the formula for estimated probabilities with logistic regression:

$$P(Y = 1) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}}$$

Plugging in the X values from above and using the regression coefficients (see below) yields:

$$P(Y=1 \mid X1 =1.5, X2=75, X3=18, X4=32) = 0.88$$

Note that this predicted probability is greater than the linear model (see 3d above).

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	3.947853	0.880139	4.485	7.27e-06	***
x1	-3.359593	0.431032	-7.794	6.48e-15	***
x2	0.025940	0.006957	3.728	0.000193	***
x3	0.088846	0.043297	2.052	0.040168	*
x4	-0.012894	0.005979	-2.157	0.031035	*

- d. Compare the fit of the linear and logistic regression models with AIC.

The AIC from the logistic model is 706.29. This is less than the AIC from the linear model which is 744.27 (see 3e).

- e. Interpret the coefficient on age as an odds ratio as discussed in lecture. (The calculation is e^{β_4} but what does it mean?)

The calculation for e^{β_4} is 0.99. An example illustrates why this is an odds ratio. As above we have

$$P(Y=1 \mid X1) =1.5, X2=75, X3=18, X4=32) = 0.88316$$

$$\text{This implies } P/(1-P) = 0.88316 / (1 - 0.88316) = 7.5587$$

Now increase $X4$ by one (so that age =33) and we can calculate

$$P(Y=1 \mid X1) =1.5, X2=75, X3=18, X4=33) = 0.881822$$

$$\text{This implies } P/(1-P) = 0.881822 / (1 - 0.881822) = 7.4618$$

$$\text{and } 7.4618 / 7.5587 = 0.99 = e^{\beta_4}$$

So we see that e^{β_4} signifies the change in the odds-of-buying-eco-apples ratio for a one unit change in $X4$.