Modeling Long Distance Relationship with a bayesian logistic regression

https://github.com/aclairekeum/ThinkBayes2/finalreport

Claire Keum - December 8, 2014

Description

There are different types of relationship. Some couples are physically close together, such as couples in same university, as known as campus couples, or some couples are far away from each other, like on the other side of the earth. However it is easy to observe people who used to be physically close together, break up when they become a long distance (relationship). Long distance relationship usually represents people who are in a relationship with a person who is 125 miles away.

This case study will do predictions on how likely the couple will break up, using bayesian multiple and logistic regressions. The main factors that are going to be taken care of are as follows:

• How frequent they meet face to face: times/month
• How much time do each spend for texting/calling: hours/week

The study utilizes the on-line references, such as statistics on long distance couples, and some personal data from acquaintances.

Methods

In logistic regressions, the results are usually calculated in terms of log-odds and use a linear model:

o = p/(1-p)

logo = logp - log(1-p)

logo = B0 + B1x1 + B2x2 + E

Our explanatory variables, B0, B1, and B2 would each represent: B0 probability of break up in a relationship, B1 frequency of meeting and B2 duration of talking to each other.

And the probability of whether or not the couple breaks up would be a dependent variable, which can be back calculated from logo.

In TestLDR.py, I first get the range of the beta values, with smf.logit function.

```
ayoungs-mbp:finalreport ayoungkeum$ python TestLDR.py
Optimization terminated successfully.
         Current function value: 0.290625
         Iterations 12
                           Logit Regression Results
==============================================================================
Dep. Variable:           relationship   No. Observations:                   12
Model:                         Logit   Df Residuals:                        9
Method:                          MLE   Df Model:                            2
Date:                Tue, 09 Dec 2014   Pseudo R-squ.:                  0.4832
Time:                        00:02:38   Log-Likelihood:                -3.4875
converged:                       True   LL-Null:                       -6.7480
                                        LLR p-value:                   0.03837
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
Intercept      3.4168      4.843      0.706      0.480        -6.074    12.908
meet          -7.8834     13.847     -0.569      0.569       -35.023    19.256
talk          -0.3799      1.549     -0.245      0.806        -3.416     2.657
==============================================================================
```

Figure(a): Result from running logit with my dataset.

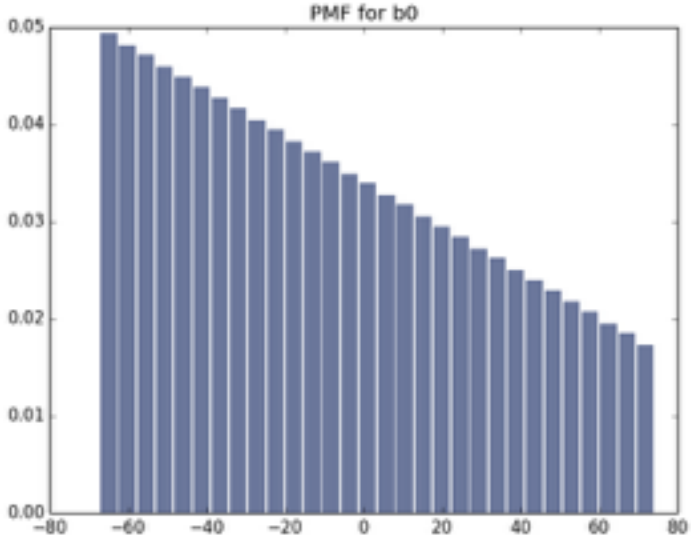
By using statsmodels's logit function, I was able to get an estimate for two explanatory parameters which are shown in the figure(a) as meet and talk and an intercept, B0 value. As it is shown in the result, how much they meet affects the probability of breaking up a lot, compared to how often they talk.

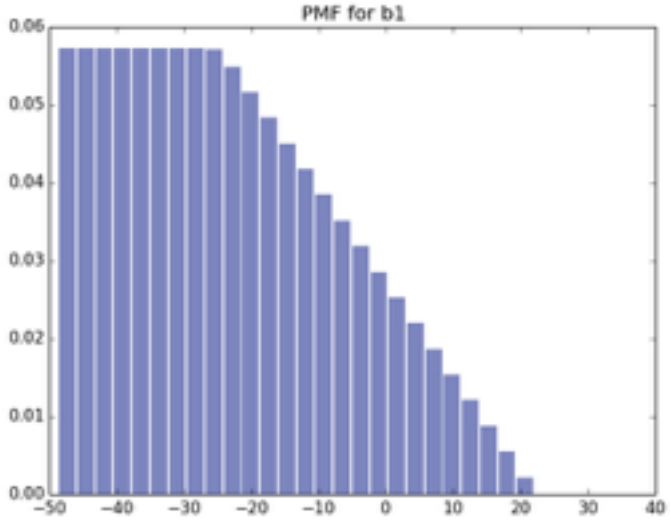
Creating Hypos

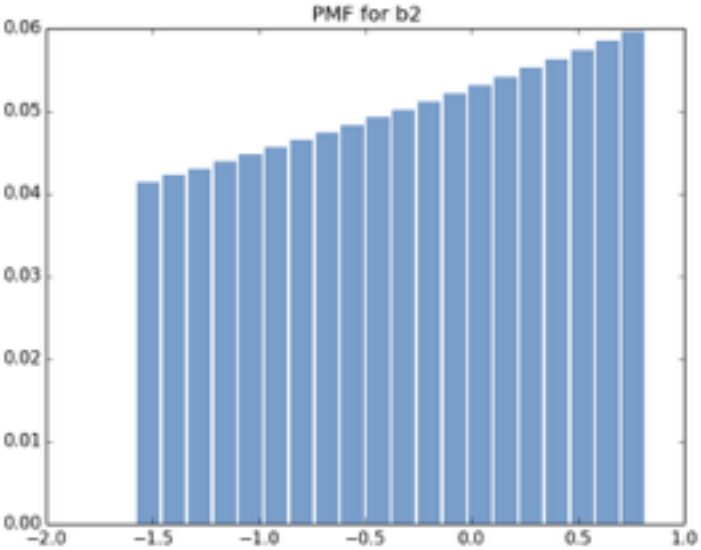
From the figure(a), we have an estimated B0, B1, and B2 values. From those values, we can have pmfs for each parameters with a certain range. The function that returns the pmf, is pmf_from_data.

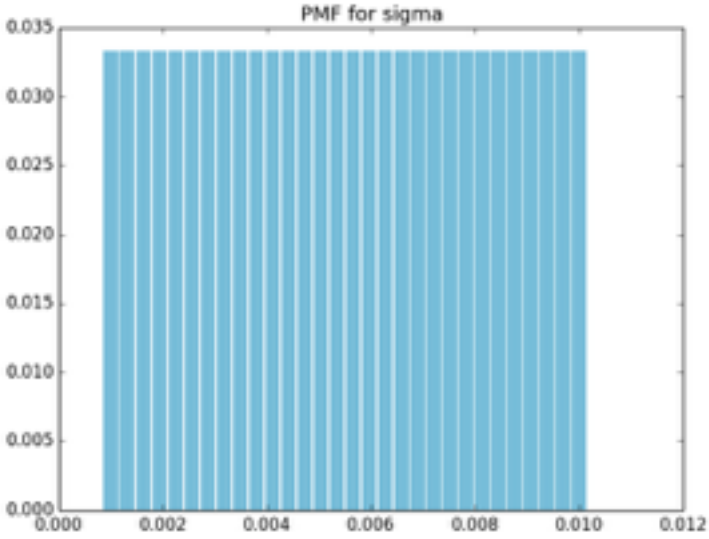
With the hypos, I have calculated to see the maximum likelihood for each beta values.

Maximum Likelihood for each parameters









These are the maximum likelihoods for each parameters. Instead of using just a fixed values for the independent variable, looking at the pmfs for each of them, and predicting how likely my model could be used in a real situation would have more applicability.

The reason why the graphs don't look like a bell curve is a lack of dataset. Dataset that I have collected was through my acquaintances that it is not always the case where the data has

correlation to each other. Relationship between people can be also varied depends on their psychological mood change, or other non-quantitative components.

How can this case study can be done better?

The biggest advantage of using thinkbayes suite is that it is easy to use actual data in order to model a certain phenomenon. In my case, however, there were only 12 data points that they barely had a correlation between them, which made it hard for me to create a bell curve, since the standard deviation is going to be very high. I sent an email to the author of the online site that I have cited for this case study, so hopefully once I get more dataset, I may able to do some more useful analysis on the relationship prediction.