*Session 16: Assignment 16.1*

**Extramarital Affairs Dataset**

The dataset is [affairs dataset](http://statsmodels.sourceforge.net/stable/datasets/generated/fair.html) and comes with [Statsmodels](http://statsmodels.sourceforge.net/" \o "Statsmodels). It was derived from a survey of women in 1974 by Redbook magazine, in which married women were asked about their participation in extramarital affairs. More information about the study is available in a [1978 paper](https://fairmodel.econ.yale.edu/rayfair/pdf/1978a200.pdf) from the Journal of Political Economy.

**Description of Variables**

The dataset contains 6366 observations of 9 variables:

* rate\_marriage: woman's rating of her marriage (1 = very poor, 5 = very good)
* age: woman's age
* yrs\_married: number of years married
* children: number of children
* religious: woman's rating of how religious she is (1 = not religious, 4 = strongly religious)
* educ: level of education (9 = grade school, 12 = high school, 14 = some college, 16 = college graduate, 17 = some graduate school, 20 = advanced degree)
* occupation: woman's occupation (1 = student, 2 = farming/semi-skilled/unskilled, 3 = "white collar", 4 = teacher/nurse/writer/technician/skilled, 5 = managerial/business, 6 = professional with advanced degree)
* occupation\_husb: husband's occupation (same coding as above)
* affairs: time spent in extra-marital affairs

**Problem Statement**

We treat this as a classification problem by creating a new binary variable affair (did the woman have at least one affair?) and try to predict the classification for each woman.

### Import Modules

## Preprocessing

## Data Exploration

We can see that on average, women who have affairs rate their marriages lower, which is to be expected. Let's take another look at the rate\_marriage variable.

## Data Visualization

take a look at the distribution of marriage ratings for those having affairs versus those not having affairs.

Use a stacked barplot to look at the percentage of women having affairs by number of years of marriage.

## Prepare Data for Logistic Regression

We add an intercept column as well as dummy variables for occupation and occupation\_husb, since we are treating them as categorial variables. The dmatrices function from the [patsy module](http://patsy.readthedocs.org/en/latest/) can do that using formula language.

## Logistic Regression

0.72588752748978946

72% accuracy may seem OK, but what's the accuracy if we simply predict no for all observations in the dataset?

0.32249450204209867

Only 32% of the women had affairs, which means that you could obtain 68% accuracy by always predicting no. So we're doing better than the null error rate, but not by much.

Increases in marriage rating and religiousness correspond to a decrease in the likelihood of having an affair. For both the wife's occupation and the husband's occupation, the lowest likelihood of having an affair corresponds to the baseline occupation (student), since all of the dummy coefficients are positive.

## Model Evaluation Using a Validation Set

We see that the classifier is predicting a 1 (having an affair) any time the probability in the second column is greater than 0.5

## Model Evaluation Using Cross-Validation

Try 10-fold cross-validation, to see if the accuracy holds up

## Predicting the Probability of an Affair

Let's predict the probability of an affair for a random woman not present in the dataset. She's a 25-year-old teacher who graduated college, has been married for 3 years, has 1 child, rates herself as strongly religious, rates her marriage as fair, and her husband is a farmer.

The predicted probability of an affair is 23%.

**Improving the Model**

The following could be tried to improve the model:

* including interaction terms
* removing features
* regularization techniques
* using a non-linear model