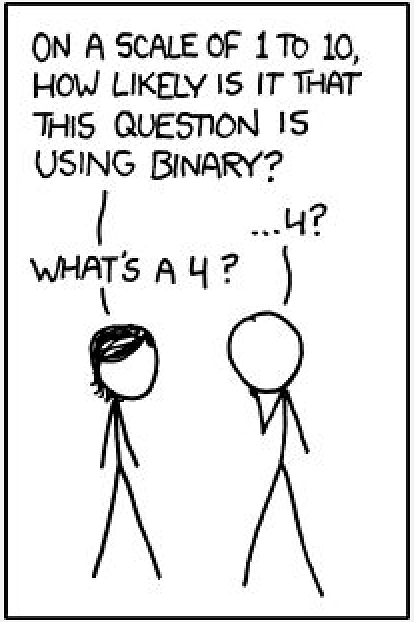
Logistic Regression

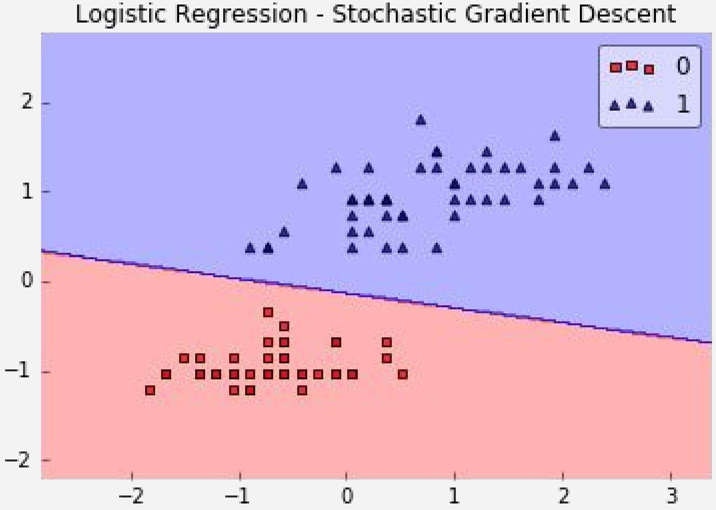

ON A SCALE OF 1 TO 10, HOW LIKELY IS IT THAT THIS QUESTION IS USING BINARY?
WHAT'S A 4?
...4?

Categorical data

- Categorical (or nominal) variable: anything with two or more categories, that have no intrinsic ordering to them (like hair colour, or occupation)
- Ordinal variable: a categorical variable with clear ordering to it (like level of schooling, opinion of current Congress on a scale of 1 - 5)
- Interval variable: Like an ordinal variable, except the intervals between values are equally spaced (like income bands)

Classification

Classification is a method of recognizing patterns in data. Classification attempts to identify to which set of observations a new data point belongs.



Logistic Regression

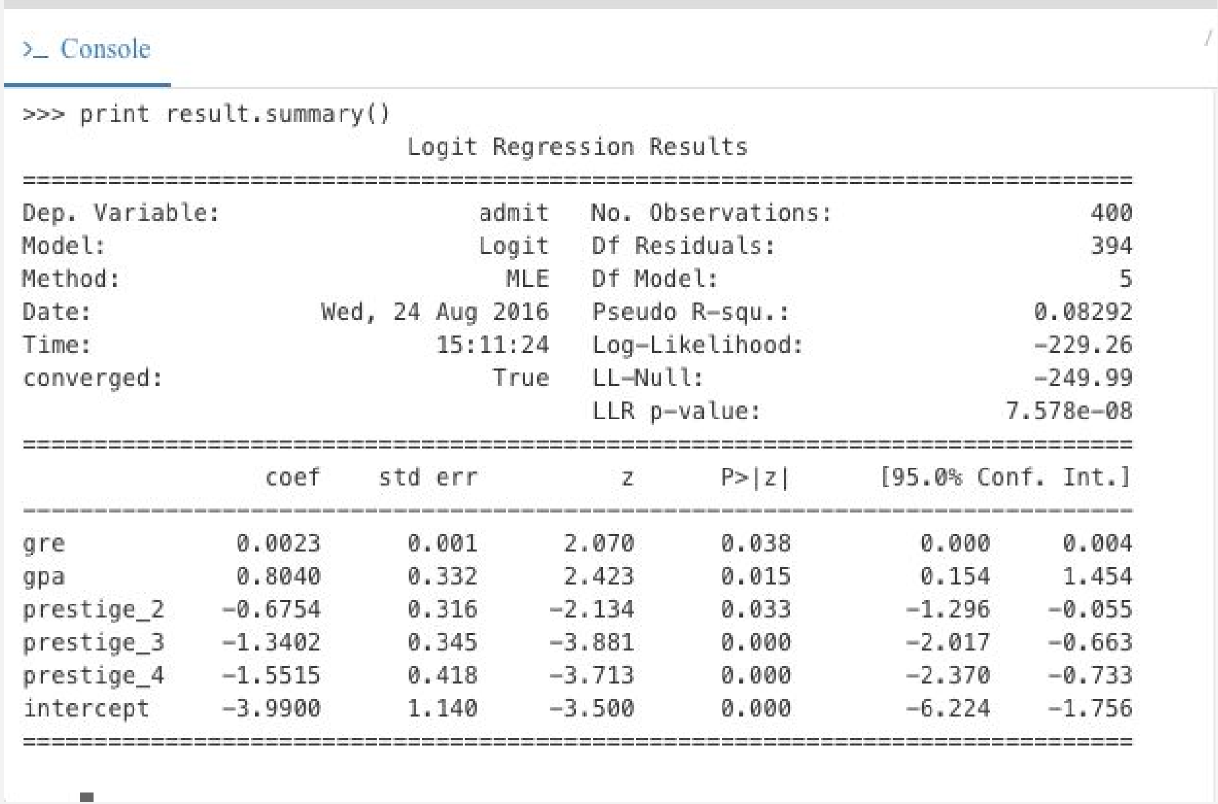
Logistic regression is a regression model where the dependent variable is categorical instead of continuous.

It learns a model for binary classification based on the independent variables we input, and outputs the probability that a sample belongs to class 0 or class 1.

Logistic Regression: a few things to keep in mind

- Binary logistic regression requires the dependent variable to be binary and ordinal logistic regression requires the dependent variable to be ordinal.
- Logistic regression assumes that $P(Y=1)$ is the probability of the event occurring, so it is necessary that the dependent variable is coded to recognise this.
- Logistic regression requires each observation to be independent. That is that the data-points should not be from any dependent samples design, e.g., before-after measurements, or matched pairings.
- All independent variables should also be independent from each other.

Interpretation



```
>>> print result.summary()
                           Logit Regression Results
==============================================================================
Dep. Variable:                  admit   No. Observations:                  400
Model:                          Logit   Df Residuals:                      394
Method:                           MLE   Df Model:                            5
Date:                Wed, 24 Aug 2016   Pseudo R-squ.:                 0.08292
Time:                        15:11:24   Log-Likelihood:                -229.26
converged:                       True   LL-Null:                       -249.99
                                        LLR p-value:                 7.578e-08
==============================================================================
                 coef    std err          z      P>|z|      [95.0% Conf. Int.]
------------------------------------------------------------------------------
gre            0.0023      0.001      2.070      0.038         0.000     0.004
gpa            0.8040      0.332      2.423      0.015         0.154     1.454
prestige_2    -0.6754      0.316     -2.134      0.033        -1.296    -0.055
prestige_3    -1.3402      0.345     -3.881      0.000        -2.017    -0.663
prestige_4    -1.5515      0.418     -3.713      0.000        -2.370    -0.733
intercept     -3.9900      1.140     -3.500      0.000        -6.224    -1.756
==============================================================================
```

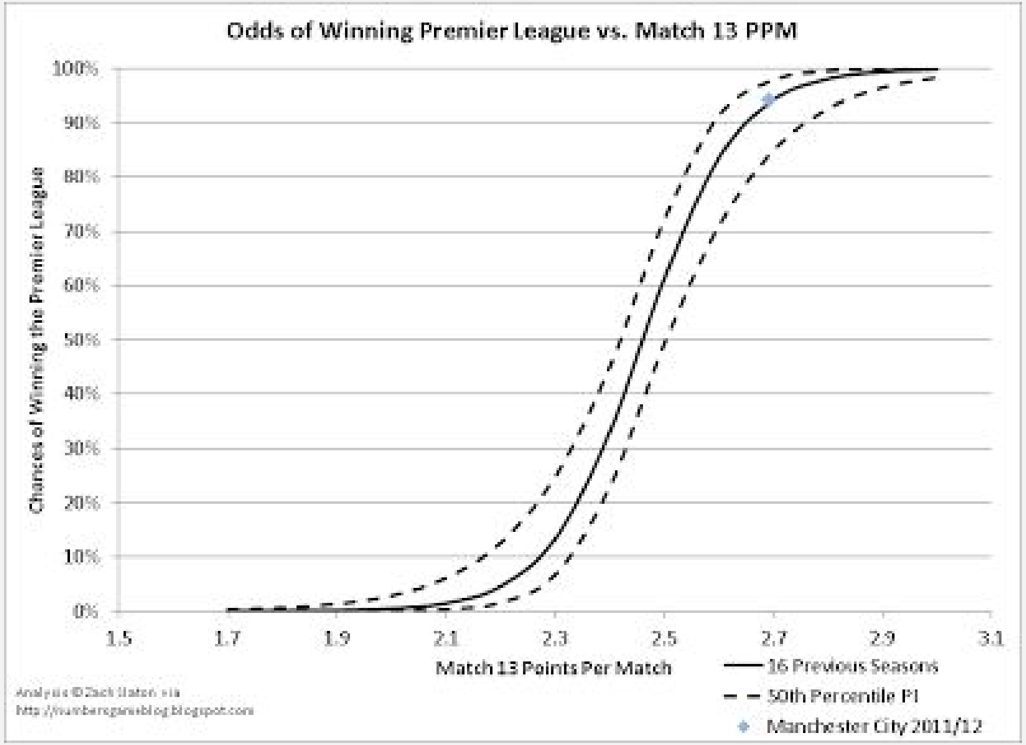
```
print metrics.confusion_matrix(y_test, predicted)
print metrics.classification_report(y_test, predicted)
```

```
[[1169  134]
 [ 382  225]]
             precision    recall  f1-score   support

        0.0       0.75      0.90      0.82      1303
        1.0       0.63      0.37      0.47       607

avg / total       0.71      0.73      0.71      1910
```

Interpretation

- If a variable has a high coefficient, the slope is a sharp one
- If a variable has a low coefficient, the slope is gradual
- Once you take the exponential of the coefficient, it tells you how a one unit increase or decrease in the variable affects the odds (probability) of the outcome occurring.

Practical examples

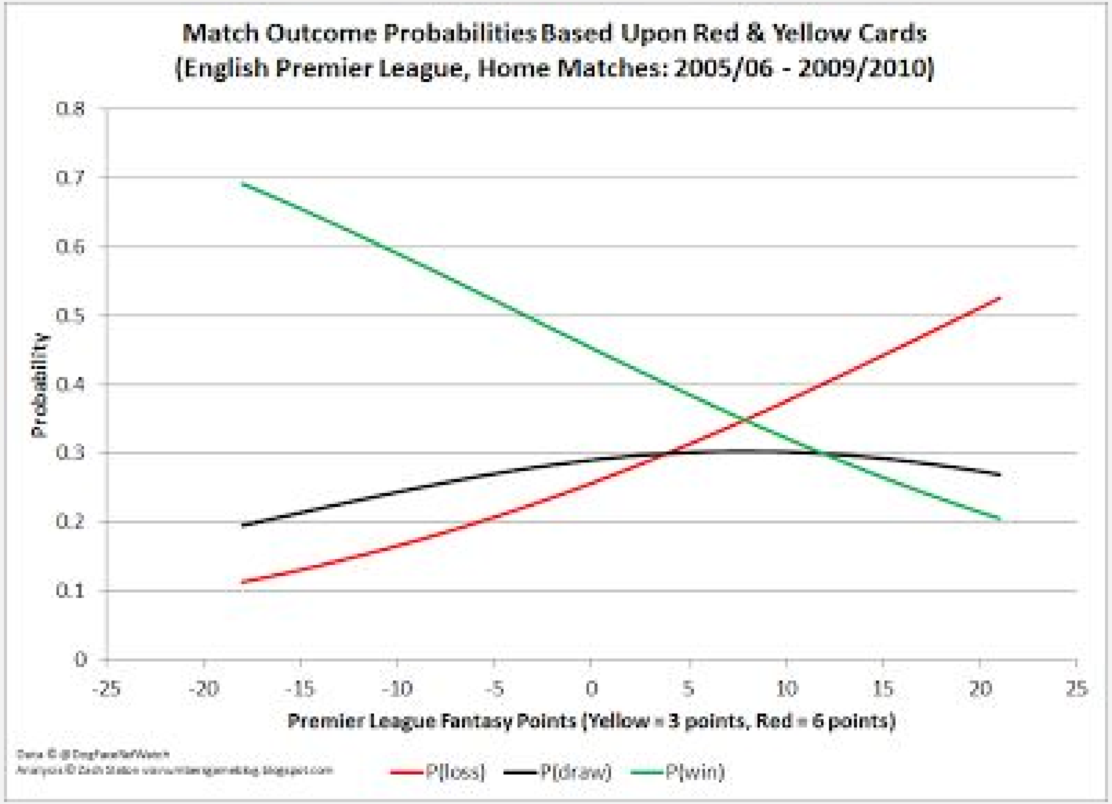


Analysis © Zach Slaton via
http://numbersgameblog.blogspot.com

Practical example #1

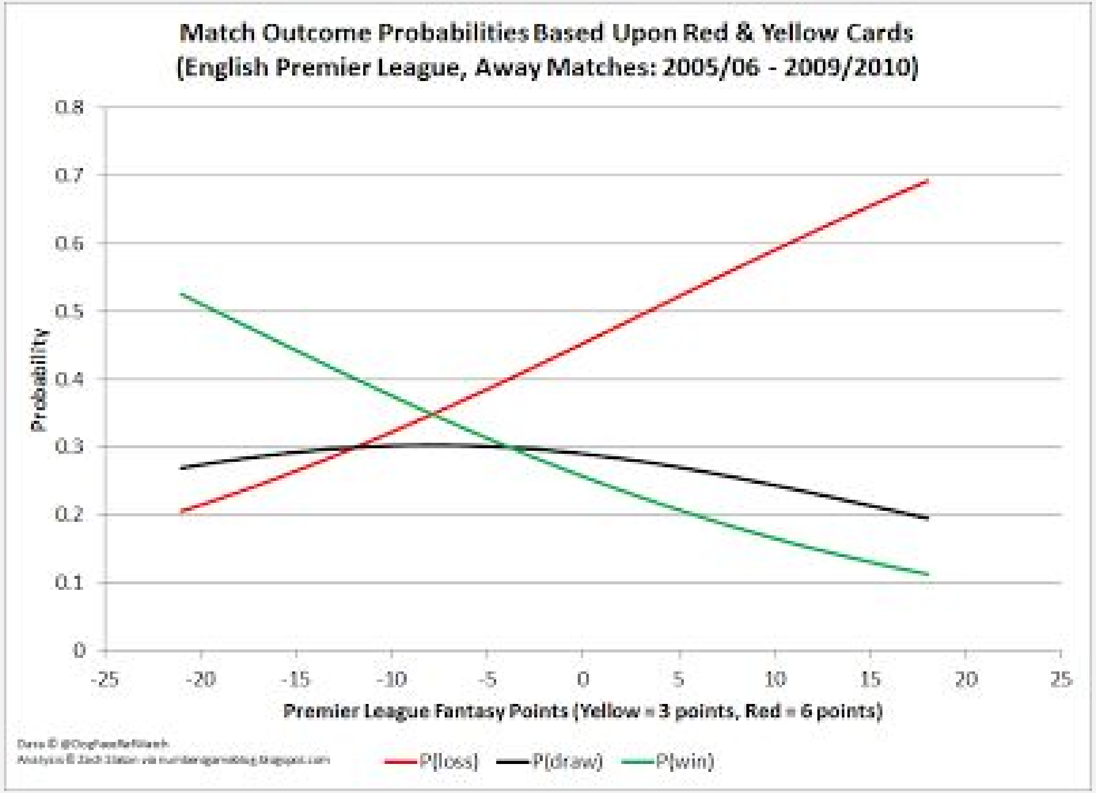
- The outcome in this case is binary: whether a team finished top or not
- The only independent variable was whether or not a team was top of the league on Match Day 13 (approximately mid-November)
- Analysis includes 50th percentile bounds
- Fun fact: the analysis shows that Manchester City had a 94% chance of winning the league in 2011/2012, which they did. It was their first league cup victory since 1968.

Practical examples



Data © @DogFaceRefWatch
Analysis © Zach Slaton www.numbersgameblog.blogspot.com

Practical examples



Practical example #2

- Looks at predicting whether a team will win, lose, or draw based on min and max values for the number of red and yellow cards the team receives (and whether the team is playing at home or away)
- Based on this analysis, it looks like venue plays a huge role in whether or not red and yellow cards play a role in a team's outcome during a game