Regression Models

Types:

Linear

Poisson

Logistic

Types:

Linear

Model relationships between two continuous (quantitative) variables

Poisson

Modeling count data

Logistic

Model relationships when response variable is categorical in nature

Linear Regression

$$Y \approx \beta 0 + \beta 1X$$

X: predictor

Y: response

$\beta 0$: Intercept

$\beta 1$: Coef

Linear Regression

Simple regression modeling:

- Python: StatsModel, SciKit - SKLearn
- R: built-in linear modeling

Python: StatsModels v. SKLearn

StatsModels

- Traditional statistical modeling
  - Estimation
  - Statistical testing
- Summary method

SKLearn

- Machine learning
- Parameter regularization
  - Reduce overfitting, greater bias
- Large, sparse data
- No summary method, need to print coef & intercept separately

Understanding the summary

The left part of the first table provides basic information about the model fit:

Dep. Variable	Which variable is the response in the model
Model	What model you are using in the fit
Method	How the parameters of the model were calculated
No. Observations	The number of observations (examples)
DF Residuals	Degrees of freedom of the residuals. Number of observations – number of parameters
DF Model	Number of parameters in the model (not including the constant term if present)

Understanding the summary

The right part of the first table shows the goodness of fit:

R-squared	The coefficient of determination. A statistical measure of how well the regression line approximates the real data points
Adj. R-squared	The above value adjusted based on the number of observations and the degrees-of-freedom of the residuals
F-statistic	A measure how significant the fit is. The mean squared error of the model divided by the mean squared error of the residuals
Prob (F-statistic)	The probability that you would get the above statistic, given the null hypothesis that they are unrelated
Log-likelihood	The log of the likelihood function.
AIC	The Akaike Information Criterion. Adjusts the log-likelihood based on the number of observations and the complexity of the model.
BIC	The Bayesian Information Criterion. Similar to the AIC, but has a higher penalty for models with more parameters.

Understanding the summary

The second table reports for each of the coefficients:

coef	The estimated value of the coefficient
std err	The basic standard error of the estimate of the coefficient. More sophisticated errors are also available.
t	The t-statistic value. This is a measure of how statistically significant the coefficient is.
P > |t|	P-value that the null-hypothesis that the coefficient = 0 is true. If it is less than the confidence level, often 0.05, it indicates that there is a statistically significant relationship between the term and the response.
[95.0% Conf. Interval]	The lower and upper values of the 95% confidence interval

Understanding the summary

Finally, there are several statistical tests to assess the distribution of the residuals:

Skewness	A measure of the symmetry of the data about the mean. Normally-distributed errors should be symmetrically distributed about the mean (equal amounts above and below the line).
Kurtosis	A measure of the shape of the distribution. Compares the amount of data close to the mean with those far away from the mean (in the tails).
Omnibus	D’Angostino’s test. It provides a combined statistical test for the presence of skewness and kurtosis.
Prob(Omnibus)	The above statistic turned into a probability
Jarque-Bera	A different test of the skewness and kurtosis
Prob (JB)	The above statistic turned into a probability
Durbin-Watson	A test for the presence of autocorrelation (that the errors are not independent.) Often important in time-series analysis
Cond. No	A test for multicollinearity (if in a fit with multiple parameters, the parameters are related with each other).