Ordinary Least Squares in Python

Linear regression, also called Ordinary Least-Squares (OLS) Regression, is probably the most commonly used technique in Statistical Learning.

Linear Regression and Ordinary Least Squares

Linear regression is one of the simplest and most commonly used modeling techniques. It makes very strong assumptions about the relationship between the predictor variables (the X) and the response (the Y). It assumes that this relationship takes the form:

$$y = \beta_0 + \beta_1 * x$$

Ordinary Least Squares is the simplest and most common estimator in which the two βs are chosen to minimize the square of the distance between the predicted values and the actual values. Even though this model is quite rigid and often does not reflect the true relationship, this still remains a popular approach for several reasons. For one, it is computationally cheap to calculate the coefficients. It is also easier to interpret than more sophisticated models, and in situations where the goal is understanding a simple model in detail, rather than estimating the response well, they can provide insight into what the model captures. Finally, in situations where there is a lot of noise, it may be hard to find the true functional form, so a constrained model can perform quite well compared to a complex model which is more affected by noise.

The resulting model is represented as follows:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 * x$$

Here the hats on the variables represent the fact that they are estimated from the data we have available. The βs are termed the parameters of the model or the coefficients. β_0 is called the constant term or the intercept.

OLS Regression Results

Dep. Variable:	Price	**R-squared:**	0.064
Model:	OLS	**Adj. R-squared:**	0.060
Method:	Least Squares	**F-statistic:**	18.11
Date:	Fri, 20 Jul 2018	**Prob (F-statistic):**	2.23e-11
Time:	00:32:35	**Log-Likelihood:**	-9207.1
No. Observations:	804	**AIC:**	1.842e+04
Df Residuals:	801	**BIC:**	1.843e+04
Df Model:	3		
Covariance Type:	nonrobust		

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

Element	Description
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)

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

Element	Description
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

Element	Description
	model.
BIC	The Bayesian Information Criterion. Similar to the AIC, but has a higher penalty for models with more parameters.

The second table reports for each of the coefficients

	coef	std err	t	P>|t|	[0.025	0.975]
Mileage	-1272.3412	804.623	-1.581	0.114	-2851.759	307.077
Cylinder	5587.4472	804.509	6.945	0.000	4008.252	7166.642
Doors	-1404.5513	804.275	-1.746	0.081	-2983.288	174.185

	Description
	The name of the term in the model
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

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

Omnibus:	157.913	**Durbin-Watson:**	0.008
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	257.529
Skew:	1.278	**Prob(JB):**	1.20e-56
Kurtosis:	4.074	**Cond. No.**	1.03

Element	Description
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

Element	Description
	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).