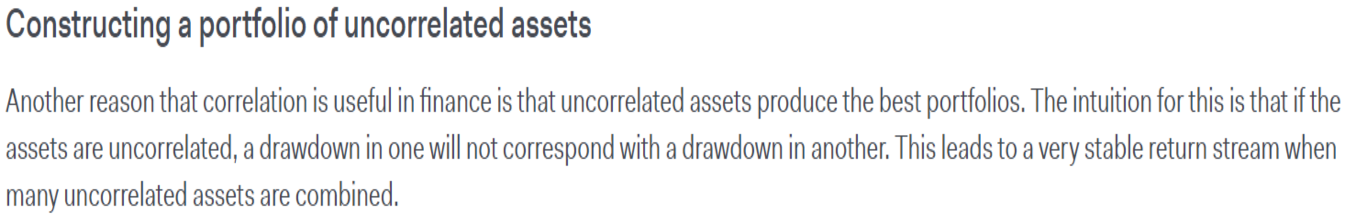
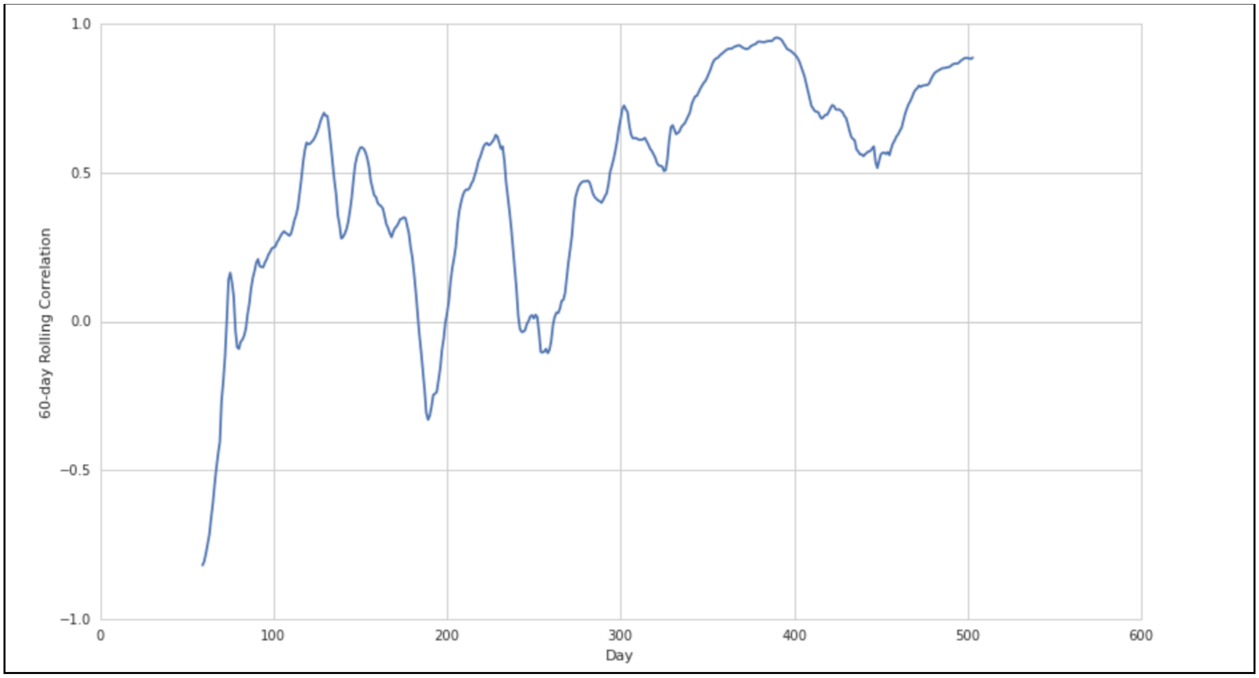
Quantopian:



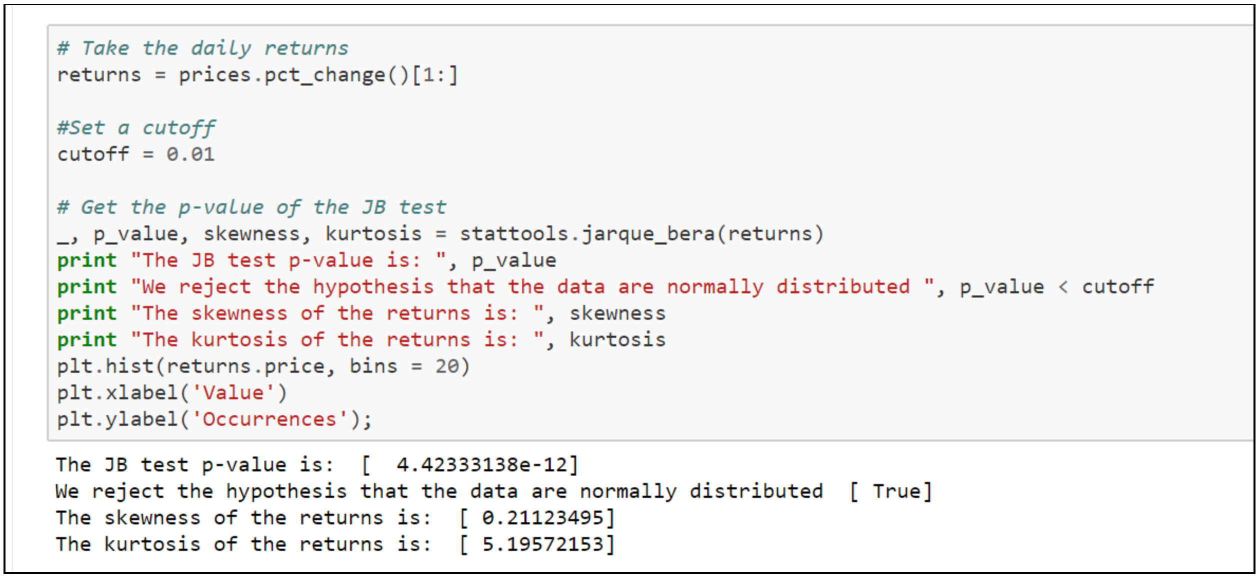
Constructing a portfolio of uncorrelated assets

Another reason that correlation is useful in finance is that uncorrelated assets produce the best portfolios. The intuition for this is that if the assets are uncorrelated, a drawdown in one will not correspond with a drawdown in another. This leads to a very stable return stream when many uncorrelated assets are combined.

Rolling correlation to find change in correlation over a time period.



TEST TO CHECK NORMALITY : JARQUE_BERA TEST



```
# Take the daily returns
returns = prices.pct_change()[1:]

#Set a cutoff
cutoff = 0.01

# Get the p-value of the JB test
_, p_value, skewness, kurtosis = stattools.jarque_bera(returns)
print "The JB test p-value is: ", p_value
print "We reject the hypothesis that the data are normally distributed ", p_value < cutoff
print "The skewness of the returns is: ", skewness
print "The kurtosis of the returns is: ", kurtosis
plt.hist(returns.price, bins = 20)
plt.xlabel('Value')
plt.ylabel('Occurrences');
```

```
The JB test p-value is:  [  4.42333138e-12]
We reject the hypothesis that the data are normally distributed  [ True]
The skewness of the returns is:  [ 0.21123495]
The kurtosis of the returns is:  [ 5.19572153]
```

@ Shubham Kotal

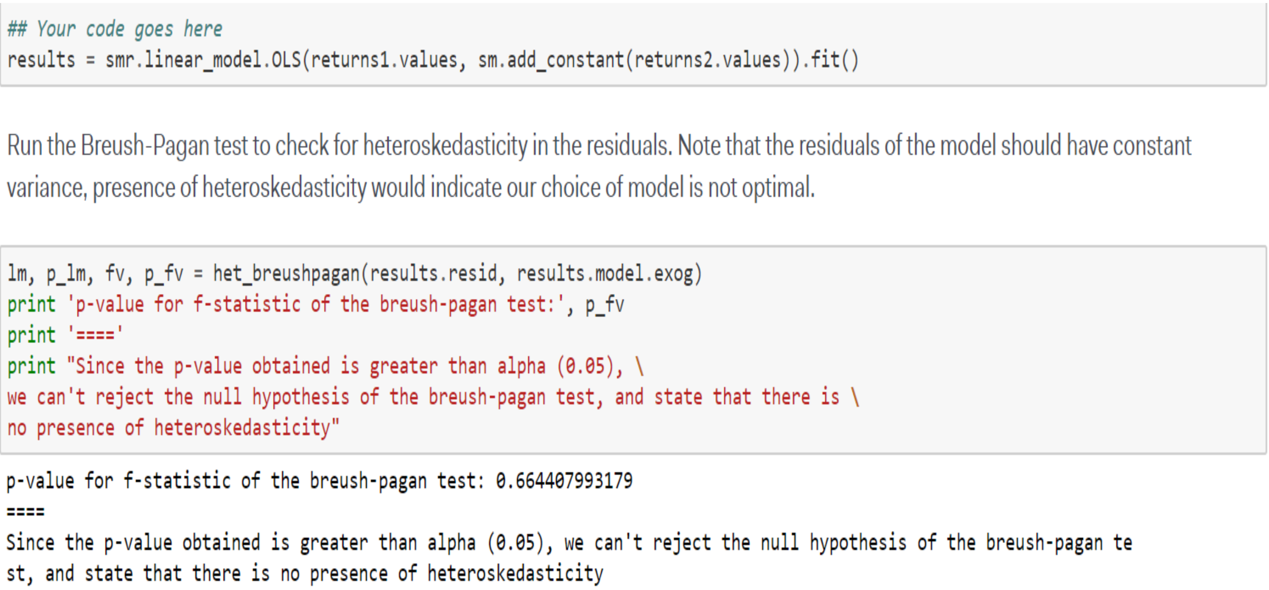
Quantopian:

Linear Regression

1. **Using OLS (Ordinary least square) - The process of iteration**

2. **Using loss function - Gradient Descent**

Test to check heteroskedasticity in residuals.
Using Breush-pagan test



```
## Your code goes here
results = smr.linear_model.OLS(returns1.values, sm.add_constant(returns2.values)).fit()
```

Run the Breush-Pagan test to check for heteroskedasticity in the residuals. Note that the residuals of the model should have constant variance, presence of heteroskedasticity would indicate our choice of model is not optimal.

```
lm, p_lm, fv, p_fv = het_breushpagan(results.resid, results.model.exog)
print 'p-value for f-statistic of the breush-pagan test:', p_fv
print '===='
print "Since the p-value obtained is greater than alpha (0.05), \
we can't reject the null hypothesis of the breush-pagan test, and state that there is \
no presence of heteroskedasticity"
```

```
p-value for f-statistic of the breush-pagan test: 0.664407993179
====
Since the p-value obtained is greater than alpha (0.05), we can't reject the null hypothesis of the breush-pagan te
st, and state that there is no presence of heteroskedasticity
```

The regression model relies on several assumptions:

- The independent variable is not random.
- The variance of the error term is constant across observations. This is important for evaluating the goodness of the fit.
- The errors are not autocorrelated. The **Durbin-Watson statistic** detects this; if it is close to 2, there is no autocorrelation.
- The errors are normally distributed. If this does not hold, we cannot use some of the statistics, such as the F-test.

@ Shubham Kotal

Quantopian:

```
slr2 = regression.linear_model.OLS(y2, sm.add_constant(xs)).fit()
slr2.summary()
```

OLS Regression Results

Dep. Variable:	**y**	**R-squared:**	**1.000**
Model:	**OLS**	**Adj. R-squared:**	**1.000**
Method:	**Least Squares**	**F-statistic:**	**3.092e+26**
Date:	**Fri, 22 Apr 2016**	**Prob (F-statistic):**	**0.00**
Time:	**20:43:27**	**Log-Likelihood:**	**1700.7**
No. Observations:	**100**	**AIC:**	**-3393.**
Df Residuals:	**96**	**BIC:**	**-3383.**
Df Model:	**3**		
Covariance Type:	**nonrobust**		

Focus on the Residuals

Rather than focusing on your model construction, it is possible to gain a huge amount of information from your residuals (errors). Your model may be incredibly complex and impossible to analyze, but as long as you have predictions and observed values, you can compute residuals. Once you have your residuals you can perform many statistical tests.

If your residuals do not follow a given distribution (usually normal, but depends on your model), then you know that something is wrong and you should be concerned with the accuracy of your predictions.

Quantopian:

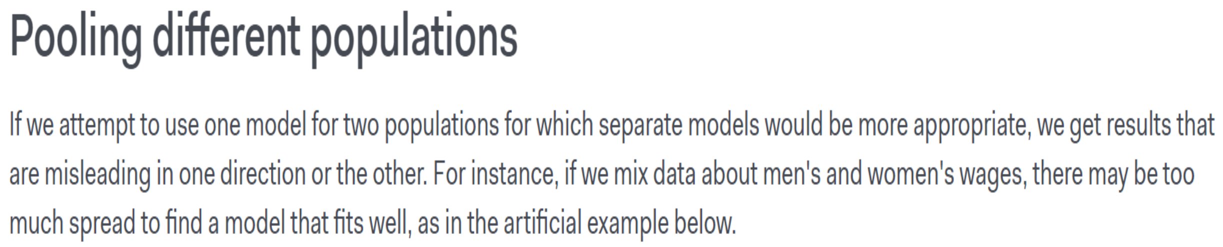
Correcting for Heteroskedasticity

How does heteroskedasticity affect our analysis? The problematic situation, known as conditional heteroskedasticity, is when the error variance is correlated with the independent variables as it is above. This makes the F-test for regression significance and t-tests for the significances of individual coefficients unreliable. Most often this results in overestimation of the significance of the fit.

Example: Sharpe ratio

One statistic often used to describe the performance of assets and portfolios is the Sharpe ratio, which measures the additional return per unit additional risk achieved by a portfolio, relative to a risk-free source of return such as Treasury bills.

Sharpe ratio is the measure of risk-adjusted return of a financial portfolio. A portfolio with a higher **Sharpe ratio** is considered superior relative to its peers. ... In **simple** terms, it shows how much additional return an investor earns by taking additional risk.

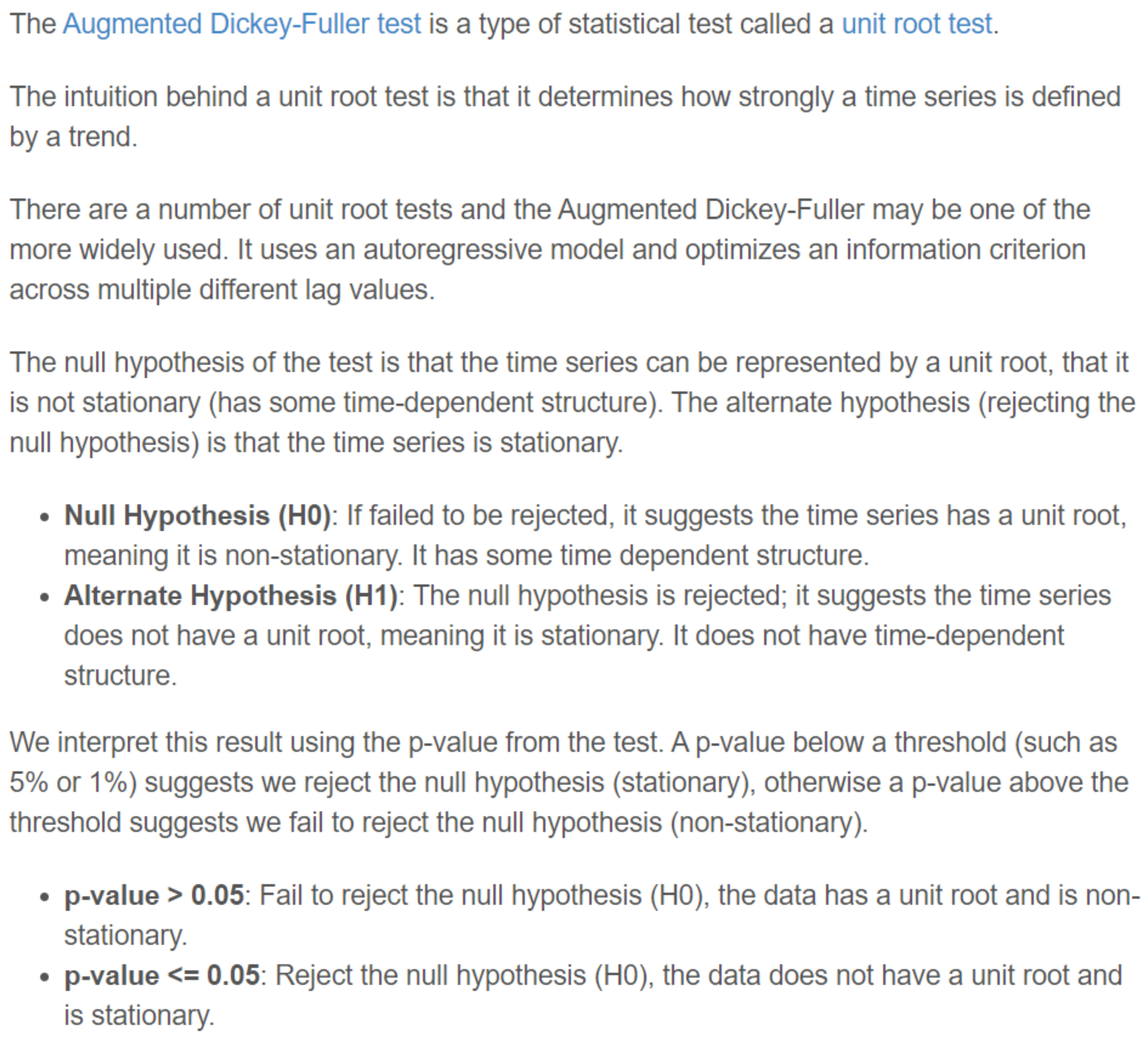


Pooling different populations

If we attempt to use one model for two populations for which separate models would be more appropriate, we get results that are misleading in one direction or the other. For instance, if we mix data about men's and women's wages, there may be too much spread to find a model that fits well, as in the artificial example below.

Quantopian:

Dickey-Fuller test



The Augmented Dickey-Fuller test is a type of statistical test called a unit root test.

The intuition behind a unit root test is that it determines how strongly a time series is defined by a trend.

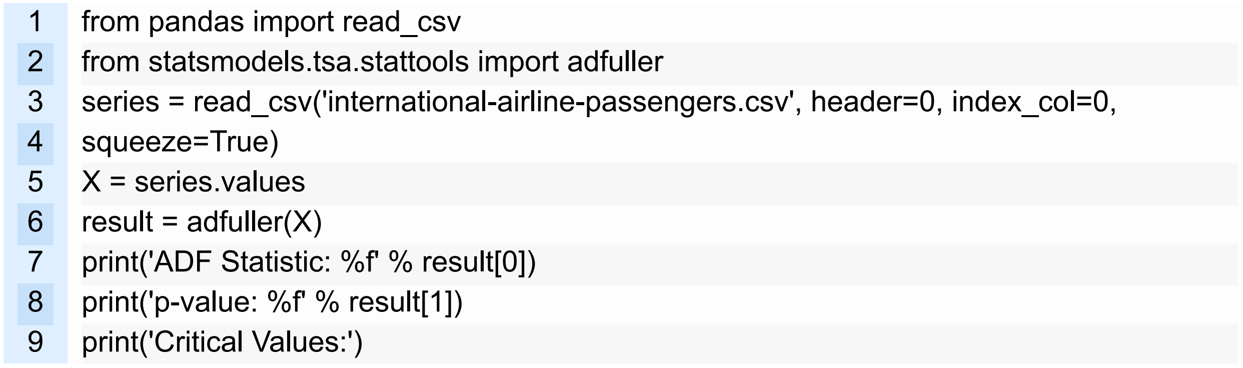
There are a number of unit root tests and the Augmented Dickey-Fuller may be one of the more widely used. It uses an autoregressive model and optimizes an information criterion across multiple different lag values.

The null hypothesis of the test is that the time series can be represented by a unit root, that it is not stationary (has some time-dependent structure). The alternate hypothesis (rejecting the null hypothesis) is that the time series is stationary.

- **Null Hypothesis (H0)**: If failed to be rejected, it suggests the time series has a unit root, meaning it is non-stationary. It has some time dependent structure.
- **Alternate Hypothesis (H1)**: The null hypothesis is rejected; it suggests the time series does not have a unit root, meaning it is stationary. It does not have time-dependent structure.

We interpret this result using the p-value from the test. A p-value below a threshold (such as 5% or 1%) suggests we reject the null hypothesis (stationary), otherwise a p-value above the threshold suggests we fail to reject the null hypothesis (non-stationary).

- **p-value > 0.05**: Fail to reject the null hypothesis (H0), the data has a unit root and is non-stationary.
- **p-value <= 0.05**: Reject the null hypothesis (H0), the data does not have a unit root and is stationary.



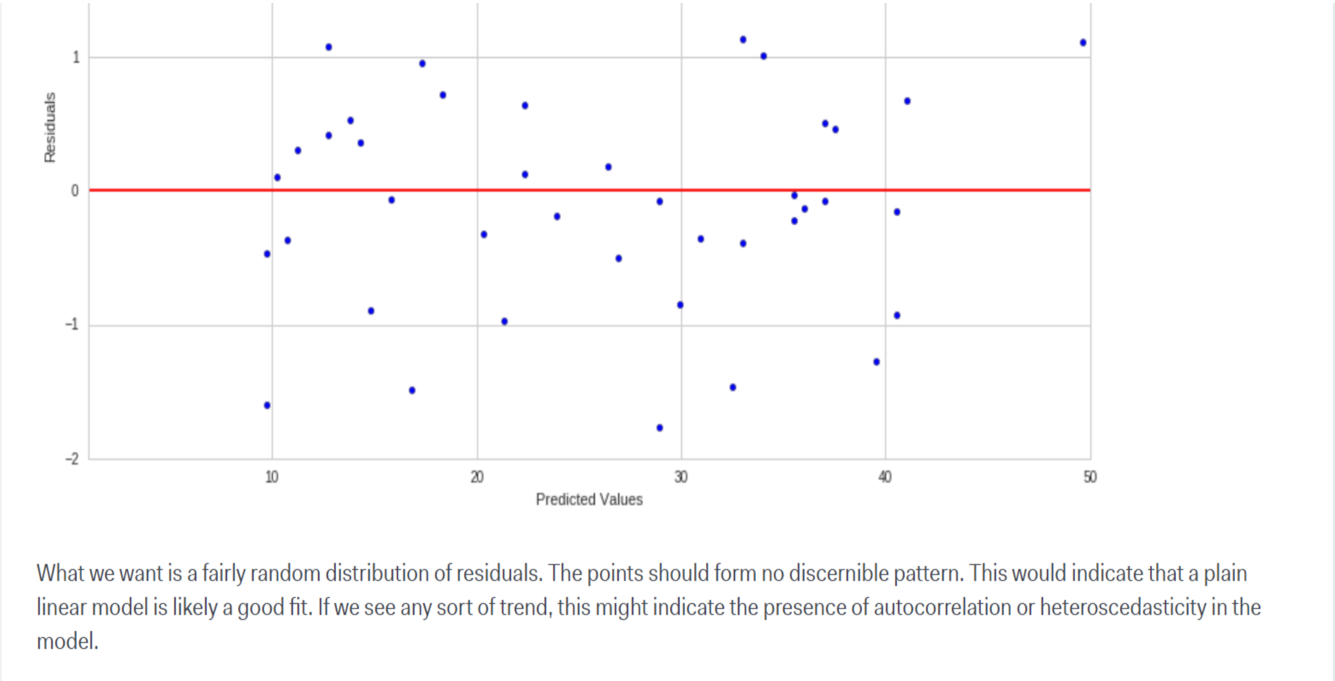
```
from pandas import read_csv
from statsmodels.tsa.stattools import adfuller
series = read_csv('international-airline-passengers.csv', header=0, index_col=0,
squeeze=True)
X = series.values
result = adfuller(X)
print('ADF Statistic: %f' % result[0])
print('p-value: %f' % result[1])
print('Critical Values:')
```

@ Shubham Kotal

Quantopian:

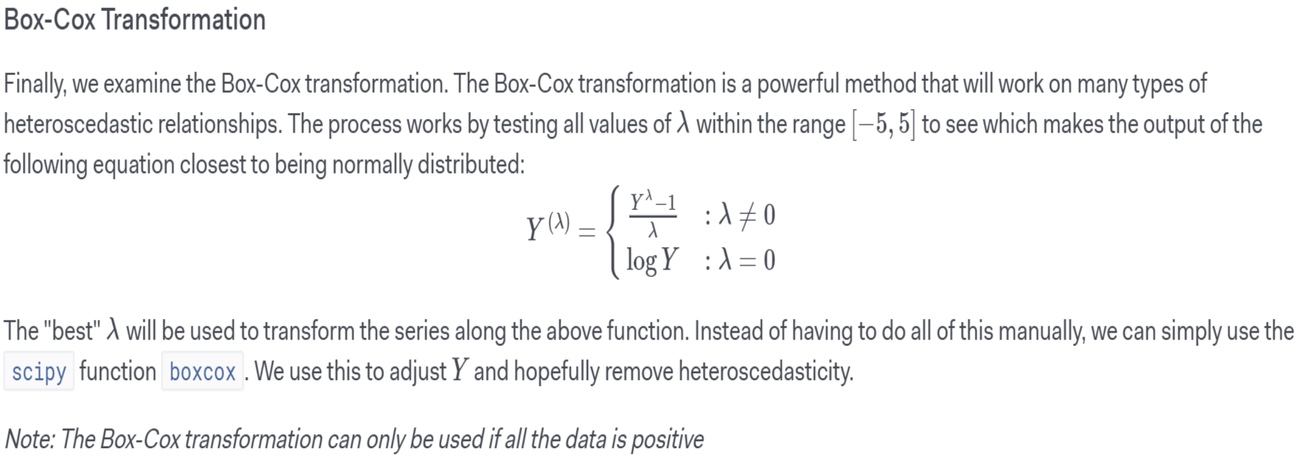
```
for key, value in result[4].items():
        print('\t%s: %.3f' % (key, value))
```

Diagnosis of Residuals and Transformations :



What we want is a fairly random distribution of residuals. The points should form no discernible pattern. This would indicate that a plain linear model is likely a good fit. If we see any sort of trend, this might indicate the presence of autocorrelation or heteroscedasticity in the model.

Further to cross check it in depth we can differentiate or can take log values and then check for heteroscedastic.



Box-Cox Transformation

Finally, we examine the Box-Cox transformation. The Box-Cox transformation is a powerful method that will work on many types of heteroscedastic relationships. The process works by testing all values of λ within the range $[-5, 5]$ to see which makes the output of the following equation closest to being normally distributed:

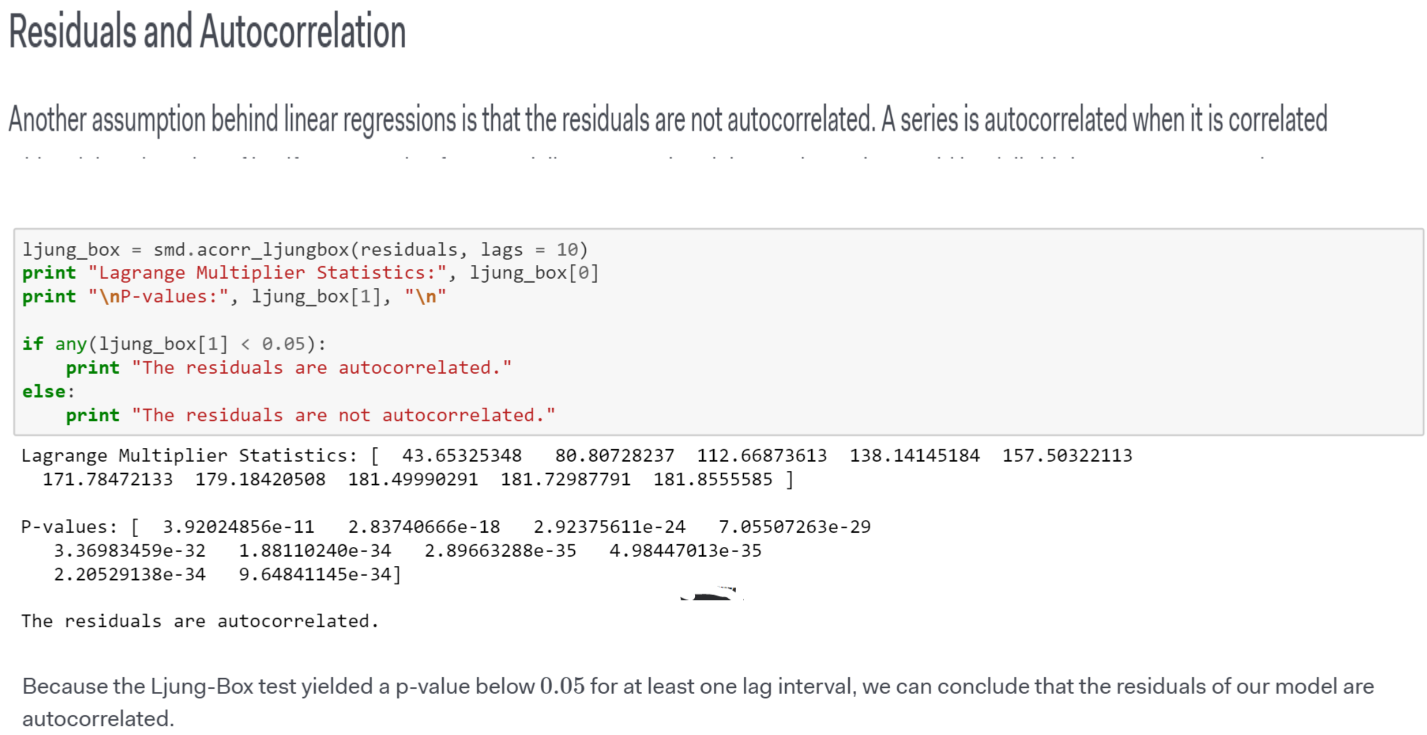
$$Y^{(\lambda)} = \begin{cases} \frac{Y^\lambda - 1}{\lambda} & : \lambda \neq 0 \\ \log Y & : \lambda = 0 \end{cases}$$

The "best" λ will be used to transform the series along the above function. Instead of having to do all of this manually, we can simply use the `scipy` function `boxcox`. We use this to adjust Y and hopefully remove heteroscedasticity.

Note: The Box-Cox transformation can only be used if all the data is positive

@ Shubham Kotal

Quantopian:



Residuals and Autocorrelation

Another assumption behind linear regressions is that the residuals are not autocorrelated. A series is autocorrelated when it is correlated

```
ljung_box = smd.acorr_ljungbox(residuals, lags = 10)
print "Lagrange Multiplier Statistics:", ljung_box[0]
print "\nP-values:", ljung_box[1], "\n"

if any(ljung_box[1] < 0.05):
    print "The residuals are autocorrelated."
else:
    print "The residuals are not autocorrelated."
```

```
Lagrange Multiplier Statistics: [  43.65325348   80.80728237  112.66873613  138.14145184  157.50322113
  171.78472133  179.18420508  181.49990291  181.72987791  181.8555585 ]

P-values: [  3.92024856e-11   2.83740666e-18   2.92375611e-24   7.05507263e-29
   3.36983459e-32   1.88110240e-34   2.89663288e-35   4.98447013e-35
   2.20529138e-34   9.64841145e-34]

The residuals are autocorrelated.
```

Because the Ljung-Box test yielded a p-value below 0.05 for at least one lag interval, we can conclude that the residuals of our model are autocorrelated.

L-jung-Box test

acorr_ljungbox function in statsmodels to test for autocorrelation in the residuals of our above model. We use a max lag interval of 10, and see if any of the lags have significant autocorrelation in the above code.

Reasons for overfitting:

1. small sample size, so that noise and trend are not distinguishable
2. choosing an overly complex model, so that it ends up contorting to fit the noise in the sample

Best way to avoid overfitting:

1. Information criteria
2. Cross validation
3. Taking large sample data

p-Hacking

p-hacking is just intentional or accidental abuse of multiple comparisons bias.
In general, the concept is simple. By running many tests or experiments and then focusing only on the ones that worked, you can present false positives as real results. Keep in mind that this also applies to running many different models or different types of experiments and on different data sets.

@ Shubham Kotal

Quantopian:

Doing correlation tests and choosing a significant level and deciding two sets are correlated or and its significant or not.

Spearman Rank Correlation

It's a variation of correlation that takes into account the ranks of the data. This can help with weird distributions or outliers that would confuse other measures. The test also returns a p-value, which is key here.

```
X = pd.Series(np.random.normal(0, 1, 100))
Y = X
r_s = stats.spearmanr(Y, X)
print 'Spearman Rank Coefficient: ', r_s[0]
print 'p-value: ', r_s[1]
```

o/p: Spearman Rank Coefficient: 1.0 ,p-value: 0.0

What is leverage?

Leverage Ratio= (Debt + Capital Base) / Capital Base

Leverage is borrowing money, then investing that money into some trading strategy so as to effectively multiply your initial capital base by some amount.

Leverage is reinvesting debt (sum of profit) to gain a greater return on an investment.

Compare Strategies by Sharpe Ratio and then Lever as Needed

NOTE: Remember that in finance, volatility is measured by the standard deviation of a time series, and the amount of future risk of a portfolio is estimated by past portfolio volatility. Be invested in as many uncorrelated assets as possible. In finance this is known as diversification. If you have a pricing model, price everything and invest accordingly.

What is Slippage?

Slippage occurs when large buy or sell orders drive the price of the market up or down respectively.

BUY 100, 50 for 1000 and 50 for 1100

Liquidity vs. Volume

Liquidity and volume are not the same thing. Liquidity is the important property that affects how easily we can trade. The quick cash out.

@ Shubham Kotal

Quantopian:

Transaction costs fall into two categories

- Direct (commissions and fees): explicit, easily measured, and in institutional trading, relatively "small"
- Indirect (market impact and spread costs): **slippage**

Slippage is when the price 'slips' before the trade is fully executed, leading to the fill price being different from the price at the time of the order. The attributes of a trade that our research shows have the most influence on slippage are:

Indirect:

1. **Volatility (**rapid change**)**
   Volatility is a statistical measure of dispersion of returns for a security. Calculated as the standard deviation of returns. The volatility of any given stock typically peaks at the open and thereafter decreases until mid-day.The higher the volatility the more uncertainty in the returns. This uncertainty is an artifact of larger bid-ask spreads during the price discovery process at the start of the trading day.
2. **Liquidity:** In general, liquidity is highest as we approach the close, and second highest at the open. Mid day has the lowest liquidity. Liquidity should also be viewed relative to your order size and other securities in the same sector and class.
3. **Relative order size :** increase relative order size at a specified participation rate, the time to complete the order increases.
4. **Bid - ask spread:** various relationships between bid-ask spread and order attributes are seen in our live trading data.

How do institutional quant trading teams evaluate transaction cost ?

Find out more here:
https://github.com/quantopian/research_public/blob/master/notebooks/lectures/Market_Impact_Model/notebook.ipynb

Quantifying Market Impact through readily available models.

1. **JPM** explicitly calls out spread impact
2. **Almgren** considers fraction of outstanding shares traded daily
3. **Q Slippage** Model does not consider volatility
4. **Kissel e**xplicit parameter to proportion temporary and permanent impact, to name a few.

Universe selection

On a high level, universe selection is the process of choosing the pool of securities upon which your algorithm will trade. Done via using pipeline and dividing equities to categories ex finance,it, commodities. And then cross validation by looking after overall turnover.

https://www.quantopian.com/lectures/universe-selection#notebook

@ Shubham Kotal

Quantopian:

Capital Asset Pricing Model

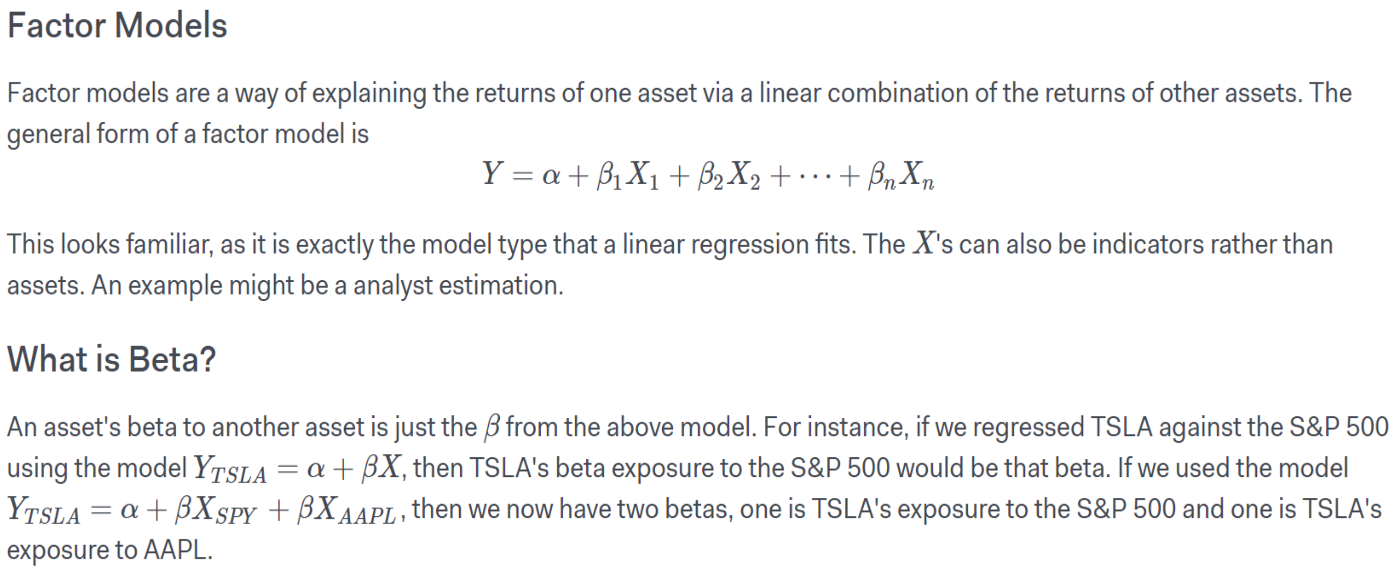
The Capital Asset Pricing Model (CAPM) is a classic measure of the cost of capital. It is used often in finance to evaluate the price of assets and to assess the impact of the risk premium from the market at large.

But it has few limitations.

Arbitrage Pricing Theory

Arbitrage is the simultaneous buying and selling of securities, currency, or commodities in different markets or in derivative forms in order to take advantage of differing prices for the same asset.

APT is a major asset pricing theory that relies on expressing the returns using a linear factor model.



Factor Models

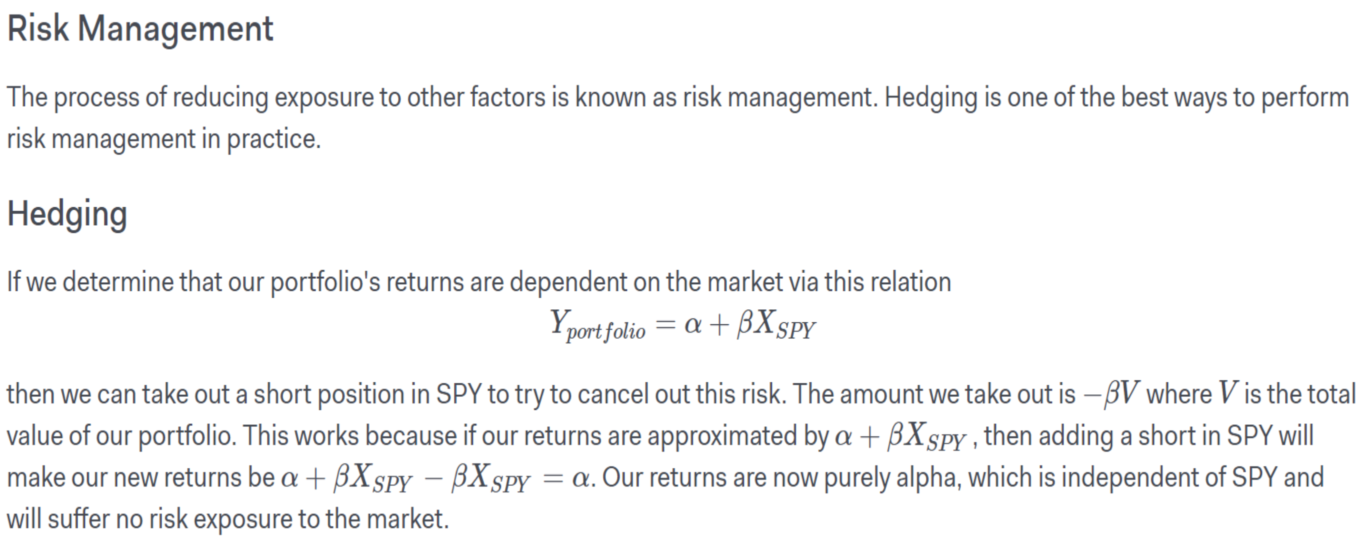
Factor models are a way of explaining the returns of one asset via a linear combination of the returns of other assets. The general form of a factor model is

$$Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

This looks familiar, as it is exactly the model type that a linear regression fits. The X's can also be indicators rather than assets. An example might be a analyst estimation.

What is Beta?

An asset's beta to another asset is just the β from the above model. For instance, if we regressed TSLA against the S&P 500 using the model $Y_{TSLA} = \alpha + \beta X$, then TSLA's beta exposure to the S&P 500 would be that beta. If we used the model $Y_{TSLA} = \alpha + \beta X_{SPY} + \beta X_{AAPL}$, then we now have two betas, one is TSLA's exposure to the S&P 500 and one is TSLA's exposure to AAPL.



Risk Management

The process of reducing exposure to other factors is known as risk management. Hedging is one of the best ways to perform risk management in practice.

Hedging

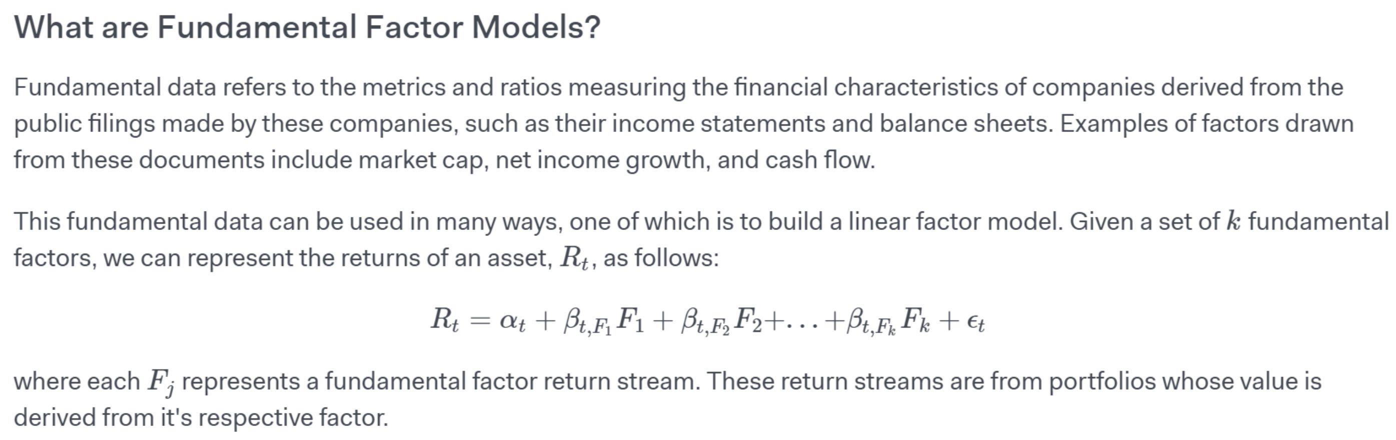
If we determine that our portfolio's returns are dependent on the market via this relation

$$Y_{portfolio} = \alpha + \beta X_{SPY}$$

then we can take out a short position in SPY to try to cancel out this risk. The amount we take out is $-\beta V$ where V is the total value of our portfolio. This works because if our returns are approximated by $\alpha + \beta X_{SPY}$, then adding a short in SPY will make our new returns be $\alpha + \beta X_{SPY} - \beta X_{SPY} = \alpha$. Our returns are now purely alpha, which is independent of SPY and will suffer no risk exposure to the market.

@ Shubham Kotal

Quantopian:



What are Fundamental Factor Models?

Fundamental data refers to the metrics and ratios measuring the financial characteristics of companies derived from the public filings made by these companies, such as their income statements and balance sheets. Examples of factors drawn from these documents include market cap, net income growth, and cash flow.

This fundamental data can be used in many ways, one of which is to build a linear factor model. Given a set of k fundamental factors, we can represent the returns of an asset, R_t, as follows:

$$R_t = \alpha_t + \beta_{t,F_1} F_1 + \beta_{t,F_2} F_2 + \ldots + \beta_{t,F_k} F_k + \epsilon_t$$

where each F_j represents a fundamental factor return stream. These return streams are from portfolios whose value is derived from it's respective factor.

Portfolio analysis using
# pyfolio
It offers many options to plot the portfolio as per varying measuring functions
Plot_rolling_sharpe, plot_turnover, etc…

PCA:
Principal components allow us to quantify the variability of the data, leading to low-dimensional projections of matrices that contain the bulk of the information contained within the original dataset.

Long/short equity

It is an investment **strategy** generally associated with hedge funds, and more recently certain progressive traditional asset managers. It involves buying **equities** that are expected to increase in value and selling **short equities** that are expected to decrease in value.
It uses a ranking system to understand the increasing and decreasing equities.

Alphalens

Building portfolios of alpha factors allows us to more carefully monitor and analyze the source and consistency of our returns. In this lecture we cover the basics of determining whether an alpha factor is suitable for a long-short equity algorithm by analyzing it using Alphalens, an open source package that we developed specifically for this purpose. We discuss the various graphs and statistics that compose an Alphalens tear sheet and provide background on how they indicate factor quality.

Quantopian:

Value at Risk (VaR) is a key concept in portfolio risk management. It uses the past observed distribution of portfolio returns to estimate what your future losses might be at different likelihood levels.

@ **Shubham Kotal**