SESSION 5: DATA RELATIONSHIPS

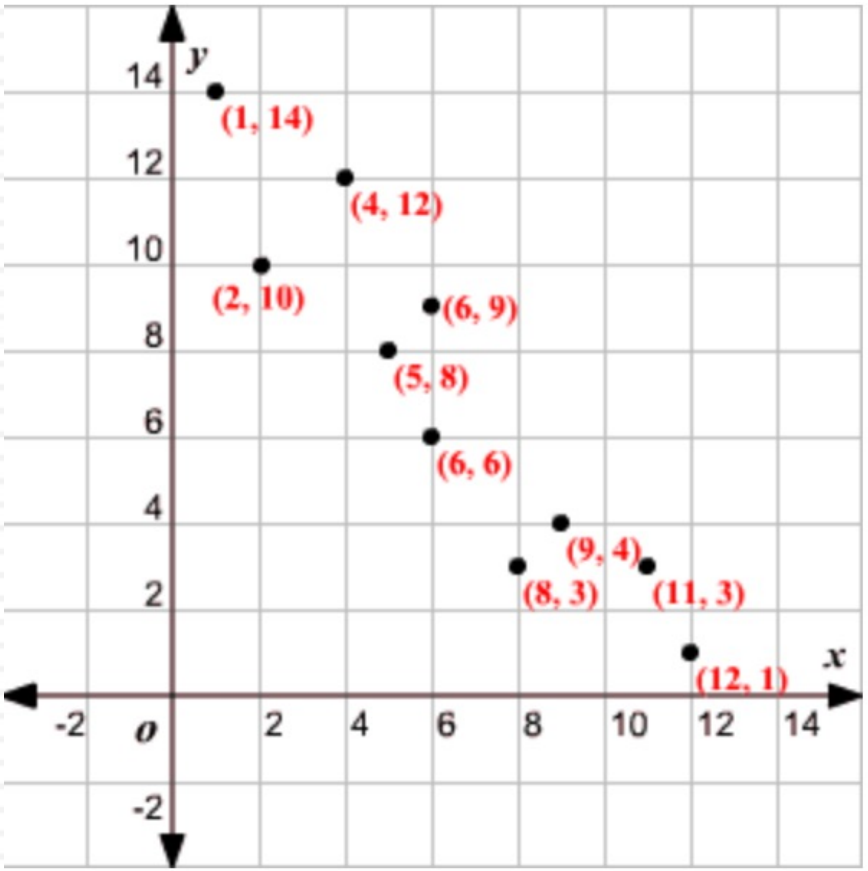
Session 5

Correlations, Covariances and Regressions

Exploring data linkages/relationships

- When you have two or more data series, you can check for linkages between the data, i.e., whether the data move together (positive co-movement), move inversely (negative co-movement) or are unrelated (no co-movement).
- If there is a linkage, you can explore further to see if
  - Time lags and leads: Changes in one data variable lead changes in the other
  - Correlation vs Causation: Changes in one data variable are ”causing” changes in the other
  - Prediction: You can predict one variable, using the other variable.

A Scatter Plot



Correlation Coefficients

- The *correlation coefficient* measures how two variables move together. The most widely used one is Pearson's correlation coefficient:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[\, n\Sigma x^2 - (\Sigma x)^2 \,][\, n\Sigma y^2 - (\Sigma y)^2 \,]}}$$

- The correlation coefficient lies *between -1 and +1*, with positive (negative) values indicating that the variables move together (inversely).
  - A correlation of zero indicates no co-movement.
  - A correlation coefficient of 1 (plus or minus) indicates perfect co-movement

And Covariances..

- The covariance, like correlation, is a measure of how two variables move together.

$$cov_{x,y} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N - 1}$$

- Like the correlation, a positive covariance indicates that two variables move together, a zero covariance that there is no relationship between the two variables, and a negative covariance an indication that they move in opposite directions.
- Unlike the correlation, the covariance is not bounded between 1 and -1, and will reflect the variable values.

Parting propositions..

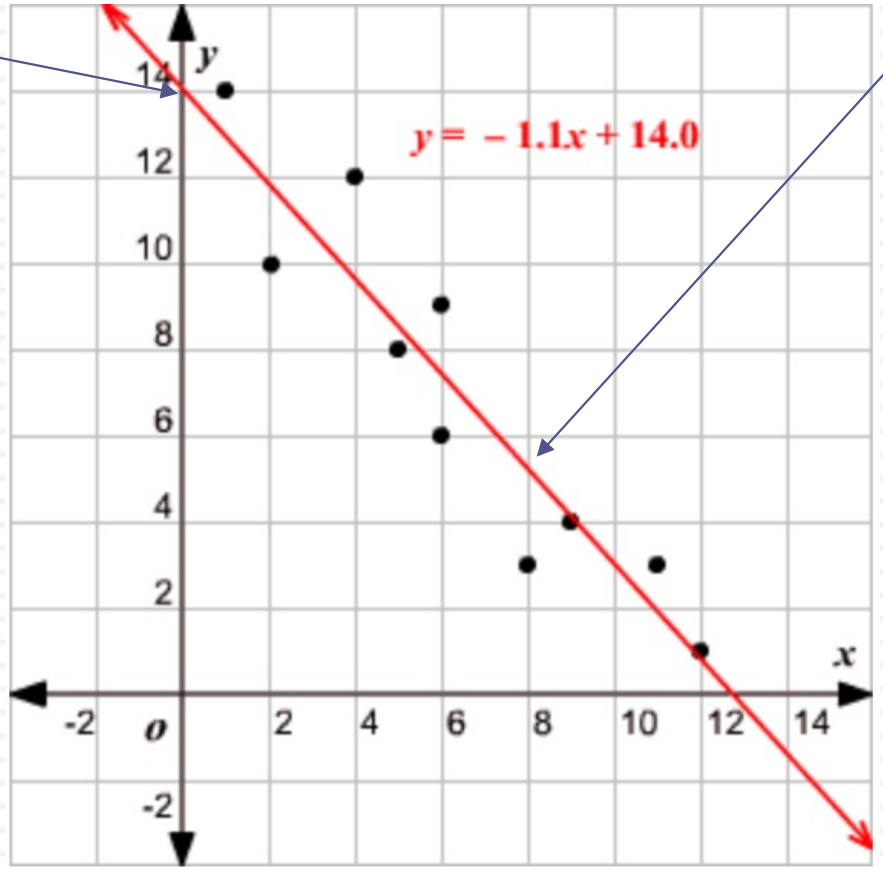
- Correlations can be spurious: If two variables are correlated, we are often tempted to create elaborate explanations for why. In many cases, that correlation can be spurious, with the two variables that are correlated both being driven by a third and often unseen variable.
- Correlation ≠ Causation: Correlation is a statistical measure of how two variables move together. To get from correlation to causation, you need to do more work (including a hypothesis of why, and an experiment designed to show that causation).
- Past versus Future: The data that we use to estimate correlation come from the past, and past correlation is not always a predictor of future correlation.

A Best Fit Line

The Intercept: In the scatter plot, the intercept is where the best-fit line crosses the Y axis. In simple terms, it is the value that the Y variable has when the X variable is zero.

The R squared: This measures the how well the line fits the data. If the line is a perfect fit (every point is on it), the R squared will be one (as will the correlation).

The Best Fit Line: In an ordinary-least-squares (OLS) regression, the best fit line is the one that minimizes the squared distances from the line.

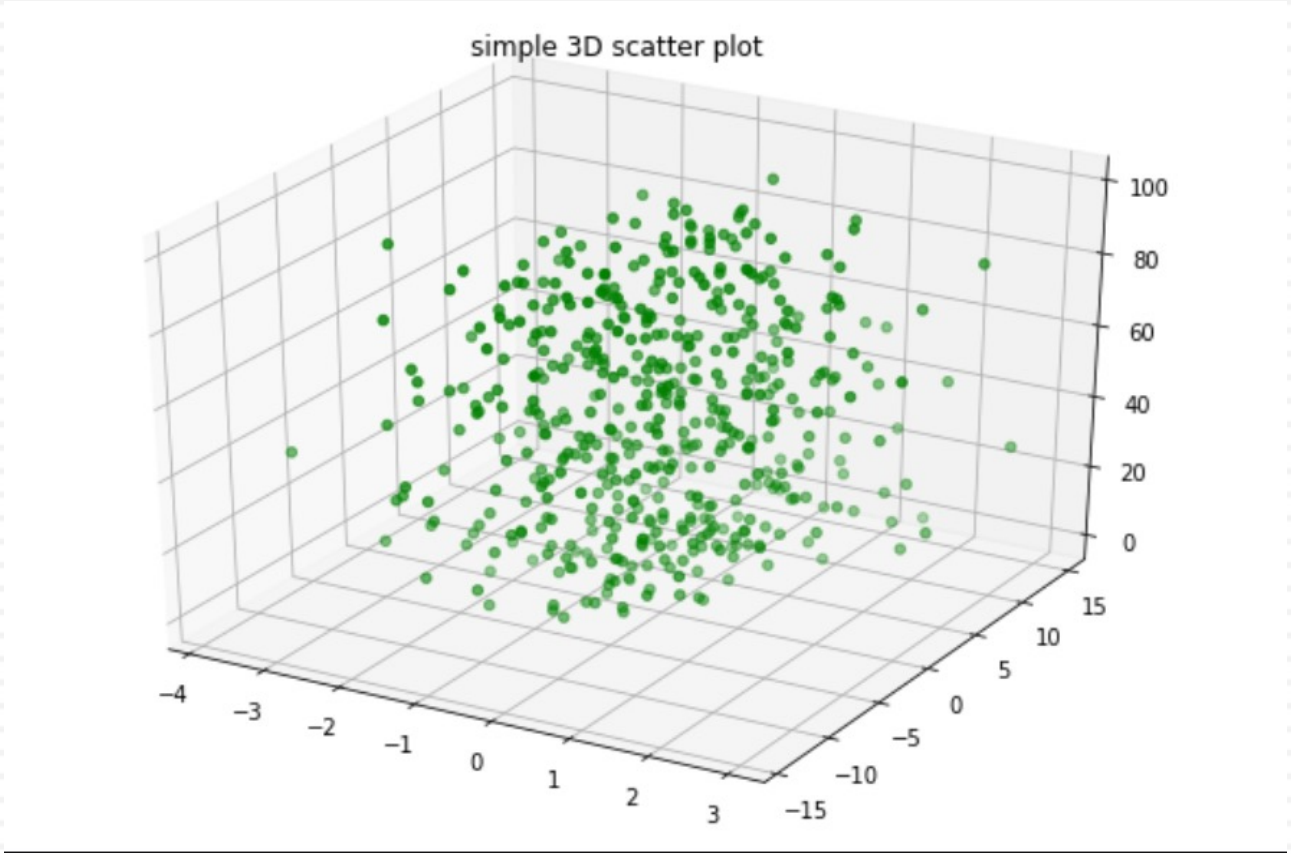


The Slope: The slope of the best fit line is also a measure of how changes in the X variable show up as changes in the Y variable. When there is no (positive, negative) relationship between the variables, the slope will be zero (positive, negative).

More than two variables?

- While a simple regression has one dependent variable and one independent variable, there are times where you may want to look at more than one independent variable in trying to explain a dependent variable.
- You can extend the tools that you use for two variables to multiple variables, by:
  - *Three dimensional plots*, if you have two independent variables, or multiple scatter plots, if you have more than two.
  - *Correlation matrices* that look at correlations between all possible pairs of variables.
  - *Multiple regressions*, where you have one dependent variable and many independent variables.

A 3-dimensional Scatter Plot…



Independent Variables Plus

- Dummy Variables: A dummy variable takes on a value of either zero or one and can be used to capture the effect of grouping in a multiple regression.
  - Discrete Variable: If the independent variable is discrete, you may have no choice but to use a dummy variable to capture its effects. (*If you have both private and publicly traded companies in a group that you are studying, and you believe that they behave differently, you could break them down on that basis).*
  - Continuous Variable: Even if a variable is continuous, there may be an advantage to converting it into a discrete dummy variable. (*Default risk is a continuous variable, but you could break companies down into investment and non-investment grade, based upon ratings.)*
- Joint Effects: If you believe that your dependent variable is affected by interactions between independent variables, you can capture the cross effects by looking at products of the independent variables. *Thus, if you are regressing PE ratios against market cap and expected growth rates, you can do the following:*

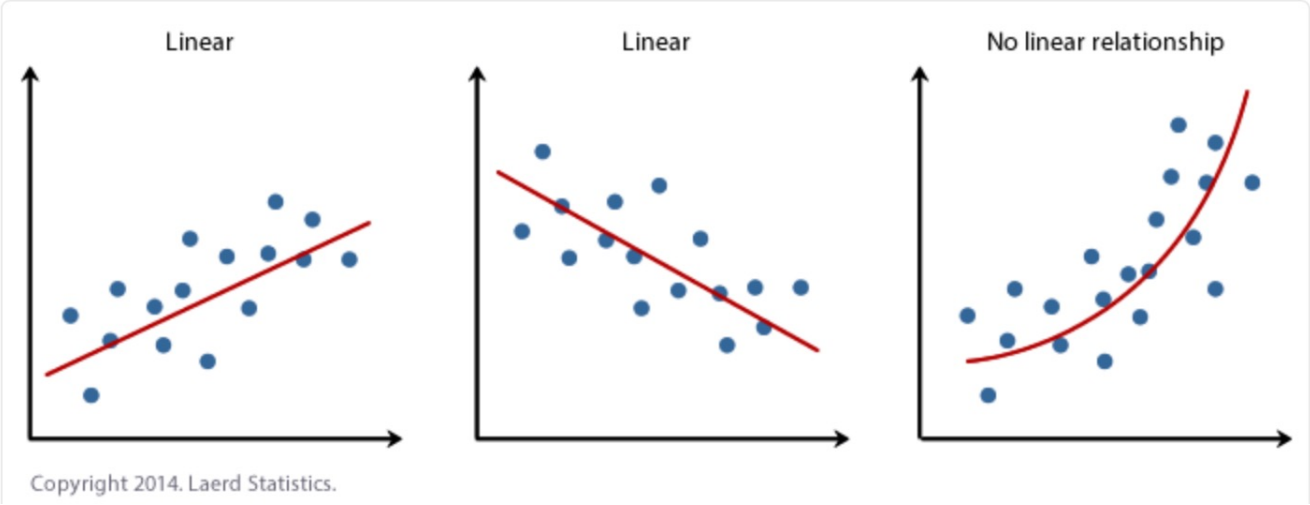
  *PE = a + b Growth Rate + c Mkt Cap + d (Growth Rate * Mkt Cap)*

Regression Diagnostics

- *Non-linearity*: In a regression, we are assuming that the relationship between the dependent and independent variables is linear. If it is not:
  - You can run a non-linear regression
  - or use mathematical transformation (square, natural log etc.) on either the dependent or independent variables to create a more linear relationship.
- *Multicollinearity*: In a multiple regression, the independent variables should, in a perfect world, be uncorrelated with each other. But we don't live in a perfect world…
- *Homoskedasticity*: The residuals (or prediction errors) should not reveal any patterns (get larger or smaller), as the independent variable(s) increase and decrease.
- *Normality*: The residuals should be normally distributed.

Linear versus Non-linear Regressions

- The easiest way to detect multicollinearity is to run scatter plots of the dependent variable against each independent variable.

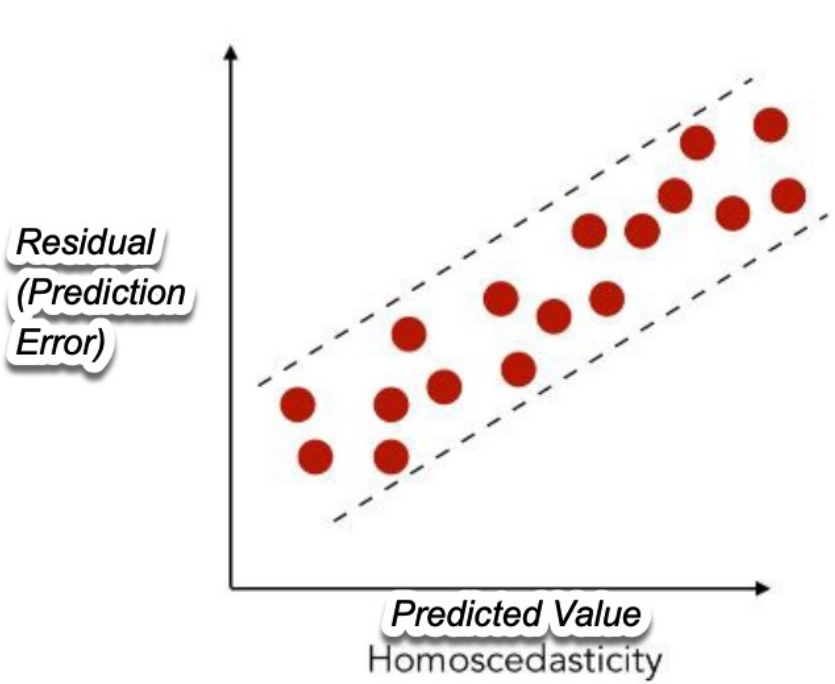


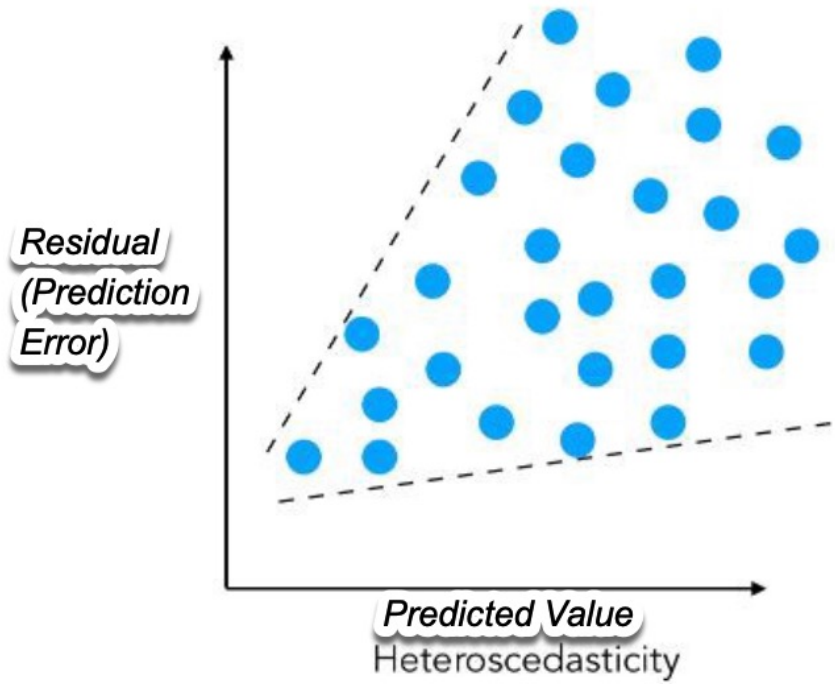
- A non-linear regression tries to find a best fit line through the non-linear plot. The process of fitting a line, though, is more complicated, and often involves trial and error (in the background, if you are using a statistics package).
- If you choose to transform a variable (dependent or independent), you may have to use trial and error (using the scatter plot) to find your best transformation.

Collinearity Diagnostics

- Multicollinearity can cause coefficient estimates on the independent variables to become unreliable and volatile.
  - It reduces the precision of these coefficients and the statistical power of the regression.
  - It generally does not influence the predictions or precision of the predictions.
- The multicollinearity in a multiple regression can be measured with a *variance inflation factor* (VIF).
  - A VIF greater than 5 indicates severe multicollinearity
  - A VIF between 1 and 5 indicates moderate multicollinearity
- If there is multicollinearity, you can
  - Try *replacing or removing one of a pair of independent variables* that are correlated.
  - Run *individual simple regressions* against the independent variables
  - Do nothing, *use the multiple regression for predictions*, and don't over read the coefficients on variables.

Prediction Errors from Regressions: Homoskedasticity and Heteroskedasticity…





Regression Variants

- Generalized Least Squares (GLS): When the residuals from your OLS regression are heteroskedastic or autocorrelated, one alternative that will yield more unbiased estimates is a *Generalized Least Squares* (GLS) regression.
  - To run a GLS regression, you usually begin with an OLS regression and use the information in the prediction error terms to build a GLS regression.
  - The output from a GLS regression resembles that from an OLS regression and can be used similarly to make predictions and test hypotheses.
- Weighted Least Squares: In a weighted least squares regression, not all observations in a sample are weighted equally. Instead, a weighting mechanism is created that weights some observations more than others.
  - In some cases, the weighting may reflect your belief that some of your observations convey more information about the population than others.
  - In others, it can be to counter the problem of standard errors of predictions varying across different values for the explanatory variable.

Regression Predictions

- Precision: In almost every regression, your predictions from a regression will be imprecise, with a standard error and a range around the prediction. If the R-squared is low, and the range is large, a regression is less actionable.
- Stationarity: Regressions are run with observed data from the past. To the extent that the process that you are observing is non-stationary, with changes in population and structure, the predictions will not hold up in future periods.
- Observability: In a regression, you need to be able to observe the dependent variable (or variables) to be able to generate predictions of the independent variables. Regressions that use independent variables that don't lend themselves to timely observation cannot be acted upon.
- Actionability: If the predicted value for Y from a regression deviates from the actual value, being able to act on the divergence makes the reliable more useful (in many cases).