

MBC 638

LIVE SESSION WEEK 9

Agenda

| Topic | Time | Thursday Section | Sunday Section |
|--|--------|------------------|----------------|
| Introduction | 5 min | 9:00 - 9:05 | 6:30 - 6:35 |
| Highlights from Week 9 Video | 30 min | 9:05 - 9:35 | 6:35 - 7:05 |
| Start on Final Review | 25 min | 9:35 - 10:00 | 7:05 - 7:30 |
| Breakout on Regression | 20 min | 10:00 - 10:20 | 7:30 - 7:50 |
| Review of Upcoming Assignments and Open Question | 10 min | 10:20 - 10:30 | 7:50 - 8:00 |

9.2 Introduction to Time Series

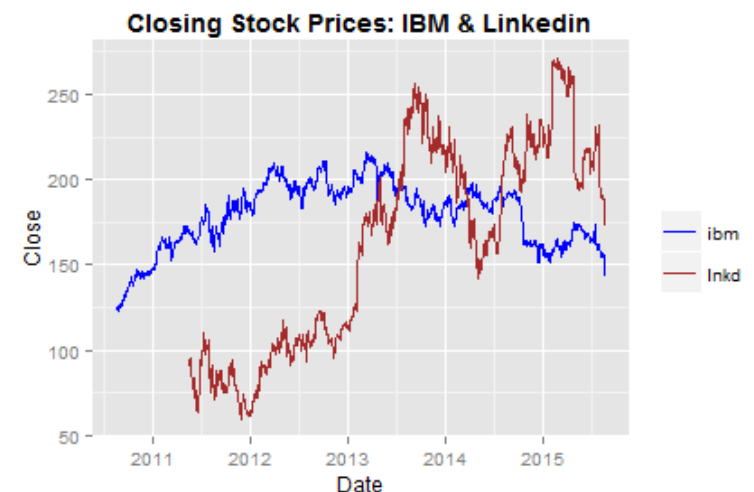
Time series analysis comprises methods for analyzing **time series** data in order to extract meaningful statistics and other characteristics of the data.

Data is collected at regular intervals over a given time period.

Time series forecasting is the use of a model to predict future values based on previously observed values. (a bit like looking in the rear view mirror to predict the future)

Time = input, Time Series = output

$Y = f(y)$, doesn't really consider x 's driving the output



9.2 Introduction to Time Series

Potential Components of Variation

1. Trend
 - Long-term rise and fall
2. Calendar cycles
 - Seasonality
3. Business cycles
 - Affected by American politics
4. Autoregressive behavior - A stochastic process used in statistical calculations in which future values are estimated based on a weighted sum of past values. **An autoregressive process operates under the premise that past values have an effect on current values.** A process considered AR(1) is the first order process, meaning that the current value is based on the immediately preceding value. An AR(2) process has the current value based on the previous two values.
<http://www.investopedia.com/terms/a/autoregressive.asp#ixzz3cTF1ZCQm>
5. Random variation

9.2 Introduction to Time Series

Time Series Analysis

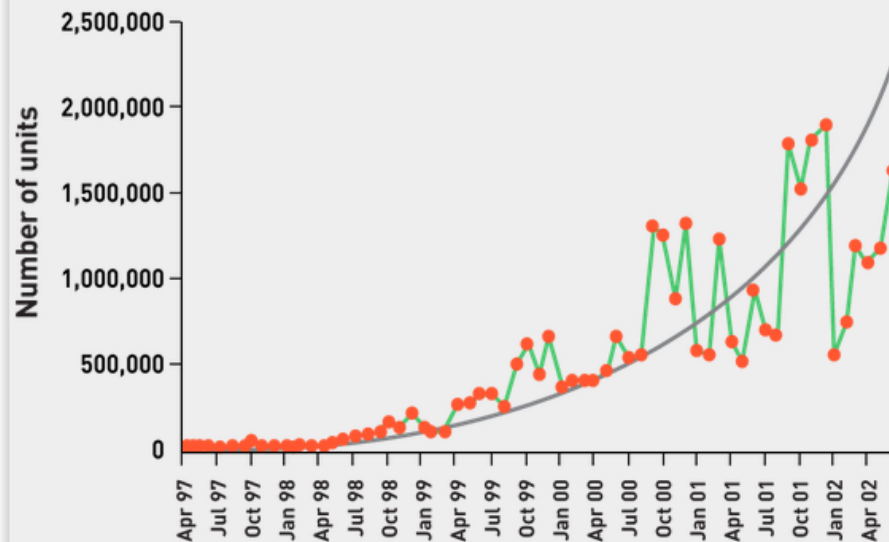
1. Time plot: tool to study/visualize time series
2. Model patterns: trends and seasonality
3. Forecast: predict future values of time series
4. Remember practical, graphical, analytical

9.2 Introduction to Time Series

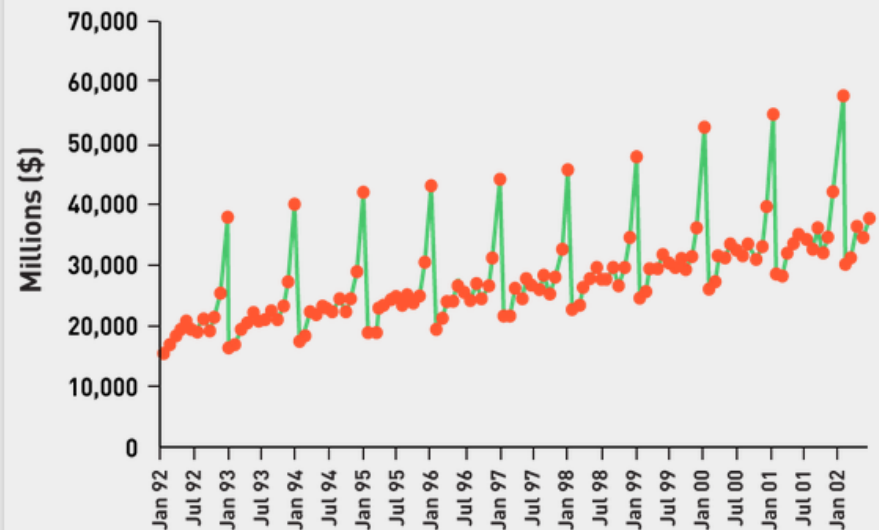
Looking for Systematic Patterns

- Trends
- Steady movement in a particular direction
- Seasonality (repeated pattern)

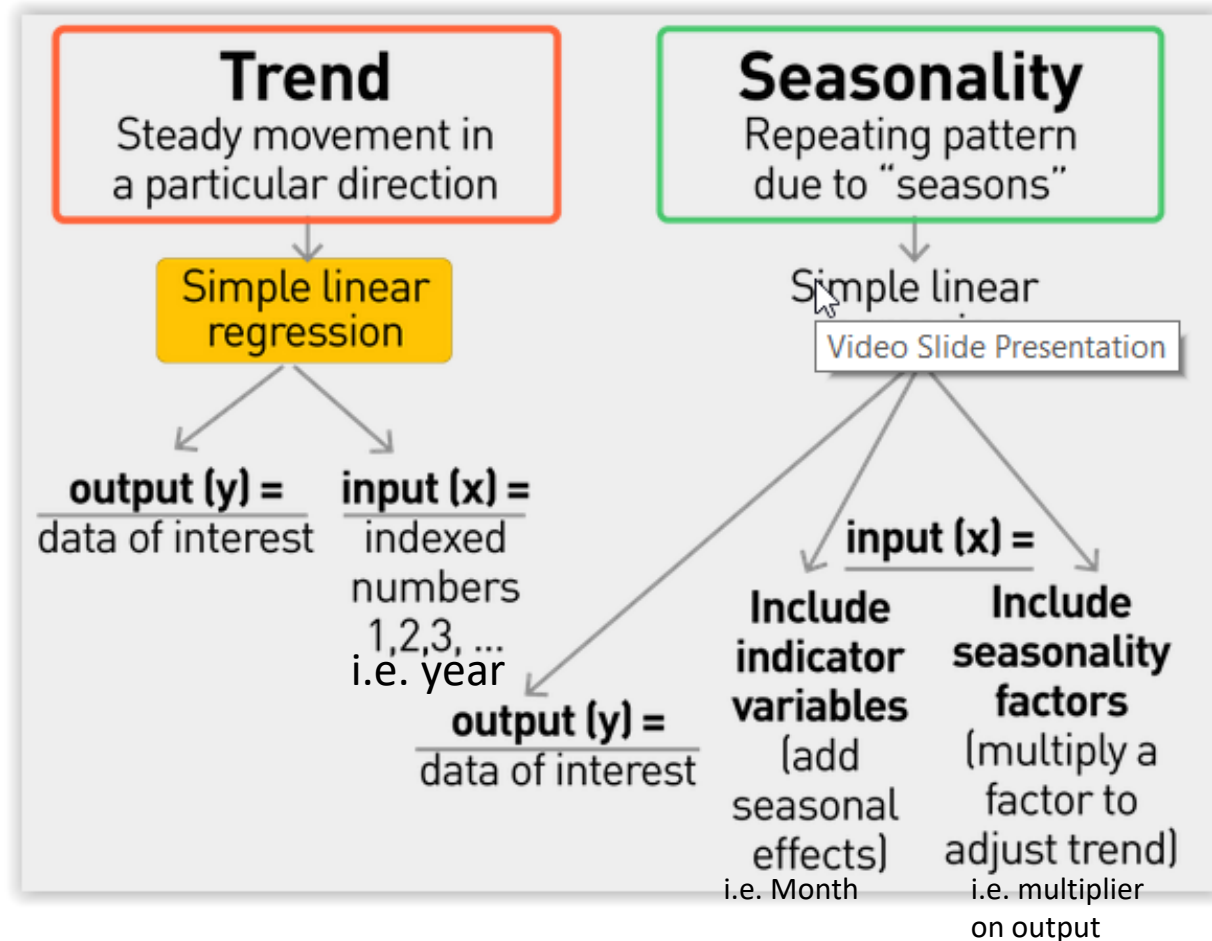
Exponential trend fitted to the number of DVD players sold



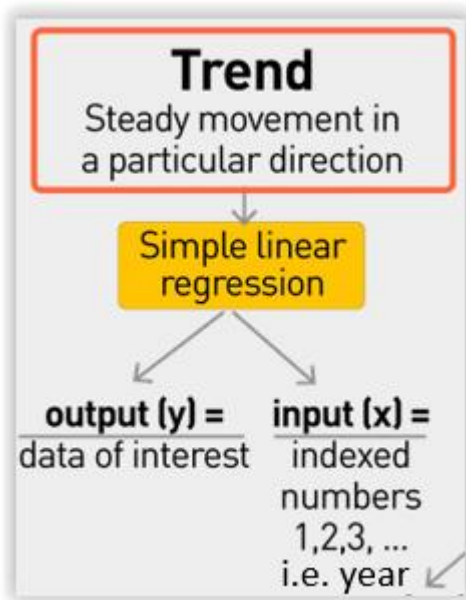
Time plot of U.S. retail sales of general merchandise stores



9.2 Introduction to Time Series



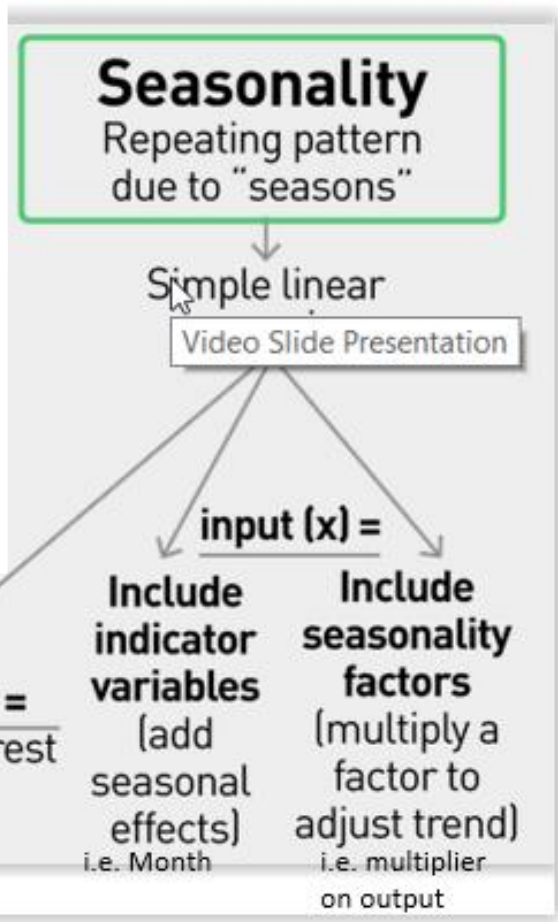
9.2 Introduction to Time Series



Trend

- Simple linear regression
 - Input (x) = indexed numbers, e.g., 1979, 1980, 1981
 - Output (y) = data of interest, e.g., budget information

9.2 Introduction to Time Series



Seasonality

- Simple linear regression
 - Input (x) =
 - Indicator variables
 - Seasonality factors
 - Output (y) = data of interest

Indicator Variables

- Months as indicator variable
 - Use $x_1 - x_{11}$ (K - 1)
- Trend + season model:
$$\hat{y} = \beta_0 + \beta_1 x + \beta_2 x_1 + \beta_3 x_2 + \dots + \beta_{12} x_{11}$$
 - Trend: $\beta_0 + \beta_1 x$
 - Seasonality: $\beta_2 x_1 + \beta_3 x_2 + \dots + \beta_{12} x_{11}$

Seasonality Factors

- Calculate adjustments, multiply by regression equation
- For each data point in time series, calculate ratio

$$\frac{\text{Actual } y}{\text{Predicted } y} = \text{Seasonality Factor (SF)}$$

- Average SF by month → 12 SFs
- Multiply regression equation by a given month's SF to account for seasonality in a trend model.

- Trend + season model: $\hat{y} = (\beta_0 + \beta_1 x) \times \text{SF}$

Video Slide Presentation

9.3 Autocorrelation

Regression on Time Series Data

Modeling trend and seasonal components may not generate random residuals

Residual plots help assess the fit of a regression line

Autocorrelation: Definition

- Relationships between neighboring points
 - E.g., January data affects February, February affects March, etc.
 - Can cause lack of randomness in data

- **Autocorrelation:** correlation between successive values

DEFINITION of 'Autocorrelation'

A mathematical representation of the degree of similarity between a given time series and a lagged version of itself over successive time intervals. It is the same as calculating the correlation between two different time series, except that the same time series is used twice - once in its original form and once lagged one
<http://www.investopedia.com/terms/a/autocorrelation.asp#ixzz3cU65w3O0>

3 Autocorrelation

Autocorrelation refers to the correlation of a time series with its own past and future values. Autocorrelation is also sometimes called “*lagged correlation*” or “*serial correlation*”, which refers to the correlation between members of a series of numbers arranged in time. Positive autocorrelation might be considered a specific form of “*persistence*”, a tendency for a system to remain in the same state from one observation to the next. For example, the likelihood of tomorrow being rainy is greater if today is rainy than if today is dry. Geophysical time series are
www.ltrr.arizona.edu/.../notes_3.pdf

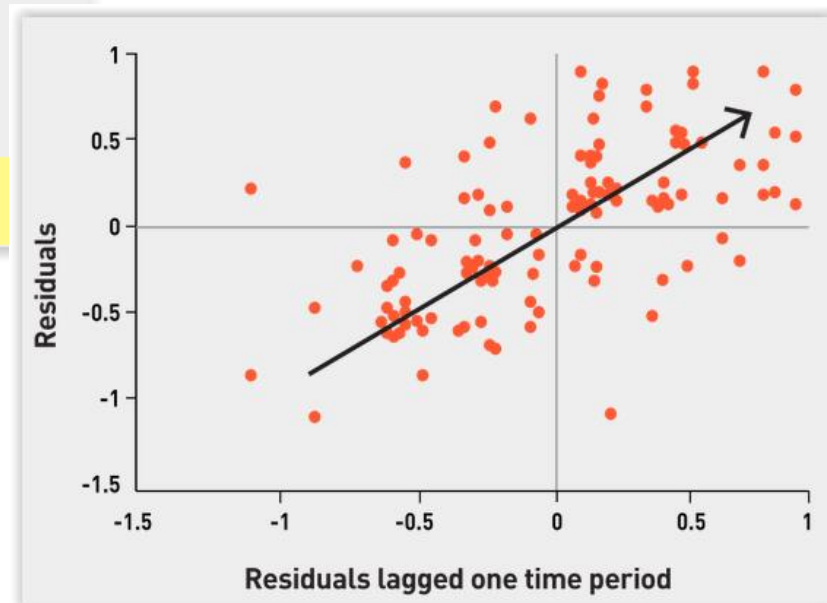
9.3 Autocorrelation

Autocorrelation: Why Do We Care?

- Can't use regression model if data violates assumption of independent residuals
- Autocorrelation in residuals indicates opportunity to improve fit
 - Add elements to model to increase predictive power

Autocorrelation: How Can We Tell?

- Test residuals by lagging, moving one time period
 - Residual = $e = y_{\text{actual}} - y_{\text{predicted}}$
 - Lagged residual plot = $(e_1, e_2), (e_2, e_3), (e_3, e_4) \dots (e_{n-1}, e_n)$
- Plot residuals, look for pattern



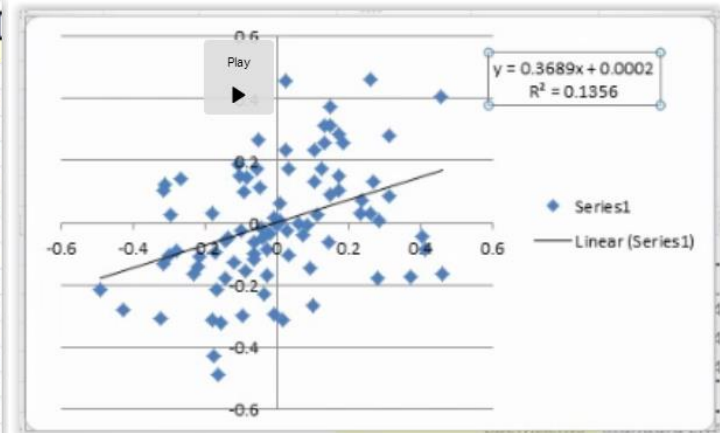
9.4 Is Autocorrelation Present?

Step 1: generate X by using lagged Y

| Great Lakes Water Level | | | |
|-------------------------|------|------------|---------------|
| | | output (y) | (x) |
| | Year | Lake Level | Lagged output |
| 1 | 1918 | 176.887 | |
| 2 | 1919 | 176.745 | 176.887 |
| 3 | 1920 | 176.625 | 176.745 |
| 4 | 1921 | 176.488 | 176.625 |
| 5 | 1922 | 176.445 | 176.488 |
| 6 | 1923 | 176.264 | 176.445 |
| 7 | 1924 | 176.187 | 176.264 |
| 8 | 1925 | 175.919 | 176.187 |
| 9 | 1926 | 175.885 | 175.919 |
| 10 | 1927 | 176.148 | 175.885 |

Step 3: Create a residual and lagged residual plot: Is there a pattern?
Y = residual, X=lagged residual

| x input | y output |
|---------------|-------------|
| lagged residu | Residual |
| | -0.06271085 |
| -0.06271085 | -0.06305948 |
| -0.06305948 | -0.09837441 |
| -0.09837441 | -0.02627936 |
| -0.02627936 | -0.17051346 |
| -0.17051346 | -0.09528201 |
| -0.09528201 | -0.29732568 |
| -0.29732568 | -0.10556241 |
| -0.10556241 | 0.18662801 |
| 0.18662801 | 0.25921724 |
| 0.25921724 | 0.46256087 |
| 0.46256087 | -0.16461966 |
| -0.16461966 | -0.49019318 |
| -0.49019318 | -0.21794482 |



Step 2: Run a regression on your X and Y above, skipping the first value

Regression

Input
Input Y Range:
Input X Range:
☐ Labels ☐ Constant is Zero
☐ Confidence Level: 95 %

Output options
☒ Output Range:
☐ New Worksheet Ply:
☐ New Workbook

Residuals
☒ Residuals ☐ Residual Plots

OK Cancel Help

| | Coefficients | Standard Error |
|--------------|--------------|----------------|
| Intercept | 27.4096104 | 10.19938281 |
| X Variable 1 | 0.844597862 | 0.057804166 |

| RESIDUAL OUTPUT | | |
|-----------------|-------------|--------------|
| Observation | Predicted Y | Residuals |
| 1 | 176.8077108 | -0.062710846 |
| 2 | 176.6880595 | -0.063059482 |
| 3 | 176.5684082 | -0.09837441 |
| 4 | 176.4487569 | -0.02627936 |
| 5 | 176.3291056 | -0.17051346 |
| 6 | 176.2094543 | -0.09528201 |
| 7 | 176.089803 | -0.29732568 |
| 8 | 175.9701517 | -0.10556241 |
| 9 | 175.8505004 | 0.18662801 |
| 10 | 175.7308491 | 0.25921724 |
| 11 | 175.6111978 | 0.46256087 |
| 12 | 175.4915465 | -0.16461966 |
| 13 | 175.3718952 | -0.49019318 |
| 14 | 175.2522439 | -0.21794482 |

Step 4: Conclusion R^2 is small, so not a strong relationship, so it is ok to use regression

9.5 Three Time Series Models

Time Series Models

1. First-order autoregressive model, a.k.a. AR(1)
2. Moving average forecast model
3. Exponential smoothing model

Autoregressive Model: AR(1)

- Takes advantage of linear relationship between successive values of time series
- **First-order autoregressive model**
 - Linear regression equation: $y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$
 - y_t = output at time t
 - Example: March; $t = 3$
 - $y_3 = \beta_0 + \beta_1 y_{3-1} + \varepsilon_3$
 - I.e., predicting March data by using February data

Video Slide Presentation

9.6 Forecast the Next Month: First Order Autoregressive Model

Forecast 2008:: using first order autoregressive model

Run Regression
plug 2008 input value into the equation to estimate y.

Assume we do have autocorrelation, assumption for regression is independence of errors this model, autoregressive, gives us permission to use regression anyways.

| Great Lakes Water Level | | | |
|-------------------------|------|------------|----------------------|
| | | output (y) | (this is y_{t-1}) |
| | Year | Lake Level | (x) Lagged output |
| 1 | 1918 | 176.887 | |
| 2 | 1919 | 176.745 | 176.887 |
| 3 | 1920 | 176.625 | 176.745 |
| 4 | 1921 | 176.488 | 176.625 |
| 5 | 1922 | 176.445 | 176.488 |
| 6 | 1923 | 176.264 | 176.445 |
| 7 | 1924 | 176.187 | 176.264 |
| 8 | 1925 | 175.919 | 176.187 |
| 9 | 1926 | 175.885 | 175.919 |
| 10 | 1927 | 176.148 | 175.885 |

| | | |
|---------------------------------|-----------|----------|
| Total | 88 | 12.42291 |
| Coefficients and Standard Error | | |
| Intercept | 27.40961 | 10.19938 |
| X Variable 1 | 0.8445979 | 0.057804 |

$y_t = \beta_0 + \beta_1 y_{t-1} + \varepsilon_t$
 $0.84459x + 27.4096 = y \text{ predicated} = \hat{y}$
 forecast 2008
 use 2007 data as the input for x
 175.943

| | | | |
|----|------|---------|---------|
| 85 | 2002 | 176.118 | 175.951 |
| 86 | 2003 | 175.892 | 176.118 |
| 87 | 2004 | 176.111 | 175.892 |
| 88 | 2005 | 176.090 | 176.111 |
| 89 | 2006 | 176.016 | 176.090 |
| 90 | 2007 | 175.943 | 176.016 |
| 91 | 2008 | | 175.943 |

$$= 0.84459 * (175.943) + 27.4096$$

| | |
|--------------------|------|
| 176.0092984 meters | 2008 |
|--------------------|------|

9.5 Three Time Series Models



Moving Average Model

A.k.a. "rolling average method"; smooths out short-term fluctuations

- Uses average of last several values of time series to forecast next value; k = number of values in span
 - Example: Monthly data, span (k) = 3
 - I.e., use average of values from January, February, and March to predict April value
- Can look back more than one time period
- Disadvantage: If, say, $n = 100$ and $k = 5$, forecast overlooks 95% of available data

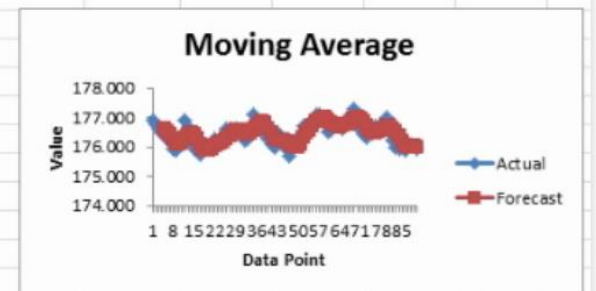
9.7 Forecast the Next Month: Moving Average Model

Moving Average

Input
 Input Range: 
☐ Labels in First Row
 Interval:
 Output options
 Output Range: 
 New Worksheet Ply:
 New Workbook
☐ Chart Output ☐ Standard Errors

OK Cancel Help

| Great Lakes Water Level | | | Forecast 2008:: using moving average model (k=5) | | | | |
|-------------------------|------|------------|--|-------------|--------------------------------|--|--|
| | Year | Lake Level | output (y) | k=5 | Data Analysis > Moving average | | |
| | | | | moving avg | | | |
| 1 | 1918 | 176.887 | | | | | |
| 2 | 1919 | 176.745 | | #N/A | | | |
| 3 | 1920 | 176.625 | | #N/A | | | |
| 4 | 1921 | 176.488 | | #N/A | | | |
| 5 | 1922 | 176.445 | | #N/A | | | |
| 6 | 1923 | 176.264 | | =AVERAGE(C5 | | | |
| 7 | 1924 | 176.187 | | 176.514 | | | |
| 8 | 1925 | 175.919 | | 176.402 | | | |
| 9 | 1926 | 175.885 | | 176.261 | | | |
| 10 | 1927 | 176.148 | | 176.140 | | | |
| 11 | 1928 | 176.443 | | 176.081 | | | |
| 12 | 1929 | 176.896 | | 176.117 | | | |



| | | | | |
|----|------|---------|-------------|---------|
| 84 | 2001 | 175.951 | | 176.514 |
| 85 | 2002 | 176.118 | | 176.373 |
| 86 | 2003 | 175.892 | | 176.200 |
| 87 | 2004 | 176.111 | | 176.035 |
| 88 | 2005 | 176.090 | | 176.010 |
| 89 | 2006 | 176.016 | | 176.032 |
| 90 | 2007 | 175.943 | | 176.045 |
| 91 | 2008 | | =AVERAGE(C9 | |

| | | | | |
|----|------|---------|--|---------|
| 90 | 2007 | 175.943 | | 176.045 |
| 91 | 2008 | | | 176.010 |

Smaller your K, the bumpier your forecast the closer to the actual data.

9.5 Three Time Series Models

Exponential Smoothing Model

- Best suited for forecasting time series without seasonal variation
- Unlike moving average model, data values not all weighted equally
- Forecasting equation: $\hat{y}_t = w y_{t-1} + (1 - w) \hat{y}_{t-1}$
 - \hat{y}_t = estimate of y at time period t
 - Example: t = February
 - w = smoothing constant
 - Your choice, pick any number between 0 and 1
 - Example: $w = 80\%$

Exponential Smoothing Model: Example

- Formula: $\hat{y}_t = w y_{t-1} + (1 - w) \hat{y}_{t-1}$
- February forecast = \hat{y}_t
- 80% of January value = $w y_{t-1} = 0.8 y_{t-1}$
- 20% of January forecast = $(1 - w) \hat{y}_{t-1}$
 $1 - 0.8)$
 \hat{y}_{t-1}
 0.2
 \hat{y}_{t-1}
- January forecast = $\hat{y}_{t-1} = \hat{y}_{2-1}$
 - January forecast incorporates data from December and prior months
- Final equation: $\hat{y}_2 = 0.8 y_{2-1} + (1 - 0.8) \hat{y}_{2-1}$

Video Slide Presenta

Exponential Smoothing Notes

- The smaller the w , the greater its smoothing effect
 - Smoothing constant always between 0 and 1
 - Larger smoothing constant → more fluctuation → closer to actual data
- Excel uses "damping constant": $(1 - w)$

Forecast Feb = 80% of Actual January + 20% of Forecasted January

Here $w=80\%$, if it were lower than you are using less of actual January

9.8 Forecast the Next Month: Exponential Smoothing

Damping Factor: $1 - W$, so if damping factor is lower, more emphasis on current data, less smooth

| Great Lakes Water Level | | | |
|-------------------------|------|--------------------------|--------------------------|
| | Year | output (y) Lake Level | smoother dampfact=0.9 |
| 1 | 1918 | 176.887 | #N/A |
| 2 | 1919 | 176.745 | 176.887 |
| 3 | 1920 | 176.625 | 176.8725 |
| 4 | 1921 | 176.488 | 176.84775 |
| 5 | 1922 | 176.445 | 176.8118083 |
| 6 | 1923 | 176.264 | 176.7751275 |
| 7 | 1924 | 176.187 | 176.7260314 |

Exponential Smoothing

Input

Input Range:

Damping factor:

☐ Labels

Output options

Output Range:

New Worksheet Ply:

New Workbook:

☒ Chart Output

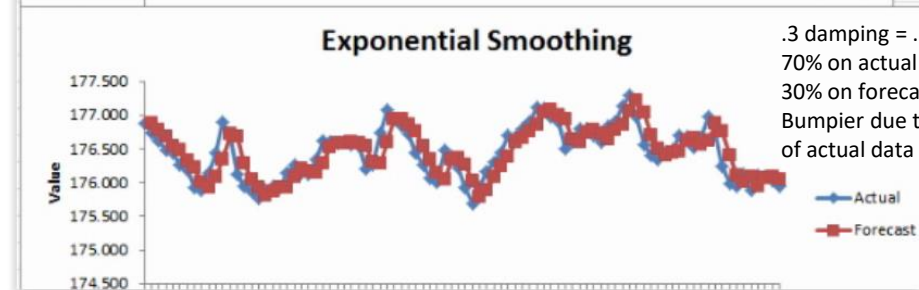
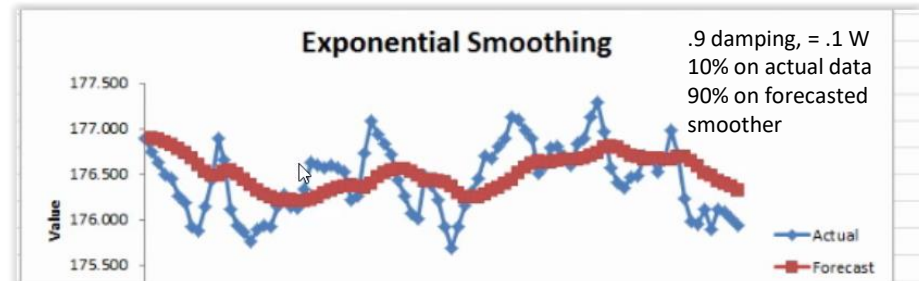
☐ Standard Errors

OK

Cancel

Help

| | | | |
|----|------|---------|-------------|
| 20 | 1937 | 175.923 | 176.2436688 |
| 21 | 1938 | 176.141 | 176.2115519 |



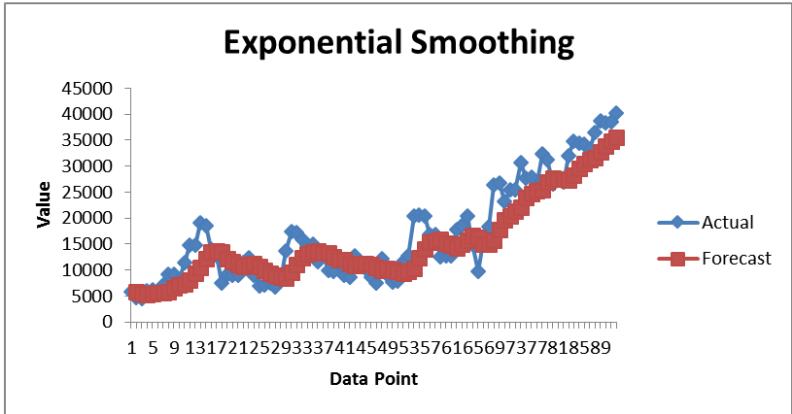
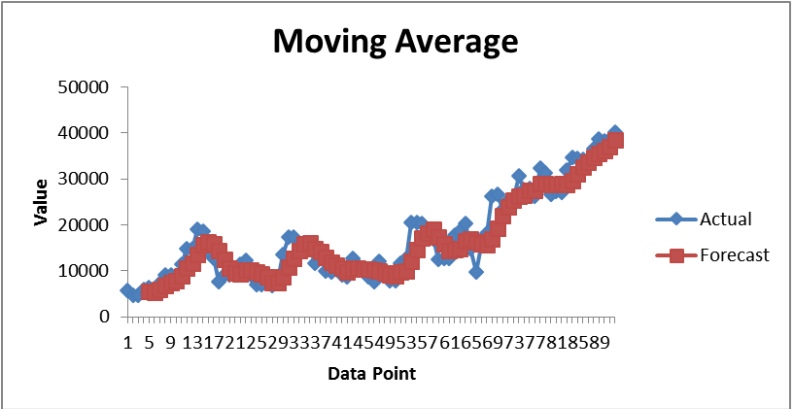
| | | | | | | |
|----|------|---------|-----------------------------|-------------|---------------------------------|--|
| 90 | 2007 | 175.943 | 176.3226931 | 176.0355809 | | |
| 91 | 2008 | | 176.2847571 | 175.9710076 | = forecast 2008 damp factor 0.3 | |
| | | | forecast 2008 with 0.9 damp | | | |

How might you determine which overall forecast was "better"?

9.9 Test Your Knowledge: Time Series Models

Forecast year 2008 attendance using these 3 methods: autoregressive AR(1), m

| Year | Attendance | AR (1) data | Moving average data (use k=5) | Exponential smoothing (use 1-w=0.8) |
|------|------------|-----------------------|----------------------------------|--|
| 1916 | 5743 | | | #N/A |
| 1917 | 4678 | 5743 | #N/A | 5743 |
| 1918 | 4558 | 4678 | #N/A | 5530 |
| 1919 | 5978 | 4558 | #N/A | 5335.6 |
| 1920 | 6244 | 5978 | #N/A | 5464.08 |
| 1921 | 5396 | 6244 | 5440 | 5620.064 |
| 1922 | 7135 | 5396 | 5371 | 5575.2512 |
| 1923 | 9139 | 7135 | 5862 | 5887.20096 |
| 1924 | 9191 | 9139 | 6778 | 6537.560768 |
| 1925 | 8086 | 9191 | 7421 | 7068.248614 |
| 1936 | 9083 | 8995 | 10494 | 11103.34224 |
| 2000 | 34438 | 34739 | 29562 | 29608.1323 |
| 2001 | 34314 | 34438 | 31121 | 30574.10584 |
| 2002 | 33248 | 34314 | 32504 | 31322.08467 |
| 2003 | 36576 | 33248 | 33746 | 31707.26774 |
| 2004 | 38660 | 36576 | 34663 | 32681.01419 |
| 2005 | 38272 | 38660 | 35447 | 33876.81135 |
| 2006 | 38558 | 38272 | 36214 | 34755.84908 |
| 2007 | 40154 | 38558 | 37063 | 35516.27927 |
| 2008 | | 40093 | 38444 | 36444 |
| | | substitute in formula | drag down | drag down |



Agenda

| Topic | Time | Thursday Section | Sunday Section |
|--|--------|------------------|----------------|
| Introduction | 5 min | 9:00 - 9:05 | 6:30 - 6:35 |
| Highlights from Week 9 Video | 30 min | 9:05 - 9:35 | 6:35 - 7:05 |
| Start on Final Review | 25 min | 9:35 - 10:00 | 7:05 - 7:30 |
| Breakout on Regression | 20 min | 10:00 - 10:20 | 7:30 - 7:50 |
| Review of Upcoming Assignments and Open Question | 10 min | 10:20 - 10:30 | 7:50 - 8:00 |

Final Review

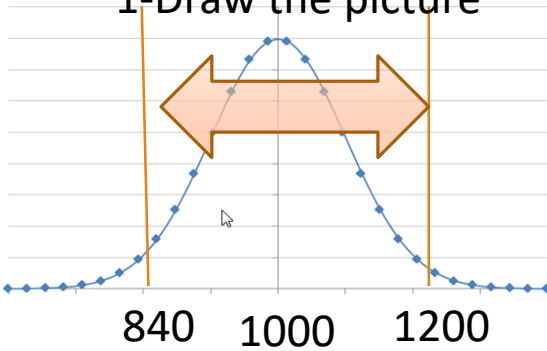
- **Regression**
 - How do you write down the formula from Excel Output?
 - How to use a regression equation to predict the output(y)?
 - How do you tell which variables are useful to have in your regression?
- **Correlation Coefficient R vs Coefficient of Determination R² – what do they represent?**
- **When is data not appropriate for regression? What are residuals?**
- **Causation vs Correlation**
- **Z calculation for the probability of a value falling between A and B**
- **Time Series - Autocorrelation and R²**
- **List of Statistical tools: Correlation, regression, hypothesis testing, scatter plots, process control charts, chi-square testing, etc. etc.**
- **Basic ways to describe data and Calculate: mean, median, mode, range, standard deviation, variance**
- **Sample size formula and manipulation impacts**
- **Margin of error and confidence intervals**
- **Process Control charts**
- **How can you determine if your measurement system is repeatable and reproducible?**
- **Hypothesis testing – at what alpha do you reject, at what p-value do you reject**

Quiz 2 Prep Question 1: Practice with Z calculations

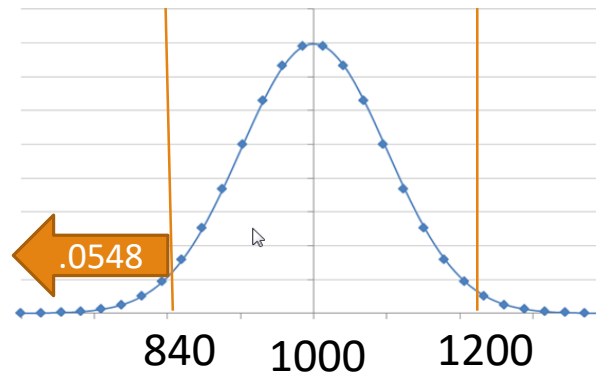
The distribution of weekly incomes of supervisors at the ABC Company follows the normal distribution, with a mean of \$1000 and a standard deviation of \$100.

What percent of the supervisors have a weekly income between \$840 and \$1200?

1-Draw the picture



2-Think about what you are calculating related to the picture



$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{840 - 1000}{100} = -1.6$$

Look up in tables, $p = .0548$

Or in Excel

=NORM.DIST(840,1000,100,TRUE)

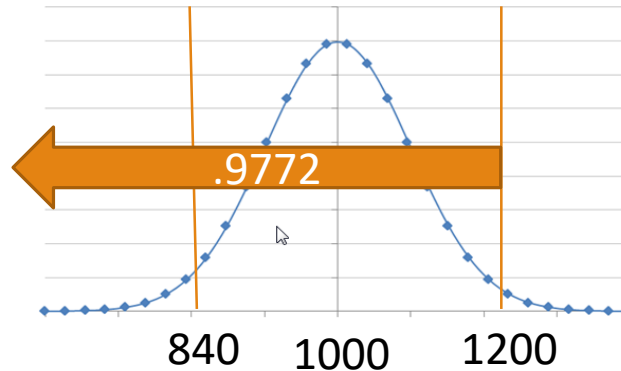
$$Z = \frac{X - \mu}{\sigma}$$

$$Z = \frac{1200 - 1000}{100} = 2$$

Look up in tables, $p = .9772$

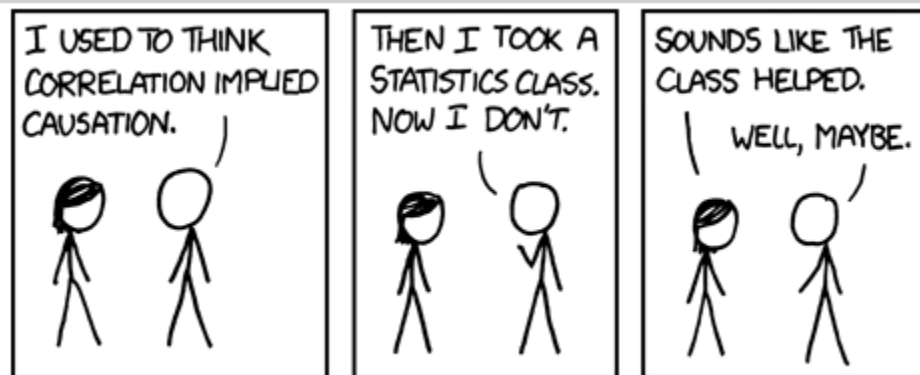
Or in Excel

=NORM.DIST(1200,1000,100,TRUE)



$.9772 - .0548 = .9224$, so 92.24% have a weekly income between \$840 and \$1200

Correlation vs Causation



From [xkcd](#), a comic by Randall Munroe

Correlation refers to the degree in which two measurements tend to vary together. Take the correlation between ice cream sales and drowning deaths. As ice cream sales increase, so do drowning deaths. Does that mean selling ice cream causes people to drown? Probably not. More likely is that people swim more and eat more ice cream the hotter it gets, so both are driven by the outside temperature.

Strong correlation doesn't mean cause and effect relationship....

Correlation has different causes

- the first caused the second
- the second caused the first
- Confounding factor– interference by a third variable distorts the association being studied between two other variables, because of a strong relationship with both of the other variables
- Common Cause – like the ice cream example
- Coincidence

Highlights: Video Segment 6.7:Correlation

Two Indices

1. Correlation coefficient (r)
2. Coefficient of determination (r^2)

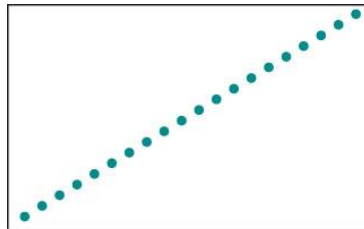
Correlation Coefficient (r)

- $-1 < r$
- -1 = perfect negative correlation
- 1 = perfect positive correlation
- 0 = no relationship
- Rule of thumb: r value of $\sim \pm 0.7$ desired
- Indicates meaningful relationship

Scatterplots provide a visual description of the relationship between two quantitative variables. The *correlation coefficient* is a numerical measure for quantifying the linear relationship between two quantitative variables.

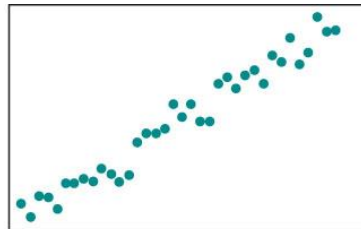
If the variability decreases, what does your correlation coefficient get closer to?
What does a correlation coefficient $r = -.72$ mean?

Properties of r



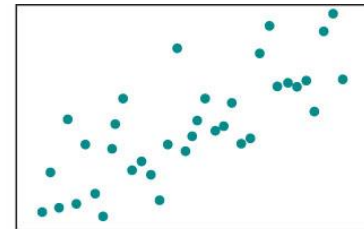
Perfect positive linear relationship, $r = 1$

(a)



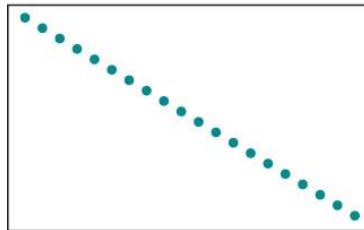
Strong positive linear relationship, $r = 0.9$

(b)



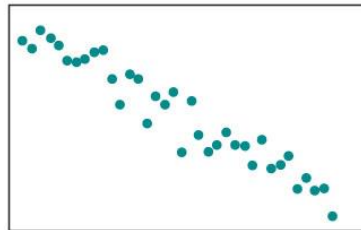
Moderate positive linear relationship, $r = 0.5$

(c)



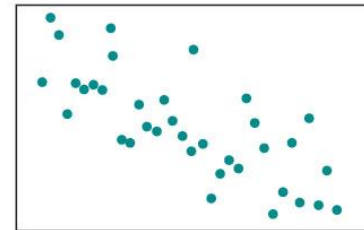
Perfect negative linear relationship, $r = -1$

(d)



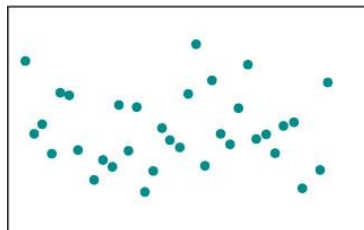
Strong negative linear relationship, $r = -0.9$

(e)



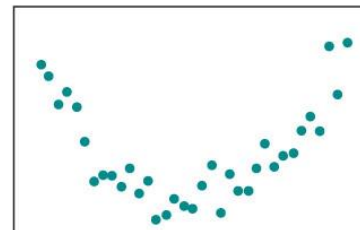
Moderate negative linear relationship, $r = -0.5$

(f)



No apparent linear relationship, $r = 0$

(g)



Nonlinear relationship but no linear relationship, $r = 0$

(h)

What if you performed a linear regression analysis on successive values of a time series analysis and you see autocorrelation.....what might your r^2 ?

Highlights: Video Segment 6.7:Correlation

Two Indices

1. Correlation coefficient (r)
2. Coefficient of determination (r^2)

Coefficient of Determination (r^2)

- Correlation coefficient squared
- Measure of the percentage of variability in y that can be accounted for by x
 - Trying to find an input x that is influencing our output y
 - x will not explain all of y
 - Recall: There is variability in everything we do.
- Metric for whether input x is really contributing to output

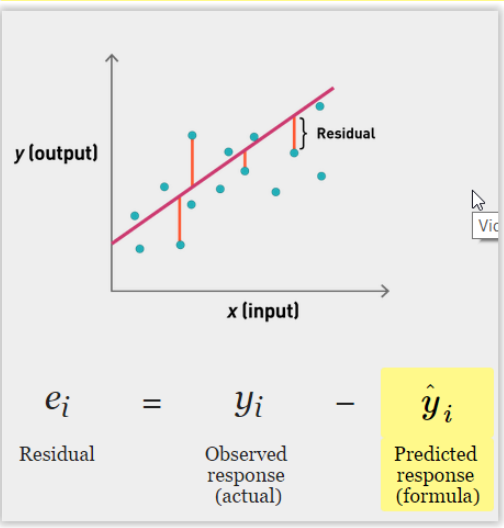
Measures the goodness of fit of the regression equation to the data. We interpret r^2 as the proportion of the variability in y that is accounted for by the linear relationship between y and x . The values that r^2 can take are $0 \leq r^2 \leq 1$.

Answer: correlation, r would be closer to 1 or -1, which would mean r^2 would be close to 1

Highlights: Video Segment 6.9:Residuals and Other Warnings

What Is a Residual?

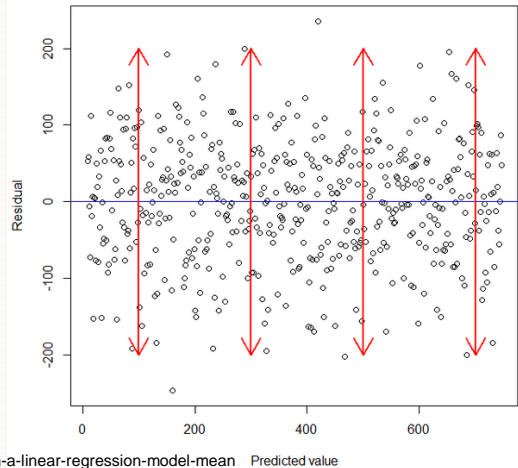
- Synonymous to error; should be random
- The distance between actual data point and the line determined by linear equation
- Determined by the difference between observed and predicted values of y
 - Ideally, points fall on regression line (i.e., perfect model)
 - Error would then be zero (rare).
- When plotted, a random series of points around a zero reference with no evidence of a pattern



Assumptions of Regression

1. Residuals are independent.
2. Residuals are normally distributed with a mean of zero.
 - The regression line will sometimes be high or low (i.e., over- or underpredicting).
3. There are equal variances (σ^2) of y .

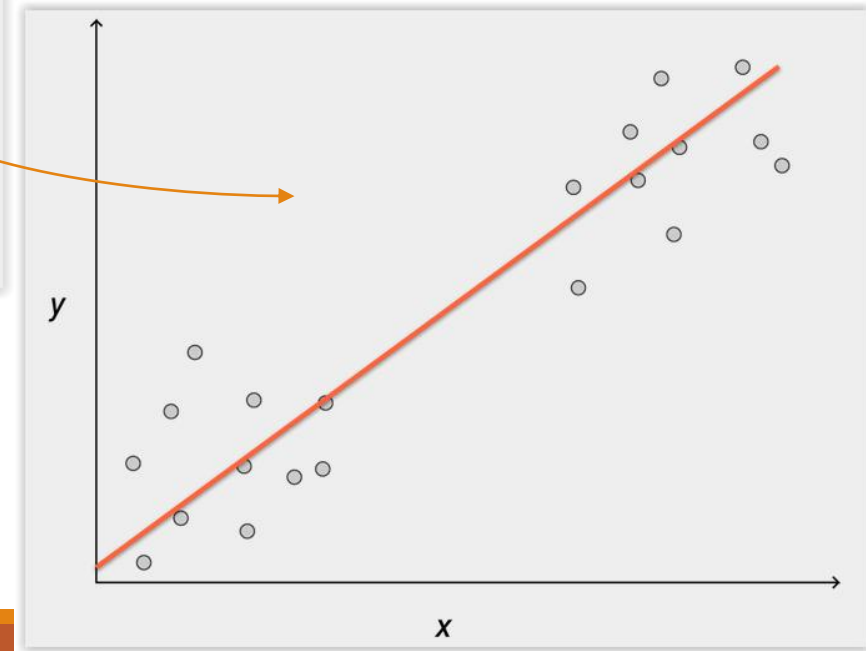
It means that when you plot the individual error against the predicted value, the variance of the error predicted value should be constant. See the red arrows in the picture below, the length of the red lines (a proxy of its variance) are the same.



Highlights: Video Segment 6.9:Residuals and Other Warnings

Other Points of Interest

- Certain data is inappropriate for a regression analysis:
 - Residuals form a pattern.
 - Large outliers are present.
 - "Clumped" data appears linear.
- Avoid extrapolating outside data.
- Beware of lurking variables, or Simpson's paradox.
- A strong correlation does not mean causation.




Agenda

| Topic | Time | Thursday Section | Sunday Section |
|--|--------|------------------|----------------|
| Introduction | 5 min | 9:00 - 9:05 | 6:30 - 6:35 |
| Highlights from Week 9 Video | 30 min | 9:05 - 9:35 | 6:35 - 7:05 |
| Start on Final Review | 25 min | 9:35 - 10:00 | 7:05 - 7:30 |
| Breakout on Regression | 20 min | 10:00 - 10:20 | 7:30 - 7:50 |
| Review of Upcoming Assignments and Open Question | 10 min | 10:20 - 10:30 | 7:50 - 8:00 |

Breakouts on Regression


Review of Upcoming Assignments: Thursday Section

- Homework #6, due Sunday, 12/2, Midnight EST in Learning Management System(2U) based on a file you will download in the assignments section for homework #6. **You will upload 1 excel file with your name in the name of the file and your answers clearly marked.**
- Process Improvement Project is due 12/10 – **You will upload 1 PowerPoint file with your name in the name of the file, you can submit a .ppt or .pdf**
- Final Exam is live, 12/13 at 9PM EST, Time limit: 90 mins, No partial credit

| | December 2018 | | | | | | |
|----------|---|--|---------|-----------|--|--------|----------|
| | Sunday | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
| Week #9 | 25 | 26 | 27 | 28 | 29 | 30 | 1 |
| | <u>Homework #5 Due:</u> 1. Problems 1-10 pg 114-116 in Understanding Variation | | | | Live Class #9  | | |
| Week #10 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | <u>Homework #6 Due:</u> 1. Time Series Problem posted in Excel | | | | Live Class #10 | | |
| Week #11 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| | | <u>Process Improvement Project DUE</u> | | | Live Class #11: FINAL EXAM | | |
| | | | | | | | |

Review of Upcoming Assignments: Sunday Section

1. Homework #6, due Wednesday, 12/5, Midnight EST in Learning Management System(2U) based on a file you will download in the assignments section for homework #6. **You will upload 1 excel file with your name in the name of the file and your answers clearly marked.**
2. Process Improvement Project is due 12/13 – **You will upload 1 PowerPoint file with your name in the name of the file, you can submit a .ppt or .pdf**
3. Final Exam is live, 12/16 at 630PM EST, Time limit: 90 mins, No partial credit

| | December 2018 | | | | | | |
|----------|--|--------|---------|---|--|--------|----------|
| | Sunday | Monday | Tuesday | Wednesday | Thursday | Friday | Saturday |
| Week #9 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| | Live Class #9  | | | <u>Homework #6 Due:</u> 1. Time Series Problem posted in Excel | | | |
| Week #10 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| | Live Class #10 | | | | <u>Process Improvement Project DUE</u> | | |
| Week #11 | 16 | 17 | 18 | 19 | 20 | 21 | 22 |
| | Live Class #11: FINAL EXAM | | | | | | |
| | | | | | | | |

Homework #6

| | | |
|----|--|---|
| 1 | The next tab of this Excel spreadsheet contains the NFL raw data for these problems. | D |
| 2 | | |
| 3 | In the National Football League, the philosophy for winning (rushing, passing, defense) seems to go through cycles. Consider a time series of the average number of rushing yards in the NFL per regular season from 1980 to 2008. | |
| 4 | | |
| 5 | 1) Make a time series plot. Is there evidence that the average rushing yards is trending in one direction? Describe the general movement of the series. | |
| 6 | | |
| 7 | 2) Fit a first order autoregressive model [AR(1)] using $y(t)$ as the response variable and $y(t-1)$ as the input variable. Record the regression equation. | |
| 8 | | |
| 9 | 3) Based on the AR(1) model , forecast the average number of rushing yards in the NFL for the 2009 regular season. | |
| 10 | | |
| 11 | 4) Calculate the exponential smoothing models using Excel damping factors 0.8 and 0.2 For each of the exponential smoothing models forecast the average number of rushing yards in the NFL for the 2009 season. | |
| 12 | | |
| 13 | 5) Calculate a moving average model using $k=5$ (Excel interval). Forecast the average number of rushing yards in the NFL for the 2009 season. | |
| 14 | | |

Project Rubric

| Process Improvement Project – Feedback – | | | |
|---|-----------------|---------------|----------|
| Content Requirements | Possible Points | Points Earned | Comments |
| Project | | | |
| A) An executive summary is provided in the storyboard format including: Is the storyboard presented in 1 PowerPoint slide? Follows DMAIC? Are tools/graphs/charts used and clearly visible? Do they support findings and conclusions Are arrows, call-out boxes, etc. used to summarize, highlight questions and key learnings? Are expected results clear? And next steps noted? | 5 | | |
| B) Is it a cohesive presentation opening with the business process and problem statement? The back-up slides (5-15) detail and support the storyboard content. | 2 | | |
| C) Was the success measure clearly identified, operationally defined and baseline identified? (Was the data identified as continuous or discrete, includes SQL?) | 3 | | |
| D) Was the data measurement plan or data stratification tree included? | 1 | | |
| E) Was the data collection method identified? | 1 | | |
| F) Was there rationale for the sample size taken? Use of the formula? Is there any reference to measurement error and how to minimize? | 1 | | |
| G) Are at least 5 different tools and techniques clearly identified? Are the tools linked/ pertinent to the data analysis? | 5 | | |
| H) Does the data analysis clearly tie to the problem conclusion? Is the “discovery” clear to the reader? | 2 | | |
| Total possible 100 points | 20 | 0.00 | |