Nassim Taleb

Nassim Nicholas Taleb is a prominent statistician, author, and stock trader. He is widely recognized for coining the term "black swan" to describe rare, unpredictable events that can have a major impact on our lives, society, and predictive models. He has written several influential books including "The Black Swan" and "Antifragile." [11,12]

The COVID-19 pandemic can be considered a black swan event due to its unprecedented nature, global impact, and the significant disruptions it caused across various sectors. Black Swan events can cause severe inaccuracies in the predictive ability of regression models, as you can see if you try to predict S.C. unemployment for 2021.

If we look at the data for South Carolina from January 2010 through January 2020, we can see a clear downward trend of the unemployment rate over time, with some minor ups and downs which are normal due to economic cycles.[10] Using January 2010 as Month 1 and January 2020 as Month 121, we will model the trend by fitting a linear regression model to the data in order to predict the unemployment rate using time (*Month*) as our only independent variable.

The least squares equation is

$$\textit{Estimated Unemployment Rate} = \hat{y} = 11.3805 - 0.07944\ \textit{Month}.$$

The computer output below tells us we have a good model of the unemployment rate using *Month* as the independent variable. Notice the very high R^2 value of almost 97%.

Coefficients

Term	Coef	SE Coef	T-Value	P-Value	VIF
Constant	11.3805	0.0908	125.35	0.000	
Period	-0.07944	0.00129	-61.51	0.000	1.00

Model Summary

S	R-sq	R-sq(adj)	R-sq(pred)
0.496235	96.95%	96.92%	96.85%

Use the model to predict the unemployment rate in South Carolina for April 2020.

Solution

Using January 2010 as Month 1, then April 2020 would be Month 124. Replacing this value into our model we get:

$$\textit{Estimated Unemployment Rate} = \hat{y} = 11.3805 - 0.07944(124) = 1.52994.$$

This value would align with the downward trend we observed in the graph. If we also consider our model is explaining almost 97% of the variability in the unemployment rate, then we would be very confident that the unemployment rate in South Carolina in April 2020 will be close to 1.53%. However, if we look at the actual value for April 2020, we find that the unemployment rate was actually 11.7%. That is a large error. What happened? As you may know, in the beginning of 2020 we had an unprecedented global pandemic, which forced a lot of companies to lay off many of their employees, resulting in skyrocketing unemployment rates. No model could have predicted this.

5.4 Exercises

Basic Concepts

1. Why is the mean not a reasonable descriptor for nonstationary time series data?
2. What is a linear time trend?
3. What is the independent variable in a linear trend model?
4. Is there a difference between the way the best fit line is determined for time series data and the way it is determined for other types of data?

Exercises

5. Using the CO_2 Emissions data set from the companion website, look at the CO_2 emissions per capita over time for Chile. Use the data to answer the following.

a. Looking at the data for Chile, do you believe the trend line will slope upward or downward?

b. Suppose we are interested in constructing a linear trend model for the data. Identify the independent and dependent variables for this model.

c. Write the general equation for the time trend model in terms of year and CO_2 emissions per capita.

d. Use statistical software to estimate the least squares model for the data.

e. Use this model to predict the CO_2 emissions per capita for Chile in 2020.

f. Can we determine the accuracy of this prediction? Explain.

Data

stat.hawkeslearning.com
Discovering Statistics and Data, Fourth Edition > Data Sets > CO_2 Emissions

6. Consider the following monthly sales data for an up-and-coming technology company.

Sales Data

Month	Sales (Thousands of Dollars)	Month	Sales (Thousands of Dollars)
1	321	7	698
2	542	8	710
3	540	9	799
4	581	10	821
5	641	11	833
6	700	12	850

a. Identify the independent and dependent variables for the linear time trend model.

b. Using statistical software, the following summary output was produced. Write the estimated regression equation.

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.949341195
R Square	0.901248704
Adjusted R Square	0.891373575
Standard Error	51.20789475
Observations	12

ANOVA

	df	*SS*	*MS*	*F*
Regression	1	239318.1818	239318.1818	91.26449427
Residual	10	26222.48485	2622.248485	
Total	11	265540.6667		

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	403.7575758	31.51628057	12.81107949	1.57569E-07
Month	40.90909091	4.282219283	9.553245222	2.41268E-06

c. What is the mean square error for this model? The standard error?

d. Using this model, predict the company's sales for the 13th month.

e. What percent of the variation in sales is explained by the linear time trend model? Does this model seem to accurately fit the data?

7. Consider the following data on beer production in the United States from 2010 to 2021.[13]

Beer Production in the U.S., 2010 to 2021	
Year	Production (in million barrels)
2010	195.14
2011	192.72
2012	195.74
2013	191.60
2014	192.56
2015	191.00
2016	190.46
2017	185.57
2018	183.28
2019	179.98
2020	179.95
2021	180.89

a. Identify the independent and dependent variables for a linear time trend model.

b. Using statistical software, determine the estimated regression equation for the data.

c. Based on the coefficient of determination, is this model a good fit to the data? Explain.

d. Predict the beer production for the next time period (*Year* = 2022).

e. What might be some external factors that would cause beer production to increase?

f. What might be some external factors that would cause beer production to decrease?

8. Consider the following time series data collected for three variables: Mean Sea Level, Air Pollution from Nitrogen Oxides, and Active Facebook Users Worldwide.

Sea Level, Air Pollution, Facebook Users			
Time Period	**Mean Sea Level (meters)**	**Air Pollution NOx(1,000 tons)**	**Active Facebook Users Worldwide (millions)**
1	0.227	26,883	1936
2	0.058	26,377	2006
3	0.065	27,079	2072
4	0.111	25,757	2129

Sea Level, Air Pollution, Facebook Users			
Time Period	Mean Sea Level (meters)	Air Pollution NOx(1,000 tons)	Active Facebook Users Worldwide (millions)
5	0.189	25,165	2196
6	0.157	24,697	2234
7	0.126	22,335	2271
8	0.079	20,261	2320
9	0.139	14,750	2375
10	0.170	11,563	2414
11	0.148	7985	2449
12	0.173	7645	2498

a. Plot each of the three variables against the *Time Period* variable.

b. What might be some confounding variables that could affect the predictive accuracy of each linear model?

c. Based on the plots in part **a.**, which set of data do you think is the best candidate to model with a linear time trend model? Explain your reasoning.

d. Fit a linear model to the data you selected in part **c.** and calculate the coefficient of determination.

e. For the data selected in part **c.**, plot the data again adding a linear trend line and predict the value for the next observation (*Time Period* = 13).

5.5 Scatterplots for More Than Two Variables

As displayed in Minard's graph of Napoleon's march in Chapter 3, it is often desirable to show more than two variables within the same graphic to determine if a relationship exists. However, a certain degree of creativity is required in order to figure out how to depict the desired variables within the spatial or dimensional limits. As each variable is added to the graph, a new method of reference must be associated with it. Scatterplots can be employed to portray relationships between more than two variables. Suppose we are interested in creating a graph that compares countries based on air pollution and rooms per person. Using data from the Organisation for Economic Cooperation and Development (OECD) Better Life Index for 2022, we created the graph in Figure 5.5.1.[14]

Data

The full OECD Better Life Index 2020 data set can be found on stat.hawkeslearning.com under **Discovering Statistics and Data, Fourth Edition > Data Sets > OECD Better Life Index 2022**.