LAB: INTRO TO STAT ANALYSIS

Zeyi Qian

zeqian@clarku.edu
Office Hours: JC 201, Tuesday 4–5 PM & Thursday 3-4 PM

September 27, 2024

Question 1 a & b

On Canvas, you will find a dataset on various measures of the 50 United States. The Murder rate is per 100,000, HS Graduation rate is in %, Income is per capita income in dollars, Illiteracy rate is per 1000, and Life Expectancy is in years. You are interested in studying determinants of Life Expectancy

- a. Write the regression equation
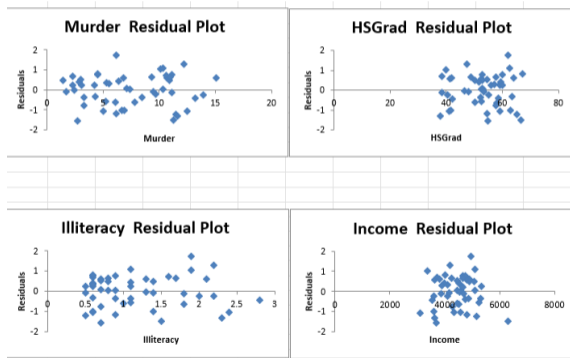 $LifeExpectancy =$
  $\beta_0 + \beta_1(MurderRate) + \beta_2(HSGraduationRate) + \beta_4(Income) + \beta_4(IlliteracyRate)$
- b. Check the conditions for multiple regression models. Show your graphs and discuss
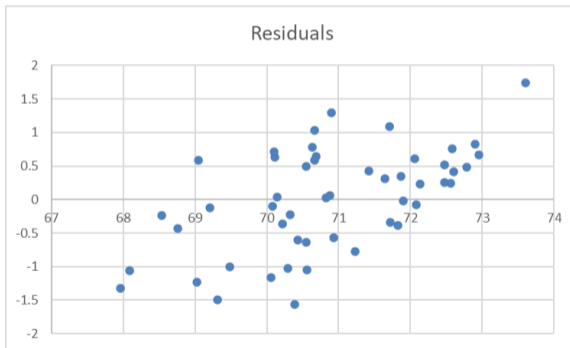 **Use "data analysis tool"**

Question 1 b

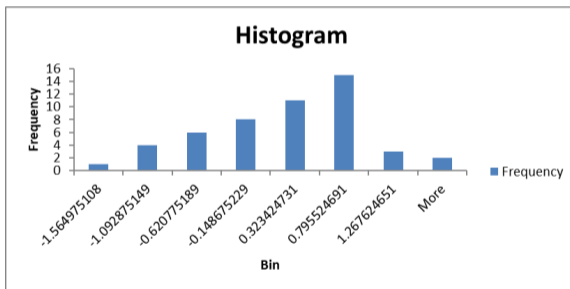Linearity Assumption & Equal Variance Assumption

Question 1 b

Independence Assumption

Question 1 b

Normal Population Assumption

Question 1 c & d

- c. What is the coefficient associated with Income? Is the sign of the relationship what we would have expected? Put the coefficient in a sentence (i.e., what does it mean)
- d. What is the coefficient associated with HS graduation rate? Is the sign of the relationship what we would have expected? Put the coefficient in a sentence (i.e., what does it mean)

	Coefficients	
Intercept	69.4833066	
Murder	-0.261940166	
HSGrad	0.046144327	
Income	0.000124948	
Illiteracy	0.276077144	

Question 1 c & d

If all other variables remain constant:

- for every additional dollar in per capita income, life expectancy increases by approximately 0.000124948 years (or about 0.0456 days)
- for every 1 percentage point increase in the high school graduation rate, life expectancy increases by approximately 0.046144327 years (about 16.8 days)

Question 1 e

e. Now drop Illiteracy rate from the regression model (i.e., run a new regression model without that variable). What is the coefficient associated with HS graduation rate in this new model?

	Coefficients
Intercept	70.14211
Murder	-0.2386
HSGrad	0.039059
Income	9.53E-05

Question 2 a

On Canvas, you will find a dataset on housing prices. You are interested in predicting housing prices from house characteristics

- a. Estimate a regression model with price as the dependent variable and living area, number of bedrooms and number of bathrooms as the explanatory variables and interpret your coefficients

	Coefficients
Intercept	7510.079074
livingarea	76.94313419
bathrooms	24474.46378
bedrooms	-8417.948057

Question 2 a

- For each additional square foot of living area, the house price is expected to increase by 76
- For each additional bedroom, the house price is expected to increase by 21141
- For each additional bathroom, the house price is expected to increase by -8417

Question 2 b

- b. Now add age to your regression model. Do any of the coefficients change?
  Explain

	Coefficients
Intercept	15344.15391
livingarea	76.97949502
bathrooms	20183.65655
bedrooms	-6972.0246
age	-150.5747329

- Some of the coefficients might change because age is correlated with other variables

Question 2 c

- c. You think the effect of number of bathrooms might not be linear so instead of using a continuous variable you decide to create an indicator variable for having 2 or more bathrooms ($=1$ if $>=2$, $=0$ if <2). Estimate this new regression model and interpret your coefficients

Question 2 c

- This coefficient represents the difference (13256) in price between homes with 2 or more bathrooms and those with fewer than 2 bathrooms

	Coefficients
Intercept	32529.10559
livingarea	84.51854967
bedrooms	-6335.650695
age	-204.0776306
b	13256.24109

Question 2 d

- d. You think the effect of number of living area on price might depend on the age of the household. Specifically, you think that bigger houses are more expensive if they are more modern (say, built within the last 20 years). Create an indicator variable for "age<=20" and estimate a regression model to test your hypothesis. Interpret your coefficients

$$Price = \beta_0 + \beta_1(LivingArea) + \beta_2(Bedrooms) + \beta_3(Bathrooms) + \beta_4(age <= 20)$$
$$+\beta_5(LivingArea \times (age <= 20))$$

Question 2 d

	A	B	C	D	E	F	G	H
							f_x	=IF(D2<=20,1,0)

	A	B	C	D	E	F	G	H
	price	livingarea	bedrooms	age	bathrooms	a		
	142212	1982	3	133	1	0		
	134865	1676	3	14	1.5	1		
	118007	1694	3	15	2	1		
	138297	1800	2	49	1	0		
	129470	2088	3	29	1	0		
	206512	1456	3	10	2	1		
	50709	960	2	12	1.5	1		

G2							f_x	=F2*B2

	A	B	C	D	E	F	G	H
	price	livingarea	bedrooms	age	bathrooms	a	Interaction	
	142212	1982	3	133	1	0	0	
	134865	1676	3	14	1.5	1	1676	
	118007	1694	3	15	2	1	1694	
	138297	1800	2	49	1	0	0	
	129470	2088	3	29	1	0	0	

Question 2 d

	Coefficients
Intercept	47405.16589
livingarea	49.67923249
bedrooms	-7778.293006
bathrooms	24211.139
age<=20	-59187.88834
Interaction	36.97704668

- β_1 represents the effect of living area on price for houses older than 20 years
- β_4 represents the price difference between modern houses and older houses
- β_5 means that larger houses have a greater impact on price (36) when they are more modern

Question 3 a

On Canvas, you will find a dataset on CO2 emissions, global temperature anomalies, and the Dow-Jones Industrial Average (DJIA) index from 1959 to 2016

- a. Estimate a regression model to test if there is correlation between global temperature anomalies and CO2 emissions **(data analysis tool)**

$$TempAnomalies = \beta_0 + \beta_1(CO2Emissions)$$

Question 3 a

	Coefficients
Intercept	-3.179329991
CO2ppm	0.009917924

- β_1 represents the change in global temperature anomalies (0.009) for each additional unit of CO2 emissions

Question 3 b

- b. Add DJIA to the regression model and interpret your results

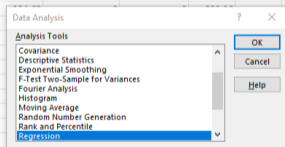|	Coefficients
Intercept	-2.870719112
CO2ppm	0.008970918
DJIA	4.81676E-06

Question 3 c

- c. Suppose you think the relationship between CO2 emissions and temperature anomalies has changed over time as CO2 emissions have increased more rapidly in the last few decades. Test your hypothesis by creating 3 indicator variables: 1) =1 if year<1970, 2) =1 if year>=1970 & year<=1990, 3) =1 if year>1990

$$TempAnomalies = \beta_0 + \beta_1(CO2Emissions) + \beta_2(pre1970) + \beta_3(1970to1990)$$
$$+ \beta_4(post1990) + \beta_5(CO2Emissions) \times (pre1970)$$
$$+ \beta_6(CO2Emissions) \times (1970to1990) + \beta_7(post1990)$$

- Excel
  pre1970 "=IF(A2<1970,1,0)"
  1970to1990 "=IF(AND(A2<=1990, A2>=1970),1,0)"
  post1990 "=IF(A2>1990,1,0)"

Question 3 c

	A	B	C	D	E	F	G	H	I	J	K	L
1	Year	Globaltempanomaly	CO2ppm	pre1970	1970to1990	post1990	pre1970_CO2	1970to1990_CO2	post1990_CO2	DJIA		
2	1959	0.0596	315.97	1	0	0	315.97	0	0	679.36		
3	1960	0.0204	316.91	1	0	0	316.91	0	0	615.89		
4	1961	0.0775	317.64	1	0	0	317.64	0	0	731.14		
5	1962	0.0888	318.45	1	0	0	318.45	0	0	652.1		
6	1963	0.1068	318.99	1	0	0	318.99	0	0	762.95		
7	1964	-0.1495	319.62	1	0	0	319.62	0	0	874.13		
8	1965	-0.078	320.04	1	0	0	320.04	0	0	969.26		
9	1966	-0.0227	321.38	1	0	0	321.38	0	0	785.69		
10	1967	-0.0131	322.16	1	0	0	322.16	0	0	905.11		
11	1968	-0.0296	323.04	1	0	0	323.04	0	0	943.75		
12	1969	0.0929	324.62	1	0	0						
13	1970	0.0372	325.68	0	1	0						
14	1971	-0.0783	326.32	0	1	0						
15	1972	0.0264	327.45	0	1	0						
16	1973	0.1641	329.68	0	1	0						
17	1974	-0.0719	330.18	0	1	0						
18	1975	0.0034	331.11	0	1	0						
19	1976	-0.0792	332.04	0	1	0						
20	1977	0.1978	333.83	0	1	0						
21	1978	0.1123	335.4	0	1	0						
22	1979	0.2273	336.84	0	1	0	0	336.84	0	838.74		
23	1980	0.2637	338.75	0	1	0	0	338.75	0	963.99		

Data Analysis

Analysis Tools

Covariance
Descriptive Statistics
Exponential Smoothing
F-Test Two-Sample for Variances
Fourier Analysis
Histogram
Moving Average
Random Number Generation
Rank and Percentile
Regression

OK
Cancel
Help

Question 3 c

	Coefficients	
Intercept	-4.54325811	
CO2ppm	0	
pre1970	6.255081604	
1970to1990	0	
post1990	1.430717388	
pre1970_CO2	-0.005307733	
1970to1990_CO2	0.013910774	
post1990_CO2	0.009740294	

- Before 1970, the average temperature anomaly is higher by 6.26 units compared to the base period when CO2 levels are zero
- After 1990, the temperature anomaly is 1.43 units higher than the base period
- For each additional ppm of CO2 before 1970, the temperature anomaly decreases slightly
- During 1970-1990, each additional ppm of CO2 increases temperature anomalies by 0.0139 units
- After 1990, each ppm of CO2 raises temperature anomalies by 0.0097 units