IP309: Data Analysis and Short Answer Report

1. Is there any evidence to suggest crime rates are higher in southern areas? Explain your approach and provide any results to support your view. (10%)

ANSWER

To answer whether crime rates are higher in southern areas, first, a t-test is conducted to see the mean comparison for only CrimeRate variable.

```
. ttest CrimeRate, by(Southern)

Two-sample t test with equal variances
```

Group	Obs	Mean	Std. err.	Std. dev.	[95% conf.	interval]
0	155	108.6129	2.52469	31.43213	103.6254	113.6004
1	80	106.35	2.710394	24.2425	100.9551	111.7449
Combined	235	107.8426	1.901594	29.15089	104.0961	111.589
diff		2.262903	4.018922		-5.655167	10.18097

```
    diff = mean(0) - mean(1)                                      t =   0.5631
H0: diff = 0                                     Degrees of freedom =      233

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.7130         Pr(|T| > |t|) = 0.5739          Pr(T > t) = 0.2870
```

Table 1. *T-test results for crime rate variable by Southern vs Non-Southern Counties*

From Table 1, it shows that the difference in means is not statistically significant. The mean crime rate for southern counties is 106.35, while the non-southern counties is higher with 108.61. Moreover, the t value is 0.5631, with p value of 0.5739. this indicates that H0 is rejected.

The second step is a regression analysis to control for other factors that may influence crime rates. By including control variables such as YouthUnemployment, Youth, Education, Wage, MatureUnemployment, and BelowWage, the results may be better due to these socioeconomic factors.

```
. regress CrimeRate Southern YouthUnemployment Youth Education Wage MatureUnemployment BelowWage
```

Source	SS	df	MS	Number of obs	=	235
				F(7, 227)	=	21.90
Model	80165.5654	7	11452.2236	Prob > F	=	0.0000
Residual	118681.609	227	522.826472	R-squared	=	0.4032
				Adj R-squared	=	0.3847
Total	198847.174	234	849.77425	Root MSE	=	22.865

CrimeRate	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Southern	-2.580569	5.04615	-0.51	0.610	-12.52385	7.362715
YouthUnemployment	-.4553372	.1286559	-3.54	0.000	-.7088497	-.2018247
Youth	.7524748	.1690868	4.45	0.000	.4192944	1.085655
Education	3.751025	1.624898	2.31	0.022	.5492134	6.952837
Wage	.3058892	.0314838	9.72	0.000	.2438514	.367927
MatureUnemployment	1.350066	.2942201	4.59	0.000	.7703146	1.929818
BelowWage	.4383093	.0815838	5.37	0.000	.2775509	.5990677
_cons	-295.6799	49.53378	-5.97	0.000	-393.2847	-198.0751

Table 2. *Regression analysis results for control variables affecting crime rates*

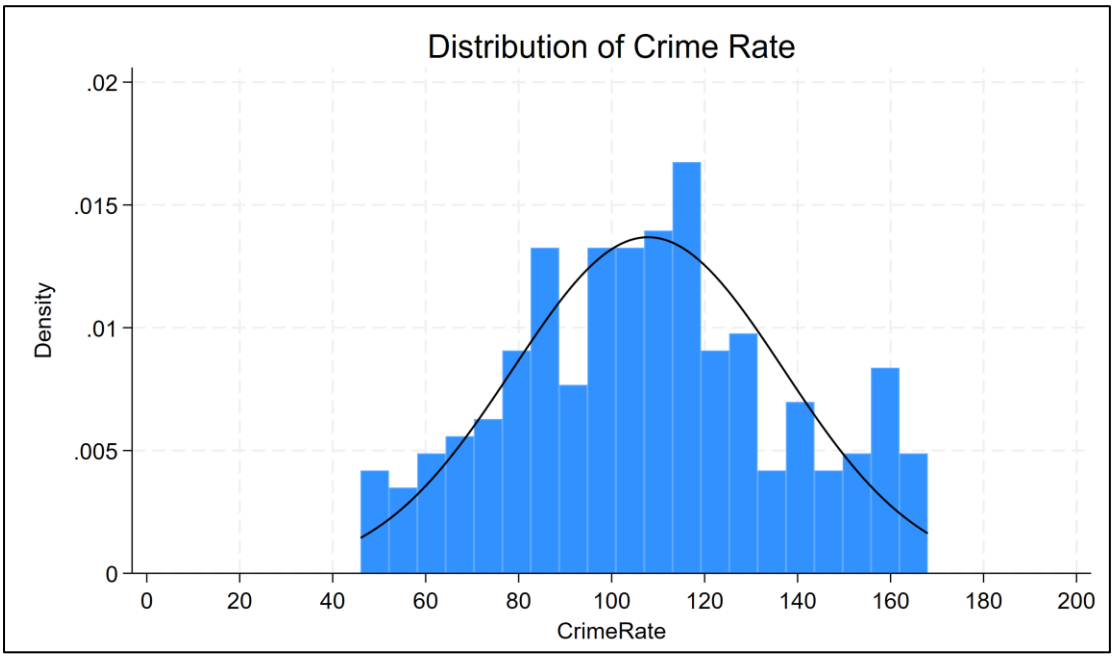
Table 2 shows that the Southern variable does not have a significant effect on crime rates with the p-value of 0.610, aligning with the results from the t-test. However, other variables indicate notable effects. YouthUnemployment is negatively associated (t = -3.54) with crime rates with the p-value of 0.000, while the Youth (t = 4.45) population shows a positive relationship with the same p value of 0.000. This suggests that an increase in the number of young males tends to correlate with higher crime rates. Furthermore, Higher education levels (t = 2.31) and Wage (t = 9.72) are also positively associated with crime rates, indicating that regions with more education and higher wages experience higher crime rates. MatureUnemployment and BelowWage also show a positive relationship with crime rates.

The results from T-test and regression analysis, when considered together, suggest that while location (Southern vs Non-Southern) may not be a significant factor for crime rates, factors like youth unemployment, education, and wages are a significant variables that majorly affect the growth of crime rates.

2. How would you characterise the distribution of the crime rate variable? Produce an appropriate visualisation to support your answer. (10%)

   ANSWER

To show the distribution of the crime rate variable, a histogram with normal distribution line must be created in STATA for the visualization. Moreover, a sktest will provide more details about the skewness and kurtosis of the histogram.



Graph 1. *The Histogram for the distribution of Crime Rate*

```
. sktest CrimeRate

Skewness and kurtosis tests for normality
                                                    ------ Joint test ------
    Variable |       Obs  Pr(skewness)  Pr(kurtosis)  Adj chi2(2)  Prob>chi2
-------------+--------------------------------------------------------------
   CrimeRate |       235        0.5337        0.0040         8.03     0.0180
```

Table 4. *Skewness and kurtosis test results for CrimeRate variable*

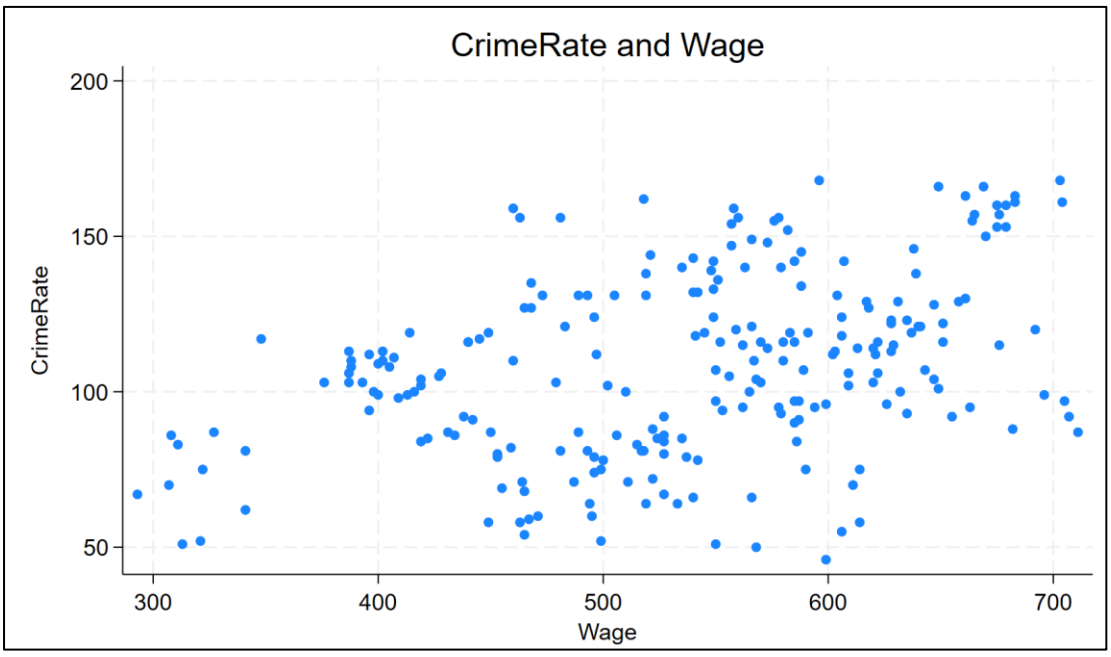
The graph shows a roughly symmetric bell shaped, with the majority of values concentrated around the 100 – 120. However, the density curve reveals slight deviations from the perfect normal distribution shape, showing minor skewness in the tail behaviour.

From the sktest, the skewness (0.5337) suggests that there is no significant skewness, meaning the distribution is relatively symmetric. On the other hand, the kurtosis (0.0040) indicates a significant deviation, which means that the distribution has heavier tails compared to a normal distribution. In addition, the joint test, combining both skewness and kurtosis, with the value of 0.0180, confirms that the crime rate distribution significantly departs from normality.

3. What if any, issues might you're answer to question 2 (above) present if trying to use regression analysis to examine crime rate? What implications can the distribution have for the interpretation of any regression results? (10%)

ANSWER

The distribution of the crime rate variable might present some problems during regression analysis due to its deviation from the normal distribution. From the sktest result in number 2, the test shows deviations if compared to a normal distribution, with high value of kurtosis and low value of skewness. With this in mind, non-normal residuals could lead to unreliable regression results such as standard errors and absurd values for the t and p values. Furthermore, there is a possibility of outliers since the heavy tails may severely impact the values of regression coefficients, leading to a biased or unstable estimation.



Graph 2. *Two-way scatter plot between CrimeRate and Wage*

Another potential issue that might be expected from the crime rates distribution is heteroscedasticity, for example, heteroscedasticity might be present since the variance in the crime rate is affected by the wages and the spread of crime

rates seem to widen as the wages increase. This implies that there is no fixed relation between wages and crime rate and might lead to the results of the regression.

4. Use Ordinary Least Squares (OLS) regression analysis to produce an appropriate model which explains the crime rate in this sample. Justify your selection of included variables, and present your results. (30%)

   ANSWER

   Prior to the regression, creating a correlation matrix is important to determine which variable is relevant for the regression, minimizing potential problems for the regression result.

```
. corr *
(obs=235)
```

	CrimeR~e	Youth	Southern	Educat~n	Expend~r	Labour~e	Males	County~e	YouthU~t	Mature~t	HighYo~y	Wage	BelowW~e
CrimeRate	1.0000												
Youth	-0.0578	1.0000											
Southern	-0.0369	0.5678	1.0000										
Education	0.1065	-0.4258	-0.4808	1.0000									
Expenditur~r	0.6218	-0.4882	-0.3756	0.2701	1.0000								
LabourForce	0.1283	-0.1510	-0.4921	0.3977	0.1092	1.0000							
Males	0.1242	-0.0085	-0.2907	0.2585	0.0456	0.5059	1.0000						
CountySize	0.2978	-0.2795	-0.0499	-0.0309	0.5119	-0.1279	-0.3876	1.0000					
YouthUnemp~t	-0.0762	-0.2360	-0.1622	0.0232	-0.0296	-0.2330	0.3129	-0.0291	1.0000				
MatureUnem~t	0.1586	-0.2389	0.0722	-0.1610	0.1653	-0.4139	-0.0295	0.2668	0.7199	1.0000			
HighYouthU~y	-0.3076	-0.0736	-0.3955	0.3111	-0.2214	0.4063	0.3656	-0.3651	0.0774	-0.4636	1.0000		
Wage	0.3956	-0.6415	-0.6195	0.4831	0.7656	0.2593	0.1617	0.3050	0.0226	0.0534	0.0297	1.0000	
BelowWage	-0.1276	0.6086	0.7219	-0.5595	-0.5943	-0.2424	-0.1348	-0.1033	-0.0382	0.0453	-0.1622	-0.8434	1.0000

Table 5. *Correlation matrix for all the variables*

Here is the list of the included and excluded variables:

a. Variable to Include

Variable Name	Value	Description
ExpenditureYear	0.6218	Strong positive correlation with crime rate, reflecting higher police funding in areas with higher crime
Wage	0.3956	Moderate positive correlation, capturing socioeconomic conditions
CountySize	0.2978	Moderate correlation, representing urbanization or population size of effects
HighYouthUnemploy	-0.3076	Moderate negative correlation, stronger correlation than Youth Unemployment or Mature Unemployment

Table 6. *Included variables*

b. Variable to Exclude

Variable Name	Value	Description
Youth	-0.0578	Weak correlation with CrimeRate, offering little explanatory
Southern	-0.0369	Insignificant correlation, suggesting no meaningful regional effect on crime
YouhUnemployment	-0.0762	Very weak correlation, redundant with HighYouthUnemploy, which is stronger.
MatureUnemployment	0.1586	Weak correlation and redundant with HighYouthUnemploy, which captures more.
BelowWage	-0.1276	Weak correlation and highly correlated with Wage, causing multicollinearity.
Males	0.1242	Weak correlation with CrimeRate and excluded based on theoretical preferences.
LabourForce	0.1283	Weak correlation, redundant when HighYouthUnemploy is included.
Education	0.1065	Weak correlation and excluded due to its limited explanatory significance.

Table 7. *Excluded variables*

```
. regress CrimeRate ExpenditureYear Wage CountySize HighYouthUnemploy
```

Source	SS	df	MS		
Model	85442.8808	4	21360.7202	Number of obs	= 235
Residual	113404.294	230	493.062146	F(4, 230)	= 43.32
				Prob > F	= 0.0000
Total	198847.174	234	849.77425	R-squared	= 0.4297
				Adj R-squared	= 0.4198
				Root MSE	= 22.205

CrimeRate	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
ExpenditureYear	.7048057	.0859313	8.20	0.000	.5354925	.8741188
Wage	-.0396721	.0249505	-1.59	0.113	-.0888328	.0094887
CountySize	-.0805246	.0470495	-1.71	0.088	-.1732277	.0121786
HighYouthUnemploy	-11.17394	3.487442	-3.20	0.002	-18.04536	-4.302519
_cons	78.50345	9.091845	8.63	0.000	60.5895	96.41741

Table 8. *Regression result for selected variables*

The regression analysis reveals that it accounts for 42.97% of the variation in crime rate ($R^2 = 0.4297$). While this might be lower compared to the regression with all the variables or a full regression, the result is still more focused and potentially avoids the risks of overfitting. Furthermore, including all variables introduces complexity due to some variable may be statistically insignificant ($p > 0.005$) or redundant due to multicollinearity, such as strong correlation between Wage and BelowWage.

5. What can regression analysis tell us about the impact of high youth unemployment on crime rates? Explain your answer. (you can use your regression model from Q4, but you do not need to use the same regression if you want to test different models or make comparisons to answer this question) (15%)

ANSWER

Comparing between the selected variables regression and the full variables regression is good to see whether high youth unemployment alone is the key variable that affects crime rates.

```
. regress *
```

Source	SS	df	MS		
Model	118503.076	12	9875.25633	Number of obs	= 235
Residual	80344.0985	222	361.910353	F(12, 222)	= 27.29
				Prob > F	= 0.0000
				R-squared	= 0.5960
				Adj R-squared	= 0.5741
Total	198847.174	234	849.77425	Root MSE	= 19.024

CrimeRate	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
Youth	.6694174	.1477574	4.53	0.000	.3782308	.960604
Southern	-.6715991	4.956993	-0.14	0.892	-10.44038	9.097183
Education	3.632584	1.413595	2.57	0.011	.846802	6.418366
ExpenditureYear	.6960498	.0763674	9.11	0.000	.545552	.8465475
LabourForce	.0990494	.0507495	1.95	0.052	-.0009631	.1990619
Males	.0077724	.0697922	0.11	0.911	-.1297677	.1453124
CountySize	-.0990221	.0481967	-2.05	0.041	-.1940038	-.0040405
YouthUnemployment	-.0973692	.1520154	-0.64	0.522	-.396947	.2022086
MatureUnemployment	.5701105	.3409429	1.67	0.096	-.1017882	1.242009
HighYouthUnemploy	-7.481765	4.414807	-1.69	0.092	-16.18206	1.218528
Wage	.1304374	.0326823	3.99	0.000	.0660302	.1948447
BelowWage	.4126495	.0756959	5.45	0.000	.2634751	.5618239
_cons	-305.4708	56.61629	-5.40	0.000	-417.0449	-193.8967

Table 9. *Regression analysis for all the variables*

In the simpler regression (Table 8), HighYouthUnemploy has a significant negative effect, which suggests that areas with high youth unemployment experience lower crime rates. On the other hand, the effect of high youth unemployment is still one of the strong variables affecting crime rate, but may be less significant in the full variables model, indicating that other variables, such as Education and Wage might help explain the relationship more effectively. Variables like these may capture the long-term opportunities or economic pressures of the people. Moreover, multicollinearity with all these variables may dilute its impact.

6. Suppose that crime rates are lower in areas with high youth unemployment, does regression analysis provide any evidence to support the idea that mature unemployment has a more negative impact on crime rates than youth unemployment? Are there any issues with how we select the variables to answer this question? Explain your answer. (same note as in Q5) (15%)

ANSWER

Hypothesis:

- **H0:** The impact of mature unemployment on crime rates is equal to the impact of youth unemployment on crime rates
- **H1:** The impact of mature unemployment on crime rates is stronger than the impact of youth unemployment

By comparing only the mature unemployment and youth unemployment variables, along with other supporting variables such as expenditure year, county size, and wage, it is more likely to see a precise result compared to using a full regression. Moreover, using test will see whether H0 is 1 or 0.

```
. regress CrimeRate MatureUnemployment YouthUnemployment ExpenditureYear CountySize Wage
```

Source	SS	df	MS		
Model	84799.5532	5	16959.9106	Number of obs	= 235
Residual	114047.621	229	498.024547	F(5, 229)	= 34.05
				Prob > F	= 0.0000
				R-squared	= 0.4265
				Adj R-squared	= 0.4139
Total	198847.174	234	849.77425	Root MSE	= 22.316

CrimeRate	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
MatureUnemployment	.781265	.2780222	2.81	0.005	.2334564	1.329074
YouthUnemployment	-.3405079	.1221002	-2.79	0.006	-.5810913	-.0999245
ExpenditureYear	.7179213	.0856987	8.38	0.000	.5490625	.8867802
CountySize	-.0821029	.0479152	-1.71	0.088	-.176514	.0123082
Wage	-.046514	.0246222	-1.89	0.060	-.0950291	.0020011
_cons	83.63669	11.59692	7.21	0.000	60.78639	106.487

Table 10. *Regression analysis for mature unemployment and youth unemployment*

```
. test MatureUnemployment = YouthUnemployment

( 1)  MatureUnemployment - YouthUnemployment = 0

       F(  1,    229) =    8.72
            Prob > F =    0.0035
```

Table 11. *Result of H0 test*

The value for MatureUnemployment (0.783) is larger and positive, whereas the value for YouthUnemployment (-0.3405) is smaller and holds a negative value. This suggests that mature unemployment has a stronger impact on crime rates than youth unemployment. The H0 test result indicates that the difference between the two

variables is statistically significant with $p < 0.05$, meaning that H0 is 0. Thus, H1 is accepted.

7. A researcher at the University of Warwick has suggested that the dataset is missing relevant socioeconomic position variables such as ethnicity, employment type and housing type. How do you respond to these claims and how might the omission of these variable affect your results? (10%)

   ANSWER

   I agree that by adding more relevant socioeconomic variables such as ethnicity, employment type, and housing type would make the analysis more accurate and significant as these variables are also one of the many important factors of crime rates. Adding these variables may help capture more aspects that affect crime rates that the current dataset does not fully address. By including them, it could help reduce bias and improve the precision of the analysis results. However, it is also important to acknowledge the limitations of the current analysis, as there is no perfect model that exists. Moreover, including more variables would make the model more comprehensive, but it could also make it more complex and might be a challenge to interpret. If in the end, these additional variables are added, we must be prepared and consider the trade-offs, understanding the positive and negative impact of our actions. It is not a sign of weakness to acknowledge these limitations, however, it demonstrates an act of caution, knowing the strength and value of improving the analysis

Appendices

STATA Do-File Code

```
//This is the Do-File for IP309: Data Analysis and Short Answer Report

//This is the Sample Answer
// X.Y
// X = The Question Number
// Y = Section of the Answer

//1.1 T-test to compare CrimeRate between Southern and Non-Southern Counties
ttest CrimeRate, by(Southern)

//1.2 Multiple Regression for Control Variables influencing CrimeRate
regress CrimeRate Southern YouthUnemployment Youth Education Wage MatureUnemployment BelowWage

//2.1 The Visualization
histogram CrimeRate, bin(20) normal ///
    title("Distribution of Crime Rate") ///
    xlabel(0(20)200, grid) ylabel(, grid)

//2.2 Skewness Test
sktest CrimeRate

//3.1 Two Way Scatter for CrimeRate and Wage
twoway scatter CrimeRate Wage, title("CrimeRate and Wage")

//4.1 Correlation Matrix for All Variables
corr *

//4.2 Regression for Selected Variables
regress CrimeRate ExpanditureYear Wage CountySize High

//5.1 Regression for Full Variables
regress *

//6.1 Regression for MatureUnemployment and YouthUnemployment
regress CrimeRate MatureUnemployment YouthUnemployment ExpanditureYear CountySize Wage

//6.2 Testing for H0
test MatureUnemployment = YouthUnemployment
```