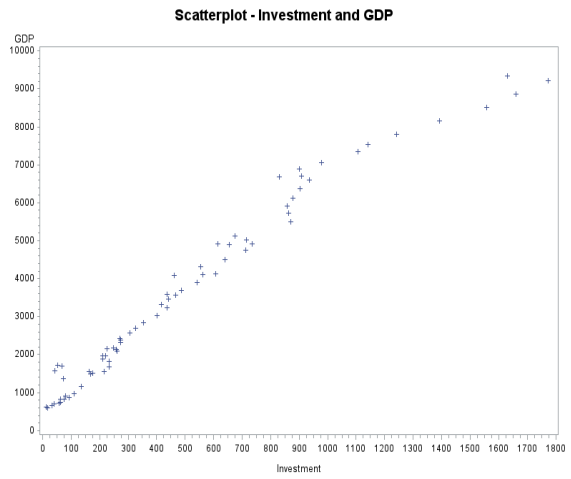
Group Assignment Econometrics 1

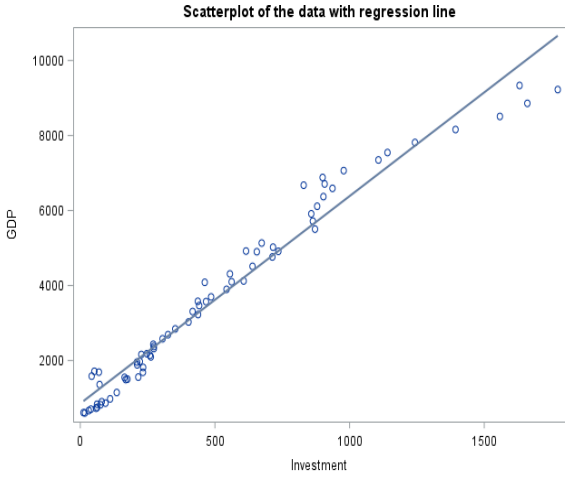
Exercise 1: Linear regression

1. Produce a scatter plot of the data. What relationship seems to exit?

It is evident from the plot below to the left that there is a quite strong positive linear relationship between the dependent variable GDP and the independent variable investment. It makes sense since investment together with consumption, government expenditure and net exports are the four major components in calculating the GDP according to macroeconomic theory.

.





2. Plot the data with regression equation line together and comment on it.

The regression equation line, above to the right, shows clearly a positive relationship between X and Y in the theoretical model:

$$GDPi = \beta_1 + \beta_2 Investment + u_i$$

3. Perform the regression analysis with the equation of line that best fit the data, plot $\widehat{ui}$ versus Xi, and comment on the plot.

Investment vs GDP

The REG Procedure
Model: MODEL1
Dependent Variable: GDP GDP

Number of Observations Read	72
Number of Observations Used	72

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	418579691	418579691	1888.24	<.0001
Error	70	15517432	221678		
Corrected Total	71	434097122			

Root MSE	470.82650	R-Square	0.9643
Dependent Mean	3668.24861	Adj R-Sq	0.9637
Coeff Var	12.83519		

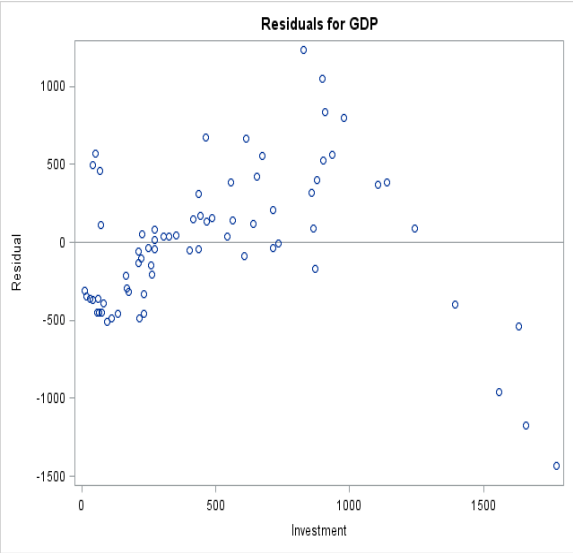
Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	855.18090	85.26268	10.03	<.0001
Investment	Investment	1	5.52881	0.12723	43.45	<.0001

The reported F-value shows that we have a highly significant model, and R^2 is 0.9643 which is very high. The estimated model is:

$$\widehat{GDPi} = 855.18090 + 5.52881 Investment$$

The reported t-values are also highly significant (at alpha=0.001). However, prior knowledge says that the variables stated in exercise 1.1) should also be included.



Looking at the residual plot for X_i, there seems to be a systematic pattern, the residual function of X_i is taking a concave form as the x-values get larger than approximately 1200 billion dollars. Possible reason to why the function has a concave appearance could be heteroscedasticity or that other explanatory variables are lacking in the model.

4. Use the model to estimate the mean value of GDP for the investment cost 966.

The estimated regression line or the sample regression function $\hat{y} = \hat{\beta}_1 + \hat{\beta}_2 x_i$ which predicts the PRF $E(Y|x_i) + u_i = Y_i = \beta_1 + \beta_2 x_i + u_i$ is:

$$\widehat{GDPi} = 855.18090 + 5.52881 * 966 \approx 6\,196$$

Using this model to estimate the mean value of GDP for the investment cost 966⇒ $\hat{y} = 855.18090 + 5.52881(966) \approx 6196$. As the investment level reaches 966 billion US dollars, the estimated value of GDP will be around 6 196 billion dollars on average.

5. Find the investment cost at which the estimated mean value of GDP is 3650.

$$3650 = 855.18090 + 5.52881 Investment$$

$$Investment = \frac{3650 - 855.18090}{5.52881} \approx 505.50$$

According to the estimated model, the investment-level needs to be approximately 505.5 billion dollars in order to reach a GDP of 3 650 billion dollars.

6. Test whether the residuals are homoscedastic.

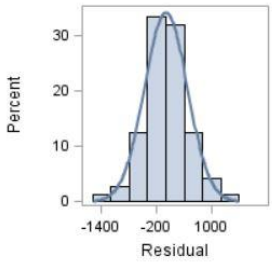
As mentioned earlier, we are suspecting heteroscedasticity since the residuals in the plot are taking a concave shape and they are also unevenly spread throughout the plot. Therefore, the error variance does not seem to be constant Symbolically we suspect this:

$$E(\widehat{u_i}) = \sigma_i^2$$

Which means the conditional variance of GDP is not the same as Investment increases. Heteroscedasticity would still give unbiased and consistent OLS-estimators. They would no longer be efficient with minimum variance, and therefore not BLUE with misleading t and F-tests.

In order to test if there is heteroscedasticity, one could do the White's and Breusch-Pagan test

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
GDP	White's Test	31.01	2	<.0001	Cross of all vars
	Breusch-Pagan	21.02	1	<.0001	1, Investment



As can be seen from the table above, the p-values are very low, so the null hypothesis about constant variance is rejected at the 1%-level, and we can assume that we have non-constant variance, heteroscedasticity, as the residual plot indicated.

One of the assumptions for the BG-test is normally distributed residuals, and this is something we can assume for the regression between Investment and GDP.

Exercise 2: Multiple regression

The data set that our group has decided to base our regression analysis on is taken from http://www.oswego.edu/~kane/econometrics/data.htm

Our dataset consists of the following variables:
crime = total crimes in 1993 / 100,000 population by state (including the District of Columbia)
pov = % of population in state living below the poverty level in 1993
metro = metroplitan population as % of state population
popdens = population per square mile

Source: Statistical Abstract of the United States, 1993.

Our data set consists of the 50 states in USA and D.C. (District of Columbia) which gives us a total number of 51 observations. D.C. is not a state nor is it included in a state, it is a federal district. Our dilemma was whether we should include D.C. in our chosen data set or not. If we exclude D.C. then we will not be able to cover all of the geographic areas of the United States. On the other hand, if it is included, then there is a great risk that D.C. will be an outlier that can distort our results in our regression analysis. Due to the fact that we want to illustrate the number of crimes in the entire country, we found it relevant to include D.C. Below we have three scatter plots, where each plot includes the regression line between the dependent variable and the independent variable for respective graph.

Furthermore, it is important to note that this is a total analysis of the entire population, since our data set consists of every state in USA and not a random sample of states.

1. **Check if there is a linear relationship between the variables.**

In the first plot where we have pov as the explanatory variable and crime as the dependent variable, we can detect a slight positive linear relationship between the variables. Here we can also detect an outlier, an observation which has a high rate of poverty, more than 25% live below the poverty level and has a number of total crimes/ 100,000 population near 12000. The observation which has the next highest total of crimes amounts to approximately 8,000 crimes, about a third less.

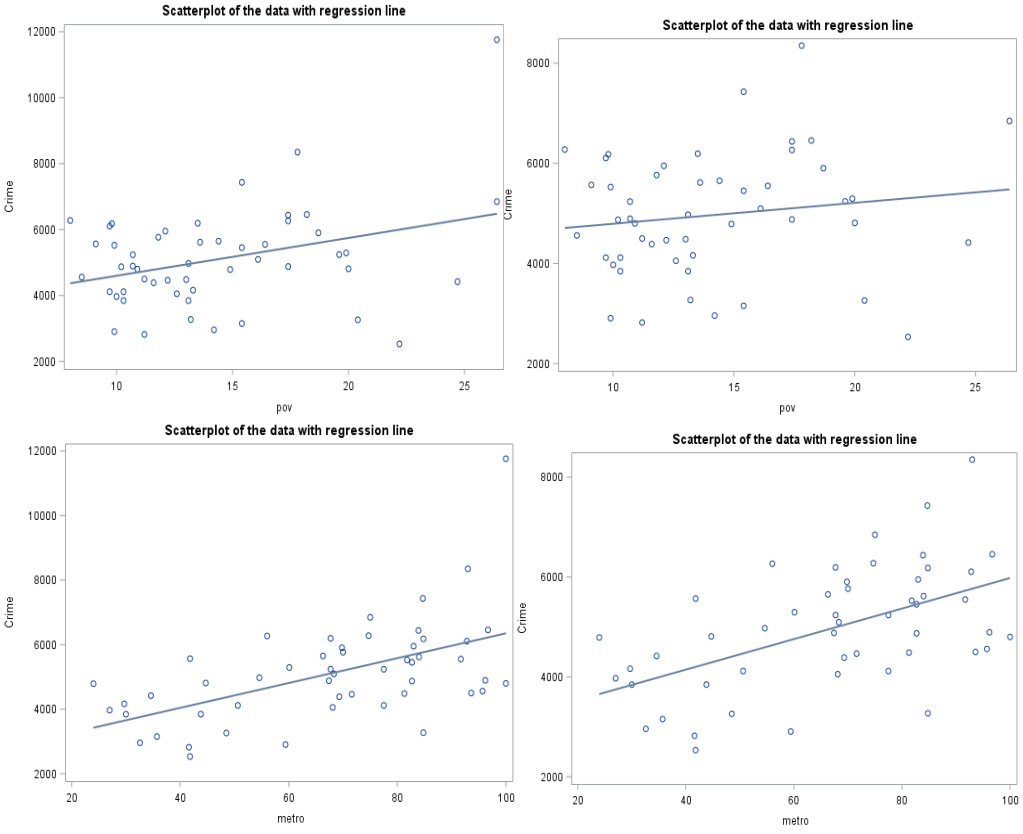
In the second plot graph where the dependent variable crime is plotted against the independent variable metro, we can observe a linear relationship between the variables that is just slightly steeper than the first one. Even here is the outlier noticeable, where we have

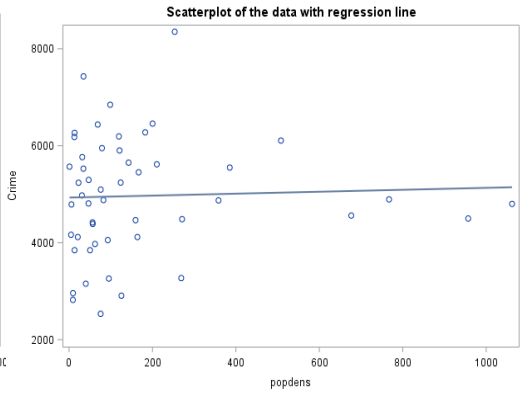
the metropolitan population as a percentage of state population plotted against crime. The observation that stands out from the rest has a percentage near 100% and 12,000 crimes for that year.

In the third and the last plot we have a much steeper regression line between the dependent variable crime and the independent variable population density (population per square mile). The probable reason for this is that the outlier draws the regression line towards its position. The plot shows for the outlier that a population between 9,000 and 10,000 per square mile committed approximately 12,000 crimes in 1993 while rest of the observations lie roughly under 8,000 crimes and a population density of 2,000.

It is now quite evident after having examined the plots above that our intuition about the high risk of D.C. being an outlier in our data set was right from the beginning. D.C. stands out remarkably in comparison to all the states in the data set due to the fact that it is much more densely populated than the rest of the states and has also a very high poverty rate.

We have also conducted a regression on the dependent variable crime for each and every independent variable separately after having removed D.C. from the data set and we could see that D.C. had a significant impact on the regression. Especially when we observed the plot here we regressed the independent variable population density against the predictand crime. The slope of the regression line became much more flat, almost horizontal. As mentioned previously, our main objective was to base our analysis on the entire country, therefore we could not exclude D.C. from our data set.





2. **Use proper transformation to transform your dataset (if needed).**

Transformation is not necessary due to the fact that heteroscedasticity does not exist, which we prove in exercise 2.3.

3. **Check if all the assumptions of OLS hold or not.**

Assumption 1 – Linearity in parameters
Linearity in parameters is what our theoretical model is based upon. Since it will be shown, all of the OLS-assumptions hold, and we can therefore be confident in our model.

Assumption 2 – The values of regressors are fixed or there is zero covariance between ui and X-variables, they are independent
All of the regressors in our model use data from a fixed dataset, therefore this assumption is valid.

Assumption 3 – Zero mean value of disturbance ui.
See assumption 9. Since we assume there is no specification bias in the model, we also assume there are no hidden factors in the disturbance term systematically affecting the mean value of crime. The mean effect of the residuals on the dependent variable therefore is assumed to be zero.

Assumption 4 – Homoskedasticity or constant variance of ui

White's Pagan test:

Heteroscedasticity Test					
Equation	Test	Statistic	DF	Pr > ChiSq	Variables
Crime	White's Test	2.64	9	0.9769	Cross of all vars
	Breusch-Pagan	0.47	3	0.9264	1, pov, metro, popdens

There is no evidence of heteroscedasticity, both the White's test and the Pagan test show a p-value greater than 0.1, so the null hypothesis that there is no heteroscedasticity cannot be rejected.

Assumption 5 – For given X's there is no autocorrelation between the disturbances.

Perhaps the most famous test to detect autocorrelation is the Durbin-Watson test.

$$d = \frac{\sum_{t=2}^{t=n}(\hat{u}_t - \hat{u}_{t-1})}{\sum_{t=1}^{t=n}\hat{u}_t^2}$$

If d_{obs}< d_L (n=50, k'=3) = 1,245, we have evidence of positive autocorrelation at the 5%-level, and if d_u (1.491) < d_{obs}< 4 - d_u (2.509) we do not reject H_0 and conclude there is no evidence of autocorrelation. A value above 4 – d_L (2.755) would indicate negative autocorrelation.

Durbin-Watson Statistics			
Order	DW	Pr < DW	Pr > DW
1	1.9940	0.4890	0.5110
2	1.7727	0.2708	0.7292
3	1.9581	0.5781	0.4219
4	1.9102	0.5589	0.4411

Our reported d-statistics are in the acceptance region, and therefore we have no evidence of autocorrelation. This is a common feature in cross-section data.

Assumption 6 – The number of observations n must be greater than the number of parameters to be estimated

We have 51 observations and 4 parameters to be estimated.

Assumption 7 – There must be sufficient variation in the values of the X variables

The independent variables have sufficient variation, see exercise 2.1.

Assumption 8 – There is no exact collinearity between the X-variables

As we can see on the tables, the VIF-values for all variables is <10 which indicates that there is no obvious multicollinearity between the variables.

Parameter Estimates								
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Tolerance	Variance Inflation
Intercept	Intercept	1	1638.80463	774.68027	2.12	0.0397	.	0
pov	pov	1	83.29978	35.40912	2.35	0.0229	0.85213	1.17353
metro	metro	1	31.17466	7.35529	4.24	0.0001	0.86086	1.16163
popdens	popdens	1	0.45630	0.12795	3.57	0.0008	0.77243	1.29462

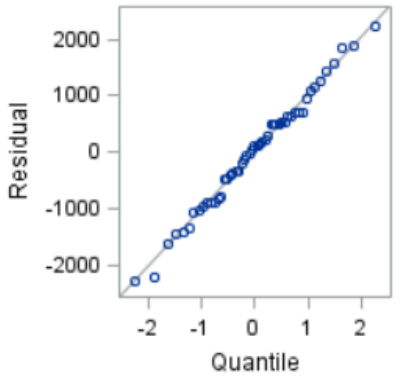
Assumption 9 - The model is correctly specified, so there is no specification bias

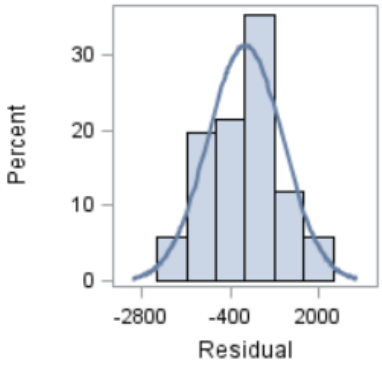
As noted earlier, transformation of the model is not necessary. Our model with three independent variables does not have a very high adjusted R-squared, and of course there could be other variables explaining crime, such as the number of police officers per 100 000 inhabitants, unemployment rate, number of years in imprisonment for a certain kind of crime etc. But, our dataset does not contain any other variables, so within the limitation of our

dataset we can assume that there is no specification bias, however theoretically this may not be true.

Assumption 10 – The stochastic (disturbance) term ui is normally distributed

The normal probability plot and the histogram of the residuals below indicate that the assumption of normally distributed residuals holds. The asymptotically chi-squared distributed Jarque-Bera test also confirms this with a very low p-value.





Miscellaneous Statistics			
Statistic	Value	Prob	Label
Normal Test	78.9239	<.0001	Pr > ChiSq

4. Estimate the model

We get the following estimated model:

$$crime = 1638.80463 + 83.29978pov + 31.17466metro + 0.45630popdens$$

The reported F-value is significant at a high level, suggesting that at least one of the estimated parameters should be included in the model. We can also see that we have significant t-values for all of our three regressors, all of them have a p-value below 0.05.

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	1638.80463	774.68027	2.12	0.0397
pov	pov	1	83.29978	35.40912	2.35	0.0229
metro	metro	1	31.17466	7.35529	4.24	0.0001
popdens	popdens	1	0.45630	0.12795	3.57	0.0008

5. Explain the relationships between the predictors and predictand

On average, holding the remaining explanatory variables constant the interpretation is the following. For every increase of one percentage-point in the poverty rate, crime increases by 83 crimes per 100 000 inhabitants, and as the metropolitan population is bigger by one percentage point (in relation to the total state population), the number of crimes (per 100 000 inhabitants) increase by around 31. An increase of one unit in population per square mile, increase the number of crimes per 100 000 inhabitants with around 0.46 crimes.

6. **Confidence intervals for parameters (or combinations of parameters) of interest.**

On the table below you can see the confidence intervals for all our variables. Since none of the intervals contains zero we can reject the null hypothesis and they are significant. The parameters are significantly different from zero.

An example of how to interpret the intervals is if pov and metro are kept fixed, but metro increases by 1, then we can say with 95% confidence that the metro-coefficient will lie somewhere between 16.38 and 45.97. Which will also be the effect of a unit increase in metro on crime, holding pov and popdens constant.

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	Intercept	1	1638.80463	774.68027	2.12	0.0397	80.34895	3197.26030
pov	pov	1	83.29978	35.40912	2.35	0.0229	12.06583	154.53374
popdens	popdens	1	0.45630	0.12795	3.57	0.0008	0.19890	0.71371
metro	metro	1	31.17466	7.35529	4.24	0.0001	16.37773	45.97159

7. Check for multicollinearity

See exercise 2.3.

8. Check for heteroscedasticity

See exercise 2.3.

Conclusion

The conclusion we arrived at after doing all these test is that the model works quite well. All of the parameters were significant. However, the R-square could be interpreted as quite low but within our limitations we are unable to get it any higher. That's because we only had access to three variables while there could be other variables explaining crime as we stated in exercise 2.3, assumption 9.

Overall, the model gave great results in the performed tests. It does not need a transformation and it fulfills all of the OLS assumptions. There is no heteroscedasticity, multicollinearity and it has a constant variance etc.

All of our three independent variables are, in our analysis, estimated to have a positive connection with crime. Increases in the poverty rate and proportion of metropolitan area of state population, on average, means more crime in any given state. The same goes for the population density. However, removing D.C. from the data set would give another interpretation.

Herman Niclazon 951027-9434
Martin Ottosson 840413-0455
Igor Stokic 730712-0035