Introduction to Applied Statistical Methods

Practical Session 9 Solutions

Please open the SPSS Data file "Practical 9 data.sav" from the Topic 9 Practical Materials folder on KEATS.

Background:

The data file contains crime rate data on 51 US states. Several continuous and categorical variables have been recorded:

- **state** – US state
- **crime** – violent crime rate (per 100,000 people)
- **murder** – murder rate (per 100,000 people)
- **metropol** – percent living in cities
- **white** – percent white people
- **edu** – percent gaining high school education
- **poverty** – percent below the poverty line
- **single** – percent lone parents
- **urban** – a categorised (binary) version of metropol [urban = 1 if metropol >= median (69.8); urban = 0 otherwise)

Task 1

First, identify the type of each variable in the dataset.

- **crime** is a *numerical continuous* variable
- **murder** is a *numerical continuous* variable
- **metropol** is a *numerical continuous* variable
- **edu** is a *numerical continuous* variable
- **poverty** is a *numerical continuous* variable
- **single** is a *numerical continuous* variable
- **urban** is a *categorical binary* variable

Task 2

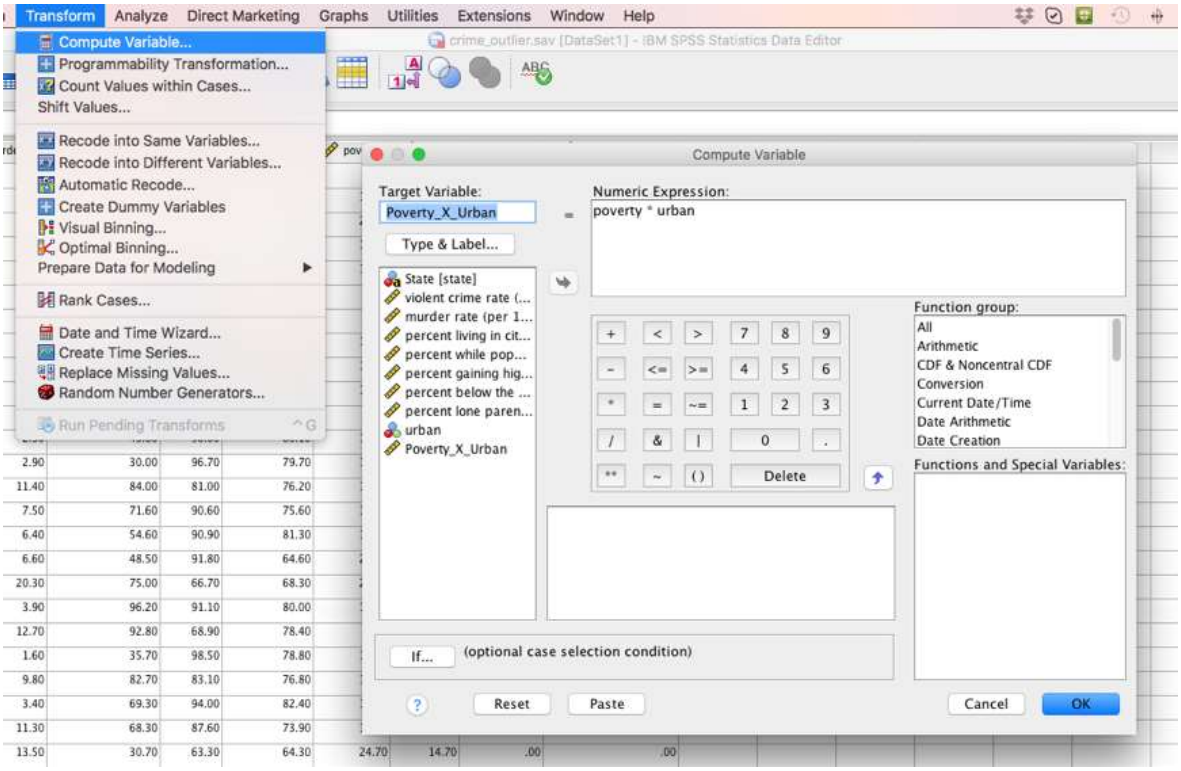
Use the appropriate descriptive indices to identify potential typos and if so, clean the dataset. Use the space below to keep a record of the typos you found and then delete them to create a "clean data set".

Typos found and deleted:
Metropol (-30.70), murder (-9)

Task 3

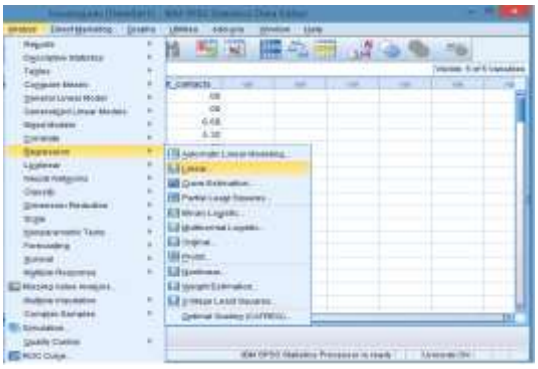
It is hypothesised that urbanicity is a modifier of the effect of poverty on crime. To investigate the effect modification please follow the next steps.

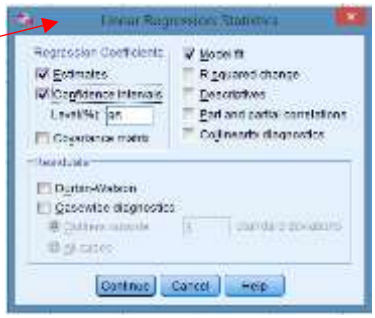
a) Use the appropriate SPSS command to create a cross-product (poverty_X_urban) term by multiplying poverty with urban.



b) Use a multiple linear regression model to assess if urbanicity is an effect modifier of the poverty-crime association.

1. 2. 3.





Output:

Coefficients[a]								
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	95.0% Confidence Interval for B	
		B	Std. Error	Beta			Lower Bound	Upper Bound
1	(Constant)	112.756	219.327		0.514	0.610	-328.473	553.986
	urban	-416.523	284.781	-0.477	-1.463	0.150	-989.427	156.382
	percent below the poverty line	21.798	14.248	0.227	1.53	0.133	-6.866	50.462
	PxU	57.326	18.95	1.005	3.025	0.004	19.203	95.449
[a] Dependent Variable: violent crime rate (per 100,000 people)								

c) What does the regression coefficient of poverty_X_urban tell you? Comment on the statistical significance of the coefficient.

The regression coefficient of poverty_X_urban represents the interaction effect between poverty and urbanicity. The p-value (p=0.004) suggests that interaction effect is statistically significant. This implies that both predictors jointly affect crime rate, but their effects are not independent of each other. Effect of poverty depends on urbanicity and vice-versa.

The estimated coefficient (57.326) can be interpreted as the difference of the effect of poverty on crime rate between highly urbanised (urban=1) and low-urbanised (urban=0) states.

d) Do the coefficients of poverty and urban carry their usual meaning? Interpret their estimated values from the fitted interaction model.

The coefficients of the predictors poverty and urban do not carry their usual interpretations because of the presence of an interaction (cross-product) term involving these predictors.

For the interaction model, the coefficient of poverty can be interpreted as the effect of poverty on crime rate when urban=0. The estimated coefficient (21.798) implies that, in low-urbanised states, one unit increase in poverty leads to 21.798 units increase in crime rate.

Similarly, the coefficient of urban (-416.523) represents the effect of urbanicity on crime rate when poverty=0. This coefficient may not be of great interest in this study as the zero poverty is an unrealistic value.

Interaction term is the difference between low urbanised and high urbanised areas in the the poverty –crime association

e) What is the estimated linear effect of poverty on crime rate for low-urbanised (urban=0) and highly urbanised (urban =1) states?

The general formula for the effect of poverty on crime rate based on the above interaction model is given by:

*Effect of poverty = Coefficient of poverty + Coefficient of interaction term * Urbanicity*

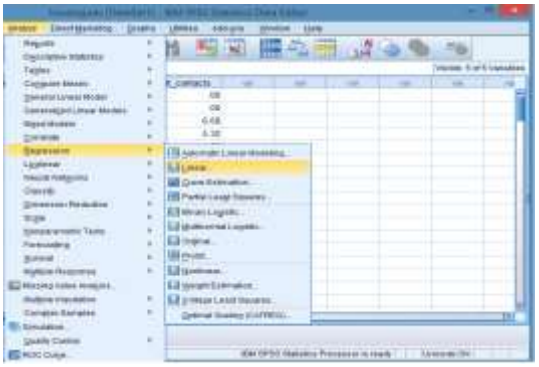
Therefore:
Effect of poverty for low urbanised states = 21.798 + 57.326 x 0 = 21.798
Effect of poverty for highly urbanised states = 21.798 + 57.326 x 1 = 79.124

Task 4

a) For the poverty by urban interaction model in Task 3, calculate the standardised DFBETA and DFFIT measures.

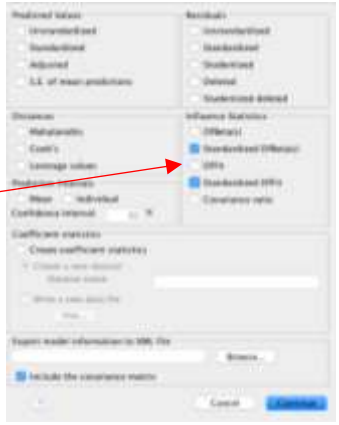
Standardised DFBETA and DFFIT measures can be calculated and saved in the data file by:

1.



2.

3.



Output:

state	SDF_1	SDB0_1	SDB1_1	SDB2_1	SDB3_1
AK	.58993	.53318	-.45897	-.41083	.34359
AL	.23629	-.08321	.12355	.04868	-.09290
AR	.05094	-.02806	.03947	.02238	-.02968
AZ	-.14263	.00000	.00000	.00000	-.03098
CA	-.05531	.00000	.00000	.01443	-.02496
CO	.07714	.00000	.00000	.04175	-.03199
CT	.09121	.00000	.00000	.05351	-.04457
DE	.15619	.00000	.00000	.08239	-.06158
FL	.09175	.00000	.00000	-.02141	.03930
GA	.22880	.12457	-.06680	-.09594	.05023
HI	-.07592	.00000	.00000	-.04535	.03857
IA	-.01174	-.01043	.00854	.00803	-.00642
ID	-.08807	-.05273	.03193	.04081	-.02401
IL	.12581	.00000	.00000	.02847	-.00249
IN	-.12199	.00000	.00000	-.04889	.02485
KS	.07225	.04427	-.02680	-.03409	.02015
KY	-.11505	.06899	-.09171	-.05313	.06895
LA	-2.11542	.00000	.00000	1.03939	-1.30583
MA	.21299	.00000	.00000	.10663	-.07576
MD	.50023	.00000	.00000	.27464	-.21380
ME	-.21806	-.16982	.15141	.14542	-.11384
MI	-.06732	.00000	.00000	.00018	-.01888
MN	-.03374	-.02694	.02029	.02075	-.01526
MO	.20255	-.00377	.05947	.00290	-.04472
MS	-.48943	.37921	-.44984	-.29206	.33823
MT	-.17039	-.04499	-.00437	.03465	.00328
NC	.17413	.08336	-.01617	-.04880	.01216
ND	-.26549	-.21249	.16513	.16365	-.12416
NE	.00180	.00159	-.00131	-.00123	.00098
NH	-.21083	-.19120	.15929	.14726	-.11977
NJ	.05393	.00000	.00000	.02634	-.01826
NM	.36440	-.09747	.19053	.07507	-.14328
NV	.36721	.00000	.00000	.20027	-.15466
NY	.06210	.00000	.00000	-.00700	.02000
OH	-.16016	.00000	.00000	-.04489	.01529
OK	.10041	-.05650	.07720	.04351	-.05805
OR	-.09227	.00000	.00000	-.03895	.02303
PA	-.21957	.00000	.00000	-.08046	.01643
RI	-.13868	.00000	.00000	-.06493	.04316
SC	-.15462	.00000	.00000	.04605	-.07362
SD	-.14841	-.06037	.02063	.04649	-.01551
TN	.24731	-.13296	.16532	.10240	-.13934
TX	-.27155	.00000	.00000	.05511	-.10926
UT	-.19626	.00000	.00000	-.09026	.06981
VA	-.08287	.00000	.00000	-.04553	.03542
VT	-.23678	-.21370	.17730	.16458	-.13331
WA	-.09957	.00000	.00000	-.03883	.02124
WI	-.09693	-.06587	.04408	.05073	-.03314
WV	-.62295	.43400	-.54042	-.35425	.40634
WY	-.08488	-.04919	.02821	.03799	-.02121
DC	4.02152	.00000	.00000	-1.97692	2.48245

b) Identify any influential observations. What do the standardised DFBETA and DFFIT measures tell you? Which states have strong influence on the fitted model? Which state is the most influential data point?

The data for the states LA and DC have absolute standardised DFFIT and DFBETA measures exceeding 1, indicating that these two states have strong influence on the fitted model. The state DC (District of Columbia) has the highest absolute DFFIT and DFBETA measures suggesting that the data for this state has the strongest influence. Such data points need to be further scrutinised for validity and may need to be removed from the analysis as they can lead to a distorted estimate of the true population relationship.