21HT-1ST817 Statistical data processing Assignment1 (Yuka Tatsumi)

a. Compare summary statistics of all the variables in the model and provide a brief interpretation.

According to Table 1, there is a similarity between SCHOOL and EXPER regarding variance and standard deviation. There is also a similarity between UNION, MAR, BLACK, HISP regarding variance and standard deviation. However, there is a slight difference in variance and standard deviation between the 2 groups.

Valid N(listwise)is 545, which means there are no missing values.

Table 1
Descriptive Statistics of all the variables in the original model

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation	Variance
BLACK	545	0	1	,12	,320	,102
EXPER	545	7	18	10,01	1,655	2,739
HISP	545	0	1	,16	,363	,132
MAR	545	0	1	,61	,487	,237
UNION	545	0	1	,26	,440	,194
SCHOOL	545	3	16	11,77	1,748	3,054
WAGE	545	-,191008716822	3,34316444397 0	1,86647923080 199	,466890021739 279	,218
Valid N (listwise)	545					

(Syntax)

```
GET DATA
  /TYPE=XLS
  /FILE='C:\Users\y_tat\Downloads\statsitics_uni\lin\Assignment1\Data (males).xls'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.
USE ALL.
COMPUTE filter_$=(YEAR = 1987).
VARIABLE LABELS filter_$ 'YEAR = 1987 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
DESCRIPTIVES VARIABLES=BLACK EXPER HISP MAR UNION SCHOOL WAGE
  /STATISTICS=MEAN STDDEV VARIANCE MIN MAX.
```

b. Estimate the parameters by OLS. Report and interpret the estimation results, including the R2. Pay attention to economic interpretation as well as statistical significance.

According to Unstandardized Coefficients in Table 2 below from SPSS, parameter can be estimated as follows.

logwagei = 0.776 + 0.088 school + -0.02 experi + 0.117 union + 0.091mari + -0.198blacki + 0.043 hispi + ɛi

However, results of t test and 95% confidential Interval for unstandardized coefficients show that unstandardized coefficients for EXPER and HISP are not statistically significant (95% Confidence).

The coefficients below also shows that UNION (positive impact) and BLACK (negative impact) have stronger coefficients than other variables. It means that economically speaking, these two factors have stronger influence on wages than other factors in the model.

Table 2
Coefficients (original model)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	,776	,261		2,969	,003	,262	1,289
	SCHOOL	,088	,013	,331	6,689	<,001	,062	,114
	EXPER	-,002	,014	-,007	-,139	,890	-,029	,025
	UNION	,117	,043	,111	2,733	,006	,033	,202
	MAR	,091	,039	,095	2,342	,020	,015	,167
	BLACK	-,198	,061	-,136	-3,253	,001	-,318	-,079
	HISP	,043	,053	,033	,806	,420	-,061	,147

a. Dependent Variable: WAGE

According to Table 3 below from SPSS, R square and Adjusted R square are not so high and implies that there is a room for improvement.

Table 3
R square and Adjusted R square of the original model

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,394[a]	,155	,145	,431592250878701

a. Predictors: (Constant), HISP, UNION, MAR, EXPER, BLACK, SCHOOL

b. Dependent Variable: WAGE

(Syntax: Original Model)

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT WAGE
  /METHOD=ENTER SCHOOL EXPER UNION MAR BLACK HISP
  /SCATTERPLOT=(*ZRESID ,WAGE)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /SAVE ZPRED ZRESID.
```

c. Test on the basis of the results in b, test the null hypothesis that being a union member, ceteris paribus, affects a persons expected wage by 5%. Also, test the joint hypothesis that race does not affect wages. In each case formulate the null and alternative hypotheses and present the test statistic.

Regarding a test to investigate whether the union member affects a person's expected wage by 5% or not, null and alternative hypothesis can be formulated as below.

> Null hypothesis: being a union member, ceteris paribus, affects a persons expected wage by 5%.($\beta 4 = 0.05$)
>
> Alternative Hypothesis: being a union member does not affect a persons expected wage by 5% ($\beta 4 \neq 0.05$)

The results regarding coefficients (see Table 4) from question b, 95,0% Confidence Interval for Unstandardized Coefficients for Union are from 0,033 to 0,202, which includes 5% (0,05) and the result is statistically significant(sig<0.05). Therefore, null hypothesis cannot be rejected.

Table 4
Coefficients (original model)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	,776	,261		2,969	,003	,262	1,289
	SCHOOL	,088	,013	,331	6,689	<,001	,062	,114
	EXPER	-,002	,014	-,007	-,139	,890	-,029	,025
	UNION	,117	,043	,111	2,733	,006	,033	,202
	MAR	,091	,039	,095	2,342	,020	,015	,167
	BLACK	-,198	,061	-,136	-3,253	,001	-,318	-,079
	HISP	,043	,053	,033	,806	,420	-,061	,147

a. Dependent Variable: WAGE

Regarding the joint hypothesis that race does not affect wages, null and alternative hypothesis can be stated as below

> Null Hypothesis: race does not affect wages (β6=0 and β7=0): restricted model
>
> Alternative Hypothesis: Race affect wages (β6≠0 or β7≠0): unrestricted model

According to the Table 5 below, unrestricted Residual Sum of Squares(URSS) is 100,214 and Adjusted R Square is 0,145 while Restricted Residual Sum of Squares(RRSS) is 102,565 and Adjusted R Square 0,129(see Table 6). F Statistics for testing the hypothesis is $F = ((RRSS - URSS)/q)/ (URSS/(N - k))$. Therefore, F= ((102,565-100,214)/2)/(100,214/(545-6)) = 6,322415032. Critical value of $F_{1\text{-}0.05,\ (2,539)}$ is 3.01244425, which is smaller than 6,322415032. Therefore, we can reject Null Hypothesis.

Table 5
ANOVA - Unrestricted model

ANOVA[a]

Model		4umo f Squares	df	Mean Square	F	Sig.
1	Regression	18,370	6	3,062	16,437	,000[b]
	Residual	100,214	538	,186		
	Total	118,585	544			

a. Dependent Variable: WAGE

b. Predictors: (Contant), MAR, EXPER, UNION, HISP, BLACK, SCHOOL

Table 6
ANOVA - Restricted model

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	16,020	4	4,005	21,086	,000[b]
	Residual	102,565	540	,190		
	Total	118,585	544			

a. Dependent Variable: WAGE

b. Predictors: (Constant), MAR, EXPER, UNION, SCHOOL

Syntax (restricted model)

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
```

```
/MISSING LISTWISE
/STATISTICS COEFF OUTS CI(95) R ANOVA
/CRITERIA=PIN(.05) POUT(.10)
/NOORIGIN
/DEPENDENT WAGE
/METHOD=ENTER SCHOOL EXPER UNION MAR
/SCATTERPLOT=(*ZRESID ,WAGE)
/RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

d. Consider a more general model that includes exper^2. Compare this model with the model given above using R2, adjusted R2 and t-test. What is your conclusion?

If exper^2 is added to the original model, the model is written as follows according to Table 7 below. Coefficients for Hispi β7(0.033) is not statistically significant. (Sig 0,530)

logwagei = β1(=1.892) + β2(=0.095) school + β3(=-0.225)exper + β4(=0.129)unioni + β5(=0.091)mari + β6(=-0.194)blacki + β7(=0.033)hispi + β8(=0.10)exper^2+ εi

Table 7
Coefficients (model that includes exper^2)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			95,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	1,892	,509		3,718	<,001	,892	2,892
	SCHOOL	,095	,013	,357	7,100	<,001	,069	,122
	EXPER	-,225	,089	-,798	-2,542	,011	-,399	-,051
	UNION	,129	,043	,122	3,006	,003	,045	,214
	MAR	,091	,039	,095	2,361	,019	,015	,167
	EXPER2	,010	,004	,813	2,551	,011	,002	,018
	HISP	,033	,053	,026	,628	,530	-,071	,137
	BLACK	-,194	,061	-,133	-3,191	,002	-,313	-,074

a. Dependent Variable: WAGE

Regarding R square(see Table 8), both R square and Adjusted R square are better than the original model, which means the model is better fit than original model.

Table 8
R square and Adjusted R square (model that includes exper ^{2})

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,406[a]	,165	,154	.4293989864

a. Predictors: (Constant), BLACK, SCHOOL, MAR, UNION, HISP, EXPER, EXPER2

b. Dependent Variable: WAGE

(Syntax)

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT WAGE
  /METHOD=ENTER SCHOOL EXPER UNION MAR EXPER2 HISP BLACK
  /SCATTERPLOT=(*ZRESID ,WAGE)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

e. Save the OLS residuals from the original model. Run a regression where you try to explain the residuals from the explanatory variables in the original regression. What do you find? Explain.

Residual statistics in Table 9 shows that maximum of Standardized residual is above 3, which indicates probabilities of outlier and needs to be checked.

Table 9
Residuals Statistics (original model)

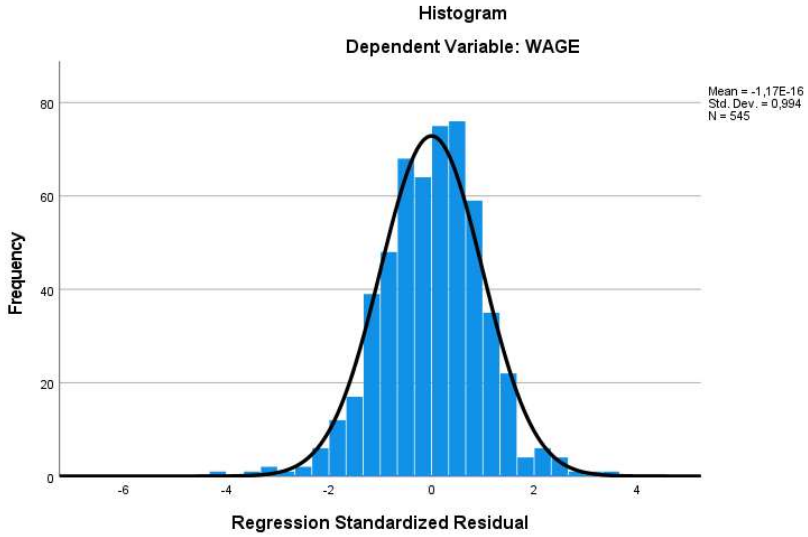
Residuals Statistics[a]

	Minimum	Maximum	Mean	Std. Deviation	N
Predicted Value	1,04891502857208	2,29388713836670	1,86647923080199	,183763137072736	545
Residual	-1,74986279010 7727	1,52272880077 3621	,000000000000	,429205547323 072	545
Std. Predicted Value	-4,449	2,326	,000	1,000	545
Std. Residual	-4,054	3,528	,000	,994	545

a. Dependent Variable: WAGE

According to Histogram below (Figure1), it can be said that residuals are normally distributed.

Figure 1
Regression Standardized Residuals Histogram (original model)

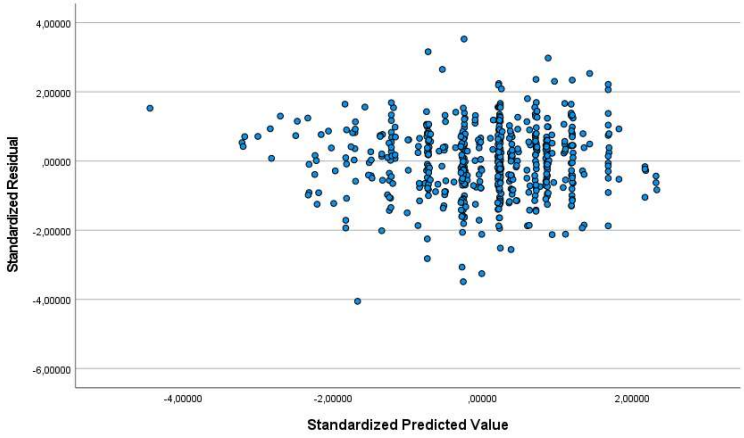



```
(Syntax)

REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT WAGE
  /METHOD=ENTER SCHOOL EXPER UNION MAR BLACK HISP
  /SCATTERPLOT=(*ZRESID ,WAGE)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /SAVE ZPRED ZRESID.
```

Looking at scatter plot below (Regression Standardized Residual vs Standardized Predicted Value), plots are not completely evenly spread (more like U form) and more plots are seen in the right-hand side of the graph. This means that heteroskedasticity is detected in the model.

Figure 2
Scatter plot (Standardised Residuals vs Standardised Predicted Value in original model)



(Syntax)

```
GRAPH
  /SCATTERPLOT(BIVAR)=ZPR_1 WITH ZRE_1
  /MISSING=LISTWISE.
```

f. Extend the model to investigate whether black union members benefit more from union membership than non-black union members. Estimate the extended model and test the hypothesis

The following model can be investigated with the hypothesis below.

[Model]
logwagei = β1 + β2school + β3experi + β4union + β5mari + β6black + β7hispi + β8(unbla*)+ εi
(*unbla = union * black)

[Hypothesis]

Null Hypothesis: black union members does not benefit more from union membership than non-black union members β7 =0

Alternative Hypothesis: black union members benefit more from union membership than non-black union members β7>0

Adjusted R stayed almost same but with small improvement. (See Table 10)

Table 10
R square and Adjusted R Square (Model where Exper2 and Black * UNION are added to the original model)

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	,409[a]	,167	,155	.4291883337

a. Predictors: (Constant), UNBLA, SCHOOL, MAR, HISP, UNION, EXPER, BLACK, EXPER2

b. Dependent Variable: WAGE

Unstandardized Coefficients for UNBLA(β7) is 0.145. But this result is not statistically significant with 95% Confidence (one tailed test / Sig 0,217). (See Table 11) Therefore, Alternative Hypothesis (black union members benefit more from union membership than non-black union members β7 >0) can be rejected.

Table 11
Coefficients (Model where Exper2 and Black * UNION are added to the original model)

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients			90,0% Confidence Interval for B	
		B	Std. Error	Beta	t	Sig.	Lower Bound	Upper Bound
1	(Constant)	1,886	,509		3,708	<,001	1,048	2,725
	BLACK	-,259	,081	-,178	-3,217	,001	-,392	-,127
	EXPER	-,223	,088	-,792	-2,524	,012	-,369	-,078
	EXPER2	,010	,004	,807	2,532	,012	,003	,016
	HISP	,035	,053	,027	,653	,514	-,053	,122
	MAR	,091	,039	,095	2,347	,019	,027	,154
	SCHOOL	,095	,013	,357	7,117	<,001	,073	,118
	UNION	,107	,047	,101	2,283	,023	,030	,183
	UNBLA	,145	,118	,072	1,236	,217	-,048	,339

a. Dependent Variable: WAGE

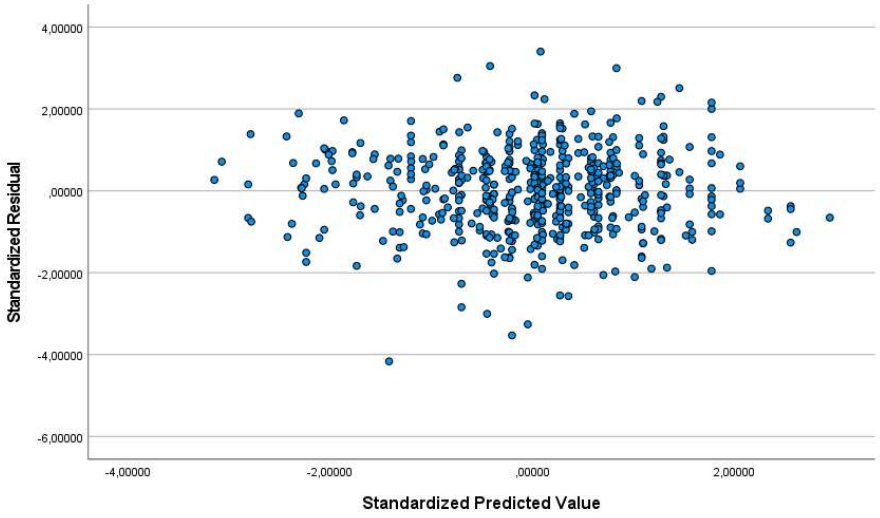
(Syntax)

```
COMPUTE UNBLA= UNION* BLACK.
EXECUTE.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(90) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  NOORIGIN
  /DEPENDENT WAGE
  /METHOD=ENTER BLACK EXPER EXPER2 HISP MAR SCHOOL UNION UNBLA
  /SCATTERPLOT=(*ZPRED ,WAGE)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /SAVE ZPRED ZRESID.
```

g. Make plots that can be used to investigate heteroskedasticity.

According to the scatter plot (Figure 3), standardized residual is more evenly spread than original model. But it still stays slightly U form and plots are gathered in the center. It means that heteroskedasticity is still detected in the model.

Figure 3
Scatter plot (Standardised Residuals vs Standardised Predicted Value in the model where Exper2 and Black * UNION are added to the original model)



(Syntax)

```
GRAPH
  /SCATTERPLOT(BIVAR)=ZPR_1 WITH ZRE_1
  /MISSING=LISTWISE.
```

Appendix

Select the data for the year 1987. Get descriptive statistics

```
GET DATA
  /TYPE=XLS
  /FILE='C:\Users\y_tat\Downloads\statsitics_uni\lin\Assignment1\Data (males).xls'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.
USE ALL.
COMPUTE filter_$=(YEAR = 1987).
VARIABLE LABELS filter_$ 'YEAR = 1987 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
DESCRIPTIVES VARIABLES=BLACK EXPER HISP MAR UNION SCHOOL WAGE
  /STATISTICS=MEAN STDDEV VARIANCE MIN MAX.
```

Run the Regression with original model. Save the residual

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT WAGE
  /METHOD=ENTER SCHOOL EXPER UNION MAR BLACK HISP
  /SCATTERPLOT=(*ZRESID ,WAGE)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /SAVE ZPRED ZRESID.
```

Run the Regression with the model where black and hisp are removed from the original model.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT WAGE
  /METHOD=ENTER SCHOOL EXPER UNION MAR
  /SCATTERPLOT=(*ZRESID ,WAGE)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

Run the Regression with the model where EXPER2 is added to original model.

```
REGRESSION
  /DESCRIPTIVES MEAN STDDEV CORR SIG N
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(95) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT WAGE
  /METHOD=ENTER SCHOOL EXPER UNION MAR EXPER2 HISP BLACK
  /SCATTERPLOT=(*ZRESID ,WAGE)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID).
```

Make plots with standards residual and standardized predicted value to investigate heteroskedasticity in original model.

```
GRAPH
  /SCATTERPLOT(BIVAR)=ZPR_1 WITH ZRE_1
  /MISSING=LISTWISE.
```

#Compute Black * UNION and run the Regression with the model where Exper2 and Black * UNION are added to the original model.

```
GET DATA
  /TYPE=XLS
  /FILE='C:\Users\y_tat\Downloads\statsitics_uni\lin\Assignment1\Data (males).xls'
  /SHEET=name 'Sheet1'
  /CELLRANGE=FULL
  /READNAMES=ON
  /DATATYPEMIN PERCENTAGE=95.0.
EXECUTE.
DATASET NAME DataSet1 WINDOW=FRONT.
USE ALL.
COMPUTE filter_$=(YEAR = 1987).
VARIABLE LABELS filter_$ 'YEAR = 1987 (FILTER)'.
VALUE LABELS filter_$ 0 'Not Selected' 1 'Selected'.
FORMATS filter_$ (f1.0).
FILTER BY filter_$.
EXECUTE.
COMPUTE UNBLA= UNION* BLACK.
EXECUTE.
REGRESSION
  /MISSING LISTWISE
  /STATISTICS COEFF OUTS CI(90) R ANOVA
  /CRITERIA=PIN(.05) POUT(.10)
  /NOORIGIN
  /DEPENDENT WAGE
  /METHOD=ENTER BLACK EXPER EXPER2 HISP MAR SCHOOL UNION UNBLA
  /SCATTERPLOT=(*ZPRED ,WAGE)
  /RESIDUALS HISTOGRAM(ZRESID) NORMPROB(ZRESID)
  /SAVE ZPRED ZRESID.
```

Make plots with standards residual and standardized predicted value to investigate heteroskedasticity in the model f (the model where Exper2 and Black * UNION are added to the original model)

```
GRAPH
  /SCATTERPLOT(BIVAR)=ZPR_1 WITH ZRE_1
  /MISSING=LISTWISE.
```