Analyzing the Settlement Patterns of North and South Koreans in Germany: A Statistical Approach

SEUNGHYUN KIM
seunghyunkim821@gmail.com

Table of Contents

[1] Topic: Statistical Analysis of the North Korean Population Residing in Germany

Germany is the 10th most populated country by South Koreans, with a total of 47,428 Koreans residing in Germany as of 2021. Occasionally, one may come across North Korean individuals while living in Germany, albeit infrequently. Therefore, to assess the population of North Koreans residing in Germany and compare it with the South Korean population, data from the Federal Statistical Office of Germany was utilized for conducting various hypothesis tests and analysis.

[2] Statistical Data Used

Original Source Data:

Statistic

This is statistical data classified by gender, residing state, and years of residence for the South Korean and North Korean populations in Germany. The data includes information for a total of five years: 1998, 2004, 2010, 2016, and 2022.

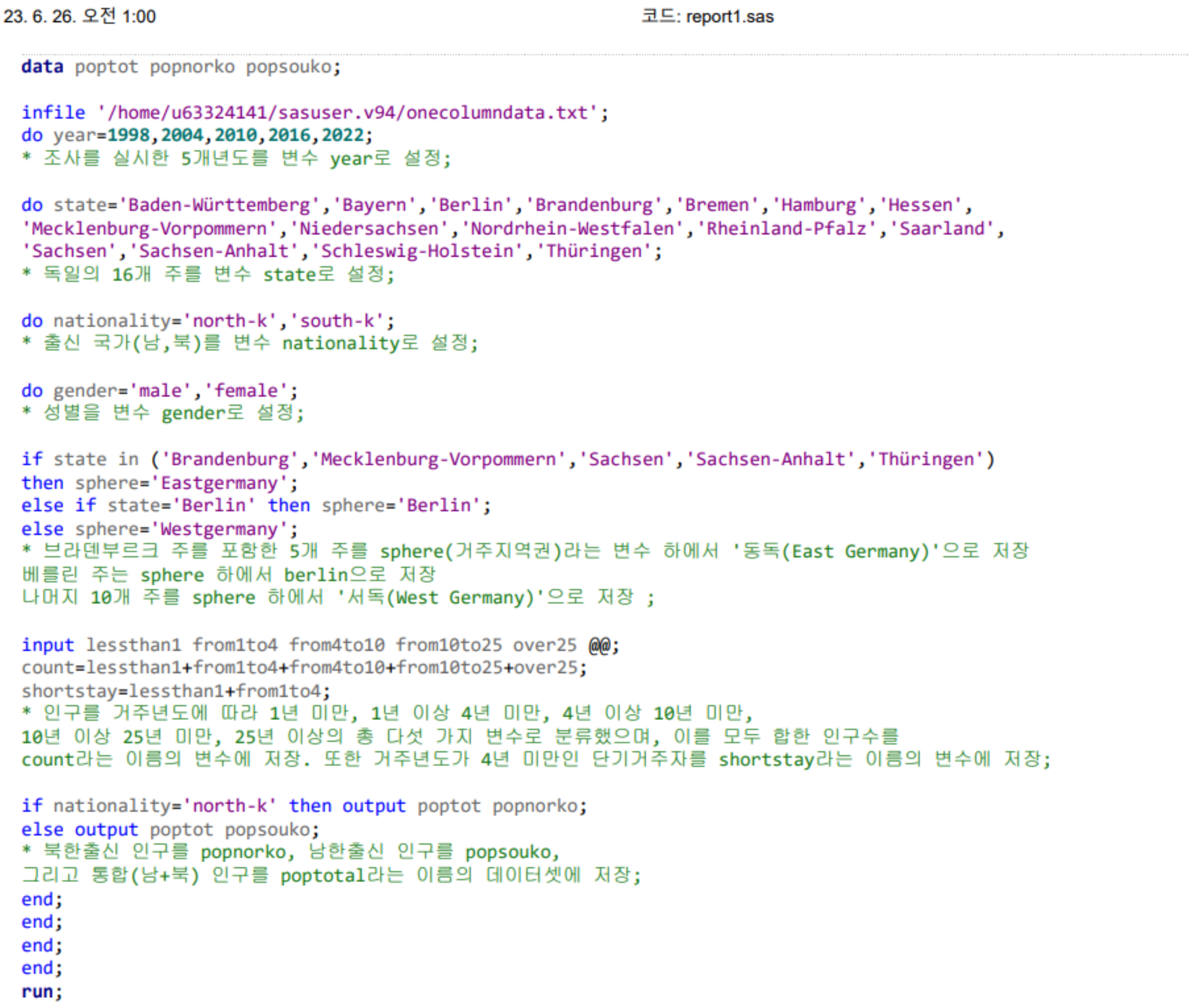
(from: Federal Statistical Office Germany - GENESIS-Online: Database of the <br/>Federal Statistical Office of Germany (destatis.de))

Translated (into Korean) and Preprocessed Data:

Translated and Preprocessed Data

[3] Analysis and Conclusions

SAS code that creates the dataset:

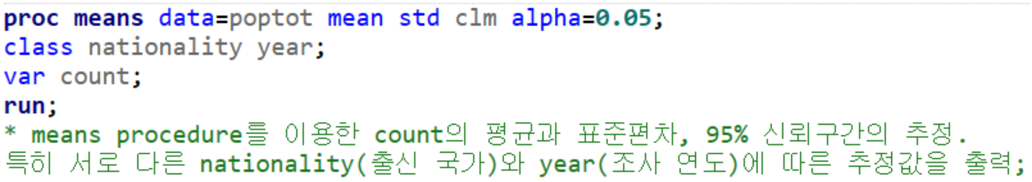


23. 6. 26. 오전 1:00 코드: report1.sas

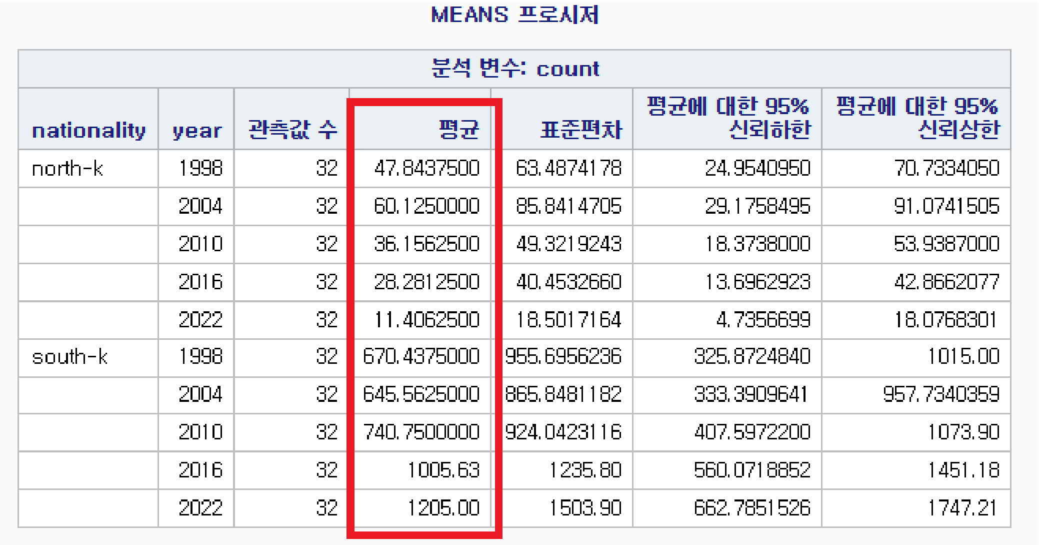
```
data poptot popnorko popsouko;

infile '/home/u63324141/sasuser.v94/onecolumndata.txt';
do year=1998,2004,2010,2016,2022;
* 조사를 실시한 5개년도를 변수 year로 설정;

do state='Baden-Württemberg','Bayern','Berlin','Brandenburg','Bremen','Hamburg','Hessen',
'Mecklenburg-Vorpommern','Niedersachsen','Nordrhein-Westfalen','Rheinland-Pfalz','Saarland',
'Sachsen','Sachsen-Anhalt','Schleswig-Holstein','Thüringen';
* 독일의 16개 주를 변수 state로 설정;

do nationality='north-k','south-k';
* 출신 국가(남,북)를 변수 nationality로 설정;

do gender='male','female';
* 성별을 변수 gender로 설정;

if state in ('Brandenburg','Mecklenburg-Vorpommern','Sachsen','Sachsen-Anhalt','Thüringen')
then sphere='Eastgermany';
else if state='Berlin' then sphere='Berlin';
else sphere='Westgermany';
* 브라덴부르크 주를 포함한 5개 주를 sphere(거주지역권)라는 변수 하에서 '동독(East Germany)'으로 저장
베를린 주는 sphere 하에서 berlin으로 저장
나머지 10개 주를 sphere 하에서 '서독(West Germany)'으로 저장 ;

input lessthan1 from1to4 from4to10 from10to25 over25 @@;
count=lessthan1+from1to4+from4to10+from10to25+over25;
shortstay=lessthan1+from1to4;
* 인구를 거주년도에 따라 1년 미만, 1년 이상 4년 미만, 4년 이상 10년 미만,
10년 이상 25년 미만, 25년 이상의 총 다섯 가지 변수로 분류했으며, 이를 모두 합한 인구수를
count라는 이름의 변수에 저장. 또한 거주년도가 4년 미만인 단기거주자를 shortstay라는 이름의 변수에 저장;

if nationality='north-k' then output poptot popnorko;
else output poptot popsouko;
* 북한출신 인구를 popnorko, 남한출신 인구를 popsouko,
그리고 통합(남+북) 인구를 poptotal라는 이름의 데이터셋에 저장;
end;
end;
end;
end;
run;
```

<1> Point Estimation

(1) Estimation of Population Mean



```
proc means data=poptot mean std clm alpha=0.05;
class nationality year;
var count;
run;
* means procedure를 이용한 count의 평균과 표준편차, 95% 신뢰구간의 추정.
특히 서로 다른 nationality(출신 국가)와 year(조사 연도)에 따른 추정값을 출력;
```



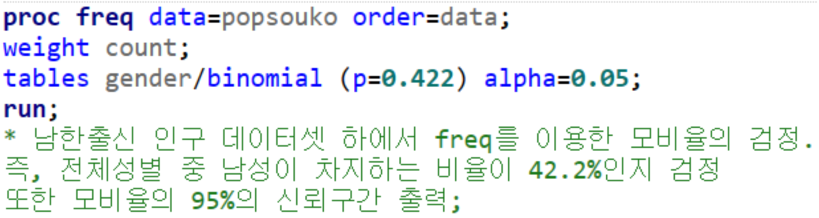
MEANS 프로시저

분석 변수: count						
nationality	year	관측값 수	평균	표준편차	평균에 대한 95% 신뢰하한	평균에 대한 95% 신뢰상한
north-k	1998	32	47.8437500	63.4874178	24.9540950	70.7334050
	2004	32	60.1250000	85.8414705	29.1758495	91.0741505
	2010	32	36.1562500	49.3219243	18.3738000	53.9387000
	2016	32	28.2812500	40.4532660	13.6962923	42.8662077
	2022	32	11.4062500	18.5017164	4.7356699	18.0768301
south-k	1998	32	670.4375000	955.6956236	325.8724840	1015.00
	2004	32	645.5625000	865.8481182	333.3909641	957.7340359
	2010	32	740.7500000	924.0423116	407.5972200	1073.90
	2016	32	1005.63	1235.80	560.0718852	1451.18
	2022	32	1205.00	1503.90	662.7851526	1747.21

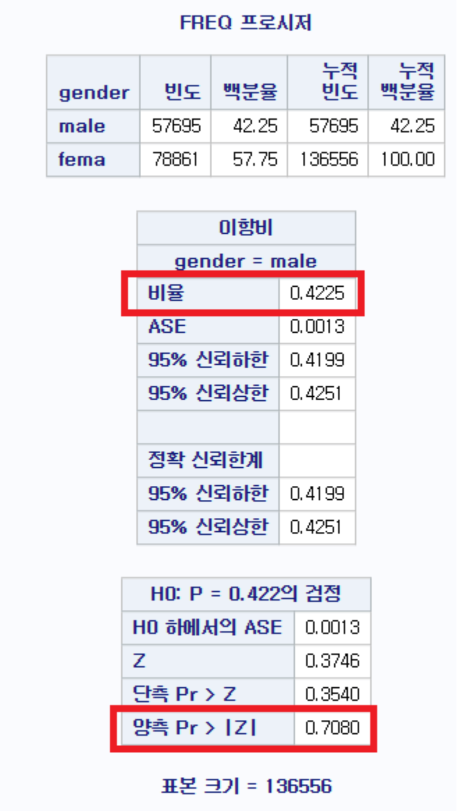
This is a table generated using the means procedure, displaying the mean and standard deviation of the 'count' variable, as well as the lower and upper bounds of the 95% confidence interval.

Examining the mean values of the 'count' variable, highlighted in red, reveals that the average population of North Korean origin in Germany generally decreases over time, while, in contrast, the average population of South Korean origin increases overall.

(2) Estimation of population proportion (+Hypothesis testing of population proportion)



```
proc freq data=popsouko order=data;
weight count;
tables gender/binomial (p=0.422) alpha=0.05;
run;
* 남한출신 인구 데이터셋 하에서 freq를 이용한 모비율의 검정.
즉, 전체성별 중 남성이 차지하는 비율이 42.2%인지 검정
또한 모비율의 95%의 신뢰구간 출력;
```



FREQ 프로시저

gender	빈도	백분율	누적 빈도	누적 백분율
male	57695	42.25	57695	42.25
fema	78861	57.75	136556	100.00

이항비	
gender = male	
비율	0.4225
ASE	0.0013
95% 신뢰하한	0.4199
95% 신뢰상한	0.4251
정확 신뢰한계	
95% 신뢰하한	0.4199
95% 신뢰상한	0.4251

H0: P = 0.422의 검정	
H0 하에서의 ASE	0.0013
Z	0.3746
단측 Pr > Z	0.3540
양측 Pr > |Z|	0.7080

표본 크기 = 136556

The results table indicates that the estimated population proportion is 0.4225.

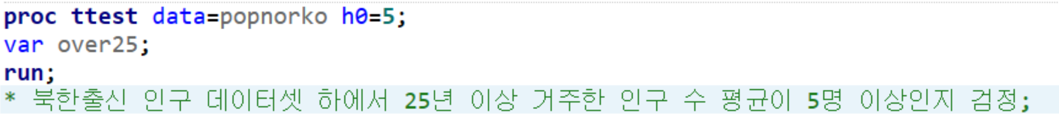
For the hypothesis testing:

H0: The proportion of males in the population of South Korean origin is 42.2%.
H1: It is not 42.2%.

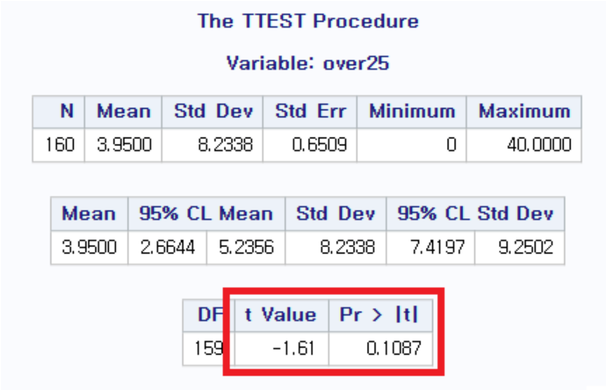
In a two-tailed test, the p-value, highlighted in red as 0.7080, is greater than the typical significance level, leading to a failure to reject the null hypothesis H0. Therefore, it can be concluded that the proportion of males is 42.2%. In other words, among South Korean residents in Germany, there are more females than males.

<2> Hypothesis Testing

(1) Hypothesis Testing of Population Mean



```
proc ttest data=popnorko h0=5;
var over25;
run;
* 북한출신 인구 데이터셋 하에서 25년 이상 거주한 인구 수 평균이 5명 이상인지 검정;
```



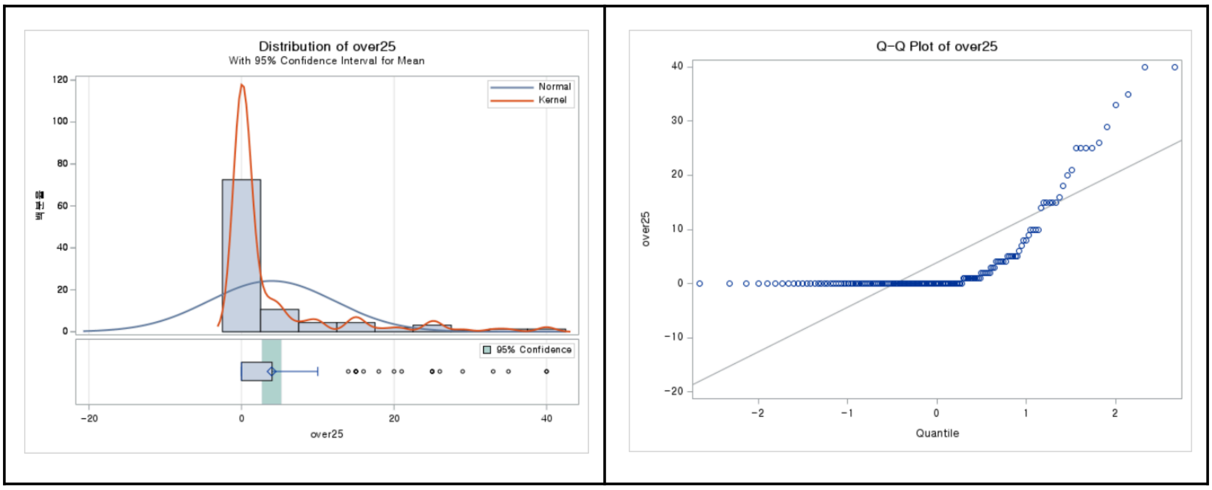
The TTEST Procedure

Variable: over25

N	Mean	Std Dev	Std Err	Minimum	Maximum
160	3.9500	8.2338	0.6509	0	40.0000

Mean	95% CL Mean		Std Dev	95% CL Std Dev	
3.9500	2.6644	5.2356	8.2338	7.4197	9.2502

DF	t Value	Pr > |t|
159	-1.61	0.1087

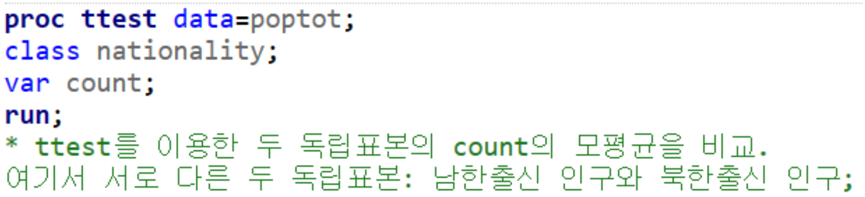


For the hypothesis testing:

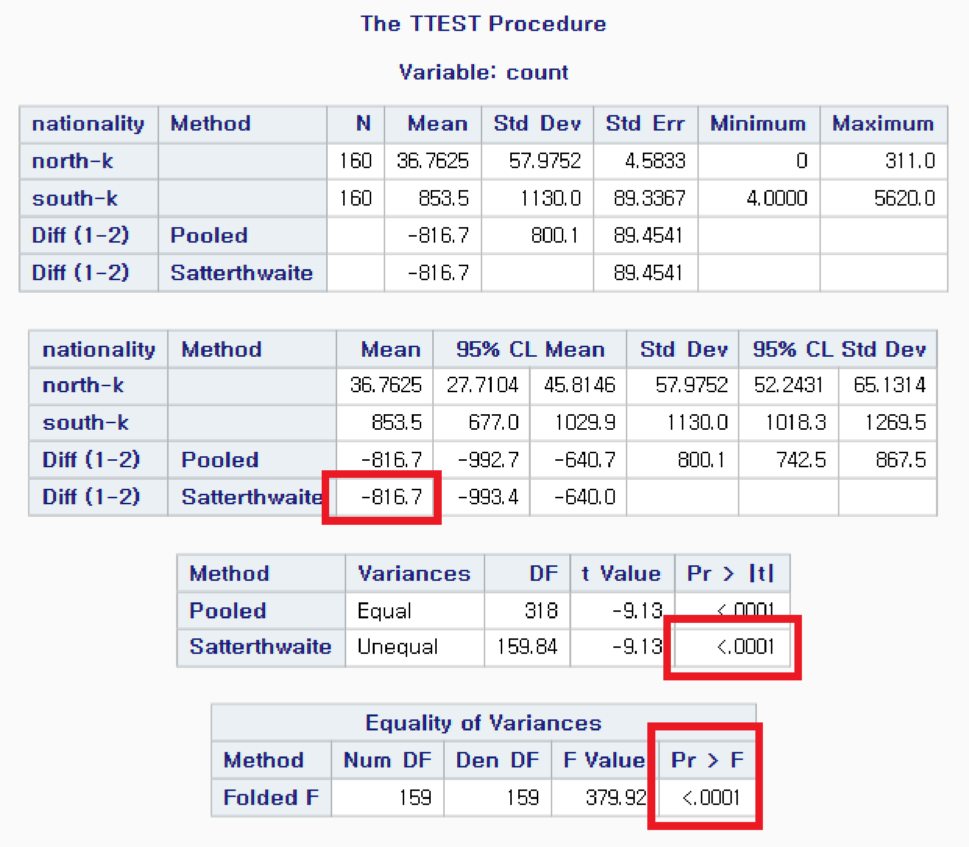
H0: The mean population of North Korean residents who have lived in Germany for 25 years or more is at least 5.
H1: It is less than 5.

Considering that the t-value is negative and that the two-tailed test's p-value is 0.1087, the one-tailed left test's p-value is approximately 0.054. In other words, with a significance level of 6% or higher, the null hypothesis H0 can be rejected. Thus, it can be concluded that the mean population of North Korean residents in Germany who have lived in the country for 25 years or more is less than 5.

(2) Independent Samples' Population Mean Test



```
proc ttest data=poptot;
class nationality;
var count;
run;
* ttest를 이용한 두 독립표본의 count의 모평균을 비교.
여기서 서로 다른 두 독립표본: 남한출신 인구와 북한출신 인구;
```



The TTEST Procedure

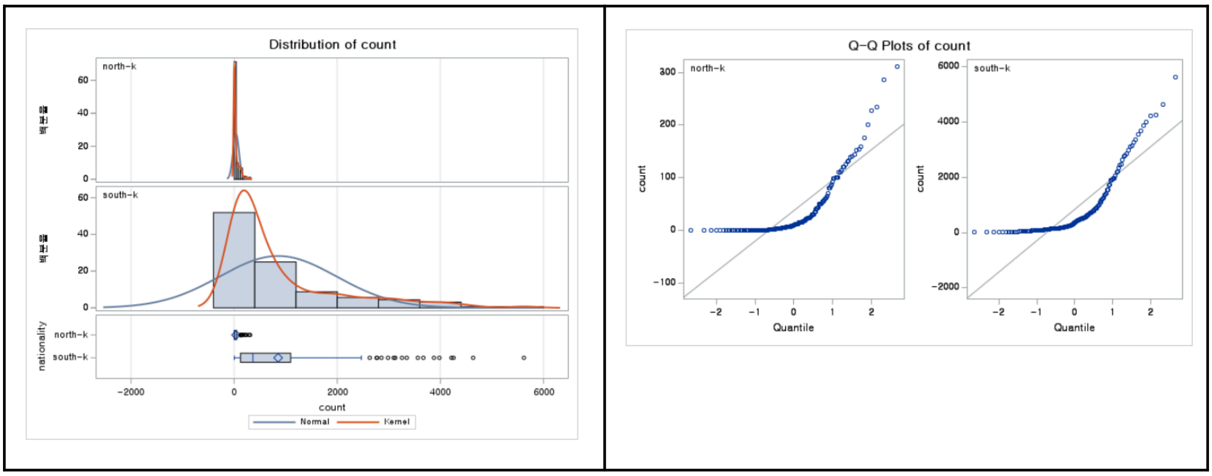
Variable: count

nationality	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
north-k		160	36.7625	57.9752	4.5833	0	311.0
south-k		160	853.5	1130.0	89.3367	4.0000	5620.0
Diff (1-2)	Pooled		-816.7	800.1	89.4541		
Diff (1-2)	Satterthwaite		-816.7		89.4541		

nationality	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
north-k		36.7625	27.7104	45.8146	57.9752	52.2431	65.1314
south-k		853.5	677.0	1029.9	1130.0	1018.3	1269.5
Diff (1-2)	Pooled	-816.7	-992.7	-640.7	800.1	742.5	867.5
Diff (1-2)	Satterthwaite	-816.7	-993.4	-640.0			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	318	-9.13	<.0001
Satterthwaite	Unequal	159.84	-9.13	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	159	159	379.92	<.0001



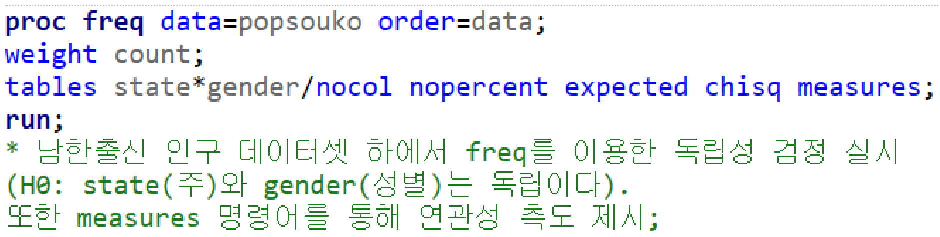
First, in the test for the equality of variances, the F-value is 379.92, and the corresponding p-value is less than 0.0001, indicating that the population variances of the two groups are significantly different. Since the assumption of equal variances is rejected, the Satterthwaite method, rather than the Pooled method should be used.
In the results of the Satterthwaite method, the estimated difference in means for the two populations is -816.7, and the p-value for a two-tailed test is less than 0.0001.

H0: The mean of the North Korean population is greater than or equal to the mean of the South Korean population.
H1: It is less.

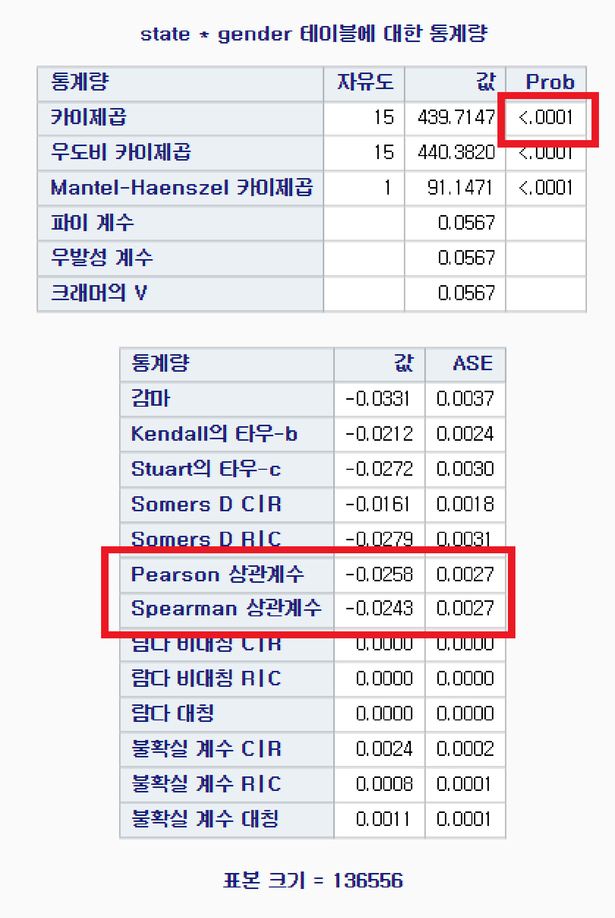
Given the hypothesis as above, the p-value is approximately 0.0001/2. Therefore, at a typical significance level, H0 can be rejected, concluding that the mean of the North Korean population in Germany is significantly lower than the mean of the South Korean population.

<3> Categorical Data Analysis

(1) Test of Independence



```
proc freq data=popsouko order=data;
weight count;
tables state*gender/nocol nopercent expected chisq measures;
run;
* 남한출신 인구 데이터셋 하에서 freq를 이용한 독립성 검정 실시
(H0: state(주)와 gender(성별)는 독립이다).
또한 measures 명령어를 통해 연관성 측도 제시;
```



state * gender 테이블에 대한 통계량

통계량	자유도	값	Prob
카이제곱	15	439.7147	<.0001
우도비 카이제곱	15	440.3820	<.0001
Mantel-Haenszel 카이제곱	1	91.1471	<.0001
파이 계수		0.0567	
우발성 계수		0.0567	
크래머의 V		0.0567	

통계량	값	ASE
감마	-0.0331	0.0037
Kendall의 타우-b	-0.0212	0.0024
Stuart의 타우-c	-0.0272	0.0030
Somers D C|R	-0.0161	0.0018
Somers D R|C	-0.0279	0.0031
Pearson 상관계수	-0.0258	0.0027
Spearman 상관계수	-0.0243	0.0027
람다 비대칭 C|R	0.0000	0.0000
람다 비대칭 R|C	0.0000	0.0000
람다 대칭	0.0000	0.0000
불확실 계수 C|R	0.0024	0.0002
불확실 계수 R|C	0.0008	0.0001
불확실 계수 대칭	0.0011	0.0001

표본 크기 = 136556

First, the hypothesis are as:

H0: State and gender are independent.
H1: They are not independent.

With a p-value for the chi-squared test of less than 0.0001, H0 can be rejected at the typical significance level. In other words, the two variables are not independent.

Furthermore, when the hypothesis are given as:

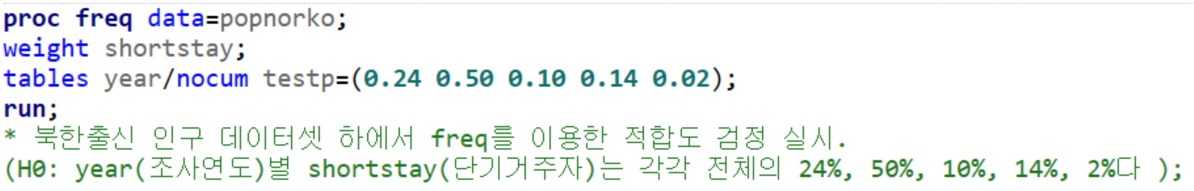
H0': There is no association between the two variables.
H1': There is an association.

The values of the Pearson correlation coefficient and Spearman correlation coefficient, divided by the ASE, are:

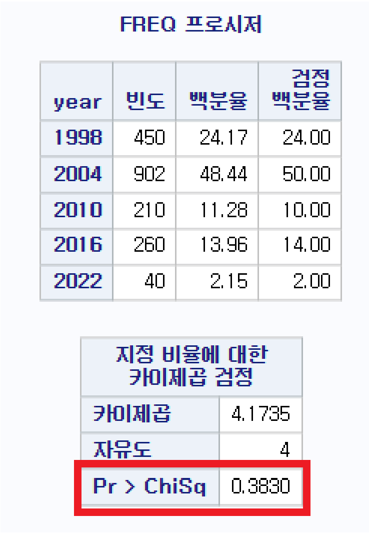
-0.0258/0.0027 ≈ -9.56
-0.0243/0.0027 ≈ -9

Thus, at a 5% significance level, H0' can be rejected, indicating that there is an association between the two variables as the correlation coefficients are not equal to zero.

(2) Goodness-of-fit test



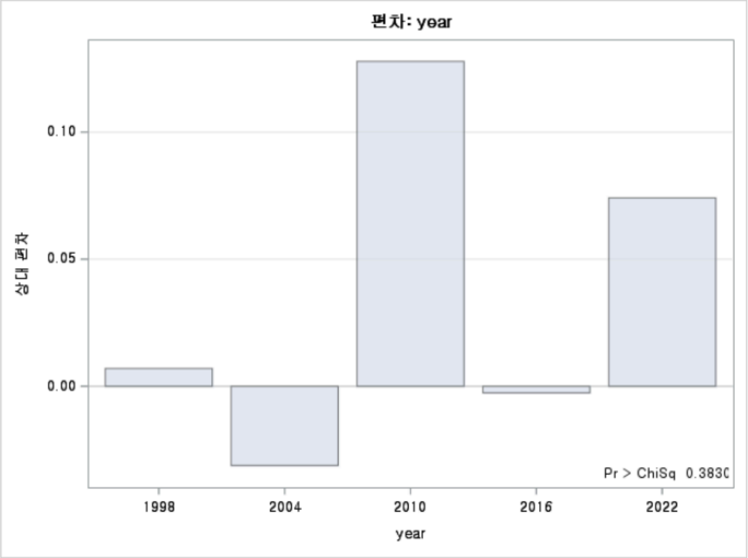
```
proc freq data=popnorko;
weight shortstay;
tables year/nocum testp=(0.24 0.50 0.10 0.14 0.02);
run;
* 북한출신 인구 데이터셋 하에서 freq를 이용한 적합도 검정 실시.
(H0: year(조사연도)별 shortstay(단기거주자)는 각각 전체의 24%, 50%, 10%, 14%, 2%다 );
```



FREQ 프로시저

year	빈도	백분율	검정 백분율
1998	450	24.17	24.00
2004	902	48.44	50.00
2010	210	11.28	10.00
2016	260	13.96	14.00
2022	40	2.15	2.00

지정 비율에 대한 카이제곱 검정	
카이제곱	4.1735
자유도	4
Pr > ChiSq	0.3830



The hypothesis are as:

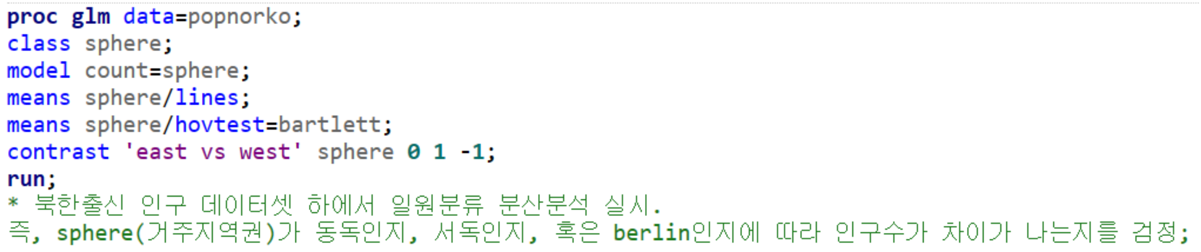
H0: The year-wise proportions of short-stay = 24:50:10:14:2.

H1: They do not satisfy the given proportions.

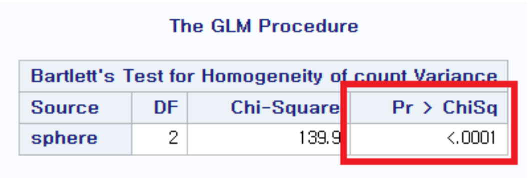
With a p-value from the chi-squared test of 0.3830, the null hypothesis cannot be rejected at the typical significance level. Therefore, it can be concluded that the year-wise proportions of short-stay are indeed 24:50:10:14:2.

<4> Analysis of Variance

(1) One-Way ANOVA



```
proc glm data=popnorko;
class sphere;
model count=sphere;
means sphere/lines;
means sphere/hovtest=bartlett;
contrast 'east vs west' sphere 0 1 -1;
run;
* 북한출신 인구 데이터셋 하에서 일원분류 분산분석 실시.
즉, sphere(거주지역권)가 동독인지, 서독인지, 혹은 berlin인지에 따라 인구수가 차이가 나는지를 검정;
```

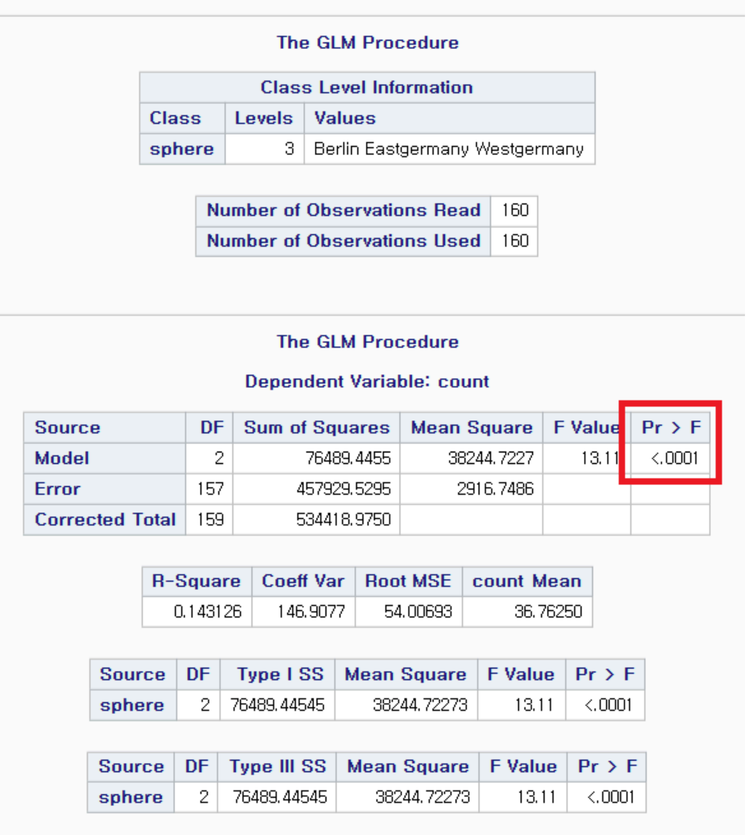


The GLM Procedure

Bartlett's Test for Homogeneity of count Variance

Source	DF	Chi-Square	Pr > ChiSq
sphere	2	139.9	<.0001

(As shown in the image above, since the assumption of homogeneity of variances was rejected, parametric tests such as GLM or ANOVA may not be appropriate. However, when non-parametric tests like Wilcoxon's rank-sum test were attempted, the computational workload became extensive, and results could not be obtained within a reasonable timeframe. Consequently, the decision was made to proceed with the testing using the GLM procedure.)



The GLM Procedure

Class Level Information		
Class	Levels	Values
sphere	3	Berlin Eastgermany Westgermany

Number of Observations Read	160
Number of Observations Used	160

The GLM Procedure

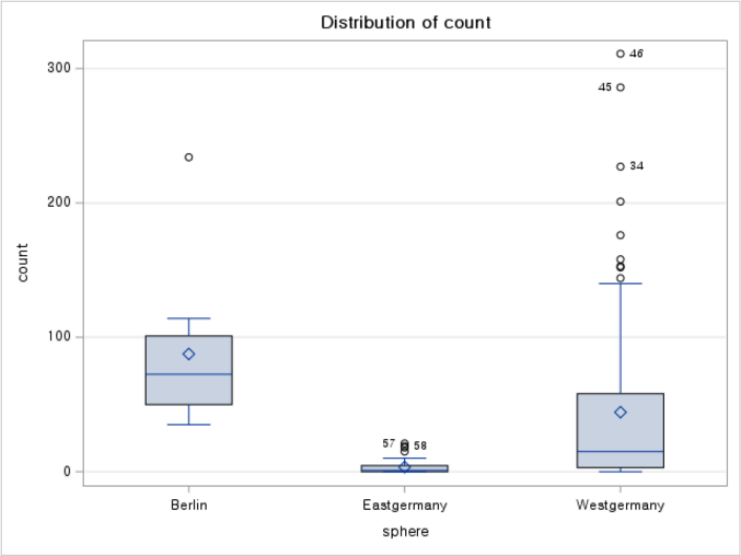
Dependent Variable: count

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	76489.4455	38244.7227	13.11	<.0001
Error	157	457929.5295	2916.7486		
Corrected Total	159	534418.9750			

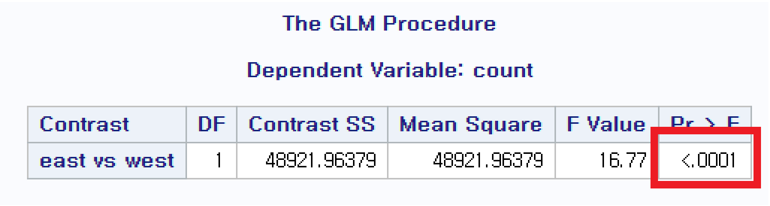
R-Square	Coeff Var	Root MSE	count Mean
0.143126	146.9077	54.00693	36.76250

Source	DF	Type I SS	Mean Square	F Value	Pr > F
sphere	2	76489.44545	38244.72273	13.11	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
sphere	2	76489.44545	38244.72273	13.11	<.0001



When examining the analysis of the variance table, the F value is 13.11, and the corresponding p-value is less than 0.0001. Therefore, at a typical significance level, the null hypothesis (H0: Regardless of sphere, the North Korean population remains constant) can be rejected. As a result, it can be concluded that there is a difference in the North Korean population based on sphere (residential area region).

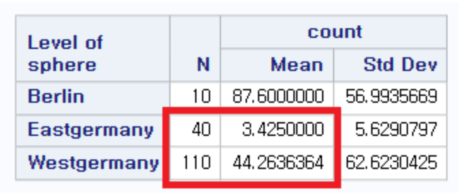


The GLM Procedure

Dependent Variable: count

Contrast	DF	Contrast SS	Mean Square	F Value	Pr > F
east vs west	1	48921.96379	48921.96379	16.77	<.0001

Furthermore, post hoc testing using contrasts reveals a significant difference between the North Korean population in East Germany and West Germany.

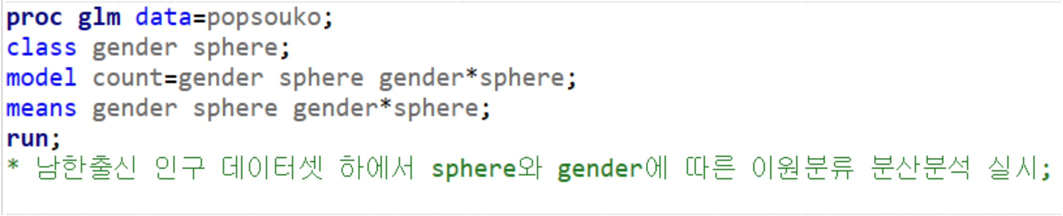


Level of sphere	N	count	
		Mean	Std Dev
Berlin	10	87.6000000	56.9935669
Eastgermany	40	3.4250000	5.6290797
Westgermany	110	44.2636364	62.6230425

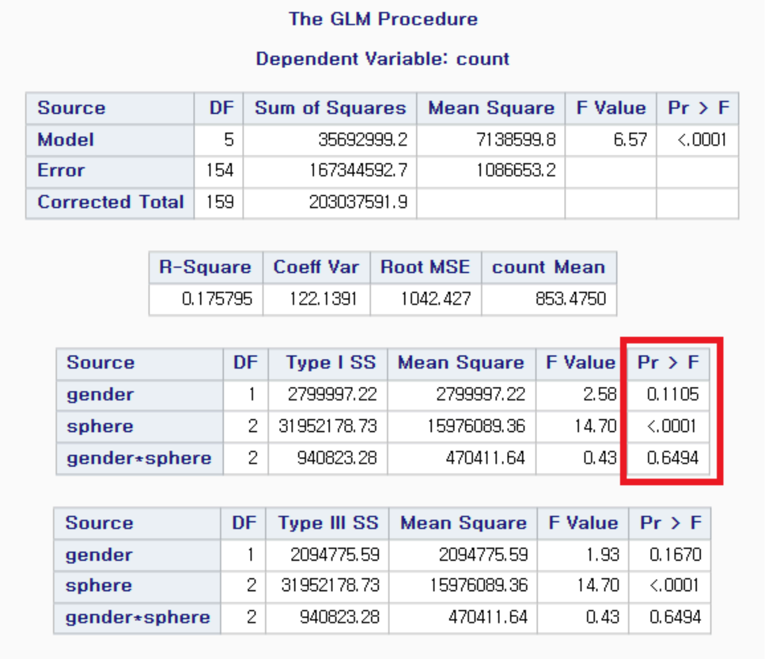
Having concluded that there is a difference in the North Korean population based on the residential region, an interesting observation is that significantly more North Korean individuals reside in West Germany than in East Germany, despite the latter's historical ties to the former Communist bloc.

Excluding Berlin, the population ratio of East Germany to West Germany is approximately **1:5.38** (as of 2021, source: "Population in Germany by federal state 2021 | Statista"). However, the North Korean population's ratio of residing in East Germany to West Germany is **1:35.54**, indicating that they reside in West Germany at a much higher rate than the general population of Germany.

(2) Two-way ANOVA



```
proc glm data=popsouko;
class gender sphere;
model count=gender sphere gender*sphere;
means gender sphere gender*sphere;
run;
* 남한출신 인구 데이터셋 하에서 sphere와 gender에 따른 이원분류 분산분석 실시;
```



The GLM Procedure

Dependent Variable: count

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	35692999.2	7138599.8	6.57	<.0001
Error	154	167344592.7	1086653.2		
Corrected Total	159	203037591.9			

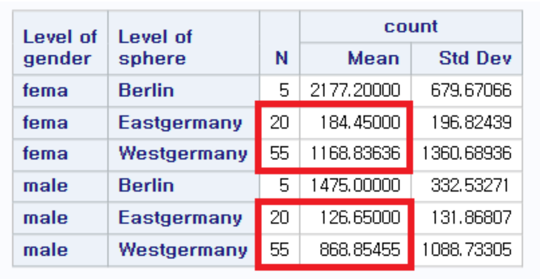
R-Square	Coeff Var	Root MSE	count Mean
0.175795	122.1391	1042.427	853.4750

Source	DF	Type I SS	Mean Square	F Value	Pr > F
gender	1	2799997.22	2799997.22	2.58	0.1105
sphere	2	31952178.73	15976089.36	14.70	<.0001
gender*sphere	2	940823.28	470411.64	0.43	0.6494

Source	DF	Type III SS	Mean Square	F Value	Pr > F
gender	1	2094775.59	2094775.59	1.93	0.1670
sphere	2	31952178.73	15976089.36	14.70	<.0001
gender*sphere	2	940823.28	470411.64	0.43	0.6494

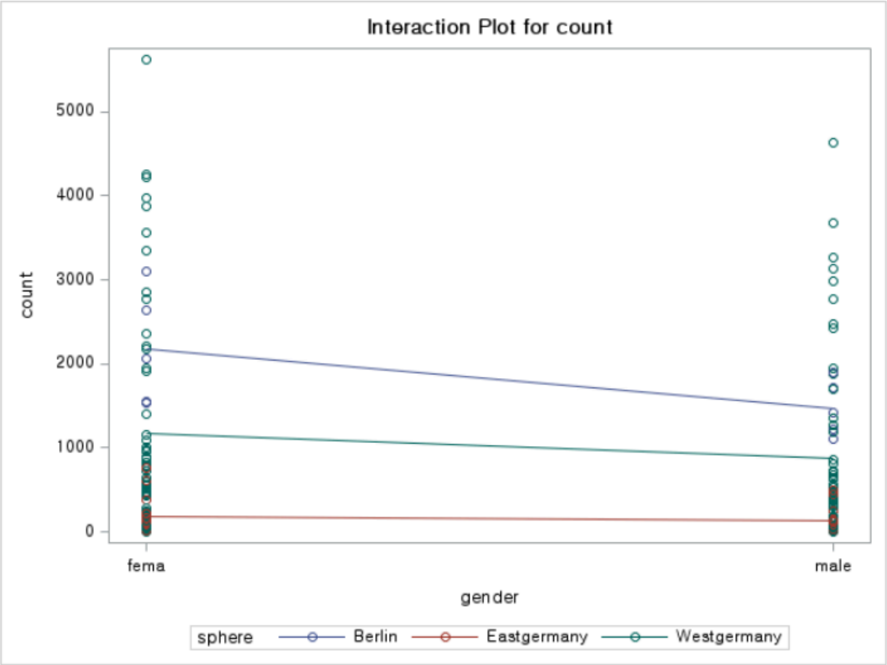
When examining the p-values from the F-tests:

For the variable "gender," the p-value is 0.1105, indicating that it can be rejected at a significance level of approximately 12% or higher. In other words, gender is statistically significant at a significance level of 12% or higher.
For the variable "sphere," the p-value is less than 0.0001, making it possible to reject it at all typical significance levels. In other words, "sphere" is always statistically significant.
On the other hand, the p-value for the interaction between "gender" and "sphere" is 0.6494, and it cannot be rejected at a typical significance level. Therefore, there is no interaction between these two variables.

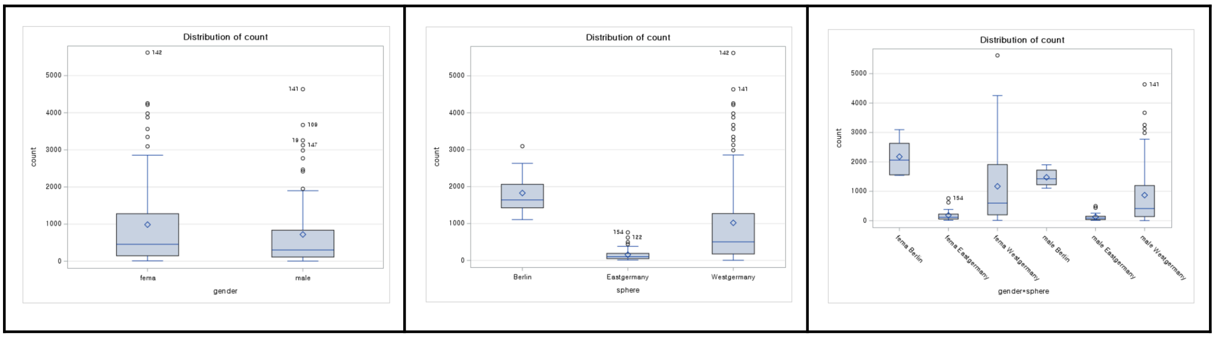


Level of gender	Level of sphere	N	count	
			Mean	Std Dev
fema	Berlin	5	2177.20000	679.67066
fema	Eastgermany	20	184.45000	196.82439
fema	Westgermany	55	1168.83636	1360.68936
male	Berlin	5	1475.00000	332.53271
male	Eastgermany	20	126.65000	131.86807
male	Westgermany	55	868.85455	1088.73305

Furthermore, it can be observed that South Korean nationals also reside more in West Germany than in East Germany.

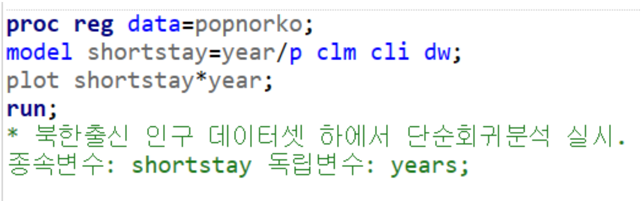


: Interaction Plot

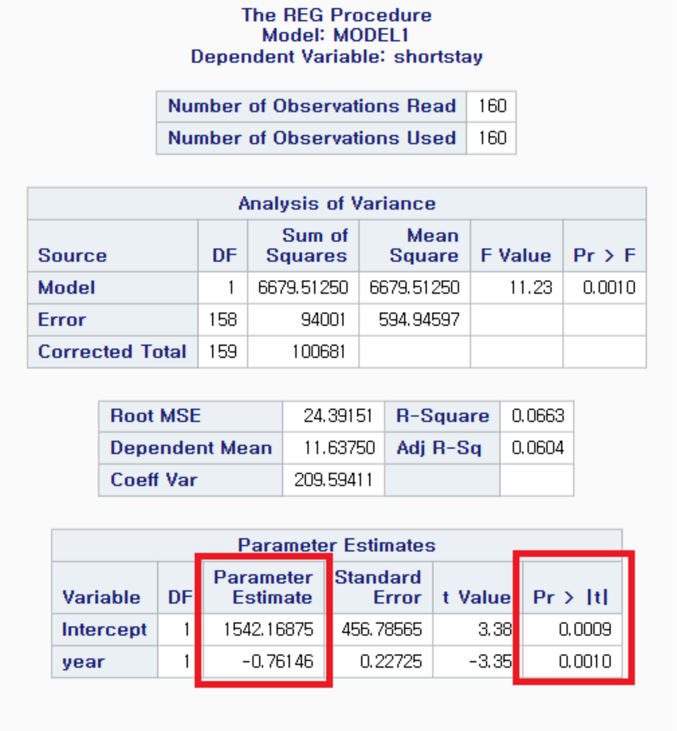


<5> Regression Analysis

(1) Simple Linear Regression Analysis



```
proc reg data=popnorko;
model shortstay=year/p clm cli dw;
plot shortstay*year;
run;
* 북한출신 인구 데이터셋 하에서 단순회귀분석 실시.
종속변수: shortstay 독립변수: years;
```



The REG Procedure
Model: MODEL1
Dependent Variable: shortstay

Number of Observations Read	160
Number of Observations Used	160

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	6679.51250	6679.51250	11.23	0.0010
Error	158	94001	594.94597		
Corrected Total	159	100681			

Root MSE	24.39151	R-Square	0.0663
Dependent Mean	11.63750	Adj R-Sq	0.0604
Coeff Var	209.59411		

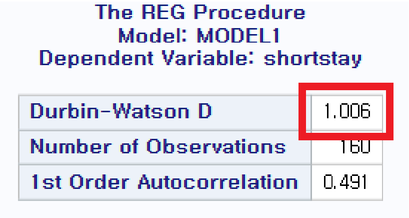
Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	1542.16875	456.78565	3.38	0.0009
year	1	-0.76146	0.22725	-3.35	0.0010

As evident from the table, the estimated parameters for β0 and β1 are 1542.17 and -0.76, respectively. Thus, the estimated regression equation can be written as y = -0.76x + 1542.17.

Here, the value β1 = -0.76 indicates that for each year that passes, the North Korean population in Germany decreases by 0.76 individuals.

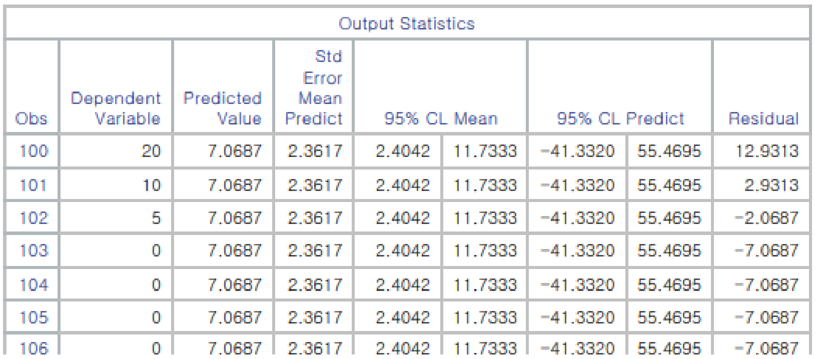
Additionally, the two-tailed p-value for the hypothesis test regarding the slope coefficient β1 is 0.001. Therefore, at the typical significance level, the null hypothesis (H0: β1 = 0) can be rejected. This implies that the North Korean population is changing significantly over time.



The REG Procedure
Model: MODEL1
Dependent Variable: shortstay

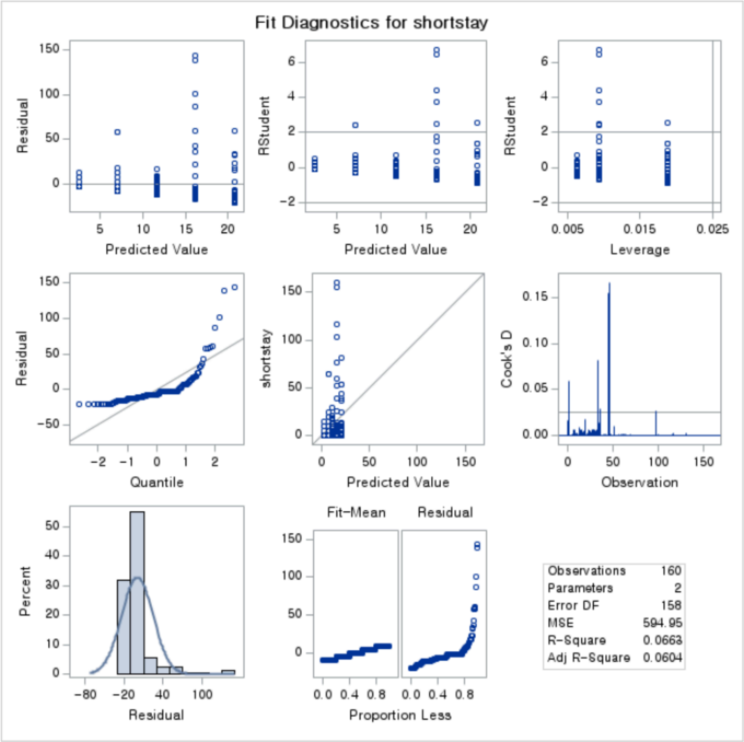
Durbin-Watson D	1.006
Number of Observations	160
1st Order Autocorrelation	0.491

With a Durbin-Watson coefficient value of 1.006, it can be concluded that the error terms satisfy the condition of independence, as they are relatively close to the value of 2.



Output Statistics								
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean		95% CL Predict		Residual
100	20	7.0687	2.3617	2.4042	11.7333	-41.3320	55.4695	12.9313
101	10	7.0687	2.3617	2.4042	11.7333	-41.3320	55.4695	2.9313
102	5	7.0687	2.3617	2.4042	11.7333	-41.3320	55.4695	-2.0687
103	0	7.0687	2.3617	2.4042	11.7333	-41.3320	55.4695	-7.0687
104	0	7.0687	2.3617	2.4042	11.7333	-41.3320	55.4695	-7.0687
105	0	7.0687	2.3617	2.4042	11.7333	-41.3320	55.4695	-7.0687
106	0	7.0687	2.3617	2.4042	11.7333	-41.3320	55.4695	-7.0687

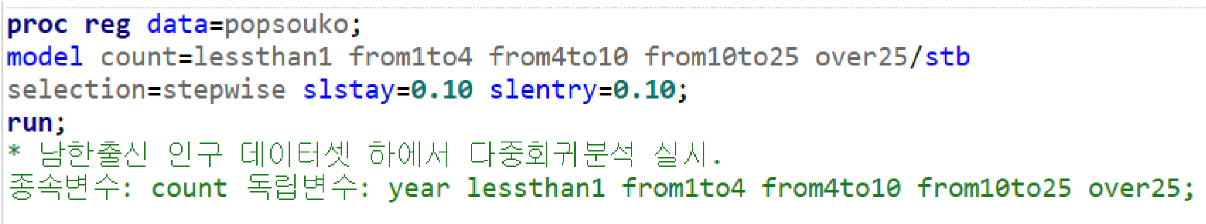
: P, CLI, CLM



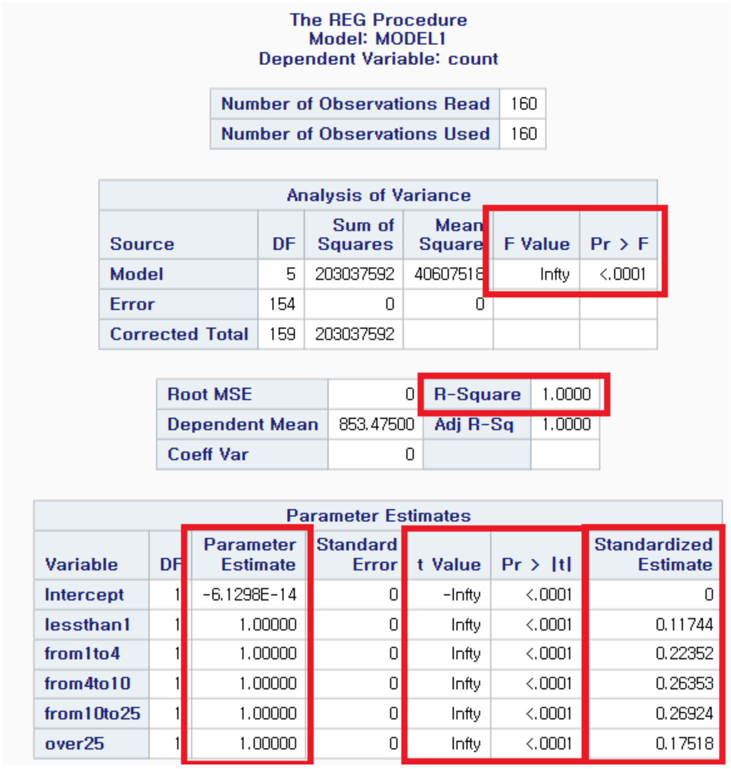
Observations	160
Parameters	2
Error DF	158
MSE	594.95
R-Square	0.0663
Adj R-Square	0.0604

: Regression Diagnostic Plot

(2) Multiple Regression Analysis



```
proc reg data=popsouko;
model count=lessthan1 from1to4 from4to10 from10to25 over25/stb
selection=stepwise slstay=0.10 slentry=0.10;
run;
* 남한출신 인구 데이터셋 하에서 다중회귀분석 실시.
종속변수: count 독립변수: year lessthan1 from1to4 from4to10 from10to25 over25;
```



The REG Procedure
Model: MODEL1
Dependent Variable: count

Number of Observations Read	160
Number of Observations Used	160

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	5	203037592	40607518	Infty	<.0001
Error	154	0	0		
Corrected Total	159	203037592			

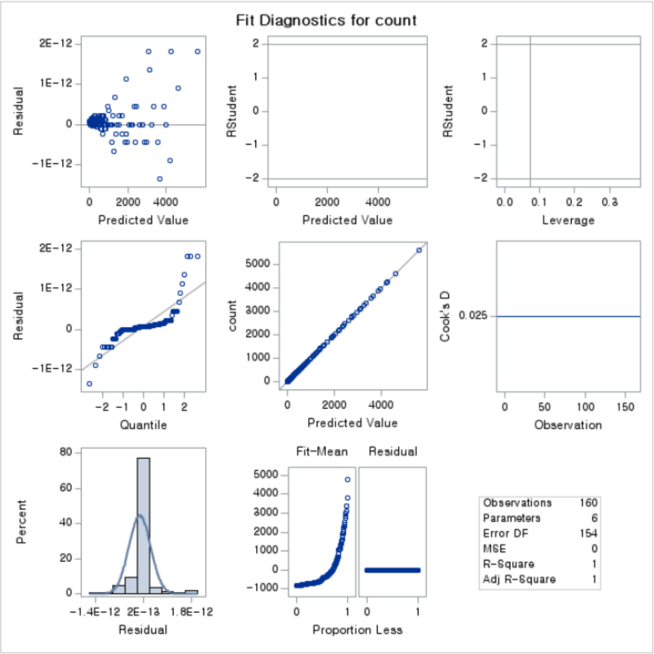
Root MSE	0	R-Square	1.0000
Dependent Mean	853.47500	Adj R-Sq	1.0000
Coeff Var	0		

Parameter Estimates						
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	1	-6.1298E-14	0	-Infty	<.0001	0
lessthan1	1	1.00000	0	Infty	<.0001	0.11744
from1to4	1	1.00000	0	Infty	<.0001	0.22352
from4to10	1	1.00000	0	Infty	<.0001	0.26353
from10to25	1	1.00000	0	Infty	<.0001	0.26924
over25	1	1.00000	0	Infty	<.0001	0.17518

First, with the stepwise variable selection method, it is shown that all five independent variables are significant at a significance level of 10% or less. With an R-squared value of 1.00, this means that the five independent variables can explain the dependent variable by 100%.

Additionally, when examining the parameter estimate column, it's evident that the estimated slopes for each independent variable are all 1.

Lastly, comparing the standardized regression coefficients for each independent variable, it is shown that "lessthan1" (residents for less than 1 year) has the smallest coefficient of 0.11744, while "from10to25" (residents for 10 to 25 years) has the largest coefficient of 0.26924.



: Regression Diagnostic Plot