SAS II

Inferential Statistics

After attending this training session, you will be able to run and interpret the following tests in SAS:

1. Chi-square test
2. Independent samples t-test and Levene's test for the equality of variances
3. One-way ANOVA
4. Bivariate Statistics: Correlation
5. Bivariate Statistics: Regression
6. Multiple regression

Introduction

Code and Dataset

For this training session, please download the following data and code:

gss2018sasi.sas7bdat

SAS II Script for Training.sas

Recap of SAS I

In the **SAS I** training session, we covered the following topics. Please feel free to refer back to the SAS I workbook using the following links if you would like a refresher.

- Introduction
  - What is SAS?
  - SAS Programs
  - The SAS Environment
  - SAS Libraries
  - Opening a Data Set
- The **Data** Step
- The **Proc** Step
- Descriptive Statistics
- Multivariate Tables
- Graphics
- Saving and Printing

As in the SAS I training session, we will be using data from the General Social Survey.

The Data for This Workshop

The SAS program below creates a temporary dataset using data from the 2018 GSS, *gss2018sasi.sas7bdat*. Create a folder called *Training* in which you will save the SAS *gss2018sasi* data file. Next, we will set up a working directory called *Training* for our analysis. Go to the *Explorer* and navigate to *Libraries*. Right click and create a new library called *Training* using the location for your training folder. For step by step instructions, see the **SAS I Workbook**. We will now create a temporary dataset for this training, *gss2018*.

```
data gss2018;
    set training.gss2018sasi;
run;
```

Chi-square Test

A chi-square test is used to determine whether there is a statistically significant association between two categorical variables. The chi-square test statistic is requested in the `proc freq` procedure.

Example 1: 2x2 Table

For example, you may want to know if there is a significant association between the variables *sex* (Respondent's gender) and *hschem* (Respondent ever took chemistry class in high school). The following program runs `proc freq` with the `chisq` option:

```
proc freq data = gss2018;
      tables hschem*sex /chisq;
run;
```

The variable name that appears immediately after tables will appear in the rows and the second variable, which appears after the *, will be placed in the columns. Either of the variables can go in the rows or columns.

The SAS System

The FREQ Procedure

Frequency Percent Row Pct Col Pct	Table of hschem by sex		
hschem(R ever took a high school chemistry course)	sex(Respondents sex)		
	1	2	Total
1	259 24.86 43.53 60.37	336 32.25 56.47 54.81	595 57.10
2	170 16.31 38.03 39.63	277 26.58 61.97 45.19	447 42.90
Total	429 41.17	613 58.83	1042 100.00
Frequency Missing = 1160			

SAS generates the following test statistics for chi-square. The levels of measurement, number of categories within the variables, and the sample size will determine which test statistic is appropriate.

Statistics for Table of hschem by sex

Statistic	DF	Value	Prob
Chi-Square	1	3.1857	0.0743
Likelihood Ratio Chi-Square	1	3.1937	0.0739
Continuity Adj. Chi-Square	1	2.9627	0.0852
Mantel-Haenszel Chi-Square	1	3.1826	0.0744
Phi Coefficient		0.0553	
Contingency Coefficient		0.0552	
Cramer's V		0.0553	

Fisher's Exact Test	
Cell (1,1) Frequency (F)	259
Left-sided Pr <= F	0.9679
Right-sided Pr >= F	0.0425
Table Probability (P)	0.0103
Two-sided Pr <= P	0.0754

Sample Size = 1042
Frequency Missing = 1160

WARNING: 53% of the data are missing.

Chi-Square: Used to test the hypothesis that the row and column variables are independent. This should not be used if any cell has an expected value less than 1, or if more than 20 percent of the cells have expected values less than 5.

Likelihood Ratio Chi-Square: A goodness-of-fit statistic similar to Pearson's chi-square. This statistic is appropriate to report for smaller sample sizes. The likelihood ratio approaches the Pearson's chi-square as *n* increases, so for large sample sizes, the two statistics are the same and either can be reported.

Continuity Adj. Chi-Square: A correction is applied in the computation of chi-square statistic for 2x2 tables to improve the approximation. Corrected chi-square values are always smaller than uncorrected values.

Mantel-Haenszel Chi-Square: A measure of the linear association between the row and column variables. This statistic should not be used for nominal data. It is also referred to as the Linear-by-Linear Association and the Mantel-Haenszel test for linear association.

The **Phi Coefficient**, **Contingency Coefficient**, and **Cramer's V** are measures of the strength of the association and are not covered in this workshop.

Fisher's Exact Test: A test for independence in a 2x2 table. This statistic is most useful when the total sample size and expected values are small.

Because this example uses a 2X2 table, we will use Fisher's Exact test. Here, the Fisher's Exact test has a probability *p-value* larger than 0.05 so we can reject the null hypothesis using the two-sided (or two-tailed) p-value. Thus, we can conclude that there is a statistically significant relationship between the respondent's gender and taking chemistry in high school.

Example 2: 3x2 Table

In this example, you test whether there is a significant association between the variables *natspac* (attitude towards the space exploration program) and *colsci* (whether respondent took college-level science courses).

The following program runs `proc freq` with the `chisq` option:

```
proc freq data = gss2018;
      tables natspac*colsci /chisq;
run;
```

The SAS System

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct	Table of natspac by colsci		
natspac(Space exploration program)	colsci(R has taken any college-level sci course)		
	1	2	Total
1	92 67.241 14.42 64.34 30.67	51 75.759 7.99 35.66 15.09	143 22.41
2	150 151.41 23.51 46.58 50.00	172 170.59 26.96 53.42 50.89	322 50.47
3	58 81.348 9.09 33.53 19.33	115 91.652 18.03 66.47 34.02	173 27.12
Total	300 47.02	338 52.98	638 100.00
Frequency Missing = 1564			

Statistics for Table of natspac by colsci

Statistic	DF	Value	Prob
Chi-Square	2	29.8814	<.0001
Likelihood Ratio Chi-Square	2	30.2973	<.0001
Mantel-Haenszel Chi-Square	1	29.4841	<.0001
Phi Coefficient		0.2164	
Contingency Coefficient		0.2115	
Cramer's V		0.2164	

Sample Size = 638
Frequency Missing = 1564

WARNING: 71% of the data are missing.

It is appropriate to use the Chi-Square test statistic in this example. The *p-value* is smaller than 0.0001. Therefore, it may be concluded that there is a statistically significant relationship between attitude towards the space exploration program and having taken college level science classes.

Additional Options for Contingency Tables

Some additional options that can be specified on the tables statement (appearing after the forward slash - `/`) in `proc freq` include:

- **all** - calculates all available statistics in a two-way table
- **chisq** - calculates chi-square and other statistics for independence in a two-way table
- **expected** - computes the expected counts for a two-way table
- **fisher** - requests Fisher's exact test for tables larger than two-by-two
- **nocol** - omits column percents
- **norow** - omits row percents
- **nopercent** - omits table total percents

Exercise 1

Perform a chi-square test to determine if educational attainment (*degree*) is significantly associated with the belief that marijuana should be legal (*grass*). (See Appendix A for answers.)

1. How many high school graduates (*degree* = 1) said Yes to legalization (*grass* = 1)?
2. Is there a statistically significant relationship between the two variables? Which chi-square test statistic is appropriate?

Independent samples t-test

An independent samples t-test determines whether there is a statistically significant difference in the mean of a continuous variable between two groups. The independent samples t-test is obtained by using the `proc ttest` procedure.

For example, this procedure can be used to find out if there is a statistically significant difference between the responses of men and women (*sex*) for r how much they have to relax per day (*hrlax*). The following program runs `proc ttest`:

```
proc ttest data = gss2018;
      class sex;
      var hrlax;
run;
```

The grouping variable (*sex*) is identified in the `class` statement. The test variable (*hrlax*) is identified in the `var` statement.

Variable: hrlax (Hours per day R have to relax)

sex	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		636	3.9748	2.8681	0.1137	0	24.0000
2		686	3.5015	2.7508	0.1050	0	20.0000
Diff (1-2)	Pooled		0.4734	2.8079	0.1546		
Diff (1-2)	Satterthwaite		0.4734		0.1548		

sex	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		3.9748	3.7515	4.1982	2.8681	2.7187	3.0351
2		3.5015	3.2952	3.7077	2.7508	2.6126	2.9047
Diff (1-2)	Pooled	0.4734	0.1702	0.7766	2.8079	2.7047	2.9192
Diff (1-2)	Satterthwaite	0.4734	0.1697	0.7771			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	1320	3.06	0.0022
Satterthwaite	Unequal	1302.1	3.06	0.0023

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	635	685	1.09	0.2834

The results of the t-test presented in the output are used to decide whether there is a statistically significant difference by sex in hours relaxed per week. There are two sets of statistics generated by SAS: (1) one set that assumes the variances in the distribution of the dependent variable are equal for both groups - Pooled method; and (2) the other set that does not assume that the variances in the distributions are equal - Satterthwaite method.

To determine which method is appropriate, check the *p-value* of the Equality of Variances test results. When the Equality of Variances test generates a non-significant F-test statistic, the assumption of equal variance is met, and the statistics for Pooled method should be used. Because the *p-value* of Equality of Variances test in this example is larger than 0.05 (p=0.2834), the Pooled statistics should be used to determine if there is a statistically significant difference.

Using the Pooled test, the *p-value* of the *t* statistic is less than 0.05, therefore we may conclude that there is a statistically significant difference in the number of hours men and women have to relax per day.

Exercise 2

Find out if there is a statistically significant difference in terms of respondent's family income(*fmi*) between respondents who took college-level science courses and those who didn't? (*colsci*). (See Appendix A for answers.)

ANOVA

A One-way ANOVA is used to determine whether there is a statistically significant difference in the mean of a continuous dependent variable among (typically more than) two groups. One method of performing ANOVA in SAS is to use the `proc glm` procedure. Unlike `proc anova`, the `proc glm` procedure is designed to handle both balanced data (data with equal numbers of observations for every combination of the classification factors) and unbalanced data.

For example, this procedure may be used to find out if there is a statistically significant difference in the average number of hours spent watching television per day (*tvhours*) by marital group (*marital*). The following program runs `proc glm` with the `hovtest` option:

```
proc glm data = gss2018;
      class marital;
      model tvhours = marital;
      means marital / hovtest;
run;
```

The dependent variable (*tvhours*) and grouping variable (*marital*) are both identified in the `model` statement in the expression `tvhours = marital`. The grouping variable (*marital*) is also identified in both the `class` and `means` statements. The means statement is used to obtain descriptive statistics for the dependent variable by the grouping variable. The `hovtest` option is specified in the means statement to request the Levene's test of the homogeneity of variance. Before we check the ANOVA results, we should check Levene's test results.

The SAS System

The GLM Procedure

Levene's Test for Homogeneity of tvhours Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
marital	4	10517.6	2629.4	2.78	0.0257
Error	1453	1374997	946.3		

From the Levene results table, P value is smaller than 0.05. We reject the null hypothesis that the variances are equal, and not-equal variances are assumed. The result confirms that We should use Welch test for non-equal variances, instead of regular ANONVA results shown from this outcome.

Welch Test

When the homogeneity of variances assumption is not met, the Welch test should be requested when performing one-way ANOVA. To request the Welch test of equality of means, specify the `welch` option in the `means` statement.

The following program runs `proc glm` with the `welch` option:

```
proc glm data = gss2018;
      class marital;
      model tvhours = marital;
      means marital / welch;
run;
```

The GLM Procedure

Dependent Variable: tvhours Hours per day watching TV

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	227.44429	56.86107	7.29	<.0001
Error	1453	11333.17917	7.79985		
Corrected Total	1457	11560.62346			

R-Square	Coeff Var	Root MSE	tvhours Mean
0.019674	95.78764	2.792821	2.915638

From the Welch table, you can conclude that the difference in the mean number of hours watching television per day among marital groups is statistically significant.

ANOVA Post Hoc Tests

Once you have determined that differences exist among the means, post hoc tests can determine which means differ. Post hoc tests are specified in the means statement. Post hoc tests available in SAS include:

Keyword	Test
`t`	Student's t
`snk`	Student-Newman-Keuls
`tukey`	Tukey-Kramer
`sidak`	Sidak
`gabriel`	Gabriel
`duncan`	Waller-Duncan
`regwq`	Ryan-Einot-Gabriel-Welch
`bon`	Bonferroni
`scheffe`	Scheffe
`dunnett`	Dunnett

> **Note**
>
> In SAS only the Dunnett test is available for pairwise comparisons for one-way ANOVA when variances in groups are not equal.

In this example, looking at differences in the average time spent watching TV per day (*tvhours*) by marital status (*marital*), the results of the Levene's Test of Homogeneity of TVHOURS Variance indicate that equal variances can not be assumed. Therefore, we can select a post hoc test that assumes equal variances. For example, specify `dunnett` in `means` statement to request the Dunnett post hoc test.

The following program runs `proc glm` with the `dunnett` option:

```
proc glm data = gss2018;
      class marital;
      model tvhours = marital;
      means marital / dunnett;
run;
```

Comparisons significant at the 0.05 level are indicated by ***.				
marital Comparison	Difference Between Means	Simultaneous 95% Confidence Limits		
2 - 1	1.3805	0.7038	2.0572	***
3 - 1	0.5791	0.0719	1.0862	***
5 - 1	0.4076	-0.0384	0.8536	
4 - 1	0.4011	-0.7728	1.5750	

> **Note**
>
> In the data, Married = 1, Widowed = 2, Divorced = 3, Separated = 4, and Never married = 5.

The Dunnet Tests for TVHOURS provides the results of post hoc test. From this table it may be concluded that:

1. Group 1 (Married) is significantly different from Group 2 (Widowed) and Group 3 (Divorced) at the 0.05 level.
2. All the other groups based on marital status have no significant different with each other. After identifying statistically significant differences, We can determine which group watched more TV hours per day: Compared with Married group, Widowed group watched 1.38 more hours and Divorced group watched 0.579 more hours per day.

Exercise 3

Using the one-way ANOVA procedure, test whether respondents' family income(*fmi*) is different among different education groups (*degree*). Remember to check the Test of Homogeneity of Variances to determine if the Welch test is required. (See Appendix A for answers.)

Bivariate Correlation

Correlation is a measure of the linear relationship between two continuous variables. A correlation coefficient ranges from −1 to 1. A positive correlation coefficient indicates a positive linear relationship (i.e., as one variable increases, the other tends to as well). On the other hand, a negative correlation coefficient indicates a negative linear relationship. A correlation coefficient of 0 indicates there is no linear relationship between the two variables. Bivariate correlations are obtained using the `proc corr` procedure.

For example, this procedure may be used to measure the relationship between the continuous variables spouse occupation prestige score (*sop*), spouse socioeconomic index score (*ssi*), and no.of hrs spouse usually works a week (*sphrs2*).

> **i** Note
>
> A bivariate correlation measures the linear relationship between two continuous variables, but it is convenient to calculate several bivariate correlations simultaneously in a matrix.

The following program runs `proc corr`:

```
proc corr data = gss2018;
      var ssi sop sphrs2;
run;
```

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
sop	1102	45.48457	13.40263	50124	16.00000	80.00000	Spouse occupational prestige score
ssi	1102	49.38348	22.71699	54421	12.60000	93.70000	Spouse socioeconomic index score
sphrs2	22	42.04545	10.28396	925.00000	15.00000	65.00000	No. of hrs spouse usually works a week

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations			
	sop	ssi	sphrs2
sop Spouse occupational prestige score	1.00000 1102	0.83556 <.0001 1102	-0.14836 0.5099 22
ssi Spouse socioeconomic index score	0.83556 <.0001 1102	1.00000 1102	-0.08875 0.6945 22
sphrs2 No. of hrs spouse usually works a week	-0.14836 0.5099 22	-0.08875 0.6945 22	1.00000 22

Each cell contains:

1. A correlation coefficient (Pearson Correlation)
2. The p-value associated with the correlation coefficient for the test of "no linear relationship"
3. The number of cases (N)

In this example, the correlation coefficient between *sop* and *ssi* is 0.83556 and the significance of the coefficient is less than 0.0001, which suggests that these two variables are significantly and positively correlated.

Note on missing values and sample size (N): According to the third row of the cell for the correlation between *sphrs2* with *ssi* or *sop* , only 22 cases are used, while there are 1,102 valid cases for both *ssi* or *sop* variables. The default treatment of missing cases in the Correlation procedure is to use the maximum number of non-missing values for each pair of variables, which is known as pairwise deletion. An alternative treatment of missing cases is listwise deletion. When this option is selected, cases with a missing value in any of the analyzed variables will be excluded from the analysis. (When only two variables are included, the sample size is the same whether pairwise or listwise deletion is used. However, these options make a difference in the sample size when three or more variables are included.)

The default is pairwise deletion, but we can change this by adding the `nomiss` option to the `proc corr` procedure. The interpretation of the correlation coefficient for listwise deletion is the same as that for pairwise deletion. The following program runs `proc corr` with the `nomiss` option:

```
proc corr data = gss2018 nomiss;
      var ssi sop sphrs2;
run;;
run;
```

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
ssi	22	60.49091	26.25897	1331	14.00000	89.30000	Spouse socioeconomic index score
sop	22	51.45455	16.61820	1132	21.00000	75.00000	Spouse occupational prestige score
sphrs2	22	42.04545	10.28396	925.00000	15.00000	65.00000	No. of hrs spouse usually works a week

Pearson Correlation Coefficients, N = 22 Prob > |r| under H0: Rho=0			
	ssi	sop	sphrs2
ssi Spouse socioeconomic index score	1.00000	0.96742 <.0001	-0.08875 0.6945
sop Spouse occupational prestige score	0.96742 <.0001	1.00000	-0.14836 0.5099
sphrs2 No. of hrs spouse usually works a week	-0.08875 0.6945	-0.14836 0.5099	1.00000

Exercise 4

Compute the correlation between: the highest year of school completed (*educ*), the respondent's father's occupational prestige index (*faos*), and the respondent's mother's occupational prestige index (*moos*) using (1) pairwise deletion and (2) listwise deletion and compare how the sample size changes. (See Appendix A for answers.)

Bivariate Regression

Bivariate linear regression is used to assess the influence of a continuous or dichotomous independent variable on a continuous dependent variable. Regression analysis is performed using the `proc reg` procedure.

This example uses bivariate regression to measure the effect of educational attainment (*educ*) on family's income (*realinc*).

```
proc reg data = gss2018;
      model realrinc = educ /stb;
run;
```

The option **stb** requests standardized coefficients.

The SAS System

The REG Procedure
Model: MODEL1
Dependent Variable: realrinc R's income in constant $

Number of Observations Read	2202
Number of Observations Used	1278
Number of Observations with Missing Values	924

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	99205782665	99205782665	129.84	<.0001
Error	1276	9.74977E11	764088567		
Corrected Total	1277	1.074183E12			

Root MSE	27642	R-Square	0.0924
Dependent Mean	25017	Adj R-Sq	0.0916
Coeff Var	110.49128		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	-19272	3963.09104	-4.86	<.0001	0
educ	Highest year of school completed	1	3131.35065	274.81160	11.39	<.0001	0.30390

Interpretation of Results:

- **Significance:** The **Analysis of Variance** table tells is if the independent variable has a significant linear relationship with the dependent variable. In this example, the F-statistic is statistically significant (*p-value* <0.001). Thus, educational attainment is a significant predictor of income.
- **Model Fit:** R-squared, also known as the **coefficient of determination**, tells you the proportion of variation in the dependent variable explained by the independent variable. R-squared ranges from 0 (no relationship) to 1 (perfect prediction of dependent variable from independent variable). In this example, 9.24% of the variation in income (dependent variable) is explained by the years of education (independent variable).

  An Adjusted R-squared is adjusted for the number of independent variables used in the model. However, since there is only one independent variable in this example, R-squared and Adjusted R-squared should be similar.

- **Parameter Estimates:** This table provides the values of the unstandardized regression coefficients (*Parameter Estimate*), standardized coefficients (*Standardized Estimate*), and a significance test of the coefficients (Pr > |t|).

The regression coefficient for the intercept is −19,272, while the coefficient for the independent variable *educ* is 3,131.35065. It can be said that the predicted level of income will increase by $3,131.35 for each additional year of education. Because this effect is measured with respect to the original scale of the independent variable, it is referred to as an **unstandardized coefficient**. When we have multiple independent variables as predictors in the model, we cannot compare the effects of independent variables using unstandardized coefficients. This will be addressed further in the section on Multiple Regression.

The standard error (SE) is used to calculate the t-statistic (β/SE). This statistic is used to test the null hypothesis that the coefficient is equal to 0. In this example, *educ* is statistically significant at the 0.0001 level. Thus, you can reject the null hypothesis that the education variable's coefficient is equal to 0 and conclude education is a significant predictor of income.

Exercise 5

Does age (*age*) have a significant linear relationship with performance on a vocabulary test (*wordsum*)? Treat *wordsum* as the dependent variable. (See Appendix A for answers.)

Multiple Regression

Multiple regression is the same as bivariate regression, except it uses two or more independent variables to explain the variation of the dependent variable. This example uses Multiple regression to measure the effect of respondents' health score (*rhs*) and age (*age*) on respondent's income (*realrinc*).

This example uses multiple regression to measure the effect of educational attainment (*educ*), sex(*male*), and age (*age*) on family's income (*realinc*).

```
proc reg data = GSS2018;
      model realrinc = rhs age /stb;
run;
```

- The dependent variable (*realrinc*) is placed in the `model` statement before the equal sign.
- The independent variables (*age*, and *rhs*) are placed in the `model` statement after the equal sign.
- The `stb` option is specified on the `model` statement to obtain standardized regression coefficients in addition to unstandardized regression coefficients.

The REG Procedure
Model: MODEL1
Dependent Variable: realrinc R's income in constant $

Number of Observations Read	2202
Number of Observations Used	553
Number of Observations with Missing Values	1649

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	15633534751	7816767376	14.64	<.0001
Error	550	2.936032E11	533824067		
Corrected Total	552	3.092368E11			

Root MSE	23105	R-Square	0.0506
Dependent Mean	20791	Adj R-Sq	0.0471
Coeff Var	111.12981		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	12491	3673.32222	3.40	0.0007	0
rhs	R health rating	1	-3809.56781	1470.92152	-2.59	0.0099	-0.10960
age	Age of respondent	1	350.43124	67.95197	5.16	<.0001	0.21824

- **Significance:** The **Analysis of Variance** table measures whether the linear relationship between the independent variables and dependent variable is significant. In this example, the F-statistic is below 0.05. Thus, the groups of three variables significantly predicts income.
- **Model Fit:** Since the denominator of R-squared is fixed, only the numerator can increase. Thus, each additional variable used in the equation increases the size of the numerator only. As a result, the introduction of additional variables produces a higher R-squared value, regardless of their usefulness. The Adjusted R-Squared value corrects this problem by adjusting both the numerator and the denominator by their respective degrees of freedom. Unlike R-squared, the Adjusted R-squared can decrease as more predictors are added to a model. It is useful when comparing models, because it balances the goals of model fit and model simplicity.

In this example, compared to the bivariate model in the previous section, both R-squared and Adjusted R-squared have increased. Hence, this new model explains more variation of the dependent variable than the old model with one independent variable, and the additional variation explained is worth the cost of making the model more complex.

- **Parameter Estimates:** The Parameter Estimates tables provide the values of the unstandardized regression coefficients (Parameter Estimate), standardized coefficients (Standardized Estimate), and a significance test of the coefficients (Pr > |t|). It shows that a unit increase in *rhs* will decrease predicted income by $3809.57 (higher rhs means worse health condition) while holding *age* constant. The t-test statistic is −2.59, and the p-value associated with the t-test statistic is`<0.01. Thus, *rhs* is statistically significant at the 0.01 level while controlling for *age*.

To compare the magnitude of the effect of each independent variable, the coefficients must be standardized using the same scale. The standardized coefficient (*Standardized Estimate*) measures change in the dependent variable (measured in standard deviations) per one standard deviation change in the independent variable.

Because the absolute value of standardized coefficient for *age* is larger than that for *rhs* we can conclude that age has a larger impact on real income than health ratings.

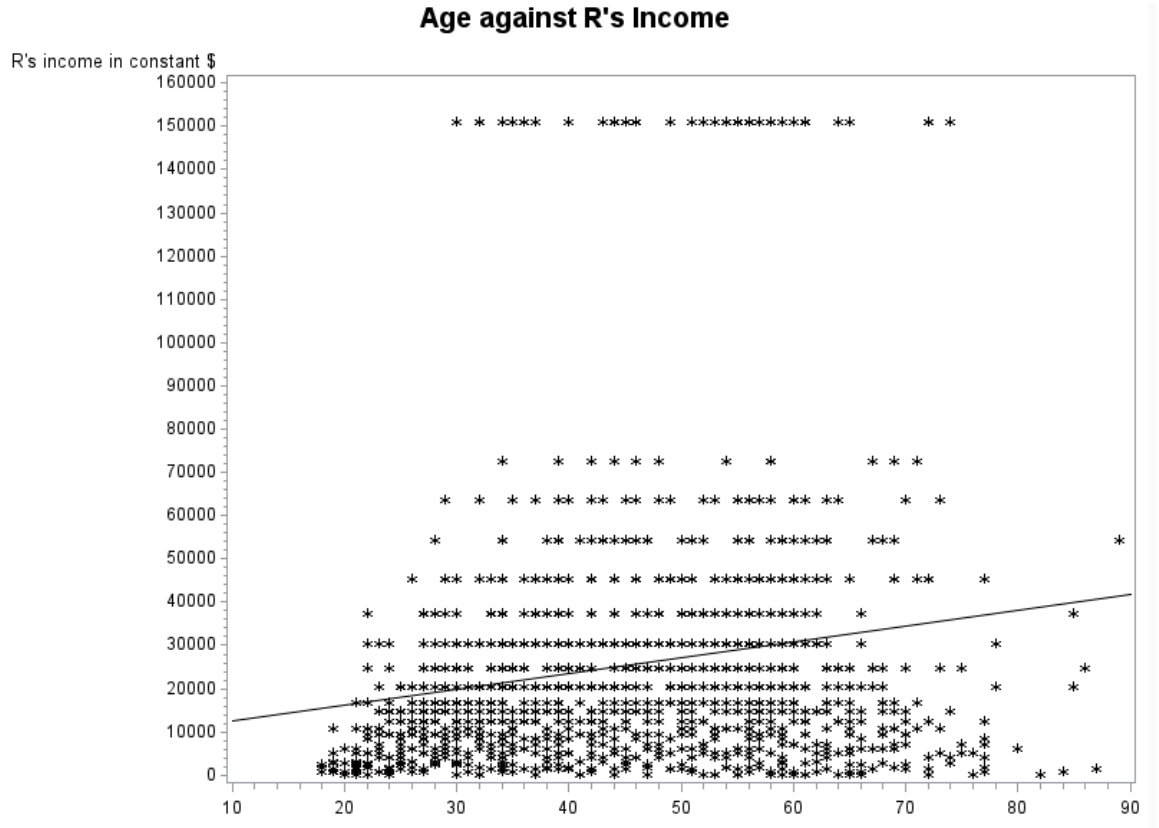
Exercise 6

Run multiple regression to assess the effect of occupational prestige index of a respondent's parents (*faos* and *moos*) and respondent's gender (*Male*) on the respondent's occupational prestige (*pres*). (See Appendix A for answers.)

Plotting a Regression Line on a Scatter Plot

The relationship between respondents' education and income can be visually portrayed in a scatter plot, using the `proc gplot` procedure. In addition, you can overlay the regression line and confidence interval range on the scatter plot.

```
symbol color = black value = star interpol = r;
title "Age against R's Income";
proc gplot data = gss2018;
      plot  realrinc*age;
run;
quit;
```



- `symbol` statements give `proc gplot` information about plot characters, plot lines, color, and interpolation (smoothing of lines).
- The color black is specified in the `color` option.
- The character `star` is specified in the `value` option
- The `interpol=r` option indicates that the plot is a regression analysis.
- The dependent variable (*realrinc*) appears first in the `plot` statement, before the *, and the independent variable (*age*) appears second, after the *.

> ! Important
>
> If a second symbol is present in a plot, the symbol statement must include a reference to which symbol is being defined by adding a number to symbol. The first symbol would be referred to as `symbol1` and the second symbol would be referred to as `symbol2`.

APPENDIX A: ANSWERS FOR EXERCISES

SAS II Complete Script

Exercise 1

Perform a chi-square test to determine if educational attainment (*degree*) is significantly associated with the belief that marijuana should be legal (*grass*). * How many high school graduate (*degree*

= 1) said Yes to legalization *(grass* = 1)? * Is there a statistically significant relationship between the two variables? Which chi-square test statistic is appropriate?

```
proc freq data = gss2018;
      tables grass*degree / chisq expected;
run;
```

The SAS System

The FREQ Procedure

Frequency Expected Percent Row Pct Col Pct	Table of grass by degree					
grass(Should marijuana be made legal)	degree(R's highest degree)					
	0	1	2	3	4	Total
1	81 98.531 5.97 9.22 53.29	438 429.13 32.30 49.83 66.16	78 77.139 5.75 8.87 65.55	190 177.62 14.01 21.62 69.34	92 96.586 6.78 10.47 61.74	879 64.82
2	71 53.469 5.24 14.88 46.71	224 232.87 16.52 46.96 33.84	41 41.861 3.02 8.60 34.45	84 96.385 6.19 17.61 30.66	57 52.414 4.20 11.95 38.26	477 35.18
Total	152 11.21	662 48.82	119 8.78	274 20.21	149 10.99	1356 100.00
Frequency Missing = 846						

Statistics for Table of grass by degree

Statistic	DF	Value	Prob
Chi-Square	4	12.4898	0.0141
Likelihood Ratio Chi-Square	4	12.2004	0.0159
Mantel-Haenszel Chi-Square	1	1.8704	0.1714
Phi Coefficient		0.0960	
Contingency Coefficient		0.0955	
Cramer's V		0.0960	

Sample Size = 1356
Frequency Missing = 846

WARNING: 38% of the data are missing.

1. 438 high school graduates said “Yes” to legalization of marijuana.
2. There is a statistically significant relationship between the two variables. The *p-value* associated with the chi-square statistic (0.0141) is less than 0.05. We know that the chi-square test is appropriate because the expected cell count for all cells exceeds five.

Exercise 2

Find out if there is a statistically significant difference in terms of respondents family income (*fmi*) between resondents who took college-level science courses and those who did not (*colsci*).

```
proc ttest data = GSS2018;
      class colsci;
      var fmi;
run;
```

Variable: fmi (Family income in constant $)

colsci	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
1		458	42888.3	34493.1	1611.8	227.0	119879
2		543	24488.5	23470.6	1007.2	227.0	119879
Diff (1-2)	Pooled		18399.8	29036.9	1842.2		
Diff (1-2)	Satterthwaite		18399.8		1900.6		

colsci	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		42888.3	39720.9	46055.7	34493.1	32394.6	36884.5
2		24488.5	22509.9	26467.0	23470.6	22152.8	24956.4
Diff (1-2)	Pooled	18399.8	14784.8	22014.9	29036.9	27817.7	30368.7
Diff (1-2)	Satterthwaite	18399.8	14669.0	22130.7			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	999	9.99	<.0001
Satterthwaite	Unequal	782.95	9.68	<.0001

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	457	542	2.16	<.0001

1. Because the Equality of Variances test is significant $(0.0001 < 0.05)$, the Satterthwaite test is appropriate.
2. The Satterthwaite *t*-test statistic is significant $(0.0001 < 0.05)$, therefore the amount of income respondents made differs significantly based on if they did or did not take college-level science courses.

Exercise 3

Using the one-way ANOVA procedure, test whether respondents family income (*fmi*) is different among different education groups (*degree*). Remember to check the Test of homogeneity of variance to determine if the Welch test is required.

```
proc glm data = gss2018;
      class degree;
      model fmi = degree;
      means degree / hovtest;
run;
```

- The dependent variable (*fmi*) and grouping variable (*degree*) are both identified in the `model` statement in the expression `fmi=degree`.
- The grouping variable (*degree*) is identified in both the `class` and `means` statements. The `means` statement is used to obtain descriptive statistics for the dependent variable by grouping variable.
- The `hovtest` option is specified in the `means` statement to request the Levene's test of the homogeneity of variance.

The GLM Procedure

Levene's Test for Homogeneity of fmi Variance
ANOVA of Squared Deviations from Group Means

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
degree	4	2.399E20	5.999E19	22.74	<.0001
Error	2010	5.302E21	2.638E18		

From the Levene's results table, *p-value* is smaller than 0.05. We reject the null hypothesis that the variances are equal, and not-equal variances are assumed. The result confirms that we should use the Welch test for non-equal variances, instead of regular ANOVA procedure results.

```
proc glm data = gss2018;
      class degree;
      model fmi = degree;
      means degree / welch;
run;
```

The GLM Procedure

Dependent Variable: fmi Family income in constant $

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	317076843856	79269210964	96.58	<.0001
Error	2010	1.6497118E12	820752158.76		
Corrected Total	2014	1.9667887E12			

R-Square	Coeff Var	Root MSE	fmi Mean
0.161216	84.60403	28648.77	33862.18

Source	DF	Type I SS	Mean Square	F Value	Pr > F
degree	4	317076843856	79269210964	96.58	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
degree	4	317076843856	79269210964	96.58	<.0001

From the results above,we can conclude that differences in family income among educational levels are statistically significant.

Exercise 4

Compute the correlation between: the highest year of school completed (*educ*), the respondent's father's occupational prestige index (*faos*), and the respondent's mother's occupational prestige index (*moos*) using (1) pairwise deletion and (2) listwise deletion and compare how the sample size changes.

- Pairwise Deletion:

```
proc corr data = gss2018;
      var educ faos moos;
run;
```

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
educ	2199	13.74488	2.94613	30225	0	20.00000	Highest year of school completed
faos	1727	44.48929	13.01015	76833	16.00000	80.00000	Father's occupational prestige score
moos	1565	41.86645	13.03117	65521	16.00000	80.00000	Mothers occupational prestige score

Pearson Correlation Coefficients Prob > |r| under H0: Rho=0 Number of Observations			
	educ	faos	moos
educ Highest year of school completed	1.00000 2199	0.25802 <.0001 1725	0.28029 <.0001 1564
faos Father's occupational prestige score	0.25802 <.0001 1725	1.00000 1727	0.23308 <.0001 1156
moos Mothers occupational prestige score	0.28029 <.0001 1564	0.23308 <.0001 1156	1.00000 1565

Educational attainment (*educ*) is significantly and positively correlated with mother's and father's occupational prestige index. Mother's and father's occupational index are also significantly and positively correlated.

- Listwise Deletion:

```
proc corr data = GSS2018 nomiss;
      var educ faos moos;
run;
```

Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
educ	1155	14.21558	2.76229	16419	0	20.00000	Highest year of school completed
faos	1155	44.47532	12.98853	51369	16.00000	80.00000	Father's occupational prestige score
moos	1155	42.64502	13.17659	49255	16.00000	80.00000	Mothers occupational prestige score

Pearson Correlation Coefficients, N = 1155 Prob > |r| under H0: Rho=0			
	educ	faos	moos
educ Highest year of school completed	1.00000	0.24787 <.0001	0.25625 <.0001
faos Father's occupational prestige score	0.24787 <.0001	1.00000	0.23308 <.0001
moos Mothers occupational prestige score	0.25625 <.0001	0.23308 <.0001	1.00000

While the directions of the relationships between the variables do not change significantly if you use either pairwise or listwise deletion, the values of the correlation coefficients do change depending on the treatment of missing cases.

Pairwise deletion should include more samples, because it includes more cases based on how it treats missing variables.

Exercise 5

Does age (*age*) have a significant linear relationship with performance on a vocabulary test (*wordsum*)? Treat *wordsum* as the dependent variable.

```
proc reg data = gss2018;
      model wordsum = age /stb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: wordsum Number words correct in vocabulary test

Number of Observations Read	2202
Number of Observations Used	1443
Number of Observations with Missing Values	759

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	38.74140	38.74140	9.53	0.0021
Error	1441	5855.46650	4.06347		
Corrected Total	1442	5894.20790			

Root MSE	2.01581	R-Square	0.0066
Dependent Mean	5.93139	Adj R-Sq	0.0059
Coeff Var	33.98537		

Parameter Estimates							
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	5.47263	0.15777	34.69	<.0001	0
age	Age of respondent	1	0.00921	0.00298	3.09	0.0021	0.08107

- Because the F-statistic is statistically significant, age is a significant predictor of vocabulary test score at the 0.0001 level.
- .6 percent of the variation in vocabulary test score is explained by age.
- The regression coefficient for independent variable age is 0.00921 and the constant for the model is 5.4726 The predicted vocabulary test score will increase by 0.00921 for each additional year of age.

Exercise 6

Run multiple regression to assess the effect of the occupational prestige index of a respondent's parents (*faos* and *moos*) and respondent's gender (*Male*) on the respondent's occupational prestige (*pres*).

```
proc reg data = gss2018;
      model pres = moos faos male /stb;
run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: pres R's occupational prestige score

Number of Observations Read	2202
Number of Observations Used	1122
Number of Observations with Missing Values	1080

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	11297	3765.54379	22.01	<.0001
Error	1118	191246	171.06076		
Corrected Total	1121	202543			

Root MSE	13.07902	R-Square	0.0558
Dependent Mean	46.28075	Adj R-Sq	0.0532
Coeff Var	28.26017		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Standardized Estimate
Intercept	Intercept	1	32.96056	1.72859	19.07	<.0001	0
moos	Mothers occupational prestige score	1	0.16619	0.03041	5.47	<.0001	0.16318
faos	Father's occupational prestige score	1	0.14244	0.03088	4.61	<.0001	0.13780
Male	Male Sex(Recoded)	1	-0.22398	0.78349	-0.29	0.7750	-0.00832

- Because the F-statistic is statistically significant, as a group, mother's occupational prestige index score, father's occupational prestige index score, and sex are significant predictors of respondent's occupational prestige score.
- The predicted occupational prestige score will increase by 0.166 for each unit increase of mother's occupational prestige score, 0.142 for each unit increase of father's occupational prestige score, and −0.224 if the respondent is male.
- The effects of *faos* on *moos* are statistically significant at the 0.001 level. The effect of sex (*male*) is not statistically significant.
- 5.58 percent of the variation in *pres* is explained by the three predictors.