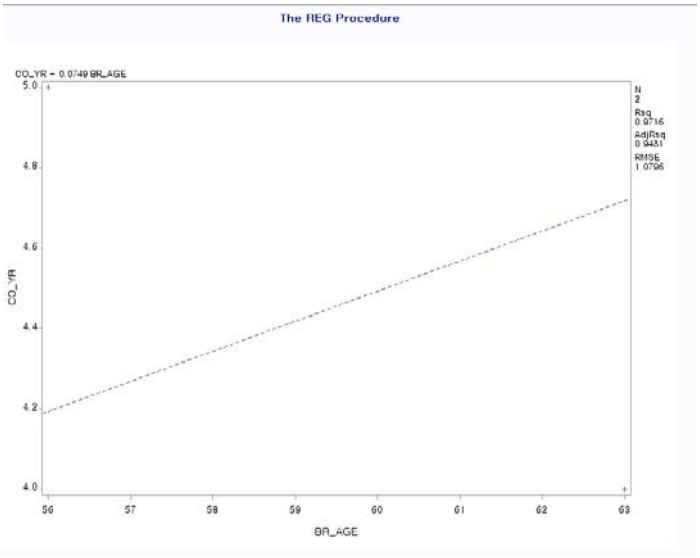
< Simple regression analysis >

```
proc contents data=a.thesis_total ; run;

proc reg data=a.thesis_total;
model CO_YR = BR_AGE/ noint;
plot CO_YR*BR_AGE;
RUN;
```



SAS 시스템

The REG Procedure
Model: MODEL1
Dependent Variable: CO_YR

Number of Observations Read	173357
Number of Observations Used	2
Number of Observations with Missing Values	173355

Note: No intercept in model. R-Square is redefined.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	39.83448	39.83448	34.18	0.1079
Error	1	1.16552	1.16552		
Uncorrected Total	2	41.00000			

Root MSE	1.07959	R-Square	0.9716
Dependent Mean	4.50000	Adj R-Sq	0.9431
Coeff Var	23.99092		

Parameter Estimates					
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
BR_AGE	1	0.07488	0.01281	5.85	0.1079

DF

- Degrees of freedom. This term refers to the number of values that are free to vary

F values

- The **value** of Prob(**F**) is the probability that the null hypothesis for the full model is true (i.e., that all of the **regression** coefficients are zero).

- The F value is a value on the F distribution. Various statistical tests generate an F value. The value can be used to determine whether the test is statistically significant.
- The F value is used in analysis of variance (ANOVA). It is calculated by dividing two mean squares. This calculation determines the ratio of explained variance to unexplained variance.
- The F distribution is a theoretical distribution. There are many of these distributions, and each of them differs based on the degrees of freedom.
- The F value and the degrees of freedom of the sources of variance are used to determine the probability of the F value. The probability is the significance value for the test.

Source : https://www.ibm.com/docs/en/cognos-analytics/11.1.0?topic=terms-f-value

R-squared

- R-squared is expressed as a percentage between 0 and 100, with 100 signaling perfect correlation and zero no correlation at all. The figure does not indicate how well a particular group of securities is performing. It only measures how closely the 'returns' align with those of the measured benchmark. It is also backwards-looking—it is not a predictor of future results.

Adjusted R-squared

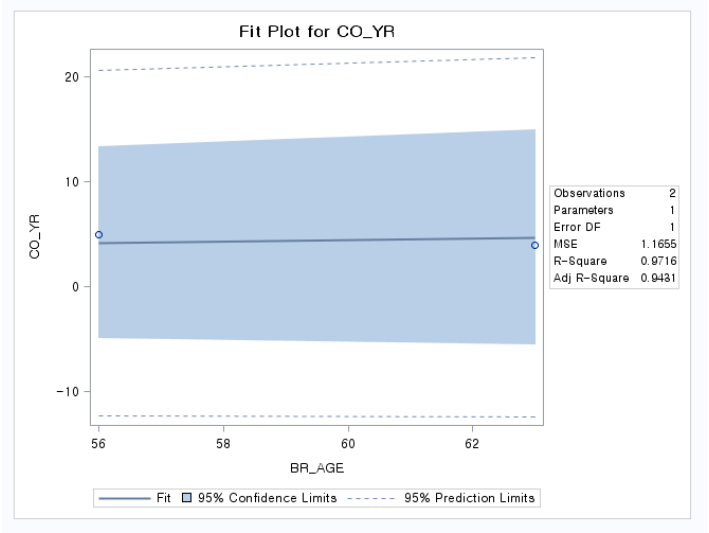
- Adjusted R-squared can provide a 'more precise' view of that correlation by also taking into account how many independent variables are added to a particular model. This is done because such additions of independent variables usually increase the reliability of that model—meaning, for investors, the correlation with the index.

Source : https://www.investopedia.com/ask/answers/012615/whats-difference-between-rsquared-and-adjusted-rsquared.asp

MSE

- The MSE either assesses the quality of a **predictor** (i.e., a function mapping arbitrary inputs to a sample of values of some random variable), or of an **estimator** (i.e., a mathematical function mapping a sample of data to an estimate of a parameter of the population from which the data is sampled). MSE measures the mean square difference between the estimated value and the actual value.

Source : https://en.wikipedia.org/wiki/Mean_squared_error#Definition_and_basic_properties



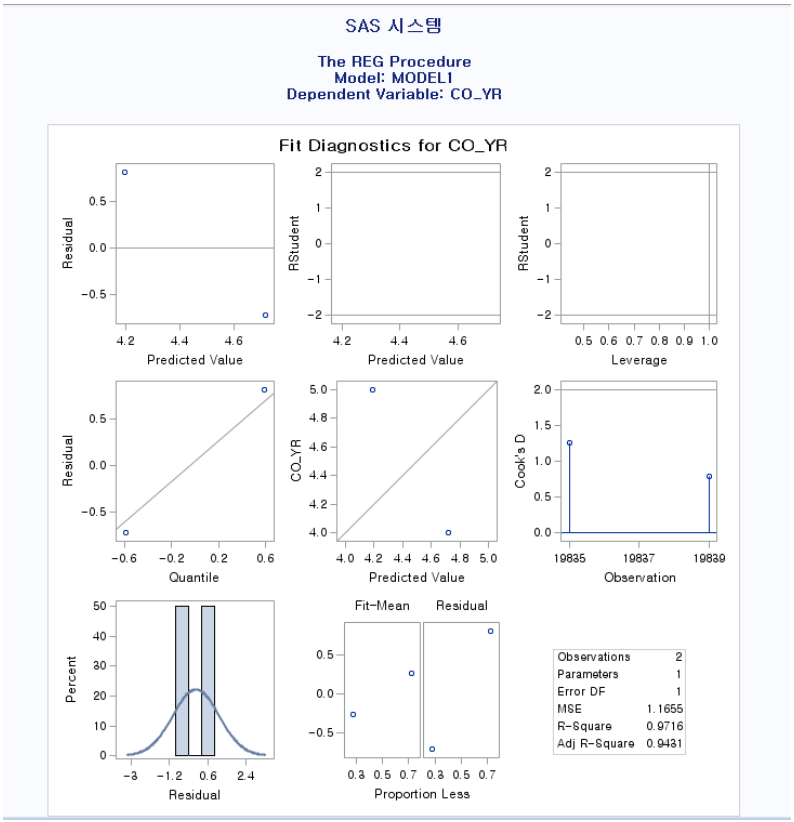
Confidence limits

- Confidence limits for the mean (Snedecor and Cochran, 1989) are an interval estimate for the mean. Interval estimates are often desirable because the estimate of the mean varies from sample to sample. Instead of a single estimate for the mean, a confidence interval generates a lower and upper limit for the mean. The interval estimate gives an indication of how much uncertainty there is in our estimate of the true mean. The narrower the interval, the more precise is our estimate.

  Confidence limits are expressed in terms of a confidence coefficient. Although the choice of confidence coefficient is somewhat arbitrary, in practice 90 %, 95 %, and 99 % intervals are often used, with 95 % being the most commonly used.

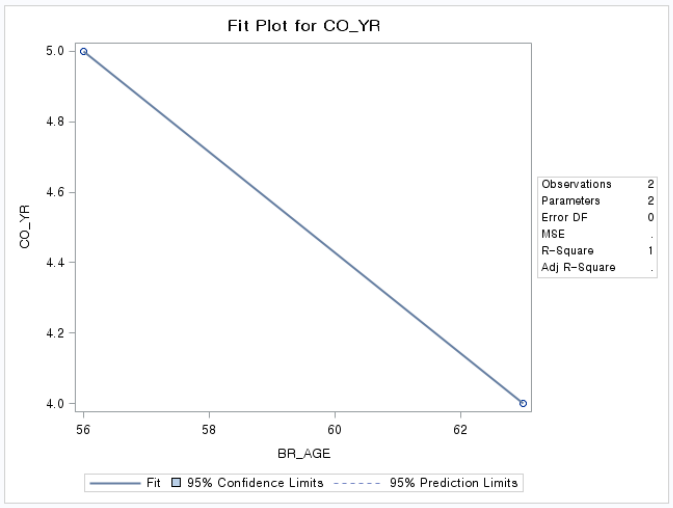
  As a technical note, a 95 % confidence interval does not mean that there is a 95 % probability that the interval contains the true mean. The interval computed from a given sample either contains the true mean or it does not. Instead, the level of confidence is associated with the method of calculating the interval. The confidence coefficient is simply the proportion of samples of a given size that may be expected to contain the true mean. That is, for a 95 % confidence interval, if many samples are collected and the confidence interval computed, in the long run about 95 % of these intervals would contain the true mean.

  Source: https://www.itl.nist.gov/div898/handbook/eda/section3/eda352.htm



< Multiple linear regression analysis >

```
proc reg data=a.thesis_total;
model CO_YR = BR_AGE  CO_AGE ;
plot CO_YR*BR_AGE*CO_AGE;
RUN;
```



< Testing differences between two group means >

Testing

- In statistics, "testing" refers to a series of processes in which observed statistics are determined whether or not statistically significant. At this time, we need to be particularly careful that we use the expression "statistically significant," not just the expression "significant."

Source: https://babilusa.tistory.com/42

```
proc ttest data=a.thesis_total;
class DS1_SEX ;
VAR cbmi;
RUN;
```

SAS 시스템

The TTEST Procedure

Variable: cbmi

DS1_SEX	N	Mean	Std Dev	Std Err	Minimum	Maximum
1	59293	629.7	7755.7	31.8507	13.0869	99999.0
2	114059	527.7	7080.5	20.9653	12.9613	99999.0
Diff (1-2)		102.0	7318.5	37.0526		

DS1_SEX	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
1		629.7	567.3	692.1	7755.7	7711.8	7800.1
2		527.7	486.6	568.8	7080.5	7051.6	7109.7
Diff (1-2)	Pooled	102.0	29.3879	174.6	7318.5	7294.2	7342.9
Diff (1-2)	Satterthwaite	102.0	27.2731	176.7			

Method	Variances	DF	t Value	Pr > |t|
Pooled	Equal	173350	2.75	0.0059
Satterthwaite	Unequal	110973	2.68	0.0075

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	59292	114058	1.20	<.0001

Pooled vs Satterthwaite

- The main difference is that the **Satterthwaite** approximation does not assume equal variances, whereas the **pooled** method does. In other words, you can always use the **Satterthwaite** method and be correct, but you can only use the **pooled** method in very specific (and rare) circumstances.

Source: https://www.bgsu.edu/content/dam/BGSU/college-of-arts-and-sciences/center-for-family-and-demographic-research/documents/Help-Resources-and-Tools/Statistical%20Analysis/Annotated-Output-T-Test-SAS.pdf

< ANOVA >

The distribution used for ANOVA is the F-distribution.

Total Variation represents the overall variation of the response variable.

$$\Sigma\Sigma(Y_{ij} - \bar{\bar{Y}})^2$$

- Calculating the allows you to obtain the total variation SST (Total Sum of Squares.)

- Between Group Variation is the variation described by an independent variable.

$$\Sigma n_i(\bar{Y}_i - \bar{\bar{Y}})^2$$

- Calculating the allows you to obtain Model Sum of Squares (SSMs) of variation between groups.

- Within Group Variation is a variation that is not described by the model.

$$\Sigma\Sigma(Y_{ij} - \bar{Y}_i)^2$$

- Calculating the allows you to obtain within-group error sum of squares (SSE).

SST=SSM+SSE

The F statistic used for ANOVA can be calculated as follows:

$$F_{(Model\ \ df, Error\ \ df)} = \frac{MSM}{MSE} = \frac{SSM/DF_M}{SSE/DF_E}$$

$$R^2 = SSM/SST$$

You can also obtain a coefficient of determination in ANOVA. The coefficient of determination can be obtained as the proportion of variation described by the model, that is, the proportion of variability described by the independent variable.

```
proc ANOVA data=a.thesis_total;
class DS1_MARRY_A;
MODEL BR_AGE = DS1_MARRY_A;
MEANS effects
RUN;
```

Dependent Variable: BR_AGE

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	14	11777326499	841237607.05	7.86	<.0001
Error	905	96800898730	106962319.04		
Corrected Total	919	108578225229			

R-Square	Coeff Var	Root MSE	BR_AGE Mean
0.108469	831.9579	10342.26	1243.123

Source	DF	Anova SS	Mean Square	F Value	Pr > F
DS1_JOB	14	11777326499	841237607	7.86	<.0001