Chapter 5 Regression Analysis in SAS

In this chapter, we will introduce how to perform correlation and regression analysis.

5.1 Correlation

Using PROC CORR to compute correlation coefficients. Example data: *sashelp.class*. We look at relationship between Height and Weight. Code:

```
*Pearson correlation analysis*;
proc corr data=sashelp.class cov;
  var height weight;
run;
```

- By default, Pearson correlation is computed using non-missing values.
- PROC CORR line: specify dataset
  - COV option provides covariance matrix.
- VAR statement: specify variables of interest to be analyzed. If there are more than two variables, it will produce correlation coefficients for each pair.

Results:

The CORR Procedure

2 Variables:	Height	Weight

Covariance Matrix, DF = 18

	Height	Weight
Height	26.2869006	102.4934211
Weight	102.4934211	518.6520468

Covariance matrix

Var (Height) Cov(Height, Weight)
Cov(Height, Weight) Var(Weight)

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum
Height	19	62.33684	5.12708	1184	51.30000	72.00000
Weight	19	100.02632	22.77393	1901	50.50000	150.00000

Descriptive statistics

Pearson Correlation Coefficients, N = 19
Prob > |r| under H0: Rho=0

	Height	Weight
Height	1.00000	0.87779 <.0001
Weight	0.87779 <.0001	1.00000

Correlation matrix with test of Pearson correlation

Cor(Height, Weight) = .88, $p < .0001$

Conclusion: There is a strong positive correlation between height and weight.

Supplemental Material

Pearson correlation coefficient

- Purpose: Examine the strength and direction of a relationship between two continuous variables.
- Equation: $r_{xy} = \frac{s_{xy}}{s_x s_y}$
- Range: $-1 \leq r_{xy} \leq 1$
  - Closer to 1 → stronger positive relationship between *x* and *y*
  - Closer to -1 → stronger negative relationship between *x* and *y*
  - Closer to 0 → weaker relationship between *x* and *y*

Test of correlation

- Hypotheses
  - Null hypothesis: population correlation = 0 (no correlation)
  - Alternative hypothesis: population correlation ≠ 0
- If p-value is smaller than pre-specified alpha (Type-I error rate), we reject the null hypothesis claiming a significant correlation between *x* and *y*.

5.2 Linear Regression

Introduce PROC REG for linear regression analysis including model diagnosis and variable selection. We will look at three examples.

Supplemental material

Regression analysis models the relationship between a response or outcome variable and another set of variables. This relationship is expressed through a statistical model equation that predicts a *response variable* (also called a *dependent variable* or *criterion*) from a function of *regressor variables* (also called *independent variables, predictors, explanatory variables, factors,* or *carriers*) and *parameters*.

- Equation: $Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \varepsilon_i$.
- Parameters of interest: $\beta_0, \beta_1, \ldots, \beta_k,$ – unknown regression coefficients
  - ε_i: Random error/residual for each observation *i*.
- Prediction equation: $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_k X_{ki}$
- Test of Coefficients
  - Null hypothesis for β_0: $\beta_0 = 0$
  - Null hypothesis for β_1: $\beta_1 = 0$
  - …
  - Null hypothesis for β_k: $\beta_k = 0$

Example I – Simple linear regression

Use height to predict weight -- $Weight_i = \beta_0 + \beta_1 Height_i + \varepsilon_i$.

```
*Simple linear regression with PROC REG*;
proc reg data=sashelp.class;
   model Weight = Height / clb cli p r;
   output out=out_set1 p=predicted r=residual;
run;
```

MODEL statement

- Specify model: Dependent variable (Y) = list of independent variables (X's)
- Options (after /)
  - CLB: $100(1-\alpha)\%$ confidence interval (CI) for parameter estimates; by default, $\alpha = .05$ → 95% CI for each β_k.
  - CLI: $100(1-\alpha)\%$ CI for an individual predicted value ($\beta_0 + \beta_1 X_{1i} + \varepsilon_i$) – prediction limits.
  - P: requests for predicted values ($\hat{Y}_i$).
  - R: produces residuals ($\varepsilon_i = Y_i - \hat{Y}_i$)
- Other options
  - CLM: $100(1-\alpha)\%$ CI for the expected value (mean) of the dependent variable for each observation *i* ($\beta_0 + \beta_1 X_{1i}$) – confidence limits.
  - ALPHA: change the default (.05) to a user-defined significance level for computing confidence and prediction intervals.
  - LCL and UCL: lower and upper limit of CLI.
  - LCLM and UCLM: lower and upper limit of CLM.
  - DW: Durbin-Watson statistic (examine autocorrelation issue)

Results:

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Number of Observations Read	19
Number of Observations Used	19

Number of valid responses

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	7193.24912	7193.24912	57.08	<.0001
Error	17	2142.48772	126.02869		
Corrected Total	18	9335.73684			

Overall usefulness of model (ANOVA)
The F-test indicates the model explains a significant amount of variation in the dependent variable, $F(1, 17) = 57.08, p < .0001$.

Root MSE	11.22625	R-Square	0.7705
Dependent Mean	100.02632	Adj R-Sq	0.7570
Coeff Var	11.22330		

Root MSE: residual standard error (estimate of standard deviation of the residuals)
R-Square: amount of variation explained by model
Adj R-Sq: adjusted R-squared (adjusted for the number of predictors in the model)

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-143.02692	32.27459	-4.43	0.0004	-211.12035	-74.93348
Height	1	3.89903	0.51609	7.55	<.0001	2.81017	4.98789

Estimates of β_0 and β_1

t-tests for parameters

95% CI for parameter estimates

Conclusion:

- Prediction equation: $Predicted\ Weight_i = -143.027 + 3.899 Height_i$.
- Both the intercept and regression coefficient for *Height* are significantly different from 0.
    - Intercept: predicted weight when height equals to zero
    - Height estimate: predicted change in weight with one unit increase in height.

More results from SAS output:

I. Case wise statistics

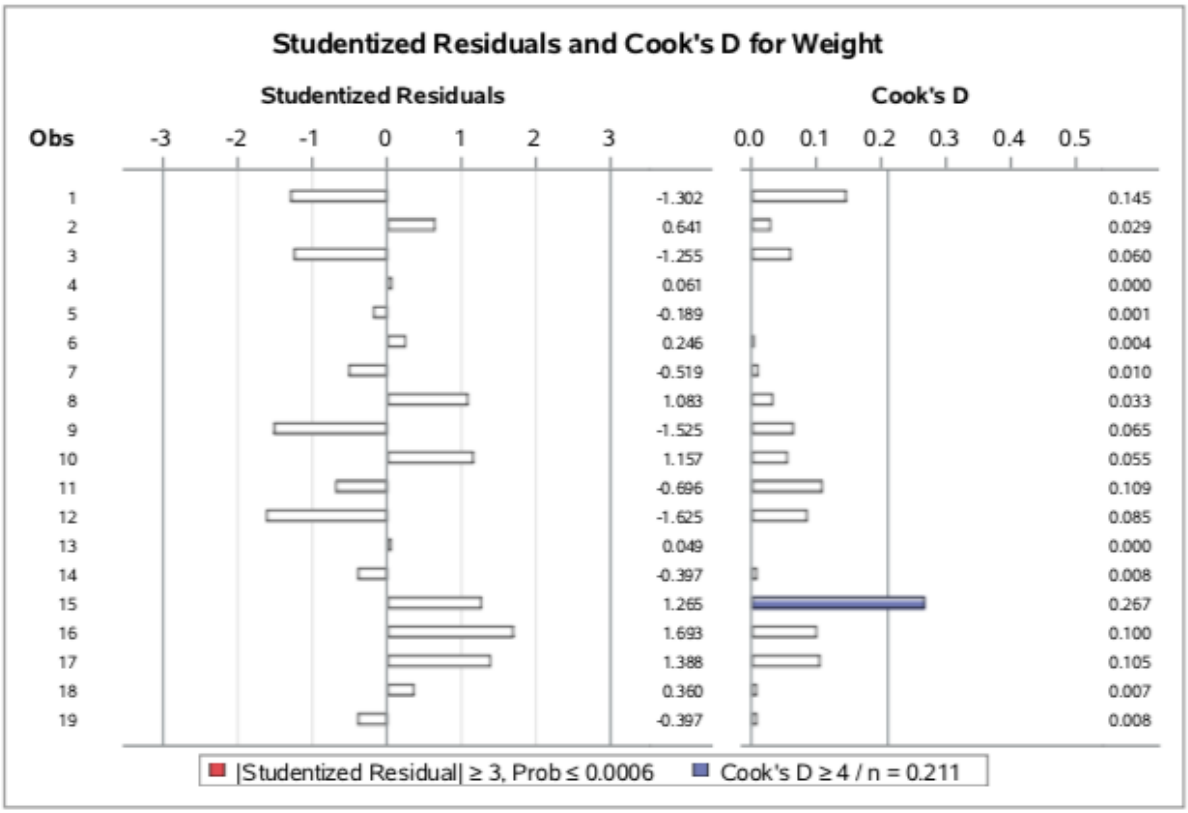
Obs – observation ID; Dependent Variable: observed *y* value (Y_i); Predicted Value: predicted *y* value ($\hat{Y}_i$); 95% CL Predict: CLI – prediction limits; Residual: $\varepsilon_i = Y_i - \hat{Y}_i$; Student Residual: studentized residuals; Cook's D: Cook's Distance for identifying outliers

The REG Procedure
Model: MODEL1
Dependent Variable: Weight

Output Statistics									
Obs	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Predict		Residual	Std Error Residual	Student Residual	Cook's D
1	112.5	126.0062	4.2963	100.6456	151.3668	-13.5062	10.372	-1.302	0.145
2	84.0	77.2683	3.9633	52.1503	102.3863	6.7317	10.503	0.641	0.029
3	98.0	111.5798	2.9953	87.0659	136.0936	-13.5798	10.819	-1.255	0.060
4	102.5	101.8322	2.5865	77.5263	126.1380	0.6678	10.924	0.061	0.000
5	102.5	104.5615	2.6445	80.2279	128.8951	-2.0615	10.910	-0.189	0.001
6	83.0	80.3875	3.6593	55.4757	105.2993	2.6125	10.613	0.246	0.004
7	84.5	90.1351	2.8892	65.6780	114.5922	-5.6351	10.848	-0.519	0.010
8	112.5	100.6625	2.5769	76.3612	124.9637	11.8375	10.927	1.083	0.033
9	84.0	100.6625	2.5769	76.3612	124.9637	-16.6625	10.927	-1.525	0.065
10	99.5	87.0159	3.0982	62.4451	111.5866	12.4841	10.790	1.157	0.055
11	50.5	56.9933	6.2512	29.8835	84.1032	-6.4933	9.325	-0.696	0.109
12	90.0	107.6807	2.7676	83.2863	132.0752	-17.6807	10.880	-1.625	0.085
13	77.0	76.4885	4.0423	51.3145	101.6624	0.5115	10.473	0.049	0.000
14	112.0	116.2586	3.3540	91.5388	140.9784	-4.2586	10.714	-0.397	0.008
15	150.0	137.7033	5.6129	111.2225	164.1840	12.2967	9.722	1.265	0.267
16	128.0	109.6302	2.8721	85.1821	134.0784	18.3698	10.853	1.693	0.100
17	133.0	118.2081	3.5249	93.3827	143.0335	14.7919	10.659	1.388	0.105
18	85.0	81.1673	3.5867	56.3025	106.0321	3.8327	10.638	0.360	0.007
19	112.0	116.2586	3.3540	91.5388	140.9784	-4.2586	10.714	-0.397	0.008

II. Displays of Studentized residuals and Cook's D for identifying outliers

The REG Procedure
Model: MODEL1
Dependent Variable: Weight



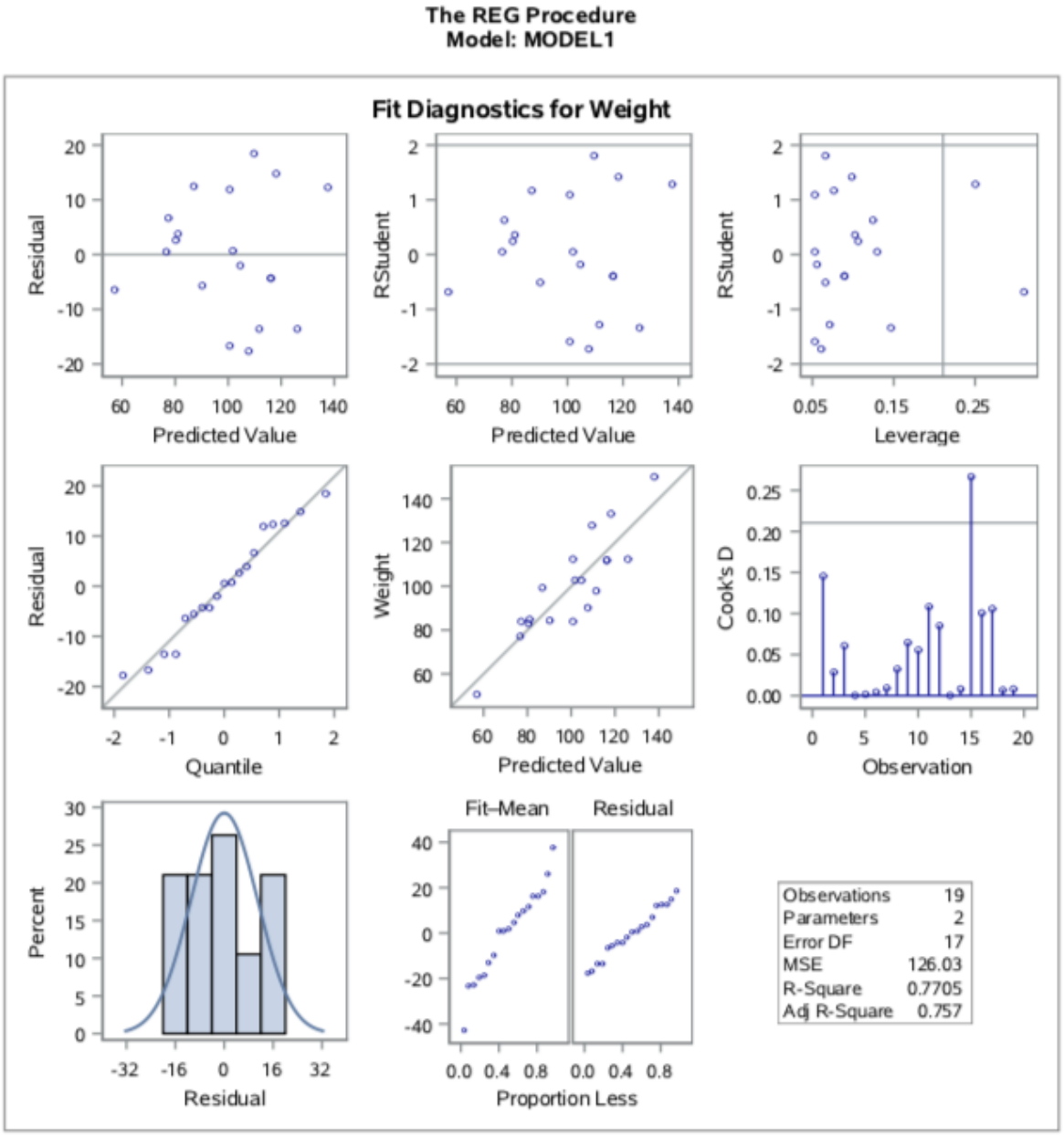
Sum of Residuals	0
Sum of Squared Residuals	2142.48772
Predicted Residual SS (PRESS)	2651.35206

III. The panel of regression diagnostics. It provides a more detailed look at the model-data fit.

- First row: raw residual plot; studentized residuals that take into account heterogeneity in the variability of the residuals; studentized residual-by-leverage plot shows that two observations have high leverage—that is, they are unusual in their height values relative to the other children.

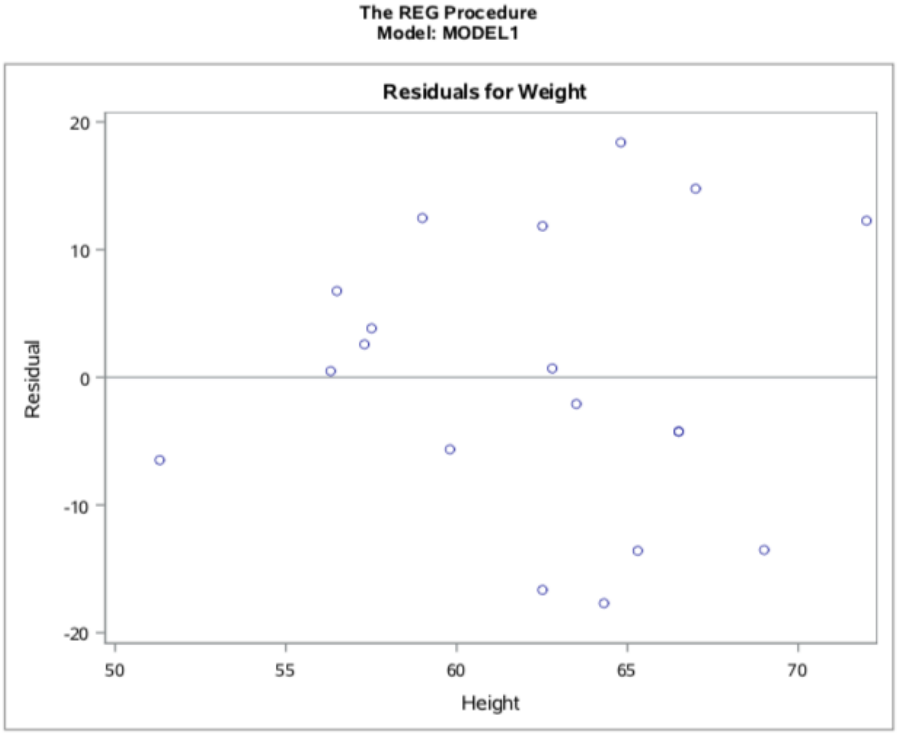
- Second row: normal Q-Q plot (examines the normality assumption); observed y against predicted y values; plot of Cook's *D*.

- Third row: histogram of residuals (assess the normality assumption).

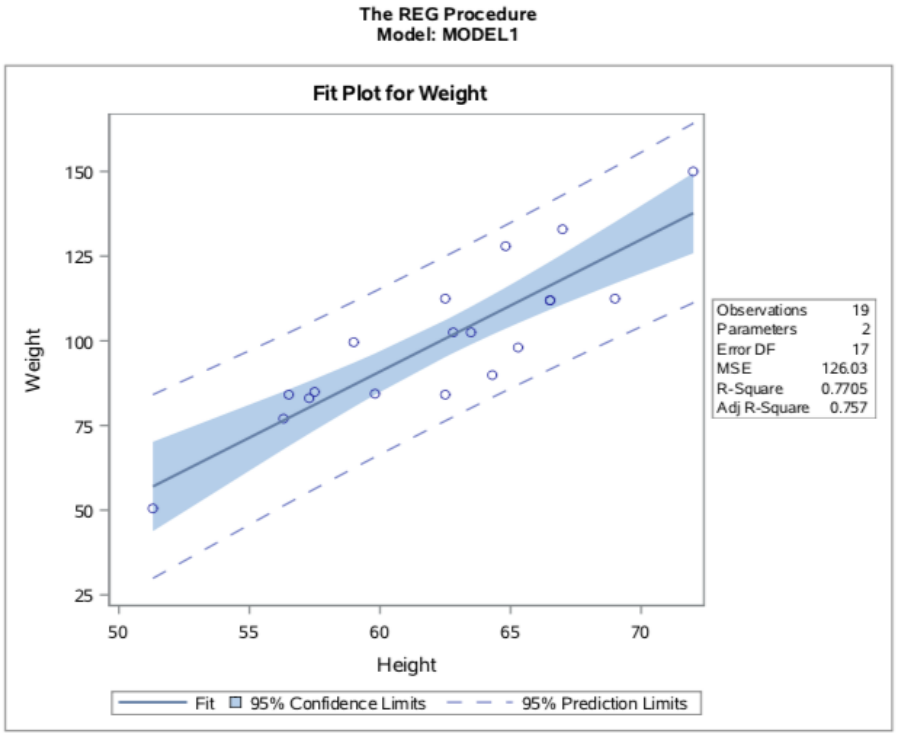


IV. Residual plot

If the children were randomly selected, the observations from different children should not be correlated. If the mean function of the model is also correctly specified, the residuals should scatter around the zero without discernible structure.



V. Prediction limits (CLI) and Confidence limits (CLM)



Example II – Multiple linear regression

Predict Weight using Height and Age

```
*Multiple linear regression with PROC REG*;
proc reg data=sashelp.class;
   model Weight = Height Age / clb cli p r;
   output out=out_set1 p=predicted r=residual;
run;
```

The only difference is the model specification in MODEL statement. Check out the results by yourself in SAS!

Example III – Variable selection

Find the best set of predictors to predict/explain the dependent variable. Use example data *sashelp.vote1980*. Specify a selection method as an option in MODEL statement.

a. Forward Selection: Add significant variables one by one until all significant predictors are included in the model. Need to specify an alpha value to determine significance – Entry criterion: SLENTRY. A variable is entered if p-value (of a F-test for R-square change) is smaller than SLENTRY.

```
*Variable selection -- Forward selection*;
proc reg data=sashelp.vote1980;
  model LogVoteRate = Pop Edu Houses Income / selection=forward slentry=.05;
run;
```

b. Backward Selection: Take out variables that are not significant predictors until only significant ones left in the model. Need to specify an alpha value to determine significance – Removal criterion: SLSTAY. A variable is removed if p-value (of a F-test for R-square change) is larger than SLSTAY.

```
*Variable selection -- Backward selection*;
proc reg data=sashelp.vote1980;
  model LogVoteRate = Pop Edu Houses Income / selection=backward slstay=.10;
run;
```

c. Stepwise Selection: Each time a variable is added to a model if it meets the entry criterion, the significance of each of the existing variables in the model is re-evaluated. That said, at each step, both a forward selection and a backward selection are performed. Therefore, both Entry criterion -- SLENTRY and Removal criterion -- SLSTAY need to be specified.

```
*Variable selection -- Stepwise selection*;
proc reg data=sashelp.vote1980;
  model LogVoteRate = Pop Edu Houses Income / selection=stepwise slentry=.05 slstay=.10;
run;
```

Output tables include model results at each step. The last step indicates the final model (best model) together with diagnostic plots. Check out results in SAS.