ST512 – Lab 01 *Solutions*

Part 2: Graphical and Numerical Summaries

1. Find the means and standard deviations for `yield` and `nclust`.

The easiest way to find the means and standard deviations for the two variables in the `grape` dataset is using PROC MEANS, although both of these statistics can be found for both variables in the PROC UNIVARIATE output as well. By adding both the `mean` and `std` options to the `proc means` statement in our code, the mean and standard deviation for both variables are provided in the table that is created after running the PROC MEANS procedure. The correct means and standard deviations for both variables can be seen in the table below.

Variable	Mean	Std Dev
yield	4.4750000	0.8518696
nclust	107.1008333	15.1407280

2. Find the five-number-summaries for `yield` and `nclust`.

The five-number-summaries can be found in the PROC UNIVARIATE output. One of the tables produced by this procedure, for each variable, is the quantiles table. All five numbers of the five-number-summary can be found in this table, along with other quantiles. The tables for both variables are shown below, with `yield` on the left and `nclust` on the right.

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	5.60
99%	5.60
95%	5.60
90%	5.40
75% Q3	5.05
50% Median	4.60
25% Q1	4.15
10%	3.20
5%	2.70
1%	2.70
0% Min	2.70

Quantiles (Definition 5)	
Quantile	Estimate
100% Max	125.240
99%	125.240
95%	125.240
90%	122.300
75% Q3	116.865
50% Median	113.280
25% Q1	95.405
10%	82.770
5%	80.190
1%	80.190
0% Min	80.190

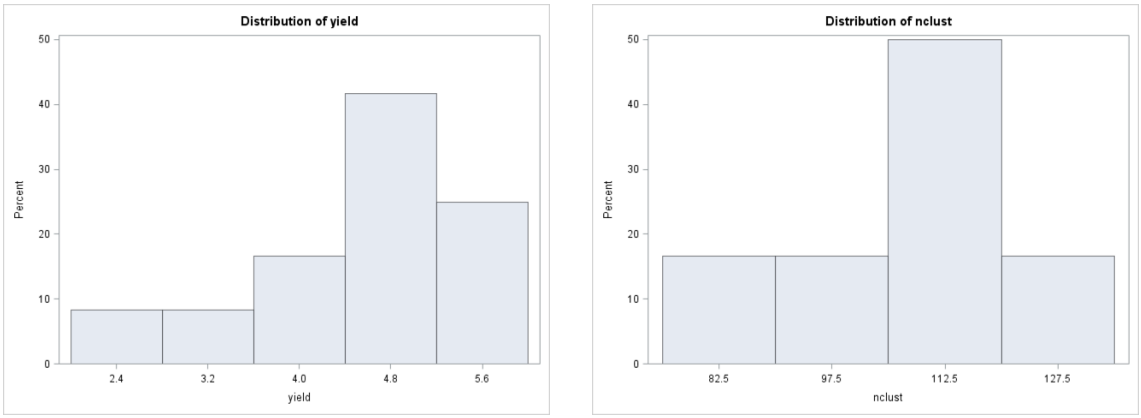
3. Find a 95% confidence interval for the mean of `yield`.

A 95% confidence interval for the mean can be found in the table generated by PROC MEANS because of the addition of the `clm` option to the `proc means` statement. By default, the `clm` option creates a 95% confidence interval, but we could instead create a confidence interval with a different confidence level by adding the `alpha=` option. The 95% confidence interval for the mean of `yield` can be seen in the table below.

Lower 95% CL for Mean	Upper 95% CL for Mean
3.9337479	5.0162521

4. Draw histograms to visualize the marginal distributions of `yield` and `nclust`. Do either of these distributions look normal?

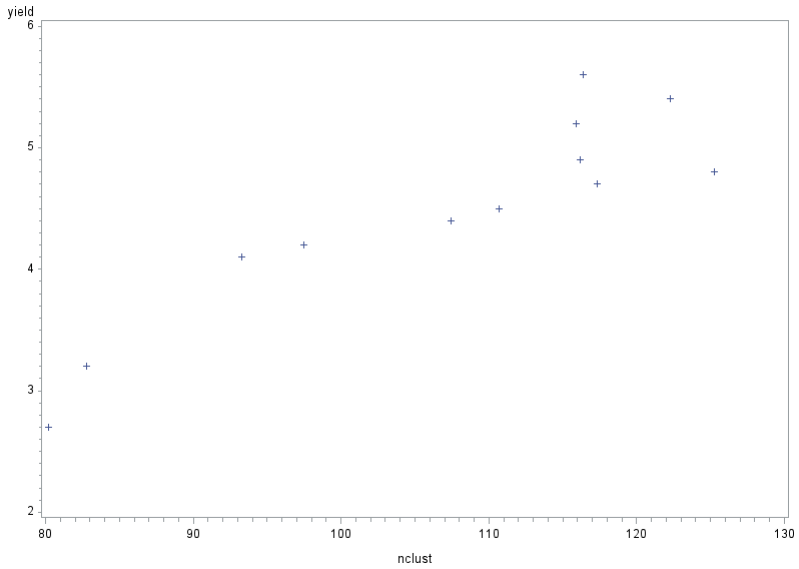
Histograms for the marginal distributions both variables are created by the PROC UNIVARIATE procedure after adding the `histogram` statement to the `proc` step and specifying the variables of which we want histograms. Both histograms can be seen below. The distribution of `yield` appears to be negatively skewed. The distribution of `nclust` looks like it could be normal, although it is also slightly negatively skewed.



Part 3: Simple Linear Regression in SAS

1. Draw a scatterplot of the data with `nclust` on the horizontal axis and `yield` on the vertical axis. Does the relationship between the two variables seem to be approximately linear?

A scatterplot can be made using the PROC PLOT procedure. This is done by specifying the vertical axis variable and the horizontal axis variable in the `plot` statement in the following manner: `plot 'vertical axis variable'*'horizontal axis variable'`. The `='o'` part of the `plot` statement specifies the character that is used to mark each observation on the plot. The scatterplot can be seen below. The relationship between the two variables does appear to be approximately linear.



2. Find the least-squares estimates of the slope and intercept parameters and the corresponding standard errors. Is the slope is significantly different from zero? Formally state your hypotheses, carry out the test, and carefully state your conclusions.

The least-squares estimates of the slope and intercept parameters and the corresponding standard errors can both be found in the 3rd and 4th columns of the Parameter Estimates table in the PROC REG output, which is shown below. The estimate and standard error of the slope parameter are found in the row corresponding to the variable `nclust`.

The formal hypotheses for testing whether the slope is significantly different from zero are $H_0 : \beta_1 = 0$ and $H_a : \beta_1 \neq 0$. This test can be done with the global F test provided in the ANOVA table included in the PROC REG output, or with a t-test that is in the Parameter Estimates table (both tests are equivalent because this is a simple linear regression). The test statistic and p-value for the t-test are in the Parameter Estimates table in the row corresponding to the variable `nclust`. Because the p-value is so small, there is significant evidence that the null hypothesis should be rejected and we should instead conclude that slope is significantly different from zero.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	1	-1.02790	0.78355	-1.31	0.2189
nclust	1	0.05138	0.00725	7.09	<.0001

3. Suppose that the number of clusters this year is `nclust` = 102. Estimate the mean grape crop yield for this value of `nclust` and give a 95% confidence interval.

Before we can estimate the mean at `nclust` = 102, we have to notice that we have an observation with this value of `nclust`. We have to add such an observation to our dataset to

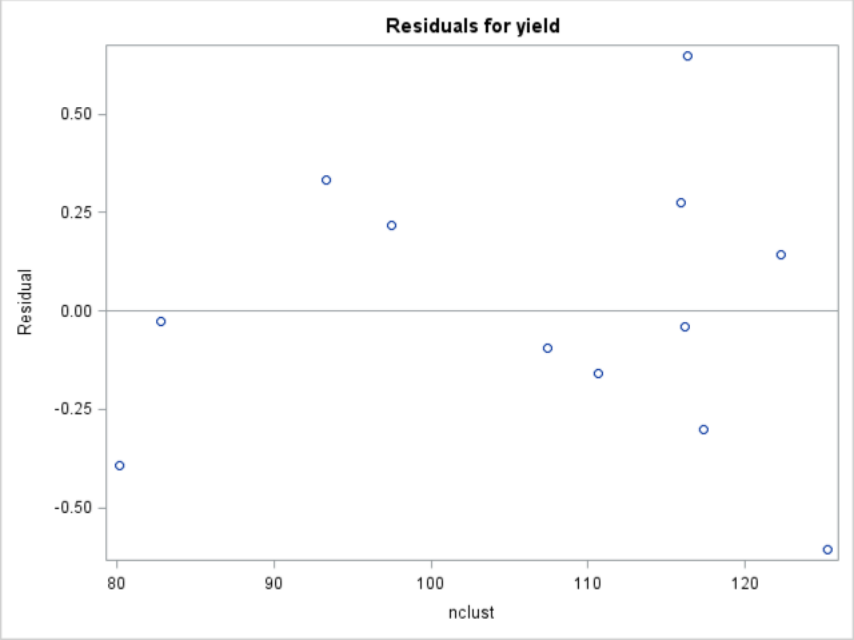
have SAS do the estimation for us at this point. This is done by creating a new dataset with only one observation and then concatenating the new dataset to our original dataset. In the new dataset, we set `nclust` = 102 and we let `yield` be missing (In SAS, missing numeric values are designated with a .). By having `yield` be missing, this observation won't be included among the observations used to fit the regression model when we run PROC REG again, as only observations with non-missing values for all variables in the regression model are used, and so our regression model has not been changed by the addition of this new "observation" at the point we would like to estimate `yield`. When PROC REG is used on a dataset in which there are missing values, or if we explicitly ask for it, an Output Statisitics table is generated. In this table, for every observation, the observed value of the response variable is shown along with the predicted value of the response variable and the confidence interval for the estimated mean response variable. The estimation of the mean grape crop yield when `nclust` = 102, as well as the 95% confidence interval, are shown in that table, in the row corresponding to the missing value for the dependent variable.

Obs	nclust	Dependent Variable	Predicted Value	Std Error Mean Predict	95% CL Mean	
13	102.0	.	4.2129	0.1114	3.9647	4.4612

*Note: This table can be easier to read if you add `id nclust;` to the PROC REG step. This statement identifies each observation in the table by the value of `nclust`, in addition to the value of `yield`. This is especially helpful if you have multiple observations with the same value for the dependent variable.

4. Draw a plot of the residuals versus the predictor variable `nclust`. What does this plot say concerning the assumptions of the simple linear regression model?

A residual plot is provided, by default, in the PROC REG output. The residual plot can be seen below. Based on the assumptions of the simple linear regression model, we would expect the residual plot to be a random cloud of points centered at zero with constant variance across the different values of the predictor variables. In this residual plot, that appears to be the case and there does not appear to be any noticeable pattern in the points.


Residuals for yield
Residual
0.50
0.25
0.00
-0.25
-0.50
80
90
100
110
120
nclust