

Lab 2 – Spring 2021

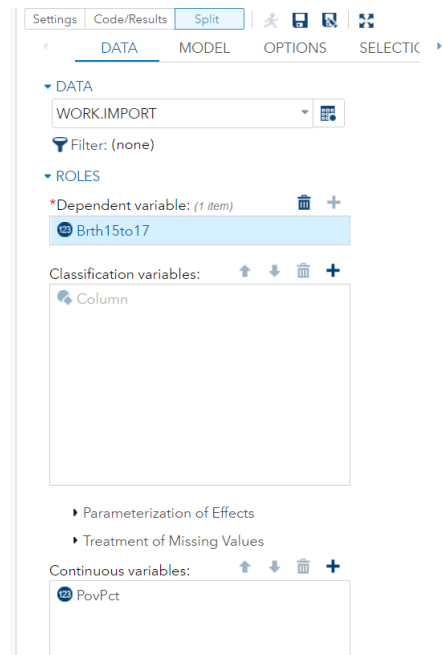
11th March 2021

cavicchia@ese.eur.nl

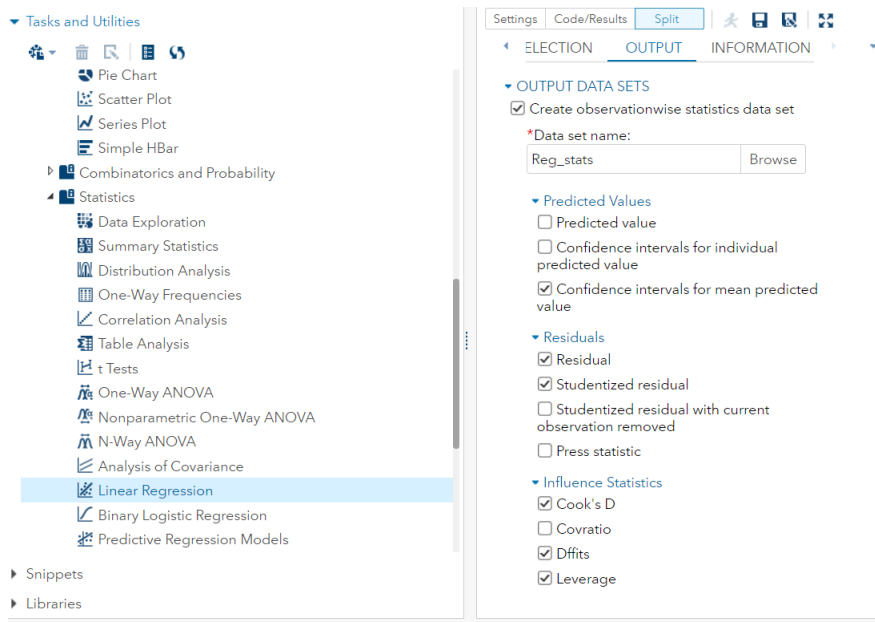
1 Linear Regression Model

We want to study the relationship between $y =$ year 2002 birth rate per 1000 females 15 to 17 years old and $x =$ poverty rate, which is the percent of the state's population living in households with incomes below the federally defined poverty level. Thus we focus on the scatterplot between PovPct and Brth15to17.

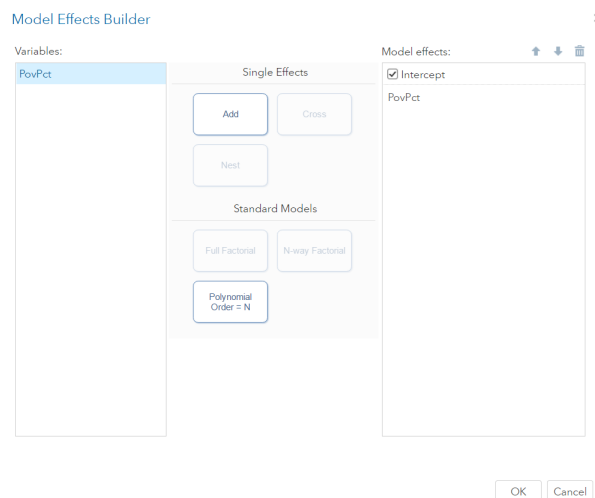
The plot shows a generally linear relationship, on average, with a positive slope. As the poverty level increases, the birth rate for 15 to 17 year old females tends to increase as well. We are now ready to estimate the linear regression model. To get the output select **Linear Regression** under the **Statistics** window. Select Birth15to17 as Dependent variable and PovPct as Continuous variable.



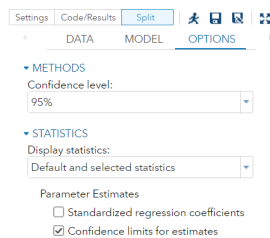
Open the Output window. Here there are some options, according to the aim of the analysis it is possible to compute the predicted values, to conduct a residual analysis or to investigate the presence of outliers and/or influential observations.



Then the model has to be built as follows, by opening the **Model** window. Add the variable **PovPct** in the Model effects column.



Finally, under the Option window, select the confidence intervals box for the estimates.



After selecting the run button, it is possible to get the estimates for the regression model.

1.1 Output

Model: MODEL1
Dependent Variable: Brth15to17

Number of Observations Read	51
Number of Observations Used	51

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1725.25949	1725.25949	56.00	<.0001
Error	49	1509.63463	30.80887		
Corrected Total	50	3234.89412			

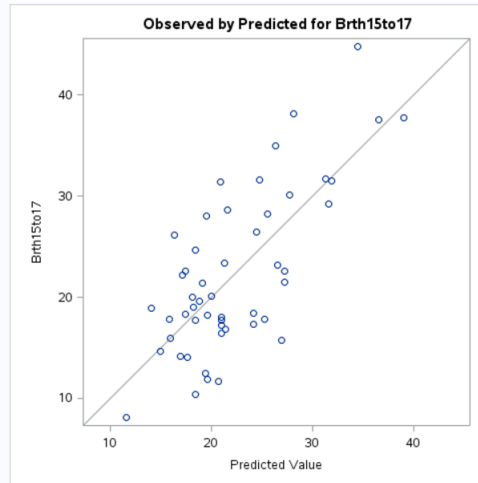
Root MSE	5.55057	R-Square	0.5333
Dependent Mean	22.28235	Adj R-Sq	0.5238
Coeff Var	24.91018		

Parameter Estimates							
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	95% Confidence Limits	
Intercept	1	4.26729	2.52975	1.69	0.0980	-0.81642	9.35101
PovPct	1	1.37335	0.18352	7.48	<.0001	1.00454	1.74215

The estimated regression line is the following one $\hat{y} = 4.267 + 1.373PovPct$, where $y = Brth15to17$

- The **interpretation of the slope** (value = 1.373) is that the 15 to 17 year old birth rate increases 1.373 units, on average, for each one unit (one percent) increase in the poverty rate.
- The **interpretation of the intercept** (value=4.267) is that if there were states with poverty rate = 0, the predicted average for the 15 to 17 year old birth rate would be 4.267 for those states. Since there are no states with poverty rate = 0 this interpretation of the intercept is not practically meaningful for this example.

From the output, we also see the information that $s = 5.55057$ (Root MSE) and $R^2 = 53.3\%$.



The value of s tells us roughly the average difference between the y -values of individual observations and predictions of y based on the regression line. The value of R^2 can be interpreted to mean that poverty rates explain 53.3% of the observed variation in the 15 to 17 year old average birth rates of the states. The R^2 (adj) value (52.4%) is an adjustment to R^2 based on the number of x -variables in the model (only one here) and the sample size. With only one x -variable, the adjusted R^2 is not important.

1.2 Simple Linear Regression Model Evaluation

Recall that we are ultimately always interested in drawing conclusions about the population, not the particular sample we observed. In the simple regression setting, we are often interested in learning about the population intercept β_0 and the population slope β_1 . Confidence intervals and hypothesis tests are two related, but different, ways of learning about the values of population parameters.

Is there a relationship between Brth15to17 and PovPct? Certainly, since the estimated slope of the line, b_1 , is 1.373, not 0, there is a relationship between Brth15to17 and PovPct in the sample of 51 data points. But, we want to know if there is a relationship between the population of all of Brth15to17 and PovPct. That is, we want to know if the population slope β_1 is unlikely to be 0.

Looking at the p -value it is possible to conclude that β_1 is significantly different from 0 (p -value $< \alpha$, where α is usually 0.05 - thus the null hypothesis $H_0 : \beta_1 = 0$ is rejected).

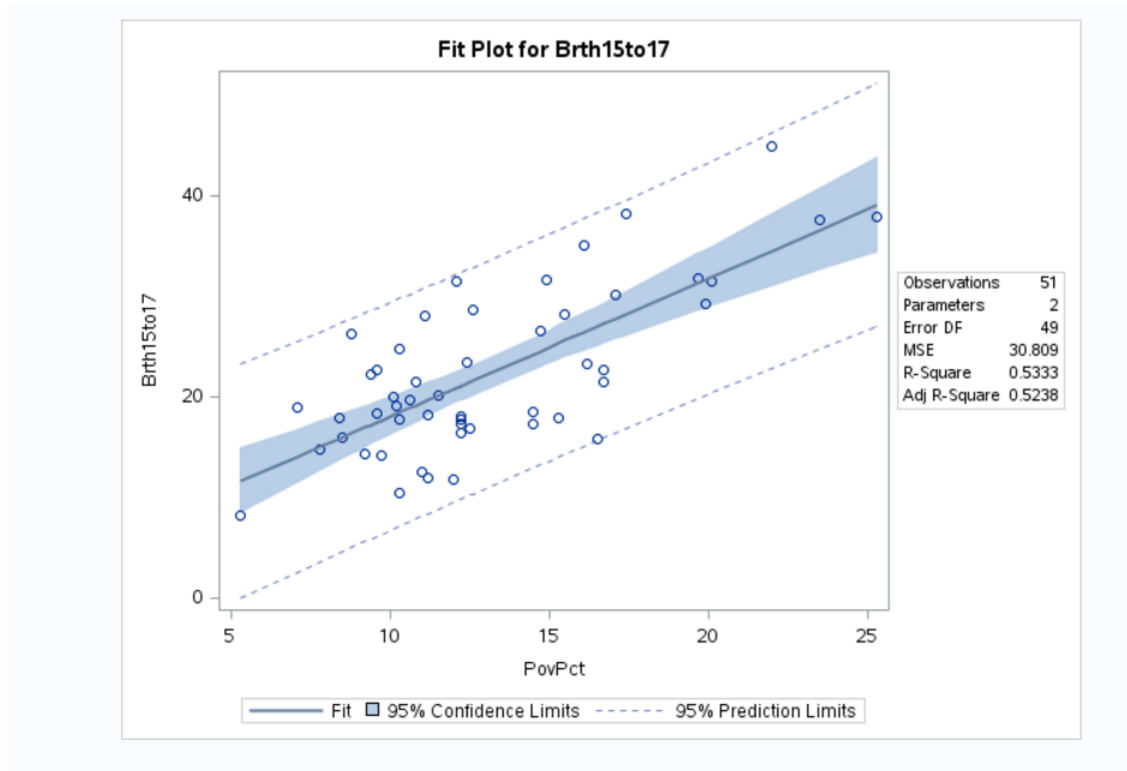
This is equivalent to look at the confidence intervals; if 0 is included in the interval, then the corresponding β is not significant. It is possible to note that 0 is not included in the CI for β_1 .

Furthermore it is very important to know the interpretation of the ANOVA table, and the meaning of all its elements (as shown in class).






The F -test is more useful for the multiple regression model when we want to test that more than one slope parameter is 0. Here F -test is equivalent to t -test (F -value= t -value² and the corresponding p -values are the same).

1.3 Simple Linear Regression Model Prediction

Typically, a regression analysis involves the following steps: model formulation, model estimation, model and model use. Here we focus on the following questions: what is the average response for a given value of the predictor \mathbf{x} ? What is the value of the response likely to be for a given value of the predictor \mathbf{x} ?



Observe that the prediction interval is always wider than the confidence interval. Furthermore, both intervals are narrowest at the mean of the predictor values (about 15). (In SAS it is possible to get the specific CI or PI for a given x value, by selecting **Confidence intervals for individual predicted value** under the window Predicted Values). The output can be seen by selecting Output data window.

Settings Code/Results Split     

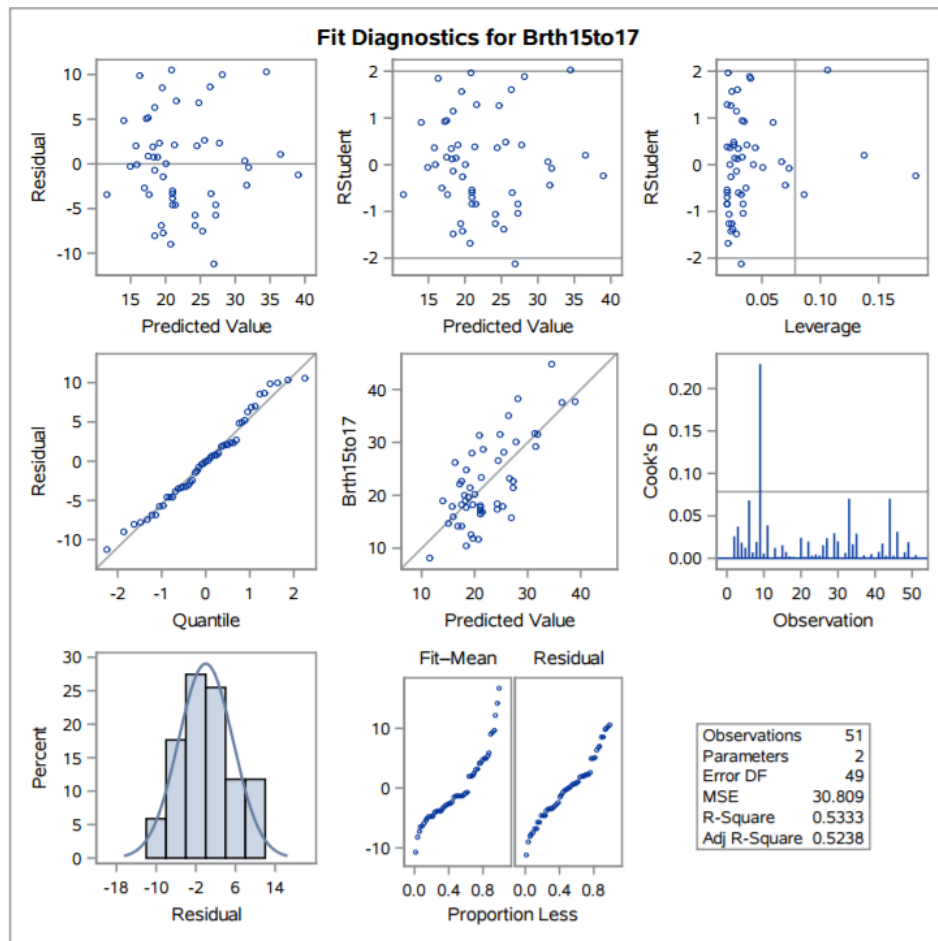
CODE LOG RESULTS OUTPUT DATA

Table: WORK.REG_STATS View: Column names Filter: (none)

Columns Total rows: 51 Total columns: 15 Rows 1-51

Columns		lclm_	uclm_	lcl_	ucl_
<input checked="" type="checkbox"/>	Select all	16.536737887	20.288762805	7.1017938769	29.723706815
<input checked="" type="checkbox"/>	Location	30.851744735	38.110038053	22.751056827	46.210725961
<input checked="" type="checkbox"/>	PovPct	24.583689119	28.44728716	15.195145346	37.835830932
<input checked="" type="checkbox"/>	Brth15to17	19.278399264	22.491144827	9.6153999685	32.154144123
<input checked="" type="checkbox"/>	ViolCrime	16.536737887	20.288762805	7.1017938769	29.723706815
<input checked="" type="checkbox"/>	TeenBrth	22.537789304	25.823812653	12.906147952	35.455454005
<input checked="" type="checkbox"/>	Brth18to19	19.712595977	22.880955348	10.030545646	32.563005678
<input checked="" type="checkbox"/>	lclm_	15.420986718	19.48183043	6.1138194821	28.788997666
<input checked="" type="checkbox"/>	uclm_	19.423948128	22.620265041	9.7539024707	32.290310698
<input checked="" type="checkbox"/>	lcl_	17.318922516	20.879923565	7.8039152776	30.394930803
<input checked="" type="checkbox"/>	ucl_	22.78809585	26.122844263	13.177241018	35.733699095
<input type="checkbox"/>	Lower Bound of 95% C.I. (Individual Pred)	28.435541031	34.208852968	19.800430808	42.843963192
<input type="checkbox"/>	Label	17.934189729	21.363332662	8.3634579738	30.934064418
<input type="checkbox"/>	Name	16.220227986	20.05593455	6.8201099522	29.456052584
<input type="checkbox"/>	Length	17.627802313	21.120381923	8.0839264159	30.66425782
<input type="checkbox"/>	Type	19.423948128	22.620265041	9.7539024707	32.290310698
<input type="checkbox"/>	Format	14.774362738	19.029778099	5.5466552577	28.257485579
<input type="checkbox"/>		22.405541260	26.676264605	24.644712606	40.427105210

1.4 Simple Linear Regression Model Assumptions



Main Remarks

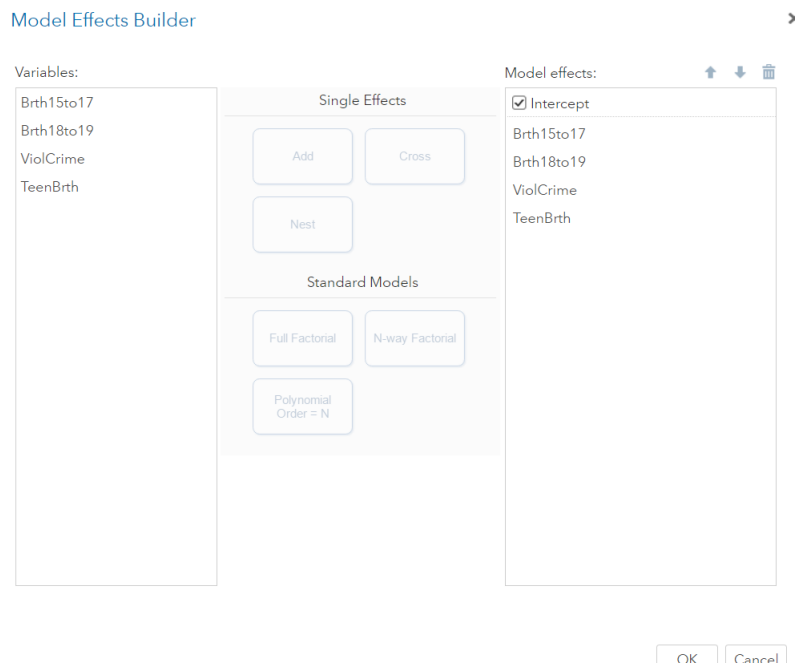
- **Residuals vs Predicted Value.** The residuals *bounce randomly* around the 0 line. This suggests that the assumption that the relationship is linear is reasonable. The residuals roughly form a *horizontal band* around the 0 line. This suggests that the variances of the error terms are equal. No one residual *stands out* from the basic random pattern of residuals. This suggests that there are no outliers.
- **Residuals vs Quantile.** Note that the relationship between the theoretical percentiles and the sample percentiles is approximately linear. Therefore, the normal probability plot of the residuals suggests that the error terms are indeed normally distributed. This is also confirmed by the histogram of the residuals.
- **Leverage and Cook's Distance.** They are useful to assess the presence of outliers and/or influential observations. Here only one observation needs further care to investigate its nature, that is observation number 9. (influential observation means that it could influence significantly the estimates of the regression line - graphically

the regression line is pulled up towards this observation. If the observation is not significantly, the estimates of the regression line with or without the observation under investigation are approximately the same. In other words, the regression line does not pull up towards this observation value).

2 Multiple Linear Regression Model

In the second part of this class we extend the linear regression model to the multiple linear regression model. We use the same data set, but in this case the response variable is PovPct. We want to study how this variable is related to all the others.

To estimate the linear regression model, select **Linear Regression** under the **Statistics** window. Select PovPct as Dependent variable and all other variables as continuous variables. Then open the Model window, as shown in the picture.



Open the Output window. Here there are some options, according to the aim of the analysis it is possible to compute the predicted values, to conduct a residual analysis or to investigate the presence of outliers and/or influential observations.

Under the Option window, select the following options: c.i. for the estimates, standardized regression coefficients and VIF values.

DATA MODEL **OPTIONS**

95%

▼ STATISTICS

Display statistics:

Default and selected statistics

Parameter Estimates

- Standardized regression coefficients
- Confidence limits for estimates

Sums of Squares

- Sequential sum of squares (Type I)
- Partial sum of squares (Type II)

Partial and Semipartial Correlations

- Squared partial correlations
- Squared semipartial correlations

Diagnostics

- Analysis of influence
- Analysis of residuals
- Predicted values

► Multiple Comparisons

▼ Collinearity

- Collinearity analysis
- Tolerance values for estimates
- Variance inflation factors

After selecting the run button, it is possible to get the estimates for the regression model.

2.1 Output - Full Model

2.2 Output

Model: MODEL1									
Dependent Variable: PovPct									
Number of Observations Read		51							
Number of Observations Used		51							
Analysis of Variance									
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F				
Model	4	560.90320	140.22580	18.23	<.0001				
Error	46	353.83092	7.69198						
Corrected Total	50	914.73412							
Root MSE		2.77344	R-Square	0.6132					
Dependent Mean		13.11765	Adj R-Sq	0.5796					
Coeff Var		21.14283							
Parameter Estimates									
Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	1	6.22349	1.82549	3.41	0.0014	0	0	2.54898	9.89801
Brth15to17	1	-0.45769	0.44681	-1.02	0.3110	-0.86071	83.95773	-1.35707	0.44168
Brth18to19	1	-0.82144	0.27311	-3.01	0.0043	-3.64426	174.57598	-1.37118	-0.27171
ViolCrime	1	-0.07786	0.06683	-1.17	0.2500	-0.16228	2.30683	-0.21238	0.05665
TeenBrth	1	1.81957	0.66635	2.73	0.0089	5.24039	437.98062	0.47827	3.16086

The estimated regression line is the following one $PovPct = 6.22 - 0.46Brth15to17 - 0.82Brth18to19 - 0.08VC + 1.82TB$.

Some remarks

- Interpretation.** Each coefficient represents the change in the mean response, per unit increase in the associated predictor variable when all the other predictors are held constant.
 For example, -0.82 represents the change in PovPct, per unit increase in Brth18to19 when all other variables are held constant.
 The intercept term, 6.22, represents the mean response, when all the predictors are all zero (which may or may not have any practical meaning).
- Significance.** The only significant variables are Brth18to19 and TeenBrth (look at the p-value and CIs) .
- R^2 vs. $adj R^2$.** R^2 is always greater than $adj R^2$. Whenever a further variable R^2 increases since the fitting improves - SSM increases, a higher proportion of variability of y is explained. However, this is not a good measure to compare two different multiple linear regression model. Indeed the model complexity increases too; so a better measure is represented by $adj R^2$, where the model complexity (i.e. the number of predictors) is taken into account.

- **t-test vs. F-test.** In a multiple linear context, t-test and F-test are not equivalent anymore. The t-test assess the significance of the single predictor, given ALL others in the model. On the other hand, F-test assess the significance of the set of predictors. In other words, it says if there is at least one predictor in the set that is significant. In this example, the p-value of the F-test is 0.0001 - it follows that $H_0 : \beta_1 = \dots = \beta_P = 0$ is rejected. This means that there exists at least one predictor that is significant, but we do not know which one. This information is provided by the single t-tests.
- **Standardized Coefficients.** Standardization of the coefficient is usually done to answer the question of which of the predictor variables have a greater effect on the dependent variable in a multiple regression analysis, when the variables are measured in different units of measurement. Here the most important variables are also the only significant ones, i.e. TeenBrth (5.24) and Brth18to19 (-3.64).
- **Multicollinearity.** VIF quantifies the severity of multicollinearity; If $VIF > 5$ then multicollinearity is high. Here VIF values are extremely high (84, 175 and 438); they correspond to very strong correlations between these variables (look at the correlation matrix). Behind the problems arisen from the multicollinearity in the statistical inference (high s.e., low t-values, and sometimes very high R^2 that lead to unreliable estimates for the coefficients), it is important to understand the intuition. If two variables are highly correlated, it may mean that they describe the same information; in other words they explain the same amount of variability of y in the same way. This means that it is not needed to have both predictors in the model. We only need one of them. So a possible solution is to remove one or more predictors that are highly correlated.

2.3 Multiple Linear Regression - Model Selection

In the multiple linear regression model, to select the best model, different procedures are available. The best model means to conduct a variable selection. From a practitioner's point of view, it is always advisable to combine these automatic procedures with some a priori knowledge. Indeed there could exist some variables that should be kept in the model since they are important for the analysis of the phenomenon under investigation.

To conduct the variable selection, open the Selection window and select the following options.

The selected models are the following ones.

Settings Code/Results Split

OPTIONS SELECTION OUTPUT

MODEL SELECTION

Selection method: Forward selection

Add/remove effects with: Significance level

Stop adding/removing effects with: Default criterion

Select best model by: Default criterion

*Significance level to add an effect to the model: 0.2

SELECTION STATISTICS

SELECTION PLOTS

DETAILS

OPTIONS SELECTION OUTPUT

MODEL SELECTION

Selection method: Backward elimination

Add/remove effects with: Significance level

Stop adding/removing effects with: Significance level

Select best model by: Default criterion

*Significance level to remove an effect from the model: 0.2

SELECTION STATISTICS

SELECTION PLOTS

DETAILS

Selection process details: Selection summary

OPTIONS SELECTION OUTPUT

MODEL SELECTION

Selection method: Stepwise selection

Add/remove effects with: Significance level

Stop adding/removing effects with: Significance level

Select best model by: Default criterion

*Significance level to add an effect to the model: 0.2

*Significance level to remove an effect from the model: 0.2

SELECTION STATISTICS

SELECTION PLOTS

DETAILS

Selection process details: Selection summary

- Forward selection

The selected model is the model at the last step (Step 1).

Effects: Intercept Brth15to17

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	487.85328	487.85328	56.00	<.0001
Error	49	426.88083	8.71185		
Corrected Total	50	914.73412			

Root MSE	2.95158
Dependent Mean	13.11765
R-Square	0.5333
Adj R-Sq	0.5238
AIC	165.35864
AICC	165.86928
SBC	116.22229

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	4.464469	1.227985	3.64	0.0007
Brth15to17	1	0.388342	0.051895	7.48	<.0001

Model: MODEL1
Dependent Variable: PovPct

Parameter Estimates										
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	4.46447	1.22799	3.64	0.0007	0	0	1.99674	6.93220
Brth15to17	Brth15to17	1	0.38834	0.05189	7.48	<.0001	0.73029	1.00000	0.28406	0.49263

- Backward selection

Selected Model

The selected model is the model at the last step (Step 2).

Effects: Intercept Brth18to19 TeenBrth

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	540.20699	270.10350	34.62	<.0001
Error	48	374.52713	7.80265		
Corrected Total	50	914.73412			

Root MSE	2.79332
Dependent Mean	13.11765
R-Square	0.5906
Adj R-Sq	0.5735
AIC	160.68577
AICC	161.55533
SBC	113.48125

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	6.416344	1.774646	3.62	0.0007
Brth18to19	1	-0.471507	0.140583	-3.35	0.0016
TeenBrth	1	0.962501	0.216556	4.44	<.0001

Model: MODEL1
Dependent Variable: PovPct

Parameter Estimates										
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	6.41634	1.77465	3.62	0.0007	0	0	2.84818	9.98451
Brth18to19	Brth18to19	1	-0.47151	0.14058	-3.35	0.0016	-2.09180	45.60216	-0.75417	-0.18885
TeenBrth	TeenBrth	1	0.96250	0.21656	4.44	<.0001	2.77202	45.60216	0.52709	1.39792

- Stepwise selection

Selected Model

The selected model is the model at the last step (Step 1).

Effects: Intercept Brth15to17

Note: The p-values for parameters and effects are not adjusted for the fact that the terms in the model have been selected and so are generally liberal.

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	487.85328	487.85328	56.00	<.0001
Error	49	426.88083	8.71185		
Corrected Total	50	914.73412			

Root MSE	2.95158
Dependent Mean	13.11765
R-Square	0.5333
Adj R-Sq	0.5238
AIC	165.35864
AICC	165.86928
SBC	116.22229

Parameter Estimates					
Parameter	DF	Estimate	Standard Error	t Value	Pr > t
Intercept	1	4.464469	1.227985	3.64	0.0007
Brth15to17	1	0.388342	0.051895	7.48	<.0001

Model: MODEL1
Dependent Variable: PovPct

Parameter Estimates										
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Standardized Estimate	Variance Inflation	95% Confidence Limits	
Intercept	Intercept	1	4.46447	1.22799	3.64	0.0007	0	0	1.99674	6.93220
Brth15to17	Brth15to17	1	0.38834	0.05189	7.48	<.0001	0.73029	1.00000	0.28406	0.49263

It is possible to note that both Forward and Stepwise select the same model as the best one. The Backward select a model with two predictors, but VIF values shows that there is still multicollinearity. Thus, the best model is the one with only Brth15to17 as predictor variable.

A further consequence of multicollinearity is to have different final models selected by different selection procedures. Anyhow, different selection procedures could lead to different models, although multicollinearity is not present.

2.4 Multiple Linear Regression - Further Remarks

Here, we did not conduct the residual analysis again. It will be the same as that conducted in Section 1.4. In the multiple linear regression model it is possible to plot the Residuals vs.

Single Predictors. No other differences exist. Also in this case, it is possible to estimate the confidence interval and prediction interval the response mean, as described in section 1.3. Finally it is possible to assess the omission of some predictors. Plot the residuals versus the variable that is not included in the model. If there is a linear pattern, it means that the variability of the residuals (the variability of y that is unexplained by the current set of predictors) can be explained by this predictor. Such pattern in the plot suggests to include the predictor in the model.

3 Multiple Linear Regression Model - Assignment

ASSIGNMENT

Use the data set Grape Juice and answer to the following questions.

Data description A company is selling a new type of grape juice in some of its stores for pilot selling. Its marketing team wants to analyse:

- Which type of in-store advertisement is more effective?
- The Price Elasticity
- The Cross-price Elasticity
- How to find the best unit price to maximize the profit and the forecast of sales with that price.

There are 5 variables:

- **Sales:** Total unit sales of the grape juice in one week in a store;
- **Price:** average unit price of the grape juice in one week;
- **Ad type:** The in-store advertisement type to promote the grape juice, ad type=0 (natural production); ad type=1 (family health caring)
- **Price apples:** average unit price of the apple juice in the same store in one week
- **Price cookies:** average unit price of the cookies in the same store in one week

Work on yourself on the following tasks:

1. Data Exploration
2. Fit Multiple Linear Regression. Provide a brief interpretation of coefficients; evaluate the statistical significance of the model (t-tests and F-test and say in what they differ); assess the model assumptions (residual analysis).
3. With the fitted model, we can analysis the Price Elasticity(PE) and Cross-price Elasticity(CPE) to predict the reactions of sales quantity to price. Price elasticity is defined as $\% \Delta Q / \% \Delta P$, which indicates the percent change in quantity divided by the percent change in price; Cross-price Elasticity is the percent change in quantity divided by the change in the price of some other product - $PE = (\Delta Q / Q) / (\Delta P / P) = (\Delta Q / \Delta P) * (P / Q)$. Calculate also the CPE on apple juice and cookies to analyze the how the change of apple juice price and cookies price influence the sales of grape juice.
4. Optimal Pricing and Sales Prediction. Usually companies want to get higher profit rather than just higher sales quantity. So, how to set the optimal price for the new grape juice to get the maximum profit based on the dataset collected in the pilot period and the regression model above? To simplify the question, we can let the ad type =

1, the price apple = 7.659 (mean value), and the price cookies = 9.738 (mean value). Assume the marginal cost(C) per unit of grape juice is 5. We can calculate the profit (Y) by the following formula - $Y = (\text{price} - C) * \text{Sales Quantity} = (\text{price} - 5) * (804.55 - 51.24 * \text{price})$. Find the optimal price.