

## CHAPTER 8

# LINEAR REGRESSION

---

### Learning Objectives:

1. *To explain estimation of regression coefficients with ordinary least squares*
  2. *To interpret coefficients of linear regression models*
  3. *To use linear regression for prediction*
- 

### Content

Lecture Notes	p. 2
- Introduction	p. 2
- Simple Linear Regression	p. 2
- Multiple Linear Regression (MLR)	p. 6
Tutorial 8	p. 15
Answers	p. 17

## 1. Introduction

We are often interested in studying the relationship among variables to determine whether they are associated with one another. The ability to understand causes and predict outcomes of events is critical in nearly every facet of our lives.

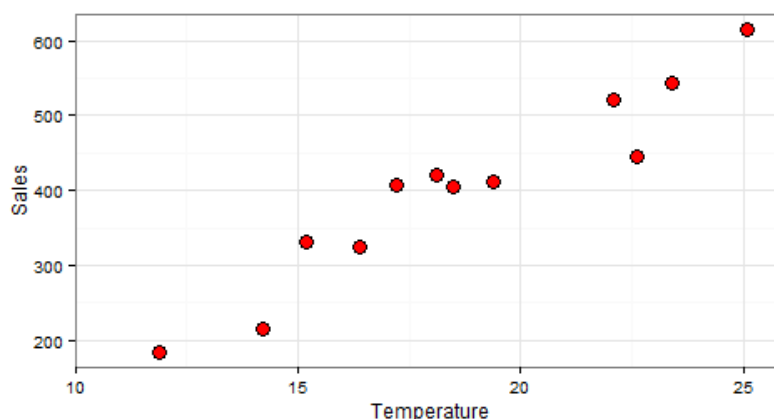
We intuitively have the ability to detect that variables are related. Linear regression analysis allows us to investigate the nature of such relationships. It is a common supervised data mining technique that data scientists use for predictive modelling.

## 2. Simple linear regression

**Example 1:** The local ice cream shop keeps track of how much ice cream they sell versus the temperature on that day. The figure for the last 12 days are given in the table.



Temp °C	Sales
14.2	\$215
16.4	\$325
11.9	\$185
15.2	\$332
18.5	\$406
22.1	\$522
19.4	\$412
25.1	\$614
23.4	\$544
18.1	\$421
22.6	\$445
17.2	\$408



Next to the table is the same set of data represented in a  $x$ - $y$  plane called a **scatterplot**. We can easily see from the scatterplot that there is an association between the sales of ice-cream and temperature on the day.

\_\_\_\_\_ is the **response variable**.

Temperature is the \_\_\_\_\_ variable.

How can the ice-cream shop owner use the relationship between temperature and sales?

We use **simple linear regression** to model the association between **one** \_\_\_\_\_ response variable and **one** predictor variable.

## 2.1 Simple linear regression model

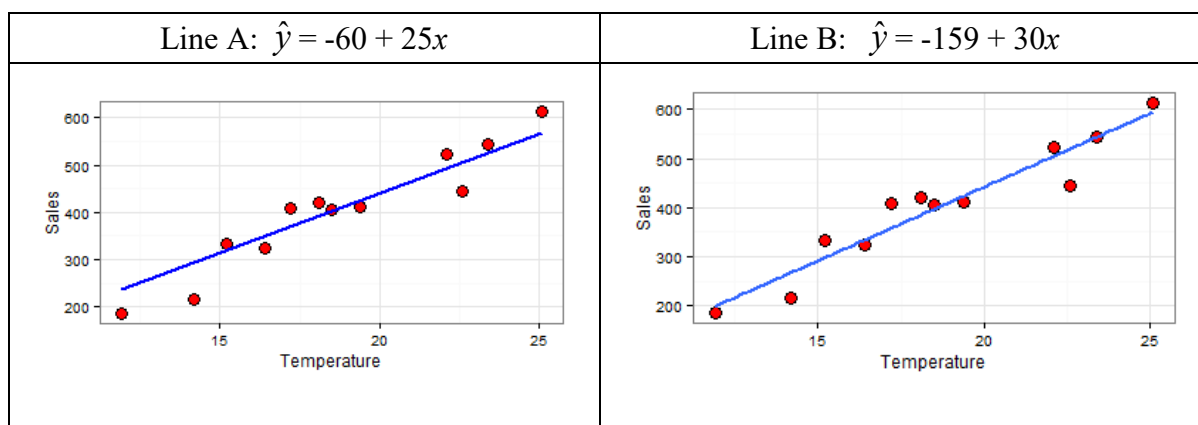
If the scatterplot shows that the variables are correlated, then it is reasonable to assume that the **conditional mean** of  $y$  given a value of  $x$  can be estimated using a straight line equation, i.e.

$$\hat{y} = E(y | x) = A + Bx$$

## 2.2 Least squares criterion

The values of  $A$  and  $B$  are unknown and have to be estimated from the sample. It turns out that the “best” estimate of  $A$  and  $B$  is found using the **least squares criterion**. This means that the total squared differences between the actual observed  $y$ -values and  $\hat{y}$  should be minimised.

Consider the ice cream sales data. Consider two lines, Line A and Line B. These two lines are just two of the many straight lines that could be fitted on the ice-cream sales and temperature data points, as shown in the following:



To decide whether Line A or Line B fits the data set better, we consider the **residuals** made in using the line to estimate the mean  $y$ -values of the data points.

\_\_\_\_\_ are defined to be  $e = y - \hat{y}$ . That is the difference between the \_\_\_\_\_ values and the \_\_\_\_\_ mean value for a given  $x$ .

Then we square each  $e$  and sum up over all  $x$  to get  $\sum e^2$ . This value represents how well our regression line fits into the sample data.

Line A: $\hat{y} = -60 + 25x$				Line B: $\hat{y} = -159 + 30x$			
Temp °C (x)	Sales (y)	$\hat{y} = -60 + 25x$	$e^2 = (y - \hat{y})^2$	Temp °C (x)	Sales (y)	$\hat{y} = -159 + 30x$	$e^2 = (y - \hat{y})^2$
14.2	\$215	\$295	6400	14.2	\$215	\$267	2704
16.4	\$325	\$350	625	16.4	\$325	\$333	64
11.9	\$185	\$238	2756.25	11.9	\$185	\$198	169
15.2	\$332	\$320	144	15.2	\$332	\$297	1225
18.5	\$406	\$403	12.25	18.5	\$406	\$396	100
22.1	\$522	\$493	870.25	22.1	\$522	\$504	324
19.4	\$412	\$425	169	19.4	\$412	\$423	121
25.1	\$614	\$568	2162.25	25.1	\$614	\$594	400
23.4	\$544	\$525	361	23.4	\$544	\$543	1
18.1	\$421	\$393	812.25	18.1	\$421	\$384	1369
22.6	\$445	\$505	3600	22.6	\$445	\$519	5476
17.2	\$408	\$370	1444	17.2	\$408	\$357	2601
			Total $e^2$ 19356.25				Total $e^2$ 14554

Since Line B has smaller sum of squared residuals than Line A, hence Line B is the one that fits the data set better.

In summary, the least squares criterion states that the line having the smallest value of total  $e^2$  is the “best” straight line. This “best” line is called the **regression line** or the **least-squares line**. The equation of the least squares line is called the **regression equation or model**.

Since the regression equation is the line that best fits the sample data, we can now use the regression equation to make estimations of  $\hat{y}$  for a given  $x$  value.

Simple linear regression is performed when: -

- the output (response) is expensive to measure while the predictor is not.
- There might be some \_\_\_\_\_ link between predictor and response. The regression coefficients \_\_\_\_\_ the effect of the \_\_\_\_\_ on the response.

**Example 2:** (a) Find the regression equation of the ice-cream sales example and interpret its regression coefficients.

(b) How much ice cream sales can you expect on a 30°C day?

## 2.3 Coefficient of determination, $R^2$

The **coefficient of determination** measures the \_\_\_\_\_ of your model.

The coefficient of determination has the following formula

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y})^2}{\sum_i (y_i - \bar{y})^2}$$

This is a measure of how good the model is able to estimate the observed data points.

Some properties of  $R^2$ :

- i)  $0 \leq R^2 \leq 1$ .
- ii) A value of  $R^2$  close to 1 suggests that the predictor variable explains the variation in response. Thus, the model is useful in making predictions.
- iii) A value of  $R^2$  close to 0 suggests that the predictor variable does not adequately explain the variation in response. Thus, the model is not useful in making predictions.
- iv)  $R^2$  values are often expressed as percentages.

**Example 3:** Determine the  $R^2$  of the ice-cream sales model. Interpret it.

### 3. Multiple linear regression (MLR)

We use \_\_\_\_\_ linear regression analysis to investigate how a continuous response variable is related to \_\_\_\_\_ predictor variables,  $X_1, X_2, \dots, X_p$ .

Mathematically speaking, a multiple linear regression model with  $p$  predictors is given by

$$\hat{y} = b_0 + b_1X_1 + \dots + b_pX_p$$

where  $b_0, b_1, \dots, b_p$  are called r\_\_\_\_\_ c\_\_\_\_\_.

#### Case study

It is now widely believed that smoking tends to impair lung function. Much of the data to support this claim arises from studies of pulmonary function in adults who are long-time smokers. A question then arises whether such deleterious effects of smoking can be detected in children who smoke. To address this question, measures of lung function were made in 654 children seen for a routine check up in a particular pediatric clinic. The children participating in this study were asked whether they were current smokers.

A common measurement of lung function is the forced expiratory volume (FEV), which measures how much air you can blow out of your lungs in a short period of time. A higher FEV is usually associated with better respiratory function. It is well known that prolonged smoking diminishes FEV in adults, and those adults with diminished FEV also tend to have decreased pulmonary function as measured by other clinical variables, such as blood oxygen and carbon dioxide levels.

Below are the explanations for the variable names

age	subject age at time of measurement (years)
fev	measured FEV (liters per second)
height	subject height at time of measurement (inches)
sex	subject sex (1 = male, 0 = female)
smoke	smoking habits (1 = yes, 0 = no)

**Example 4:** Consider the FEV dataset in the case study which is a regression task to predict Forced Exhalation Values (FEV) from a features of observations.

(a) Identify the predictor variables and the response variable.

(b) For each variable, state whether it is a qualitative or quantitative variable.

We use MLR when

- there are \_\_\_\_\_ predictor variable related to \_\_\_\_\_ response variable.
- We need to predict the behaviour of output when controlling for input values.

Do the predictor variables need to be continuous or categorical variables?

### 3.1 Interpreting regression coefficients

#### *Intercept term*

The **intercept term**,  $b_0$  is known as the baseline value of the regression. This estimates the response value if all predictors equals zero.

$b_0$  may or may not have a meaningful interpretation. For example, in a situation where we are predicting sales, a negative value of the intercept doesn't make much sense. So it has no interpretation.

#### *Predictor coefficient terms*

The other **regression coefficients** are interpreted then as the average \_\_\_\_\_ in the response if  $X_i$  \_\_\_\_\_ by one unit holding other predictors constant.

#### **Controlling for the effect of multiple predictors**

When a dataset has more than one predictor for a response, we cannot fit simple linear regression models separately for each predictor.

For example, for the FEV dataset,

$$\text{FEV} = 0.4316 + 0.222\text{age and}$$

$$\text{FEV} = -5.433 + 0.132\text{ht}$$

However, the MLR model for FEV on age and height together is

#### **Regression Equation**

$$\text{FEV} = -4.610 + 0.10971 \text{ ht} + 0.05428 \text{ age}$$

Notice that the coefficients are different. And this is because by fitting individual simple linear regression models, we ignore the effect of the other predictor. This gives the illusion that a single predictor is solely responsible for the effect on the response when it actually had “help” from the other predictor.

Multiple linear regression allows us to properly assign “responsibility for effects” to each individual predictor.

### 3.2 Overfitting and adjusted $R^2$

When there are a lot of predictors in a MLR model, some predictors may appear significant by chance. In fact, population wise, these predictors have no effect on the response. This leads to an inflated  $R^2$ . A high  $R^2$  does not necessarily mean that the response variables is correlated with its attribute variables.

Why is this undesirable?

This phenomena is called **overfitting**.

The adjusted  $R^2$  is a way to compensate for overfitting by \_\_\_\_\_ any extra variables in the model. Hence, it is important to compare with  $R_{adj}^2$  is obtain a more realistic estimate of model fit.

The formula for adjusted  $R^2$  is

$$R_{adj}^2 = 1 - \frac{(1 - R^2)(n - 1)}{n - p - 1}$$

where  $n$  is the sample size and  $p$  is the number of predictor variables in your MLR model.

If there is a large difference between  $R^2$  and  $R_{adj}^2$ , then your MLR model suffers from overfitting.

**Example 5:** Given the resulting regression equation of the MLR model for the FEV data from KNIME:

File				
Statistics on Linear Regression				
Variable	Coeff.	Std. Err.	t-value	P> t
age	0.0702	0.0103	6.8104	2.71E-11
ht	0.1	0.0051	19.5445	0.0
sex=1	0.1793	0.0353	5.0743	5.43E-7
smoke=1	-0.0801	0.064	-1.2527	0.2109
Intercept	-4.2579	0.2385	-17.8504	0.0
Multiple R-Squared: 0.7869				
Adjusted R-Squared: 0.7852				



(a) Interpret the coefficients of the model.

Variable	Value of regression coefficient	Interpretation
Age		
Ht		
Sex=1		
Smoke=1		

(b) A boy aged 8 years old does not smoke and is 52 inches tall. What is his estimated FEV?

(c) Assess whether this model suffers from overfitting. Why or why not?

### 3.3 Assessing the effect of each predictor

A MLR model is estimated from a sample of data drawn from a population. The underlying relationship between predictors and response *in the population* is given by

$$y_i = \beta_0 + \beta_1 X_1 + \dots + \beta_p X_p + \epsilon_i$$

where  $\beta_i$  are considered the true parameter values of this relationship *in the population*. The  $\epsilon_i$  are known as *errors*.

The regression coefficients in the MLR

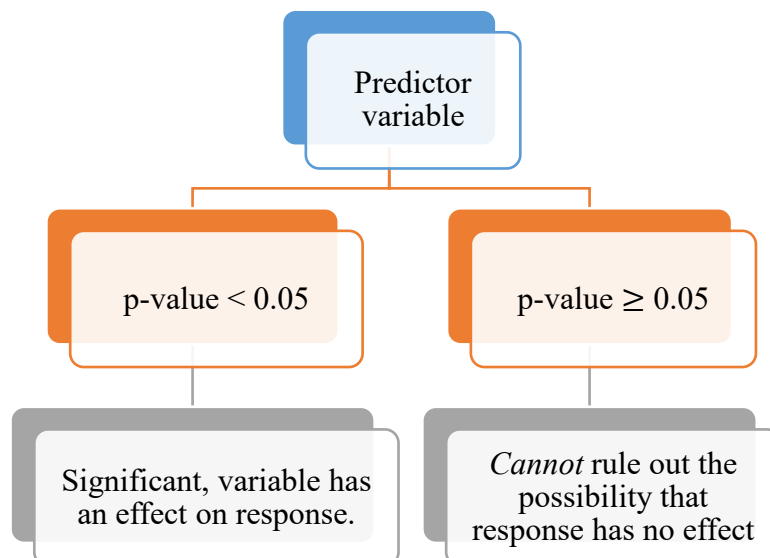
$$y_i = b_0 + b_1 X_1 + \dots + b_p X_p + e_i$$

produced by sample data via software is an \_\_\_\_\_ of the true regression model.

If predictor variable  $X_i$  has no effect on the response  $Y$ , then  $\beta_i = 0$ . Such variables are known as \_\_\_\_\_ variables. But estimated from our sample,  $b_i$  isn't necessarily equals zero! This means that we may end up thinking  $X_i$  has an effect on  $Y$  when actually it has no effect.

We use **p-values** to detect insignificant variables. If  $p$ -value is \_\_\_\_\_ 0.05, then we cannot rule out the possibility that this variable is insignificant.

Variables whose  $p$ -value is \_\_\_\_\_ than 0.05 are considered significant. Significant variables are predictors which we have very strong reason to believe that  $\beta_i \neq 0$ .



**Example 6:** Recall the MLR model built earlier in KNIME with the following output:

Statistics on Linear Regression				
Variable	Coeff.	Std. Err.	t-value	P> t
age	0.0702	0.0103	6.8104	2.71E-11
ht	0.1	0.0051	19.5445	0.0
sex=1	0.1793	0.0353	5.0743	5.43E-7
smoke=1	-0.0801	0.064	-1.2527	0.2109
Intercept	-4.2579	0.2385	-17.8504	0.0

Multiple R-Squared: 0.7869  
Adjusted R-Squared: 0.7852

(a) Which of the variables in the FEV case are significant?

(b) Would you be able to conclude that smoking impairs lung function? Why or why not?

### p-values

So what are  $p$ -values?

$p$ -values are the probability that a sample of data points  $(\mathbf{x}, \mathbf{y})$  is drawn from the population in such a way that  $b_i \neq 0$  even when  $\beta_i = 0$ . Thus if  $p$ -value is small ( $< 0.05$ ), it is unlikely that  $b_i \neq 0$  even when  $\beta_i = 0$ . Or in other words, it is unlikely that our sample tells us that a predictor has an effect when actually there is no effect (in the population).

This is why we are justified in concluding that a predictor is significant when its  $p$ -value is  $< 0.05$ .

### 3.4 Variable selection

To prevent overfitting, we would like a MLR model with the minimal amount of predictors to explain variation in the response variable. Deciding which variables to keep (and which to discard) in the model is called variable selection.

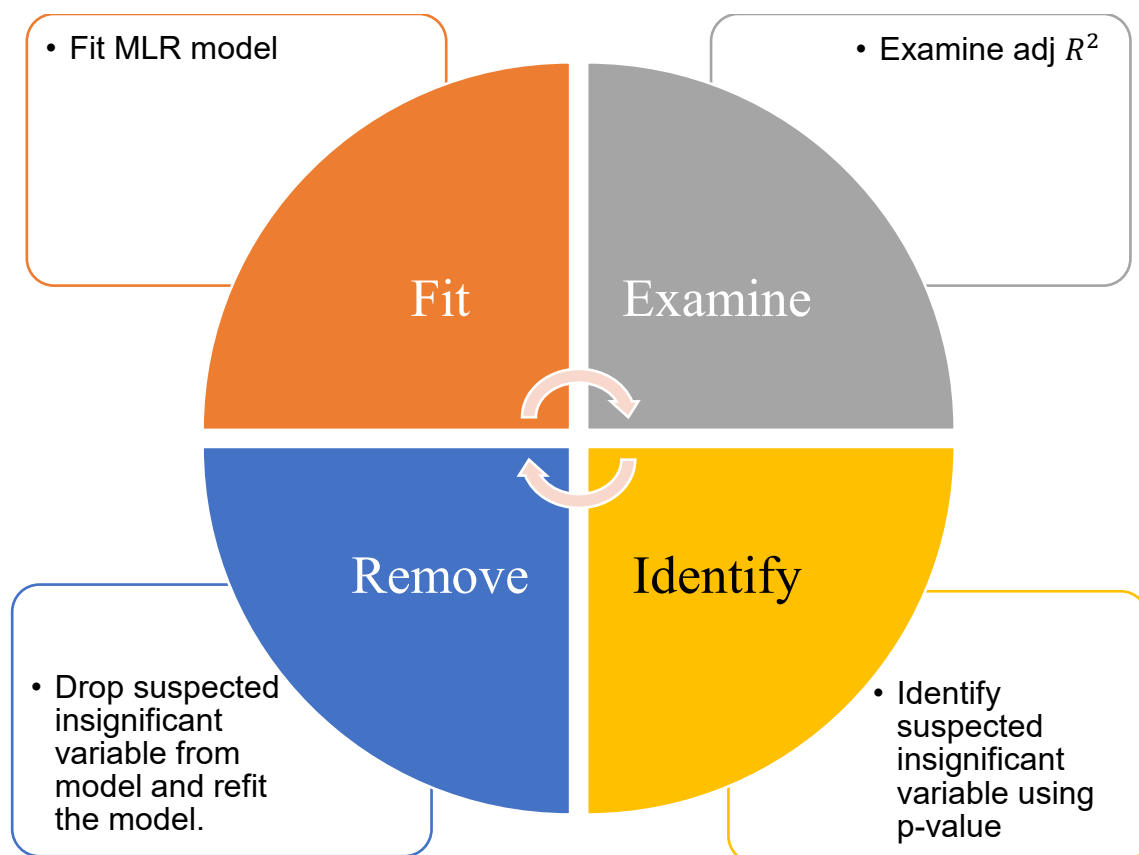


Figure 1: Backward elimination stepwise process

Figure 1 shows the main steps to drop variables from the model. Variables should be dropped one at a time. We continue to do this until there are no insignificant variables to drop.

**Example 7:** Perform backward elimination on the FEV model you build earlier. State the variable(s) that you will consider dropping from the model.

Linear Regression Result Vie...				
File				
Statistics on Linear Regression				
Variable	Coeff.	Std. Err.	t-value	P> t
age	0.0702	0.0103	6.8104	2.71E-11
ht	0.1	0.0051	19.5445	0.0
sex=1	0.1793	0.0353	5.0743	5.43E-7
smoke=1	-0.0801	0.064	-1.2527	0.2109
Intercept	-4.2579	0.2385	-17.8504	0.0
Multiple R-Squared: 0.7869				
Adjusted R-Squared: 0.7852				

### 3.5 Nonlinear effects and interactions

Sometimes the original set of predictor variables / features are insufficient to properly represent the relationships with the response variable.

This can happen when

1. We suspect a **nonlinear relationship** between predictor and response.

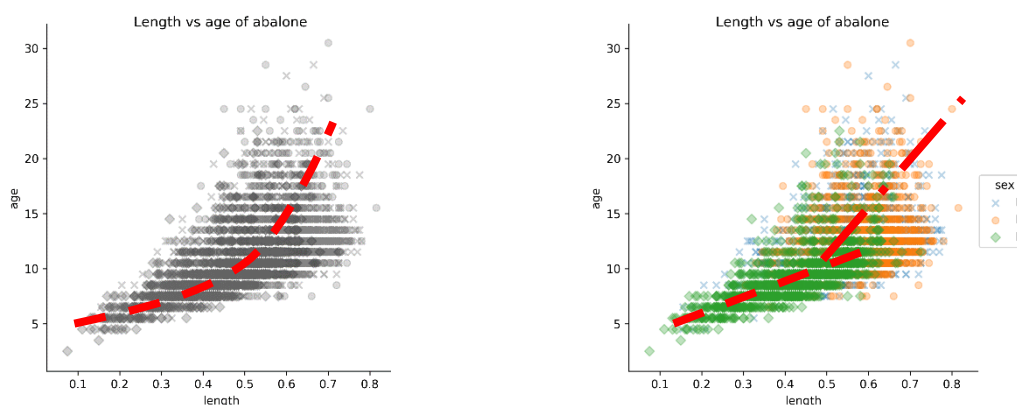


Figure 2: (Left) Nonlinear relationship between *length* and *age* of abalone. (Right) Interaction of *gender* affecting the relationship between *length* and *age* of abalone.

For example, the age of abalone is non-linearly related to the length of abalone (suggested by a dashed curve in Figure 2 (Left)).

2. There is evidence that there is an **interaction** between two predictors.

#### What is an interaction?

Interaction is needed to account for the way one predictor variable affects the relationship between another predictor and the dependant variable. For example (refer to Figure 2 (Right)), suppose we are trying to predict the age of abalone from the two variables:

Gender	The gender of abalone (M-male, F-female, I-infant)
Length	The length of abalone in cm

It may be the case that infant abalone grows at a different rate (suggested by the two different slopes “- -” and “- . -” lines in Figure 2 (Right)) with respect to length compared to male or female abalone. Thus, gender affects the relationship between age and length. These two variables (gender and *length*) are said to have an **interaction**.

If we do not add more terms to our model, our model may not give accurate predictions. This is simply because our model does not fully describe all relationships. Hence we need to fit a better model that incorporates more features.

We can expand the original set of variables by creating *calculated fields*.

1. If non-linearity is present with a particular variable,  $x$ , we can add higher powers of that feature to the dataset:  $x^2, x^3, \dots$
2. If an interaction is present between a variable,  $x_1$  and another variable  $x_2$ . Then we can add the field:  $x_1 x_2$  to the dataset. (i.e. multiply the two columns together).

**Example 8:** The following features in the FEV dataset are suspected to have an interaction: *Age* (continuous) and *Smoke* (categorical).

- (a) Interpret, in context, what the interaction term means in terms of how *Age* and *Smoke* influences the outcome, *FEV*.

- (b) The model is refitted in KNIME with non-linear and interaction terms with the following result:

Statistics on Linear Regression				
Variable	Coeff.	Std. Err.	t-value	P> t
age	0.079	0.0108	7.3031	1.07E-12
ht	-0.2877	0.0535	-5.3803	1.13E-7
sex=1	0.1247	0.0348	3.5796	0.0004
smoke=1	0.2437	0.349	0.6983	0.4853
ht_sq	0.0032	0.0004	7.2563	1.47E-12
age_smoke	-0.0284	0.0258	-1.1008	0.2715
Intercept	7.3251	1.6065	4.5597	6.41E-6
Multiple R-Squared: 0.8069				
Adjusted R-Squared: 0.8046				

Is there a significant non-linear relationship in this model? Is the interaction term statistically significant? Interpret this result in context.

## Tutorial 8

1. A scientist planted alfalfa on several plots of land, identical except for the soil pH. The data collected and shown below give the yields (in kilograms per acre) for each

pH	Yield
4.6	479
4.8	831
5.2	739
5.4	840
5.6	809
5.8	1201
6.0	967

Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P> t
<b>PH</b>	334.7826	118.7606	2.819	0.0372
<b>Intercept</b>	-950.6957	637.0201	-1.4924	0.1958

Multiple R-Squared: 0.6138

Adjusted R-Squared: 0.5366

- Which variable is the response variable and which is the predictor variable?
- Write down the regression equation for the model above in (a). Interpret this model.
- For every increase of 0.1 in soil pH, how much would you expect the yield to change?
- Estimate the yield from a plot of land with soil pH 5.0.
- Explain why the yield estimation from soil pH = 9.0 with this model may not be useful for prediction.

2. [☐] The skincancer.txt file contains data on skin cancer mortality (Mort) in states in the USA. The longitude and latitude of the states are given and the Ocean variable indicates whether or not the state has a coastline or not.

We want to determine if latitude has any effect on skin cancer mortality. This is because we hypothesize that the sunnier parts of the US leads to higher deaths due to skin cancer.

- What is the predictor and response variable?
- Fit a linear regression model and interpret the regression coefficient of the predictor variable on the response. Give an interpretation in terms of the context of the problem.
- Interpret the value of  $R^2$  for this model.
- Scientists postulate that being next to the ocean has an effect on the response. Build a multiple linear regression model to test this hypothesis. Based on the model you have built, do you agree with the scientists postulate?
- Compare the model you built in (b) and (d). Which is a preferred model and why?

3. [☐] Various companies have budgeted portions of their advertising budget to different media: TV, radio and newspaper. Their total sales for next quarter was recorded. This data can be found in the advertising.xlsx file.

A marketing consultant wants to find out which media (or combination of media) is a more effective marketing channel to raise sales.

- (a) Using KNIME's scatter matrix node, identify the variable which has
- (i) nonlinear effect upon sales
  - (ii) no effect upon sales.
- (b) Having done some visual analysis, the consultant builds a regression model to validate his findings. Build a model with the following attributes: TV, radio, newspaper and  $TV^2$ . Then perform backwards elimination to remove insignificant variables and interpret the final model.
- (c) Extend your model in (b) by adding an **interaction** term:  $TV \times radio$  and assess whether this term has a significant impact on sales. Interpret this interaction and hence help the consultant craft a marketing recommendation.

## ANSWERS

- 1.
- (a) Response – Yield, predictor – pH
  - (b)  $Yield = -950.7 + 334.8pH$
  - (c) 33.47
  - (d) 723.217
  - (e) No, because the  $pH = 9.0$  is too alkaline and it is unlikely that plants can survive in such conditions too. Data collected is from moderate acidic to neutral range, so model may not be applicable to soil in alkali range

2. (*suggested*)

- (a) predictor-Lat, response-Mort
- (d) Yes
- (e) Ocean coefficient is not zero, Ocean explains (statistically) significant amount of variance.

3.

(*Suggested*)

- (a) (i) TV (ii) newspaper



# LAB 8A : Linear Regression using KNIME (FEV dataset)

## Learning Objectives:

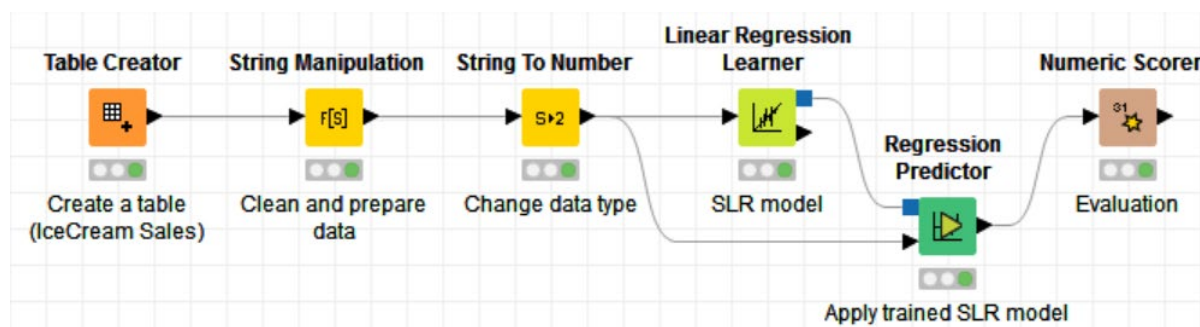
1. Build a simple linear regression model using KNIME.
2. Interpret regression result in KNIME.

### Task: The Ice Cream Dataset

In this lab, we want to build a simple linear regression model to predict ice cream sales given the day's temperature.

S Tempera...	S Sales
14.2	\$215
16.4	\$325
11.9	\$185
15.2	\$332
18.5	\$406
22.1	\$522
19.4	\$412
25.1	\$614
23.4	\$544
18.1	\$421
22.6	\$445
17.2	\$408

Your completed workflow might appear as follows:



## A. Create a KNIME Workflow

1. Download the **ice\_cream\_sales.knwf** KNIME workflow from the BlackBoard. You should see a workflow prepared with a single node.



When you open the node, you should see this:

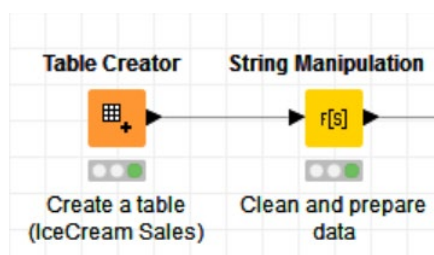
The screenshot shows the 'Table Creator' dialog box with the 'Table Creator Settings' tab selected. The 'Input line' is set to '14.2'. The table below shows the data for 13 rows (Row0 to Row12).

	S Tempera...	S Sales
Row0	14.2	\$215
Row1	16.4	\$325
Row2	11.9	\$185
Row3	15.2	\$332
Row4	18.5	\$406
Row5	22.1	\$522
Row6	19.4	\$412
Row7	25.1	\$614
Row8	23.4	\$544
Row9	18.1	\$421
Row10	22.6	\$445
Row11	17.2	\$408
Row12		

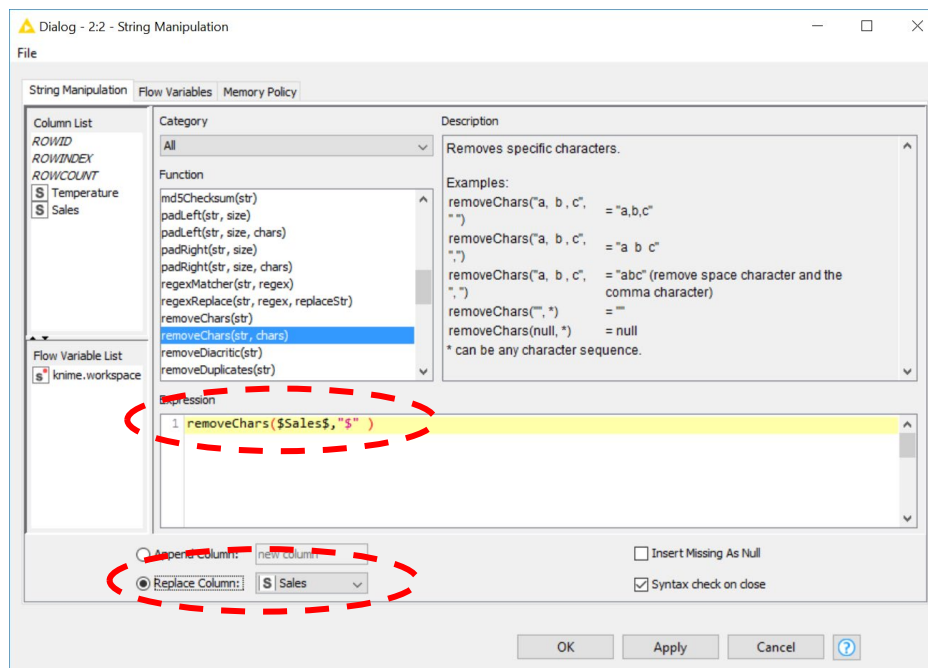
## B. Cleaning and Preparing the Dataset

This data is not clean. We will need to clean it first.

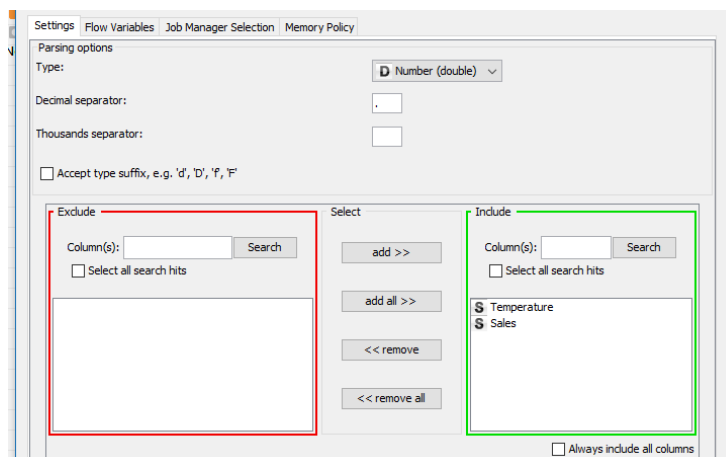
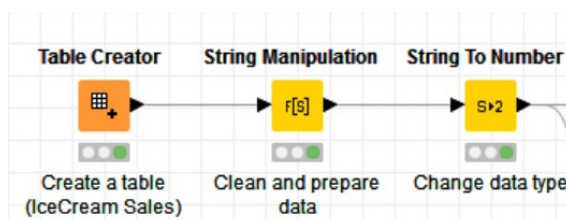
- (a) Search for the **String Manipulation** node and join it up to the Table creator node.



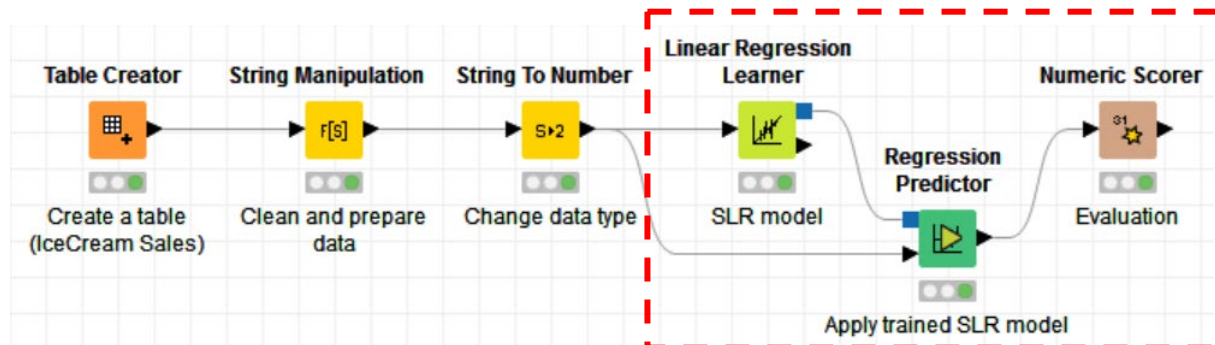
- (b) To remove the “\$” sign from *Sales*, select **removeCharacter(str, chars)** from the *Function* menu. Then double click on the *Sales* field and then type “\$” in the text area as shown below. Enable the checkbox “Replace Column:” and that *Sales* is selected as the column to replace. The “\$” will be removed and replace back to the *Sales* column. Click OK and run this node.



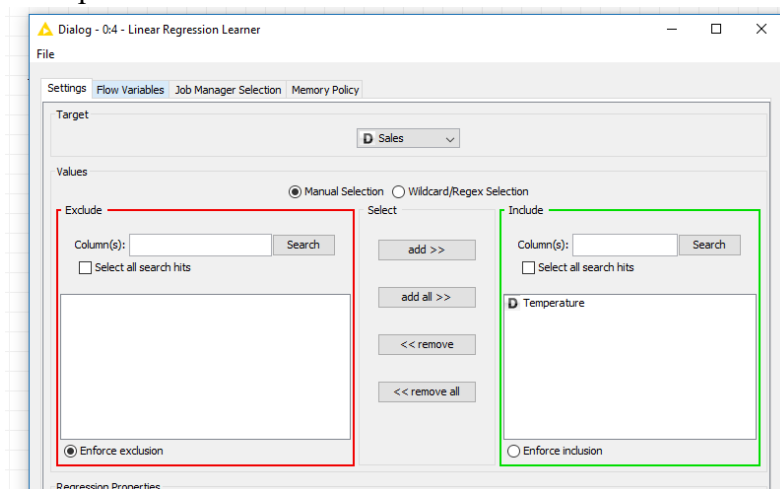
- (c) Next we need to convert the String into Numerical values. Select **String to Number** node and open it. Ensure that both *Temperature* and *Sales* are in the green box



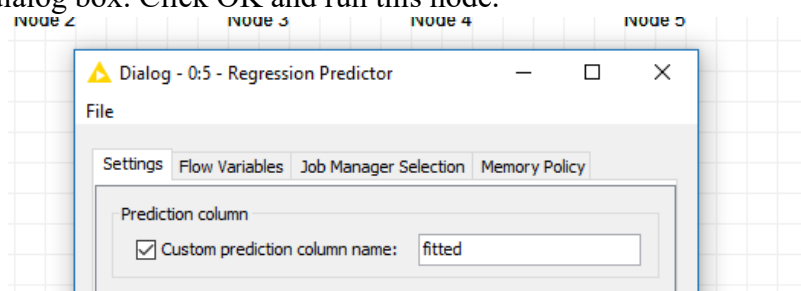
(d) Click OK. Now set up with linear regression model workflow as seen below.



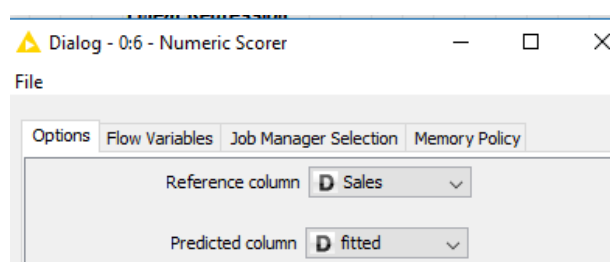
- i. For the **Linear Regression Learner** node, ensure that *Sales* is the *Target* and *Temperature* is placed in *Include*. Click OK and run this node.



- ii. Now that we have our model. Let us obtain the fitted predictions of this model. Connect the **Regression Predictor** node as shown above (ensure the blue data port is connected). Check the *Custom prediction column name:* and type “fitted” in the dialog box. Click OK and run this node.

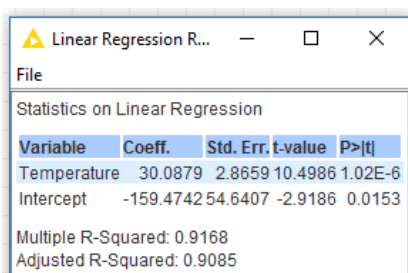


- iii. Next search for the **Numeric Scorer** node. Configure the node as shown above. Click OK and run the node.



**Investigative tasks**

- (a) In a Linear Regression Learner node configuration, should we place the response or predictor variable at the *Target*? Can this target be categorical?
- (b) What is the  $R^2$  that you obtain? Interpret this value in the context.
- (c) Interpret the results of the linear regression as shown below.



Variable	Coeff.	Std. Err.	t-value	P> t
Temperature	30.0879	2.8659	10.4986	1.02E-6
Intercept	-159.4742	54.6407	-2.9186	0.0153

Multiple R-Squared: 0.9168  
Adjusted R-Squared: 0.9085

# LAB 8B : Linear Regression using KNIME (FEV dataset)

## Learning Objectives:

1. To build a multiple linear regression (MLR) model using KNIME.
2. To build MLR models via the backward elimination stepwise process.
3. To include nonlinear and interaction terms into model.

### **Task: The FEV Dataset**

It is now widely believed that smoking tends to impair lung function. Much of the data to support this claim arises from studies of pulmonary function in adults who are long-time smokers. A question then arises whether such deleterious effects of smoking can be detected in children who smoke. To address this question, measures of lung function were made in 654 children seen for a routine check up in a particular pediatric clinic. The children participating in this study were asked whether they were current smokers.

A common measurement of lung function is the forced expiratory volume (FEV), which measures how much air you can blow out of your lungs in a short period of time. A higher FEV is usually associated with better respiratory function. It is well known that prolonged smoking diminishes FEV in adults, and those adults with diminished FEV also tend to have decreased pulmonary function as measured by other clinical variables, such as blood oxygen and carbon dioxide levels.

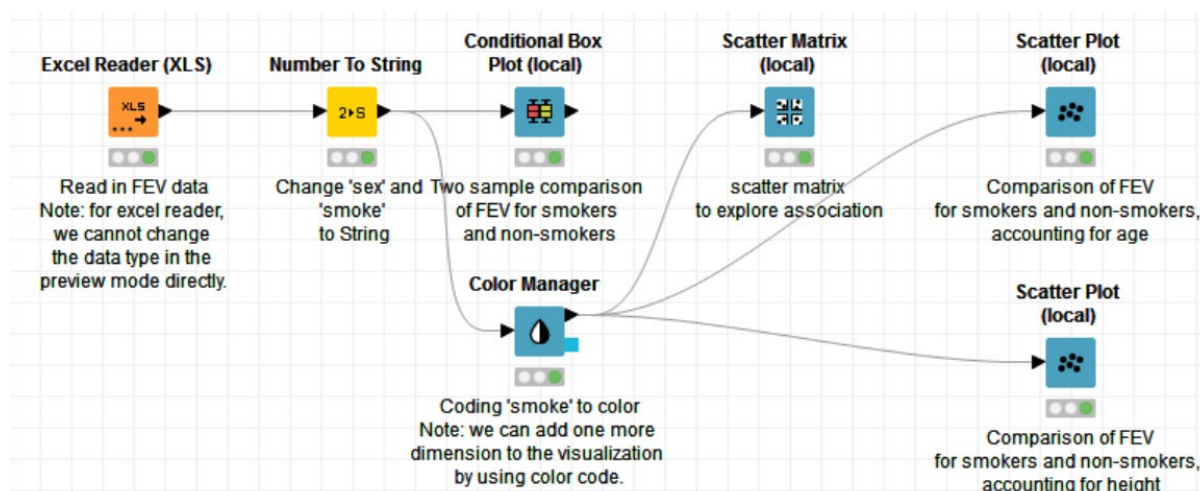
Below are the description for the variables:

Variable	Description
<b>age</b>	subject age at time of measurement (years)
<b>fev</b>	measured FEV (liters per second)
<b>height</b>	subject height at time of measurement (inches)
<b>sex</b>	subject sex (1 = male, 0 = female)
<b>smoke</b>	smoking habits (1 = yes, 0 = no)

Problem: We want to investigate if FEV is associated with the various variables: age, height, sex and smoke. Next, we want to figure out the quantitative effect of each variable on FEV, and try to get a true estimation of the effect smoking has on FEV with a MLR model.

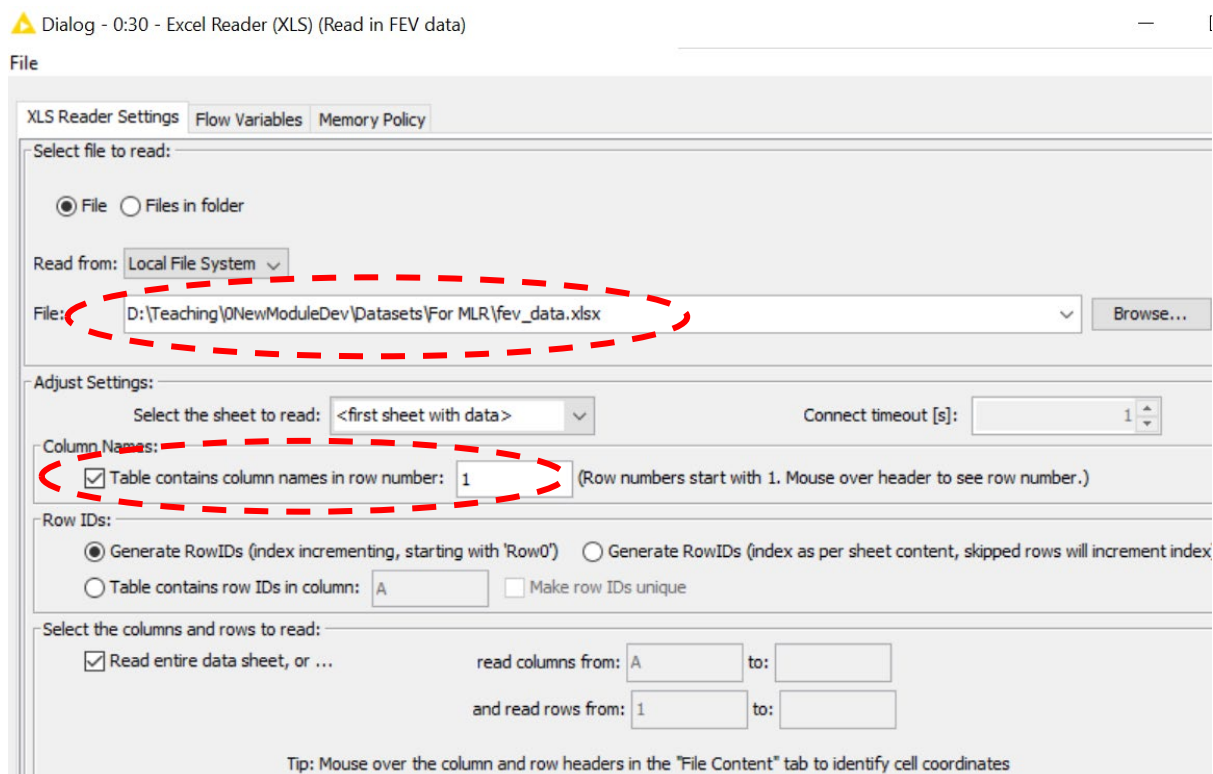
## A. Visualization

Let us first make some preliminary investigation of the dataset using KNIME's visualization tools. The completed workflow for the investigation will look like this:

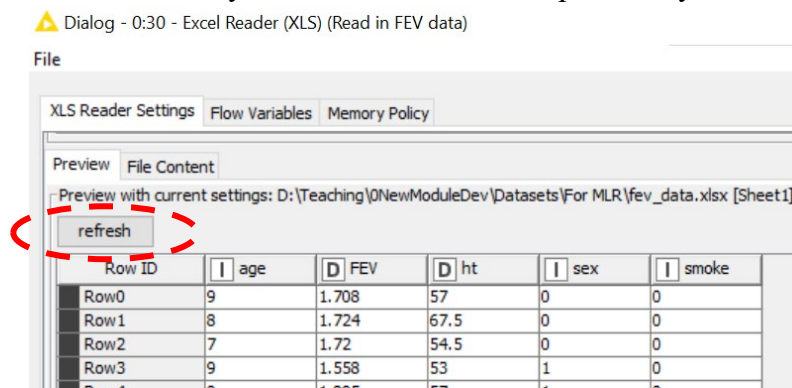


## B. Data preparation

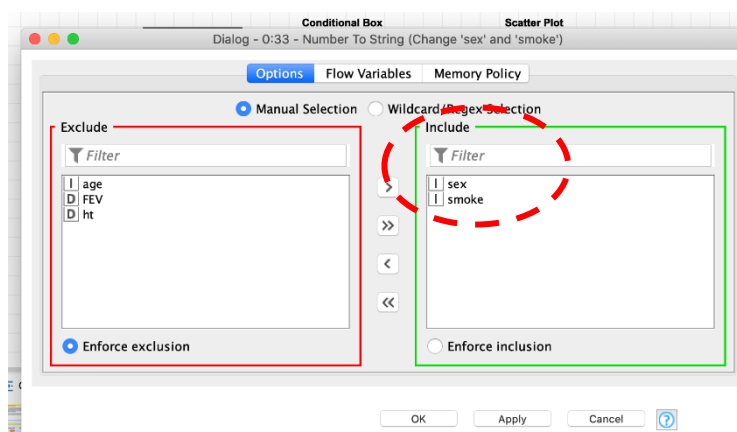
1. Read the FEV data using the **Excel Reader (XLS)** node, since the FEV data is '.xlsx' file. Check that the 'Table contains column names in row number' is checked and '1' is entered since Row 1 in the file contains the column header. If this is not checked, the data will not be loaded correctly and the column header will be read as Row 1.



- Click on the Preview > 'refresh' button and check that the data to be imported is correct.  
Note: For Excel Reader node, we cannot change the data type in the preview mode directly (we can do this directly in the File Reader node previously for '.csv' files).

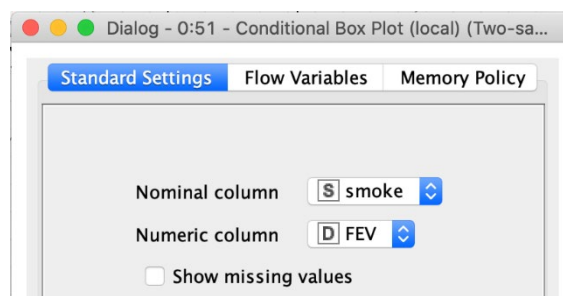


- Next, use the **Number to String** node to change *sex* and *smoke* variables into String variables.



### C. Investigating the relationship between FEV for smokers and non-smokers

We will use **Conditional Box Plot (local)** node for this. We want to see whether there is any visual evidence that FEV is affected by smoking.

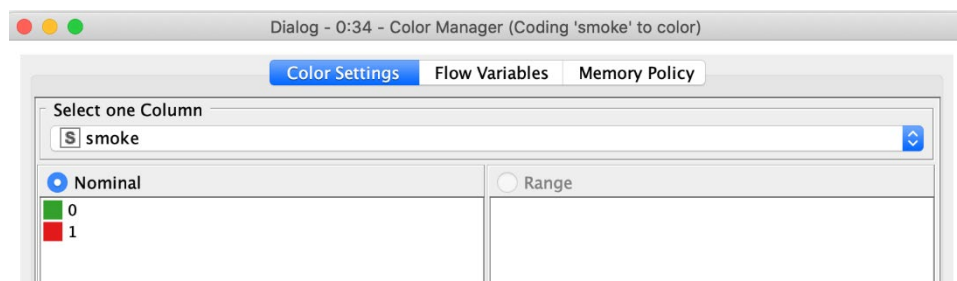




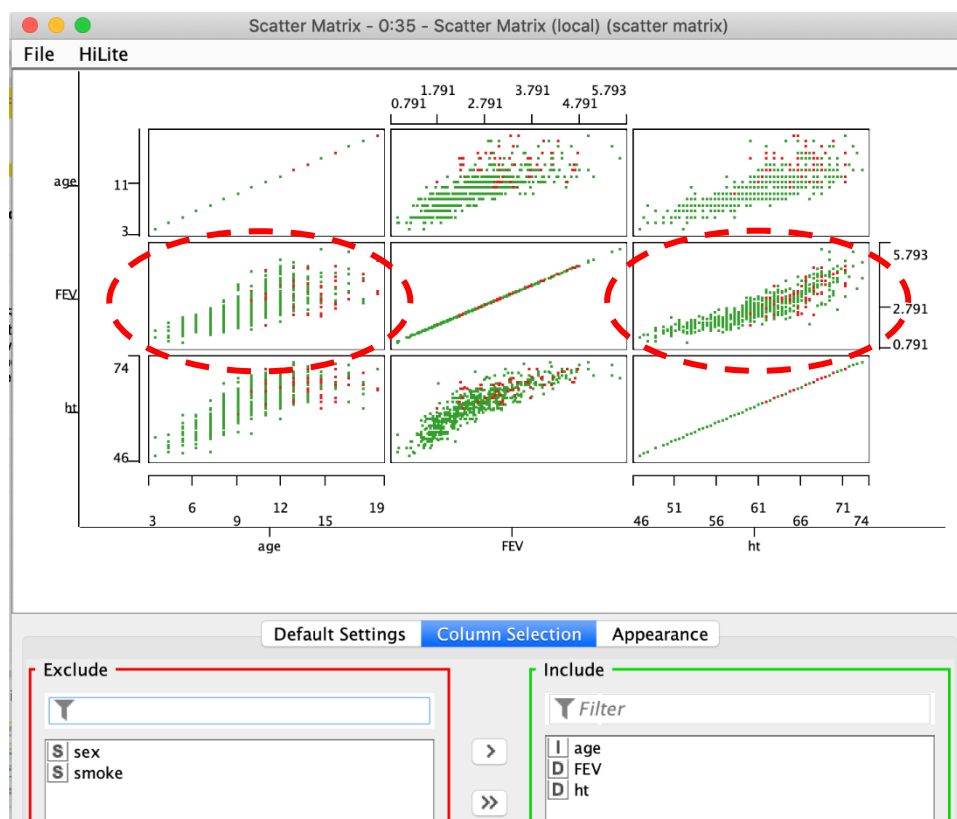
## D. Investigating effect of other variables on FEV

Let's see how FEV is affected by other variables by using scatterplots. This allows us to detect nonlinear relationship and interactions amongst variables.

1. Before that, we assign different colors to the values of the smoke variable using the **Color Manager** node.



2. Next, let's produce a scatterplot matrix using the **Scatter Matrix (local)** node. Accept the default configuration. After running the node, this is what you should see:



3. In order to see the scatterplot of FEV vs. age and FEV vs. ht in more details, generate a scatterplot using Scatter Plot (local) node respectively. (Refer to the completed workflow above for the connection.)

**Investigative tasks**

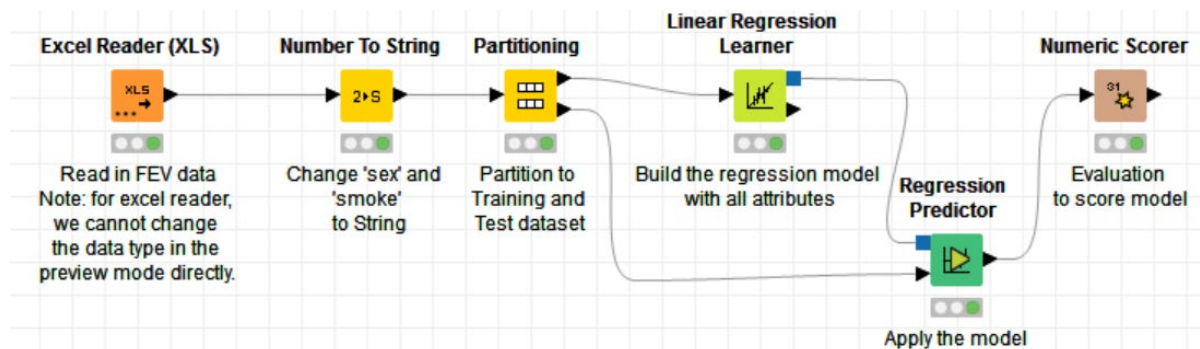
- (a) What does a 'Number To String' node do? Explain it in the context.
- (b) Write down your observations from the conditional boxplot and highlight whether smoking makes a difference to the FEV values.
- (c) Identify those variables with a nonlinear relationship with FEV (indicated by 'curve').
- (d) Use the output window of the Scatter matrix to answer the following questions:
- Is there a strong correlation between *age* with *FEV*?
  - Is there a strong correlation between *ht* with *FEV*?
  - Is the prevalence of smoking amongst older kids or younger kids?
- (e) Write down your observations from the scatterplots and highlight the relationship of *age* with *FEV* and *ht* with *FEV*.
- (f) If we change the variable in Color Manager node to 'sex', what do you observe in all the scatterplots: *FEV* vs. *age* and *FEV* vs. *ht*?

Note: It would seem that because smokers tend to be amongst older kids, the median FEV for smokers is correspondingly higher as well. Hence, we suspect that *ht* has **nonlinear** relationship with *FEV* and *age* **interacts** with *smoke*.

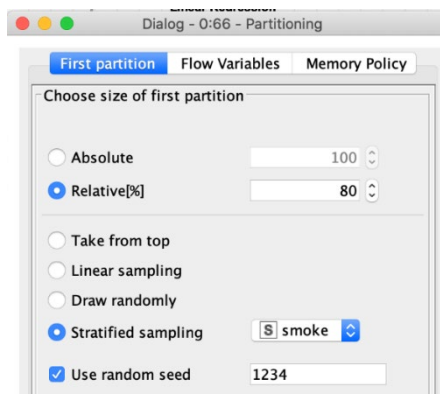
## E. Modelling (1) – MLR model with all variables

Because of the interrelationship between *age*, *smoking* and *FEV*, we need to use Multiple Linear Regression (MLR) to figure out the quantitative effect of each variable on FEV. Only then can we get a true estimation of the effect smoking has on FEV.

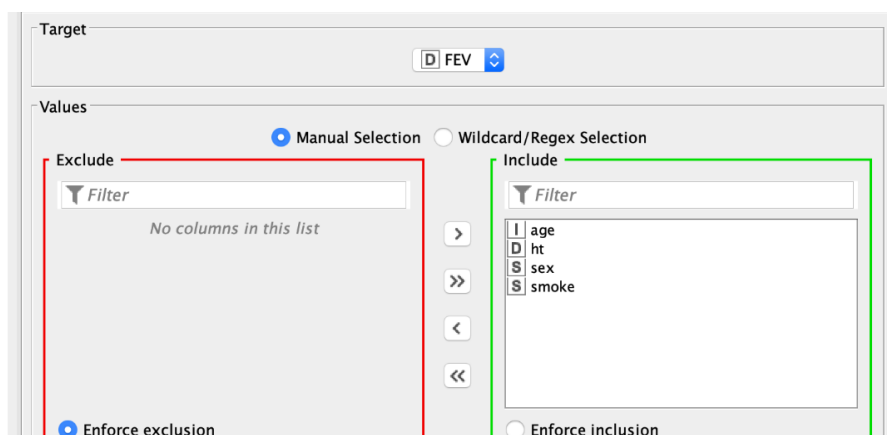
The workflow we will produce will look something like this:



1. If you continue from the previous section, ensure that the **Partitioning** node is connected to the **Number to String** node.
2. In order to gain a realistic approximation on the predictive accuracy of our model, we first use the **Partitioning** node to split the dataset according to a 80:20 ratio of training to testing dataset. We select *Stratified sampling* according to the variable *smoke* and *Use random seed* of 1234.



3. Configure the **Linear Regression Learner** as follows:



4. This is our baseline model. Run the node and note the output below:

Linear Regression Result View - 0:...

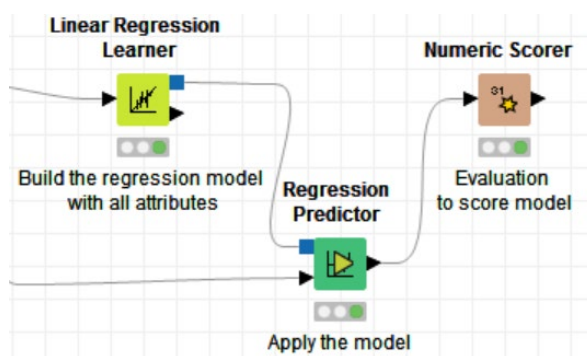
File

### Statistics on Linear Regression

Variable	Coeff.	Std. Err.	t-value	P> t
age	0.0702	0.0103	6.8104	2.71E-11
ht	0.1	0.0051	19.5445	0.0
sex=1	0.1793	0.0353	5.0743	5.43E-7
smoke=1	-0.0801	0.064	-1.2527	0.2109
Intercept	-4.2579	0.2385	-17.8504	0.0

Multiple R-Squared: 0.7869  
Adjusted R-Squared: 0.7852

5. Check the predictive accuracy of our model by connecting **Regression Predictor** and **Numeric Scorer** node as follows.



6. Configure **Numeric Scorer** as follows:

Dialog - 0:68 - Numeric Scorer

Options | Flow Variables | Memory Policy

Reference column: FEV

Predicted column: Prediction (FEV)

Output column

☐ Change column name

Output column name: Prediction (FEV)

Provide scores as flow variables

Prefix of flow variables:

☐ Output scores as flow variables

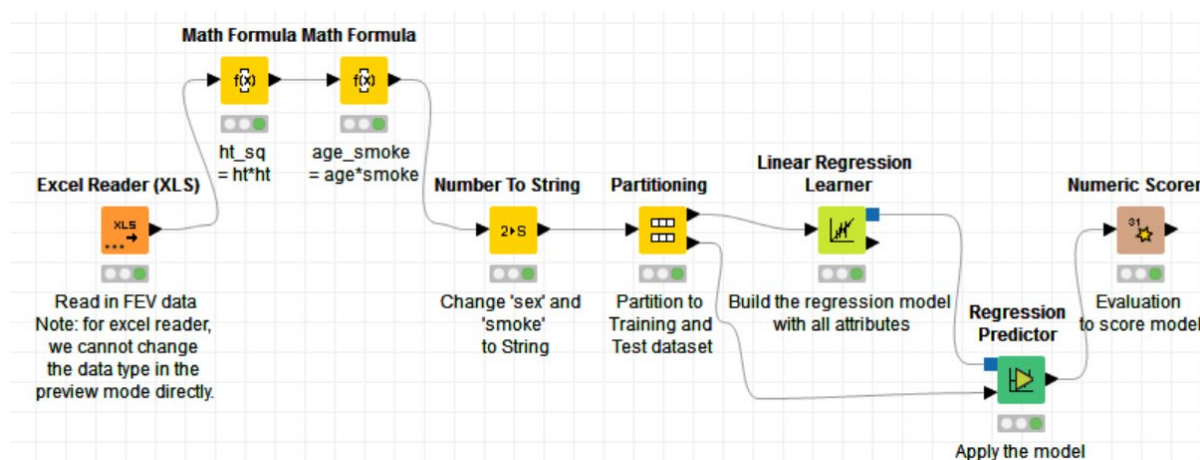
7. Run the nodes and note down the  $R^2$  value. This is the standard setup to predict accuracy for the other models we will construct.

**Investigative tasks**

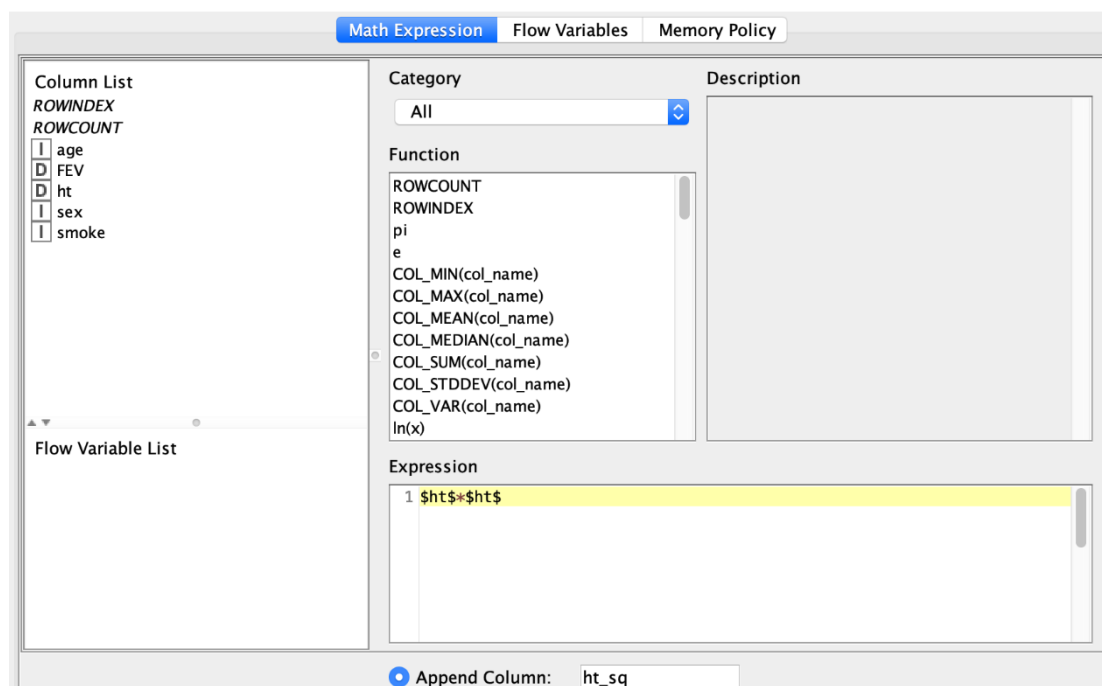
- (a) Compare the Multiple R-Squared value with the Adjusted R-Squared value of this model. Does the model suffer from overfitting?
- (b) Interpret the regression coefficient for *smoke=1*.
- (c) Does the p-value for the coefficient *smoke=1* indicate that it is statistically significant? Can we then conclude that on average, smoking lowers FEV of a child?
- (d) What is the R-Squared value for the model applied to the test data? Why do you think this value is good to predict accuracy of the model?

## 8. Modelling (2) – MLR model to include non-linear and interaction terms

There is a possibility that we have neglected **non-linear** and **interaction** terms in our model. Let's add those in.

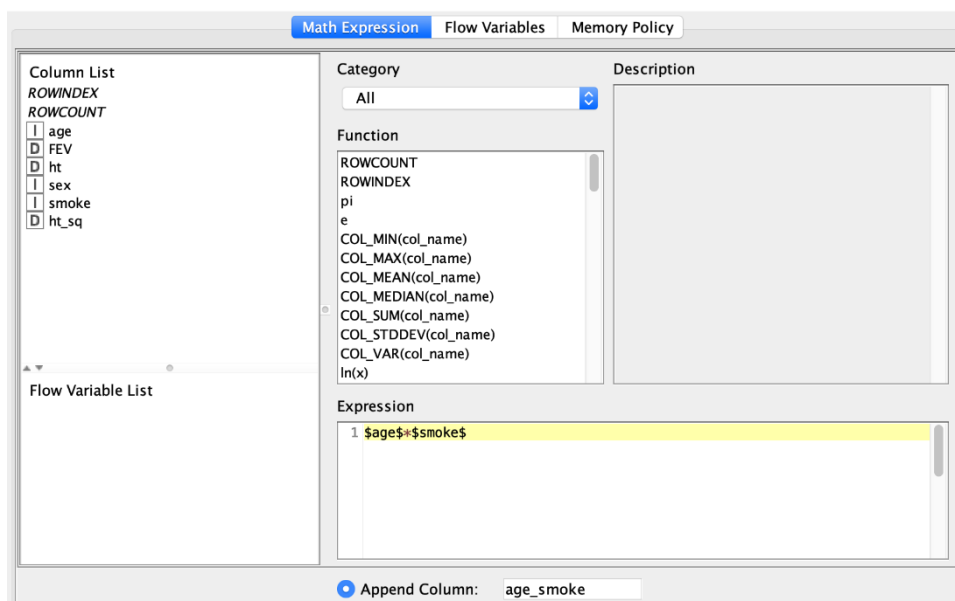


1. Use the **Math Formula** node to create a non linear term.



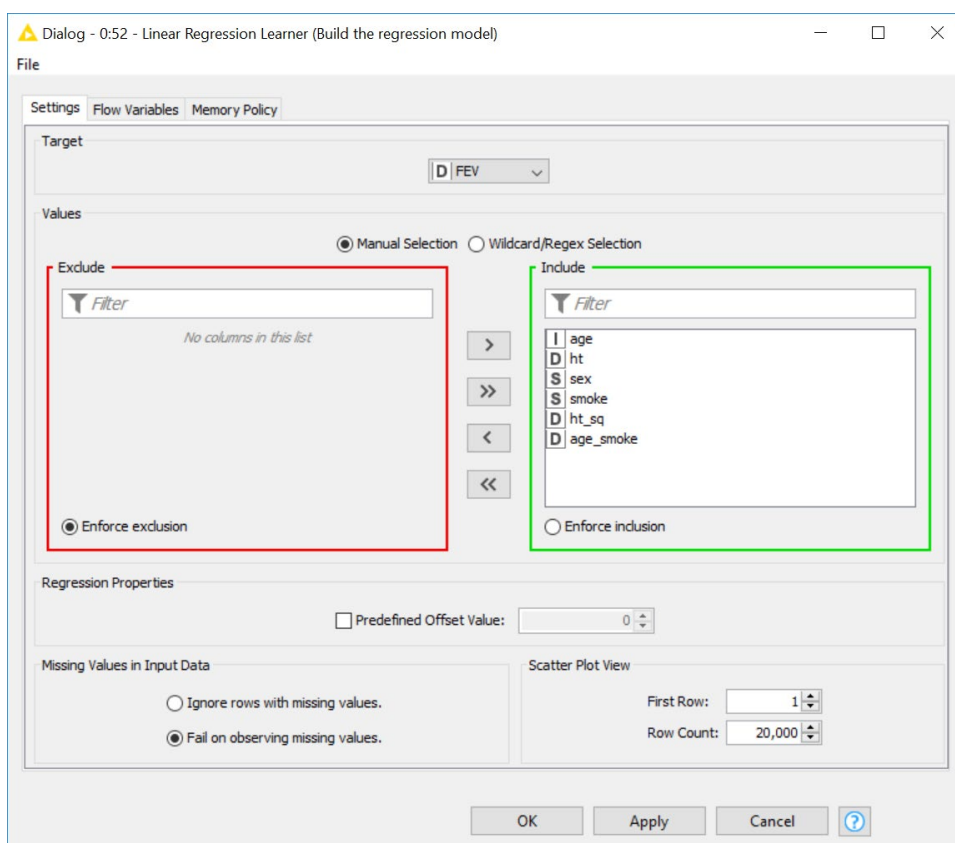
Type “\$ht\$\*\$ht\$” in the *Expression* field. Then, select *Append Column* and type “ht\_sq” in the field there.

- Now let's create an interaction term.



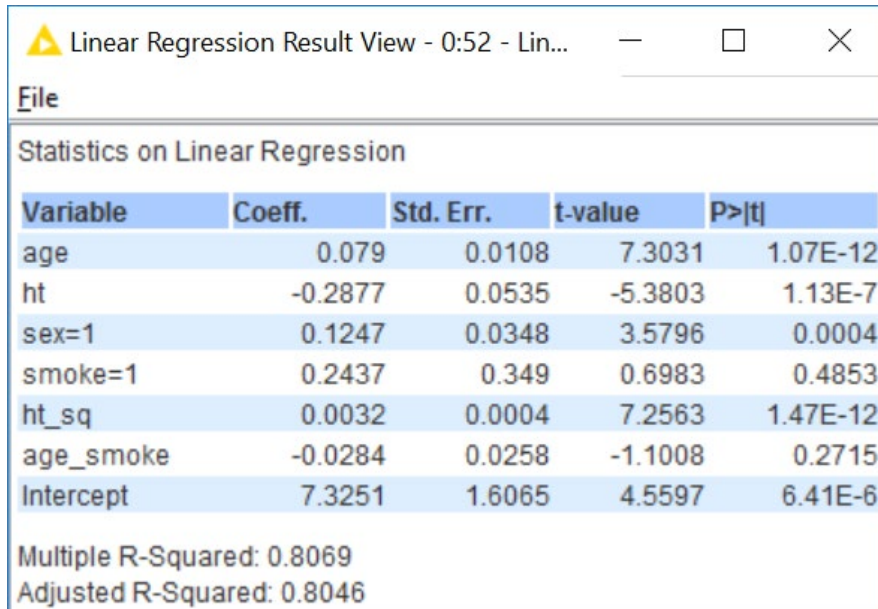
As above, type “\$age\$\*\$smoke\$” in the *Expression* field and select *Append Column* typing “age\_smoke” in the field there.

- Connect the **Math Formula** nodes to each other and then to the **Number to String** node followed by the **Partition** node configured as above.
- Now connect the training output port to the **Linear Regression Learner** node configured as below:



**Important note:** When we include the interaction term in our model, we must also include the individual predictor variables that make up the factors in the interaction. So for example, if `age_smoke` is an interaction term, both `age` and `smoke` must be part of the model as well.

5. Run the learner node. You should get a result like this:



Variable	Coeff.	Std. Err.	t-value	P> t
age	0.079	0.0108	7.3031	1.07E-12
ht	-0.2877	0.0535	-5.3803	1.13E-7
sex=1	0.1247	0.0348	3.5796	0.0004
smoke=1	0.2437	0.349	0.6983	0.4853
ht_sq	0.0032	0.0004	7.2563	1.47E-12
age_smoke	-0.0284	0.0258	-1.1008	0.2715
Intercept	7.3251	1.6065	4.5597	6.41E-6

Multiple R-Squared: 0.8069  
Adjusted R-Squared: 0.8046

### **Investigative tasks**

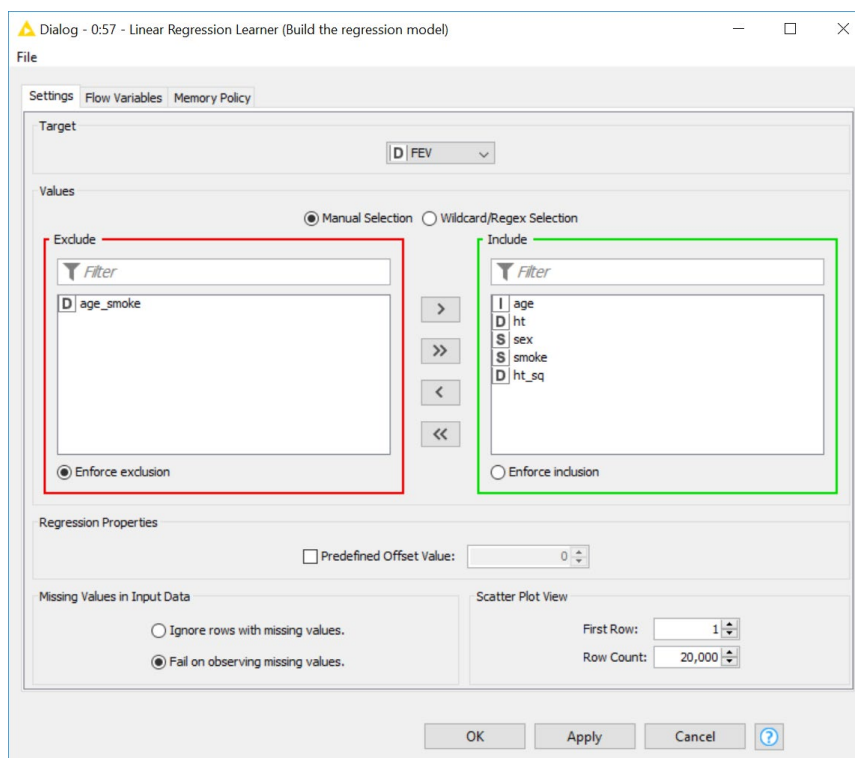
- Based on the R-Squared values you see here, is this model an improvement over the previous one?
- Observe the p-value of the `ht_sq` variable. Is there a non-linear dependence between height and FEV?
- Observe the p-value of the `age_smoke` variable. Can we conclude that smoking affects the relationship between age and FEV?
- Check the predictive accuracy of this model and compare it to the previous one. Is this a better model?



## 6. Modelling (3) – MLR model to drop variable that is not statistically significant via the backward elimination process

As seen above, *age\_smoke* is not statistically significant. Therefore, we would like to drop this term from our model. This is known as *backward elimination*. We can only drop one term at a time for every step of backward elimination. After dropping the term, we must retrain the model and examine the R2 of the model for its fit.

1. Go back to the learner node and reconfigure it as follows.



The following results are obtained

Statistics on Linear Regression				
Variable	Coeff.	Std. Err.	t-value	P> t
age	0.0741	0.0098	7.5244	2.37E-13
ht	-0.2839	0.0534	-5.3185	1.56E-7
sex=1	0.1208	0.0347	3.4849	0.0005
smoke=1	-0.1345	0.0615	-2.1874	0.0292
ht_sq	0.0032	0.0004	7.2219	1.84E-12
Intercept	7.2003	1.6028	4.4923	8.70E-6
Multiple R-Squared: 0.8064				
Adjusted R-Squared: 0.8045				

**Investigative tasks**

- (a) There is a possibility that by dropping terms from a model, the fit becomes worse. In your judgement, is the model fit still comparable to the previous model before we dropped the *age\_smoke* term?
- (b) Summarize your findings for the models (model (1), model (2) and model (3)) created in the lab.