SIMPLE REGRESSION

1. **Scatter plot:**

To do a scatter plot in R between xData and yData, use:

plot(xData, yData, col= “color Name”, main= “title of the plot”, xlab= “x name”, ylab= “y name”, pch=19)

*Note, pch= 19 means you want to plot a filled circle. You can change this number for different shape (see figure 1).*

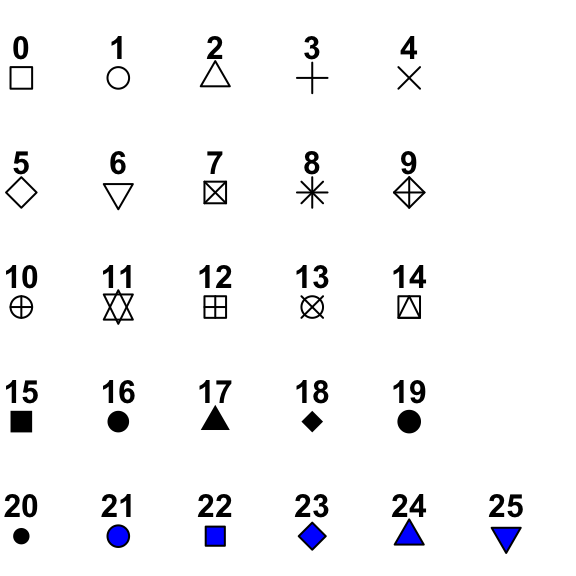


Figure 1: The different point symbols in R.

1. **Correlation and covariance:**

To find correlation between xData and yData, use:

cor(xData, yData)

To find corvariance between xData and yData, use:

cov(xData, yData)

1. **Simple regression:**

To compute a simple regression model in R (y=mx+b), use:

model= lm(yData ~ xData, the name of you data)

Then, call the following command to see the full output (see figure 2) of your model

summary(model)

*Note: If you plan to have more than one models, make sure to give each a different name (like model\_1, model\_2, etc).*

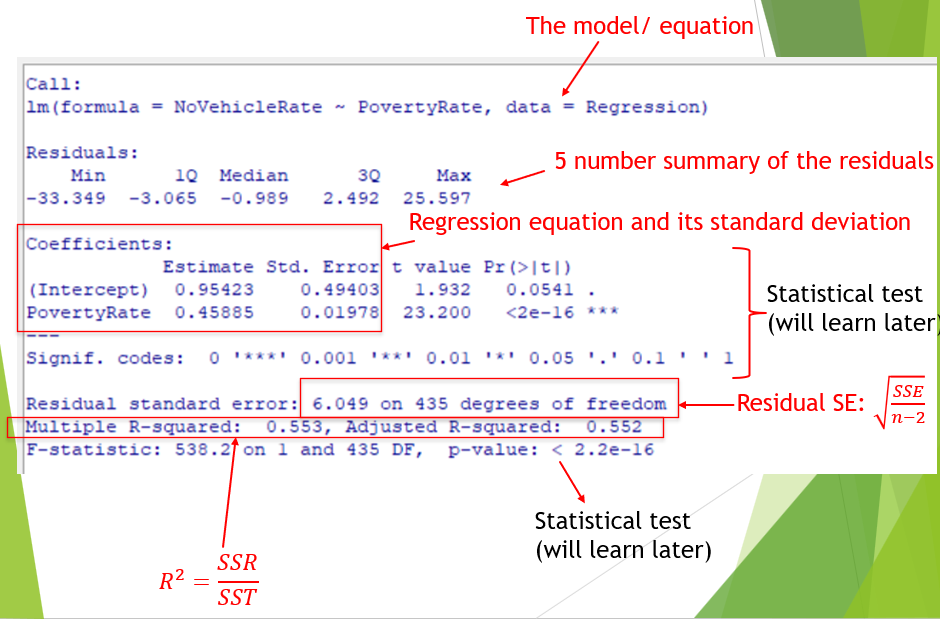


Figure 2: Simple regression output from R.

1. **Plot regression line and the scatter plot:**

To impose the regression line into the scatter plot, first you need to do the scatter plot just like before (see the Scatter plot section). Then, to add the regression line, use the command:

abline(name of the model, col= “color Name”, lwd= 3)

The option “lwd” tells you how thick the line is. If you want a thick line, use bigger number.

Example 1: Import the data HealthData in R. Do a quick EDA (summary statistics, histogram and QQ plot) for Smoking rate and Life expectancy. Then, fit a regression model that has smoking rate as predictor and life expectancy as the response.

Write down the equation of the model you got from R:

What is the slope? What does it mean in the context of this problem?

What is the intercept? What does it mean in the context of this problem?

Is there a significant linear relationship? Explain. State the t test hypothesis, t-value, and p-value.

What is the value? What is the correlation of the two variables?

Example 1 (Cont’d): Plot the regression line together with the scatter plot.

1. **Access important numerical values of the model:**

To access the intercept, use:

model\_int= modelName$coefficients[1]

To access the slope, use:

model\_slope= modelName$coefficients[2]

To access the fitted y values and save it in y\_predicted, use:

y\_predicted= Fitted(modelName)

To access the residual and save it in y\_residual, use:

y\_residual= resid(model)

It is also good to now put your original data together with the fitted values and the residual into a single data set. To do that, use the command:

CompleteDataName=dataFrame(Name of the original data, y\_predicted, y\_residual)

To save this data frame into a .csv file, use the command:

write.csv(CompleteDataName, "Name of your new file.csv", row.names=TRUE)

Make sure to give it a different name from the name of your original file.

Example 2: Compute the fitted values and the residual. Then, put them in a new data frame together with your original data. Save this new data frame as a csv file. Make sure to give it a different name from your original data.

Example 2 (Cont’d): Compute SST, SST, SSE, and residual standard error (enter the math formula in R for these). Compare your computed residual standard error with the R output from the model (they should be the same).

1. **Prediction interval and mean response interval:**

To find mean response interval for x=a, use the following command:

NameOfDataFrame= data.frame(xName=a)

predict(model, NameOfDataFrame, interval= “confidence”, level=.95)

If you want a different confidence level, you can change that value in level.

To find prediction interval for x=a, use the following command:

NameOfDataFrame= data.frame(xName=a)

predict(model, NameOfDataFrame, interval= “prediction”, level=.95)

If you want a different confidence level, you can change that value in level.

Example 3: Find the prediction interval and mean response interval at 95% CI for current smoking rate = 17.08. Note that when you call your data frame, you need to have the correct name for your variable. In this case, it has to be: *data.frame(Intercept=1, CurrentSmoker.=17.08)*

Interpret your intervals.

1. **Drawing prediction interval and mean response interval:**

To draw lines of mean response interval and prediction interval for all values of x, you first need to define a vector of those values using the command:

*library(pracma)*

*N=100* # How many values you want

*zInt=linspace(1, 1, n=N)* # this is for the intercept

*zSmoke=linspace(min(HealthData$CurrentSmoker.),max(HealthData$CurrentSmoker.), n=N)* #this is for the smoking rate

*zVec=data.frame(Intercept=zInt, CurrentSmoker.=zSmoke)* # put them both in the data frame.

Then, compute the prediction interval and the mean response interval just like before. Make sure to give them a name.

To add lines of mean response interval and prediction interval, use the command:

*lines(zVec[,2], Name of your prediction interval[,2], col= ‘colorName) #lower prediction value*

*lines(zVect[,2], Name of your prediction interval [,3], col= ‘colorName) #upper prediction value*

*lines(zVect[,2], Name of your mean response interval[,2], col= ‘colorName) #lower mean response value*

*lines(zVect[,2], Name of your mean response interval [,3], col= ‘colorName) #upper mean response value*

Here you use zVect[,2] for the smoking rate values.

Example 3 (Cont’d): Plot the data together with the regression line (just as Example 1 above), the mean response intervals, and the prediction intervals.

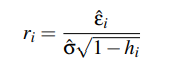
1. **Diagnostic plots:**

To do diagnostic plots, you can enter

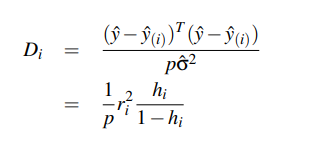
*plot(the name of your model)*

This will give you 4 plots.

Here, standardized residual means



Cook statistics as a function of standardize residual and the leverage is given below. Note that this is a function of two variables (in 3D). Thus, the plot is the contour plot.



Example 4: Do a quick diagnostic plot for your model. How is the normality, independent, and equal variance assumption? Is there any potential unusual observation? What state is it?

Example 4 (Cont’d): If there is an unusual observation, redo the regression model without that value.

Name of your new model= lm(yData ~ xData, Name of you data, subset=-c(rowIndex))

Then, see if there is a drastic difference (in the intercept values, the slope values, the t statistics, the p value, and the value) between this new fit and the previous fit

Example 4 (Cont’d): Plot this new regression line and the previous regression line together with the data on the same graph. Make sure to have different colors for the regression lines. Is there a big difference? Is this data point influential?

What if you have many potential problematic observations? It will not be too fun to check all of them. Use this to do a quick check.

*plot(dfbeta(Name of you original model)[,parameter index],ylab="Change in the parameter")*

*abline(h=0)*

Here, R will do automatically remove each data point and do regression without each of them and plot the change in the parameter for each case. The parameter index is 1 if you want to check the change in intercept. The parameter index is 2 if you want to check the change in slope.

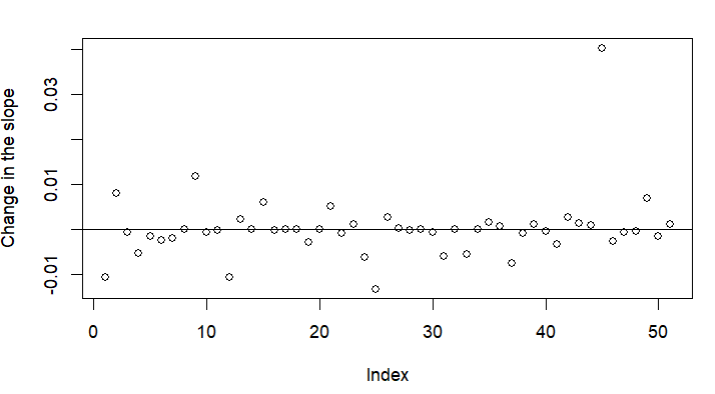




Figure 2: This is an output from R. Here you see that if we remove data with index 45, there is a big change (relatively compared to removing other data) in the slope parameter.

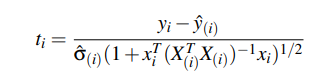
Example 5: Check and plot the change of the intercept and the slope if we remove each of the data. Is there any datapoint that stands out? If so, which state is this? Is this consistent with what you see in the diagnostic plot above?

If you do not like automatic diagnostic plot in R, you can plot them yourselves. Also, the automotic R plots do not include all the plots. You can compute those values as

*hatvalues(Name of your model)* # hat matrix

*rstandard(Name of your model)* # standardized residuals

*rstudent(Name of your model)* # studentized residuals

**

*cooks.distance(Name of your model)* # Cook distance

*resid(Name of your model)* # residuals

Example 6: Compute the hat matrix (for leverage values), the standardize residual values, the student dized residual values, the Cook distance, and the residuals.

Example 6 (Cont’d): Plot the residual values against the predictor values. Do you see any pattern in your plot?