# Topic: Predictive Data Analytics in R

## Sub topic: Regression Analysis

### Exercise 1: Understanding Regression

**Table of Contents**

1. **What is Regression analysis?**
2. **Why do we use regression analysis?**
3. **What are the types of regressions?**
4. **Linear Regression**
5. **Logistic Regression**
6. **How to select the correct regression model?**

**Key terms**

1. **Predictor variable**

**Predictor variable is also known as the independent variable**

1. **Response variable**

**Response variable is also known as the dependent variable or sometimes the target variable**

1. **What is Regression analysis?**

**Regression analysis is a type of predictive modeling technique that investigates the relationship between a dependent (response/target) and independent(s) (predictor) variables. This method is useful for forecasting time series modeling and finding the causal effect relationship between the variables. For example, the relationship between rash driving and the number of road accidents can best be studied through regression. Better put, regression analysis helps in degerming relationship between the variables.**

**Regression is widely used for prediction and forecasting in field of machine learning. Focus of regression is on the relationship between dependent and one or more independent variables. The *“dependent variable” represents the output or effect, or is tested to see if it is the effect*. The *“independent variables” represent the inputs or causes, or are tested to see if they are the cause*. Regression analysis helps to understand how the value of the dependent variable changes when any one of the independent variables is varied, while the other independent variables are kept unchanged. In the regression, dependent variable is estimated as function of independent variables which is called regression function.**

**Regression model involves following variables.**

* **Independent variables X.**
* **Dependent variable Y**
* **Unknown parameter θ**

1. **Why do we use regression analysis?**

**There are multiple benefits of using regression analysis. They are as follows:**

* **It indicates the significant relationships between dependent variable and independent variable.**
* **It indicates the strength of impact of multiple independent variables on a dependent variable.**

**Regression analysis also allows us to compare the effects of variables measured on different scales, such as the effect of price changes and the number of promotional activities. These benefits help market researchers / data analysts / data scientists to eliminate and evaluate the best set of variables to be used for building predictive models.**

**3.a. Simple Linear Regression**

**It is one of the most widely known modeling technique. Linear regression is usually among the first few topics which people pick while learning predictive modeling. In this technique, the dependent variable is continuous, independent variable(s) can be continuous or discrete, and nature of regression line is linear.**

**Linear Regression establishes a relationship between dependent variable (Y) and one or more independent variables (X) using a best fit straight line (also known as regression line).**

**It is represented by an equation Y=a+b\*X + e, where a is intercept, b is slope of the line and e is error term. This equation can be used to predict the value of target variable based on given predictor variable(s).**

**The difference between simple linear regression and multiple linear regression is that, multiple linear regression has (>1) independent variables, whereas simple linear regression has only 1 independent variable.**

**Important Points:**

* **There must be linear relationship between independent and dependent variables**
* **Multiple regression suffers from multicollinearity, autocorrelation, heteroskedasticity.**
* **Linear Regression is very sensitive to Outliers. It can terribly affect the regression line and eventually the forecasted values.**
* **Multicollinearity can increase the variance of the coefficient estimates and make the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable**
* **In case of multiple independent variables, we can go with forward selection, backward elimination and step wise approach for selection of most significant independent variables.**

**Within multiple types of regression models, it is important to choose the best suited technique based on type of independent and dependent variables, dimensionality in the data and other essential characteristics of the data. Below are the key factors that you should practice to select the right regression model:**

1. **How to select the correct regression model?**

* **Data exploration is an inevitable part of building predictive model. It should be you first step before selecting the right model like identify the relationship and impact of variables**
* **To compare the goodness of fit for different models, we can analyse different metrics like statistical significance of parameters, R-square, Adjusted r-square, AIC, BIC and error term. Another one is the Mallow’s Cp criterion. This essentially checks for possible bias in your model, by comparing the model with all possible submodels (or a careful selection of them).**
* **Cross-validation is the best way to evaluate models used for prediction. Here you divide your data set into two group (train and validate). A simple mean squared difference between the observed and predicted values give you a measure for the prediction accuracy.**
* **If your data set has multiple confounding variables, you should not choose automatic model selection method because you do not want to put these in a model at the same time.**
* **It’ll also depend on your objective. It can occur that a less powerful model is easy to implement as compared to a highly statistically significant model.**
* **Regression regularization methods(Lasso, Ridge and ElasticNet) works well in case of high dimensionality and multicollinearity among the variables in the data set.**

### **Linear Regression in R**

|  |
| --- |
| > str(women)  'data.frame': 15 obs. of 2 variables:  $ height: num 58 59 60 61 62 63 64 65 66 67 ...  $ weight: num 115 117 120 123 126 129 132 135 139 142 ...  > fit <-lm(weight ~ height, data=women)  > summary(fit)  Call:  lm(formula = weight ~ height, data = women)  Residuals:  Min 1Q Median 3Q Max  -1.7333 -1.1333 -0.3833 0.7417 3.1167  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) -87.51667 5.93694 -14.74 0.0000000017110819 \*\*\*  height 3.45000 0.09114 37.85 0.0000000000000109 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 1.525 on 13 degrees of freedom  Multiple R-squared: 0.991, Adjusted R-squared: 0.9903  F-statistic: 1433 on 1 and 13 DF, p-value: 0.00000000000001091 |

Understanding the output

Values of coefficients(θs) are -87.51667 and 3.45000, hence prediction equation for model is as below

Weight = -87.52 + 3.45\*height

Now, we will look at real values of weight of 15 women first and then will look at predicted values. Actual values of weight of 15 women are as below

|  |
| --- |
| > women$weight  [1] 115 117 120 123 126 129 132 135 139 142 146 150 154 159 164 |

We can see that predicted values are nearer to the actual values.

|  |
| --- |
| > fitted(fit)  1 2 3 4 5 6 7 8 9 10 11 12 13  112.5833 116.0333 119.4833 122.9333 126.3833 129.8333 133.2833 136.7333 140.1833 143.6333 147.0833 150.5333 153.9833  14 15  157.4333 160.8833 |

**Remember: Correlation does not imply Causation**

In the regression, dependent variable is correlated with the independent variable. This means, as the value of the independent variable changes, value of the dependent variable also changes. But, this does not mean that independent variable cause to change the value of dependent variable. Causation implies correlation, but reverse is not true.

Example 2:

Problem Statement: Load the Boston housing dataset from the MASS library in R. Create a simple linear regression model to predict the median value of owner-occupied homes (mdev) over lower status of the population (lstat)

|  |
| --- |
| > library(MASS)  > str(Boston)  'data.frame': 506 obs. of 14 variables:  $ crim : num 0.00632 0.02731 0.02729 0.03237 0.06905 ...  $ zn : num 18 0 0 0 0 0 12.5 12.5 12.5 12.5 ...  $ indus : num 2.31 7.07 7.07 2.18 2.18 2.18 7.87 7.87 7.87 7.87 ...  $ chas : int 0 0 0 0 0 0 0 0 0 0 ...  $ nox : num 0.538 0.469 0.469 0.458 0.458 0.458 0.524 0.524 0.524 0.524 ...  $ rm : num 6.58 6.42 7.18 7 7.15 ...  $ age : num 65.2 78.9 61.1 45.8 54.2 58.7 66.6 96.1 100 85.9 ...  $ dis : num 4.09 4.97 4.97 6.06 6.06 ...  $ rad : int 1 2 2 3 3 3 5 5 5 5 ...  $ tax : num 296 242 242 222 222 222 311 311 311 311 ...  $ ptratio: num 15.3 17.8 17.8 18.7 18.7 18.7 15.2 15.2 15.2 15.2 ...  $ black : num 397 397 393 395 397 ...  $ lstat : num 4.98 9.14 4.03 2.94 5.33 ...  $ medv : num 24 21.6 34.7 33.4 36.2 28.7 22.9 27.1 16.5 18.9 ...  > lm.fit=lm(medv~lstat, data = Boston)  > summary(lm.fit)  Call:  lm(formula = medv ~ lstat, data = Boston)  Residuals:  Min 1Q Median 3Q Max  -15.168 -3.990 -1.318 2.034 24.500  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 34.55384 0.56263 61.41 <2e-16 \*\*\*  lstat -0.95005 0.03873 -24.53 <2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 6.216 on 504 degrees of freedom  Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432  F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16 |

Visualizing the linear regression results

|  |
| --- |
| > par(mfrow=c(2,2)) # The R function par() tells R to split the screen into separate panels so that multiple plots can be viewed simultaneously  > plot(lm.fit) |

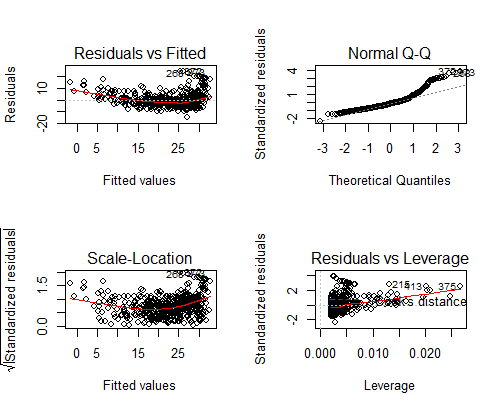


Fig. 1. Simple Linear Regression plot

**3.b. Multiple Linear Regression**

As the name indicates, there are more than one independent (or predictors) variables and one dependent (or response) variable

Problem Statement: Using the Boston housing dataset from the MASS library, determine the most significant predictors to predict the median value of a house

|  |
| --- |
| # Multiple linear regression  > library(MASS)  > lm.fit=lm(medv~., data = Boston) # here ~. indicates to use all the predictors  > summary(lm.fit)  Call:  lm(formula = medv ~ ., data = Boston)  Residuals:  Min 1Q Median 3Q Max  -15.595 -2.730 -0.518 1.777 26.199  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 36.4594884 5.1034588 7.144 0.000000000003283 \*\*\*  crim -0.1080114 0.0328650 -3.287 0.001087 \*\*  zn 0.0464205 0.0137275 3.382 0.000778 \*\*\*  indus 0.0205586 0.0614957 0.334 0.738288  chas 2.6867338 0.8615798 3.118 0.001925 \*\*  nox -17.7666112 3.8197437 -4.651 0.000004245643808 \*\*\*  rm 3.8098652 0.4179253 9.116 < 2e-16 \*\*\*  age 0.0006922 0.0132098 0.052 0.958229  dis -1.4755668 0.1994547 -7.398 0.000000000000601 \*\*\*  rad 0.3060495 0.0663464 4.613 0.000005070529023 \*\*\*  tax -0.0123346 0.0037605 -3.280 0.001112 \*\*  ptratio -0.9527472 0.1308268 -7.283 0.000000000001309 \*\*\*  black 0.0093117 0.0026860 3.467 0.000573 \*\*\*  lstat -0.5247584 0.0507153 -10.347 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 4.745 on 492 degrees of freedom  Multiple R-squared: 0.7406, Adjusted R-squared: 0.7338  F-statistic: 108.1 on 13 and 492 DF, p-value: < 2.2e-16 |

The \* besides the predictor indicates its significance. More the number of \* (asterisks) greater the predictor significance. Therefore, the variables *zn, nox, rm, dis, rad, ptratio, black and lstat are highly significant predictors*.

|  |
| --- |
| > summary(lm.fit)$r.sq # gives the Rsquared  [1] 0.7406427  > summary(lm.fit)$sigma # gives the RSE  [1] 4.745298 |

**3.c. Multicollinearity Detection in R**

Variance Inflation Factor (VIF) is used to detect multicollinearity amongst the predictors. Use the car package to use the VIF function in R.

The smallest possible value of VIF is 1 which indicates complete absence of multicollinearity. As a rule of thumb, a VIF value that exceeds 5 or 10 indicates a problematic amount of multicollinearity. The predictors rad (index of accessibility to radial highways) and tax (full-value property-tax rate per \$10,000.) are highly multicollinear with values greater than 5.

|  |
| --- |
| > library(car) # to use the VIF()  > vif(lm.fit) # VIF is used to detect multicollinearity among the predictors.  crim zn indus chas nox rm age dis rad tax ptratio  1.792192 2.298758 3.991596 1.073995 4.393720 1.933744 3.100826 3.955945 7.484496 9.008554 1.799084  black lstat  1.348521 2.941491 |

**Method 1: remove one of the highly multicollinear variable and then compute the lm and vif**

|  |
| --- |
| > boston.new<-Boston[,c(1:9,11:14)] # here dropping the tax variable  > lm.fit.new<- lm(medv~.,data = boston.new)  > summary(lm.fit.new)  Call:  lm(formula = medv ~ ., data = boston.new)  Residuals:  Min 1Q Median 3Q Max  -16.1449 -2.9143 -0.5661 1.7438 26.3113  Coefficients:  Estimate Std. Error t value Pr(>|t|)  (Intercept) 34.6286431 5.1228018 6.760 0.000000000039178 \*\*\*  crim -0.1067339 0.0331863 -3.216 0.001384 \*\*  zn 0.0363707 0.0135129 2.692 0.007354 \*\*  indus -0.0677783 0.0558291 -1.214 0.225317  chas 3.0292314 0.8636515 3.507 0.000494 \*\*\*  nox -18.7012125 3.8466152 -4.862 0.000001566976404 \*\*\*  rm 3.9116902 0.4208752 9.294 < 2e-16 \*\*\*  age -0.0006054 0.0133339 -0.045 0.963804  dis -1.4883027 0.2013809 -7.390 0.000000000000631 \*\*\*  rad 0.1345757 0.0412534 3.262 0.001182 \*\*  ptratio -0.9851286 0.1317385 -7.478 0.000000000000348 \*\*\*  black 0.0095464 0.0027115 3.521 0.000470 \*\*\*  lstat -0.5222095 0.0512087 -10.198 < 2e-16 \*\*\*  ---  Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1  Residual standard error: 4.792 on 493 degrees of freedom  Multiple R-squared: 0.735, Adjusted R-squared: 0.7285  F-statistic: 113.9 on 12 and 493 DF, p-value: < 2.2e-16 |

|  |
| --- |
| > vif(lm.fit.new) # multicolinearity is under control now  crim zn indus chas nox rm age dis rad ptratio black  1.791940 2.184240 3.226015 1.058220 4.369271 1.923075 3.098044 3.954446 2.837494 1.788839 1.347564  lstat  2.940800 |

After removing the highly collinear predictors, we can now see that multicollinearity is under control now

**Method 2: Use Partial Least Squares (PLS) or Principal Component Analysis (PCA) regression methods that cut the number of predictors to a smaller set of uncorrelated components.**

|  |
| --- |
| > boston.pca<- princomp(boston.new) |

To view the proportion of the total variance explained by each component, use the command summary()

|  |
| --- |
| > summary(boston.pca)  Importance of components:  Comp.1 Comp.2 Comp.3 Comp.4 Comp.5 Comp.6 Comp.7  Standard deviation 91.963548 32.1435632 16.57664377 10.18868804 7.764898294 5.476052224 4.121750109  Proportion of Variance 0.846288 0.1033891 0.02749666 0.01038779 0.006033347 0.003000695 0.001700004  Cumulative Proportion 0.846288 0.9496771 0.97717373 0.98756152 0.993594870 0.996595565 0.998295570  Comp.8 Comp.9 Comp.10 Comp.11 Comp.12 Comp.13  Standard deviation 3.595809132 1.6499240096 1.04927248 0.46639314286 0.244010751630 0.0540811193853  Proportion of Variance 0.001293838 0.0002724047 0.00011017 0.00002176662 0.000005958057 0.0000002926699  Cumulative Proportion 0.999589408 0.9998618127 0.99997198 0.99999374927 0.999999707330 1.0000000000000 |

From the summary, we can see that 94% (Add the Proportion of variance for Comp.1 & Comp.2) is shown by the first two components.

**Question:** So, how do we decide how many components should we select for modeling stage?

**Answer: Scree plot** to determine the number of relevant predictors

A scree plot is used to access components or factors which explains the most of variability in the data. It represents values in descending order

|  |
| --- |
| > screeplot(boston.pca) |