Polynomial Regression

Lab 8 R Notes: EXST 7014/15

Contents

0.1 Objectives

1. Use the **lm** function to fit polynomial regression models regression models

Polynomial regression is a statistical modeling technique to fit the curvilinear data that either shows a maximum or a minimum in the curve, or that could show a max or min if you extrapolated the curve beyond your data. The ability to determine a minimum or maximum point based on the experimental data is a useful application of polynomials. The simple polynomial regressions are multiple regression that use power terms of the independent variable (X_i) with the form of $Y = \beta_0 + \beta_1 X_i + \beta_2 X_i^2 + ... + \beta_k X_i^k + \epsilon_i$. Notice the subtle difference from multiple linear regression model, here the numbers $2, 3, , k$ represent the powers of the same variable.

When fitting polynomial regression models, if a particular model term is significant, all terms of lower order should be assumed significant and retained in the regression model. For this reason, the sequentially adjusted **Type I Sums of Squares** should be used when one attempts to test whether a polynomial model is as good as the one with a higher order term. There is also a null hypothesis for each equation that says that it does not fit the data significantly better than a horizontal line; in other words, that there is no relationship between the X and Y variables.

It should be noticed that the fully adjusted regression coefficients are still used to fit the polynomial regression, which usually leads to no practical explanation of regression coefficients. Further, extrapolation outside the range of the fitted experimental data is untenable, and should not be attempted. This is because the shape of the regression function, as predicted by the model, may not at all accurately represent reality outside the range of the experimental data.

All of the assumptions for regression apply to polynomial regression analyses. The assumptions of normality and homogeneity can and should be verified via use of customary diagnostic techniques (Shapiro-Wilk test, residual plots). Residual and influence statistics still work with these models.

0.2 Getting help with the lm function and packages in R

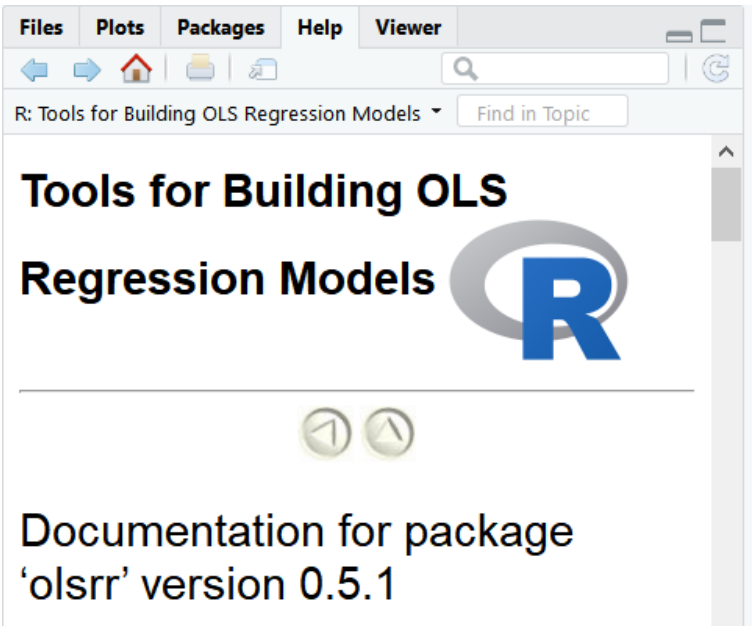
You can get by quite well (for a basic user) in R by simply knowing how to seek help regarding R functions and packages. R has a vibrant help or user community - as such, has exhaustive documentation especially for the most common packages.

You can get illustrative examples as well as argument definitions of most functions in R. Likewise you can get documentation on all the various functions that come with packages. See examples below:

```r
#' olsrr  - package that assesses model fit and variable diagnostics
library(olsrr)

#' use the @help funtion to get documentation on functions
#'    that come with ANY package
help(package = "olsrr")
```

If you are using RStudio (highly recommended), you can see a list of all the functions of the **olsrr** package with their respective usage information after running the code above. The results would be displayed in the help sub-menu pane situated most likely to your bottom right of the RStudio window.




```r
# get help with the lm function and its various methods/attributes
help(formula.lm)
```

The results can also be seen in the **help** sub-menu pane.

0.3 The Data

The data set that we use contains the per capita state and local public expenditures and associated state demographic and economic characteristics for 48 states during the year of 1960. Detailed information can be found at http://lib.stat.cmu.edu/DASL/Datafiles/pubexpendat.html. The variables in the dataset are:

- **EX**: Per capita state and local public expenditures ($)

- **ECAB**: Economic ability index, in which income, retail sales, and the value of output (manufactures, mineral, and agricultural) per capita are equally weighted.
- **MET**: Percentage of population living in standard metropolitan areas
- **GROW**: Percent change in population, 1950-1960
- **YOUNG**: Percent of population aged 5-19 years
- **OLD**: Percent of population over 65 years of age
- **WEST**: Western state (1) or not (0)

In this lab, we will use MET, Percentage of population living in standard metropolitan area, as the independent variable and the expenditure (EX) as the dependent variable. Other variables are dropped from the dataset.

```
#' Download the data_lab8.txt file to your working directory
#' Create an object to host the data set
#'
#' @sep="" because the columns are seperated by 'space'

Expenditure <- read.table('data_lab8.txt', header = TRUE)

str(Expenditure)  # get a structure (description) of your dataset

#View(Expenditure) # to view the liver dataset in RStudio's GUI pane

with(Expenditure, plot(MET, EX)) # Plot model variables
```

0.4 Fitting a Polynomial Regression Model with a Cubic term of MET

To fit the polynomial, we are going to use the **lm** function. This implies that the model object can be subjected to any function that takes an object of class **lm**. This means that you can use all the various methods and functions explored in the previous labs to further dissect the lm object for relevant information.

The are many ways to fit polynomial models in **R**, however for the sake of output-readability regarding the Type I SS we are going to use a 'less-efficient' method to fit the polynomial model.

```
#' Use the identity function @I() to create polynomial terms

cubicModel <- lm(EX ~ MET + I(MET^2) + I(MET^3), data=Expenditure)

options(scipen=999) # prevents R from reporting numbers with scientific notation
summary(cubicModel)

#' Extract Type I SS
anova(cubicModel)
```

Carefully examine the **F-tests of parameter estimates**. Keep in mind that the sequentially adjusted

type I sums of squares are used for polynomial regression.

```
#' plot fitted values against model residuals
plot(cubicModel, which=1)

#' Tests the assumption of the normality of the residuals
ols_test_normality(cubicModel)
```

Here **I( )** is the identity function, which instructs R to "leave the original data frame (or dataset without the polynomial terms) alone". We use it here because the usual symbol for raising to a power,^, has a special meaning in linear-model formulas - that is, it creates interaction terms. So "leaving the data frame alone" allows for polynomial terms void of the "interaction" connotation.

Alternatively, you can use the **poly** function, and specify the order or degree of the polynomial using the **degree** argument as below:

```
cubicModel2 <- lm(EX ~ poly(MET, degree = 3, raw=TRUE), data=Expenditure)
```

Here, **degree=3** because we are fitting the cubic model $Y \sim X, X^2, X^3$

0.5 Fitting a Polynomial Regression Model with a Quadratic term of MET

```
quadModel <- lm(EX ~ MET + I(MET^2), data=Expenditure)
summary(quadModel)

#' Extract Type I SS
anova(quadModel)

#' plot fitted values against model residuals
plot(quadModel, which=1)

#' Tests the assumption of the normality of the residuals
ols_test_normality(quadModel)
```

0.6 Lab Assignment

Your assignment is to perform necessary analysis using either SAS or R to answer the following questions Only print the graphs and tables that you think are relevant to your answers.

0.6.1 Question 1

Describe the trend in the scatterplot of the raw data: what is the relationship between variables EX and MET?

0.6.2 Question 2

Fit a polynomial regression model with cubic term of MET. When you decide whether the cubic term and quadratic term should be included in the model, do you use the Type I SS or Type II SS? Why?

0.6.3 Question 3

Is the cubic effect significant? How about quadratic and linear effects?

0.6.4 Question 4

Based on your answers to the above questions and the SAS or R output, which polynomial model do you consider the best? Write down the polynomial model with the estimated coefficient values. Do you keep the linear term in the model? Why?

0.6.5 Question 5

Now assume that there is a state where 80 percent of its residents live in standard metropolitan areas. Use the best model to predict the per capita public expenditure of this state. Is there any problem in doing so?