RegressIt

Free Excel add-in for regression and multivariate data analysis

Linear regression analysis of auto-mpg data set

Data preparation

Visit regressit.com for the software and data and links to related videos

First step in regression analysis: *obtain and prepare the data*

- For this exercise: the source data is the famous "auto-mpg" data set
  - Originally released in 1983 for the American Statistical Association Data Expo
  - Widely used on the internet for demonstrating regression
  - Popular on R sites
  - *The analysis shown here is different from the usual ones*.
- Sample consists of 392 cars manufactured between 1970 and 1982
  - Objective: predict a car's fuel economy from its physical parameters such as weight and engine size and power
  - Variables are **mpg**, **cylinders**, **displacement**, **horsepower**, **weight**, **acceleration**, **year**, **origin**, and **car name**.

Top of data spreadsheet:

mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin	name
18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10	70	1	ford galaxie 500
14	8	454	220	4354	9	70	1	chevrolet impala
14	8	440	215	4312	8.5	70	1	plymouth fury iii
14	8	455	225	4425	10	70	1	pontiac catalina
15	8	390	190	3850	8.5	70	1	amc ambassador dpl
15	8	383	170	3563	10	70	1	dodge challenger se
14	8	340	160	3609	8	70	1	plymouth 'cuda 340
15	8	400	150	3761	9.5	70	1	chevrolet monte carlo

First lesson of preparing data: ***you don't have to use the names and units of variables that you are given***

- Think ahead to what your output is going to look like and whether it will be easy to interpret and present.
- Variable names should be self-descriptive.
- A 1-unit change should be neither negligible nor impossible.
- You may want to make some changes if they have not been cast in stone by your instructor or institution.
- Most important output: the **model equation** and **coefficient table**.
- Coefficient = predicted # units of change in Y per unit of change in X
- *Good practice: the coefficient values should not have too many digits to the left of the decimal point or too many zeros to the right of it.*

Original car-size variable names and units

- **weight** is measured in **pounds, displacement** is measured in **cubic inches**, and **engine power** is measured in **horsepower**.
- A change of one pound or one cubic inch or one horsepower is an insignificant difference in a car.
- Here is the equation for a simple regression model in which MPG is predicted from weight in this data set:

  Predicted MPG = 46.2 - 0.0076*weight

- *An additional pound of weight is predicted to reduce the distance traveled per gallon by 76 thousands of a mile, which is about 40 feet.*
- It might be better to choose larger units for weight (and the others too).

Renamed and rescaled car-size variables to be used

Weight1000lbs: 1000's of pounds

Displacement100ci: 100's of cubic inches,

Horsepower100: 100's of horsepower

- A one-unit change in any of them is roughly the difference between a large car and a small or medium sized car.
- Liters would also be a good choice for units of displacement.
- The names include the units for clarity.
- Let's also look at other variables….

What is "acceleration"?

- Does a bigger number mean faster?
- No, slower!
- It's the elapsed time in seconds to go from 0 to 60 miles per hour.
- New name: **Seconds0to60**

What is "origin"?

- It's a numeric code: 1 = US, 2 = Europe, 3 = Japan
- Let's add a text variable **Country** with values **1_US**, **2_Europe**, and **3_Japan**.
- It will be used later for creating self-descriptive **dummy variables.**
- They will sort in the same order as the numeric codes.

What is "fuel economy"?

- Is miles-per-gallon the only measure?
- What about the reciprocal, gallons per mile?
- Or with better scaling, gallons-per-100 miles.
- A similar measure, liters-per-100-kilometers is standard in the metric world.
- This is an example of a **nonlinear transformation** of a variable, something that is often needed when fitting linear models.
- New variable: **GallonsPer100Miles = 100/MPG**

New variables and names:

GallonsPer100Miles	MPG	Cylinders	Displacement100ci	Horsepower100	Weight1000lb	Seconds0to60	Year	Origin	Country	Name
5.56	18	8	3.07	1.30	3.504	12.0	70	1	1_US	chevrolet chevelle malibu
6.67	15	8	3.50	1.65	3.693	11.5	70	1	1_US	buick skylark 320
5.56	18	8	3.18	1.50	3.436	11.0	70	1	1_US	plymouth satellite
6.25	16	8	3.04	1.50	3.433	12.0	70	1	1_US	amc rebel sst
5.88	17	8	3.02	1.40	3.449	10.5	70	1	1_US	ford torino
6.67	15	8	4.29	1.98	4.341	10.0	70	1	1_US	ford galaxie 500
7.14	14	8	4.54	2.20	4.354	9.0	70	1	1_US	chevrolet impala
7.14	14	8	4.40	2.15	4.312	8.5	70	1	1_US	plymouth fury iii
7.14	14	8	4.55	2.25	4.425	10.0	70	1	1_US	pontiac catalina
6.67	15	8	3.90	1.90	3.850	8.5	70	1	1_US	amc ambassador dpl
6.67	15	8	3.83	1.70	3.563	10.0	70	1	1_US	dodge challenger se
7.14	14	8	3.40	1.60	3.609	8.0	70	1	1_US	plymouth 'cuda 340
6.67	15	8	4.00	1.50	3.761	9.5	70	1	1_US	chevrolet monte carlo

Missing values?

- Most regression software automatically drops whatever rows have missing values of any of the variables used in a given model.
- Adding a variable could lead to the loss of a large amount of data if it has missing values where the others don't.
- **Look at the sample size reported in each model's output** so that you know whether it varies from one model to another.
- The original data file for this analysis actually contained 408 rows but 16 of them had missing values for one or more variables.
- The file to be used here only contains the 392 rows with complete cases for all variables, so that the sample is the same for all models.

Sorting of the data?

- Rows should be sorted in time order if the variables are time series.
- Even if they aren't, it may be helpful to sort on a particular variable in order to group rows with similar characteristics.
- Patterns you see in graphs of the variables and graphs of the model's predictions and errors may be more informative that way.
  - It's often good to look at plots of actual and predicted values versus row number and errors versus row number even for non-time-series data.
  - Your software ought to make this easy and routine.
- In this data set the rows have been sorted into blocks by year of manufacture in increasing order: 1970 cars, then 1971 cars, etc.
- It will be meaningful to watch for **year-by-year time trends** in graphs of variables and errors versus the row number.