Multiple Regression Analysis

In multiple linear correlation and regression we use additional independent variables (denoted *X1, X2,* … , and so on) that help us better explain or predict the dependent variable (Y). Almost all of the ideas we saw in simple linear correlation and regression extend to this more general situation. However, the additional independent variables do lead to some new considerations. Multiple regression analysis can be used either as a descriptive or as an inferential technique

Multiple Regression and Correlation Analysis

The dependent variable Y is related to independent variables $X_1, X_2, \ldots, X_k$ and the error term e through the linear relationship

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_k X_k + e$$

where $a_0, a_1, \ldots, a_k$ are unknown population parameters.

- There is no linearity among $X_1, \ldots, X_k$ (no collinearity assumption)
- The error term e has normal distribution with expectation ($Ee=0$) (normality)
- Variance of e does not depend on $X_1, \ldots, X_k$ (homoscedasticity)
- In cross-sectional data analysis we infer about the unknown parameters using random sample

$$(Y_1, X_{11}, \ldots, X_{1k}), (Y_2, X_{21}, \ldots, X_{2k}), \ldots, (Y_n, X_{n1}, \ldots, X_{nk})$$

Multiple Regression and Correlation Analysis

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_k X_k + e$$

The parameters are estimated by the least squares method. The least square relationship is written as

$$\hat{Y} = \hat{a}_0 + \hat{a}_1 X_1 + \hat{a}_2 X_2 + \ldots + \hat{a}_k X_k$$

Steps in multivariate regression analysis

1. Preliminary variable selection: if the correlation absolute value between any two continuous explanatory variables exceeds 0.7 skip one of them - the one having smaller correlation with Y
2. Check whether loglinear model gives better correlations
3. Check model's validity verifying the hypotheses.

$$H_0 : a_1 = a_2 = \ldots = a_k = 0, \quad H_1 : some\ a_i \neq 0$$

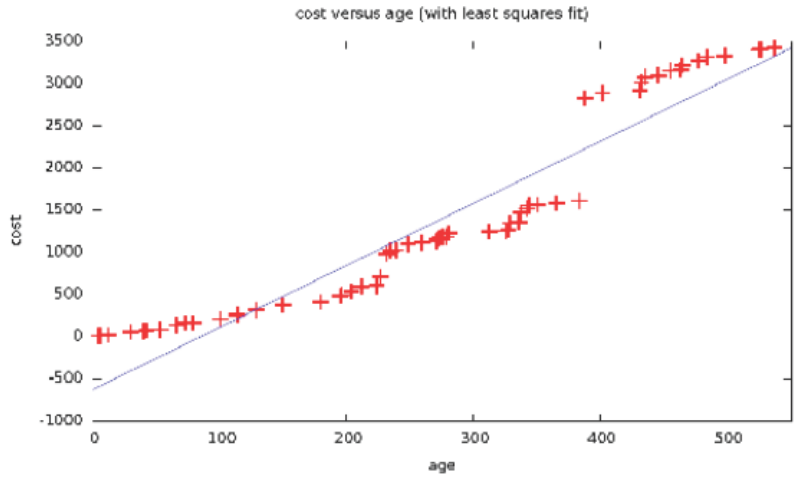
4. Skip non significant independent variables using t-test

$$H_0 : a_i = 0 \quad H_1 : a_i \neq 0$$

5. Redo the least squares for the final model
6. Verify model assumptions (homoscedasticity and normality)
7. Check if there are influential outlying observations - influential outliers.

Laboratory example

DATA3-7: Data for a Toyota station wagon (57 observations)
cost = cumulative repair cost in actual dollars (11 - 3425)
age = age of car in weeks of ownership (Range 5 - 538)
miles = miles driven in thousands (Range 0.8 - 74.4)



Add logarithms of variables (Gretl:Add\logs of selected vars) and compute correlation matrix

```
Correlation Coefficients, using the observations 1 - 57
5% critical value (two-tailed) = 0.2609 for n = 57

      cost        age       miles
    1.0000     0.9488     0.9265  cost
               1.0000     0.9965  age
                          1.0000  miles
```

```
Correlation Coefficients, using the observations 1 - 57
5% critical value (two-tailed) = 0.2609 for n = 57

     l_cost     l_miles      l_age
     1.0000      0.9711     0.9822  l_cost
                 1.0000     0.9971  l_miles
                            1.0000  l_age
```

We choose model with logarithms.

```
Model 1: OLS, using observations 1-57. Dependent variable: l_cost

                  coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------------
  const       -0.898719    0.199045    -4.515    3.38e-05  ***
  l_age        1.41680     0.0365646    38.75     1.30e-041 ***

R-squared          0.964662,   F test   P-value(F) =1.30e-41
White's test for heteroscedasticity - Null hypothesis: heteroscedasticity not
present

with p-value = P(Chi-square(2) > 32.2411) = 9.97572e-008
```

Make diagnostic plots:

Graphs\Fitted actual plot\against l_age

Graphs\Residual plot\against l_age

Conclusion: First observation is an outlier and it should be removed

After outlier removal

```
Model 3: OLS, using observations 2-57 (n = 56)
Dependent variable: l_cost

          coefficient   std. error   t-ratio   p-value
  -------------------------------------------------------
  const     -1.49829     0.186139    -8.049    8.21e-11 ***
  l_age      1.52391     0.0339184   44.93     1.85e-44 ***

R-squared           0.973946   Adjusted R-squared   0.973463
F(1, 54)            2018.580   P-value(F)           1.85e-44


White's test for heteroskedasticity - Null hypothesis:
heteroskedasticity not present
with p-value = P(Chi-square(2) > 9.91542) = 0.00702901
```

Your task

Choose multivaraite data of your interest

and

make a complete data analysis using multivariate regression model