QMM: Exercise Sheet 2 - Multiple Linear Regression

Fabien Baeriswyl, Jérôme Reboulleau, Tom Ruszkiewicz

Exercise 1. For this exercise, use the `McCracken.csv` datafile. The J.J. McCracken Company has authorised its marketing research department to make a study of customers who have been issued a McCracken charge card. The marketing research department hopes to identify the significant variables explaining the variation in the purchase volume. Once these variables are identified, the department intends to try to attract new (profitable) customers, that is those who are predicted to purchase high volumes.

25 customers were selected at random and values for the following variables were recorded in the datafile above:

$y =$ average monthly purchase (in dollars) at McCracken
$x_1 =$ customer age
$x_2 =$ customer family income
$x_3 =$ family size.

a) Create scatterplots illustrating the relations between the average monthly purchases and customer age, customer family income and family size. Use them to assess the correlation between the dependent variable and the independent variables. Then compute the correlation coefficients between all the variables and compare it to your visual assessment.

b) Test whether the correlation coefficients are significantly different from 0, using the significance test for correlation. Report the associated rejection region.

c) Perform a stepwise regression with forward selection using the p-value as criteria for the selection and significance level $\alpha = 5\%$. Report the selected model. What is its coefficient of determination ? Give an interpretation to it. What is the adjusted R_a^2 of your model? Compare it to R^2.

d) Test the overall significance of the selected model in c).

e) Plot the residuals of your model in an histogram. Do they look normally distributed? Compute the mean of the residuals and discuss it.

f) Compute the MSE of the model and discuss it.

g) Plot the residuals of the model against its predicted values. Can you observe a specific pattern?

h) Assuming that the residuals of your selected model in c) are normally distributed, compute a 90% confidence interval for the coefficients associated with family income (the customer family income). Does 0 belong to this interval? If so, what can you say about it?

i) Discuss why the coefficient associated to the age changes if you add the family size variable.

Exercise 2. For this exercise, use the `RealEstate.csv` datafile. A real estate agent wishes to determine the selling price of residences using as predicting factors the size (in square feet) and whether the residence is a condominium or a single-family home. A sample of 20 residences was obtained.

a) Fit a regression to predict the selling price for residences using a model of the form

$$y_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \epsilon_i$$

where $x_{1,i}$ is a variable representing the type of the ith residence and $x_{2,i}$ is the square footage of residence i. Observe, in `R`, how "character" variables are treated in a regression equation. Then, interpret the parameters β_1 and β_2 in this model.

b) Is the estimator b_2 of β_2 significantly different from 0 at a significance level of $\alpha = 5\%$? What about b_1?

c) Fit a model with interaction between the explanatory variables. Interpret the regression coefficients.

d) Your friends lives in a single-family home of 1750 square feet. Using your model in a), what is its predicted price under your model in a)? Under your model in b)?

e) As the type of residence might have no impact on its price, you decide to stop using it, but you fit a new model consisting of a polynomial regression of degree 4 of the square footage instead. Write the equation of the model. Comment on your results.

f) Apply a backward selection method with the AIC as criterion to the model fitted in e). Report the selected model equation. Report the estimated price for your friend's house under this model.

g) Suppose you are given a new explanatory variable for your linear fit from part a), the size of the residence in square meters. Should you add it to your initial model? If not, why?

h) Add the square meters as a new column in your data frame. If you were to fit the same model as in a) but replacing square feet by square meters, what do you expect the regression coefficient and its associated standard error to be? (*Recall that 1 square meter corresponds to 10.7639 square foot.)*