KU LEUVEN

FACULTY OF SCIENCE

Project

Regression Analysis

Daniel Gerardo GIL SANCHEZ
daniel.gilsanchez@student.kuleuven.be

Prof. Pieter Segaert
Prof. Kris Peremans

Academic year 2017-2018

Earn data

In the Panel Study of Income Dynamics done in 1982, different information regard to salary was collected from 358 individuals[1]. The goal of this study is to reveal which are the key factors that determine the wage earned. In the following, a descriptive analysis is performed to identify patterns in the data. Then, different regression models are conducted to establish the factors that present significant differences in the salary received.

On average, the *income* received by the people surveyed is approximately $1,155$ dollars, the *years of experience* is about 23 years, the *weeks worked in the past year* is almost 47 and the *years of education* is around 12. The correlation between these variables is not very high, the larger is 0.44 between *income* and *education*, and the smaller is -0.1 between *income* and *weeks worked.* As it is expected the distribution of the response variable, *income*, is skewed to the right. Regarding the remain variables collected, all dichotomous, the results are shown in Table 1, where it can be seen that the differences in salary between categories are not high.

	Proportion		Mean of Salary		SD of Salary	
	No	Yes	No	Yes	No	Yes
Blue collar occupation	0.48	0.52	1360.31	966.44	634.79	332.74
Manufacturing Industry	0.6	0.4	1140.11	1176.07	565.05	493.55
Resides in the South	0.71	0.29	1190.95	1065.75	537.58	527.69
Resides in a Metropolitan Area	0.35	0.65	992.97	1242.34	412.49	575.87
Married	0.2	0.8	897.17	1217.14	478.25	532.58
Female	0.89	0.11	1206.96	738.1	538.69	284.33
Union contract	0.65	0.35	1177.57	1112.23	638.91	257.79
Black	0.93	0.07	1174.59	888.04	540.94	402.81

Table 1: Descriptive statistics

So, in order to evaluate the significance of the variables collected, an Ordinary Least Squares (OLS) model is fitted. As a result, the overall null hypothesis of the model is rejected with a significance level of 0.05, but only *years of experience*, *blue collar occupation*, *living in a metropolitan area*, *female* and *education* are significant. The residual analysis shows problems with normality and heteroscedasticity (not displayed here). An output expected because of the distribution of the response variable. Therefore, a transformation of the variables is needed.

In the literature is well known that the variable *income* should be analyzed after using the logarithmic transformation, so a new model is performed using the logarithmic on the response variable. As a consequence, the variables mentioned before are significant again and the predictors *contract union* and *black* are now significant. This is an improvement of the model because the transformation used, refines the relationship between the last variables mentioned and the response.

[1] The original dataset have 376 individuals.

Regarding the variables that are not significant and before removing them from the model, a stepwise regression is performed. The output suggests that *living in the south* and *weeks worked in the past year* can be deleted from the model without losing significance. Consequently, a new model is fitted where as a result, it can be seen that *marital status* and working in a *manufacturing industry* are factors that do not present significant changes on *Income*. Hence, before removing both factors from the model, two different models are fitted removing one of them at a time. In this way, one can be sure that after removing one variable, the results do not present considerable changes in the other one. As a result, both variables are removed from the model.

In this perspective, the variables that have significant results are: *years of experience*, *blue collar occupation*, *living in a metropolitan area*, *female*, *contract union*, *black* and *education*. Thus, in order to give precise conclusions, the residuals have to be analyzed. In Figure 1 the different plots of the residuals are shown.

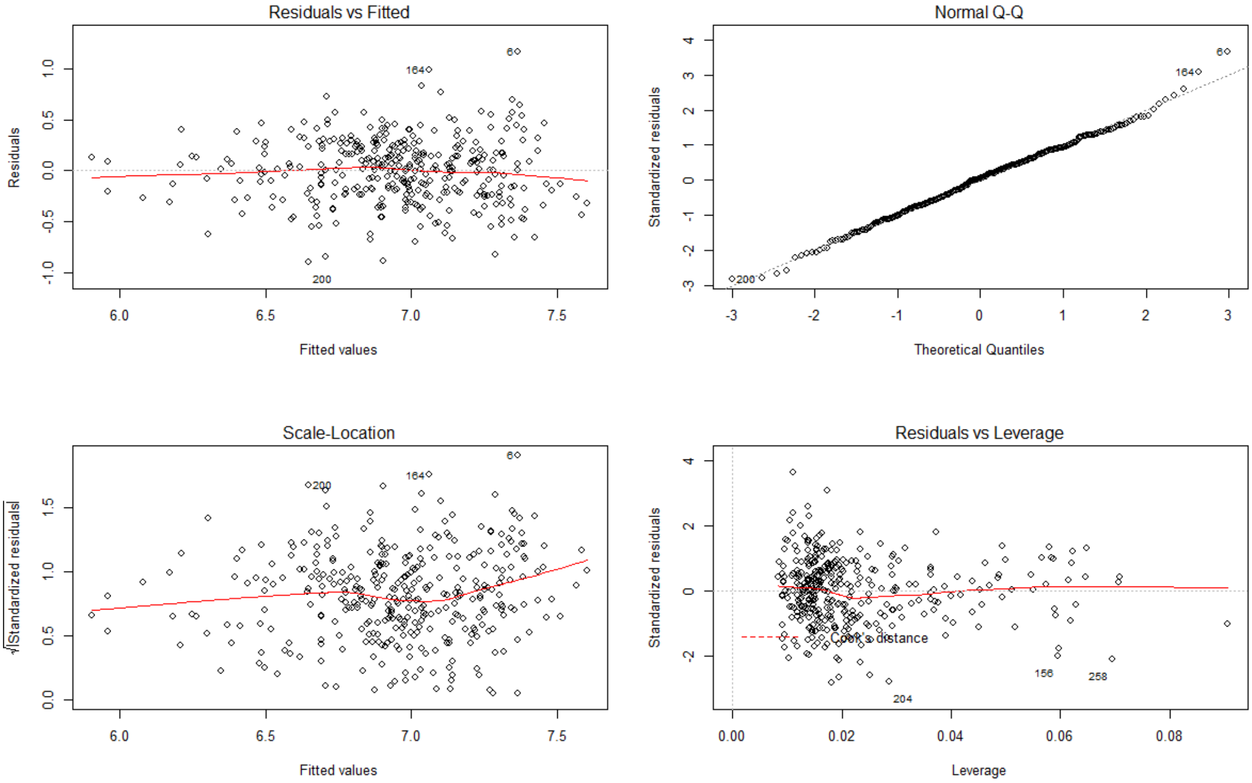


Figure 1: Residuals plots

Figure 1, presents 4 different plots. From the plots on the left, it can be seen that there is a funnel behaviour indicating heteroscedastic residuals. From the upper right plot, it can be seen that the residuals do not present departures from normality. From the lower right plot, it can be seen that there are not outliers in the model.

As a consequence, in order to be completely sure about the inference of the parameters, the model has to be improved in terms of heteroscedasticity. So, since a transformation has already performed on the model, Weighted Least Squares is the right choice to solve the problem. To calculate the weights, the absolute residuals of the model are used as the response variable and the fitted values are taken as independent variable. The fitted values of the latter model are plugged in the last *income* model that presented heteroscedastic residuals, and the estimation is done once again.

As a result, the final model is improved in every way and now it accomplishes all assumptions (normality, homocedasticity and independence). Regarding the effect that each independent variable has on the salary, the conclusions can be completely reliable. In Table 2 the results are displayed.

	Estimation	**Original Scale**	**As percentage**
Intercept	6.04	421.16	
Experience	0.006	1.006	0.63
Blue collar occupation	-0.19	0.83	-17.36
Resides in Metropolitan Area	0.16	1.17	17.46
Female	-0.49	0.61	-38.77
Union contract	0.19	1.21	20.59
Education	0.06	1.06	6.19
Black	-0.13	0.88	-12.32

Table 2: Estimation

From Table 2[2], it can be seen that *women* are paid less than men under the same conditions. In fact, the final estimation leads to conclude that *women* are receiving on average about 39% less money than men, if the other variables are constant. It can also be seen that as an average, *black* people receive approximately 12% less salary that non-black people, if the other variables remain constant. In addition, workers indeed benefit from *union contracts*; the estimation leads to conclude that on average they earn around 20% more salary that people that do not set their salary by a *union contract*. Besides, it is important to notice that *marital status* does not affect the *salary*; this variable is not significant in any of the models fitted.

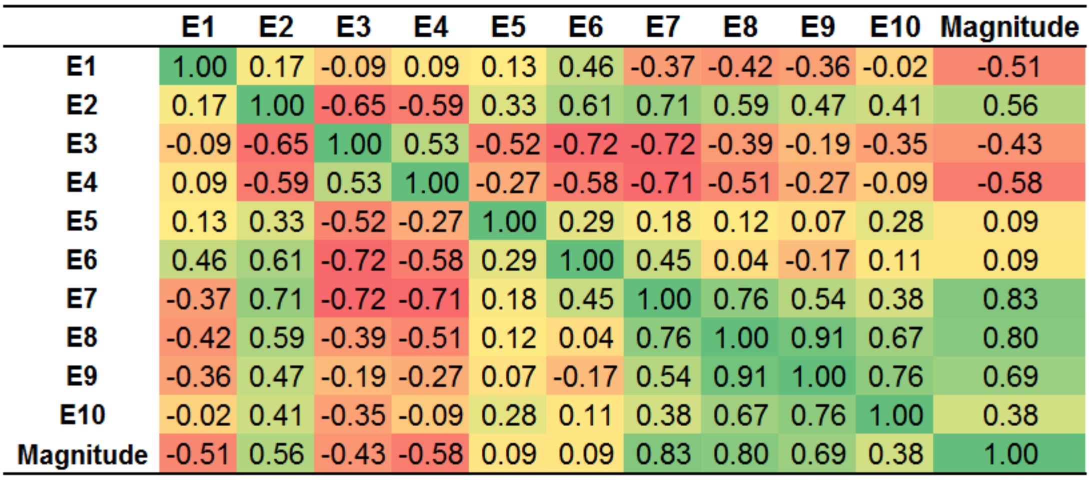
To sum up, several models are fitted to reveal the effect that some factors have on the wage earned. Because of the distribution of the response variable, *income*, a log-transformation is used to improve the linear relationship. The selection of the significant factors is performed through stepwise regression, then two models are used to assess the significance of two specific factors (*marital status* and *manufacturing industry*). After having the significant factors, residuals plots suggest a problem of heteroscedasticty that is solved by Weighted Least Squares. The final model leads to conclude that as an average, *women* earn 39% less salary than men. *Black* people earn 12% less salary than non-black people and people who set their wage by a *union contract* received 20% more salary than people who do not do it. It is important to mention that these conclusions are valid if the other variables remain constant. Finally, it is concluded that *marital status* does not affect *income*.

[2] To express each parameter as percentage, the estimation is transformed to its original scale using the exponential, then an unit is subtracted from it and multiplied by 100. For instance, the percentage of *female* is $-38.77 = (exp(-0.49) - 1) * 100$

Supernova data

A physicist has access to a dataset that contains measurements from 38 Type Ia supernovas[3]. For each supernova, the absolute *magnitude* and the *spectral energy* in 10 different frequency bands are recorded and then standardized to have mean zero and variance one[4]. The purpose of this study is to construct a prediction rule based on the frequency vector taking into account that for distant supernovas it is impossible to obtain precise measurements. In the following, a descriptive analysis is performed to identify patterns in the data. Then, different regression models are conducted to construct different prediction rules, to finally conclude which one is better.

Considering that the variables are standardized, the interpretation of univariate analysis does not give much information about the characteristics present in the dataset. Moreover, a bivariate analysis gives more information because it leads to see how the behaviour of a *spectral energy* measured in one frequency band is related to the behaviour in other frequency band. In Table 3 the correlation matrix is presented. It can be seen a strong relationship between the *magnitude* and some of the measurements. It can also be seen that the linear relationship of *spectral energy* between different frequency bands are high enough to predict a multicollinearity problem in the models.



	E1	E2	E3	E4	E5	E6	E7	E8	E9	E10	Magnitude
E1	1.00	0.17	-0.09	0.09	0.13	0.46	-0.37	-0.42	-0.36	-0.02	-0.51
E2	0.17	1.00	-0.65	-0.59	0.33	0.61	0.71	0.59	0.47	0.41	0.56
E3	-0.09	-0.65	1.00	0.53	-0.52	-0.72	-0.72	-0.39	-0.19	-0.35	-0.43
E4	0.09	-0.59	0.53	1.00	-0.27	-0.58	-0.71	-0.51	-0.27	-0.09	-0.58
E5	0.13	0.33	-0.52	-0.27	1.00	0.29	0.18	0.12	0.07	0.28	0.09
E6	0.46	0.61	-0.72	-0.58	0.29	1.00	0.45	0.04	-0.17	0.11	0.09
E7	-0.37	0.71	-0.72	-0.71	0.18	0.45	1.00	0.76	0.54	0.38	0.83
E8	-0.42	0.59	-0.39	-0.51	0.12	0.04	0.76	1.00	0.91	0.67	0.80
E9	-0.36	0.47	-0.19	-0.27	0.07	-0.17	0.54	0.91	1.00	0.76	0.69
E10	-0.02	0.41	-0.35	-0.09	0.28	0.11	0.38	0.67	0.76	1.00	0.38
Magnitude	-0.51	0.56	-0.43	-0.58	0.09	0.09	0.83	0.80	0.69	0.38	1.00

Table 3: Correlation Matrix

Given the strong relationship between *magnitude* and the measurements of *spectral energy*, an Ordinary Least Squares (OLS) regression is performed to establish a possible prediction rule for new observations. As a result, the overall null hypothesis of the model[5] is rejected with a significance level of 0.05, but no variable is significant, a contradictory result.

This incongruity can be explained by a multicollinearity problem, as it is seen in Table 3. Therefore, the condition number (CN)[6] and the variance inflation factor (VIF) are analyzed. As a result, the VIF for variables *E6*, *E7*, *E8* and *E9* are larger than

[3] The original dataset have information of 39 supernovas.

[4] *Magnitude* is only centered, its mean is also zero but its variance is not one

[5] $H_0: \ \beta_1 = \beta_2 = \ldots = \beta_{10} = 0$

[6] This calculation is based on the eigenvalues of the correlation matrix

10 and the CN is approximately 16. This basically means that the OLS estimation is compromised because of multicollinearity problem.

To solve this issue, several remedies can be taken into account. Principal Component Regression and Ridge Regression are chosen to re-estimate the model, because even though both methodologies have biased estimators the variability is smaller, this is known as *bias-variance trade-off*.

In Principal Component Regression, the data should be first reduced from a p-dimensional space to a k-dimensional space ($k << p$), then a new model is fitted with the projection of each observation in the new space. In this case, for the ten measures in different frequency bands, three principal components are retained based on the total amount of variance explained and the number of eigenvalues larger than one.

A new model is then fitted with these principal components. The overall hypothesis is statistically significant as well as each component. In regard to multicollinearity problem, it can be easily seen that is solved because by the definition the principal components are orthogonal and so the correlation matrix is the identity. Thus, the VIF and CN are one for each variable.

On the other hand, the estimation of Ridge Regression is obtained by introducing a constant to the diagonal of the correlation matrix. The larger the constant, the more biased is the estimation and so the betas are closer to zero. There are several ways to choose the constant, in this case it is selected such that the multicollinearity problem is solved. This model is then fitted using a constant equal to 0.14.

Now it is necessary to see which model constructs the better prediction rule based on the frequency vector. Two measures are considered to make this comparison: The Mean Squared Error (MSE) to compare how each model fits the data and the PRESS$_p$ criterion to estimate the mean squared error of the prediction (see Table 4).

	PCR	Ridge
MSE	1.13	1.03
PRESS$_p$	36.69	42.43

Table 4: Comparison

Given that the model proposed is to predict new observations, the decision is based mostly on the PRESS$_p$ criterion. It works under the idea of leave-one-out cross validation, where a single observation is taken out to fit the model and then it is predicted, storing its residual. Then, the PRESS$_p$ statistic is calculated as the sum of these residuals[7]. From Table 4 can be seen that the Ridge regression has lower MSE, which means that fits the data better compared to the other. However, it can also be seen that the Principal Component Regression has the lower PRESS$_p$, which means that

[7] It is important to mention that even though there are different ways to measure the predictive power of a model, like k-fold cross validation, in this case only PRESS$_p$ is used because of the sample size, is too small to use different criteria

predicts better new observations. Thus, the Principal Component model constructs the best prediction rule based on the frequency vector.

It is important to mention that the predictions should only be done if the new measures are in the range of the data. So, given that these measures were done in supernovas close to the earth, the prediction of the magnitude of new supernovas should only be done if they are close enough as the data collected for the model. The predictions will not be reliable if this is not taken into account.

To conclude, three models are fitted to construct prediction rules based on the frequency vector. Because of the high correlation between the *spectral energy* measures in different frequency bands, multicollinearity is a problem in the estimation of the model. Principal Components Regression and Ridge Regression are used to solve the problem. Then a comparison between the models is done based on their power to fit the data and to predict new observations. MSE and PRESS$_p$ are the statistics used to make the comparison. The final results lead to conclude that even though Ridge regression model fits the data best, the Principal Component model constructs the best decision rule to predict new observations. These predictions are reliable only if the new observations are close enough to the earth as it was in the data used to construct the decision rule.

BigMac data

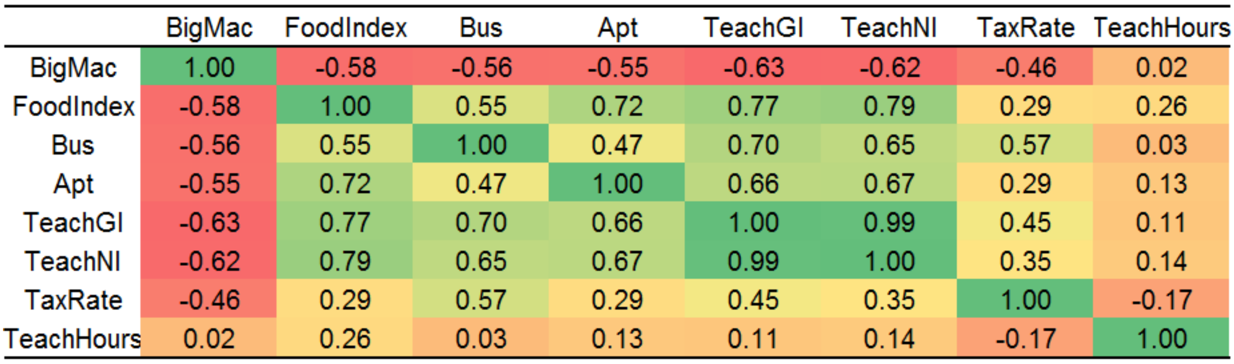
The Union Bank of Switzerland published a report entitled Prices and Earnings Around the Globe, where a price comparison is done between 66 cities[8]. In the version of 2003, some variables were collected such as minutes of labor to purchase a Big Mac hamburger, food price index, the cost for a one-way 10 km ticket, monthly rent of a three-room apartment, primary teachers' gross income, primary teachers' net income, tax rate paid by a primary teacher and primary teachers hours of work per week. The goal of this study is to explain the behaviour of the minutes of labor to purchase a *BigMac* hamburger in terms of the variables collected. In the following, a descriptive analysis is performed to identify patterns in the data. Then, different regression models are conducted to explain the response variable *BigMac*.

On average, the number of minutes of labor to purchase a *BigMac* is approximately 38, the *food index* is about 62, the monthly rent of a three-room *apartment* is 718 US dollars, the *primary teachers' net income* is around $15,850$ US dollars and the primary teachers' *hours of work* per week is about 37. The dispersion in *primary teachers' gross income* and *primary teachers' net income* is similar, whereas the *monthly rent apartment* is approximately 471 (see Table 5 for more detail). In the correlation matrix can be seen that the response variable is highly negative correlated with the independent variables except the number of hours of work per week. It can also be noticed that some of the independent variables are highly correlated, so a stepwise regression is a good idea to select an efficient set of explanatory variables (see Table 6).

[8] The original dataset have 69 cities.

	BigMac	FoodIndex	Bus	Apt	TeachGI	TeachNI	TaxRate	TeachHours
Mean	37.95	62.02	1.06	718.03	21.39	15.85	21.6	36.56
SD	31.97	24.97	0.81	471.07	19.59	14.37	10.51	7.32

Table 5: Summary statistics



	BigMac	FoodIndex	Bus	Apt	TeachGI	TeachNI	TaxRate	TeachHours
BigMac	1.00	-0.58	-0.56	-0.55	-0.63	-0.62	-0.46	0.02
FoodIndex	-0.58	1.00	0.55	0.72	0.77	0.79	0.29	0.26
Bus	-0.56	0.55	1.00	0.47	0.70	0.65	0.57	0.03
Apt	-0.55	0.72	0.47	1.00	0.66	0.67	0.29	0.13
TeachGI	-0.63	0.77	0.70	0.66	1.00	0.99	0.45	0.11
TeachNI	-0.62	0.79	0.65	0.67	0.99	1.00	0.35	0.14
TaxRate	-0.46	0.29	0.57	0.29	0.45	0.35	1.00	-0.17
TeachHours	0.02	0.26	0.03	0.13	0.11	0.14	-0.17	1.00

Table 6: Correlation Matrix

Given the strong relationship between *BigMac* and the independent variables, an Ordinary Least Squares (OLS) regression is performed to evaluate the significance of each explanatory variable. As a result, the overall null hypothesis of the model is rejected with a significance level of 0.05 and the R^2 is about 52%, but only one variable is significant (*tax rate paid by primary teachers*), a contradictory result.

This incongruity can be explained by different factors: a multicollinearity problem, as it is seen in Table 6 or a strong nonlinear relationship between the response variables and the independent variables. In Figure 6 in the appendix the scatter plots between the variable *BigMac* and each of the independent variables are shown. It can be easily seen that the correlation is nonlinear with almost all independent variables.

To solve this issue, several remedies can be taken into account. Transform the independent variables or the response variable. In this case, given that the nonlinear relationship is mostly affected by the response variable, a transformation is considered for it. Thus, according to Tukey and Mosteller's 'bulge rule' and given the shape of the relationship, the logarithm and the square root of the variable *BigMac* are conducted.

From the model where the square root transformation is used, the results do not seem to improve compared with the previous model, the overall null hypothesis is significant and the R^2 is 65%, but only a single variable is significant, *tax rate paid by primary teachers*. In addition, the residuals plots of this model show that they are not normal nor homoscesdastic (not displayed here). So no improvement of the model is achieved with this transformation.

On the other hand, from the model where the logarithmic transformation is performed, the results are better compared with the previous models, the overall null hypothesis is significant and the total of variance explained is about 74%. Regard to the significance of the independent variables, only the *tax rate paid by primary teachers* is significant at an alpha of 0.05, but the *monthly rent of a three-room apartment* and the *primary teachers' net income* are significant at an alpha of 0.1. This, of course, does not mean that the significance level is changed for the next models, on the contrary, it says that a combination of this transformation with a stepwise regression can improve the results. Additionally, the residuals plots of this model show that there are not departures from normality and the heteroscedasticity problem found before is improved (not displayed here).

As a consequence, the logarithmic transformation is chosen to proceed with the analysis. Now, given that the percentage of explained variance is acceptable and that only few variables seems to be significant, stepwise regression is conducted to select an efficient set of independent variables without losing the explanatory power that the model already has.

As a result of the stepwise regression, the *hours of work per week of the primary teachers* and the *food index* can be deleted from the model without losing significance. Consequently, a new model is fitted where, as a result, the *cost for a one-way 10 km ticket* and the *gross income of primary teachers* are variables that do not explain significantly the response variable. Hence, before removing both variables from the model, two different models are fitted removing one of them at a time. In this way, one can be sure that after removing one variable, the results do not present considerable changes. Therefore, both variables are removed from the model.

In this perspective, the variables that are significant in the model are: *monthly rent of a three-room apartment*, *tax rate paid by primary teachers* and *net income of primary teachers*. Thus, in order to give precise conclusions, the residuals have to be analyzed. In Figure 2 the different plots of the residuals are shown.

From the plots on the left, it can be seen that there is a little funnel behaviour indicating possible heteroscedastic residuals. From the upper right plot, it can be seen that the residuals do not present important departures from normality. From the lower right plot, it can be seen that there are not outliers in the model. Therefore, given that visually there is still doubt about the good behaviour of the residuals, different tests are used in order to be completely sure about these conclusions. Shapiro-Wilk test is used to assess normality of the residuals; it suggests that the residuals indeed follow a normal distribution. Breusch-Pagan test is used to assess the homocedasticity of the residuals; it suggests that the variance is indeed constant. Hence, it can be concluded that this model is optimal to explain the response variable *BigMac* and so the conclusions can be completely reliable. In table 7 the results are displayed.

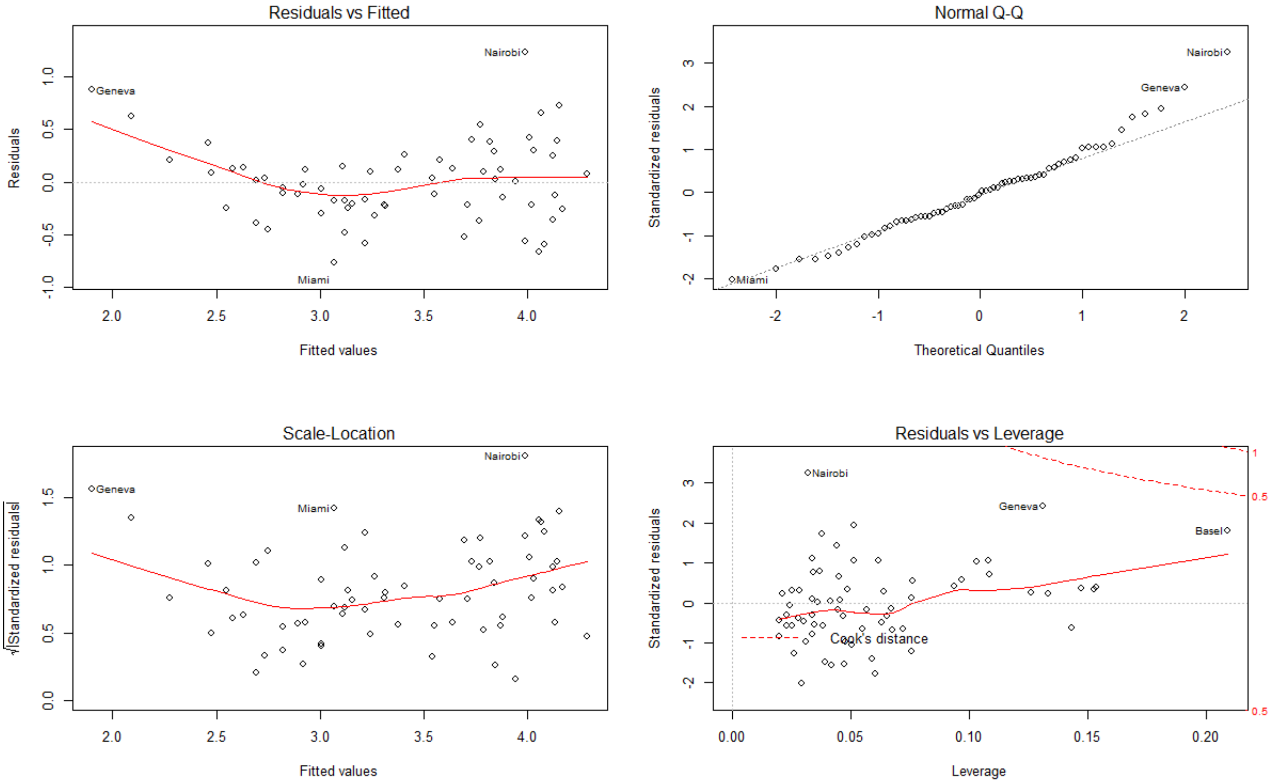


Figure 2: Residuals plots

From Table 7[9], it can be seen that an increment of 1 dollar in the *monthly rent of a three-room apartment*, decrease on average between 0.006% and 0.07% the number of minutes needed to purchase a *BigMac*. It can also be seen that an increment of 1000 dollars in the *net income of primary teachers*, decrease on average between 1.78% and 3.55% the number of minutes needed to purchase a *BigMac*. Finally, it can be concluded that an increment of one unit in the *tax rate paid by primary teachers* leads to an average decrease in the number of minutes to purchase a *BigMac* between 0.62% and 2.54%. It is important to mention that these conclusions are valid if the other variables remain constant.

	Estimation	**Original scale**	**Confidence interval**	**Confidence interval as percentage**
Intercept	4.394	80.997	[63.873, 102.713]	
Apt	-0.0003	1	[0.999, 1.000]	[-0.061, -0.007]
TeachNI	-0.027	0.973	[0.964, 0.982]	[-3.559, -1.784]
Tax Rate	-0.016	0.984	[0.975, 0.994]	[-2.542, -0.629]

Table 7: Estimation

In summary, several models are fitted to explain the response variable *BigMac*. Because of the nonlinear relationship between the response variable and the independent

[9] The results are already transformed in the original scale. To express each confidence interval as a percentage, the estimation is transformed to its original scale using the exponential, then a unit is subtracted from it and multiplied by 100.

variables, square root and logarithmic transformation are used to improve the linear relationship, being the latter the one that performed best. The selection of the significant variables is performed through stepwise regression, then two models are used to assess the significance of two specific explanatory variables (*cost for a one-way 10 km ticket* and *gross income of primary teachers*). After having the significant variables in the model, residuals plots, Shapiro-Wilks test and Breusch-Pagan test suggest that there are not problems of normality nor heteroscedasticity. The final model leads to conclude that only the *monthly rent of a three-room apartment*, the *net income of primary teachers* and the *tax rate paid by primary teachers* are significant to explain the response variable.

Fish data

In 1998 in Lake Erie, the age in years and the length in millimeters of a sample of 75 perch was collected[10]. The age is determined by counting the number of rings on a scale of the fish. The goal of this study is to propose different models that describe the relationship between length and age.

The *length* presents an average of 140.3 millimeters with a standard deviation of 31.29 and the *age* have a mean of 3.72 years and a standard deviation of 0.99. In Figure 3 can be seen that there is a positive relationship between the variables but this relationship does not look linear. This plot suggests that the older the fish is, the bigger it is until it reaches a plateau after 4 years.

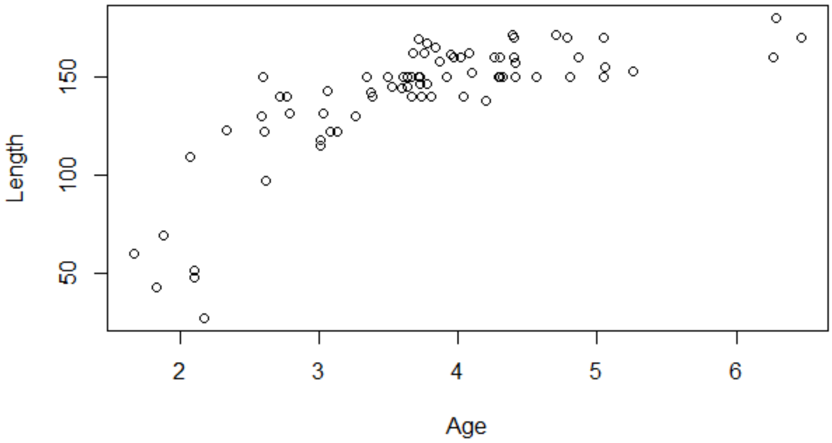


Figure 3: Scatter plot: Age vs Length

Although the relation is not linear, an Ordinary Least Squares (OLS) model is fitted setting *length* as a response variable and *age* as independent. The overall null hypothesis of the model is rejected with a significance level of 0.05 as well as *age*. The coefficient of determination is 0.5669 which means that about 57% of the variance of the data is explained by the model. These measures of adequacy of the model are quite good, however the residual analysis indicates the opposite (not displayed). In the first place, the residuals present large deviations from a normal distribution. Secondly

[10] The original dataset have measurements of 78 perch.

the residuals seem to be correlated with the fitted values. Finally, some outliers are identified from the standardized residuals. As a consequence and given the shape of the relationship between the variables, a transformation of the independent variable is considered.

The variable age is transformed using the logarithm. The overall null hypothesis of the model is rejected with a significance level of 0.05 as well as the independent variable. The coefficient of determination is 0.6878 which means that the variance of the data explained by the model has increased approximately 11%. In regard to residuals analysis, the normality of the model is improved as well as the presence of outliers. However, it seems that there is still a correlation between the residuals and the fitted values, see Figure 4.

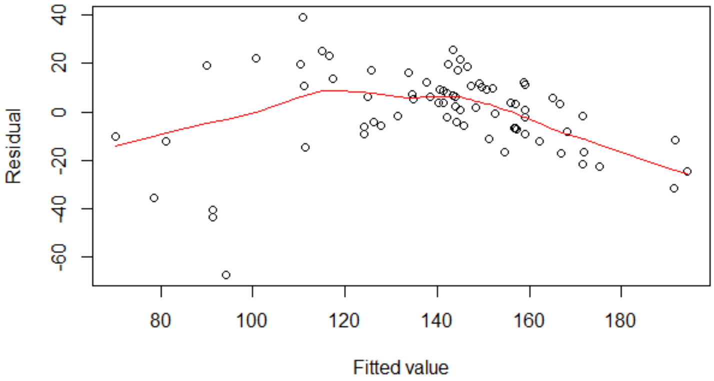


Figure 4: Residual plot - Log(Age) model

Consequently, a nonlinear regression is considered. An exponential regression model is chosen because of the shape of the relationship between the variables. To obtain starting values for the numerical procedure, the model is linearized as follows:

$$Length = \beta_0 e^{\beta_1 Age} \quad => \quad log(Length) = \beta_0' + \beta_1 Age$$

As a result, $\beta_0' = 4.05$ and $\beta_1 = 0.22$, and transforming to the original model $\beta_0 = 57.57$. These initial estimates are used as starting values for the nonlinear estimation method.

The overall null hypothesis of the model is rejected with a significance level of 0.05 as well as each component. In regard to residuals analysis, the normality of the model does not improve compared with the previous models. In addition, it seems that there is still correlation between the residuals and the fitted values. Therefore, other expression of an exponential regression is used.

In the literature the growth curves are used to model the growth of an individual over time. The model structure is as follows:

$$Length = \beta_0 + \beta_1 e^{\beta_2 Age}$$

In this case to obtain the initial estimates through the linearization of the model is not as straightforward as it was in the previous one. Thus, taking into account the shape of the relationship and how the $Betas$ are usually interpreted[11], the initial estimates are:

[11] This interpretation is taken from the course notes

- β_0: If $\beta_2 < 0$, β_0 is the maximum length. In this case β_0 is set to 180.
- β_1: $\beta_0 + \beta_1$ the length at time 0, thus $\beta_1 < 0$. In this case the length at time 0 is considered to be 27, so β_1 is set to -153
- β_2: Is assumed to be negative and close to zero, so β_2 is set to -0.25

As a result, the overall null hypothesis of the model is rejected with a significance level of 0.05 as well as each component. In regard to residuals analysis, the normality of the model has improved a little bit as well as the presence of outliers. Additionally, the correlation between the residuals and the fitted values is gone.

So, in order to decide which model fits the data best the Mean Squared Error (MSE) is used to compare the four models explained above as well as the plot of each fit, see Table 8 and Figure 5). From the table, it is easy to see that the last model is the one that has the smallest MSE. Actually, in the plot is very clear how the last model improved the fit of the data. For this reason, the model chosen is the growth curve.

	OLS	Log(Age)	Exponential 1	Exponential 2
MSE	429.84	309.9	505.49	212.95

Table 8: MSE comparison

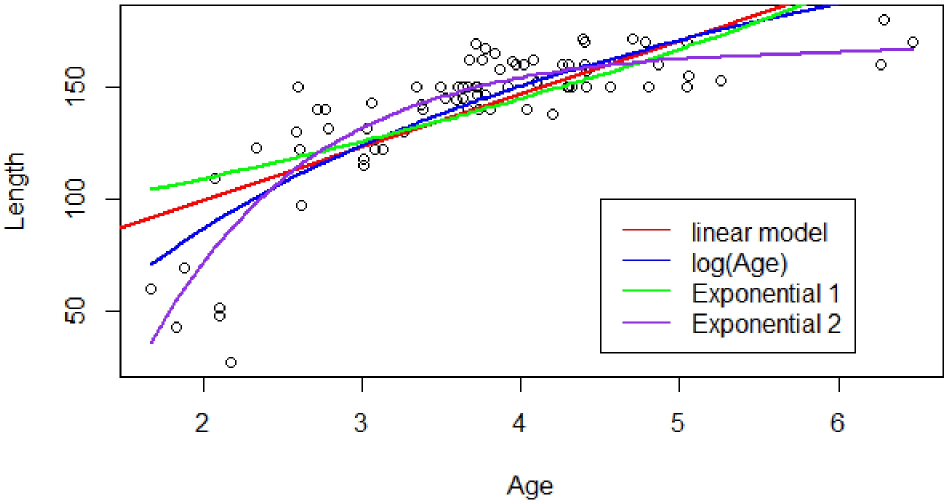


Figure 5: Fit comparison

To conclude, several models are fitted to assess the relationship between the length and the age of 75 perch from Lake Erie. Because of the shape of the correlation between the variables, the linearity of the model is compromised. The logarithm of the independent variable as well as two different nonlinear models are used to improved the fit. The growth curve model is the one that fits the data best using the mean squared error as selection criteria of the final model.

Appendix

BigMac - Scatter plots

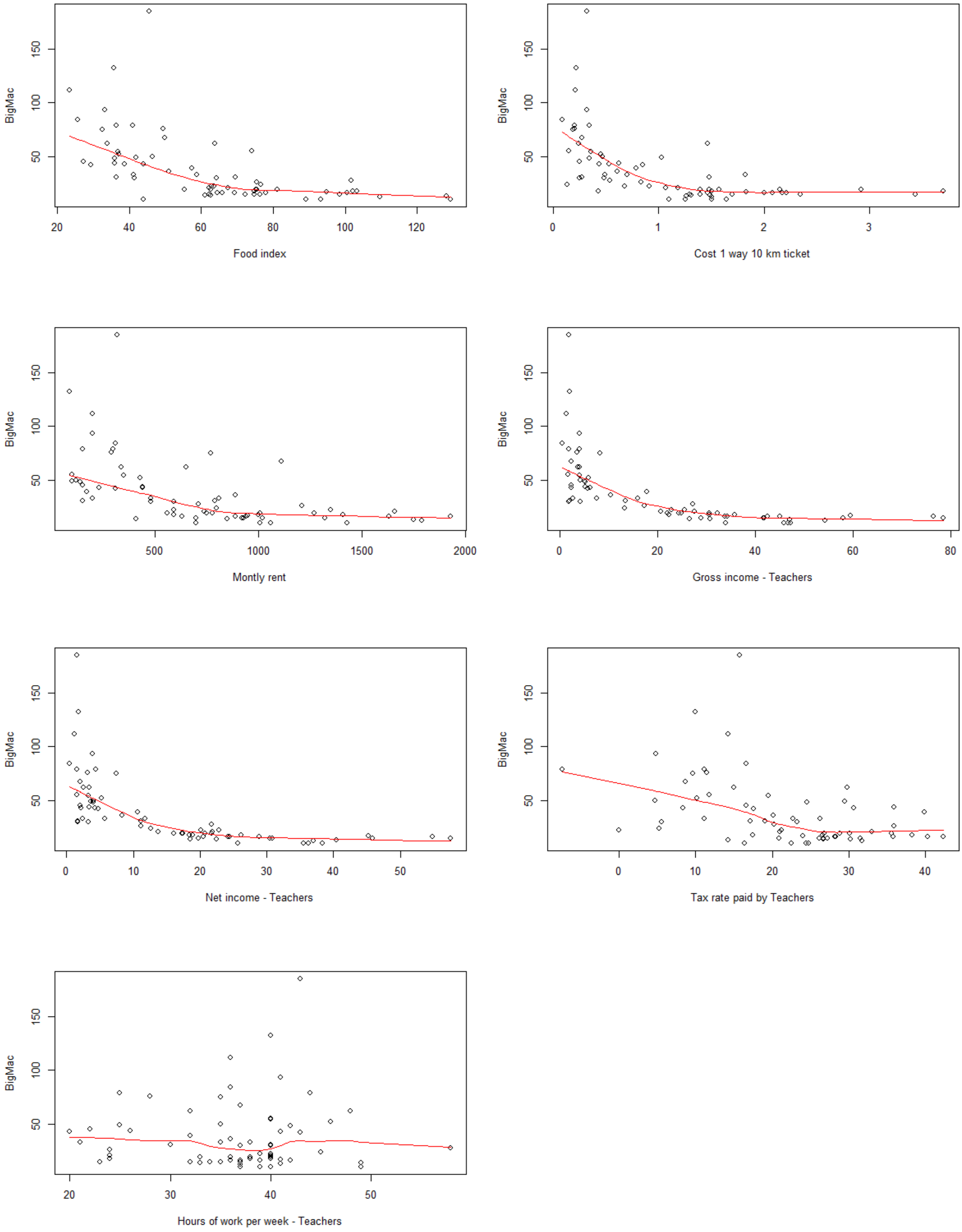


Figure 6: Scatter plots

Earn data - R Code

```
###  Clear workspace
rm(list=ls())

###   Set directory
setwd("...")

###   Load packages
library(MASS)   # Stepwise regression and standardized residuals
library(olsrr)  # Normality
library(lmtest) # Breush Pagan Test for Heteroscedasticity
library(car)    # NCV Test for Heteroscedasticity
library(dplyr)  # Summarise function

###   Read data
earn_full = read.table("earn.txt",header=T)
str(earn_full)
head(earn_full)

###   Remove 5% of the observations at random
set.seed(689432)
rownumbers = sample(1:nrow(earn_full),size=0.05*nrow(earn_full))
earn = earn_full[-rownumbers,]
earn[rownumbers,]
n = dim(earn)[1]

#   Transform dummies to factor variables
earn$OCC = as.factor(earn$OCC)
earn$IND = as.factor(earn$IND)
earn$SOUTH = as.factor(earn$SOUTH)
earn$SMSA = as.factor(earn$SMSA)
earn$MS = as.factor(earn$MS)
earn$FEM = as.factor(earn$FEM)
earn$UNION = as.factor(earn$UNION)
earn$BLK = as.factor(earn$BLK)

###   Exploratory Analysis
summary(earn)
cor(earn[,c(1,2,10,12)])
plot(density(earn$WAGE))
boxplot(earn$WAGE~earn$OCC,main="OCC")
boxplot(earn$WAGE~earn$IND,main="IND")
boxplot(earn$WAGE~earn$SOUTH,main="SOUTH")
boxplot(earn$WAGE~earn$SMSA,main="SMSA")
boxplot(earn$WAGE~earn$MS,main="MS")
boxplot(earn$WAGE~earn$FEM,main="FEM")
boxplot(earn$WAGE~earn$UNION,main="UNION")
boxplot(earn$WAGE~earn$BLK,main="BLK")
plot(earn$WAGE~earn$EXP)
plot(earn$WAGE~earn$WKS)
plot(earn$WAGE~earn$ED)

##  summary by factor
summarise(group_by(earn,OCC),freq=n(),
    prop=freq/sum(n),salary=mean(WAGE),sd=sd(WAGE))
summarise(group_by(earn,IND),freq=n(),
    prop=freq/sum(n),salary=mean(WAGE),sd=sd(WAGE))
summarise(group_by(earn,SOUTH),freq=n(),
    prop=freq/sum(n),salary=mean(WAGE),sd=sd(WAGE))
summarise(group_by(earn,SMSA),freq=n(),
    prop=freq/sum(n),salary=mean(WAGE),sd=sd(WAGE))
summarise(group_by(earn,MS),freq=n(),
    prop=freq/sum(n),salary=mean(WAGE),sd=sd(WAGE))
summarise(group_by(earn,FEM),freq=n(),
    prop=freq/sum(n),salary=mean(WAGE),sd=sd(WAGE))
summarise(group_by(earn,UNION),freq=n(),
    prop=freq/sum(n),salary=mean(WAGE),sd=sd(WAGE))
summarise(group_by(earn,BLK),freq=n(),
    prop=freq/sum(n),salary=mean(WAGE),sd=sd(WAGE))

###   Model 1
fit1 = lm(WAGE~EXP+WKS+OCC+IND+SOUTH+SMSA+MS+FEM+UNION+ED+BLK,data=earn)
fit1_sum = summary(fit1);fit1_sum

#   Residuals: Model 1
fit1_res <- residuals(fit1)
fit1_fitted <- fitted.values(fit1)
fit1_stdres <- fit1_res/fit1_sum$sigma
sum(fit1_res)

#   Check model assumptions: Model 1
par(mfrow = c(2,2))
qqnorm(fit1_stdres, main = "")
qqline(fit1_stdres)
plot(fit1_res, xlab = "Index", ylab = "Residual")
plot(fit1_fitted, fit1_res, xlab = "Fitted value", ylab = "Residual")
lines(lowess(fit1_res ~ fit1_fitted), col = "red")
plot(fit1_stdres, xlab = "Index", ylab = "Standardized residual", ylim = c(-4,4))
abline(h = -2.5, lty = 2)
abline(h = 2.5, lty = 2)

###   Model 2: Transform the response variable (log)
fit2 = lm(log(WAGE)~EXP+WKS+OCC+IND+SOUTH+SMSA+MS+FEM+UNION+ED+BLK,
    data=earn)
fit2_sum = summary(fit2);fit2_sum

#   Residual: Model 2
fit2_res <- residuals(fit2)
fit2_fitted <- fitted.values(fit2)
fit2_stdres <- fit2_res/fit2_sum$sigma
sum(fit2_res)

#   Check model assumptions: Model 2
par(mfrow = c(2,2))
qqnorm(fit2_stdres, main = "")
qqline(fit2_stdres)
plot(fit2_res, xlab = "Index", ylab = "Residual")
plot(fit2_fitted, fit2_res, xlab = "Fitted value", ylab = "Residual")
lines(lowess(fit2_res ~ fit2_fitted), col = "red")
```

```
plot(fit2_stdres, xlab = "Index", ylab = "Standardized residual", ylim = c(-4,4))
abline(h = -2.5, lty = 2)
abline(h = 2.5, lty = 2)

#   Check assumptions
fit2_stdres = stdres(fit2)
shapiro.test(fit2_stdres)
ols_norm_test(fit2_stdres)
#   Residuals are normal

# Heteroscedasticity
bptest(fit2)
ncvTest(fit2)

# Independence
dwtest(fit2)

#   Before trying to fix heteroscedasticity
#   remove variables in the model
step <- stepAIC(fit2, direction="both")
step$anova # display results

#   Fit the model with the log transformation
fit3 = lm(log(WAGE)~EXP+OCC+IND+SMSA+MS+FEM+UNION+ED+BLK,data=earn)
fit3_sum = summary(fit3);fit3_sum

#   Remove one variable at a time
fit4 = lm(log(WAGE)~EXP+OCC+SMSA+MS+FEM+UNION+ED+BLK,data=earn)
fit4_sum = summary(fit4);fit4_sum

fit5 = lm(log(WAGE)~EXP+OCC+IND+SMSA+FEM+UNION+ED+BLK,data=earn)
fit5_sum = summary(fit5);fit5_sum

#   Final model
fit6 = lm(log(WAGE)~EXP+OCC+SMSA+FEM+UNION+ED+BLK,data=earn)
fit6_sum = summary(fit6);fit6_sum

#   Check assumptions
par(mfrow = c(2,2))
plot(fit6)
fit6_stdres = stdres(fit6)
shapiro.test(fit6_stdres)
ols_norm_test(fit6_stdres)

# Heteroscedasticity
bptest(fit6)
ncvTest(fit6)

# Independence
dwtest(fit6)

# Weighted least squares
fit6_res = residuals(fit6)
fit6_fitted = fitted.values(fit6)
stdev = lm(abs(fit6_res)~fit6_fitted)
```

```
weight = 1/(stdev$fitted.values^2)
fit7 = lm(log(WAGE)~EXP+OCC+SMSA+FEM+UNION+ED+BLK,data=earn,weights = weight)
fit7_sum = summary(fit7);fit7_sum  #  Not big difference wiht OLS
par(mfrow = c(1,1))
plot(fit7$fitted.values,residuals(fit7)*sqrt(weight))

fit7_coef = coefficients(fit7)
fit7_coef_orig = exp(fit7_coef)
(fit7_coef_orig-1)*100

#--> Are women paid less than men in a comparable situation?
(fit7_coef_orig[5]-1)*100

#-->   Is wage affected by marital status?
# The variable is not significant in any model

#-->   Do blacks earn more than non-blacks?
(fit7_coef_orig[8]-1)*100

#-->   Do workers benefit from union contracts?
(fit7_coef_orig[6]-1)*100
```

Supernova data - R Code

```
#   Clear workspace
rm(list=ls())

#   Set directory
setwd("...")

#   Load packages
library(MASS)
library(qpcR)
library(ridge)

#   Read data
supernova_full = read.table("supernova(1).txt",header=T)
str(supernova_full)
head(supernova_full)

#   Remove 5% of the observations at random
set.seed(689432)
rownumbers = sample(1:nrow(supernova_full),size=0.05*nrow(supernova_full))
supernova = supernova_full[-rownumbers,]
supernova[rownumbers,]
n = dim(supernova)[1]
p = dim(supernova)[2]

#   Exploratory Analysis
summary(supernova)
cor(supernova)
pairs(supernova, panel = function(x,y) {points(x,y); lines(lowess(x,y), col =
 ↪  "red")})
```

```
boxplot(supernova)

###   Model 1
fit1 = lm(Magnitude~E1+E2+E3+E4+E5+E6+E7+E8+E9+E10-1,data=supernova)
fit1_sum = summary(fit1);fit1_sum
fit1$coefficients

#   VIF calculation
cor_matrix = cor(supernova[,-11])
vif = diag(solve(cor_matrix));vif
mean(vif)

#   Eigenvalues calculations
cor_eigen = eigen(cor_matrix)$values
#   Compute the condition number
cn = sqrt(cor_eigen[1]/cor_eigen);cn

# PCR
pca = princomp(supernova[,-11])
summary(pca)
plot(pca$sdev,type="l")
#   Add scores columns to dataset
supernova_2 = cbind(supernova,pca$scores[,1:3])

#   Model 2: PCR
fit2_pca = lm(Magnitude ~ Comp.1 + Comp.2 + Comp.3 - 1, data = supernova_2)
fit2_pca_sum = summary(fit2_pca);fit2_pca_sum
fit2_pca_coef = colSums(coefficients(fit2_pca) * t(pca$loadings[,1:3]))
fit2_pca_coef

#   Model 3: Ridge Regression
#   Choose the constant with lm.ridge
#   Criteria: GCV
plot(lm.ridge(Magnitude~E1+E2+E3+E4+E5+E6+E7+E8+E9+E10-1,
    data=supernova,lambda = seq(0,10,0.001)))
select(lm.ridge(Magnitude~E1+E2+E3+E4+E5+E6+E7+E8+E9+E10-1,
    data=supernova,lambda = seq(0,10,0.001)))
fit3_ridge = lm.ridge(Magnitude~E1+E2+E3+E4+E5+E6+E7+E8+E9+E10-1,
    data=supernova,lambda = 9.419)
fit3_ridge

#   VIF calculation
cor_matrix_ridge = cor_matrix+diag(9.419/n,10,10)
new_cor = solve(cor_matrix_ridge) %*% cor_matrix %*% solve(cor_matrix_ridge)
vif_ridge = diag(new_cor);vif_ridge
mean(vif_ridge)

#   Model 4: Ridge regression VIF criteria
#   Choose the new constant
#   Criteria: VIF
constant <-NULL
for(i in seq(0, 1, 0.01)){
  vif = diag(solve(cor_matrix+i*diag(10))%*%cor_matrix%*%
        solve(cor_matrix+i*diag(10)))
  vif_c = c(i,vif)
  constant = rbind(constant,vif_c)
}
constant
c=0.14

fit4_ridge = lm.ridge(Magnitude~E1+E2+E3+E4+E5+E6+E7+E8+E9+E10-1,
         data=supernova, lambda = n*c)
fit4_ridge
fit4_ridge_coef = fit4_ridge$coef

summary(linearRidge(Magnitude~E1+E2+E3+E4+E5+E6+E7+E8+E9+E10-1,
                    data=supernova, lambda = c))
#   Fitted values for ridge
fit4_ridge_fitted = as.matrix(supernova[,-11]) %*% fit4_ridge_coef

# Comparison coefficients
data.frame("LS" = fit1$coefficients, "PCR" = fit2_pca_coef, "RR" =
 ↪  fit4_ridge_coef)


#   Model Comparison: MSE
#   OLS
fit1_MSE = (sum((fit1$residuals)^2))/(n-p)
fit1_MSE

#   PCR
fit2_pca_MSE = (sum((fit2_pca$residuals)^2))/(n-p)
fit2_pca_MSE

#   Ridge
fit4_ridge_MSE = (sum((supernova$Magnitude - fit4_ridge_fitted)^2))/(n-p)
fit4_ridge_MSE

#   PRESS
#   OLS
res=rep(0,38)
for(i in 1:n){
  data=supernova[-i,]
  y_i = supernova[i,]
  model = lm(Magnitude~E1+E2+E3+E4+E5+E6+E7+E8+E9+E10-1,data=data)
  pre = predict(model,y_i[-11])
  res[i] = as.numeric(y_i[11] - pre)
}
fit1_press = sum(res^2);fit1_press
#   To validate
PRESS(fit1)

#   PCA
res=rep(0,38)
for(i in 1:n){
  data=supernova_2[-i,]
  y_i = supernova_2[i,]
  model = lm(Magnitude~Comp.1 + Comp.2 + Comp.3 -1,data=data)
  pre = predict(model,y_i[12:14])
  res[i] = as.numeric(y_i[11] - pre)
```

```
}
fit2_pca_press = sum(res^2);fit2_pca_press
#   To validate
PRESS(fit2_pca)


#   PRESS for Ridge
res=rep(0,38)
for(i in 1:n){
  data=supernova[-i,]
  y_i = supernova[i,]
  model = lm.ridge(Magnitude~E1+E2+E3+E4+E5+E6+E7+E8+E9+E10-1,
                   data=data,lambda = n*c)
  model_coef = model$coef
  pre = as.matrix(y_i[,-11]) %*% model_coef
  #pre = predict(model,y_i[-11])
  res[i] = as.numeric(y_i[11] - pre)
}
fit4_ridge_press = sum(res^2);fit4_ridge_press
```

BigMac data - R Code

```
#   Clear workspace
rm(list=ls())

#   Set directory
setwd("...")

#   Load packages
library(leaps)   # All possible regression
library(MASS)    # Stepwise regression and standardized residuals
library(olsrr)   # Normality
library(lmtest)  # Breush Pagan Test for Heteroscedasticity
library(car)     # NCV Test for Heteroscedasticity

#   Read data
bigmac_full = read.table("BigMac.txt",header=T)
str(bigmac_full)
head(bigmac_full)

#   Remove 5% of the observations at random
set.seed(689432)
rownumbers = sample(1:nrow(bigmac_full),size=0.05*nrow(bigmac_full))
bigmac = bigmac_full[-rownumbers,]
bigmac_full[rownumbers,]  # Athens, Kuala_Lumpur and Singapore are out of my
 ↪  analysis
rownames(bigmac) = bigmac$City
bigmac = bigmac[,-1]
n = dim(bigmac)[1]
p = dim(bigmac)[2]

#   Exploratory Analysis
summary(bigmac)
```

```
apply(bigmac,2,mean)
apply(bigmac,2,sd)
cor(bigmac)
pairs(bigmac, panel = function(x,y) {points(x,y); lines(lowess(x,y), col =
 ↪  "red")})
boxplot(bigmac)
boxplot(bigmac[,-4])
plot(density(bigmac$BigMac))

par(mfrow = c(2,2))
plot(bigmac$FoodIndex,bigmac$BigMac,xlab = "Food
 ↪  index",ylab="BigMac");lines(lowess(bigmac$BigMac ~ bigmac$FoodIndex), col =
 ↪  "red")
plot(bigmac$Bus,bigmac$BigMac,xlab = "Cost 1 way 10 km
 ↪  ticket",ylab="BigMac");lines(lowess(bigmac$BigMac ~ bigmac$Bus), col = "red")
plot(bigmac$Apt,bigmac$BigMac,xlab = "Montly
 ↪  rent",ylab="BigMac");lines(lowess(bigmac$BigMac ~ bigmac$Apt), col = "red")
plot(bigmac$TeachGI,bigmac$BigMac,xlab = "Gross income -
 ↪  Teachers",ylab="BigMac");lines(lowess(bigmac$BigMac ~ bigmac$TeachGI), col =
 ↪  "red")
plot(bigmac$TeachNI,bigmac$BigMac,xlab = "Net income -
 ↪  Teachers",ylab="BigMac");lines(lowess(bigmac$BigMac ~ bigmac$TeachNI), col =
 ↪  "red")
plot(bigmac$TaxRate,bigmac$BigMac,xlab = "Tax rate paid by
 ↪  Teachers",ylab="BigMac");lines(lowess(bigmac$BigMac ~ bigmac$TaxRate), col =
 ↪  "red")
plot(bigmac$TeachHours,bigmac$BigMac,xlab = "Hours of work per week -
 ↪  Teachers",ylab="BigMac");lines(lowess(bigmac$BigMac ~ bigmac$TeachHours), col
 ↪  = "red")

#   Model with Raw data
fit = lm(BigMac~FoodIndex+Bus+Apt+TeachGI+TeachNI+TaxRate+TeachHours,
        data=bigmac)
fit_sum = summary(fit);fit_sum

#   Check assumptions
par(mfrow = c(2,2))
plot(fit)
fit_stdres = stdres(fit)
shapiro.test(fit_stdres)

# Nothing is significant and the correlations between the response and the
 ↪  independent variables are considerable, so there is a problem of
 ↪  multicollinearity
# Residuals are not normal

#   Non linear relationship so transformation is necessary
#   Log transformation
fit_log = lm(log(BigMac)~FoodIndex+Bus+Apt+TeachGI+TeachNI+TaxRate+
        TeachHours,data=bigmac)
fit_log_sum = summary(fit_log);fit_log_sum

#   Check assumptions
par(mfrow = c(2,2))
plot(fit_log)
```

```
fit_log_stdres = stdres(fit_log)
shapiro.test(fit_log_stdres)
#   Residuals are normal
#   Homoscedasticity

#   Sqrt transformation
fit_sqrt = lm(sqrt(BigMac)~FoodIndex+Bus+Apt+TeachGI+TeachNI+TaxRate+
        TeachHours,data=bigmac)
fit_sqrt_sum = summary(fit_sqrt);fit_sqrt_sum

#   Check assumptions
par(mfrow = c(2,2))
plot(fit_sqrt)
fit_sqrt_stdres = stdres(fit_sqrt)
shapiro.test(fit_sqrt_stdres)
#   The residuals are not normal

#   Box Cox Transformation
par(mfrow = c(1,1))
box_cox <- boxcox(BigMac~FoodIndex+Bus+Apt+TeachGI+TeachNI+TaxRate+TeachHours,
                  data=bigmac, plotit = TRUE)
lambda <- box_cox$x[which(box_cox$y == max(box_cox$y))]
fit_box <- lm(((BigMac)^lambda - 1)/lambda ~
 ↪  FoodIndex+Bus+Apt+TeachGI+TeachNI+TaxRate+TeachHours,
              data=bigmac)
fit_box_sum = summary(fit_box);fit_box_sum
#   Check assumptions
par(mfrow = c(2,2))
plot(fit_box)
fit_box_stdres = stdres(fit_box)
shapiro.test(fit_box_stdres)
# Residuals are normal
#   Homoscedasticity

#   Log-transformation is chosen
#   Stepwise selection
step <- stepAIC(fit_log, direction="both")
step$anova # display results

fit_log_step = lm(log(BigMac)~Bus+Apt+TeachGI+TeachNI+TaxRate,data=bigmac)
fit_log_step_sum = summary(fit_log_step);fit_log_step_sum

#   Check assumptions
par(mfrow = c(2,2))
plot(fit_log_step)
fit_log_step_stdres = stdres(fit_log_step)
shapiro.test(fit_log_step_stdres)
#   Residulas are normal
#   Homoscedasticity

#   Remove one and then the other...not significant
fit_log_wbus = lm(log(BigMac)~Bus+Apt+TeachNI+TaxRate,data=bigmac)
fit_log_wbus_sum = summary(fit_log_wbus);fit_log_wbus_sum
fit_log_wGI = lm(log(BigMac)~Apt+TeachGI+TeachNI+TaxRate,data=bigmac)
fit_log_wGI_sum = summary(fit_log_wGI);fit_log_wGI_sum
```

```
# Model after checking significances
fit_log_fin = lm(log(BigMac)~Apt+TeachNI+TaxRate,data=bigmac)
fit_log_fin_sum = summary(fit_log_fin);fit_log_fin_sum
#   Check assumptions
par(mfrow = c(2,2))
plot(fit_log_fin)
fit_log_fin_stdres = stdres(fit_log_fin)
# Normality
shapiro.test(fit_log_fin_stdres)
ols_norm_test(fit_log_fin)

# Heteroscedasticity
bptest(fit_log_fin)
ncvTest(fit_log_fin)

# Independence
dwtest(fit_log_fin)

#   To interpret
fit_log_fin_sum
CI = confint(fit_log_fin)
beta = exp(fit_log_fin_sum$coefficients[,1])
beta_CI = exp(CI)
(beta_CI-1)*100
```

Fish data - R Code

```
#   Clear workspace
rm(list=ls())

#   Set directory
setwd("...")

#   Load packages
library(MASS)   # standardized residuals
library(olsrr)  # Normality
library(lmtest) # Breush Pagan Test for Heteroscedasticity
library(car)    # NCV Test for Heteroscedasticity
library(nlstools)    # residuals and normality in non linear

#   Read data
fish_full = read.table("fish.txt",header=T)
str(fish_full)
head(fish_full)

#   Remove 5% of the observations at random
set.seed(689432)
rownumbers = sample(1:nrow(fish_full),size=0.05*nrow(fish_full))
fish = fish_full[-rownumbers,]
fish[rownumbers,]   # Observation 3, 33 and 68 are out of my analysis
n = dim(fish)[1]
p = dim(fish)[2]
```

```
#   Exploratory Analysis
summary(fish)
apply(fish,2,mean)
apply(fish,2,sd)
cor(fish)
pairs(fish, panel = function(x,y) {points(x,y); lines(lowess(x,y), col = "red")})
boxplot(fish$age)
boxplot(fish$length)
plot(density(fish$age))
plot(density(fish$length))
plot(fish$age,fish$length,xlab = "Age",ylab="Length")

#   Model with Raw data
par(mfrow = c(1,1))
lfit = lm(length~age,data=fish)
lfit_sum = summary(lfit);lfit_sum
plot(fish$age,fish$length,xlab = "Age",ylab="Length")
abline(lfit, col = "red")

#   Check assumptions
par(mfrow = c(2,2))
plot(lfit)
lfit_stdres = stdres(lfit)
shapiro.test(lfit_stdres)
#   Residuals not normal

#   Model log of age
lfit_loga = lm(length~log(age),data=fish)
lfit_loga_sum = summary(lfit_loga);lfit_loga_sum
par(mfrow = c(1,1))
plot(fish$age,fish$length,xlab = "Age",ylab="Length")
abline(lfit, col = "red")
lines(sort(fish$age), lfit_loga$fitted.values[order(fish$age)], col = "blue")
legend(5, 80, c("linear model", "log(Age)"), lty = 1, col = c("red", "blue"))

#   Check assumptions
par(mfrow = c(2,2))
plot(lfit_loga)
lfit_loga_stdres = stdres(lfit_loga)
shapiro.test(lfit_loga_stdres)
ols_norm_test(lfit_loga)
#   Residuals are not normal

#   Correlation between residuals and fitted values
par(mfrow = c(1,1))
plot(lfit_loga$fitted.values, lfit_loga$residuals, xlab = "Fitted value", ylab =
 ↪  "Residual")
lines(lowess(lfit_loga$residuals ~ lfit_loga$fitted.values), col = "red")


# Non linear model
lfit_log_coef = coefficients(lm(log(length)~age,data=fish))
lfit_non = nls(length~A*exp(B*age),data=fish,
    start=list(A=exp(lfit_log_coef[1]),B=lfit_log_coef[2]),
    trace=T)
lfit_non_sum = summary(lfit_non,correlation = T);lfit_non_sum
A = lfit_non_sum$parameters[1]
B = lfit_non_sum$parameters[2]
xx = seq(min(fish$age),max(fish$age),length=n)
yy = A*exp(B*sort(fish$age))

# Plot regression curve
par(mfrow = c(1,1))
plot(fish$age,fish$length,xlab = "Age",ylab="Length")
abline(lfit, col = "red")
lines(sort(fish$age), lfit_loga$fitted.values[order(fish$age)], col = "blue")
lines(sort(fish$age),yy,col = "green")
legend(4.5, 80, c("linear model", "log(Age)","non-linear model"),
       lty = 1, col = c("red", "blue","green"))

#   Growth curve
lfit_non2 = nls(length~A+B*exp(C*age),data=fish,
               start=list(A=180,B=-153,C=-0.25),trace=T)
lfit_non2_sum = summary(lfit_non2,correlation = T);lfit_non2_sum
A2 = lfit_non2_sum$parameters[1]
B2 = lfit_non2_sum$parameters[2]
C2 = lfit_non2_sum$parameters[3]
xx2 = seq(min(fish$age),max(fish$age),length=n)
yy2 = A2+B2*exp(C2*sort(fish$age))

# Plot regression curve
par(mfrow = c(1,1))
plot(fish$age,fish$length,xlab = "Age",ylab="Length")
abline(lfit, col = "red",lwd=2)
lines(sort(fish$age), lfit_loga$fitted.values[order(fish$age)], col =
 ↪  "blue",lwd=2)
lines(sort(fish$age),yy,col = "green",lwd=2)
lines(sort(fish$age),yy2,col = "purple2",lwd=2)
legend(4.5, 100, c("linear model", "log(Age)","Exponential 1","Exponential 2"),
       lty = 1, col = c("red", "blue","green","purple2"))

#   Comparison: MSE
# Linear model
lfit_MSE = (sum((lfit$residuals)^2))/(n-p)
lfit_MSE

# Log(Age)
lfit_loga_MSE = (sum((lfit_loga$residuals)^2))/(n-p)
lfit_loga_MSE

# Non linear 1
lfit_non_MSE = (sum(residuals(lfit_non)^2))/(n-p)
lfit_non_MSE

# Non linear 2
lfit_non2_MSE = (sum(residuals(lfit_non2)^2))/(n-p)
lfit_non2_MSE
```