Multiple Regression

Tanja

20 February, 2021

In the previous section we have learn about *simple linear regression* as a very simple tool for predicting quantitative measured response variable. Its simple application, easy interpretation and predictive capabilities make it one of the most popular approach for supervised learning. Developing good understanding of the linear regression serves as a solid base for learning how to use and adopt other more sophisticating modelling procedures in the context of machine learning.

**Additive Model**

The basic idea in regression modelling is that there is a phenomenon of interest whose behaviour we seek to account for or explain. For example a company may be interested in why its sales figures have followed a particular pattern; or Government officials may be interested in explaining the behaviour of unemployment statistics.

Let the phenomenon of interest be denoted by ; as discussed in the previous sections, is also known as the response variable, or the dependent variable. The fundamental nature of is that it displays variability and it is this inherent variability that the data analyst seeks to explain by using regression modelling. Intuitively, if we can account for a large amount of the movements in then in some sense this is *good*. Conversely, if we put forward an argument which explains only a very small amount of the behaviour of the phenomenon of interest then in some sense this is *bad*.

In trying to explain, or account for the behaviour of we often build a regression model. As a first step in building such a model we specify a set of variables, called the explanatory or predictor variables, that we believe to be important in explaining the behaviour of, or the variability in . Specifying this set of variables is in effect the first step in developing as the data analysts our viewpoint, our model, our theory. As such we are entitled to ask where such a viewpoint comes from - for example, it may be the logical outcome of some theoretical argument or it may be a replication of a previous study.

{{% notice warning %}} 💡 We do well to remember that this statement of a model is a viewpoint, a belief, a theory and need not be correct, i.e. true. {{% /notice %}}

Indeed, it is the very essence of the regression model validation to judge if a viewpoint, a belief, a theory is in any sense acceptable. So, when developing a regression model all we argue is: , which represents a belief that the response variable depends upon a set of explanatory variables .

Specifying a set of variables is not in itself a complete model. We have to be prepared to indicate and argue more precisely how the explanatory variables are supposed to relate to : that is, we must specify the functional form that links the set of variables to . The simplest relationship is a linear relationship or a straight line. As we have already learnt for a single explanatory variable, , this is given by:

* (Within this structure it is important that and can be interpreted clearly.)

and for a set of -variables this is given by:

* (Within this structure it is important that can be interpreted clearly.)

Those linear models specified so far are **exact** or **deterministic**. Such structures do not represent statistical problems. The very nature of statistical data is that there is inherent variability in that there is some part of the responses variable that we cannot explain. Thus there is a random element whose behaviour is by nature and definition unpredictable. We know that a random component is going to impact on , but we do not know the size or sign of such random happenings.

This random/stochastic part of the model is captured as follows:

* , for a single explanatory variable model, and
* , for the general explanatory variable model

where is also known as the error term.

{{% notice warning %}} 💡 These expressions need to be fully understood and the role and significance of the random/stochastic element has to be fully recognised. {{% /notice %}}

In order to complete the specification of the statistical regression model some assumption has to be made about the underlying process that produces . The standard approach is to adopt the following distribution structure:

, with the error term from a normal distribution with a mean of , and a variance of .

Knowing this, we should now be in a position to understand and identify the structure of the standard **multiple linear regression model**.

**Prior Knowledge**

As a statement of belief, the specified model just suggests a list of important explanatory variables within a linear structure. We need to be willing to take an extra conceptual step by specifying our expectation on how the individual explanatory variables are connected to the response variable. We should make assertions about:

1. the expected sign of each slope (or marginal response coefficient) for each explanatory variable?
2. the expected size of each slope?

Again knowledge of these two potential tricky areas are connected to theoretical arguments and are often the results of previous similar studies. Such knowledge is sometimes called **prior knowledge**, or **prior views**. In order to give a viewpoint some evidential or empirical support is required:

* a sample of data has to be collected and
* the model has to be put to a scrutany of the rigour diagnostic checking

That is, the proposed theory has to be tested against the available evidence.

{{% notice note %}} It needs to be fully understood that the model developed so far is no more than a belief. {{% /notice %}}

We may believe passionately that is caused by the structure put forward but we do not know the structure in that we do not know the values of the parameters. These values are held to exist and have particular true numerical values which we do not know what they are. In an attempt to give our theory/model/viewpoint empirical content we have to collect a sample of pertinent data. On the basis of the information contained in this particular data we have to estimate the values and draw valid conclusions about the nature of the true but unknown underlying structure which is claimed to be responsible for the behaviour of .

**Building a Model**

The sample, of size , evidence is seen as coming from the following structure:

Selecting a sample by definition means that we have imperfect information, that is we do not have information on the whole population or we do not see the whole picture. We only see a possibly disjoint part of the whole picture and thus taking a sample gives obvious difficulties in making sense of what we can see. Taking a sample introduces an interpretation problem of trying to say something sensible about the whole picture, the true model from a part of the picture.

The data handling process involves two major steps, namely:

1. model estimation and
2. model validation

**Model Estimation**

As we have discussed, the model developed so far is a viewpoint. It represents our view as to the true underlying structure of the world. We believe that this structure exists but unfortunately, we do not know the values of the parameters. These parameters are held to exist but their true values are unknown. Hence there is a need to develop estimates of the true but unknown parameters. These estimates are based on a sample of data and it is to be hoped that they are in some sense acceptable, i.e. good.

The most common estimating approach is the **principle of Ordinary Least Squares** (**OLS**). As we have already learnt, this idea is easiest to understand when there is only one explanatory variable and the problem can be easily depicted with a scatter diagram onto which we can superimpose the line of the best fit. This process of finding the line of best fit has been programmed in R through the lm() function. Problems involving more than one explanatory variable are handled in a similar fashion. As part of the model fitting process R produces a great deal of output which should be used to judge the validity, or usefulness, of the fitted model.

**Model Validation**

On deriving coefficient estimates from a sample of data we end up with a fitted regression model. We must now take this estimated model and ask a series of questions to decide whether or not our estimated model is good or bad. That is, we have to subject the fitted model to a set of tests designed to check the validity of the model, which is in effect a test of our viewpoint.

**Test a)**: Does the fitted model make sense?

This type of testing is based on the *prior knowledge*. We need to be in a position to judge if the parameter estimates are meaningful. This comes in two guises:

1. Do the estimated coefficients have the correct **sign**?
2. Do the estimated coefficients have the correct **size**?

This test assesses if the *marginal response coefficients* display both the correct signs and sizes. To assist the analyst with these issues theoretical reasoning and information gained from past studies should be used. For example, if one was relating sales to advertising then a positive connection would be expected and a simple plot of sales against advertising would reveal information concerning the direction and strength of any sample relationship. Similarly, in a study of sales against price then a negative relationship would be expected. The issue of “correct size” is normally more troublesome. Possibly some clue as to size can be drawn from similar, past studies and this can be compared with the current estimated values.

By way of addressing this test the modeller may adopt the following approach:

1. Fit a regression model for each of the explanatory variables one at a time. Check each one for Test a).
2. Fit a regression model to all of the explanatory variables together. Check each explanatory variable for Test a).
3. Check if Test a) from 1. and 2. are consistent. If *Yes* then great? 😅; if *No* then…? 😟

**Test b)** : Overall is the model a good fit?

The coefficient of determination, , is a single number that measures the extent to which the set of explanatory variables can explain, or account for, the variability in . That is, how well does the set of explanatory variables explain the behaviour of the phenomenon we are trying to understand. It is looking at the set of explanatory variables as a whole and is making no judgement about the contribution or importance of any individual explanatory variable. By construction is constrained to lie in the following range:

Intuitively the closer is to then the better is the model in the sense that the set of explanatory variables can explain a lot of the variability in . Conversely, a value of close to implies a weak, i.e. poor model in that the set of explanatory variables can only account for a limited amount of the behaviour of . But what about non-extreme values of ?

The adequacy of can be judged both formally and informally.

The formal test is given by the result:

* (that is, the set of explanatory variables are *insignificant*, or in other words: useless)
* (that is, at least one explanatory variable is *significant*, or in other words: important)

and the appropriate test statistic is such that , obtained using a data set of size for a fitted model with number of explanatory variables.

The mechanics of an **F-test** are as follows:

* Two vital pieces of information are required for an F-test: and .
* is displayed within the R’s standard output. however is a figure derived from degrees of freedom elements and a designated level of significance. For hypothesis testing the significance () level is commonly used. The degrees of freedom elements are
  1. *Regression*: number of explanatory variables in the model
  2. *Error*: number of observations minus the number of estimated coefficients in the fitted regression model
* The all important **decision making rule** is: if is less than then the null hypothesis is accepted; if is larger than then the alternate hypothesis is accepted:
* if
* if

This formal test involves a rather weak alternative hypothesis, which says only that is significantly bigger than . If is accepted we will have to make a judgement, without the aid of any further formal test, about the usefulness of and hence the model being studied.

can also be used in comparing competing models. If two or more models have been put forward to explain the same response variable then a reasonable rule is to rank the explanatory power of the models in terms of their values. Thus the model with the highest value would be the best model.

The information gathered form the individual regressions carried out in *Test a)* can be used to give an initial judgement and ranking of the relative importance of the various explanatory variables.

However this rule should be used carefully. It is a valid rule when the competing models have the same number of explanatory variables. It is not valid when comparing models that have different numbers of explanatory variables. It is intuitively obvious that a model with more explanatory variables will have a better chance of explaining the variable than a model with fewer variables. Thus the highest rule is biased in favour of those models with more explanatory variables even when some of these explanatory variables are not very useful. In an attempt to redress the imbalance when comparing models with different numbers of explanatory variables a different version of called has been developed as follows:

💡 Note that mathematically adopted way of writing is .

By examining it can be seen that when goes up it is possible for the value of to fall. Thus this statistic gives a better way of comparing models with different values for . As before the rule for comparing models is straightforward: **choose the model with the highest** .

**Test c) t-tests**: Individually are the explanatory variables important?

Using to judge the goodness of the set of explanatory variables does not tell us anything about the importance of any one single explanatory variable. Just because a set of variables is important does not necessarily mean that each individual variable is contributing towards explaining the behaviour of . What is needed is a test than enables us to check the validity of each variable one at a time. Such a test procedure is available as follows:

* (explanatory variable is not important)
* (explanatory variable has a positive influence) or
  + (explanatory variable has a negative influence) or
  + (explanatory variable has an influence)

The appropriate for any particular variable is crucially connected with the prior view as to the nature of the connection between any one proposed explanatory variable and the response variable.

The appropriate test statistic involves a **t-test** based on degrees of freedom.

The critical region, denoted by , is crucially dependent upon the nature of the alternate hypothesis as put forward by the data analyst. is generated from the R output. is a combination of the Degrees of Freedom from Error and the level of test significance. is the benchmark helping to determine whether one accepts or rejects the *null* Hypothesis.

If lies in the critical region then the *alternative* hypothesis is accepted and the modeller accepts that the explanatory variable does have an influence on the behaviour of . If this is not the case and lies in the area, we will accept that the explanatory variable is useless and that it should be eliminated from the model.

The hypothesis testing methodology typically adopts the level of significance and can be illustrated as follows:

Once all the variables have been individually analysed a further regression command should be given on all remaining acceptable explanatory variables in order to ascertain the best fitted model.

Some of the decision making outcomes form the series of **t-test** may be uncomfortable in terms of the information form other tests and we may have to experiment with various competing fitted models using subtle combinations of prior views, , values, and , outcomes before selecting a best model or a set of equally good models.

There are many other, more sophisticated, tests that can be used to judge the validity, i.e. usefulness of a fitted regression model. Before we get familiar with them and in order to appreciate their practicality and effectiveness we will conduct an analysis for fitting multiple regression model based on the procedure explained above.

**-Value Approach for Hypothesis Testing**

The -value is a probability of obtaining a value of the test statistic or a more extreme value of the test statistic assuming that the null hypothesis is true. Thus, the -value approach involves determining “likely” or “unlikely” by determining the probability, assuming the null hypothesis were true of observing a more extreme test statistic in the direction of the alternative hypothesis than the one observed.

* if the -value accept (“unlikely”)
* if the -value reject in favour of (“likely”)

As such, in the context of the inference about the regression parameters, the **-values** help determine whether the relationships that you observe in your sample also exist in the larger population. The -value for each independent variable tests the null hypothesis that the variable has no effect on the dependent variable. In other words, there is insufficient evidence to conclude that there is effect at the population level.

**Case Study**

**carData::Salaries**: The 2008-09 nine-month academic salary for Assistant Professors, Associate Professors and Professors in a college in the U.S. The data were collected as part of the on-going effort of the college’s administration to monitor salary differences between male and female faculty members.

**Format:** A data frame with 397 observations on the following 6 variables:

* rank: a factor with levels:
  + AssocProf
  + AsstProf
  + Prof
* discipline: a factor with levels:
  + A (“theoretical” departments) or
  + B (“applied” departments)
* yrs.since.phd: years since PhD
* yrs.service: years of service
* sex: a factor with levels
  + Female
  + Male
* salary: nine-month salary, in dollars

First set of questions is:

* which of the measured variable is the response variable?
* which are the explanatory variables?
* are the explanatory variables measured or attribute, or a mixture of both?

Considering that the data is collected for the purpose of the analysis of academic salary we are going to identify salary as the response variable. The rest of the variables we can use as possible factors that can influence the behaviour of the response variable.

“Standard” regression approach assumes modelling a relationship between measured response and measured explanatory variables. However, often when building multiple regression models we do not wont to be limited to the choice of only measured explanatory variables. We want to be able to include attribute explanatory variables in the multiple regression modelling too. In consequence, it is important to learn about supplementary steps that are required to make such model interpretable.

**Dummy Variables**

As we have learnt previously, attribute variables are vectors of class factor in R. This vector is encoded numerically, with information about the levels of the variable saved in the levels attribute.

# If you don't have carData installed yet, uncomment and run the line below  
# install.packages(carData)  
library(carData)  
data(Salaries)  
attach(Salaries)  
class(sex)

## [1] "factor"

unclass(sex)

## [1] 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 1 1 2  
## [38] 2 2 2 2 2 2 2 2 2 2 1 1 2 2 2 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 1 2 2 2 2 2  
## [75] 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2  
## [112] 2 2 2 1 2 2 2 2 1 2 2 2 1 2 2 2 1 2 2 2 2 1 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [149] 1 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2  
## [186] 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2  
## [223] 2 2 2 2 2 2 2 2 1 1 2 1 2 2 2 1 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 1 1 2 2 2 2  
## [260] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [297] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 1  
## [334] 2 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 1 2 2 2 2 2 2 2 2  
## [371] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## attr(,"levels")  
## [1] "Female" "Male"

Note, that the unclass() function removes the attributes of the sex variable above and prints it using default print method, allowing for easier examination of the internal structure of the object.

However, when using attribute variable in a linear regression model it would make no sense to treat it as a measured explanatory variable because of its numeric levels. In the context of linear modelling we need to code them to represent the levels of the attribute.

Two-level attribute variables are very easy to code. We simply create an **indicator** or **dummy** variable that takes on two possible *dummy* numerical values. Consider the sex indicator variable. We can code this using a dummy variable :

$$\begin{equation}
d=\left\{
\begin{array}{@{}ll@{}}
0, & \text{if female} \\
1, & \text{if male}
\end{array}\right.
\end{equation}$$

{{% notice warning %}} 💡 This is the default coding used in R. Zero value is assigned to the level which is first alphabetically, unless it is changed by using the releveld() function for example, or by specifying the levels of the factor variable specifically. {{% /notice %}}

For a simple regression model of salary verses sex:

this results in the model

$$\begin{equation}
salary\_i = b\_0 + b\_1sex\_i + e\_i=\left\{
\begin{array}{@{}ll@{}}
b\_0 + b\_1 \times 1 + e\_i = b\_0 + b\_1 + e\_i, & \text{if the person is male} \\
b\_0 + b\_1 \times 0 + e\_i = b\_0 + e\_i, & \text{if the person is female}
\end{array}\right.
\end{equation}$$

where can be interpreted as the average for females, and as the average for males. The value of represents the average difference in between females and males.

We can conclude that dealing with an attribute variable with two levels in a linear model is straight forward. In this case, a dummy variable indicates whether an observation has a particular characteristic: yes/no. We can observe it as a “switch” in a model, as this dummy variable can only assume the values 0 and 1, where 0 indicates the absence of the effect, and 1 indicates the presence. The values 0/1 can be seen as off/on.

The way in which R codes dummy variables is controlled by the ***contrast*** option:

options("contrasts")

## $contrasts  
## unordered ordered   
## "contr.treatment" "contr.poly"

The output points out the conversion of the factor into appropriate set of contrasts. In particular, the first one: for unordered factors, and the second one: the order for order factors. The former one is applicable in our context.

To explicitly identify the coding of the factor, i.e. dummy variable used by R, we can use the contrasts() function.

contrasts(sex)

## Male  
## Female 0  
## Male 1

contrasts(discipline)

## B  
## A 0  
## B 1

contrasts(rank)

## AssocProf Prof  
## AsstProf 0 0  
## AssocProf 1 0  
## Prof 0 1

Note that applied “contr.treatment” conversion takes only the value 0 or 1 and that for an attribute variable with k levels it will create k-1 dummy variables. One can argue that the printout of the function could have been more informative by putting indexed letter d as a header for each of the printed columns. For example:

* attribute variable sex, where k=2

|  |  |
| --- | --- |
| attribute |  |
| Female | 0 |
| Male | 1 |

* attribute variable rank, where k=3

|  |  |  |
| --- | --- | --- |
| attribute |  |  |
| AsstProf | 0 | 0 |
| AssocProf | 1 | 0 |
| Prof | 0 | 1 |

There are many different ways of coding attribute variables besides the dummy variable approach explained here. All of these different approaches lead to equivalent model fits, what differ are the coefficients, i.e. model parameters as they require different interpretations, arranged to measure particular contrasts. This 0/1 coding implemented in R’s default **contr.treatment** contrast offers straightforward interpretation of the associated parameter in the model, which often is not the case when implementing other contrasts.

**Interpreting coefficients of attribute variables**

In the case of measured predictors, we are comfortable with the interpretation of the linear model coefficient as a *slope*, which tells us what is a *unit increase in the response variable, i.e. outcome per unit increase in the explanatory variable*. This is not necessarily the right interpretation for attribute explanatory variables.

# average salary values for each sex group  
suppressPackageStartupMessages(library(dplyr))  
Salaries %>%   
 select(salary, sex) %>%   
 group\_by(sex) %>%   
 summarise(mean=mean(salary))

## # A tibble: 2 x 2  
## sex mean  
## <fct> <dbl>  
## 1 Female 101002.  
## 2 Male 115090.

If we obtain mean of the salary for each sex group we will get that for female professors average salary is and for male professors the average of . That is, a difference of . If we now look at the parameters of the regression model for salary vs sex where female are coded as zero and male as one we get exactly same information, implying that the coefficient is the estimated difference in average between the two groups.

# regression model   
lm(salary ~ sex)

##   
## Call:  
## lm(formula = salary ~ sex)  
##   
## Coefficients:  
## (Intercept) sexMale   
## 101002 14088

For more on this topic check the following link: [Categorical variables and interaction terms in linear regression](https://www.andrew.cmu.edu/user/achoulde/94842/lectures/lecture10/lecture10-94842.html#interpreting-coefficients-of-factor-variables)

**Fitting a Moultivariate Regression Model**

We are interested in the extent to which variation in the response variable salary is associated with variation in the explanatory variables available in the Salaries data set, that is we want to fit a multiple linear regression model to the given data. The model we wish to construct should be plentiful enough to explain relations in the data and at the same time simple enough to understand, explain to others, and use.

For the convenience we will adopt the following notation:

* : salary
* : yrs.since.phd
* : yrs.service
* : discipline
* : sex
* : rank

Generally, in multiple regression we have a continuous response variable and two or more continuous explanatory variables, however in this dataset we have three attribute variables that we wish to include as the explanatory variables in the model.

Next, we need to specify the model that embodies our mechanistic understanding of the factors involved and the way that they are related to the response variable. It would make sense to expect that all of the available variables may impact the behaviour of , thus the model we wish to built should reflect our view point, i.e. :

Our view point states a belief that all explanatory variables have positive impact on the response. For example, more years in service will cause higher salary.

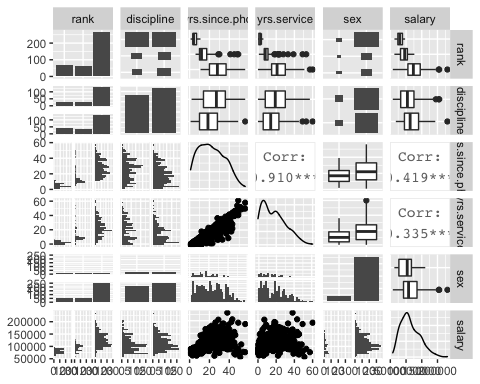
Our objective now is to determine the values of the parameters in the model that lead to **the best fit of the model to the data**. That is, we not only trying to estimate the parameters of the model, but we are also seeking for the minimal adequate model to describe the data.

{{% notice tip %}} The best model is the model that produces the least unexplained variation following the principle of parsimony rather than complexity. That is the model should have as few parameters as possible, subject to the constraint that the parameters in the model should all be statistically significant. {{% /notice %}}

For regression modelling in R we use the lm() function, that fits a linear model assuming normal errors and constant variance. We specify the model by a formula that uses arithmetic operators: +, -, \*, / and ^ which enable different functionalities from their ordinary ones. But, before we dive into statistical modelling of the given data, we need to make a fist step and conduct the most fundamental task of data analysis procedure: **Get to Know Our Data**.

Examining multivariate data is intrinsically more challenging than examining univariate and bivariate data. To get the most in-depth vision into multivariate data behaviour we construct scatter plot matrices that enable display of pairwise relationships. In R, the scatter plot matrices are composed by the pairs() function, which comes as a part of the default graphics package. Since we wish to include attribute variables in our analysis we are going to use the GGally::ggpairs() function that produces a pairwise comparison of multivariate data for both data types: measured and attribute.

# If you don't have GGally installed yet, uncomment and run the line below  
# install.packages(GGally)  
suppressPackageStartupMessages(library(GGally))  
ggpairs(Salaries)



It is an information rich visualisation that includes pairwise relationships of all the variables we want to consider for our model. By focusing on the last column of the plots, we can notice influence from all explanatory variables onto the response, expect maybe for disscipline and sex. We also notice unbalanced representation of the groups for variables rank and sex, but for the purpose of our practice in fitting a multi factor model this isn’t that problematic. We need to be especially concerned with the extent of correlations between the explanatory variables and what is of particular interest to us is the **high multicollinearity** between rank, yrs.since.phd and yrs.service, which happens when the variables are **highly linearly related**. As a consequence, we will need to keep an eye on the significance of using all of those variables in the model.

**Fitting the Model**

There are no fixed rules when fitting a linear modelles, but there are adopted standards that have proven to work well in practice. We start off by fitting a maximal model then we carry on simplifying it by removing non-significant explanatory variables. This needs to be done with caution, making sure that the simplifications make good scientific sense, and do not lead to significant reductions in explanatory power. Although this should be adopted strategy for fitting a model it is not a warrant in finding all the important structures in a complex data frame.

We can summarise our model building procedure algorithm as following:

1. Fit the **maximal model** that includes all the variables
   * Asses the overall significance of the model by checking how big is the /. If statistically significant, carry on with the model fitting procedure, otherwise stop (F-test)
2. Remove the least significant terms **one at the time**
   * Check variables’ values and perform one or two tail *t-test* in accordance to your prior view
   * If the deletion causes an insignificant increase in leave that term out of the model
3. Keep removing terms from the model until the model contains nothing but significant terms.

# model\_1 <- lm(salary ~ yrs.since.phd + yrs.service + discipline + sex + rank, data = Salaries) #long handed way  
model\_1 <- lm(salary ~ ., data = Salaries) # full stop, . , implies: all other variables in data that do not already appear in the formula  
summary(model\_1)

##   
## Call:  
## lm(formula = salary ~ ., data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -65248 -13211 -1775 10384 99592   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 65955.2 4588.6 14.374 < 2e-16 \*\*\*  
## rankAssocProf 12907.6 4145.3 3.114 0.00198 \*\*   
## rankProf 45066.0 4237.5 10.635 < 2e-16 \*\*\*  
## disciplineB 14417.6 2342.9 6.154 1.88e-09 \*\*\*  
## yrs.since.phd 535.1 241.0 2.220 0.02698 \*   
## yrs.service -489.5 211.9 -2.310 0.02143 \*   
## sexMale 4783.5 3858.7 1.240 0.21584   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22540 on 390 degrees of freedom  
## Multiple R-squared: 0.4547, Adjusted R-squared: 0.4463   
## F-statistic: 54.2 on 6 and 390 DF, p-value: < 2.2e-16

1. Overall is the model a good fit? How big is the /?

The and the are well above the value of zero allowing us to accept this as a valid model without having to formally test it to assess its statistical significance. It manages to explain almost half of the variability in the response variable salary.

1. Individually are the explanatory variables important?

We identify sex explanatory variable as clearly not significant, which is in line with the conclusion we could draw from the boxplot in the pairwise comparison plot for salary vs. sex. We will remove it to begin the process of model simplification.

#model\_1 <- lm(salary ~ yrs.since.phd + yrs.service + discipline + sex + rank, data = Salaries) # long handed way  
model\_2 <- update(model\_1,~. - sex) # refitting by removing the least significant term  
summary(model\_2)

##   
## Call:  
## lm(formula = salary ~ rank + discipline + yrs.since.phd + yrs.service,   
## data = Salaries)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -65244 -13498 -1455 9638 99682   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 69869.0 3332.1 20.968 < 2e-16 \*\*\*  
## rankAssocProf 12831.5 4147.7 3.094 0.00212 \*\*   
## rankProf 45287.7 4236.7 10.689 < 2e-16 \*\*\*  
## disciplineB 14505.2 2343.4 6.190 1.52e-09 \*\*\*  
## yrs.since.phd 534.6 241.2 2.217 0.02720 \*   
## yrs.service -476.7 211.8 -2.250 0.02497 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22550 on 391 degrees of freedom  
## Multiple R-squared: 0.4525, Adjusted R-squared: 0.4455   
## F-statistic: 64.64 on 5 and 391 DF, p-value: < 2.2e-16

We note slight reduction in from to which we can regard as an insignificant decrease. The next step is to check the coefficients and assess for the effect of the remaining variables. We identify yrs.since.phd and yrs.service as the least influential in explaining the variability of salary. To illustrate how to formally assess their effect we will conduct the *t-test* for the yrs.since.phd variable:

=========================

* If
* If

qt(0.95, 391)

## [1] 1.64876

As we will keep the remaining variable and stop with the model simplification and focus on its interpretation.

We can take a closer look at the coefficients of our fitted model:

coef(model\_2)

## (Intercept) rankAssocProf rankProf disciplineB yrs.since.phd   
## 69869.0110 12831.5375 45287.6890 14505.1514 534.6313   
## yrs.service   
## -476.7179

The structure of our final fitted model is:

where

* : salary
* : yrs.since.phd
* : yrs.service
* : discipline
* : rank

We can take a closer look at the coefficients of our fitted model:

coef(model\_2)

## (Intercept) rankAssocProf rankProf disciplineB yrs.since.phd   
## 69869.0110 12831.5375 45287.6890 14505.1514 534.6313   
## yrs.service   
## -476.7179

Examining the output we can realise that R has created tree “sub” dummy variables for the variable rank:

It has chosen to use the model:

where:

* is salary
* is yrs.since.phd
* is yrs.servic
* and are the dummy variables define above for the purpose of coding variable rank
* is a dummy variable used in coding of variable discipline as explained earlier

Note that R doesn’t need to use , to create three models it only needs two dummy variables since it is using as a reference level, also known as the *base line*. This subsequently allows R to create three models in respect to rank variable:

* AsstProf:
* AssocProf:
* Prof:

telling us that:

* is average salary for Assistant Professor who works in a “theoretical” department and average salary for Assistant Professor who works in an “applied” department
* is average salary for Associate Professor who works in a “theoretical” department and average salary for Associate Professor who works in an “applied” department
* is average salary for Professor who works in a “theoretical” department and average salary for Professor who works in an “applied” department

Learning this we can make an interpretation of our final fitted model as following:

* For every year since PhD (yrs.since.phd) on average salary (salary) will go up by assuming the rest of the variables are fixed in the model
* For every year in service (yrs.service) on average salary (salary) will go down by assuming the rest of the variables are fixed in the model
* Average salary of an Assistant Professor (rank: AsstProf) who works in a “theoretical” department is and who works in an “applied” department is ; This can vary for the number of years in service and since PhD
* Average salary of an Associate Professor (rank: AssocProf) who works in a “theoretical” department is and who works in an “applied” department is ; This can vary for the number of years in service and since PhD
* Average salary of a Professor (rank: Prof) who works in a “theoretical” department is and who works in an “applied” department is ; This can vary for the number of years in service and since PhD

This model explain around of the variability in the response variable salary.

Adding ~ 0 to lm() formula enables R to suppress the intercept. Note that if we remove the intercept, then we can directly obtain all “three intercepts” without a base level to fit the final fitted model:

model\_2\_1 <- lm(salary ~ 0 + rank + discipline + yrs.since.phd + yrs.service)  
summary(model\_2\_1)

##   
## Call:  
## lm(formula = salary ~ 0 + rank + discipline + yrs.since.phd +   
## yrs.service)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -65244 -13498 -1455 9638 99682   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## rankAsstProf 69869.0 3332.1 20.968 < 2e-16 \*\*\*  
## rankAssocProf 82700.5 3916.7 21.115 < 2e-16 \*\*\*  
## rankProf 115156.7 4350.9 26.467 < 2e-16 \*\*\*  
## disciplineB 14505.2 2343.4 6.190 1.52e-09 \*\*\*  
## yrs.since.phd 534.6 241.2 2.217 0.0272 \*   
## yrs.service -476.7 211.8 -2.250 0.0250 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 22550 on 391 degrees of freedom  
## Multiple R-squared: 0.9638, Adjusted R-squared: 0.9633   
## F-statistic: 1736 on 6 and 391 DF, p-value: < 2.2e-16

**Adding a Complexity**

So far we have seen that by fitting an additive regression model to the data we wish to identify and understand how the value of a dependent response variable changes when any one of the independent explanatory variables is changed while the other independent variables stay the same. This is a restrictive form of a model as it only allows for linear relationships between the response and the explanatory variables and the way in which one explanatory variable affects the response is the same for any value of the other explanatory variables used in the model.

We need to add a flexibility to accommodate for those limitations that will allow:

* the use of linear models for **non-linear relationships** and
* in which the effect of one explanatory variable can be different for different values of the other explanatory variable by introducing the concept of **interaction**

that command more complexity into the multivariate regression model.

**Case Study: The Quality of red Bordeaux Vintages**

To illustrate multivariate model fitting procedure with interactions we are going to use [wine.csv](https://raw.githubusercontent.com/egarpor/handy/master/datasets/wine.csv) available from [Eduardo García Portugués’s book: Notes for Predictive Modeling](https://bookdown.org/egarpor/PM-UC3M/lm-i-lab-wine.html). The dataset is formed by the auction Price of 27 red Bordeaux vintages, five vintage descriptors (WinterRain, AGST, HarvestRain, Age, Year), and the population of France in the year of the vintage, FrancePop.

* Year: year in which grapes were harvested to make wine
* Price: logarithm of the average market price for Bordeaux vintages according to 1990–1991 auctions. The price is relative to the price of the 1961 vintage, regarded as the best one ever recorded
* WinterRain: winter rainfall (in mm)
* AGST: Average Growing Season Temperature (in Celsius degrees)
* HarvestRain: harvest rainfall (in mm)
* Age: age of the wine measured as the number of years stored in a cask
* FrancePop: population of France at Year (in thousands)

We would like to analyse the quality of a vintage that has been quantified as the price and make the interpretation of our statistical findings.

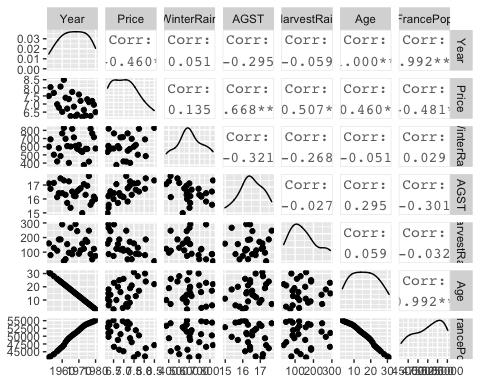
{{% notice warning %}} Don’t forget!!! 🤔 First things first: **Get to Know Data** 🤓 {{% /notice %}}

We will start this analysis by examining the pairwise plot.

wine = read.csv("https://raw.githubusercontent.com/egarpor/handy/master/datasets/wine.csv")  
summary(wine)

## Year Price WinterRain AGST HarvestRain   
## Min. :1952 Min. :6.205 Min. :376.0 Min. :14.98 Min. : 38.0   
## 1st Qu.:1960 1st Qu.:6.508 1st Qu.:543.5 1st Qu.:16.15 1st Qu.: 88.0   
## Median :1967 Median :6.984 Median :600.0 Median :16.42 Median :123.0   
## Mean :1967 Mean :7.042 Mean :608.4 Mean :16.48 Mean :144.8   
## 3rd Qu.:1974 3rd Qu.:7.441 3rd Qu.:705.5 3rd Qu.:17.01 3rd Qu.:185.5   
## Max. :1980 Max. :8.494 Max. :830.0 Max. :17.65 Max. :292.0   
## Age FrancePop   
## Min. : 3.00 Min. :43184   
## 1st Qu.: 9.50 1st Qu.:46856   
## Median :16.00 Median :50650   
## Mean :16.19 Mean :50085   
## 3rd Qu.:22.50 3rd Qu.:53511   
## Max. :31.00 Max. :55110

ggpairs(wine)



*What conclusions can we draw:*

* We can notice a perfect relationship between the variables Yeear and Age. This is to expect since this data was collected in 1983 and Age was calculated as: Age = 1983 - Year. Knowing this, we are going to remove Year from the analysis and use Age as it will be easier to interpret.
* There is a strong relationship between Year, ie. Age and FrancePop and since we want to impose our viewpoint that the total population does not influence the quality of wine we will not consider this variable in the model.
* We are going to investigate possible interactions between the rainfalls (WinterRain) and growing season temperature (AGST). In R this will be create automatically by using the \* operator.

*Let us build a model:*

We will start with the most complicated model that includes the highest-order interaction.

{{% notice note %}} In R we will specify the three-way interaction, which will automatically add all combinations of two-way interactions. {{% /notice %}}

model1 <- lm(Price ~ WinterRain + AGST + HarvestRain + Age + WinterRain \* AGST \* HarvestRain, data = wine)  
summary(model1)

##   
## Call:  
## lm(formula = Price ~ WinterRain + AGST + HarvestRain + Age +   
## WinterRain \* AGST \* HarvestRain, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.35058 -0.19462 -0.02645 0.17194 0.52079   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 8.582e+00 1.924e+01 0.446 0.6609   
## WinterRain -1.858e-02 2.896e-02 -0.642 0.5292   
## AGST -1.748e-01 1.137e+00 -0.154 0.8795   
## HarvestRain -4.713e-02 1.540e-01 -0.306 0.7631   
## Age 2.476e-02 8.288e-03 2.987 0.0079 \*\*  
## WinterRain:AGST 1.272e-03 1.712e-03 0.743 0.4671   
## WinterRain:HarvestRain 7.836e-05 2.600e-04 0.301 0.7665   
## AGST:HarvestRain 3.059e-03 9.079e-03 0.337 0.7401   
## WinterRain:AGST:HarvestRain -5.446e-06 1.540e-05 -0.354 0.7278   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2833 on 18 degrees of freedom  
## Multiple R-squared: 0.8621, Adjusted R-squared: 0.8007   
## F-statistic: 14.06 on 8 and 18 DF, p-value: 2.675e-06

The model explains well over of variability and is clearly a strong model, but the key question is can we simplify it. We will start the process of this model simplification by removing the three-way interaction as it is clearly not significant.

model2 <- update(model1, ~. -WinterRain:AGST:HarvestRain, data =wine)  
summary(model2)

##   
## Call:  
## lm(formula = Price ~ WinterRain + AGST + HarvestRain + Age +   
## WinterRain:AGST + WinterRain:HarvestRain + AGST:HarvestRain,   
## data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.35245 -0.19452 0.01643 0.17289 0.51420   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2.980e+00 1.066e+01 0.279 0.78293   
## WinterRain -9.699e-03 1.408e-02 -0.689 0.49930   
## AGST 1.542e-01 6.383e-01 0.242 0.81168   
## HarvestRain 6.496e-03 2.610e-02 0.249 0.80610   
## Age 2.441e-02 8.037e-03 3.037 0.00678 \*\*  
## WinterRain:AGST 7.490e-04 8.420e-04 0.890 0.38484   
## WinterRain:HarvestRain -1.350e-05 7.338e-06 -1.840 0.08144 .   
## AGST:HarvestRain -1.032e-04 1.520e-03 -0.068 0.94656   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2767 on 19 degrees of freedom  
## Multiple R-squared: 0.8611, Adjusted R-squared: 0.8099   
## F-statistic: 16.83 on 7 and 19 DF, p-value: 6.523e-07

The has slightly increased in value. Next, we remove the least significant two-way interaction term.

model3 <- update(model2, ~. -AGST:HarvestRain, data = wine)  
summary(model3)

##   
## Call:  
## lm(formula = Price ~ WinterRain + AGST + HarvestRain + Age +   
## WinterRain:AGST + WinterRain:HarvestRain, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.35424 -0.19343 0.01176 0.17161 0.51218   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.518e+00 6.946e+00 0.507 0.61802   
## WinterRain -1.017e-02 1.195e-02 -0.851 0.40488   
## AGST 1.218e-01 4.138e-01 0.294 0.77147   
## HarvestRain 4.752e-03 4.553e-03 1.044 0.30901   
## Age 2.451e-02 7.710e-03 3.179 0.00472 \*\*  
## WinterRain:AGST 7.769e-04 7.166e-04 1.084 0.29119   
## WinterRain:HarvestRain -1.342e-05 7.059e-06 -1.902 0.07174 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2697 on 20 degrees of freedom  
## Multiple R-squared: 0.8611, Adjusted R-squared: 0.8194   
## F-statistic: 20.66 on 6 and 20 DF, p-value: 1.35e-07

Again, it is reassuring to notice an increase in the , but we still simplify the model further by removing another least significant two-way interaction term.

model4 <- update(model3, ~. -WinterRain:AGST, data = wine)  
summary(model4)

##   
## Call:  
## lm(formula = Price ~ WinterRain + AGST + HarvestRain + Age +   
## WinterRain:HarvestRain, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.5032 -0.1934 0.0109 0.1771 0.4621   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.812e+00 1.598e+00 -2.386 0.026553 \*   
## WinterRain 2.747e-03 9.471e-04 2.900 0.008560 \*\*   
## AGST 5.586e-01 9.495e-02 5.883 7.71e-06 \*\*\*  
## HarvestRain 4.717e-03 4.572e-03 1.032 0.313877   
## Age 2.785e-02 7.094e-03 3.926 0.000774 \*\*\*  
## WinterRain:HarvestRain -1.349e-05 7.088e-06 -1.903 0.070835 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2708 on 21 degrees of freedom  
## Multiple R-squared: 0.8529, Adjusted R-squared: 0.8179   
## F-statistic: 24.35 on 5 and 21 DF, p-value: 4.438e-08

There is an insignificantly decrease in . We notice HarvestRain now being the least significant term, but as it is used for the WinterRain:HarvestRain interaction, which is significant at and therefore we should keep it. However, as the concept of parsimony prefers a model without interactions to a model containing interactions between variables, we will remove the remaining interaction term and see if it significantly affects explanatory power of the model.

model5 <- update(model4, ~. -WinterRain:HarvestRain, data = wine)  
summary(model5)

##   
## Call:  
## lm(formula = Price ~ WinterRain + AGST + HarvestRain + Age, data = wine)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.46024 -0.23862 0.01347 0.18601 0.53443   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.6515703 1.6880876 -2.163 0.04167 \*   
## WinterRain 0.0011667 0.0004820 2.420 0.02421 \*   
## AGST 0.6163916 0.0951747 6.476 1.63e-06 \*\*\*  
## HarvestRain -0.0038606 0.0008075 -4.781 8.97e-05 \*\*\*  
## Age 0.0238480 0.0071667 3.328 0.00305 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.2865 on 22 degrees of freedom  
## Multiple R-squared: 0.8275, Adjusted R-squared: 0.7962   
## F-statistic: 26.39 on 4 and 22 DF, p-value: 4.057e-08

The is reduced by around , but it has all significant terms and it is easier to interpret. For those reasons and in the spirit of parsimony that argues that a model should be as simple as possible we can suggest that this should be regarded as the best final fitted model.

We realise that for the large numbers of explanatory variables, and many interactions and non-linear terms, the process of model simplification can take a very long time. There are many algorithms for automatic variable selection that can help to chose the variables to include in a regression model. **Stepwise** regression and **Best Subsets** regression are two of the more common variable selection methods.

The *stepwise* procedure starts from the saturated model (or the maximal model, whichever is appropriate) through a series of simplifications to the minimal adequate model. This progression is made on the basis of deletion tests: F tests, AIC, t-tests or chi-squared tests that assess the significance of the increase in deviance that results when a given term is removed from the current model.

The *best subset* regression (*BREG*), also known as “all possible regressions”, as the name of the procedure indicates, fits a separate least squares regression for each possible combination of the predictors, i.e. explanatory variables. After fitting all of the models, BREG then displays the best fitted models with one explanatory variable, two explanatory variables, three explanatory variables, and so on. Usually, either adjusted R-squared or Mallows Cp is the criterion for picking the best fitting models for this process. The result is a display of the best fitted models of different sizes up to the full/maximal model and the final fitted model can be selected by comparing displayed models based on the criteria of parsimony.

“These methods frequently are abused by naive researchers who seek to interpret the order of entry of variables into the regression equation as an index of their ‘‘importance.’’ This practice is potentially misleading.” [J. Fox and S. Weisberg book: An R Companion to Applied Regression, Third Edition, Sage (2019)](https://socialsciences.mcmaster.ca/jfox/Books/Companion/)

{{% notice tip %}} 💡 When selecting a model one should remember the important concept of parsimony! {{% /notice %}}

As [M.J. Crawley](https://royalsociety.org/people/michael-crawley-11279/) points out in his well know editions of [“The R Book”](http://www.imperial.ac.uk/bio/research/crawley/therbook) we need to remember that models are portrayal of phenomena that should be both “**accurate and convenient**” and the principle of parsimony is an essential tool for model exploration. As he suggests: “just because we go to the trouble of measuring something does not mean we have to have it in our model.”

"Parsimony says that, other things being equal, we prefer:

* a model with parameters to a model with n parameters
* a model with explanatory variables to a model with k explanatory variables
* a linear model to a model which is curved
* a model without a hump to a model with a hump
* a model without interactions to a model containing interactions between factors" [Crawley, M.J. 2013, The R Book. 2nd Edition. John Wiley, New York](http://www.imperial.ac.uk/bio/research/crawley/therbook)

**Useful links:**

* [How To Interpret R-squared in Regression Analysis](https://statisticsbyjim.com/regression/interpret-r-squared-regression/#:~:text=R%2Dsquared%20evaluates%20the%20scatter,multiple%20determination%20for%20multiple%20regression.&text=R%2Dsquared%20is%20the%20percentage,that%20a%20linear%20model%20explains.)
* [Dummy-Variable Regression](https://socialsciences.mcmaster.ca/jfox/Courses/SPIDA/dummy-regression-notes.pdf)
* [Least-Squares Regression](http://www.stat.yale.edu/Courses/1997-98/101/linreg.htm)
* [The F-test for Linear Regression](http://facweb.cs.depaul.edu/sjost/csc423/documents/f-test-reg.htm)
* [Best Subset](http://www.science.smith.edu/~jcrouser/SDS293/labs/lab8-r.html), <https://afit-r.github.io/model_selection>, <https://bookdown.org/tpinto_home/Regularisation/best-subset-selection.html>

**Summary**

For problems ranging from bioinformatics to marketing, many analysts desire to develop “classifiers” instead of developing predictive models

Data to practice:

<https://www.kdnuggets.com/datasets/index.html>

<http://archive.ics.uci.edu/ml/index.php>

**Further Reading**

Claeskens, G. and Hjort, N.L. (2008) Model Selection and Model Averaging, Cambridge University Press, Cambridge.

Fox, J. (2002) An R and S-Plus Companion to Applied Regression, Sage, Thousand Oaks, CA.

**YOUR TURN 👇**

1. Go back to Saloaries data Case Study:
2. Adding ~0 to lm() formula enables R to suppress the intercept. If we remove the intercept, then we can directly obtain all “three intercepts” without a reference level try to fit the following model:

model\_2\_1 <- lm(salary ~ 0 + rank + discipline + yrs.since.phd + yrs.service)  
summary(model\_2\_1)

* Does this model differ from the previously fitted model\_2? Provide justified explanation
* Interpret the model

1. Can the final fitted model developed for Salaries data Case Study be further simplified? Justify your answer
2. Use Prestige data available from the carData package of datasets designed to accompany [J. Fox and S. Weisberg book: An R Companion to Applied Regression, Third Edition, Sage (2019)](https://socialsciences.mcmaster.ca/jfox/Books/Companion/). Fit a multivariate model that explains variation in the response variable prestige, explainig the resanos behind the steps taken with appropriate interpretation of the final fitted model.

carData::Prestige: The Canadian occupational prestige data, where the observations are occupations. Justification for treating the 102 occupations as a sample implicitly rests on the claim that they are “typical” of the population, at least with respect to the relationship between prestige and income.

* education: Average education of occupational incumbents, years, in 1971.
* income: Average income of incumbents, dollars, in 1971.
* women: Percentage of incumbents who are women.
* prestige: Pineo-Porter prestige score for occupation, from a social survey conducted in the mid-1960s.
* census: Canadian Census occupational code.
* type: Type of occupation. A factor with levels (note: out of order):
  + bc: Blue Collar
  + prof: Professional, Managerial, and Technical
  + wc: White Collar

1. The dataset Prostate available in the package lasso2 contains information on 97 men who were about to receive a radical prostatectomy. These data come from a study examining the correlation between the prostate specific antigen (lpsa) and a number of other clinical measures.

* Select the best subset model to predict lpsa using all the available predictors.

Take a look at [vignette for the lasso2 package](<https://cran.r-project.org/web/packages/lasso2/vignettes/Manual-vignette.pdf>

<https://www.theguardian.com/data>

<https://data.cdrc.ac.uk/search/field_tags/tutorial-1679/type/dataset?sort_by=changed>

<https://data.cdrc.ac.uk/dataset/introduction-spatial-data-analysis-and-visualisation-r>

© 2021 Tatjana Kecojevic