Review of Multiple Regression

Jerome Reboulleau[1]

[1] HEC, Lausanne

Quantitative Methods for Management

Outline

PART I:
TARGETS

Multiple cases

Outcome in multiple linear analysis

- Understand the general concepts behind model building.
- Analyze the model output.
- Test hypotheses about the significance of a multiple regression model and independent variables.
- Understand the uses of stepwise regression.

PART I: MULTIPLE REGRESSION MODEL

Model Equation

- We wish to explain y thanks to several independant variables such as:

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \ldots + \beta_k \cdot x_k + \epsilon \quad (1)$$

where

$$\begin{aligned} \beta_0 &= \text{Population's regression constant} \\ \beta_j &= \text{Population regression coefficient for each variable j} \\ k &= \text{Number of indepedent variables } x_i \\ \epsilon &= \text{Model Error} \end{aligned}$$

Model Assumptions

- The error terms ϵ are statistically independent of one another.
- The distribution of ϵ is normal.
- For all values of x ,the ϵ have equal variance.
- The means of the dependant variable, y, for all specified values of x can be connected with a line called the population regression line.

Model building and Model Diagnosis

- Model building is the process of constructing a mathematical regression model where some independent variable x are selected to explain variations of y.
- Model diagnosis is the analysis of the quality of the model including: output analysis, model quality...
- The best model will be the simplest one explaining a satisfying level of y variations.

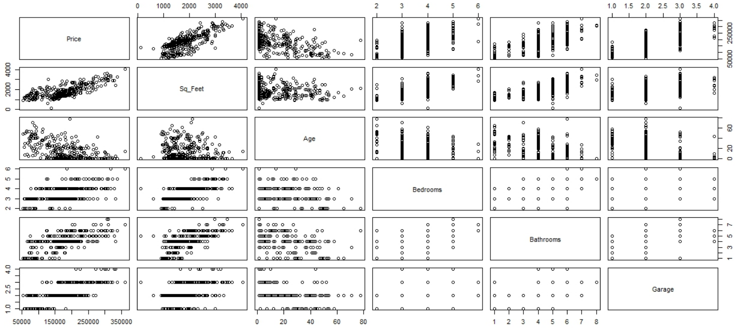
Example: Developing a Multiple Regression Model

- We are interested in developing a model for house prices with following variables:

$$
\begin{aligned}
y &= \text{House price in dollars} \\
x_1 &= \text{Home size in square feet} \\
x_2 &= \text{Age of the house} \\
x_3 &= \text{Number of bedrooms} \\
x_4 &= \text{Number of bathrooms} \\
x_5 &= \text{Garage size in number of cars}
\end{aligned}
$$

Example: Developing a Multiple Regression Model

- Step 1: Scatterplot We start by developing an intuitive analysis of the various



- The linear relationship seems satisfying between y and each x_j. For instance, the relationship between House Price and Square Feet is quite good.

Example

- Step 2: Correlation Matrix We compute the correlation matrix

	Price	Sq Feet	Age	Bedrooms	Bathrooms	Garage
Price	1.000	0.747	-0.485	0.540	0.665	0.693
Sq Feet	0.747	1.000	-0.072	0.705	0.629	0.416
Age	-0.485	-0.072	1.000	-0.202	-0.387	-0.437
Bedrooms	0.540	0.705	-0.202	1.000	0.599	0.312
Bathrooms	0.665	0.629	-0.387	0.599	1.000	0.464
Garage	0.693	0.416	-0.437	0.312	0.464	1.000

- Our graphical intuition is confirmed by the good relationship between y and all x_j.
- We can also notice that of course, the size of the house is related to the number of bedrooms for example. As a result, we can have correlation between some of the dependant variable x_j (here x_1 and x_3).
- One would apply the correlation test seen in the previous Chapter to conclude about the significance of the various correlation.
- The priority would be to check the correlation between Price and the other variables.

Example

- Step 3: Computing the regression equation We apply the OLS technique to estimate the model coefficients

```
> mymultipleregression <- lm(Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage)
> summary(mymultipleregression)

Call:
lm(formula = Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage)

Residuals:
    Min      1Q  Median      3Q     Max 
-106752  -15052    2587   17602   77565 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 31127.602   9539.669   3.263  0.00122 ** 
Sq_Feet        63.066      4.017  15.700  < 2e-16 ***
Age         -1144.437    112.780 -10.148  < 2e-16 ***
Bedrooms    -8410.379   3002.511  -2.801  0.00541 ** 
Bathrooms    3521.954   1580.997   2.228  0.02661 *  
Garage      28203.542   2858.692   9.866  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27350 on 313 degrees of freedom
Multiple R-squared:  0.8161,	Adjusted R-squared:  0.8131 
F-statistic: 277.8 on 5 and 313 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\hat{y} &= 31127.6 + 63.1 \cdot \text{Sq Feet} - 1144.4 \cdot \text{Age} - 8410.4 \cdot \text{Bedrooms} \\ &\quad + 3522.0 \cdot \text{Bathrooms} + 28203.5 \cdot \text{Garage}\end{aligned}$$

$$\Leftrightarrow \hat{y} = 31127.6 + 63.1 \cdot x_1 - 1144.4 \cdot x_2 - 8410.4 \cdot x_3 + 3522.0 \cdot x_4 + 28203.5 \cdot x_5$$

Example

- Step 4: Overall analysis of the model We analyze the regression quality for the overall model
- To achieve such analysis, we start by using the R^2 coefficient, Multiple Coefficient of Determination:
$$R^2 = \frac{SSR}{SST}$$
- We find $R^2 = 0.8161$. It means that more than 81.6% of the price variations are explained by the linear relationship.
- Next question is : is it significant ?

PART III: MODEL TESTING

Is the overall model significant ?

- Step 5: Overall significance of the model Because our analysis is based on a sample, we have to test for the overall model significance:
- Assumptions:

$$\begin{cases} H_0 : & \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0 \\ H_1 : & \text{At least one} \beta_i \neq 0 \end{cases}$$

Is the overall model significant ?

- Step 5: Overall significance of the model Because our analysis is based on a sample, we have to test for the overall model significance:
- F-Test Statistic in the case of k independant variables:

$$\begin{aligned} F &= \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} \\ &= \frac{MSR}{MSE} \end{aligned}$$

- We notice that the F statistic is here just a generalization of the one presented in the previous Chapter.
- We have defined here the "Mean Square" values:

$$\begin{aligned} MSE &= \frac{SSE}{n-k-1} \\ MSR &= \frac{SSR}{k} \end{aligned}$$

Example

- In our example, we find that

$$F = \frac{\frac{SSR}{k}}{\frac{SSE}{n-k-1}} = 2.07 \cdot 10^{11} / 7.48 \cdot 10^{8} = 277.76$$

- In order to draw test conclusions, we now compare this statistic with the following critical value of the F Distribution that has 2 degrees of freedom:
  1. The first degree of freedom is $D_1 = k$, i.e. the number of independent variables.
  2. The second degree of freedom is $D_2 = n - k - 1$.
- If we decide to do test at the level $\alpha = 1\%$, we got

$$F_{k,n-k-1,\alpha} = F_{5,319-5-1,1\%} = 3.076$$

- Since $F > F_{k,n-k-1,\alpha}$, F belongs to the rejection region and we reject H_0. One could use the p-value as well.
- The overall regression is significant.

Adjusted R-Squared

- The R-Squared is not that efficient to estimate the model quality as it does not take into account the number of dependant variables.
- If we simply used the R-Squared, we can not really compare the model quality.
- To correct this weakness, we use instead the Adjusted R-Squared, a measure based on the R-Squared but taking into account the sample size and the number of variables:

$$R_a^2 = 1 - (1 - R^2) \cdot (\frac{n-1}{n-k-1})$$

- The formula idea is that even when we add one dependent variable to our mode, R_a^2 can increase or decrease while R^2 will always increase.
- In our example, we get

$$R_a^2 = 1 - (1 - 81.6\%) \cdot (\frac{319-1}{319-1-5}) = 81.3\%$$

t-Test for Significance of each regression coefficient

- Our previous step consisted in analyzing the overall regression model.
- If the overall model is satisfying, then one would be interested in trying to optimize this model: are all dependent variables significant or not ?
- We had previously been testing that "at least" one β_j was different from 0.
- We are now going to test each one based on the same t-test as in the simple linear regression

t-Test for Significance of each regression coefficient

1. Assumptions of a two-sided test related each regression coefficient (here β_j)

$$\begin{cases} H_0: & \beta_j = 0 \\ H_1: & \beta_j \neq 0 \end{cases}$$

2. t-test Statistic

$$t = \frac{b_j - 0}{s_{b_j}}$$

3. We then compare this statistic with the critical value defined by $\pm t_{n-k-1,\alpha/2}$

t-Test for Significance of each regression coefficient

1. In our example, for the variable x_1, i.e. Square Feet, we have
$$t = \frac{b_j - 0}{s_{b_j}} = \frac{63.1 - 0}{4.02} = 15.70$$
2. We then compare this statistic with the critical value defined by $\pm t_{n-k-1,\alpha/2} = 1.97$.
3. We clearly reject H_0

Conclusions

- Following the R output

```
> mymultipleregression <- lm(Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage)
> summary(mymultipleregression)

Call:
lm(formula = Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage)

Residuals:
    Min      1Q  Median      3Q     Max
-106752  -15052    2587   17602   77565

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 31127.602   9539.669   3.263  0.00122 **
Sq_Feet        63.066      4.017  15.700  < 2e-16 ***
Age         -1144.437    112.780 -10.148  < 2e-16 ***
Bedrooms    -8410.379   3002.511  -2.801  0.00541 **
Bathrooms    3521.954   1580.997   2.228  0.02661 *
Garage      28203.542   2858.692   9.866  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27350 on 313 degrees of freedom
Multiple R-squared:  0.8161,    Adjusted R-squared:  0.8131
F-statistic: 277.8 on 5 and 313 DF,  p-value: < 2.2e-16
```

- All independent variables are significant in our model.

PART IV: MODEL DISCUSSION

One step further: Model discussion

- A further way to discuss about the model quality is associated to the Standard Error of the estimate s_ϵ.
- We remember that it is in some way the standard deviation of the regression model.
- In the context of Multiple Regression

$$s_\epsilon = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{MSE}$$

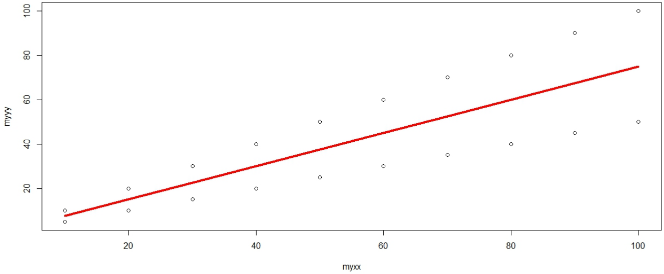
- In our example, we got

$$s_\epsilon = \sqrt{\frac{SSE}{n-k-1}} = \sqrt{\frac{234135351519}{319-5-1}} = 27350\$$$

- It means that in our model, we generally deviate from the regression by 27350$.

One step further: Model discussion, standard error

- One can notice that despite a high R-squared, we can have a very large s_ϵ



- Here the R^2 is close to 70% but the we see that the deviation of the regression is going to be quite high, specially for large values of x.

One step further: Model discussion, standard error

- This discussion is important as we see that despite a high R^2, due to the large s_ϵ, we can have poor predictions.
- As a result, any satisfying regression model does not ensure in any case having good predictions.
- The question of the model quality and of predictions are a bit disconnected.

One step further: Model discussion, multicollinearity

- Coming back to our model results, we see a strange value related to the Bedrooms

```
> mymultipleregression <- lm(Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage)
> summary(mymultipleregression)

Call:
lm(formula = Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage)

Residuals:
    Min      1Q  Median      3Q     Max
-106752  -15052    2587   17602   77565

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 31127.602   9539.669   3.263  0.00122 **
Sq_Feet        63.066      4.017  15.700  < 2e-16 ***
Age         -1144.437    112.780 -10.148  < 2e-16 ***
Bedrooms    -8410.379   3002.511  -2.801  0.00541 **
Bathrooms    3521.954   1580.997   2.228  0.02661 *
Garage      28203.542   2858.692   9.866  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 27350 on 313 degrees of freedom
Multiple R-squared:  0.8161,	Adjusted R-squared:  0.8131
F-statistic: 277.8 on 5 and 313 DF,  p-value: < 2.2e-16
```

- We note that the regression coefficient for the number of Bedrooms within an house is negative (-8410.379).
- It does not make sense: why would the house price decrease if we have more Bedrooms ? It is related to multicollinearity.

One step further: Model discussion, multicollinearity

- Coming back to the fundamental aspect of linear regression, we know that we are looking for correlation between y and many x_j
- By doing so, it may happens that we have correlation between some of the dependant variables x_i and x_j.
- In our example, we have that $\text{Cor}(\text{Sq Feet}, \text{Bedrooms}) = 70\%$, so quite high.
- It means that most of the information contained into Bedrooms is already taken into account into Sq Feet.
- Due to the redundancy of the contained information, at some point the model is lost when fitting: he does not know where the information is coming from: Sq Feet or Bedrooms.
- It affects the regression results.

One step further: Model discussion, multicollinearity

- Identifying multicollinearity:
- Potentially incorrect regression coefficient.
- A significant change in the estimate coefficient when a new variable is added to the model.
- A variable that was previously significant becomes insignificant when a new independent variable is added.
- The estimate of the standard deviation of the the model error increases when a variable is added to the model.

One step further: Model discussion, multicollinearity

- In order to see the degree of multicollinearity, we use the Variance Inflation Factor, *VIF*.
- The greater is the *VIF*, the more severe is the multicollinearity

$$VIF = \frac{1}{1 - R_j^2}$$

where R_j^2 is the coefficient of determination of the regression model

$$x_j = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_{j-1} \cdot x_{j-1} + \beta_{j+1} \cdot x_{j+1} + \ldots + \beta_k \cdot x_k$$

- If x_j is highly linearly related to the other dependant variable, the R_j^2 is large, and the *VIF* is large.
- Generally, we consider the multicollinearity to be severe if $VIF \geq 5$.
- In our case, we will not consider the multicollinearity to be too severe.

Confidence Interval for Regression Coefficients

- The formula is exactly the same as in the simple linear case up to the degrees of freedom of the Student distribution (here $n - k - 1$).

$$b_j \pm t_{n-k-1,\alpha/2} \cdot s_{b_j}$$

- In the case of the Sq Feet, we get

$$63.91 \pm 1.97 \cdot 4.017 = [56; 72]$$

- If all other variables remains constant, it means that the house price will increase on average between 56$ and 72$.

Model including Dummy Variable

- A dummy variable is a variable that is assigned 2 values to represent 2 categories: 0 for Male or 1 for Female.
- The important point is that we always need 1 variable less than the number of categories.
- To represent being a male or a female (2 categories), we need 1 variable.
- To represent 4 categories (Never Married, Married, Divorced, Widow), we need 3 variables.
- Let's introduce in our example, the dummy variable related to the area "Suburb", "Not suburb".
- We definethe Area variable such as x_6= 1 if Suburb, 0 if not.
- We are now willing to fit

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_4 \cdot x_4 + b_5 \cdot x_5 + b_6 \cdot x_6$$

Model including Dummy Variable

- The model fit after adding these new variables

```
> mymultipleregression <- lm(Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage + Area)
>
> summary(mymultipleregression)

Call:
lm(formula = Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage +
    Area)

Residuals:
   Min     1Q Median     3Q    Max
-97212 -10810   2133  12010  53857

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6817.339   7273.961  -0.937 0.349368
Sq_Feet        63.333      2.912  21.747  < 2e-16 ***
Age          -333.836     94.883  -3.518 0.000499 ***
Bedrooms    -8444.831   2176.762  -3.880 0.000128 ***
Bathrooms    -949.195   1176.549  -0.807 0.420418
Garage      26246.435   2075.752  12.644  < 2e-16 ***
Area        62040.983   3684.608  16.838  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19830 on 312 degrees of freedom
Multiple R-squared:  0.9036,    Adjusted R-squared:  0.9018
F-statistic: 487.7 on 6 and 312 DF,  p-value: < 2.2e-16
```

- We immediately note the variable Bathrooms became not significant.
- We refresh our model without this variable.

Model including Dummy Variable

- We are now fitting

$$\hat{y} = b_0 + b_1 \cdot x_1 + b_2 \cdot x_2 + b_3 \cdot x_3 + b_5 \cdot x_5 + b_6 \cdot x_6$$

```
> mymultipleregressiondummyvariable2 <- lm(Price ~ Sq_Feet + Age + Bedrooms + Garage + Area)
>
> summary(mymultipleregressiondummyvariable2)

Call:
lm(formula = Price ~ Sq_Feet + Age + Bedrooms + Garage + Area)

Residuals:
   Min     1Q Median     3Q    Max
-96167 -10557   1762  11732  53980

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7050.235   7264.176  -0.971 0.332523
Sq_Feet        62.494      2.719  22.988  < 2e-16 ***
Age          -321.988     93.688  -3.437 0.000668 ***
Bedrooms    -8830.006   2122.574  -4.160 4.11e-05 ***
Garage      26053.864   2060.832  12.642  < 2e-16 ***
Area        61370.082   3587.536  17.106  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19820 on 313 degrees of freedom
Multiple R-squared:  0.9034,    Adjusted R-squared:  0.9019
F-statistic: 585.7 on 5 and 313 DF,  p-value: < 2.2e-16
```

Model including Dummy Variable

- Model results

```
> mymultipleregressiondummyvariable2 <- lm(Price ~ Sq_Feet + Age + Bedrooms + Garage + Area)
>
> summary(mymultipleregressiondummyvariable2)

Call:
lm(formula = Price ~ Sq_Feet + Age + Bedrooms + Garage + Area)

Residuals:
   Min      1Q Median      3Q     Max
-96167 -10557   1762  11732   53980

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -7050.235   7264.176  -0.971 0.332523
Sq_Feet        62.494      2.719  22.988  < 2e-16 ***
Age          -321.988     93.688  -3.437 0.000668 ***
Bedrooms    -8830.006   2122.574  -4.160 4.11e-05 ***
Garage      26053.864   2060.832  12.642  < 2e-16 ***
Area        61370.082   3587.536  17.106  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 19820 on 313 degrees of freedom
Multiple R-squared:  0.9034,    Adjusted R-squared:  0.9019
F-statistic: 585.7 on 5 and 313 DF,  p-value: < 2.2e-16
```

$$\begin{aligned}\hat{y} \quad &= \quad -7050 + 62.5 \cdot \text{Sq Feet} - 322 \cdot \text{Age} - 8830 \cdot \text{Bedrooms} \\ &\quad +26053 \cdot \text{Garage} + 61370 \cdot \text{Area}\end{aligned}$$

- According to our coding method, a House located in a Suburb worth 61370$ more than if not.

Model including Dummy Variable

- According to the previous fitting, Bedrooms is still an issue regarding the multicollinearity.
- One may want to compare the results without the variable Bedrooms:

```
> mymultipleregressiondummyvariable3 <- lm(Price ~ Sq_Feet + Age + Garage + Area)
>
> summary(mymultipleregressiondummyvariable3)

Call:
lm(formula = Price ~ Sq_Feet + Age + Garage + Area)

Residuals:
    Min      1Q  Median      3Q     Max
-101248   -9585    1376   11633   57750

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -25617.326   5878.261  -4.358 1.78e-05 ***
Sq_Feet         54.832      2.051  26.737  < 2e-16 ***
Age           -261.297     94.917  -2.753  0.00625 **
Garage       26753.303   2106.618  12.700  < 2e-16 ***
Area         60578.045   3674.322  16.487  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20330 on 314 degrees of freedom
Multiple R-squared:  0.8981,    Adjusted R-squared:  0.8968
F-statistic: 691.9 on 4 and 314 DF,  p-value: < 2.2e-16
```

PART V: MODEL SELECTION

Stepwise Regression: Backward selection based on p-value

- Start with all predictors in the model.

$$\hat{y} = \text{constant} + \beta_1 x_1 + \ldots + \beta_p x_p$$

- Remove the predictor with the highest p-value greater than α level (for instance 5%).
- Continue until all p-values are smaller than α level.

Stepwise Regression: Backward selection based on AIC

- The backward selection starts with the most complex model and stopping at the most efficient one.

$$\hat{y} = \text{constant} + \beta_1 x_1 + \ldots + \beta_p x_p$$

- The next step is eliminating a useless independent variable x_j. The x_j that is selected is the one that explain the less variation, i.e. the less significant variable according to a criteria.
- The process continues until all non-significant variables have been eliminated.
- Several criterions can be used. The most used criteria is the AIC "Akaike Information Criteria".
- Without going too much into details, AIC computes the quality of the various models.

Stepwise Regression: Backward selection

```
> step(myfull, scope=list(lower=mynull, upper=myfull), direction="backward")
Start:  AIC=6319.85
Price ~ Sq_Feet + Age + Bedrooms + Bathrooms + Garage + Area

            Df  Sum of Sq        RSS    AIC
- Bathrooms  1 2.5590e+08 1.2292e+11 6318.5
<none>                    1.2267e+11 6319.8
- Age        1 4.8670e+09 1.2753e+11 6330.3
- Bedrooms   1 5.9175e+09 1.2858e+11 6332.9
- Garage     1 6.2859e+10 1.8553e+11 6449.8
- Area       1 1.1147e+11 2.3414e+11 6524.1
- Sq_Feet    1 1.8594e+11 3.0860e+11 6612.2

Step:  AIC=6318.51
Price ~ Sq_Feet + Age + Bedrooms + Garage + Area

           Df  Sum of Sq        RSS    AIC
<none>                   1.2292e+11 6318.5
- Age       1 4.6388e+09 1.2756e+11 6328.3
- Bedrooms  1 6.7965e+09 1.2972e+11 6333.7
- Garage    1 6.2770e+10 1.8569e+11 6448.1
- Area      1 1.1492e+11 2.3785e+11 6527.1
- Sq_Feet   1 2.0753e+11 3.3046e+11 6632.0

Call:
lm(formula = Price ~ Sq_Feet + Age + Bedrooms + Garage + Area,
    data = mydata)

Coefficients:
(Intercept)      Sq_Feet          Age     Bedrooms       Garage         Area
   -7050.23        62.49      -321.99     -8830.01     26053.86     61370.08
```

Stepwise Regression: Forward selection based on p-value

- Start with no variables in the model.
- For all predictors not in the model, compute their p-values for adding them to the model. Choose the one with the lowest p-value less than α (for instance 5%).
- Continue until no new predictors can be added.

Stepwise Regression: Forward selection based on AIC criteria

- The forward selection starts with the simplest model wihtout any variable

$$\hat{y} = \text{constant}$$

- The next step is adding an independent variable x_j among all that are available. The x_j that is selected is the one that explain the most variation, i.e. the most significant variable.

$$\hat{y} = \text{constant} + x_j$$

- The process continues until all significant variables have been selected.
- Several criterions can be used. The most used criteria is the AIC "Akaike Information Criteria".
- Without going too much into details, AIC computes the quality of the various models. The one with the lowest AIC is the best one.

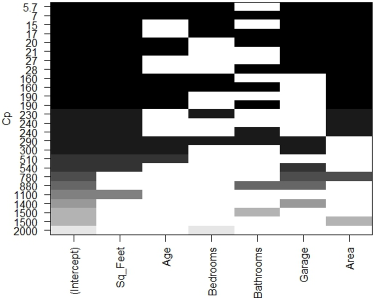
Stepwise Regression: F

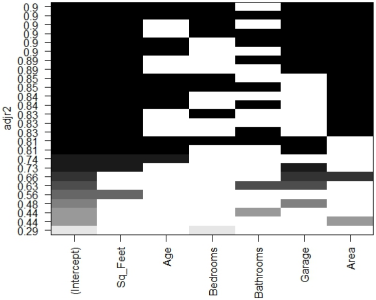
```
> step(mynull, scope=list(lower=mynull, upper=myfull), direction="forward")
Start:  AIC=7054.21
Price ~ 1

            Df  Sum of Sq        RSS    AIC
+ Sq_Feet    1 7.1172e+11 5.6131e+11 6795.0
+ Garage     1 6.1232e+11 6.6071e+11 6847.0
+ Bathrooms  1 5.6382e+11 7.0921e+11 6869.6
+ Area       1 5.6362e+11 7.0941e+11 6869.7
+ Bedrooms   1 3.7134e+11 9.0170e+11 6946.2
+ Age        1 2.9972e+11 9.7331e+11 6970.6
<none>                    1.2730e+12 7054.2

Step:  AIC=6794.99
Price ~ Sq_Feet

            Df  Sum of Sq        RSS    AIC
+ Area       1 3.4379e+11 2.1753e+11 6494.6
+ Age        1 2.3744e+11 3.2387e+11 6621.6
+ Garage     1 2.2505e+11 3.3627e+11 6633.5
+ Bathrooms  1 8.0126e+10 4.8119e+11 6747.9
<none>                    5.6131e+11 6795.0
+ Bedrooms   1 3.8433e+08 5.6093e+11 6796.8

Step:  AIC=6494.59
Price ~ Sq_Feet + Area

            Df  Sum of Sq        RSS    AIC
+ Garage     1 8.4676e+10 1.3285e+11 6339.3
+ Age        1 2.1178e+10 1.9635e+11 6463.9
+ Bedrooms   1 6.7741e+09 2.1075e+11 6486.5
<none>                    2.1753e+11 6494.6

+ Bathrooms  1 5.7107e+08 2.1696e+11 6495.7

Step:  AIC=6339.29
Price ~ Sq_Feet + Area + Garage

            Df  Sum of Sq        RSS    AIC
+ Bedrooms   1 5288505488 1.2756e+11 6328.3
+ Age        1 3130798506 1.2972e+11 6333.7
<none>                    1.3285e+11 6339.3
+ Bathrooms  1  514983221 1.3234e+11 6340.0

Step:  AIC=6328.33
Price ~ Sq_Feet + Area + Garage + Bedrooms

            Df  Sum of Sq        RSS    AIC
+ Age        1 4638812670 1.2292e+11 6318.5
<none>                    1.2756e+11 6328.3
+ Bathrooms  1   27691404 1.2753e+11 6330.3

Step:  AIC=6318.51
Price ~ Sq_Feet + Area + Garage + Bedrooms + Age

            Df Sum of Sq        RSS    AIC
<none>                   1.2292e+11 6318.5
+ Bathrooms  1 255897412 1.2267e+11 6319.8

Call:
lm(formula = Price ~ Sq_Feet + Area + Garage + Bedrooms + Age,
    data = mydata)
```

Stepwise Regression: Other selection criteria (C_p and R^2_{adj})

- In statistics, Mallows's Cp is used to assess the fit of a regression model that has been estimated using ordinary least squares. It is applied in the context of model selection.
- The C_p Mallows coefficient is associated to the *MSE*. Of course, we wish to have the smallest *MSE* as possible, so that we are looking for the smallest C_p coefficient.
- Generally, the C_p is going to be close to $k + 1$ for a good model.
- In the case of the R^2_{adj}, we are going to select the model with the highest R^2_{adj}.

Stepwise Regression: Forward selection





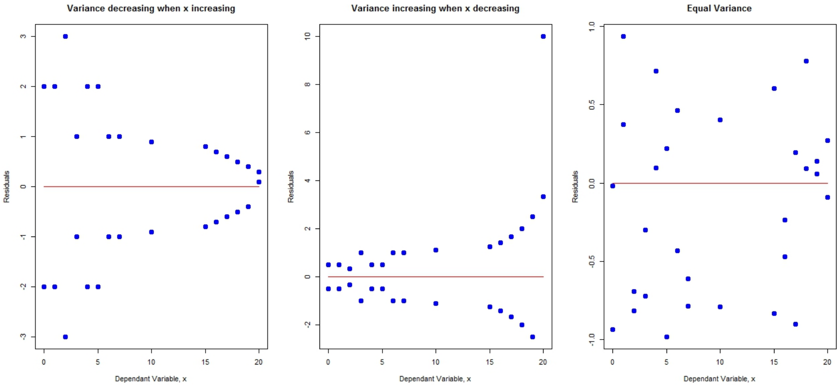
PART VI: MODEL DIAGNOSTICS

Model Diagnostic

- After all the design and the building work, we have to perform the diagnostic analysis in order to check the model's assumptions.
- We have to work on the main assumptions behind the model that are related to the residuals: equal variance, independency and normality.

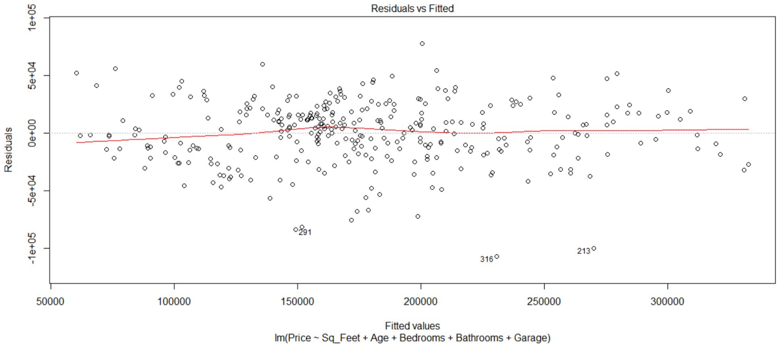
Model Diagnostic: Equal Variance

- Equal variance



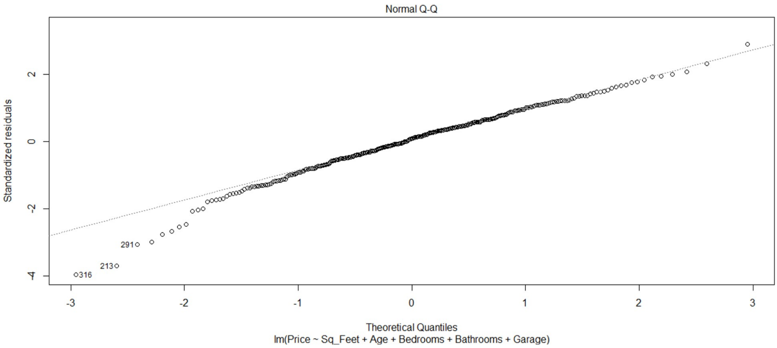
Model Diagnostic: Equal Variance

- Equal variance: We wish not to see any structure in the residuals



Model Diagnostic: Normality

- Normality is checked through a qqplot



Model Diagnostic: What to do when assumptions are not verified

- When either the variance nor the normality assumptions are not verified, we generally tends to transform the data in order to smooth the variations.
- Smoothing the variations is generally achieved by applying the log function to one, several or all variables.
- Transformation can be applied to the response variable or the independent variables.
- If such transformation is achieved, never forget to apply converse transformation to interpret the results.

Executive Summary

- Multiple Linear regression: model building and coefficients estimate.
- Overall model analysis and individual variable selection.
- Selection process: Backward, Forward.
- Dummy variable specific case.
- Model discussion: prediction quality and multicollinearity.
- Model diagnostics and data transformation.