**LINEAR/MULTIPLE LINEAR REGRESSION**

1. **Multiple Linear Regression Model**

Example: Use the “Viscosity of a Polymer” data set in lecture (construct an excel table named “Viscosity of a Polymer.xlsx” with three columns: Reaction.Temperature, Catalyst.Feed.Rate, and Viscosity)

Load the library “readxl” to read excel file

Command:

> setwd("D:/Program Problem/EU Plus Project/WP3/Course 9 \_ Data Analytic/Power Point")

 ! Change the working environment to the folder where the data file is located

> ViscosityData <- read\_excel("Viscosity of a Polymer.xlsx")

 ! Read data file and assign a name

> RegressionResult = lm(Viscosity~Reaction.Temperature + Catalyst.Feed.Rate, data = ViscosityData)

 ! Create regression model with the command “**lm**”

> summary(RegressionResult) ! Show the summarized result

Result:

Call:

lm(formula = Viscosity ~ Reaction.Temperature + Catalyst.Feed.Rate,

 data = ViscosityData)

Residuals:

 Min 1Q Median 3Q Max

-21.4972 -13.1978 -0.4736 10.5558 25.4299

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) 1566.0778 61.5918 25.43 1.80e-12 \*\*\*

Reaction.Temperature 7.6213 0.6184 12.32 1.52e-08 \*\*\*

Catalyst.Feed.Rate 8.5848 2.4387 3.52 0.00376 \*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.36 on 13 degrees of freedom

Multiple R-squared: 0.927, Adjusted R-squared: 0.9157

F-statistic: 82.5 on 2 and 13 DF, p-value: 4.1e-08

* Use the following command to see the detailed graphical results

 > plot(RegressionResult, pch = "+", col = "red")

Focus on the following plots



**Observations**: It seems that the residual plot is not structureless (the trend line is not a horizontal straight line), i.e., it seems that the residual is dependent on the fitted value. This means that the variance of the error term in the regression model is not a constant. However, the pattern observed is not so strong! It can be ignored!



**Observations**: It seems that there is no violation of the assumption that the error terms follow normal distribution

* Use the following command to see the graphical representations of the data (this plot is not important, just to show observed relationships in raw data!)

 > plot(ViscosityData, pch = 16, col = "blue")



Note: **pch** stands for “plot character”, it can be from 0 to 25 or a string which specify the character you want to use

* Use the following command to plot the residuals

 > plot(RegressionResult$residuals, pch = 16, col = "blue")



**Observations**: This plot shows the relationship between residual and the order of data points in the data set. The plot shows no specific pattern (i.e., structureless), so the error terms are independent, they do not depend on each other

1. **Analyze and Select the Appropriate Regression Model**

Example: Construct an excel table named “Income.xlsx” with the data given below to analyze the effect of age and height (in cm) on income (in Baht) of an IT engineer

|  |  |  |
| --- | --- | --- |
| **Age** | **Height** | **Income** |
| 31 | 156 | 26946 |
| 40 | 174 | 38065 |
| 45 | 170 | 45114 |
| 27 | 169 | 22422 |
| 43 | 156 | 41481 |
| 58 | 159 | 64770 |
| 46 | 164 | 46652 |
| 58 | 166 | 64804 |
| 43 | 165 | 42125 |
| 57 | 161 | 63325 |
| 51 | 164 | 53697 |
| 33 | 158 | 29361 |
| 32 | 171 | 28025 |
| 29 | 160 | 24744 |
| 30 | 163 | 25388 |
| 28 | 166 | 22427 |
| 45 | 173 | 44654 |
| 39 | 168 | 36619 |
| 42 | 171 | 39889 |
| 37 | 172 | 33280 |

Remember to load the library “readxl” to read excel file

Initial commands:

> setwd("D:/Program Problem/EU Plus Project/WP3/Course 9 \_ Data Analytic/Power Point")

 ! Change the working environment to the folder where the data file is located

> IncomeData <- read\_excel("Income.xlsx")

 ! Read data file and assign a name

* Let first consider the model:

$$Income= β\_{0}+β\_{1}\*Age+β\_{2}\*Height+ε$$

> RegressionResult = lm(Income~Age + Height, data = IncomeData)

 ! Create regression model with the command “**lm**”

> summary(RegressionResult) ! Show the summarized result

Result:

Call:

lm(formula = Income ~ Age + Height, data = IncomeData)

Residuals:

 Min 1Q Median 3Q Max

-1993.6 -460.6 -125.0 703.3 1813.0

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) -5174.86 6534.65 -0.792 0.439

Age 1374.62 22.04 62.367 <2e-16 \*\*\*

Height -67.05 38.83 -1.727 0.102

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 957.8 on 17 degrees of freedom

Multiple R-squared: 0.9957, Adjusted R-squared: 0.9952

F-statistic: 1958 on 2 and 17 DF, p-value: < 2.2e-16

**Observations**:

1. The whole model is OK due to:

 Multiple R-squared: 0.9957, Adjusted R-squared: 0.9952

 F-statistic: 1958 on 2 and 17 DF, p-value: < 2.2e-16

1. But it seems that “Height” is not an appropriate predictor due to the p-value of the test is 0.102
* Let remove the predictor “Height” and consider the model:

$$Income= β\_{0}+β\_{1}\*Age+ε$$

> RegressionResult = lm(Income~Age, data = IncomeData)

 ! Create regression model with the command “**lm**”

> summary(RegressionResult) ! Show the summarized result

Result:

Call:

lm(formula = Income ~ Age, data = IncomeData)

Residuals:

 Min 1Q Median 3Q Max

-1590.2 -730.2 24.7 749.0 1594.1

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) -16344.47 970.26 -16.84 1.82e-12 \*\*\*

Age 1376.75 23.19 59.38 < 2e-16 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1009 on 18 degrees of freedom

Multiple R-squared: 0.9949, Adjusted R-squared: 0.9946

F-statistic: 3526 on 1 and 18 DF, p-value: < 2.2e-16

**Observations**: The whole model looks OK! **But should we use it?**

Let analyze the graphs:

 > plot(RegressionResult, pch = 16, col = "blue")



**Observations**: The plot of residual vs. fitted value is not structureless (This means that the variance of the error term in the regression model is not a constant). The “Residual vs. Fitted Value” plot possesses a “quadratic” pattern. So, it may be better to incorporate the component “$Age^{2}$” in the regression model to model this pattern!



**Observations**: It seems that there is no violation of the assumption that the error terms follow normal distribution

 > plot(RegressionResult$residuals, pch = 16, col = "blue")



**Observations**: This plot shows the relationship between residual and the order of data points in the data set. The plot is also not structureless, it shows a curvature pattern.

* Now, let consider the model:

$$Income= β\_{0}+β\_{1}\*Age+β\_{2}\*Age^{2}+ε$$

> RegressionResult = lm(Income~Age + I(Age^2), data = IncomeData)

 ! Create regression model with the command “**lm**”

> summary(RegressionResult) ! Show the summarized result

Note:

Remember to declare I! The normal regression method will be used instead of the least square

method

Result:

Call:

lm(formula = Income ~ Age + I(Age^2), data = IncomeData)

Residuals:

 Min 1Q Median 3Q Max

-832.64 -113.15 96.91 264.78 476.13

Coefficients:

 Estimate Std. Error t value Pr(>|t|)

(Intercept) -114.9339 1696.9214 -0.068 0.947

Age 568.0036 82.8554 6.855 2.79e-06 \*\*\*

I(Age^2) 9.5287 0.9701 9.822 2.01e-08 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 401.9 on 17 degrees of freedom

Multiple R-squared: 0.9992, Adjusted R-squared: 0.9991

F-statistic: 1.116e+04 on 2 and 17 DF, p-value: < 2.2e-16

**Observations**: The whole model is improved! Let compare the results:

 Multiple R-squared: 0.9992, Adjusted R-squared: 0.9991

 F-statistic: 1.116e+04 on 2 and 17 DF, p-value: < 2.2e-16

With the former ones

 Multiple R-squared: 0.9949, Adjusted R-squared: 0.9946

 F-statistic: 3526 on 1 and 18 DF, p-value: < 2.2e-16

It can be observed that “R-squared”, “Adjusted R-squared”, and “F-statistic” are better!

Let analyze the graphs:

 > plot(RegressionResult, pch = 16, col = "blue")



**Observations**: The plot of residual vs. fitted value is somewhat structureless (This means that the variance of the error term in the regression model is a constant). It is better than the model without the quadratic term considered before.

 > plot(RegressionResult$residuals, pch = 16, col = "blue")



**Observations**: In term of “structureless”, this graph is also better than the graph when the quadratic term is not considered

**FINAL MODEL:** $Income= -114.9339+568.0036\*Age+9.5287\*Age^{2}$