post01-xinran-liu

Xinran Liu

October 30, 2017

Multiple Linear Regression and Its Related Application in R

1. Introduction & Motivation

- In homework1 we have used simple linear regression to predict the salary of a NBA player based on the scored points. But in real life study, it is very likely that we also want to analyze mutiple factors related to a result, and it is possible that simple linear regression model is not the most accurate one to study the data. In order to solve these problems, we would need to learn more advanced regression models that are commonly used in data analysis and how they could be applied using R.
- In this post, I would take a small step further from the simple linear regression to introduce multiple linear regression and some related tools to analyze the linear regression model. Let's employ some examples and see how the built-in functions in R help us achieve the goals!

2. Multiple Linear Regression (with example)

A very useful built-in function in R for linear regression is lm(). It is used to fit linear models, and it turns out that it could also be used for multiple linear regression. Let's take a look at some of the NBA players' data:

```
# Load necessary package
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

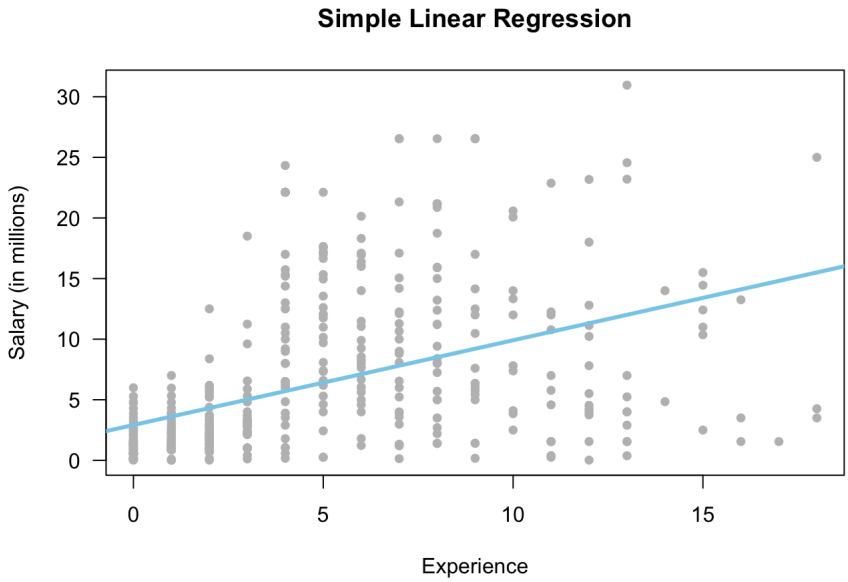
```
# Import data file
dat <- read.csv("/Users/xinran/Desktop/stat133/stat133-hws-fall17/post01/Data/nba2017-roster.csv")
# Turn salary into millions
dat <- mutate(dat, salary = round(salary / 1000000, 2))
```

If we want to do a simple linear regression between experience and salary:

```
# Use lm() to do the linear modeling
linear_model <- lm(salary ~ experience, data = dat)
summary(linear_model)
```

```
##
## Call:
## lm(formula = salary ~ experience, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.260  -2.813  -1.725   1.696  18.949
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.91511    0.41663   6.997 9.83e-12 ***
## experience   0.69969    0.06605  10.593  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.872 on 439 degrees of freedom
## Multiple R-squared:  0.2036, Adjusted R-squared:  0.2017
## F-statistic: 112.2 on 1 and 439 DF,  p-value: < 2.2e-16
```

```
# Use scatter plot show the data and linear regression line
plot(dat$experience, dat$salary, las = 1, pch = 19, cex = 0.8,
     col = "grey", xlab = "Experience", ylab = "Salary (in millions)",
     main = "Simple Linear Regression")
abline(linear_model, col = "skyblue", lwd = 3)
```



Now try to combine age and experience to do multiple linear regression:

```
# Use lm() to do multiple linear modeling
multi_linear_mod <- lm(salary ~ experience + age, data = dat)
# Use summary() to get the results
summary(multi_linear_mod)
```

```
##
## Call:
## lm(formula = salary ~ experience + age, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.352  -3.113  -1.527   2.239  18.863
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  19.3564     3.2102   6.030 3.49e-09 ***
## experience    1.3884     0.1480   9.378  < 2e-16 ***
## age          -0.7478     0.1448  -5.163 3.70e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.708 on 438 degrees of freedom
## Multiple R-squared:  0.2493, Adjusted R-squared:  0.2458
## F-statistic: 72.71 on 2 and 438 DF,  p-value: < 2.2e-16
```

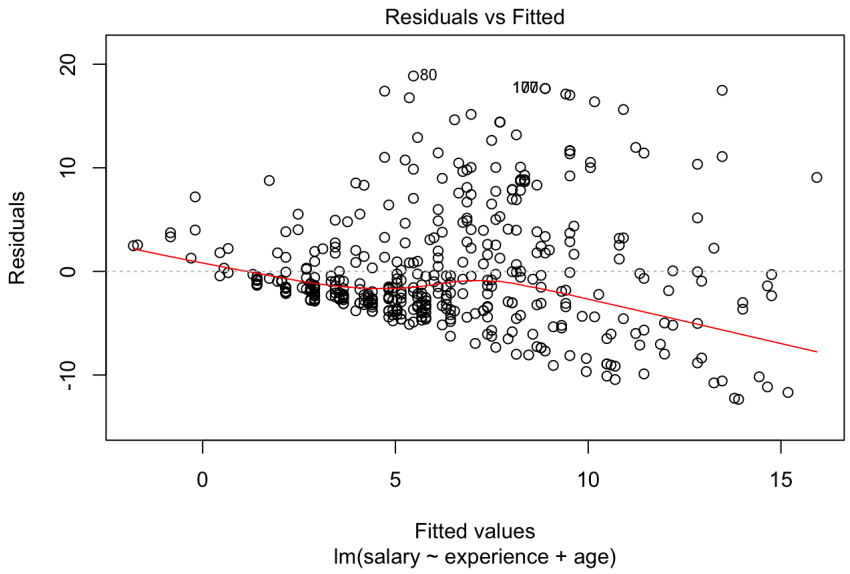
In the outcome of summary(multi_linear_mod), we could see several coefficients. Let's take a closer look at them!

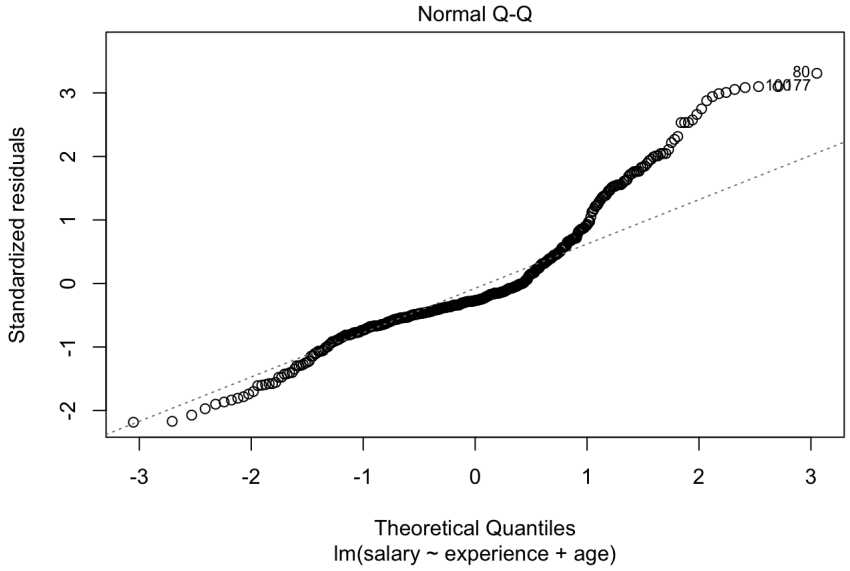
- ***Estimates*** are the regression coefficients. In this case, we could see that the linear regression equation is just $salary = 19.3564 + 1.3884 experience - 0.7478 age$.
- ***Standard errors*** are the standard errors of the regression coefficients. They could be used to construct confidence interval in the following way:
  - The confidence intervals for experience are $(1.3884 - k0.1480, 1.3884 + k0.1480)$, where k would vary based on the level of confidence.
- ***t value*** is the coefficient divided by the standard error. It can be regarded as "a measure of the precision with which the regression coefficient is measured" (From Princeton University: Interpreting regression output).
- ***Pr(>|t|)***, which is also known as the p value for each term, indicates whether the predictors that we use really have effect on response variable. If p value is less than 0.05, it is very likely that the changes in predictors' values are related to the changes in the response variable. On the contrary, if p value is large, then it indicates that the changes in predictors do not have influence on response variable. In our example, p values for experience and age are both less than 0.05, which suggest these two terms could be kept in our model.

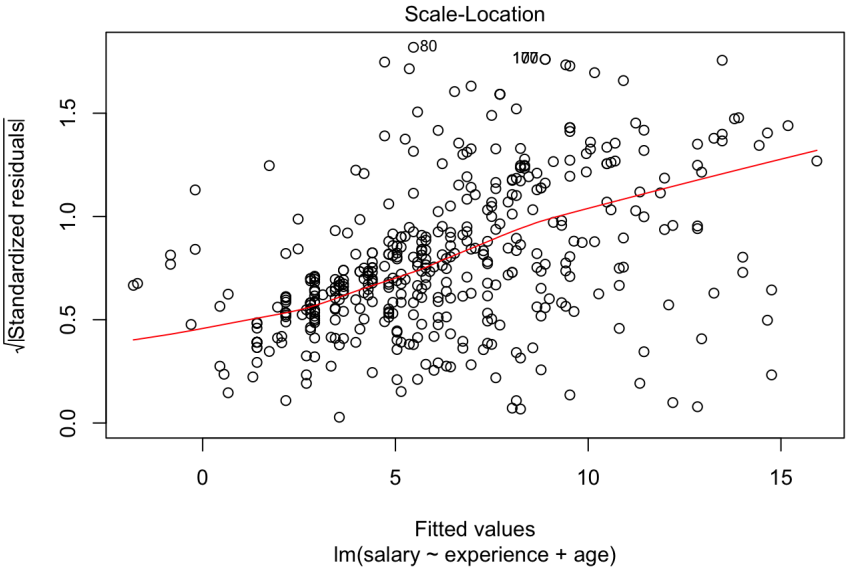
3. Useful Tools to Analyze Models

1. Diagnostic plots To begin with, let's look at the outcomes for calling plot() on multi_linear_mod we just created:

```
# diagnostic plots
plot(multi_linear_mod)
```


Residuals vs Fitted
Residuals
Fitted values
lm(salary ~ experience + age)


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(salary ~ experience + age)


Scale-Location
√|Standardized residuals|
Fitted values
lm(salary ~ experience + age)

These plots are diagnostic plots. They provide checks for heteroscedasticity, normality, and influential observerations (From Quick-R).

- The first plot depicts residuals versus predicted values (fitted values). If our data meets the linear model, the residuals should spread equally around a horizontal line without other clear patterns. In our example, the residuals are not distrbuted equally in the plot, suggesting our model does not fit the linear relation perfectly. Also notice that No.80 and 100 are marked out in the plot, suggesting they are potential problems in our modelling.
- The second plot shows if the residuals distribute normally. If the residuals are lined well in a straight line, then it suggests a good modelling. In our example the residuals line pretty well at first but go off in later points. Again notice the marked out No.80 and 100.
- The third plot checks if there is a pattern in the residuals. If the linear modelling is precise then the residuals should again distribute equally along a horizontal line. Clearly in our example the residuals show a linear pattern, which suggests the linear modelling might miss out some non-linear relationship between predictors and response variable.
- The fourth plot help us to find influential points. Those points may be extreme cases in the regression line and will result in great difference if we exclude them from analyzing. So in this plot, we need to look at outliers and cook's distance. If everything is within the line of cook's distance, then there are no extreme cases in our data. According to the plot we could not even see the dash line representing cook's distance, suggesting our result would not alter too much if we exclude anything.

2. anova() Nested models could be compared by calling anova() function.

```
# Take height into consider as well
multi_linear_mod2 <- lm(salary ~ experience + age + height,
                        data = dat)
# Call anova() to compare two models
anova(multi_linear_mod, multi_linear_mod2)
```

```
## Analysis of Variance Table
##
## Model 1: salary ~ experience + age
## Model 2: salary ~ experience + age + height
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1    438 14269
## 2    437 14147  1     122.1 3.7718 0.05276 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

To interpret the result we basically need to look at the last column. If it is less than 0.05 then the two models differ. In our case it does not suggest that adding height into consideration would be a significantly different model.

3. Robust linear regression Robust linear regression can be used in any situation where ordinary least square regression can be used. It is especially useful when there are potential outliers in data and there is no good reason to exclude them. We can do robust linear regression by calling rlm(). In general, rlm() performs better than lm().

```
# Load package
library(MASS)
```

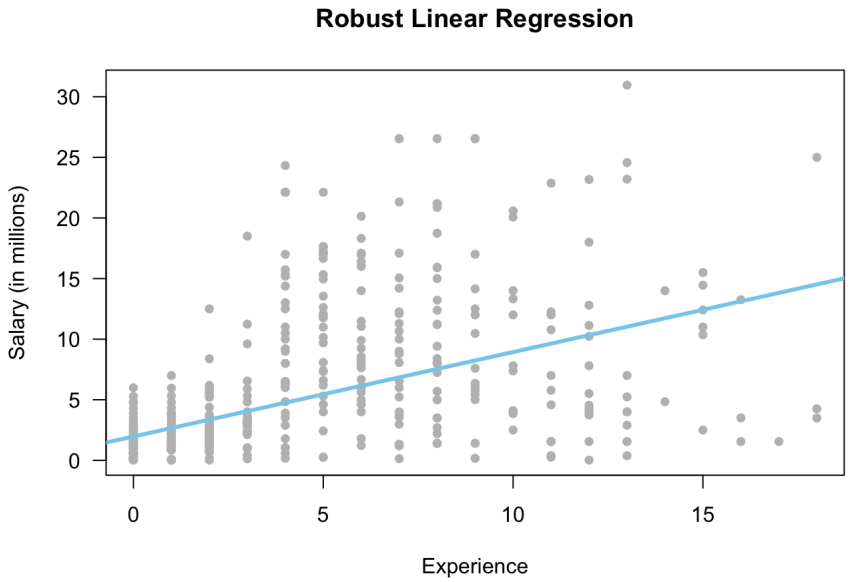
```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```r
# Use robust linear regression
rlm_mod <- rlm(salary ~ experience, data = dat)
summary(rlm_mod)
```

```
##
## Call: rlm(formula = salary ~ experience, data = dat)
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.274  -1.846  -0.776   2.649  19.926
##
## Coefficients:
##             Value   Std. Error t value
## (Intercept)  1.9660  0.2927     6.7157
## experience   0.6975  0.0464    15.0287
##
## Residual standard error: 3.056 on 439 degrees of freedom
```

```r
# Plot robust linear regression line
plot(dat$experience, dat$salary, las = 1, pch = 19, cex = 0.8,
     col = "grey", xlab = "Experience", ylab = "Salary (in millions)",
     main = "Robust Linear Regression")
abline(rlm_mod, col = "skyblue", lwd = 3)
```



4. End for now…?

From these discussions we could see how multiple linear regression is applied and analyzed using different functions in R. It is important to know not only the linear model, but also methods to analyze the precision of it. Beyond that, we should notice that linear regression is only one of the modelling choices, there are many other regression models that are non-linear and are used commonly in many real life cases. Hopefully we will get to learn about them in the future!

5. References

Dallel, Gerard E. *How to Read the Output From Multiple Linear Regression Analyses*. 2000, www.jerrydallal.com/lhsp/regout.htm. Accessed 31 Oct. 2017.

Frost, Jim. *How to Interpret Regression Analysis Results: P-values and Coefficients*. The Minitab Blog, 1 July 2013, blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients. Accessed 31 Oct. 2017.

Interpreting Regression Output. Princeton University Library Data and Statistical Services, 2007, dss.princeton.edu/online_help/analysis/interpreting_regression.htm. Accessed 31 Oct. 2017.

Kabacoff, Robert I. *Multiple (Linear) Regression*. Quick-R, www.statmethods.net/stats/regression.html. Accessed 31 Oct. 2017.

Kabacoff, Robert I. *Robust Regression*. Quick-R, r-statistics.co/Robust-Regression-With-R.html. Accessed 31 Oct. 2017.

Kim, Bommae. *Understanding Diagnostic Plots for Linear Regression Analysis*. U.Va. Research Data Services + Sciences, 21 Sep. 2015, data.library.virginia.edu/diagnostic-plots/. Accessed 31 Oct. 2017.

ROBUST REGRESSION | R DATA ANALYSIS EXAMPLES. UCLA: Statistical Consulting Group, stats.idre.ucla.edu/r/dae/robust-regression/. Accessed 31 Oct. 2017.