Linear Regression

Truong Thi An Hai

11/29/2020

Gather and Prepare Data

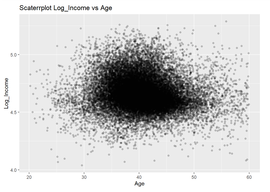
```
data = read.csv("C:/Users/hp/Desktop/zeta.csv")
#Remove all meanhouseholdincome duplicates (only females records should be in the dataset)
data = subset(data, data$sex == "F")
#Remove the columns zcta and sex
data = subset(data, select = -c(zcta, sex))
#Remove outliers
### < meaneducation < 18
data = subset(data, meaneducation <18 & meaneducation >6)
#### 200 < meanhouseholdincome < 200,000
data <- subset(data, meanhouseholdincome <200000 & meanhouseholdincome >10000)
### < meanemployment < 3
data <- subset(data, meanemployment <3 & meanemployment >0)
#### < meanage < 60
data <- subset(data, meanage <60 & meanage >20)
#Create a variable called log_income = log10(meanhouseholdincome)
data$log_income <- log10(data$meanhouseholdincome)
#Rename the columns
names(data)[names(data)=="meanage"] <- "age"
names(data)[names(data)=="meaneducation"] <- "education"
names(data)[names(data)=="meanemployment"] <- "employment"
```

Linear Regression Analysis

a. Create a scatter plot showing the effect age has on log_income and paste it here. Do you see any linear relationship between the two variables?

```
library(ggplot2)

ggplot(data,aes(x= age, y=log_income)) +geom_point(alpha=0.2) +labs(x="Age",y="Log_Income",title="Scatterplot Log _Income vs Age")
```




```
#correlation
cor(data$age, data$log_income)
```

```
## [1] -0.108893
```

From the scatter plot We can see, there seems to appear to be a very weak inverse linear relationship between the two variables.
In addition, the correlation corr- 0.108803 between the two variables is low, indicating that there is only a weak relationship between them.

b. Create a linear regression model between log_income and age. What is the interpretation of the t-value? What kind of t-value would indicate a significant coefficient?

```
linearMod <- lm(log_income ~ age, data)
print(linearMod)
```

```
##
## Call:
## lm(formula = log_income ~ age, data = data)
##
## Coefficients:
## (Intercept)          age
##     4.78774     -0.003074
```

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = log_income ~ age, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.65733 -0.08296 -0.01620  0.07178  0.67202
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7877484  0.0064657  740.5   <2e-16 ***
## age         -0.0030730  0.0001584  -19.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1366 on 31427 degrees of freedom
## Multiple R-squared:  0.01184,    Adjusted R-squared:  0.01181
## F-statistic: 376.5 on 1 and 31427 DF,  p-value: < 2.2e-16
```

The t-value tests whether or not there is a statistically significant relationship between the dependent variable and the independent variable, that is whether or not the beta coefficient of the independent variable is significantly different from zero.

Mathematically, for a given beta coefficient (b), the t-test is computed as t = (b - 0)/SE(b), where SE(b) is the standard error of the coefficient b. The t-value measures the number of standard deviations that b is away from 0. The higher the t-value, the more significant independent variable.

In our exercise, both the t-values for the intercept and age are highly significant, which means that there is a significant association between age and income.

c. What is the interpretation of the R-squared value? What kind of R-squared value would indicate a good fit?

The R-squared value is a goodness of fit measure. The R-squared ranges from 0 to 1 (i.e. a number near 0 represents a regression that does not explain the variance in the dependent variable well and a number close to 1 does explain the observed variance in the dependent variable).

$$R^2 = 1 - \frac{SSE}{SST}$$

where, SSE is the *sum of squared errors* given by $SSE = \sum_i^n (y_i - \hat{y}_i)^2$ and $SST = \sum_i^n (y_i - \bar{y}_i)^2$
is the *sum of squared total*. Here, $\hat{y}_i$ is the fitted value for observation i and $\bar{y}$ is the mean of Y.

A high value of R-squared is a good indication.
In our exercise, the R-squared we get is 0.01184. Or roughly 1.2% of the variance found in the dependent variable (income) can be explained by the independent variable (age).

d. What is the interpretation of the F-statistic? What kind of F-statistic indicates a strong linear regression model?

F-statistic is a good indicator of whether there is a relationship between our independent and the dependent variables. The further the F-statistic is from 1 the better it is. However, how much larger the F-statistic needs to be depends on both the number of data samples and the number of model parameters.

$$F = \frac{\frac{\sum_i \hat{y}_{model}^2}{p-1}}{\frac{\sum_i y_{data}^2}{n-p}}$$

Formula of F-statistic

The F-statistic is used to determine if the model is actually doing better than just guessing the mean value of y as the prediction (the "null model").

If the linear model is really just estimating the same as the null model, then the F-statistic should be about 1.

A F-statistic that is much larger than 1 indicates a strong linear regression model.

e. View a detailed summary of the previous model. What is the R-squared value? Does this suggest that the model is a good fit? Why?

```
summary(linearMod)
```

```
##
## Call:
## lm(formula = log_income ~ age, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.65733 -0.08296 -0.01620  0.07178  0.67202
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.7877484  0.0064657  740.5   <2e-16 ***
## age         -0.0030730  0.0001584  -19.4   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1366 on 31427 degrees of freedom
## Multiple R-squared:  0.01184,    Adjusted R-squared:  0.01181
## F-statistic: 376.5 on 1 and 31427 DF,  p-value: < 2.2e-16
```

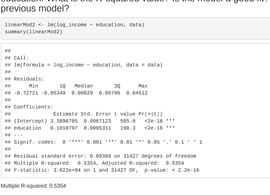
Multiple R-squared:0.01184

Adjusted R-squared: 0.01181

This R-squared value is very far from 1 and near to 0 suggests that the model is not a good fit.

f. Create a scatter plot showing the effect education has on log_income. Do you see any linear relationship between the two variables?

```
ggplot(data,aes(x= education, y=log_income)) +geom_point(alpha=0.2) +labs(x="Education",y="Log_Income",title="Sca tterplot Log_Income vs Education")
```



This scatter plot seems to suggest that there is some sort of linear relationship between the two variables.The intercept seems to be positive.

g. Analyze a detailed summary of a linear regression model between log_income and education. What is the R-squared value? Is the model a good fit? Is it better than the previous model?

```
linearMod2 <- lm(log_income ~ education, data)
summary(linearMod2)
```

```
##
## Call:
## lm(formula = log_income ~ education, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.72721 -0.05349  0.00029  0.05796  0.64512
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.3896705  0.0067123   505.0   <2e-16 ***
## education   0.1018797  0.0005311   190.3   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09369 on 31427 degrees of freedom
## Multiple R-squared:  0.5354, Adjusted R-squared:  0.5354
## F-statistic: 3.622e+04 on 1 and 31427 DF,  p-value: < 2.2e-16
```

Multiple R-squared: 0.5354

Adjusted R-squared: 0.5354

This R-squared value is much closer to 1 than our first model and suggests that the model is a decent fit. It is a better fit than the first model.

h. Analyze a detailed summary of a linear regression model between the dependent variable log_income, and the independent variables age, education, and employment. Is this model a good fit? Why? What conclusions can be made about the different independent variables?

```
linearMod3 <- lm(log_income ~ education + age + employment, data)
summary(linearMod3)
```

```
##
## Call:
## lm(formula = log_income ~ education + age + employment, data = data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.70315 -0.05623  0.00060  0.05213  0.64821
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.5123333  0.0078328  448.21   <2e-16 ***
## education    0.0912653  0.0005980  152.61   <2e-16 ***
## age         -0.0026030  0.0001109  -23.48   <2e-16 ***
## employment  0.0663722  0.0019559   33.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.09017 on 31425 degrees of freedom
## Multiple R-squared:  0.5697, Adjusted R-squared:  0.5697
## F-statistic: 1.387e+04 on 3 and 31425 DF,  p-value: < 2.2e-16
```

This model appears to be a good, but not perfect, fit because the R-squared value is somewhat close to 1.

The F-statistic is much larger than 1, and the p-value is extremely small, which indicates a strong model.

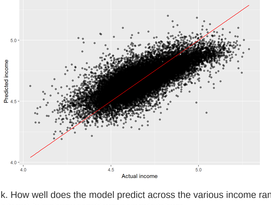
The independent variable age seems to have the weakest linear relationship because its coefficient and t-value are small.

i. Based on the coefficients of the multiple regression model, by what percentage would income increase/decrease for every unit of education completed, while all other independent variables remained constant?

For every unit of education completed, income increase 9.13%.

j. Create a graph that contains a y = x line and uses the multiple regression model to plot the predicted data points against the actual data points of the training set.

```
ggplot() + geom_point(aes(x= data$log_income, y=fitted(linearMod3)), alpha=0.5) + geom_line(aes(x=data$log_income e, y= data$log_income), col = 'red') +labs(x="Actual income", y="Predicted income")
```



k. How well does the model predict across the various income ranges?

In the graph, for lower incomes our model seems to over predict the income.

For higher incomes, our model seems to slightly under predict the income.

This graph indicates that our model provides reliable predictions around the median income range.

*** THE END ***