APS606 - Lecture Note (Week 5)

Norman Lo, M.S. in Quantitative Economics

9/30/2020

Introduction to Regression Analysis

I. Covariance & Correlation Coefficient II. Simple Linear Regression Model III. Interpret and Evaluate the Estimated Parameters IV. Evaluate the Overall Model Fit V. Multiple Linear Regression Analysis

I. Covariance & Correlation Coefficient

In the previous sections, we have examined numerical methods used to summarize the data for one variable at a time. Often a researcher is interested in the relationship between two variables. Two descriptive measures of the relationship between two variables are **covariance** and **correlation coefficient**.

Covariance: The covariance is a measure of the linear association between two variables. Positive values indicate a positive relationship. Negative values indicate a negative relationship.

The covariance is computed as follows:

For populations:

$$\sigma_{xy} = \frac{\sum(x_i - \mu_x)(y_i - \mu_y)}{N}$$

For samples:

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

Example:

Calculate the covariance of the following sample dataset.

x	y
5	10
7	17
4	9
9	18
2	8

Step 1: Calculate the average of x ($\bar{x}$) and y ($\bar{y}$).

```r
# Create the x and y variables
x <- c(5, 7, 4, 9, 2)
y <- c(10, 17, 9, 18, 8)

# Calculate the mean for x and y
xBar <- mean(x)
yBar <- mean(y)

# Print the result
print(paste("The average of x is ", xBar , "and the average of y is ", yBar))
```

```
## [1] "The average of x is  5.4 and the average of y is  12.4"
```

Step 2: Calculate the deviations of x and y.

```r
# Calculate the deviation of x and y
xDev <- x - xBar
xDev
```

```
## [1] -0.4  1.6 -1.4  3.6 -3.4
```

```r
yDev <- y - yBar
yDev
```

```
## [1] -2.4  4.6 -3.4  5.6 -4.4
```

Step 3: Sum the multiplication of the two deviations

```r
# Calculate the sum of the deviations
sumDev <- sum(xDev*yDev)
sumDev
```

```
## [1] 48.2
```

Step 4: Caculate the covariance by dividing the sum of the deviations by the degree of freedom (n - 1)

```r
# Calculate the covariance for x and y
cov <- sumDev / (5 - 1)

# Print the result
print(paste("The covariance of x and y is ", cov))
```

```
## [1] "The covariance of x and y is  12.05"
```

Correlation Coefficient: Correlation is a measure of linear association and not necessarily causation. Just because two variables are highly correlated, it does not mean that one variable is the cause of the other. The coefficient can take on values between -1 and +1. Values near -1 indicate a strong negative linear relationship. Values near +1 indicate a strong positive linear relationship. The closer the correlation is to zero, the weaker the relationship.

The correlation coefficient is computed as follows:

For populations:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

For samples:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Example:

Calculate the correlation coefficient of the following sample dataset.

x	y
5	10
7	17
4	9
9	18
2	8

Step 1: Calculate the covariance of x and y, s_{xy}, which we already did in the previous example.

$$s_{xy} = 12.05$$

Step 2: Calculate the sample standard deviation for x (s_x) and y (s_y).

```
# Calculate the standard deviation for x and y
xSD <- round(sd(x), 3)
ySD <- round(sd(y), 3)

# Print the result
print(paste("The standard deviation for x is ", xSD, "and the standard deviation for y is ", ySD))
```

```
## [1] "The standard deviation for x is  2.702 and the standard deviation for y is  4.722"
```

Step 3: Calculate the correlation coefficient

```
# Calculate the correlation coefficient
corr <- round(cov / (xSD * ySD), 3)

# Print the result
print(paste("The correlation coefficient for x and y is ", corr))
```

```
## [1] "The correlation coefficient for x and y is  0.944"
```

Correlation between sets of data is a measure of how well they are related. However, it does not provide us the effect size of a variable to the variable of interest. For instance, we may know the sales of a company is highly correlated to their advertising expenses. But how much the sales will be effected by a marginal change of the advertising expenses? We need a new tool to model this kind of relationship and that's what we are going to discover in the next section, the "**Simple Linear Regression Model**".

II. Simple Linear Regression Model

The goal in this section is to introduce linear regression, the standard tool that statisticians rely on when analysing the relationship between interval scale predictors and interval scale outcomes. Stripped to its bare essentials, linear regression models are basically a slightly fancier version of the Pearson correlation though as we'll see, regression models are much more powerful tools.

The aim of linear regression is to model a continuous variable y as a mathematical function of one or more x variable(s), so that we can use this regression model to predict the y when only the x is known. This mathematical equation can be generalized as follows:

$$y = \beta_0 + \beta_1(x) + \epsilon$$

where, β_0 is the intercept and β_1 is the slope. Collectively, they are called regression coefficients. ϵ is the error term, the part of y the regression model is unable to explain.

Model Assumptions

1. Linear in Parameter: The relationship between x and the mean of y is linear
2. Zero Conditional Mean: The error ϵ is a random variable with mean of zero
3. Homoskedasticity: The variance of ϵ, denoted by σ^2, is the same for all values of the independent variable
4. Independence of Errors: There is not a relationship between the residuals and the independent variable
5. Normality: The error ϵ is a normally distributed random variable

Least Squares Method:

To find the best fit for a set of data points, we, generally, apply a statistical procedure called "**least squares method**". Teh least squres method is a procedure for using sample data to find the esimtated regression coefficients by minimizing the sum of the offsets or residuals of points (y_i) from the fitted curve ($\hat{y}_i$, which is called the "**least squares criterion**".

The least square method uses the sample data to provide the values of β_0 and β_1 that minimize the sum of the squares of the deviations between the observed values of the dependent variable, y_i, and the predicted values of the dependent variable, $\hat{y}_i$.

Least Squares Criterion:

$$min \sum (y_i - \hat{y}_i)^2$$

where

y_i is the observed value of the dependent variable for the i^{th} observation.
$\hat{y}_i$ is the perdicted value of the dependent variable for the i^{th} observation.

Estimated y-Intercept (β_0) and Slope (β_1):

Using the differential calculus we can derive the estimated y-intercept, β_0, and slope, β_1, of the linear equation.

$$\beta_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$$

$$\beta_0 = \bar{y} - \beta_1(\bar{x})$$

where

x_i = value of the independent variable for the i^{th} observation
y_i = value of the dependent variable for the i^{th} observation
$\bar{x}$ = mean value for the independent variable
$\bar{y}$ = mean value for the dependent variable
n = total number of observations

Let's demonstrates the algorithem in this simple example:

$\bar{x} = 7$
$\bar{y} = 17$

Observation	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
1	10	22	3	5	15	9
1	5	14	-2	-3	6	4
1	6	15	-1	-2	2	1

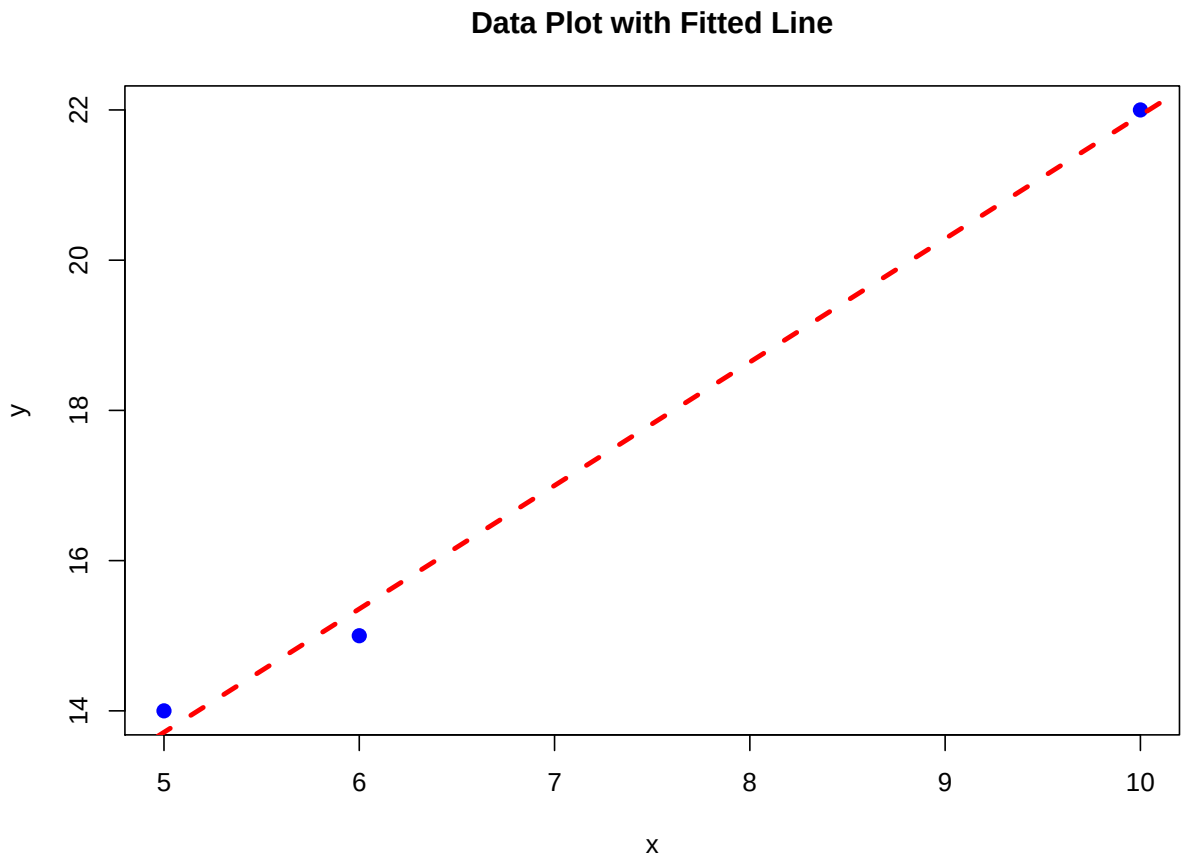
$$\beta_1 = \frac{23}{14} = 1.643$$

$$\beta_0 = 17 - 1.643(7) = 5.5$$

The estimated regression equation is

$$\hat{y} = 5.5 + 1.643(x)$$

Graphical Solution:



III. Interpret and Evaluate the Estimated Parameters

Testing for Significance

In some cases, the mean value of y does not depend on the value of x and hence we would conclude taht x and y are not linearly related. Alternatively, if the value of β_1 is not equal to zero, we would conclude that the two variables are related. To test for a significant regression relationship, we must conduct a hypothesis test to determine whether the value of β_1 is zero. Two tests are commonly used. Both require an estimate of σ^2, the variance of ϵ in the regression model.

t Test: (Test of the Estimated Parameter)

t test is designed to hypothesize about the value of β_1 and then use statistical inference to test our hypothesis.

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Estimate of σ^2:

Mean Square Error (Estimate of σ^2)

$$\hat{\sigma}^2 = MSE = \frac{SSR}{Degrees\ of\ Freedom} = \frac{SSR}{n-2}$$

where

MSE = mean square error or estimated variance of the residual, σ^2
SSR = residual sum of squares, $\sum (y_i - \hat{y}_i)^2$
Degrees of Freedom = total observations minus total number of variable in the model, $n - 2$

Standard Error of the Estimate ($\hat{\sigma}$)

$$\hat{\sigma} = \sqrt{\hat{\sigma}^2} = \sqrt{MSE}$$

The properties of the sampling distriubtion of the least squares estimateor β_1 provide the basis for the hypothesis test. Generally speaking, the t test for a significant relationship is basedo n the fact that the test statistic follows a t distriubtion with n - 2 degrees of freedom.

Sampling Distribution of β_1

$$E(\beta_1) = \hat{\beta}_1$$
$$\sigma_{\beta_1} = \frac{\sigma}{\sqrt{\sum (x_i - \bar{x})^2}}$$

Since population σ is not known, we develop an estimated of σ_{β_1} by standard error of the estimate, $\hat{\sigma}$.

$$\hat{\sigma}_{\beta_1} = \frac{\hat{\sigma}}{\sqrt{\sum (x_i - \bar{x})^2}}$$

The test statistic follows a distribution with n - 2 degrees of freedom.

$$t = \frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_{\beta_1}}$$

where

β_1 is the hypothesized value, $H_0 : \beta_1 = 0$
$\hat{\beta}_1$ is the least squares estimated parameter
$\hat{\sigma}_{\beta_1}$ is the standard error of the estimated parameter

Rejection Rule:

1. p-value Approach: Reject H_0 if p-value $\leq \alpha$
2. Critical Value Approach: Reject H_0 if $t \leq -t_{\alpha/2}$ *or* $if\ t \geq t_{\alpha/2}$

where $t_{\alpha/2}$ is based on a t distribution with n - 2 degrees of freedom.

Confidence Interval for β_1:

$$\beta_1 \pm t_{\alpha/2}\hat{\sigma}_{\beta_1}$$

F test (Test of Significance in Regression)

With only one independent variable, the F test will provide the same conclusion as the t test; that is, if the t test inducates $\beta_1 \neq 0$ and hence a significant relationship, the F test will also inducate a significant relationship. But with more than one independent variable (next section, multiple regression model), only the F test can be used to test for an overall signficant relationship.

$$H_0 : \beta_1 = \beta_2 = ... = \beta_j = 0$$
$$H_a : At\ least\ one\ \beta\ is\ not\ zero$$

F Statistic:

$$F = \frac{(SSR_r - SSR_{ur})/q}{SSR_{ur}/(n-k-1)}$$

where

SSR_r is the sum of squares residuals from the restricted model (all βs equal 0)
SSR_{ur} is the sum of squares residuals from the unrestricted model
q is numerator degrees of freedom = $df_r - df_{ur}$ $n-k-1$ is the denominator degrees of freedom = df_{ur}

Fortunately, most of the modern statistical softwares report the F statistic of the regression model, so we are not going into the details to explain the F distribution and its properties in this course.

Rejection Rule:

1. p-value Approach: Reject H_0 if p-value $\leq \alpha$
2. Critical Value Approach: Reject H_0 if $F \geq F_\alpha$

where F_α is based on an F distriubtion with mode at 1, large value of F lead to the rejection of H_0 and the conclusion that the relationship between x and y is statistically significant.

IV. Evaluate the Overall Model Fit

Measuring the Goodness of Fit (Coefficient of Determination)

The least square method allows us to find the best estimated parameters for the regression model. However, how well does the estimated regression equation fit the data? In this section, we introduce three measures of from the estimated regression model.

1. Total Sum of Squares: A measure of the error in using the estimated regression equation to predict the value of the dependent variable in the sample.

$$SST = \sum (y_i - \bar{y}_i)^2$$

2. Explained Sum of Squares: A measure of how well the observations cluster about the $\hat{y}$ fitted line.

$$SSE = \sum (\hat{y}_i - \bar{y})^2$$

3. Residual Sum of Squares: A measure of the error in using the estimated regresion equation to predict the values of the dependent variable in the sample.

$$SSR = \sum (y_i - \hat{y}_i)^2$$

As you can imagine, these three measures are related and the relationship among these three sums of squares provides one of the most important results in statistics. The following equation shows that the total sumof squares can be partitioned into two components, the explained sum of squares (due to the model) and residual sum of square (due to the random error).

$$SST = SSE + SSR$$

If we divide the explained sum of squares (SSE) by the total sum of squares (SST), it turns out the ratio is the explained variation from the model compared to the total variation of the data; thus, it is interpreted as the fraction of the sample variation in y that is explained by x, sometimes we called this the **coefficient of determination** or r^2, is dfined as

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

The value of r^2 is always between zero and one, because SSE can be no greater than SST. When interpreting r^2, we usually multiply it by 100 to change it into a percent, so we can say "*the precentage of the sample variation in y that is explained by x*".

If the data points all lie on the same line, OLS provides a perfect fit to the data. In this case, $r^2 = 1$. A value of r^2 that is nearly equal to zero indicates a poor fit of the OLS line: very little of the variation in the y_i is captured by the variation in the $\hat{y}_i$.

Example:

In this example, we are using the USArrests dataset from base R library. You can check the documentation about the dataset by the code "*?USArrests*". Suppose we are interested to explain the murder rate at the state level using the linear regression model. Let's take a look of the dataset here.

```
# Loading the Base R USArrests dataset
head(USArrests)
```

```
##            Murder Assault UrbanPop Rape
## Alabama      13.2     236       58 21.2
## Alaska       10.0     263       48 44.5
## Arizona       8.1     294       80 31.0
## Arkansas      8.8     190       50 19.5
## California    9.0     276       91 40.6
## Colorado      7.9     204       78 38.7
```

```
# Statistic Summary of the dataset
summary(USArrests)
```

```
##      Murder          Assault         UrbanPop          Rape
##  Min.   : 0.800   Min.   : 45.0   Min.   :32.00   Min.   : 7.30
##  1st Qu.: 4.075   1st Qu.:109.0   1st Qu.:54.50   1st Qu.:15.07
##  Median : 7.250   Median :159.0   Median :66.00   Median :20.10
##  Mean   : 7.788   Mean   :170.8   Mean   :65.54   Mean   :21.23
##  3rd Qu.:11.250   3rd Qu.:249.0   3rd Qu.:77.75   3rd Qu.:26.18
##  Max.   :17.400   Max.   :337.0   Max.   :91.00   Max.   :46.00
```

```r
# Dimension of the dataset
d <- dim(USArrests)

print(paste("The dataset has ", d[1], "rows and ", d[2], "columns."))
```

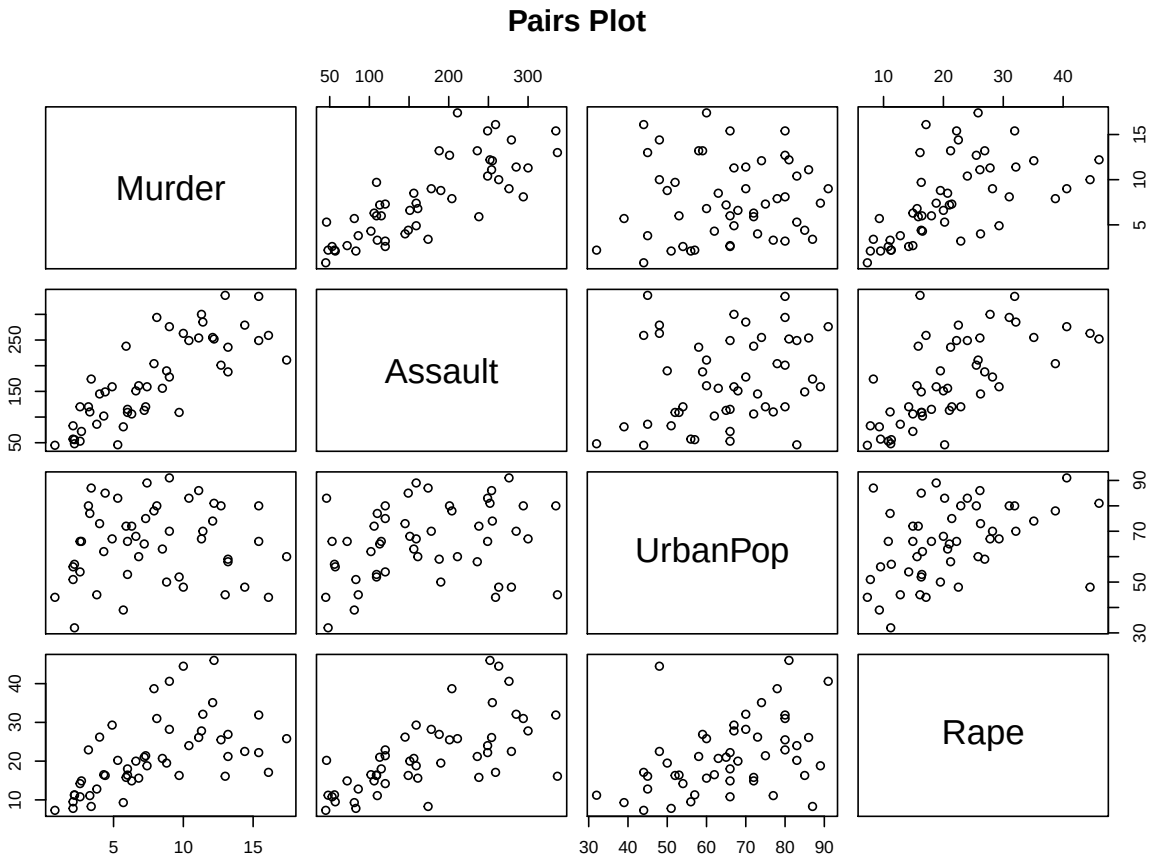
```
## [1] "The dataset has  50 rows and  4 columns."
```

The next step before building the linear model is to check the correlation between each variable to the murder rate in the dataset.

```r
# Create the correlation table
cor(USArrests)
```

```
##              Murder   Assault   UrbanPop      Rape
## Murder   1.00000000 0.8018733 0.06957262 0.5635788
## Assault  0.80187331 1.0000000 0.25887170 0.6652412
## UrbanPop 0.06957262 0.2588717 1.00000000 0.4113412
## Rape     0.56357883 0.6652412 0.41134124 1.0000000
```

```r
# Using the pairs plot to identify any trend in the data
pairs(USArrests,
      main = "Pairs Plot")
```



According to the correlation table and pairs plot, we observe a strong correlation between "Murder" and "Assault", so the next step is to build the simple linear regression model and layout the model specification. Here is what we are trying to create in R.

$$Murder = \beta_0 + \beta_1(Assault) + \epsilon$$

Once we state the model specification, then we can start building the model in R.

```
# Creat the simple linear regression model
fit1 <- lm(USArrests$Murder ~ USArrests$Assault, data = USArrests)

# Print the model summary
summary(fit1)
```

```
##
## Call:
## lm(formula = USArrests$Murder ~ USArrests$Assault, data = USArrests)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8528 -1.7456 -0.3979  1.3044  7.9256
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.631683   0.854776   0.739    0.464
## USArrests$Assault 0.041909   0.004507   9.298  2.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.629 on 48 degrees of freedom
## Multiple R-squared:  0.643,  Adjusted R-squared:  0.6356
## F-statistic: 86.45 on 1 and 48 DF,  p-value: 2.596e-12
```

Model Summary:

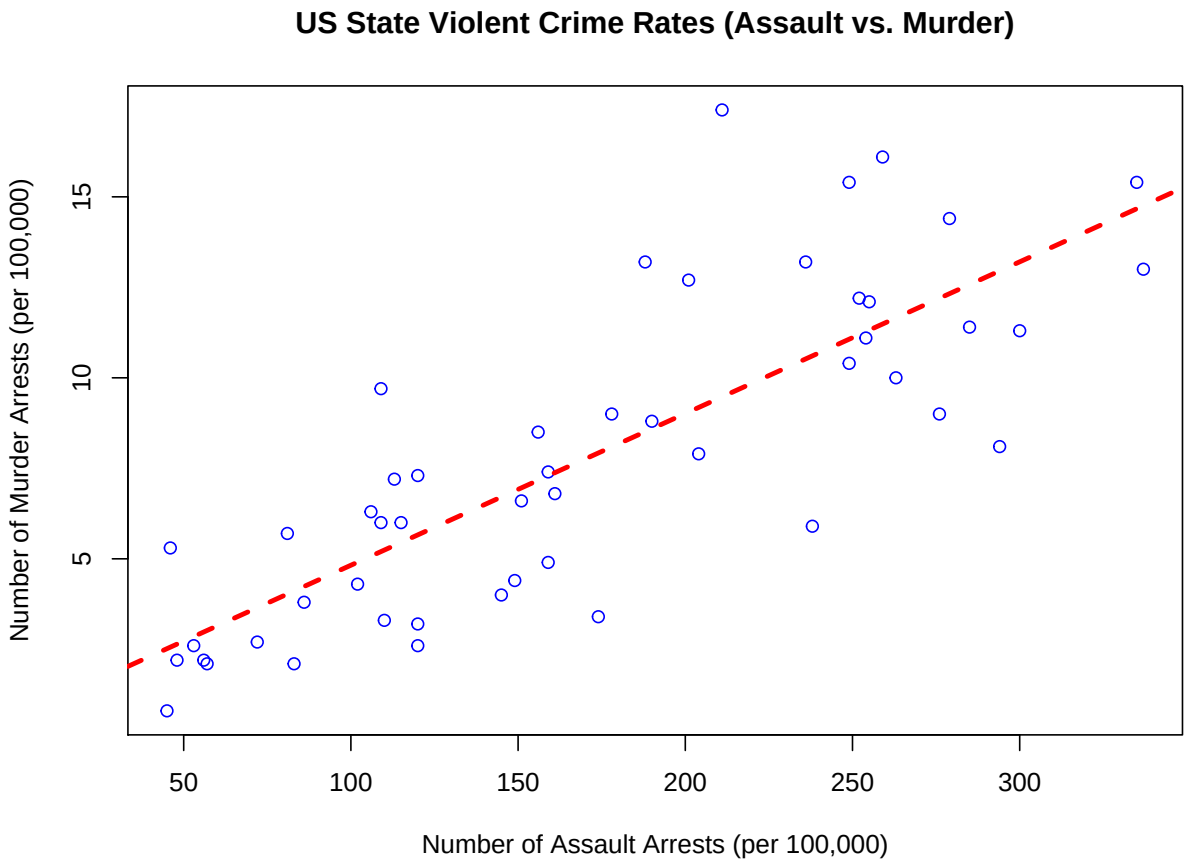
Interpretation of the Coefficient: The estimated coefficient for "Assualt" is 0.0419, which means for every 1 additional assault arrest (per 100,000), the murder arrest (per 100,000) will increase by 0.0419 case.

Evaluation of the Coefficient: The p-value for the estimated coefficient for "Assualt" is very small (at least less than 0.01), which means the coefficient is at least 99% statisticaly significant in this model.

Evaluation of the Model: The p-value for the F test is very small (at least less than 0.01), which means at least one of the estimated coefficients in the model is not equal to zero.

Model Explanatory Power: The adjusted R-square is 0.6356, which means this model can explain 63.56% of the variation of murder arrests (per 100,000) in the U.S.

Visualization of the Model:




```
## integer(0)
```

V. Multiple Linear Regression Analysis

Multiple regression analysis isthe study of how a dependent variable y is realted to two or more independent variables. In the general case, we will use j to denote the number of independent variables.

The concept of a regression model and a regression equation introduced in the preceding section are applicable in the multiple regression case. The equation that describes how the dependent variable y is related to the independent variables $x_1, x_2, x_3, ...x_j$ and an error term is called the **multiple regression model**. We begin with the assumption that the multiple regression model takes the following form.

$$y = \beta_0 + \beta_1(x_1) + \beta_2(x_2) + ... + \beta_j(x_j) + \epsilon$$

Model Assumptions

Basically all the assumptions we mentioned in the simle regression model apply to the multiple regression model. Since we have multiple independent variables in a model, we need to add one more assumption about the relationship between the independent variables.

6. No Perfect Collinearity: The least squares method does not allow the independent variables perfectly correlated. If two independent variables are perfectly or highly correlated, the model cannot be estimated by the least squares method and we say the model suffers **multicollinearity**.

Estimated Multiple Regression Equation

Generally speaking, we can apply the least squares method to develop the estimated multiple regresion equation that best approximated the linear relationship between the dependetn and independent variables.

Least Squares Criterion:

$$min \sum (y_i - \hat{y}_i)^2$$

Multiple Coefficient of Determination

In the simple regression, we showed that the total sum of squares can be partitioned into two components: the sum of squares due to regression and the sum of squares due to random error. The same procedure applies to the sum of squares in multiple regression.

$$SST = SSE + SSR$$

where

SST = Total Sum of Squares, $SST = \sum (y_i - \hat{y}_i)^2$

SSE = Explained Sum of Squares, $SSE = \sum (\hat{y}_i - \bar{y})^2$

SSR = Residual Sum of Squares, $\sum (y_i - \hat{y}_i)^2$

For simple regression, we use the coefficient of determination, $r^2 = \frac{SSR}{SST}$, to measure the goodness of fit for the estimated regression equation. The same concept applies to multiple regression. The term **multiple coefficient of determination** indicates that we are measuring the goodness of fit for the estiated multiple regression equation. The multiple coefficient of determination, denoted R^2, is computed as follows:

$$R^2 = \frac{SSR}{SST}$$

Generally speaking, as more indpendent variables are added to the model, more of the variation of the dependent variable, y, should be explained. However, the model is also being punished by reducing the degrees of freedom. Many analysts perfer adjusted R^2 for the number of independent variables to avoid overestimating the impact of adding an independent variable on the amount of variability explained by the estimated regression equation.

Adjusted Multiple Coefficient of Determination

$$R^2_{adj} = 1 - (1 - R^2)\frac{n-1}{n-p-1}$$

The R^2_{adj} should always be smaller than the R^2 because of the variable adjustment. However, the interpretation of the statistic is the same as its percedent.

Testing for Significance

In simple linear regression, the F and t tests provide the same conclusion. In multiple regression, the F and t tests have different purposes.

The **t test** is used to determine whether each of the individual independent variables is significant. A separate t test is conducted for each of the independent variables in the model. We refer to each of these t tests as a test for individual significance.

Hypotheses:

$$H_0 : \beta_j = 0$$
$$H_a : \beta_j \neq 0$$

Test Statistics:

$$t = \frac{\hat{\beta}_j - \beta_j}{\hat{\sigma}_{\beta_j}}$$

Rejection Rule:

p-value Approach: Reject H_0 if p-value $\leq \alpha$
Critical Value Approach: Reject H_0 if $t \leq -t_{\alpha/2}$ *or if* $t \geq t_{\alpha/2}$

The **F test** is used to determine whether a significant relationship exists between the dependent variable and the set of all the independent variables. The F test is referred to as the test for overall significance.

Hypotheses:

$$H_0 : \beta_1 = \beta_2 = ... = \beta_j = 0$$
$$H_a : \textit{At least one } \beta \textit{ is not zero}$$

Test Statistics:

$$F = \frac{MSR}{MSE}$$

Rejection Rule:

p-value Approach: Reject H_0 if p-value $\leq \alpha$
Critical Value Approach: Reject H_0 if $F \geq F_\alpha$

Example:

Let's use the previous example dataset "USArrests" to build a multiple linear regression model. In the previous example, we use only "Assault" to explain "Murder". May be we can try to add "Rape" into our model to see if it helps explain the State level murder rate.

```
# Creat the simple linear regression model
fit2 <- lm(USArrests$Murder ~ USArrests$Assault + USArrests$Rape, data = USArrests)

# Print the model summary
summary(fit2)
```

```
##
## Call:
## lm(formula = USArrests$Murder ~ USArrests$Assault + USArrests$Rape,
##     data = USArrests)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.8667 -1.7653 -0.3746  1.3030  7.8864
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       0.418863   0.976184   0.429    0.670
## USArrests$Assault 0.040029   0.006087   6.576 3.59e-08 ***
## USArrests$Rape    0.025142   0.054157   0.464    0.645
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.651 on 47 degrees of freedom
## Multiple R-squared:  0.6446, Adjusted R-squared:  0.6295
## F-statistic: 42.63 on 2 and 47 DF,  p-value: 2.76e-11
```

Model Summary:

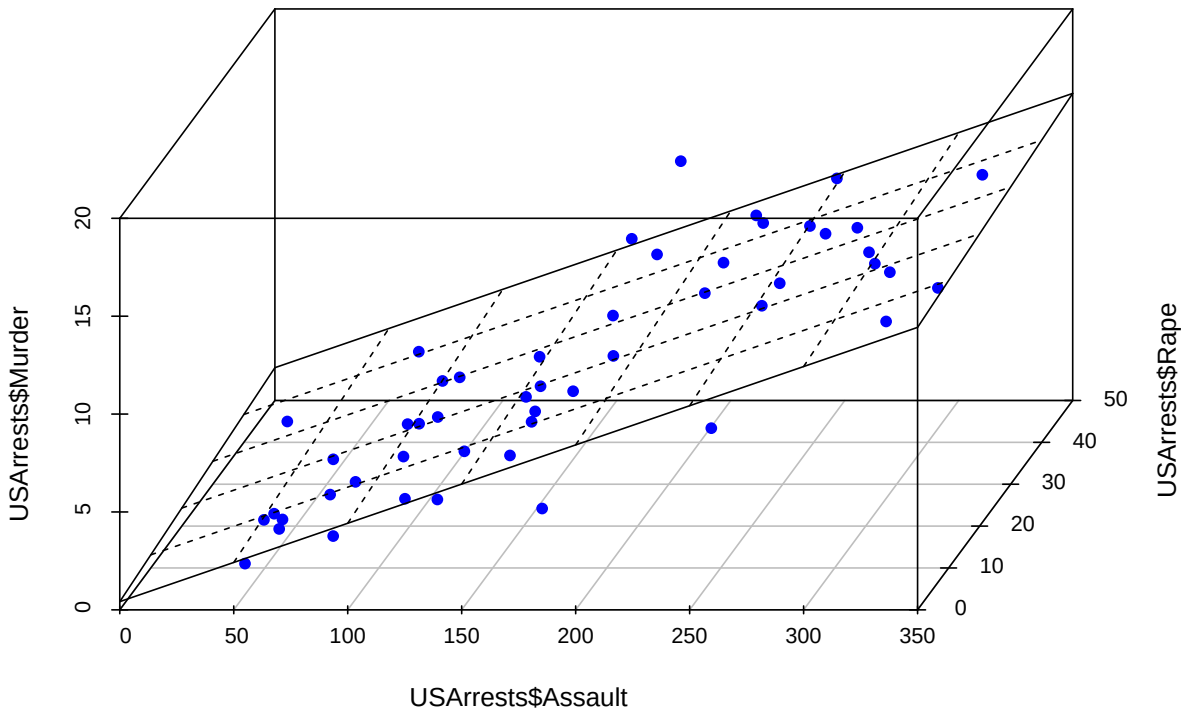
Interpretation of the Coefficient: The estimated coefficient for "Assualt" is 0.0419, which means for every 1 additional assault arrest (per 100,000), the murder arrest (per 100,000) will increase by 0.04 case. The coefficient for "Rape" is 0.0251, which suggests for every 1 additional rape arrest (per 100,000), the murder arrest (per 100,000) will increase by 0.0251 case.

Evaluation of the Coefficient: The p-value for the estimated coefficient for "Assualt" is very small (at least less than 0.01), which means the coefficient is statisticaly significant at 1% signficiant level in this model. However, the p-value for "Rape" coefficient is very high (at least greater than 0.10), which means the coefficient fails to be statistically significant at 10% significant level in this model.

Evaluation of the Model: The p-value for the F test is very small (at least less than 0.01), which means at least one of the estimated coefficients in the model is not equal to zero.

Model Explanatory Power: The adjusted R-square is 0.6295, which means this model can explain 62.95% of the variation of murder arrests (per 100,000) in the U.S.

Visualization of the Model:


Regression Plane
USArrests$Murder
0
5
10
15
20
0
50
100
150
200
250
300
350
USArrests$Assault
0
10
20
30
40
50
USArrests$Rape