BUSAD40 - Busniess Statistics

Lecture Note 4

Norman Lo

Fall 2020

I believe there is a direct correlation between love and laughter.

- Yakov Smirnoff

Linear Regression Analysis

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.

Simple Linear Regression

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

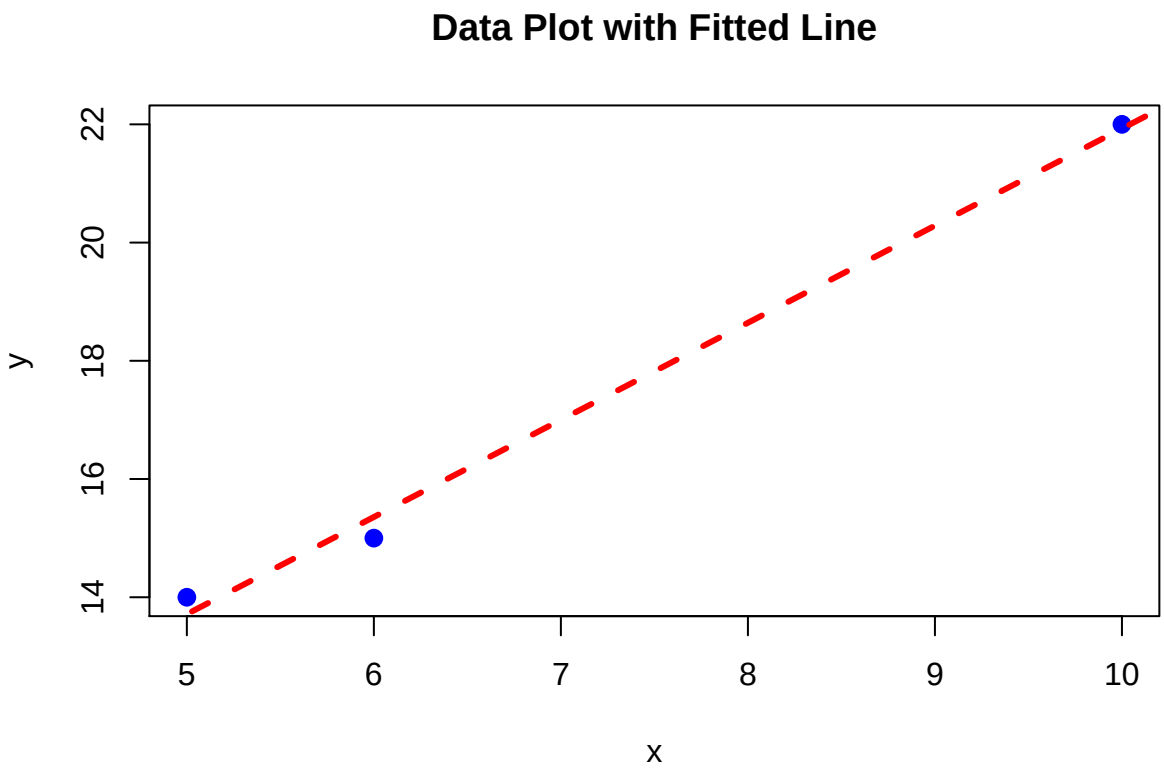
$$\beta_1 = \frac{53}{14} = 3.786$$

$$\beta_0 = 17 - 3.786(7) = -9.502$$

The estimated regression equation is

$$\hat{y} = -9.502 + 3.786(x)$$

Graphical Solution:



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.

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

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 **cofficient 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 $\hat{y}_i$.

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.

Example:

Given the *Advertising* data set, which consists of the sales of the product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: *TV*, *radio*, and *newspaper*.

Suppose it's not possible for our clients to directly increase sales of the product. On the other hand, they can control the advertising expenditure in each of the three media. Therefore, if we determine that there is an association between advertising and sales, then we can instruct our client to adjust advertising budgets, thereby indirectly increasing sales. Our goal is to develop an accurate model that can be used to predict samesl on the basis of the three media budgets.

Loading the Data:

```
## Parsed with column specification:
## cols(
##   TV = col_double(),
##   radio = col_double(),
##   newspaper = col_double(),
##   sales = col_double()
## )
```

```
## [1] "TV"        "radio"     "newspaper" "sales"
```

```
# Print the first 6 rows of the data
head(advertise)
```

```
## # A tibble: 6 x 4
##      TV radio newspaper sales
##   <dbl> <dbl>     <dbl> <dbl>
## 1 230.   37.8      69.2  22.1
## 2  44.5  39.3      45.1  10.4
## 3  17.2  45.9      69.3   9.3
## 4 152.   41.3      58.5  18.5
## 5 181.   10.8      58.4  12.9
## 6   8.7  48.9      75     7.2
```

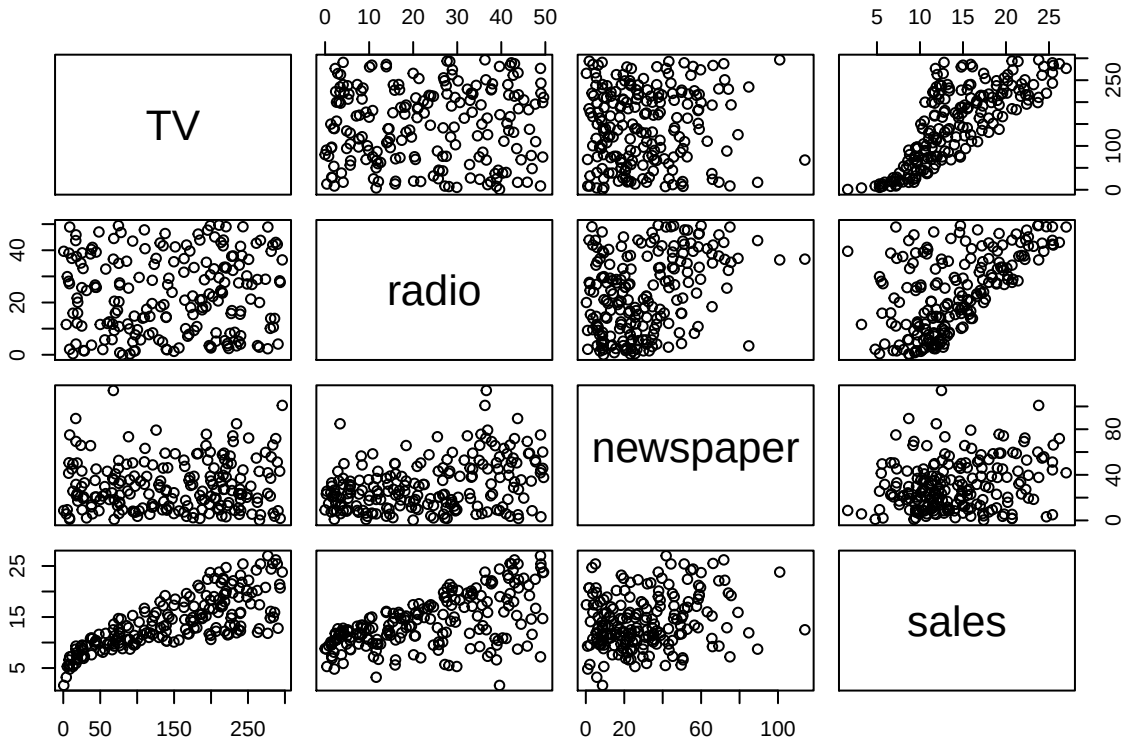
Descriptive Statistics and Basic Visualization:

To have a better idea of the data set, we can apply some base R functions to explore the data.

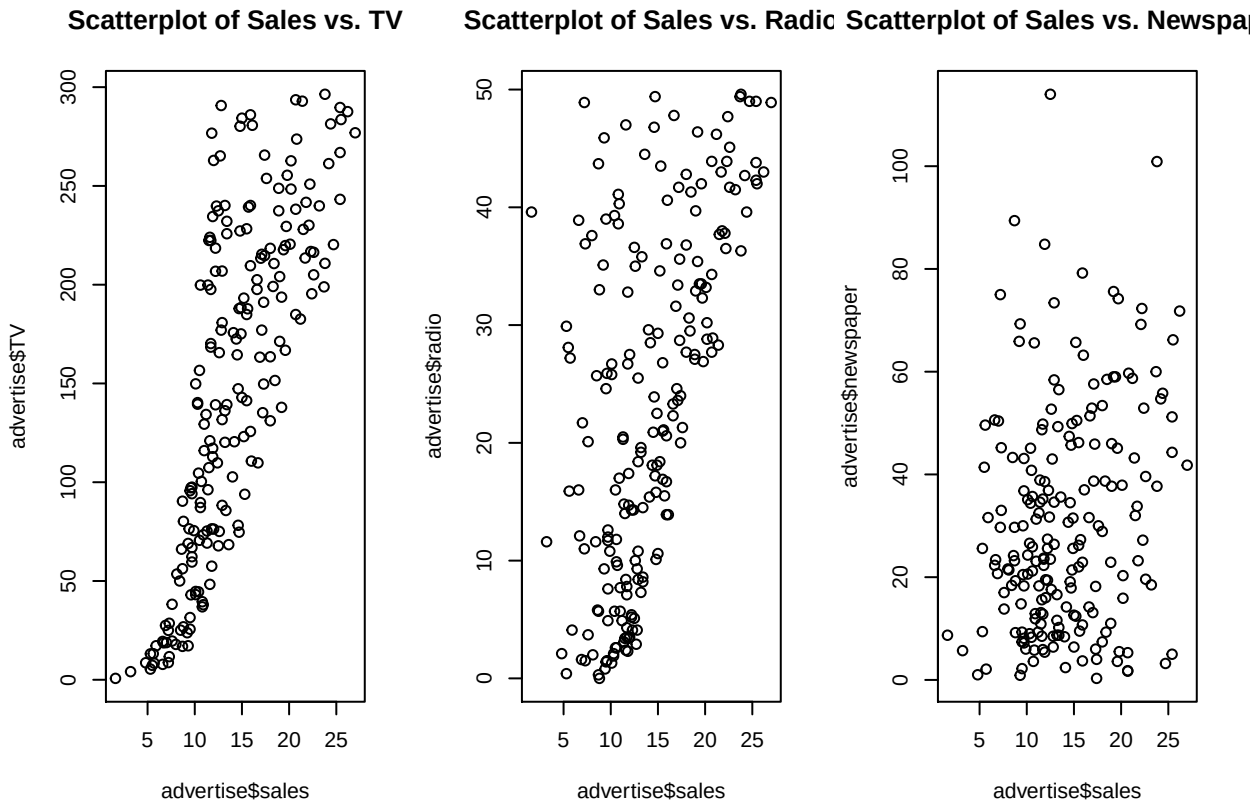
```r
# Descriptive summary of the data set
summary(advertise)
```

```
##        TV              radio          newspaper          sales
##  Min.   :  0.70   Min.   : 0.000   Min.   :  0.30   Min.   : 1.60
##  1st Qu.: 74.38   1st Qu.: 9.975   1st Qu.: 12.75   1st Qu.:10.38
##  Median :149.75   Median :22.900   Median : 25.75   Median :12.90
##  Mean   :147.04   Mean   :23.264   Mean   : 30.55   Mean   :14.02
##  3rd Qu.:218.82   3rd Qu.:36.525   3rd Qu.: 45.10   3rd Qu.:17.40
##  Max.   :296.40   Max.   :49.600   Max.   :114.00   Max.   :27.00
```

```r
# Pairs Plot
pairs(advertise)
```




```r
# Plot the TV, radio, and newspaper against sales
par(mfrow=c(1,3))
plot(advertise$sales, advertise$TV, main="Scatterplot of Sales vs. TV")
plot(advertise$sales, advertise$radio, main="Scatterplot of Sales vs. Radio")
plot(advertise$sales, advertise$newspaper, main="Scatterplot of Sales vs. Newspaper")
```




```
# Check the correlation between the three advertising expenditures with sales
cor(advertise$sales, advertise$TV)
```

```
## [1] 0.7822244
```

```
cor(advertise$sales, advertise$radio)
```

```
## [1] 0.5762226
```

```
cor(advertise$sales, advertise$newspaper)
```

```
## [1] 0.228299
```

```
# Note that the cor() function does not return result for categorical variables, such as Gender, size,
```

Linear Regression Model:

As observed from the visualization, we notice there is a clear linear relationship between Sales and TV expenditure. We are going to fit the data with the linear regression model and interpret the results. Furthermore, we are going to evaluate the model by different statistics.

```
# Fitting the simple linear regression model
fit1 <- lm(sales ~ TV, data = advertise)

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

```
##
## Call:
## lm(formula = sales ~ TV, data = advertise)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## TV          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 on 198 degrees of freedom
## Multiple R-squared:  0.6119, Adjusted R-squared:  0.6099
## F-statistic: 312.1 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
## [1] "Intercept is 7.033 and slope is 0.048 ."
```

Estmiated regression Sales = 7.033 + 0.0475(TV)
Slope (0.0475): For each additonal $1,000 spent in TV advertising, the sales would go up by $47.5.
Intercept (7.033): If the advertising expediture is $0, the sales would be $7.033.

How good is this regression line?

```
## [1] "The R square of the model is  0.611875050850071"
```

r^2 is 0.6119, which means 61.19% of variation in sales (dependent variable) in the sample can be explained by the TV advertising budget.

Is Our Linear Regression Model Significant Statistically?

t Test:

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

From the result of summary(fit1), we know the t test returns a p-value $p = 2e - 16$.

The estimated parameter β_1 is very statistically significant. We can confidently reject the null hypothesis that $\beta_1 = 0$.

F test:

H_0: The regression model does not explain any of the variation of sales ($\beta_1 = 0$).
H_a: The regression model does explain a proportion of the variation of sales ($\beta_1 \neq 0$).

From the result of summary(fit1), we know that the F test returns a p-value $p = 2.2e - 16$.

Assume that $\alpha = 0.05$. Since p-value $< \alpha$, reject H_0 at the 0.05 level of significance. Thus, the regression model does explain a proportion of the total variation in the sales amount over the range of values for the advertising amount observed in the sample. In other words, the model is significant at the 0.05 level of significance.

Confidence Interval for the Slope (β_1)

```r
# Creating a 95% confidence interval for the slope parameter
confint(fit1, level = 0.95)
```

```
##                  2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## TV          0.04223072 0.05284256
```

As the TV advertising expenditure increases by $1000, we are 95% sure that the sales would increase between $42.23 and $52.84.

Estimation and Prediction Using Regression Model

How much is your predicted sales if our client puts in an TV advertisment budget of $200? Before the calcuate, first make sure that 4 is between the forecating range (the range of the independent variable).

```r
# Find the range of TV budget spending in the data
range(advertise$TV)
```

```
## [1]   0.7 296.4
```

Method 1:

```r
# Calculate the predicted sales based on $200 spending on TV advertising
predSale <- intercept + slope * 200
print(paste("The predicted sales when spending $200 on TV advertisment is ", predSale))
```

```
## [1] "The predicted sales when spending $200 on TV advertisment is  16.5399216357316"
```

Method 2:

```r
# Using the predict.lm() function to predict the sales based on the given TV budget
predict.lm(fit1, newdata = data.frame(TV = 200))
```

```
##        1
## 16.53992
```

Note: we can also predict for a series of numbers. For example,

```r
# Make prediction for multiple TV budget values
predict.lm(fit1, newdata = data.frame(TV = c(100, 150)))
```

```
##        1        2
## 11.78626 14.16309
```

```r
# Combining the predictions and the given TV budget values
a = data.frame(TV = seq(80, 120, 160))
pred = predict.lm(fit1, newdata = a)
cbind(a, pred)
```

```
##   TV     pred
## 1 80 10.83552
```

Confidence intervals (95%)

```r
# Create 95% confidence intervals for the prediction
predict.lm(fit1, newdata = data.frame(TV = 200), interval = "confidence", level = 0.95)
```

```
##        fit      lwr      upr
## 1 16.53992 16.00567 17.07418
```

95% confidence interval is between $16.01 and $17.07.

Prediction intervals (99%)

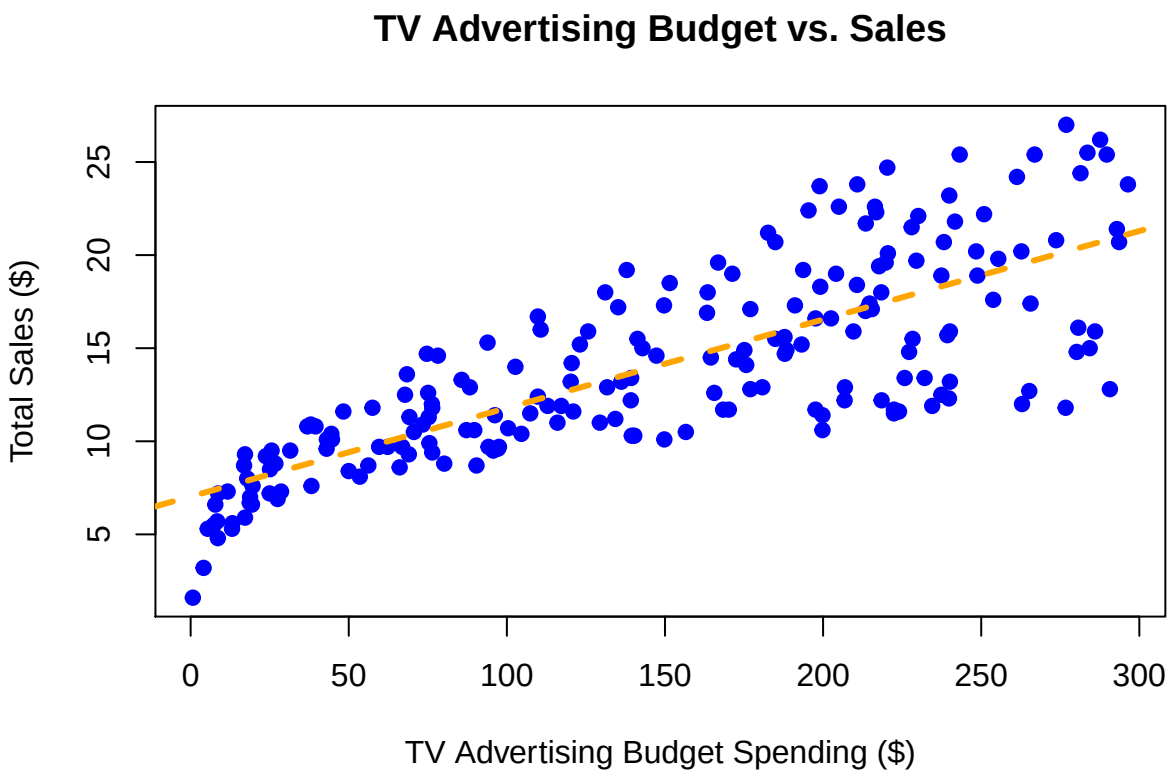
```r
# Create 99% confidence intervals for the prediction
predict.lm(fit1, newdata = data.frame(TV = 200), interval = "prediction", level = 0.99)
```

```
##        fit      lwr      upr
## 1 16.53992 8.035283 25.04456
```

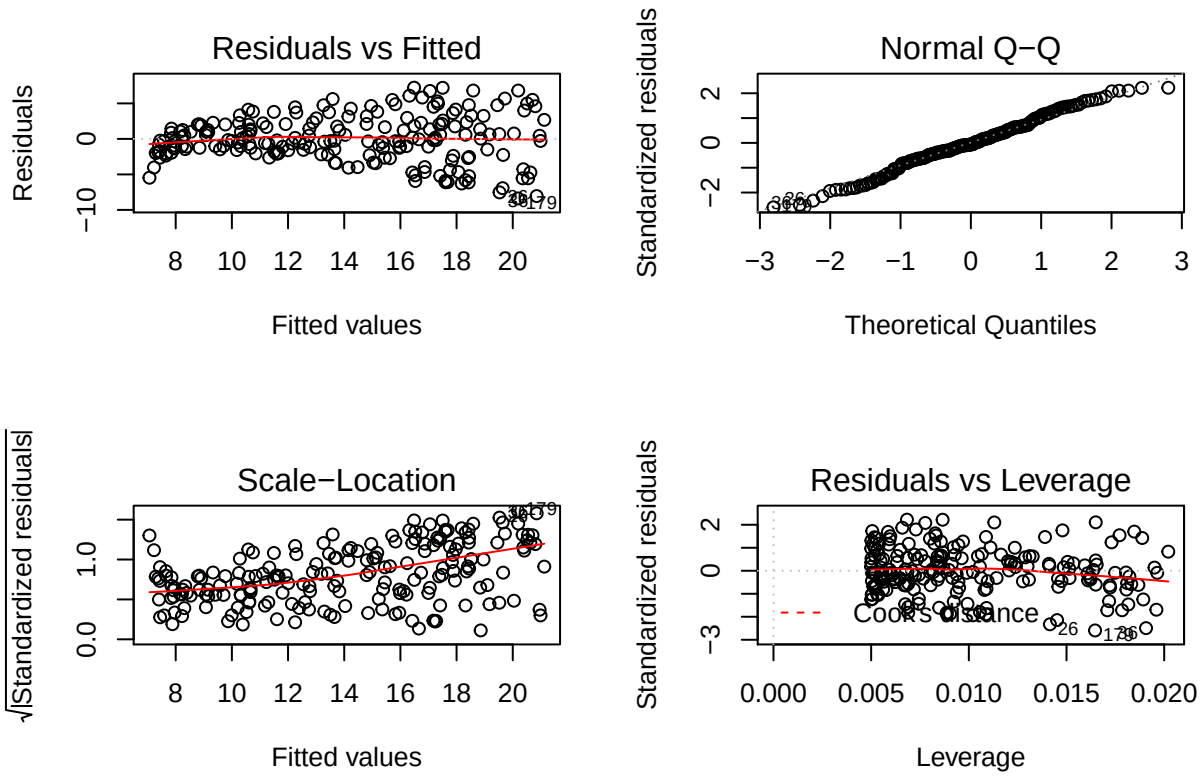
95% prediction interval is between $8.04 and $25.04.

Linear Regression Visualization:

Once we fit the data with the linear regression model, we can visualize the prediction of the model and compare it to the data set.




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



Multiple Linear Regression:

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:

According the previous advertising data set, we can construct a multiple linear regression model with the following specification.

$$sales = \beta_0 + \beta_1(TV) + \beta_2(Radio) + \epsilon$$

Model Fit:

```
# Fitting the multiple regression model
fit2 <- lm(sales~TV+radio, data=advertise)

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

```
##
## Call:
## lm(formula = sales ~ TV + radio, data = advertise)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.7977 -0.8752  0.2422  1.1708  2.8328
##
```

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.92110    0.29449   9.919   <2e-16 ***
## TV           0.04575    0.00139  32.909   <2e-16 ***
## radio        0.18799    0.00804  23.382   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.681 on 197 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8962
## F-statistic: 859.6 on 2 and 197 DF,  p-value: < 2.2e-16
```

Model Visualization:

Making prediction for sales based on the final fitted model.

```
# Make a prediction of sales with 99% confidence intervals based on our fitted model.
sales_hat <- predict(fit2, interval="confidence", level=0.99)

# Print the first 3 predicted sales based on TV and Radio
head(sales_hat, 3)
```

```
##        fit      lwr      upr
## 1 20.55546 20.03755 21.07338
## 2 12.34536 11.74601 12.94472
## 3 12.33702 11.58565 13.08838
```