# Data603: Statistical Modelling with Data

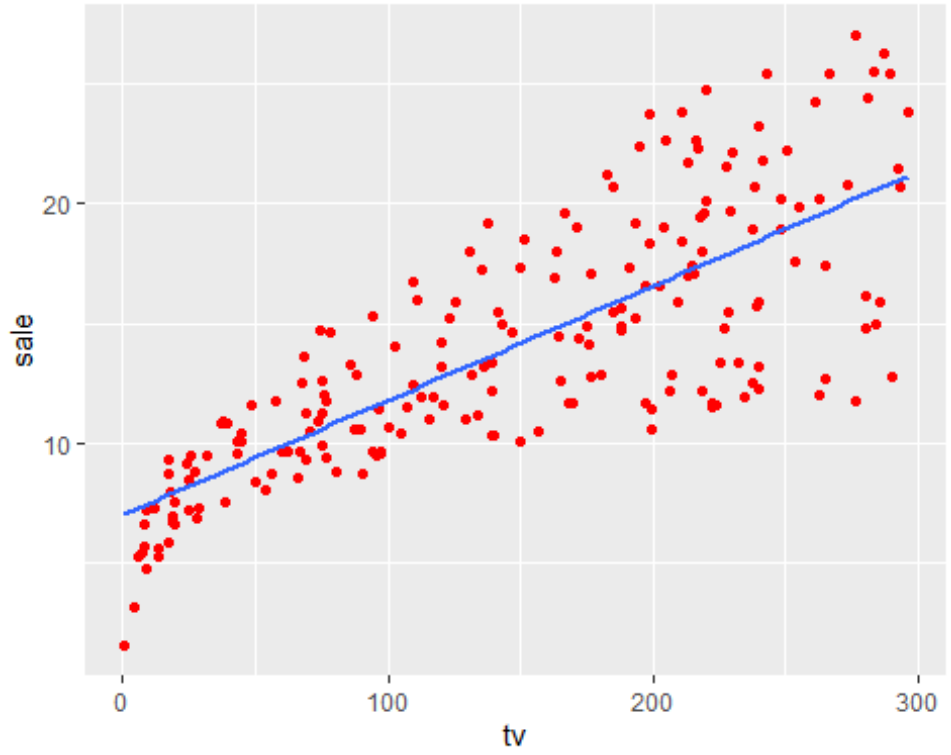
## MULTIPLE LINEAR REGRESSION

## PART I: FIRST ORDER MODELS WITH QUANTITATIVE INDEPENDENT VARIABLES

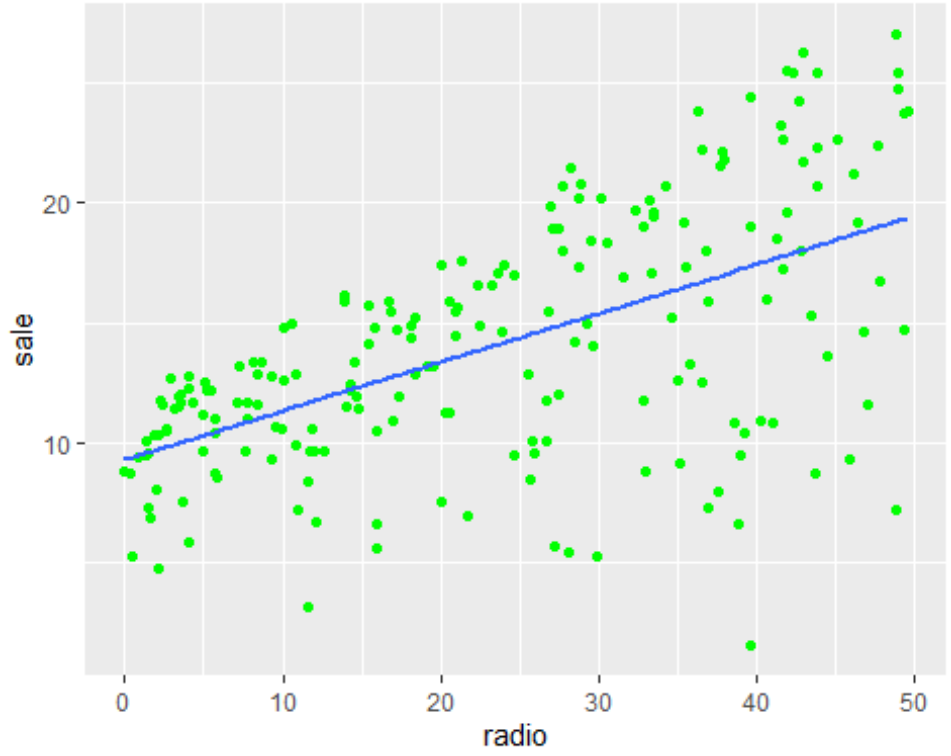
Suppose that we are statistical consultants hired by a client to provide advice on to improve sales of a particular product. The advertising data set consists of the *Sales* (in thousands of units) of that product in 200 different markets, along with advertising budgets(in thousands of dollars) for the product in each of those markets for three different media: *TV*, *radio*, and *newspaper*.

In this setting, the advertising budgets(TV, radio,and newspaper) are independent variables(predictor variables) and Sales are the dependent variable(response variable). The least square fit for the simple linear regressions of sales onto TV,radio, and newspapers are shown as following,

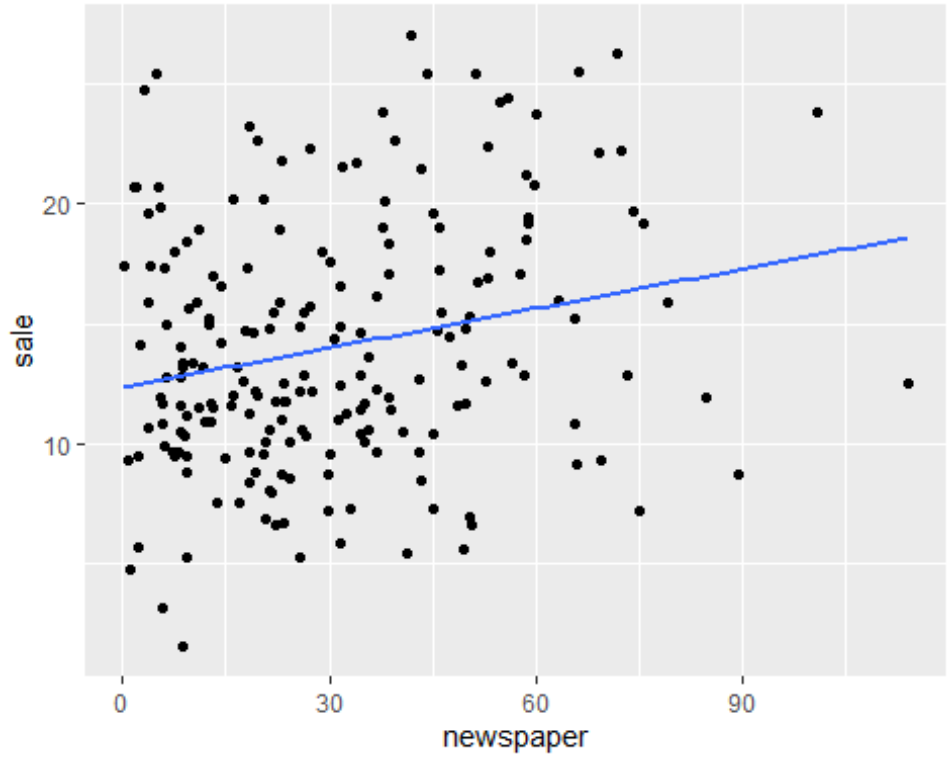
```
library(ggplot2)  #using ggplot2 for data visualization

Advertising=read.table("/Users/thuntida.ngamkham/OneDrive - University of Cal
gary/dataset603/Advertising.txt",header = TRUE,  sep ="\t" )
#attach(Advertising)


ggplot(data=Advertising,mapping= aes(x=tv,y=sale))+geom_point(color='red')+
  geom_smooth(method = "lm", se = FALSE)
```




```
ggplot(data=Advertising,mapping= aes(x=radio,y=sale))+geom_point(color='green
')+
  geom_smooth(method = "lm", se = FALSE)
```



```r
ggplot(data=Advertising,mapping= aes(x=newspaper,y=sale))+geom_point(color='b
lack')+
  geom_smooth(method = "lm", se = FALSE)
```




```r
summary(lm(sale~tv,data=Advertising))
```

```
##
## Call:
## lm(formula = sale ~ tv, data = Advertising)
##
## 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
```

```r
summary(lm(sale~radio,data=Advertising))
```

```
##
## Call:
## lm(formula = sale ~ radio, data = Advertising)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.7305  -2.1324   0.7707   2.7775   8.1810
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  9.31164    0.56290  16.542   <2e-16 ***
## radio        0.20250    0.02041   9.921   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.275 on 198 degrees of freedom
## Multiple R-squared:  0.332,  Adjusted R-squared:  0.3287
## F-statistic: 98.42 on 1 and 198 DF,  p-value: < 2.2e-16
```

```
summary(lm(sale~newspaper,data=Advertising))
```

```
##
## Call:
## lm(formula = sale ~ newspaper, data = Advertising)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -11.2272  -3.3873  -0.8392   3.5059  12.7751
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.35141    0.62142   19.88  < 2e-16 ***
## newspaper    0.05469    0.01658    3.30  0.00115 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.092 on 198 degrees of freedom
## Multiple R-squared:  0.05212,    Adjusted R-squared:  0.04733
## F-statistic: 10.89 on 1 and 198 DF,  p-value: 0.001148
```

$$\widehat{Sale} = 7.032594 + 0.047537tv$$
$$\widehat{Sale} = 9.31164 + 0.20250radio$$
$$\widehat{Sale} = 12.35141 + 0.05469newspaper$$

Simple linear regression is a useful approach for predicting a response on the basis of a single predictor variable. However, it is unclear how to make a single prediction of sales given levels of three advertising media budgets, since each of the budgets is associated with a separate regression equation.

*R functions:*

*ggplot() : is used to construct the initial scatter plot.*

*geom_point(): the point geom is used to create scatterplots.*

*geom_smooth() : aids the eye in seeing patterns in the presence of overplotting.*

# INTRODUCTION

Does a regression with one independent variable even make sense? It does or It does not. The world is might too complex a place for simple regression alone to model it, **A Regression with two or more independent variables is called Multiple Regression.** It can be looked upon as an extension of straight-line regression analysis (which involves only one independent variable) to the situation where more than one independent variable must be considered.Dealing with several independent variables simultaneously in a regression analysis is considerably more difficult than dealing with a single independent variable, for following reasons:

1. It is more difficult to choose the best model, since several reasonable candidates may exist.
2. It is more difficult to visualize what the best fitted model looks like (especially if there are more than two independent variables), since it is not possible to plot either the data or the fitted model directly in more than three dimensions.
3. Computations are virtually impossible without access to a high speed computer and a reliable packaged computer program.

# PART I: FIRST ORDER MODELS WITH QUANTITATIVE INDEPENDENT VARIABLES

## The General Multiple Linear Regression Model

A model that includes only terms denoting quantitative independent variable, called a **first-order model**,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon$$

*where*

$Y$ = the dependent variable

$X_1, X_2, \ldots, X_p$ = the independent variables, predictors

$E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p$ the deterministic portion of the model

$\beta_i$ = regression coefficients, $i = 1, \ldots, p$

From the Advertising example, Instead of fitting a separate simple linear regression model for each predictor, a better approach is to extend to the multiple linear regression model so that it can directly accommodate multiple predictors.

$$Sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspapers + \epsilon$$

## Estimating the Regression point estimates

Since we cannot know the true values of the parameters $\beta_0, \beta_1, \ldots, \beta_p$ relating $\mu$ to $X_1$, $X_2$, $X_3$,...,$X_p$ in the regression model

$$\begin{aligned} Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \epsilon \\ &= \mu + \epsilon \\ &where \\ \mu &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p \end{aligned}$$

By using the method of least squares, the estimated model is

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p$$

We can calculate the least squares point estimates $\hat{\beta}_0, \hat{\beta}_1, \widehat{\beta_2}, \ldots, \widehat{\beta_p}$ of the parameters $\beta_0, \beta_1, \ldots, \beta_p$ in the model by using the following matrix algebra formula.

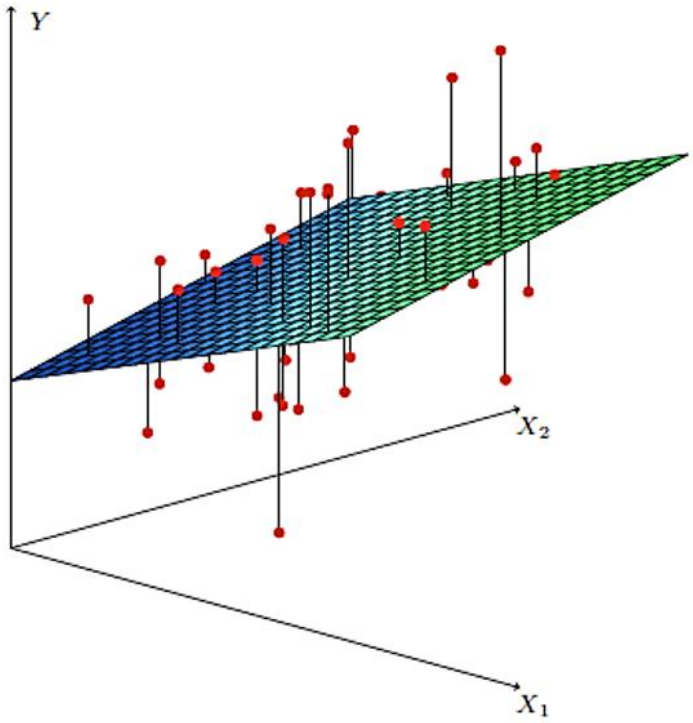
$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ . \\ . \\ . \\ \hat{\beta}_p \end{bmatrix} = \hat{\beta} = (X'X)^{-1}X'y$$

where $y$ and $X$ are the following column vector and matrix respectively:

$$y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ . \\ . \\ . \\ y_p \end{bmatrix} \text{ and } X = \begin{pmatrix} 1 & x_{11} & x_{12} & . & . & . & x_{1p} \\ 1 & x_{21} & x_{22} & . & . & . & x_{2p} \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ . & . & . & . & . & . & . \\ 1 & x_{n1} & x_{n2} & . & . & . & x_{np} \end{pmatrix}$$

The **least squares estimates (LSE)** $\hat{\beta}_0, \hat{\beta}_1, \widehat{\beta_2}, \ldots, \widehat{\beta_p}$ are obtained by **minimizing the sum of the squared residuals:**

$$\text{SSE} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_p X_{pi}))^2$$




*Regresson line become a plane for 2 independent variables*

*In a three-dimensional setting, with 2 independent variables and one dependent varibale, the least square regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.*

The values $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$ are called **the multiple least squares regression coefficient estimates.**

## Point Estimating for Multiple Regression Coefficients

In this class, we use a statistical *R* program on a computer to do all the calculations for multiple regression coefficients.

### What is the meaning of these point estimates?

### Interpreting the Intercept

$\widehat{\beta_0}$, the y-intercept, can be interpreted as the value you would predict for y when all $X_1, X_2, \ldots X_p = 0$.

### Interpreting Coefficients of Predictor Varibales

$\widehat{\beta_i}$, the regression coefficient, describes how much change in response y for every unit change in $X_i$ when other predictor variables are held constant.

From the Advertising example, the following code displays the multiple regression coefficient estimates when TV, radio, and newspaper advertising budgets are used to predict product sales.

```
reg1<-lm(sale~tv+radio+newspaper, data=Advertising)
coefficients(reg1)
```

```
##  (Intercept)           tv        radio    newspaper
##  2.938889369  0.045764645  0.188530017 -0.001037493
```

*R codes:*

*lm() : "linear model" is used to create a simple or multiple regression model.*

*coefficients(): is used to extract model coefficients from a simple or multiple regression model.*

The estimated model is $\widehat{Sale} = 2.939 + 0.046tv + 0.189radio - 0.001newspaper$

We interpret these results as following:

$\widehat{\beta_1} = 0.046$ means that for a given amount of radio and newspaper advertising, spending additional \$1,000 on TV advertising leads to an *increase* in sales by approximately 46 units.

$\widehat{\beta_2} = 0.189$ means that for a given amount of TV and newspaper advertising, spending additional \$1,000 on radio advertising leads to an *increase* in sales by approximately 189 units.

$\widehat{\beta_3} = -0.001$ means that for a given amount of TV and radio advertising, spending additional \$1,000 on newspapers advertising leads to a *decrease* in sales by approximately 1 unit !!!!

## Interval Estimate for Multiple Regression Coefficients (Confidence Interval for the individual regression coefficient)

$$A\ 100\ (1-\alpha)\%\ Confidence\ Interval\ for\ parameter\ \beta_i\ is\ \widehat{\beta_i} \pm t_{\alpha/2} S_{\widehat{\beta_i}}$$

where

$n =$ number of observations

$p =$ number of regression coefficients

```
reg1<-lm(sale~tv+radio+newspaper, data=Advertising)
confint(reg1) # a 95% confidence interval for coefficients
```

```
##                   2.5 %     97.5 %
## (Intercept)  2.32376228 3.55401646
## tv           0.04301371 0.04851558
## radio        0.17154745 0.20551259
## newspaper   -0.01261595 0.01054097
```

```r
confint(reg1, level = 0.99) # a 99% confidence interval for coefficients
```

```
##                   0.5 %     99.5 %
## (Intercept)  2.12757072 3.75020802
## tv           0.04213632 0.04939297
## radio        0.16613095 0.21092909
## newspaper   -0.01630884 0.01423386
```

*R functions:*

*confint() :Computes a 95 % confidence interval for one or more parameters in a fitted model.*

*confint(model, level=...) :Computes a specific confidence interval for one or more parameters in a fitted model.*

From the Advertising example, the output displays the multiple regression 95% confidence Interval for coefficient estimates when TV, radio, and newspaper advertising budgets are used to predict product sales.

Thus, we can interpret that sales increase between 43.01 units to 48.51 units for every $1000 increase in TV advertising budget, holding radio and newspaper advertising budget (with 95% of chance).

## Inclass Practice Problem

How do real estate agents decide on the asking price for a newly listed condominium? A computer data base in a small community contains the listed selling price Y (in thousand of dollars), the amount of living area (in hundreds of square metres), and the number of floors, bedrooms, and bathroom are recorded for 15 randomly selected condos currently on the market. The data file is provided in **condominium.csv**. Use R software package to fit the model and construct a 95% confidence interval for regression coefficients.

We notice that the multiple regression coefficient estimates for TV and radio are in the same direction but the coefficient estimate for newspaper is close to zero. Moreover, a 95% confidence interval for newspaper also includes zero (-0.0126,0.0105). In this case, is there a relationship between newspaper and sales? In general when we perform multiple regression, we usually are interested in answering a few important questions.

1. Is this multiple regression model any good at all? Is at least one of the predictors useful in predicting the response?

2. Do all the predictors help to explain Y , or is only a subset of the predictors useful?

Now we address these questions as following topics,

## Evaluating Overall Model Utility

## Testing a Relationship Between the Response and Predictors

## Full Model Test

We ask the global question,"Is this multiple regression model any good at all??" The answer is that we can test some hypotheses to see the relationship between the response and predictors. The first of these hypotheses is an overall F-test or a global F test which tells us if the multiple regression model is useful. To address the overall question, we will test

$$H_0 \quad : \beta_1 = \beta_2 = \ldots = \beta_p = 0$$
$$H_a \quad : \text{at least one } \beta_i \text{ is not zero } (i = 1,2,\ldots,p)$$

## The Analysis of Variance for Multiple Linear Regression

**The ANOVA table for Multiple Linear Regression**

| Source of Variation | Df | Sum of Squares | Mean Square | F-Statistic |
|---|---|---|---|---|
| Regression | p | SSR | MSR | MSR/MSE |
| Residual | n-p-1 | SSE | MSE | |
| Total | n-1 | SST | | |

This hypothesis test is performed by computing the F-statistic,

$$F_{cal} = \frac{MSR}{MSE} = \frac{\frac{SSR}{p}}{\frac{SSE}{(n-p-1)}}$$

*where*

Sum of squares for error or residual $= SSE = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$= \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_p X_{pi}))^2$

Sum of squares for regression $= SSR = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$

Total corrected sum of squares of the Y's $= SST = \sum_{i=1}^{n} (y_i - \overline{y})^2$

$n =$ the sample size
$p =$ the number of predictors or the number of regression coefficients

$SST = SSR + SSE$

```
reg1<-lm(sale~tv+radio+newspaper, data=Advertising) # (Full) model with all v
ariables
reg2<-lm(sale~1, data=Advertising) # Model with only intercept
summary(reg1)
```

```
##
## Call:
## lm(formula = sale ~ tv + radio + newspaper, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## tv           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

```r
anova(reg2,reg1) # We compare the NULL model with the full model
```

```
## Analysis of Variance Table
##
## Model 1: sale ~ 1
## Model 2: sale ~ tv + radio + newspaper
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    199 5417.1
## 2    196  556.8  3    4860.3 570.27 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

*R functions:*

*summary() :is used to produce result summaries of the results of various model fitting functions.*

*anova() :is used to compute the analysis of variance (or deviance) tables for one or more fitted model objects.*

**The ANOVA table for Advertising data example**

| Source of Variation | Df | Sum of Squares | Mean Square | F-Statistic |
|---|---|---|---|---|
| **Regression** | **3** | **4860.3** | **1620.1** | **570.295** |
| **Residual** | **196** | **556.8** | **2.84081** | |
| **Total** | **199** | **5417.1** | | |

*Tha Anova table for Advertising data example*

From the Advertising example, the output shows that Fcal=570.3 with df= 3,196 (p-value< 2.2e-16 < $\alpha = 0.05$ ),indicating that we should clearly reject the null hypothesis. It provides compelling evidence against the null hypothesis $H_0$. In other word, the large F-test suggests that at least one of the advertising media must be related to sales. Based on the p-value, we also have extremely strong evidence that at least one of the media is associated with increased sales.

Once we check the overall F-test and reject the null hypothesis, we can move on to checking the test statistics for the individual coefficients and particular subsets of the full model test.

# Partial Test

## Individual Coefficients Test (t-test)

$$H_0 \quad : \beta_i = 0$$
$$H_a \quad : \beta_i \neq 0 \ (i = 1,2,\ldots,p)$$

$$t_{cal} \quad = \frac{\widehat{\beta_i} - \beta_i}{SE(\widehat{\beta_i})} \text{ which has } df = n - p \text{ degree of freedom}$$

```
reg1<-lm(sale~tv+radio+newspaper, data=Advertising)
summary(reg1)
```

```
##
## Call:
## lm(formula = sale ~ tv + radio + newspaper, data = Advertising)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.8277 -0.8908  0.2418  1.1893  2.8292
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.938889   0.311908   9.422   <2e-16 ***
## tv           0.045765   0.001395  32.809   <2e-16 ***
## radio        0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.686 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

From the Advertising example, the output shows that the newspaper has tcal=-0.177 with the p-value= 0.86 > 0.05,indicating that we should clearly not to reject the null hypothesis that the newspaper advertising has not significantly influence on sales at $\alpha = 0.05$.

## Partial F test

The goal is to investigate the contribution of a subset of predictors given that a different set of predictors is already in the model. We define:

Full Model to be the model with the whole set of predictors

Reduced Model to be the model with the whole set of predictors less the subset to be tested.

For example, if we want to test X1 given X2 and X3 are in the model, then the Full Model has the predictors X1 , X2 and X3, and the Reduced Model has the predictors X2 and X3. This will test the effect of X1 in the full model with all 3 predictors.$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$. The hypotheses are:

$$H_0 \quad : \beta_1 = 0 \text{ in the model } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$
$$H_a \quad : \beta_1 \neq 0 \text{ in the model } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$$

In general,to test that a particular subset of q of the coefficients are zero, the hypotheses are

$$H_0 \quad : \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0$$
$$H_a \quad : \text{at least one } \beta_i \neq 0$$

This can be achieved using an F-test. Let SSE(Full model) be the residual sum of squares under the full model and SSE(Reduced model) be the residual sum of squares under the reduced model. Then the F-statistic is

$$F_{cal} = \frac{\frac{SSE_{\text{reduced model}} - SSE_{\text{full model}}}{df_{reduced} - df_{full}}}{\frac{SSE_{\text{full model}}}{df_{full}}}$$

```
full<-lm(sale~tv+radio+newspaper, data=Advertising)
reduced<-lm(sale~tv+radio, data=Advertising) # dropping a newspaper variable
anova(reduced,full) # test if Ho: newspaper = 0
```

```
## Analysis of Variance Table
##
## Model 1: sale ~ tv + radio
## Model 2: sale ~ tv + radio + newspaper
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    197 556.91
## 2    196 556.83  1  0.088717 0.0312 0.8599
```

*R function:*

*anova(reduced,full) :is used to compute the analysis of variance (or deviance) for comparing between reduced and full nodel.*

From the Advertising example, after dropping the variable newspaper off the full model, the reduced output shows that

$$F_{cal} = \frac{\frac{SSE_{\text{reduced model}} - SSE_{\text{full model}}}{df_{reduced} - df_{full}}}{\frac{SSE_{\text{full model}}}{df_{full}}}$$
$$= \frac{(556.9140 - 556.8253)/(197 - 196)}{(556.8253/196)} = 0.031$$

with df=1,196 (p-value=0.8599 > $\alpha = 0.05$ ), indicating that we should clearly not to reject the null hypothesis which mean that we definately drop the variable newspaper off the model.

At this point,

From the initial estimated regression model is $\widehat{Sale} = 2.939 + 0.046tv + 0.189radio - 0.001newspaper$

After checking individual coefficients test, the final regression model is $\widehat{Sale} = 2.92110 + 0.04575tv + 0.18799radio$

## Inclass Practice Problem

From the condominium problem, use the method of Partial F test to fit the model. How many possible fitted models would you suggest for predictive purpose?

## Model Fit

How well does the regression model fit?? Two of the most common numerical measures of model fit are RSE(Residual Standard Error: $s$) and $R^2$ (Coefficient of Determination), the fraction of variation explained. These quantities are computed and interpreted in the same fashion as for simple linear regression.

## $R^2$(the Coefficient of Determination)

Recall that in simple linear regression, $R^2$ is the square of the correlation of the response and the variable. **In multiple regression**, it turns out that it equals to $Cor(Y, \hat{Y})^2$, the square of the correlation between the response and the fitted linear model, $R^2$ is the proportion of the total variation that is explained by the regression model of $Y$ on $X_1, X_2, \ldots X_p$ that is,

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

An $R^2$ value close to 1 indicates that the model explains a large portion of the variance in the response variable. For example, if $R^2$ is 0.7982 for the model, then 79.82% of the variation of the response variable is explained by the model.

It turns out that $R^2$ will always increase when more variables are added to the model, even if those variables are only weakly associated with the response. To compensate for this one can define **an adjusted coefficient of determination**, $R^2_{adj}$

$$R^2_{adj} = 1 - \frac{\frac{SSE}{n-p-1}}{\frac{SST}{n-1}}$$

```r
full<-lm(sale~tv+radio+newspaper, data=Advertising)
reduced<-lm(sale~tv+radio, data=Advertising)
summary(full)$r.squared
```

```
## [1] 0.8972106
```

```r
summary(reduced)$r.squared
```

```
## [1] 0.8971943
```

```r
summary(full)$adj.r.squared
```

```
## [1] 0.8956373
```

```r
summary(reduced)$adj.r.squared
```

```
## [1] 0.8961505
```

*R functions:*

*summary(model)$r.squared :extract the coefficient of determination from the r.squared attribute of its summary*

*summary(model)$adj.r.squared :extract the coefficient of determination from the adj.r.squared attribute of its summary*

From the Advertising example, the model containing all predictors has a $R^2_{adj} = 0.8956$. In contrast, the model that contains only TV and radio as predictor has a $R^2_{adj} = 0.8962$ . This implies that a model that uses TV and radio expenditures to predict sales is substantially better than one that use the full model.

## The estimation of Standard error of residuals

One way to assess strength of fit is to consider how far off the model is for a typical case. That is, for some observations, the fitted value will be very close to the actual value, while for others it will not. The magnitude of a typical residual can give us a sense of generally how close our estimates are. Some of the residuals are positive, while others are negative.Thus, it makes more sense to compute the square root of the mean squared residual and to make this estimate unbiased, we have to divide the sum of the squared residuals by the degrees of freedom in the model. In general, RMSE or $s$ is defined as

$$s = \quad RMSE = \sqrt{\frac{1}{n-p-1}SSE} = \sqrt{MSE}$$

*where*

$$SSE \quad = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_p X_{pi}))^2$$

RMSE can be interpreted as the standard deviation of the unexplained variance, and has the useful property of being in the same units as the response variable. **Lower values of RMSE indicate better fit.**

```
full<-lm(sale~tv+radio+newspaper, data=Advertising)
reduced<-lm(sale~tv+radio, data=Advertising)
sigma(full) # RMSE for the full model

## [1] 1.68551

sigma(reduced) # Rmse for the reduced model

## [1] 1.681361
```

*R functions:*

*sigma(model) :extract the "standard error of residuals" from a fitted model).*

Looking at the reduced model that contains only TV and radio as predictors has an RMSE of 1.681, and the model that also contains newspaper as a predictor (full model) has an RMSE=1.686. This corroborates our previous conclusion that a model that uses TV and radio expenditures to predict sale is much more accurate than one that use the full model. Therefore, there is no point in using newspaper spending as a predictor in the model.

In many computer printouts and textbooks, $s^2$ is called the mean square for error (MSE). This estimate of $s^2 = MSE = \frac{1}{n-p-1}SSE$. The units of the estimated variance are squared units of the dependent variable y. Since the dependent variable y in the adverstising data example is sales in units, the units of $s^2$ are units$^2$. This makes meaningful interpretation of $s^2$ difficult, so we use the standard deviation $s$ to provide a more meaningful measure of variability.

***Output from the full model,***

*Residual standard error: 1.686 on 196 degrees of freedom*

*Multiple R-squared: 0.8972, Adjusted R-squared: 0.8956*

***Output from the reduced model***

*Residual standard error: 1.681 on 197 degrees of freedom*

*Multiple R-squared: 0.8972, Adjusted R-squared: 0.8962*

##Model Prediction Once we have fit the multiple regression model, it is straightforward to predict the response Y on the basis of a set of values for the predictors $X_1, X_2, \ldots X_p$. We usually use a **prediction interval** to predict the response y

```
reduced<-lm(sale~tv+radio, data=Advertising)
newdata = data.frame(tv=200, radio=20)
predict(reduced,newdata,interval="predict")
```

```
##        fit     lwr     upr
## 1 15.83195 12.5042 19.1597
```

*R function predict() : use for prediction of the response. We also set the interval type as "predict", and use the default 0.95 confidence level*

The 95% confidence interval of the sale with the given parameters is between 12.5042 (thousand units )and 19.1597(thousandunits) when the TV and Radio advertising budgets are 200 thousand dollars and 20 thousand dollars, respectively.

## Inclass Practice Problem

From the condominium problem, use the method of Model Fit to calculate $R^2_{adj}$ and RMSE for all possible models. Which model or set of models would you suggest for predictive purpose?

## Exercise1

The amount of water used by the production facilities of a plant varies. Observations on water usage and other,possibility related,variables were collected for 250 months. The data are given in **water.csv file** The explanatory variables are

TEMP= average monthly temperature(degree celsius)

PROD=amount of production( in hundreds of cubic)

DAYS=number of operationing day in the month (days)

HOUR=number of hours shut down for maintenance (hours)

The response variable is USAGE=monthly water usage (gallons/minute)

a. Fit the model containing all four independent variables. What is the multiple regression equation?

b. Test the hyphthesis for the full model. Use significance level 0.05.

c. Would you suggest the model in part b for predictive purposes? Which model or set of models would you suggest for predictive purposes? Hint: Use Individual Coefficients Test (t-test) to find the best model.

d. Use Partial F test to confirm that the independent variable should be out of the model at significance level 0.05.

e. Obtain a 95% confidence interval of regression coefficient for TEMP from the model in part c.

f. Use the method of Model Fit to calculate $R^2_{adj}$ and RMSE to compare the full model and the model in part c. Which model or set of models would you suggest for predictive purpose?

## References

*-Gareth James & Daniela Witten & Trevor Hastie Robert Tibshirani, An Introduction to Statistical Learning with Applications in R: Springer New York Heidelberg Dordrecht London.*

*-Wickham and Grolemund, R for Data Science: O'Reilly Media*