STA302H5
Part 3

Al Nosedal

Fall 2018

"Simple can be harder than complex: You have to work hard to get your thinking clean to make it simple. But it's worth it in the end because once you get there, you can move mountains."

Steve Jobs.

MULTIPLE LINEAR REGRESSION (cont.)

Result

Let $Y = X\beta + \epsilon$, where X has full rank $r+1$ and ϵ is distributed as $N_n(0, \sigma^2 I)$. Then,

$$\hat{\beta} = (X^T X)^{-1} X^T Y \quad \text{is distributed as} \quad N_{r+1}(\beta, \sigma^2 (X^T X)^{-1})$$

and is distributed independently of the residuals $\hat{\epsilon} = Y - X\hat{\beta}$. Further,

$$\hat{\epsilon}^T \hat{\epsilon} \quad \text{is distributed as} \quad \sigma^2 \chi^2_{n-r-1}$$

Result

Let $Y = X\beta + \epsilon$, where X has full rank $r+1$ and ϵ is $N_n(0, \sigma^2 I)$. Then a $100(1-\alpha)\%$ confidence region for β is given by

$$(\hat{\beta} - \beta)^T X^T X(\hat{\beta} - \beta) \leq (r+1)s^2 F_{r+1,n-r-1}(\alpha)$$

where $F_{r+1,n-r-1}(\alpha)$ is the upper (100α)th percentile of an F-distribution with $r+1$ and $n-r-1$ d.f.

Proof

Under the assumption that ϵ_i's are identically and independently distributed as $N(0, \sigma^2)$, the hypothesis $H_0 : C\beta = \gamma$, where C is an $m \times (r+1)$ matrix of rank m with $m < (r+1)$, is rejected if

$$\frac{m^{-1}(C\hat{\beta} - \gamma)^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta} - \gamma)}{s^2} \geq F_{m,n-r-1}(\alpha)$$

where $F_{m,n-r-1}(\alpha)$ is the upper (100α)th percentile of an F-distribution with m and $n - r - 1$ d.f.

Proof

Useful facts for proof

Let $M = C(X^T X)^{-1} C^T$.

- If A is a covariance matrix, then A is nonnegative definite.
- M is invertible, then its eigenvalues are different from zero.
- M is a real and symmetric matrix, then it can be diagonalized by an orthogonal matrix containing normalized eigenvectors of M, and the resulting diagonal matrix contains eigenvalues of M. ($B^T M B = D$).
- M is positive definite.
- $M^{1/2} = BD^{1/2}B^T$.
- The square root matrix $M^{1/2}$ is symmetric.

Example

Fit the model $Y = \beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$ to the data points given in the following table

y	x_0	x_1	x_2
0	1	-2	2
0	1	-1	-1
1	1	0	-2
1	1	1	-1
3	1	2	2

$X^T X$

$$X^T X = \begin{pmatrix} 5 & 0 & 0 \\ 0 & 10 & 0 \\ 0 & 0 & 14 \end{pmatrix}$$

$X^T Y$

$$X^T Y = \begin{pmatrix} 5 \\ 7 \\ 3 \end{pmatrix}$$

$$\hat{\beta}_{3\times 1} = (X^T X)^{-1} X^T Y = \begin{pmatrix} 1 \\ 7/10 \\ 3/14 \end{pmatrix}$$

$$\hat{y}_{5\times 1} = (X^T X)^{-1} X^T Y = \begin{pmatrix} 0.028 \\ 0.086 \\ 0.572 \\ 1.486 \\ 2.828 \end{pmatrix}$$

Test the validity of the regression model. Use $\alpha = 0.05$. That is, test the hypothesis $H_0 : \beta_1 = \beta_2 = 0$ against the alternative hypothesis H_a: at least one of the parameters β_1, β_2, differs from zero.

One way (using our result)

$C\beta = \gamma$, where

$$C_{2\times 3} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

and

$$\gamma_{2\times 1} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

$$\frac{m^{-1}(C\hat{\beta}-\gamma)^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta}-\gamma)}{s^2} \geq F_{m,n-r-1}(\alpha)$$

or

$$\frac{2^{-1}(C\hat{\beta})^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta})}{s^2} \geq F_{2,5-2-1}(\alpha)$$

$$C\hat{\beta} = \begin{pmatrix} 7/10 \\ 3/14 \end{pmatrix}$$

and

$$C(X^T X)^{-1} C^T = \begin{pmatrix} 1/10 & 0 \\ 0 & 1/14 \end{pmatrix}$$

$$(C\hat{\beta})^T [C(X^T X)^{-1} C^T]^{-1} (C\hat{\beta}) = 5.5428$$

$$SSE = \hat{\epsilon}^T \hat{\epsilon} = 0.457144$$

$s^2 = \frac{SSE}{n-r-1} = \frac{0.457144}{5-2-1} = 0.228572$
Finally,

$$\frac{2^{-1}(C\hat{\beta})^T[C(X^TX)^{-1}C^T]^{-1}(C\hat{\beta})}{s^2} = \frac{5.5428/2}{0.228572} \approx 12.1248$$

Another way

Consider the situation where we have fit a model with r independent variables and wish to test the null hypothesis
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0$
that none of the independent variables in the model contribute substantial information for the prediction of Y.
The appropriate reduced model is of the form

$$Y = \beta_0 + \epsilon$$

This reduced model contains $g = 0$ independent variables.

Thus, a test for
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0$
can be based on the statistic

$$F = \frac{(SSE_R - SSE_C)/(r - g)}{(SSE_C)/(n - r - 1)}$$

Example

```
y=c(0,0,1,1,3);

x0=c(1,1,1,1,1);

x1=c(-2,-1,0,1,2);

x2=c(2,-1,-2,-1,2);

# Reduced Model = modR
modR=lm(y~x0-1);

# Complete Model = modC
modC=lm(y~x0+x1+x2-1);
```

Example

```
anova(modR)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value Pr(>F)
## x0         1      5     5.0  3.3333 0.1419
## Residuals  4      6     1.5
```

Example

```
anova(modC)

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## x0         1 5.0000  5.0000 21.8750 0.04280 *
## x1         1 4.9000  4.9000 21.4375 0.04362 *
## x2         1 0.6429  0.6429  2.8125 0.23553
## Residuals  2 0.4571  0.2286
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

$$F = \frac{(SSE_R - SSE_C)/(r-g)}{(SSE_C)/(n-r-1)} = \frac{(6 - 0.457144)/2}{0.457144/2} = 12.1249672$$

ANOVA (Analysis of Variance)

Consider the following model

$$\begin{pmatrix} y_{11} \\ y_{12} \\ y_{13} \\ y_{21} \\ y_{22} \\ y_{23} \end{pmatrix} = \begin{pmatrix} \mu_1 \\ \mu_1 \\ \mu_1 \\ \mu_2 \\ \mu_2 \\ \mu_2 \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \epsilon_{13} \\ \epsilon_{21} \\ \epsilon_{22} \\ \epsilon_{23} \end{pmatrix}$$

where the ϵ_{ij}'s are independent $N(0, \sigma^2)$ values.
a) Find the least squares estimates of μ_1 and μ_2.
b) Test the hypothesis $H_0 : \mu_1 = \mu_2$ vs $H_a : \mu_1 \neq \mu_2$, when $y^T = [29.80, 28.57, 29.97, 30.33, 31.27, 30.37]$.

Solution a)

$$\hat{\beta}_{2\times 1} = \left(\begin{array}{c} \sum_{j=1}^{3} y_{1j} \\ \sum_{j=1}^{3} y_{2j} \end{array} \right) = \left(\begin{array}{c} \bar{y}_{1.} \\ \bar{y}_{2.} \end{array} \right)$$

Solution b)

$$F^*_{1,4} = \frac{(29.447 - 30.657)^2}{(2/3)(0.4330835)} \approx 5.0709$$

Problem

How does an MBA major affect the number of job offers received? An MBA student randomly sampled four recent graduates, one each in finance, marketing, and management, and asked them to report the number of job offers. Can we conclude at the 5% significance level that there are differences in the number of job offers between the three MBA majors?

Finance	Marketing	Management
3	1	8
1	5	5
4	3	4
1	4	6

HW?

A consumer organization was concerned about the differences between the advertised sizes of containers and the actual amount of product. In a preliminary study, six packages of three different brands of margarine that are supposed to contain 500ml were measured. The differences from 500 ml are listed here. Do these data provide sufficient evidence to conclude that differences exist between the three brands? Use $\alpha = 0.05$.

Brand 1	Brand 2	Brand 3
1	2	1
3	2	2
3	4	4
0	3	2
1	0	3
0	4	4

Solution

Step 1. State Hypotheses.
μ_i = population mean for differences from 500 ml (brand i, where $i = 1, 2, 3$).
$H_0 : \mu_1 = \mu_2 = \mu_3$
H_a : At least two means differ.

ANOVA Table (Step 2. Compute test statistic)

Source of Variation	Degrees of Freedom (df)	Sum of Square (SS)	Mean Sum of Squares (MSS)	F Ratio
Treatments	2	6.39	$\frac{6.39}{2} = 3.195$	$\frac{3.195}{1.88} = 1.70$
Error	15	28.20	$\frac{28.20}{15} = 1.88$	
Total	17	34.59		

Solution

Step 3. Find Rejection Region.
We reject the null hypothesis only if

$$F > F_{\alpha,k-1,n-k}$$

If we let $\alpha = 0.05$, the rejection region for this exercise is

$$F > F_{0.05,\ 2,15} = 3.682$$

Solution

Step 4. Conclusion.

We found the value of the test statistic to be $F = 1.70$. Since $F = 1.70 < F_{0.05,\ 2,15} = 3.682$, we **can't** reject H_0. Thus, there is **not** evidence to infer that the average differences differ between the three brands.

R Code

```
# Step 1. Entering data;

brand1=c(1,3,3,0,1,0);

brand2=c (2,2,4,3,0,4);

brand3=c(1,2,4,2,3,4);

differences=c(brand1,brand2,brand3);

brand=c(rep(1,6),rep(2,6),rep(3,6));
```

R Code

```r
# Step 2. ANOVA;

oneway.test(differences~brand,var.equal=TRUE);

##
##  One-way analysis of means
##
## data:  differences and brand
## F = 1.6864, num df = 2, denom df = 15, p-value = 0.2185
```

Example

Because of foreign competition, North American automobile manufacturers have become more concerned with quality. One aspect of quality is the cost of repairing damage caused by accidents. A manufacturer is considering several new types of bumpers. To test how well they react to low-speed collisions, 10 bumpers of each of four different types were installed on mid-size cars, which were then driven into a wall at 5 miles per hour. The cost of repairing the damage in each case was assessed. The data are shown below.

a. Is there sufficient evidence at the 5% significance level to infer that the bumpers differ in their reactions to low-speed collisions?

b. If differences exist, which bumpers differ? Apply Fisher's LSD method with the Bonferroni adjustment.

Data

Bumper 1	Bumper 2	Bumper 3	Bumper 4
610	404	599	272
354	663	426	405
234	521	429	197
399	518	621	363
278	499	426	297
358	374	414	538
379	562	332	181
548	505	460	318
196	375	494	412
444	438	637	499

Solution a)

The test statistic is $F_* = 4.06$ and the $P - value = 0.0139$. There is enough statistical evidence to infer that there are differences between some of the bumpers. The question is now, Which bumpers differ?

Fisher's Least Significant Difference Method

The test statistic to determine whether μ_i and μ_j differ is

$$t_{i\ j} = \frac{(\bar{x}_i - \bar{x}_j)}{\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}$$

with degrees of freedom $\nu = n - k$, where $n =$ total number of observations and $k =$ number of samples (groups).

Fisher's Least Significant Difference Method

The confidence interval estimator is

$$(\bar{x}_i - \bar{x}_j) \pm t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

Least Significant Difference (definition)

We define the least significant difference LSD as

$$LSD = t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}$$

A simple way of determining whether differences exist between each pair of population means is to compare the absolute value of the difference between their two sample means and LSD. In other words, we will conclude that μ_i and μ_j differ if

$$|\bar{x}_i - \bar{x}_j| > LSD$$

LSD will be the same for all pairs of means if all k sample sizes are equal. If some sample sizes differ, LSD must be calculated for each combination. It can be argued that this method is flawed because it will increase the probability of committing a Type I error. That is, it is more likely that the analysis of variance to conclude that a difference exists in some of the population means when in fact none differ.

The true probability of making at least one Type I error is called the experimentwise Type I error rate, denoted α_E. The experimentwise Type I error rate can be calculated as

$$\alpha_E = 1 - (1 - \alpha)^C$$

Here C is the number of pairwise comparisons, which can be calculated by $C = \frac{k(k-1)}{2}$. It can be shown that

$$\alpha_E \leq C\alpha$$

which means that if we want the probability of making at least one Type I error to be no more than α_E, we simply specify $\alpha = \frac{\alpha_E}{C}$. The resulting procedure is called the **Bonferroni adjustment**.

Solution b)

Let's use our example to illustrate Fisher's LSD method and the Bonferroni adjustment. The four sample means and standard deviations are

$\bar{y}_1 = 380$ and $s_1 = 130.0931$

$\bar{y}_2 = 485.9$ and $s_2 = 90.5396$

$\bar{y}_3 = 483.8$ and $s_3 = 102.1086$

$\bar{y}_4 = 348.2$ and $s_4 = 118.5268$

Solution b)

The pairwise absolute differences are

$|\bar{y}_1 - \bar{y}_2| = |380 - 485.9| = 105.9$
$|\bar{y}_1 - \bar{y}_3| = |380 - 483.8| = 103.8$
$|\bar{y}_1 - \bar{y}_4| = |380 - 348.2| = 31.8$
$|\bar{y}_2 - \bar{y}_3| = |485.9 - 483.8| = 2.1$
$|\bar{y}_2 - \bar{y}_4| = |485.9 - 348.2| = 137.7$
$|\bar{y}_3 - \bar{y}_4| = |483.8 - 348.2| = 135.6$

Solution b)

We have that $MSE = 12,399$ and $\nu = n - k = 40 - 4 = 36$.
If we perform the LSD procedure with the Bonferroni adjustment, the number of pairwise comparisons is 6. We set $\alpha = 0.05/6 = 0.0083$. Thus $t_{\alpha/2,n-k} = t_{0.00415,36} = 2.7935$ (using R) and

```
qt(0.00415,36)

## [1] -2.793555
```

$$LSD = t_{\alpha/2}\sqrt{MSE\left(\frac{1}{n_i} + \frac{1}{n_j}\right)} \approx 2.7935\sqrt{12399\left(\frac{1}{10} + \frac{1}{10}\right)} = 139.1095$$

Now no pair of means differ because all the absolute values of the differences between sample means are less than 139.1095.
The drawback to the LSD procedure is that we increase the probability of at least one Type I error. The Bonferroni adjustment corrects this problem.

HW?

Use Fisher's LSD method with the Bonferroni adjustment to determine which population means differ given the following statistics:
$k = 5,\ n_1 = 5,\ n_2 = 5,\ n_3 = 5,\ n_4 = 5,\ n_5 = 5$
$MSE = 125,\ \bar{x}_1 = 227,\ \bar{x}_2 = 205,\ \bar{x}_3 = 219,\ \bar{x}_4 = 248,\ \bar{x}_5 = 202$

Categorical Independent Variables.

Example

Johnson Filtration, Inc., provides maintenance service for water-filtration systems throughout southern Florida. Customers contact Johnson with requests for maintenance service on their water-filtration systems. To estimate the service time and the service cost, Johnson's managers want to predict the repair time necessary for each maintenance request. Hence, repair time in hours is the dependent variable. Repair time is believed to be related to two factors, the number of months since the last maintenance service and the type of repair problem (mechanical or electrical). Data for a sample of 10 service calls are are as follows:

Data

Service Call	Months since Last Service	Type of Repair	Repair time in Hours
1	2	electrical	2.9
2	6	mechanical	3
3	8	electrical	4.8
4	3	mechanical	1.8
5	3	electrical	2.9
6	7	electrical	4.9
7	9	mechanical	4.2
8	8	mechanical	4.8
9	4	electrical	4.4
10	6	electrica	4.5

Let y denote the repair time in hours and x_1 denote the number of months since the last maintenance service. The regression model that uses only x_1 to predict y is

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

R Code

```
x0=c(1,1,1,1,1,1,1,1,1,1);

x1=c(2,6,8,3,2,7,9,8,4,6);

y=c(2.9,3,4.8,1.8,2.9,4.9,4.2,4.8,4.4,4.5);

mod0=lm(y~x0-1);

mod1=lm(y~x1);
```

R Code (ANOVA)

```
anova(mod0);

## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x0         1 145.924 145.924  125.36 1.386e-06 ***
## Residuals  9  10.476   1.164
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

R Code (summary)

```
summary(mod0);

##
## Call:
## lm(formula = y ~ x0 - 1)
##
## Residuals:
##    Min      1Q  Median      3Q     Max
## -2.020 -0.895  0.480  0.905  1.080
##
## Coefficients:
##    Estimate Std. Error t value Pr(>|t|)
## x0   3.8200     0.3412    11.2 1.39e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
```

R Code (ANOVA)

```
anova(mod1);

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value  Pr(>F)
## x1         1  5.596   5.596  9.1739 0.01634 *
## Residuals  8  4.880   0.610
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

R Code (summary)

```
summary(mod1);

##
## Call:
## lm(formula = y ~ x1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.2597 -0.4772  0.1821  0.4509  1.0362
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.1473     0.6050   3.549  0.00752 **
## x1            0.3041     0.1004   3.029  0.01634 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
##
```

To incorporate the type of repair into the regression model, we define the following variable:

$$x_2 = \begin{cases} 1 & \text{if the type of repair is electrical} \\ 0 & \text{otherwise} \end{cases}$$

In regression analysis x_2 is called a **dummy** or **indicator variable**. Using this dummy variable, we can write the multiple regression model as

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

R Code

```
x1=c(2,6,8,3,2,7,9,8,4,6);

y=c(2.9,3,4.8,1.8,2.9,4.9,4.2,4.8,4.4,4.5);

x2=c(1,0,1,0,1,1,0,0,1,1);

mod2=lm(y~x1+x2);
```

R Code (ANOVA)

```
anova(mod2);

## Analysis of Variance Table
##
## Response: y
##           Df Sum Sq Mean Sq F value   Pr(>F)
## x1         1 5.5960  5.5960  26.556 0.001319 **
## x2         1 3.4049  3.4049  16.158 0.005062 **
## Residuals  7 1.4751  0.2107
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

R Code (summary)

```
summary(mod2);

##
## Call:
## lm(formula = y ~ x1 + x2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.49412 -0.24690 -0.06842 -0.00960  0.76858
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.93050    0.46697   1.993 0.086558 .
## x1           0.38762    0.06257   6.195 0.000447 ***
## x2           1.26269    0.31413   4.020 0.005062 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Recall that a test for
$H_0 : \beta_1 = \beta_2 = \cdots = \beta_r = 0$
can be based on the statistic

$$F_* = \frac{(SSE_R - SSE_C)/(r-g)}{(SSE_C)/(n-r-1)}$$

In this case,

$$F_* = \frac{(10.476 - 1.4751)/(2)}{(1.4751)/(7)} \approx 21.3566$$

Interpreting parameters

The multiple regression equation for the Johnson Filtration example is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

To understand how to interpret the parameters β_0, β_1, and β_2 when a categorical variable is present, consider the case when $x_2 = 0$ (mechanical repair).

$$E(y|\text{mechanical}) = \beta_0 + \beta_1 x_1 + \beta_2(0) = \beta_0 + \beta_1 x_1 \tag{1}$$

Interpreting parameters

Similarly, for an electrical repair ($x_2 = 1$), we have

$$E(y|\text{electrical}) = \beta_0 + \beta_1 x_1 + \beta_2(1) = (\beta_0 + \beta_2) + \beta_1 x_1 \qquad (2)$$

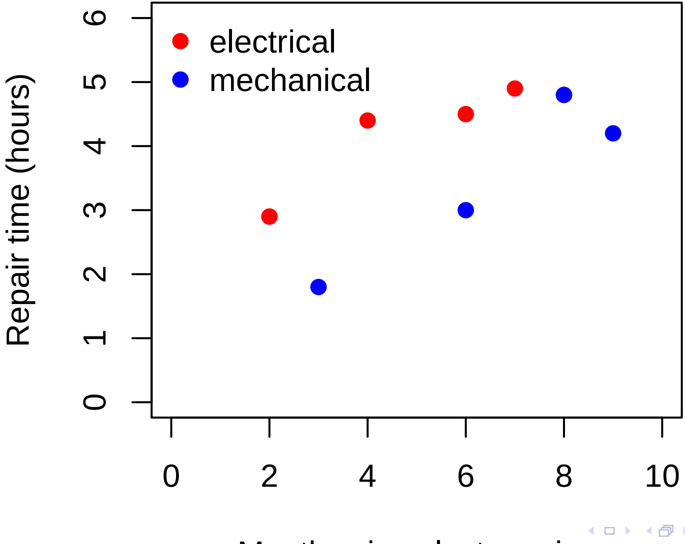
Comparing equations (1) and (2), se see that the slope of both equations is β_1, but the y-intercept differs. The interpretation of β_2 is that it indicates the difference between the mean repair time for an electrical repair and the mean repair time for a mechanical repair.

R Code

```
plot(x1[x2==1],y[x2==1],pch=19,col="red",
ylim=c(0,6),xlim=c(0,10),
xlab="Months since last service",
ylab="Repair time (hours)");

points(x1[x2==0],y[x2==0],pch=19,col="blue");

legend("topleft",c("electrical","mechanical"),pch=c(19,19),
col=c("red","blue"),bty="n");
```


electrical
mechanical
Repair time (hours)
0
1
2
3
4
5
6
0
2
4
6
8
10

R Code

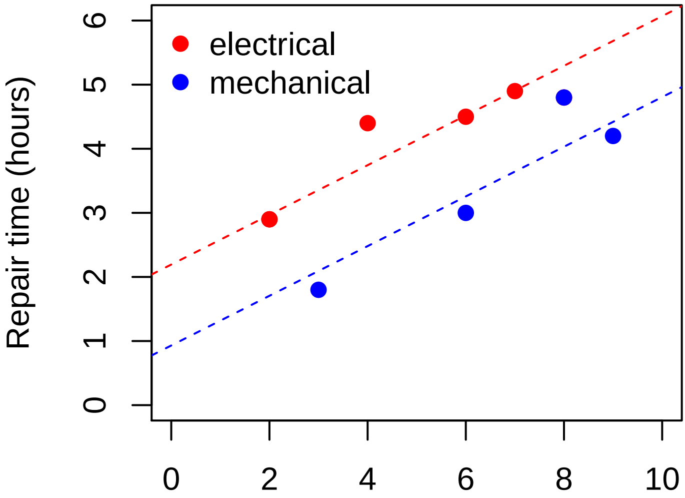
```
plot(x1[x2==1],y[x2==1],pch=19,col="red",
ylim=c(0,6),xlim=c(0,10),
xlab="Months since last service",
ylab="Repair time (hours)");

abline(a=2.1932, b= 0.3876,col="red",lty=2);

points(x1[x2==0],y[x2==0],pch=19,col="blue");

abline(a=0.9305,b=0.3876,col="blue",lty=2);

legend("topleft",c("electrical","mechanical"),pch=c(19,19),
col=c("red","blue"),bty="n");
```


electrical
mechanical
Repair time (hours)
0 1 2 3 4 5 6
0 2 4 6 8 10

Example

Car dealers across North America use the so-called Blue Book to help them determine the value of used cars that their customers trade in when purchasing new cars. It provides alternative values for each car model according to its condition and optional features. The values are determined on the basis of the average paid at recent used-car auctions, the source of supply for many used-car dealers. However, the Blue Book does not indicate the value determined by the odometer reading, despite the fact that a critical factor for used-car buyers is how far the car has been driven. To examine this issue, a used-car dealer randomly selected 100 3-year old Toyota Camrys that were sold at auction during the past month. The dealer recorded the price ($1,000) and the number of miles (thousands) on the odometer. Suppose that the dealer also believed that the color of a car is a factor in determining its auction price. Suppose the dealer believes the colors that are most popular, white and silver, are likely to lead to different prices than other colors.

$$I_1 = \begin{cases} 1 & \text{if color is white} \\ 0 & \text{if color is not white} \end{cases}$$

$$I_2 = \begin{cases} 1 & \text{if color is silver} \\ 0 & \text{if color is not silver} \end{cases}$$

R Code

```
#Step 1. Entering data;
# importing data;
# url of camrys;
camrys_url =
"https://mcs.utm.utoronto.ca/~nosedal/data/camrys.txt"
camrys= read.table(camrys_url,header=TRUE);
names(camrys);
camrys[1:4, ];
```

R Code

```
## [1] "Price"    "Odometer" "I.1"      "I.2"
##   Price Odometer I.1 I.2
## 1  14.6     37.4   1   0
## 2  14.1     44.8   1   0
## 3  14.0     45.8   0   0
## 4  15.6     30.9   0   0
```

R Code

```
# Step 2. Fitting model;

mod=lm(camrys$Price~.,data=camrys);

summary(mod);
```

R Code

```
##
## Call:
## lm(formula = camrys$Price ~ ., data = camrys)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.7047 -0.2022 -0.0133  0.1961  0.6450
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 16.837248   0.197105  85.423  < 2e-16 ***
## Odometer    -0.059123   0.005065 -11.672  < 2e-16 ***
## I.1          0.091131   0.072892   1.250 0.214257
## I.2          0.330368   0.081650   4.046 0.000105 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Thus, for a nonwhite and nonsilver car, the equation becomes

$$\hat{y} = 16.837 - 0.0591x + 0.0911(0) + 0.3304(0)$$

which is

$$\hat{y} = 16.837 - 0.0591x$$

For a white car, the regression equation is

$$\hat{y} = 16.837 - 0.0591x + 0.0911(1) + 0.3304(0)$$

which is

$$\hat{y} = 16.928 - 0.0591x$$

For a silver car, the regression equation is

$$\hat{y} = 16.837 - 0.0591x + 0.0911(0) + 0.3304(1)$$

which is

$$\hat{y} = 17.167 - 0.0591x$$

Test of coefficient of I_1

$$H_0 : \beta_2 = 0$$

$$H_a : \beta_2 \neq 0$$

Test statistic: $t = 1.25$
P-value: 0.2143.

Test of coefficient of I_2

$$H_0 : \beta_3 = 0$$

$$H_a : \beta_3 \neq 0$$

Test statistic: $t = 4.046$
P-value: 0.000105.

Another example

File *sbp.txt* contains observations on systolic blood pressure, age, and gender for a sample of 69 individuals. These data seem to support the commonly found observation that blood pressure increases with age. Another question that can be answered by such data is whether an interaction exists between age and sex: Does the slope of the straight line relating systolic blood pressure to age significantly differ for males and for females?

R Code

```
#Step 1. Entering data;
# importing data;
# url of systolic blood pressure;
sbp_url =
"https://mcs.utm.utoronto.ca/~nosedal/data/sbp.txt"
sbp= read.table(sbp_url,header=TRUE);
names(sbp);
sbp[1:4, ];
y=sbp$SBP;
x1=sbp$AGE;
x2=sbp$SEX;
x3=x1*x2;
```

R Code

```
## [1] "SEX" "SBP" "AGE"
##   SEX SBP AGE
## 1   0 158  41
## 2   0 185  60
## 3   0 152  41
## 4   0 159  47
```

Complete Model

Consider the following regression model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x2 + \beta_3 x_3 + \epsilon$$

where x_1 = AGE, x_2 = SEX ($x_2 = 0$, if male, $x_2 = 1$ if female) and x_3 is the product between AGE and SEX.

Complete Model

When $x_2 = 0$,

$$E(y) = \beta_0 + \beta_1 x_1.$$

When $x_2 = 1$,

$$E(y) = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)x_1.$$

Fitting Complete Model

```r
modC=lm(y~x1+x2+x3);

summary(modC);
```

Fitting Complete Model

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.647  -3.410   1.254   4.314  21.153
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 110.03853    4.73610  23.234  < 2e-16 ***
## x1            0.96135    0.09632   9.980 9.63e-15 ***
## x2          -12.96144    7.01172  -1.849   0.0691 .
## x3           -0.01203    0.14519  -0.083   0.9342
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Fitting Complete Model

When $x_2 = 0$:
$\hat{y}_m = 110.0385 + 0.9614x_1$

When $x_2 = 1$:
$\hat{y}_f = 97.0771 + 0.9494x_1$

Test of Parallelism

We know the null hypothesis that the two regression lines are parallel is equivalent to $H_0 : \beta_3 = 0$, then the slope for females simplifies to β_1.

Fitting Reduced Model

```
## Test of parallelism: beta3=0;

mod2=lm(y~x1+x2);

anova(mod2)
```

Fitting Reduced Model

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x1         1 14951.3 14951.3 189.693 < 2.2e-16 ***
## x2         1  3058.5  3058.5  38.805 3.701e-08 ***
## Residuals 66  5202.0    78.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Computing Test Statistic

```
anova(modC);
```

Computing Test Statistic

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq  F value    Pr(>F)
## x1         1 14951.3 14951.3 186.8390 < 2.2e-16 ***
## x2         1  3058.5  3058.5  38.2210 4.692e-08 ***
## x3         1     0.5     0.5   0.0069    0.9342
## Residuals 65  5201.4    80.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

In this case,

$$F_* = \frac{(5201.99 - 5201.44)/(1)}{(5201.44)/(65)} \approx 0.0069$$

Computing Test Statistic (another way)

```
anova(mod2,modC);
```

Computing Test Statistic (another way)

```
## Analysis of Variance Table
##
## Model 1: y ~ x1 + x2
## Model 2: y ~ x1 + x2 + x3
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     66 5202.0
## 2     65 5201.4  1   0.54936 0.0069 0.9342
```

Test of Equal Intercepts

We know the null hypothesis that the two regression lines are parallel is equivalent to $H_0 : \beta_2 = 0$. The test compares the model

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$$

to the reduced model

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

(Note that this test presumes equal slopes.)

Fitting Reduced Model

```r
## Test of equal intercepts: beta2=0;

mod1=lm(y~x1);

anova(mod1);
```

Fitting Reduced Model

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x1         1 14951.3 14951.3  121.27 < 2.2e-16 ***
## Residuals 67  8260.5   123.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Computing Test Statistic

```
anova(mod2);
```

Computing Test Statistic

```
## Analysis of Variance Table
##
## Response: y
##           Df  Sum Sq Mean Sq F value    Pr(>F)
## x1         1 14951.3 14951.3 189.693 < 2.2e-16 ***
## x2         1  3058.5  3058.5  38.805 3.701e-08 ***
## Residuals 66  5202.0    78.8
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

In this case,

$$F_* = \frac{(8260.51351 - 5201.99)/(1)}{(5201.99)/(66)} \approx 38.8049$$

Computing Test Statistic (another way)

```
anova(mod1,mod2);
```

Computing Test Statistic (another way)

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x2
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     67 8260.5
## 2     66 5202.0  1    3058.5 38.805 3.701e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Test of Coincidence

The hypothesis that the two regression lines coincide is $H_0 : \beta_2 = \beta_3 = 0$. When both β_2 and β_3 are 0, the model for females reduces to $y_f = \beta_0 + \beta_1 x_1 + \epsilon$, the model for males (i.e. the two lines coincide). The two models being compared are

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

and

$$y = \beta_0 + \beta_1 x_1 + \epsilon$$

Computing Test Statistic (another way)

```
anova(mod1,modC)
```

Computing Test Statistic (another way)

```
## Analysis of Variance Table
##
## Model 1: y ~ x1
## Model 2: y ~ x1 + x2 + x3
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     67 8260.5
## 2     65 5201.4  2    3059.1 19.114 2.96e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

Linear Algebra (background)

Definition

A real symmetric matrix A is said to be

1. Positive definite if $v^T A v > 0$ for all nonzero v in $\Re^n$
2. Positive semidefinite if $v^T A v \geq 0$ for all nonzero v in $\Re^n$

Theorem

Let x be a random n-vector. The matrix $\Sigma = cov(x)$ is at least positive semi-definite.

Definition

A matrix B is said to be a **square root** of a matrix A if $BB = A$.

Our last example...

La Quinta Motor Inns is a moderately priced chain of motor inns located across the United States. Its market is the frequent business traveler. The chain recently launched a campaign to increase market share by building new inns. The management of the chain is aware of the difficulty in choosing locations for new motels. Moreover, making decisions without adequate information often results in poor decisions. Consequently, the chain's management acquired data on 100 randomly selected inns belonging to La Quinta. The objective was to predict which sites are likely to be profitable. To measure profitability, La Quinta used operating margin, which is the ratio of the sum of profit, depreciation, and interest expenses divided by total revenue. La Quinta defines profitable inns as those with an operating margin in excess of 50%.

Column 1: Operating margin, in percent.
Column 2: Total number of motel and hotel rooms within 3 miles of La Quinta inn
Column 3: Number of miles to closest competition.
Column 4: Office space in thousands of square feet in surrounding community
Column 5: College and university enrollment (in thousands) in nearby university or college
Column 6: Median household income (in thousands) in surrounding community
Column 7: Distance (in miles) to the downtown core.

a. Develop a regression analysis.
b. Test to determine whether there is enough evidence to infer that the model is valid.
c. Test each of the slope coefficients.
d. Interpret the coefficients.
e. Predict with 95% confidence the operating margin of a site with the following characteristics. There are 3815 rooms within 3 miles of the site, the closest other hotel or motel is 0.9 miles away, the amount of office space is 476000 square feet, there is one college and one university with a total enrollment of 24500 students, the median income in the area is $35000, and the distance to the downtown core is 11.2 miles.
f. Refer to part e). Estimate with 95% confidence the mean operating margin of all La Quinta inns with those characteristics.

R Code

```r
#Step 1. Entering data;
# importing data;
# url of La Quinta Inns;
quinta_url =
"https://mcs.utm.utoronto.ca/~nosedal/data/quinta.txt"
quinta= read.table(quinta_url,header=TRUE);
names(quinta);
quinta[1:4, ];
```

R Code

```
## [1] "Margin"      "Number"      "Nearest"     "OfficeSpace'
## [6] "Income"      "Distance"
##   Margin Number Nearest OfficeSpace Enrollment Income Dista
## 1   55.5   3203     4.2         549        8.0     37
## 2   33.8   2810     2.8         496       17.5     35      1
## 3   49.0   2890     2.4         254       20.0     35
## 4   31.9   3422     3.3         434       15.5     38      1
```

R Code

```r
## a) Develop a regression model

model<-lm(quinta$Margin~.,data=quinta);
```

R Code

```
## b) test to determine whether there is enough evidence to in

model0<-lm(quinta$Margin~1,data=quinta);

anova(model0,model)
```

R Code

```
## Analysis of Variance Table
##
## Model 1: quinta$Margin ~ 1
## Model 2: quinta$Margin ~ Number + Nearest + OfficeSpace + E
##     Income + Distance
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1     99 5949.5
## 2     93 2825.6  6    3123.8 17.136 3.034e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '
```

R Code

```r
## c) test each of the slope coefficients

summary(model)
```

R Code

```
##
## Call:
## lm(formula = quinta$Margin ~ ., data = quinta)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -12.267  -3.022  -0.086   4.234  13.596
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.138575   6.992948   5.454 4.04e-07 ***
## Number      -0.007618   0.001255  -6.069 2.77e-08 ***
## Nearest      1.646237   0.632837   2.601   0.0108 *
## OfficeSpace  0.019766   0.003410   5.796 9.24e-08 ***
## Enrollment   0.211783   0.133428   1.587   0.1159
## Income       0.413122   0.139552   2.960   0.0039 **
```

R Code

```r
## e) predict with 0.95 confidence the operating margin of
## a site with the following characteristics.

attach(quinta);

x0=data.frame(Number=3815,Nearest=0.9,
OfficeSpace=476,Enrollment=24.5,
Income=35,Distance=11.2);

predict(model,x0,interval="prediction");
```

R Code

```
##         fit      lwr      upr
## 1 37.09149 25.39525 48.78772
```

R Code

```
## f) Estimate with 0.95 confidence the mean operating margin

##La Quinta Inns with those characteristics.attach(quinta);

x0=data.frame(Number=3815,Nearest=0.9,
OfficeSpace=476,Enrollment=24.5,
Income=35,Distance=11.2);

predict(model,x0,interval="confidence");
```

R Code

```
## The following objects are masked from quinta (pos = 3):
##
##     Distance, Enrollment, Income, Margin, Nearest,
Number,
##     OfficeSpace

##          fit      lwr      upr
## 1 37.09149 32.96972 41.21326
```