Multiple Regression

SLR provides for a linear relation between a response variable, y, and a single explanatory variable, x.

Multiple regression extends this to allow for multiple explanatory variables: $x_1, x_2, \dots, x_K$

The data table takes the form

y	x_1	x_2	x_3	$\dots$	x_K
y_1	x_{11}	x_{21}	x_{31}		x_{K1}
y_2	x_{12}	x_{22}	x_{32}	$\dots$	x_{K2}
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
y_n	x_{1n}	x_{2n}	x_{3n}		x_{Kn}

response — explanatory variables

The basic multiple regression model provides for a linear relation between the y and the x's

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_K x_K + e$$

where β_K is the regression coefficent for the K^{th} explanatory variable.

- The goal is to estimate the regression coefficients from the data
- Least-squares principles that minimize $\sum_{i=1}^{n} (y - \hat{y})^2$ are used to determine $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2 \cdots \hat{\beta}_K$

(The formulae are best developed using matrix algebra, and beyond the scope of this course. Statistical software carries out the necessary computations).

- The end result is a relation that allows us to predict y using all of the X's, i.e.

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \cdots + \hat{\beta}_K X_K$$

- As with SLR, we assume $e \sim N(0, \sigma^2)$ and check these assumptions with plots of the residuals $\hat{e} = y - \hat{y}$ (and QQ plots):

$\hat{e}$ vs $\hat{y}$
$\hat{e}$ vs X_1
$\hat{e}$ vs X_2
:
$\hat{e}$ vs X_K

} look for constant variance, independence and normality.

• Inference in Multiple Regression

We can test the overall significance of the regression using an F-test (and ANOVA table)

→ Is their a significant relation between the response (y) and ANY of the explanatories (x)

This makes use of slightly modified ANOVA-table:

Source	df	SS*	MS	F
Regression	k	SSR	MSR	F_{obs}
Residual	$n-k-1$	SSE	MSE	
Total	$n-1$	SST		

only the degrees of freedom change from the SLR case

To test H_0: β's $= 0$ vs H_a: atleast one of the β different from zero

→ compare F_{obs} to $F_{k,n-k-1}$ distribution

*(SSR, SSE, SST are defined as before)

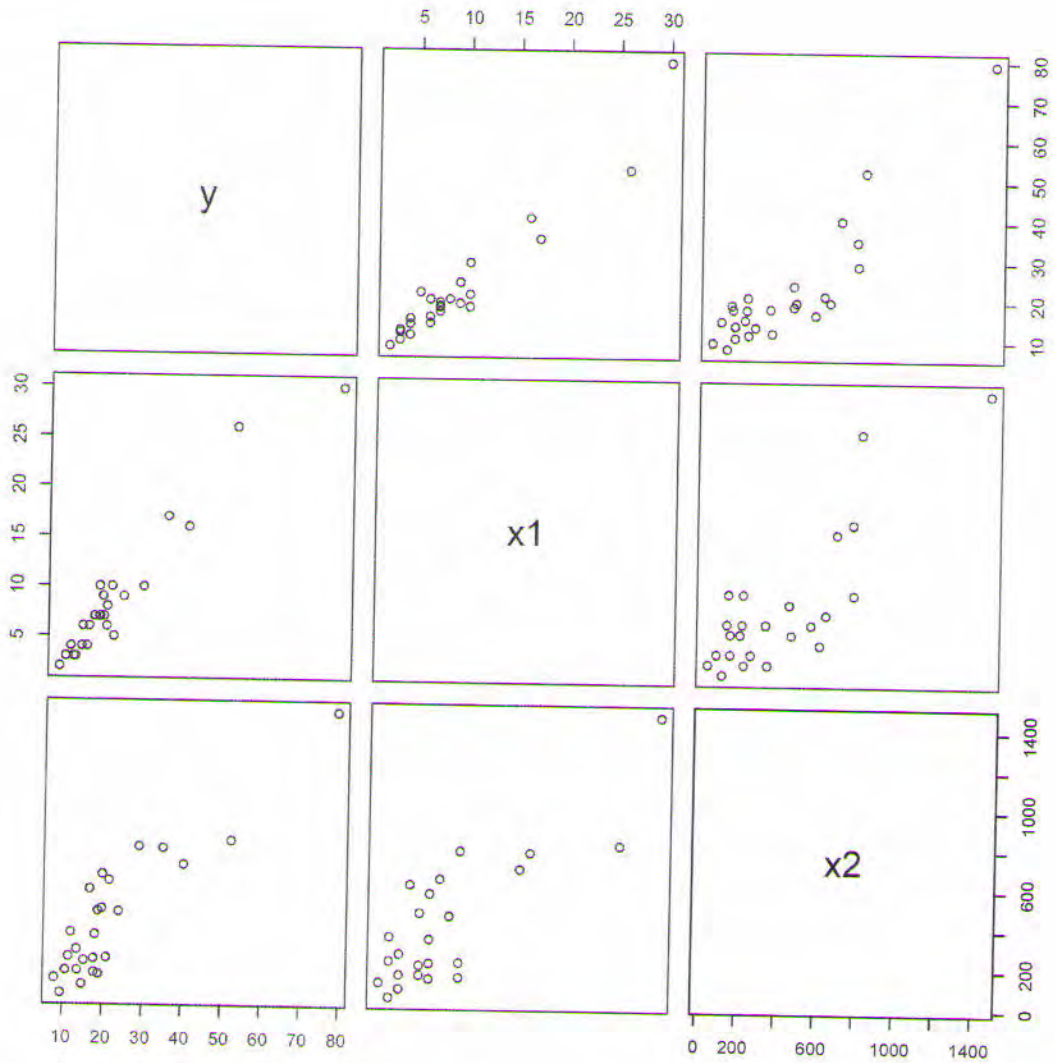
EXAMPLE: Multiple Regression for the Delivery Time Data

The variables are:
y: the Delivery Time (minutes) for cases of pop
x1: the number of Cases delivered
x2: the distance walked (feet) from the delivery truck to the dispensing machine

The goal of the multiple regression is to predict Delivery Time from # of Cases, and Distance walked.

```
          y x1   x2
 [1,] 16.68  7  560
 [2,] 11.50  3  220
 [3,] 12.03  3  340
 [4,] 14.88  4   80
 [5,] 13.75  6  150
 [6,] 18.11  7  330
 [7,]  8.00  2  110
 [8,] 17.83  7  210
 [9,] 79.24 30 1460
[10,] 21.50  5  605
[11,] 40.33 16  688
[12,] 21.00 10  215
[13,] 13.50  4  255
[14,] 19.75  6  462
[15,] 24.00  9  448
[16,] 29.00 10  776
[17,] 15.35  6  200
[18,] 19.00  7  132
[19,]  9.50  3   36
[20,] 35.10 17  770
[21,] 17.90 10  140
[22,] 52.32 26  810
[23,] 18.75  9  450
[24,] 19.83  8  635
[25,] 10.75  4  150
```

Scatter plots of the data are:



It looks like it is plausible that there is a positive relation between y and x1, as well as y and x2. (A note of interest is that there is also a weakly positive correlation between x1 and x2).

Let's fit a multiple regression model: $y = b0 + b1*x1 + b2*x2 + e$

The results from the software package R are the following:

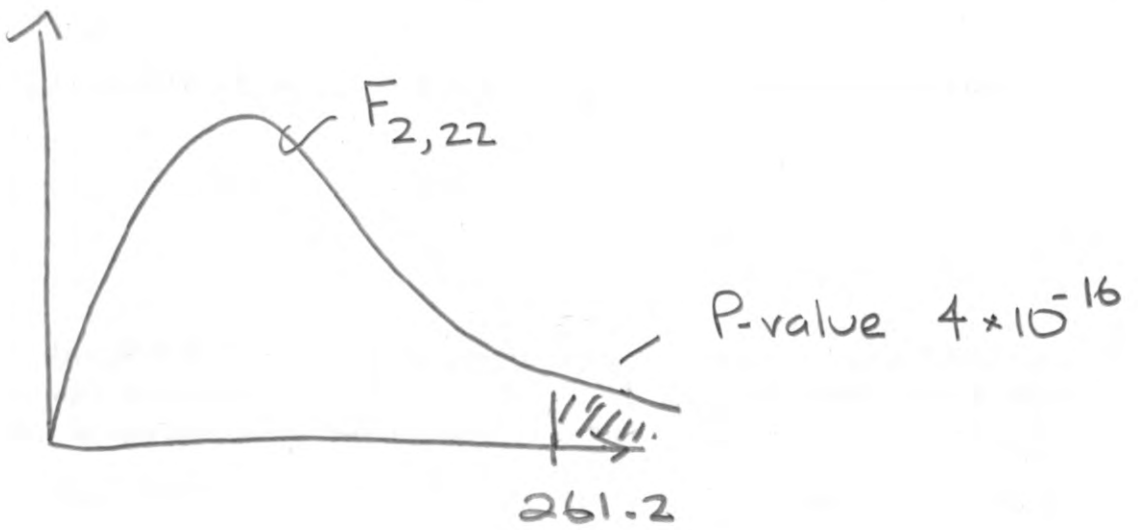
```
> dt.lm=lm(y~x1+x2)
> summary(dt.lm)

Call:
lm(formula = y ~ x1 + x2)

Residuals:
    Min      1Q  Median      3Q     Max
-5.7880 -0.6629  0.4364  1.1566  7.4197

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2.341231   1.096730   2.135 0.044170 *
x1          1.615907   0.170735   9.464 3.25e-09 ***
x2          0.014385   0.003613   3.981 0.000631 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
1

Residual standard error: 3.259 on 22 degrees of freedom
Multiple R-squared: 0.9596,   Adjusted R-squared: 0.9559
F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

EXAMPLE: Based on the regression output for the Delivery Time Data, can we reconstruct the ANOVA Table?

Source	df	SS	MS	F_{obs}
Regression	2	5547.8	2773.9	261.2
Residual	22	233.64	10.62	
Total	24	5781.44		

- degrees of freedom : $n = 25$, $K = 2$
- $F_{obs} = \frac{MSR}{MSE} \rightarrow MSR = (261.2)(10.62)$
  $= 2773.9$
  $261.2 \approx 3.259^2 = 10.62$

- Is the overall regression Significant?



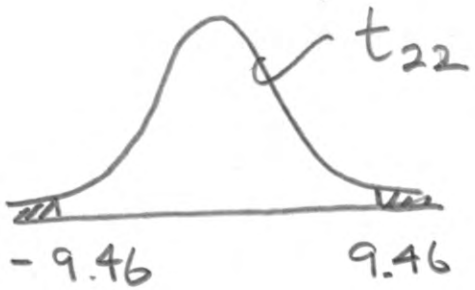
↳ there is a relation between y and at least one of the x's (x_1 and/or x_2)

- Is the β_1 coefficient significantly different from zero?

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

from regression output

$$t_{obs} = \frac{\hat{\beta}_1}{SE(\hat{\beta}_1)} = 9.464$$ compare to t_{n-k-1}



→ P-value $< 10^{-8}$

→ → reject H_0