MATH 423/533 – MULTIPLE LINEAR REGRESSION EXAMPLE IN R

The following code in R fits the simple linear regression model

$$\mathbb{E}[Y_i|\mathbf{x}_i] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \qquad i = 1, 2, \ldots, n$$

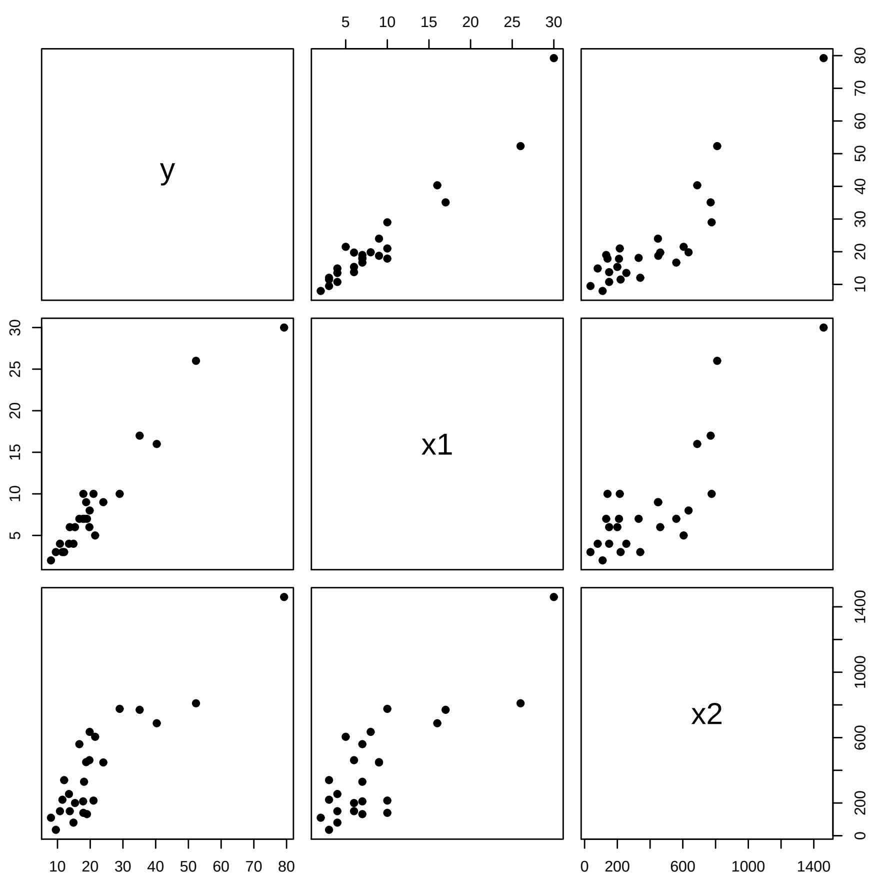
to the Delivery Time data from the textbook, where the delivery time for a vending machine service is to be predicted as a function of two continuous variables. We have $n = 25$ and

- x_{i1} is number of cases stocked, the first continuous predictor;
- x_{i2} is distance walked (in feet), the second continuous predictor;
- $\mathbf{x}_i = [1\ x_{i1}\ x_{i2}]$, a (1×3) row vector.
- y_i is the delivery time (in minutes) outcome random variable;

Multiple Linear Regression: Plot Data

```
> library(MPV) #load the textbook data sets library
> Delivery<-p8.3
> x1<-Delivery$x1
> x2<-Delivery$x2
> y<-Delivery$y
> pairs(cbind(y,x1,x2),pch=19)
```

The resulting plot of the data is given below:



Multiple Linear Regression: Find plane of best fit

```
7  > fit.Del12<-lm(y ~ x1+x2,data=Delivery)
8  > summary(fit.Del12)
9  Coefficients:
10              Estimate Std. Error t value Pr(>|t|)
11 (Intercept) 2.341231   1.096730   2.135 0.044170 *
12 x1          1.615907   0.170735   9.464 3.25e-09 ***
13 x2          0.014385   0.003613   3.981 0.000631 ***
14 ---
15 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
16
17 Residual standard error: 3.259 on 22 degrees of freedom
18 Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9559
19 F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

The output is of the same form as for simple linear regression:

- Lines 9–13: The coefficients table containing $\widehat{\beta}_j$, e.s.e.$(\widehat{\beta}_j)$, t_j and the associated p-value for $j = 0, 1, 2$.
- Line 12 contains information on $\widehat{\beta}_1$, line 13 contains information on $\widehat{\beta}_2$.
- Line 17 contains the estimate of $\widehat{\sigma}$ (3.259).
- Line 18 contains the R^2 and R^2_{adj} quantities that record how well the two predictors combine to explain the variation in y
- Line 19 contains the information on the 'global' F-test information for the test of

  $$\text{H}_0 \; : \; \beta_1 = \beta_2 = 0$$

  against the general alternative that at least one of these parameters is non zero.

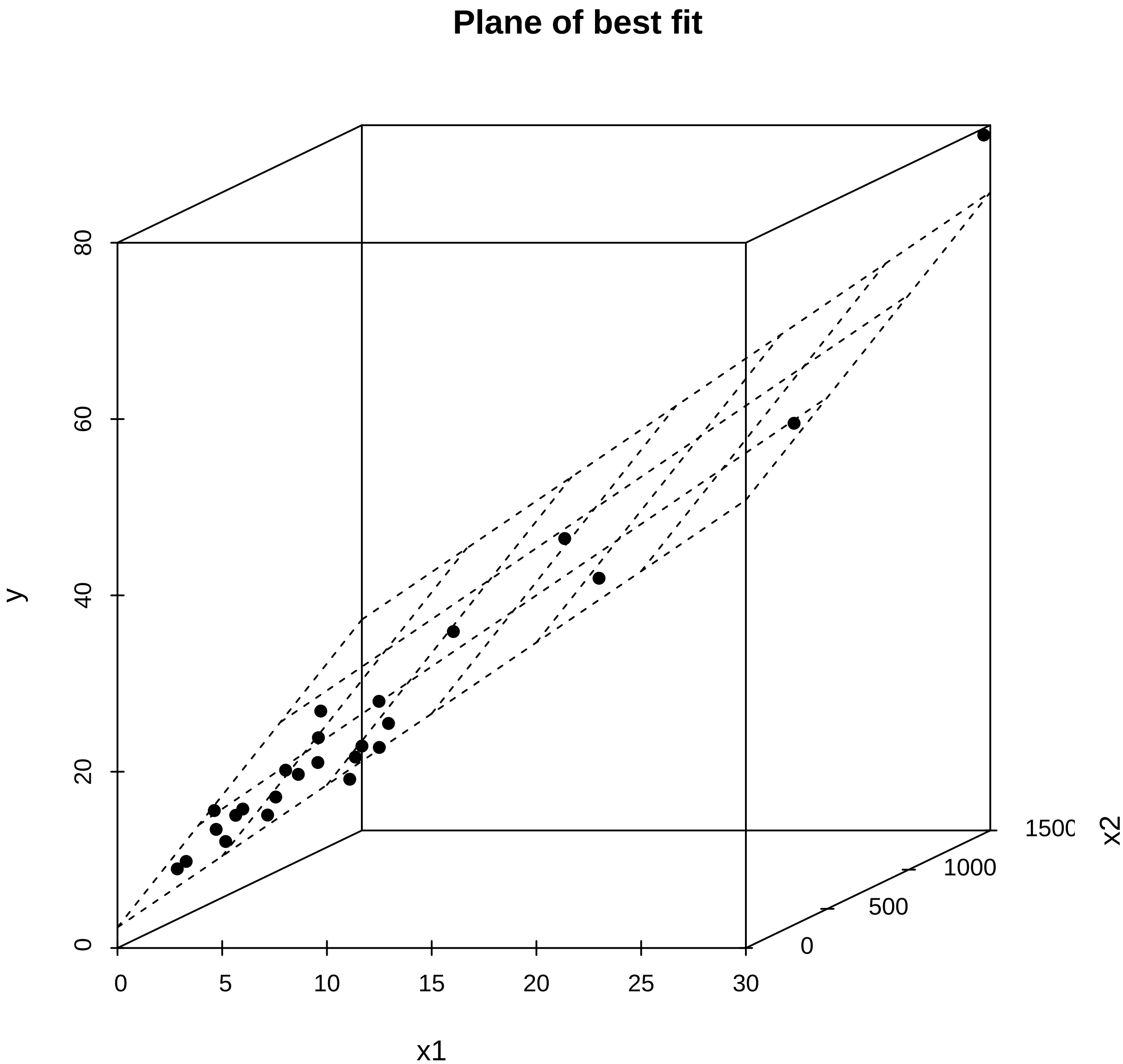
Note that line 7 that computes the model fit could be implemented with the formula rearranged

```
20 > fit.Del21<-lm(y ~ x2+x1,data=Delivery)
21 > summary(fit.Del21) #gives same regression results
22 Coefficients:
23              Estimate Std. Error t value Pr(>|t|)
24 (Intercept) 2.341231   1.096730   2.135 0.044170 *
25 x2          0.014385   0.003613   3.981 0.000631 ***
26 x1          1.615907   0.170735   9.464 3.25e-09 ***
27 ---
28 Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1     1
29
30 Residual standard error: 3.259 on 22 degrees of freedom
31 Multiple R-squared:  0.9596,    Adjusted R-squared:  0.9559
32 F-statistic: 261.2 on 2 and 22 DF,  p-value: 4.687e-16
```

The answers are identical, apart from the fact that the rows in the coefficients table are swapped (lines 25–26).

The resulting plane of best fit given below:

```
> library(scatterplot3d)
> s3d <-scatterplot3d(x1,x2,y, pch=16, grid=FALSE,
+ main="Plane of best fit",angle=20)
> s3d$plane3d(fit.Del12,col='black')
```



The residual plots reveal a possible non-constant variance, but no systematic variation in mean, for the residuals.