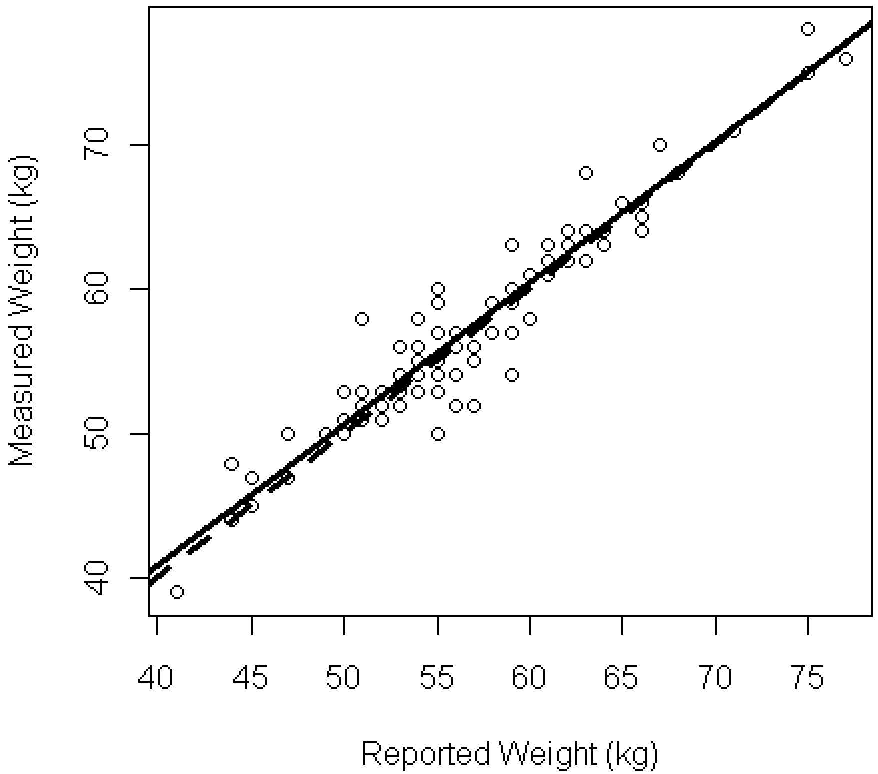
-Chapter 5: Linear Least Squares Regression

Text sections 5.1, 9.1, 10.1

5.1 Simple Regression



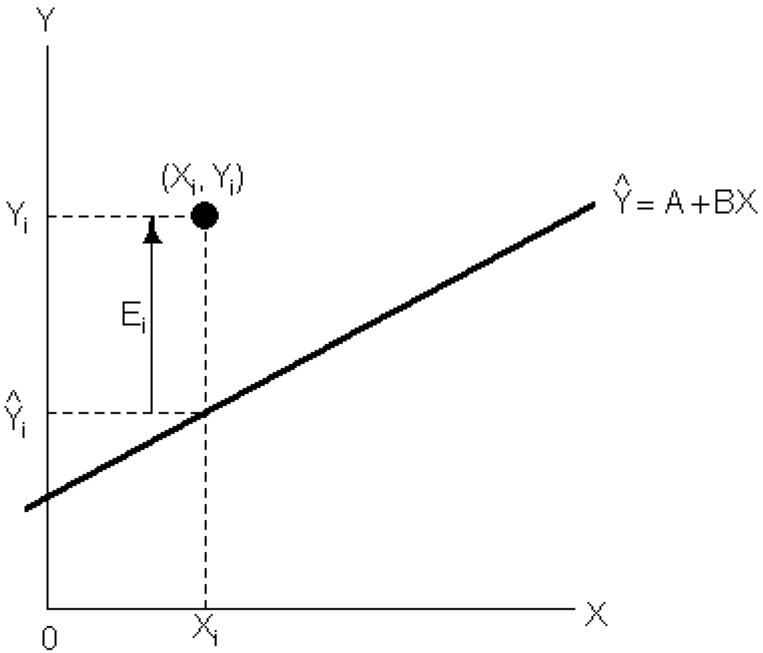
$$y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1,\ldots,n$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

(Note the bold font notation for vectors.)



A further note on notation: In Chap 5 Fox doesn't distinguish between parameters and estimates. That doesn't happen until Chap 6. I will make some distinction here.

$$Y_i = A + BX_i + E_i = \hat{Y}_i + E_i$$

Any line through the means has sum of errors = 0:

$$\bar{Y} = A + B\bar{X}$$

implies that we can write

$$Y_i - \underline{Y} = A + B(X_i - \bar{X}) + E_i$$

and

$$\sum_{i=1}^{n} E_i = \sum (Y_i - \underline{Y}) - B \sum (X_i - \bar{X}) = 0 - B \times 0 = 0$$

Two possibilities:

- Find A and B to minimize sum of absolute residuals: $\sum |E_i|$
- Find A and B to minimize sum of squared residuals: $\sum E_i^2$ --- *<u>least squares criterion</u>* (sensitive to outliers)

$$S(A,B) = \sum_{i=1}^{n} E_i^2 = \sum \left(Y_i - A - BX_i \right)^2$$

Differentiate with respect to A and B to find the minimum:

$$\frac{\partial S(A,B)}{\partial A} = -2\sum\left(Y_i - A - BX_i\right) = 0$$

$$\frac{\partial S(A,B)}{\partial B} = -2\sum X_i\left(Y_i - A - BX_i\right) = 0$$

These are the ***normal equations***:

$$An + B\sum X_i = \sum Y_i$$

$$A\sum X_i + B\sum X_i^2 = \sum X_i Y_i$$

with solution:

$$\hat{A} = \overline{Y} - B\overline{X}$$

$$\hat{B} = \frac{n\sum X_i Y_i - \sum X_i \sum Y_i}{n\sum X_i^2 - \left(\sum X_i\right)^2} = \frac{\sum\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sum\left(X_i - \overline{X}\right)^2}$$

(Everyone but Fox uses the "hat" notation for the estimators.) The first result means that the line with this slope and intercept goes through the point $\left(\overline{X}, \overline{Y}\right)$, meaning also that the sum of ***residuals*** $\hat{E}_i = Y_i - \left(\hat{A} - \hat{B}X_i\right)$ is zero.

The 2nd normal equation leads to $\sum X_i \hat{E}_i = 0$, and similarly, $\sum \hat{Y}_i \hat{E}_i = 0$, so the residuals are ***uncorrelated*** with the values of the explanatory variable X and with the ***fitted values*** $\hat{Y}_i = \hat{A} + \hat{B}X_i$. (See Fox Exercise 5.1, p. 96.)

[Note my use of hat notation for estimates, fitted values, and residuals.]

5.1.2 Simple Correlation

How close does the fitted line fits the scatter of points?

Standard error of the regression or *residual standard error* is defined as square root of the following variance of residuals

$$s_E^2 = \frac{1}{n-2}\sum E_i^2$$

$$s_E = \sqrt{\frac{1}{n-2}\sum E_i^2}$$

In contrast to standard error of the regression, the *correlation coefficient* is a *relative* measure of fit of the straight line. We could write down the formula you know for a correlation coefficient, but we'll express it differently here.

Consider the model without the explanatory variable X:

$$Y_i = A' + E_i'$$

Least squares estimation means minimize $S(A') = \sum\left(Y_i - A'\right)^2$ which leads to $\hat{A}' = \bar{Y}$.

This is the *null model* and the residual sum of squares for this model will actually be called the *total sum of squares*:

$$\text{TSS} = \sum \hat{E}_i'^2 = \sum\left(Y_i - \bar{Y}\right)^2$$

while the *residual sum of squares* for the linear fit will be written

$$\text{RSS} = \sum \hat{E}_i^2 = \sum\left(Y_i - \hat{Y}_i\right)^2 = \sum\left(Y_i - (\hat{A} + \hat{B}X_i)\right)^2$$

The difference between these is the *regression sum of squares*

$$\text{RegSS} = \text{TSS} - \text{RSS}$$

Finally, the ratio of RegSS to TSS is the *reduction in* (residual) *sum of squares* due to the linear regression and it defines the *square of the correlation coefficient*:

$$r^2 = \frac{\text{RegSS}}{\text{TSS}}$$

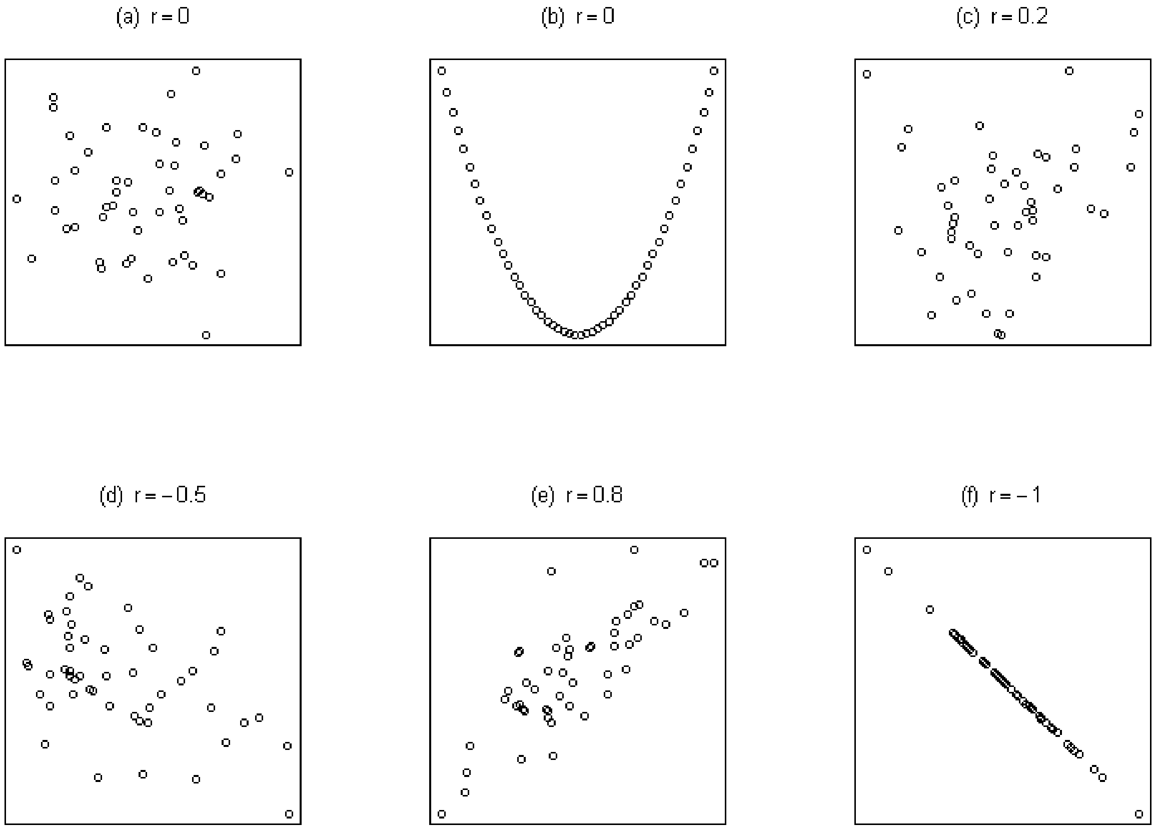


Fig 5.4 Scatterplos illustrating different levels of correlation.

By writing

$$Y_i - \bar{Y} = \left(Y_i - \hat{Y}_i\right) + \left(\hat{Y}_i - \bar{Y}\right)$$

Summing the square of both sides, we find the cross-product on the right sums to zero so that we are left with

$$\sum\left(Y_i - \bar{Y}\right)^2 = \sum\left(Y_i - \hat{Y}_i\right)^2 + \sum\left(\hat{Y}_i - \bar{Y}\right)^2$$

So that the regression sum of squares is the last term is RegSS so that this is:

$$\text{TSS} = \text{RSS} + \text{RegSS}$$

This is the ***analysis of variance*** for a regression model.

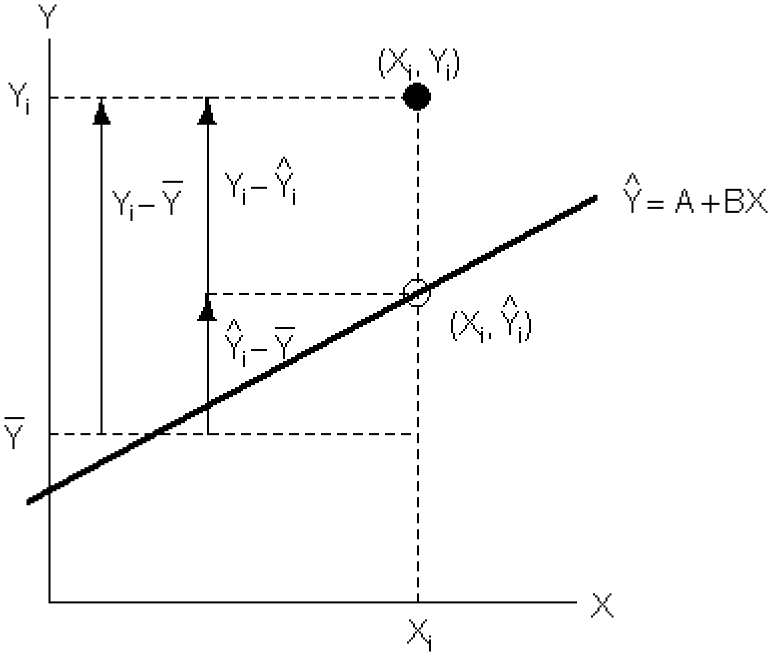


Fig 5.5 Decomposition of the total deviation $Y_i - \bar{Y}$ into components $Y_i - \hat{Y}_i$ (residual) and $\hat{Y}_i - \bar{Y}$ (regression)

We have expressed a correlation coefficient as the square root of the ratio of an "explained sum of squares" due to linear regression, RegSS, over a "total sum of squares". It can also be computed by analogy with the usual correlation coefficient for a pair of random variables $\rho = \sigma_{xy} \big/ \left(\sigma_x \sigma_y\right)$:

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{\sqrt{\sum\left(X_i - \bar{X}\right)^2 \sum\left(Y_i - \bar{Y}\right)^2}}$$

where

$$s_{XY} = \frac{\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right)}{n-1}$$

It is also interesting to relate this to the regression slope:

$$r = \frac{\left(\sum\left(X_i - \bar{X}\right)\left(Y_i - \bar{Y}\right) \big/ \sum\left(X_i - \bar{X}\right)^2\right)\sqrt{\sum\left(X_i - \bar{X}\right)^2}}{\sqrt{\sum\left(Y_i - \bar{Y}\right)^2}} = \hat{B}\frac{\sqrt{\sum\left(X_i - \bar{X}\right)^2}}{\sqrt{\sum\left(Y_i - \bar{Y}\right)^2}}$$

Ex. Davis reported and measured weights

```
names(Davis)
summary(Davis)

frepwt <- repwt[sex=="F"]
fweight <- weight[sex=="F"]
fweight <- fweight[!is.na(frepwt)]
frepwt <- frepwt[!is.na(frepwt)]
TSS <- sum( (fweight-mean(fweight))^2 )

Tlmfit <- lm( fweight ~ frepwt )
fweight.hat <- fitted(Tlmfit)  # Ahat + Bhat * X
RSS <- sum( (fweight - fweight.hat)^2 )
RegSS <- sum( (fweight.hat - mean(fweight))^2 )

summary(Tlmfit)

anova(Tlmfit)

cor(fweight,frepwt)^2
```

9.1-9.2 Linear models in matrix form

(We are not now covering everything in these first sections of 9.1 and 9.2, just a "taste".)

$$y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, \ldots, n$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

The residual sum of squares can be written

$$RSS = \|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

Then

$$\frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

So, the normal equations are obtained by setting this to zero:

$$\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$$

and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$

The fitted values can then be written

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

which is the *orthogonal projection* of $\mathbf{y}$ onto the plane spanned by the columns of $\mathbf{X}$.

The vector geometry of least squares

$$y_i = \alpha + \beta x_i + \varepsilon_i,\ i = 1, \ldots, n$$

$$\boldsymbol{y} = \alpha + \beta \boldsymbol{x} + \boldsymbol{\varepsilon}$$

$$\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

This is an *n-dimensional observation space* in which the variables are represented as *vectors*. Fig 10.1 illustrates this model.

10.1. Simple Regression 221

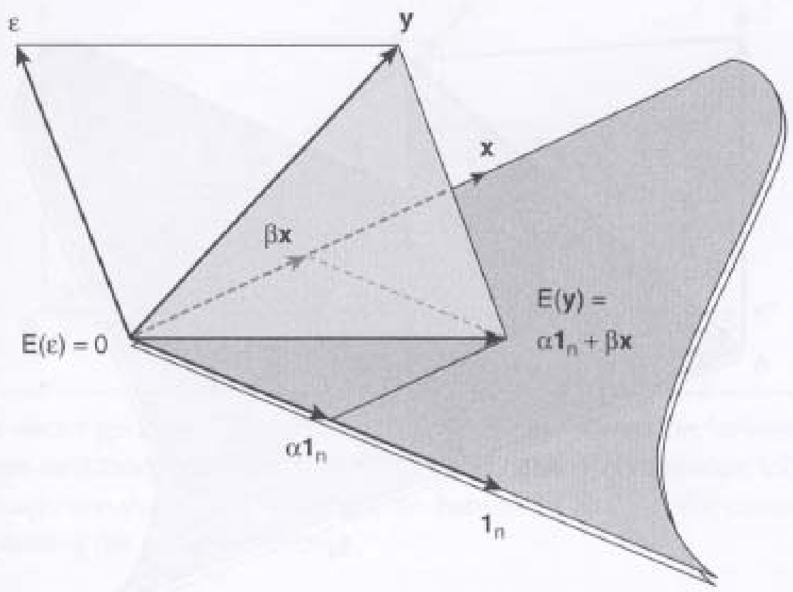


Figure 10.1 The vector geometry of the simple-regression model, showing the three-dimensional subspace spanned by the vectors **x**, **y**, and $\mathbf{1}_n$. Because the expected error is **0**, the expected-Y vector, $E(\mathbf{y})$, lies in the plane spanned by $\mathbf{1}_n$ and **x**.

Note: If you need some mathematical review, see the text Appendix material available online, especially section B: Matrices, Linear Algebra and Vector Geometry.

Figure 10.1 showed the *model*. Figure 10.2 illustrates the *least squares fit*.

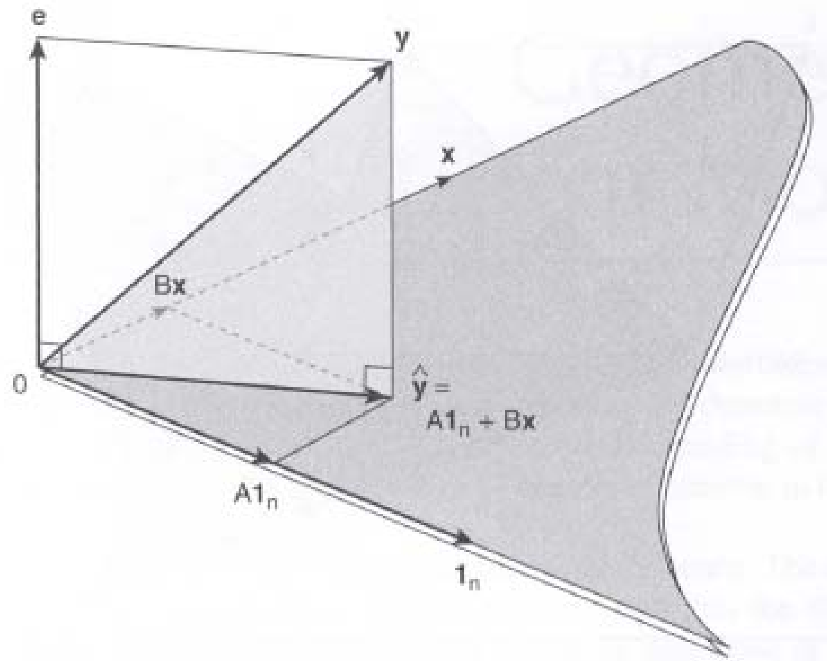


Figure 10.2 The vector geometry of least-squares fit in simple regression. Minimizing the residual sum of squares is equivalent to making the **e** vector as short as possible. The $\hat{\mathbf{y}}$ vector is, therefore, the orthogonal projection of **y** onto the $\{\mathbf{1}_n, \mathbf{x}\}$ plane.

The ANOVA decomposition

The ANOVA decomposition derives from the *mean-deviation form* of the model

$$Y_i = A + Bx_i + E_i$$

Substituting the estimate $A = \overline{Y} + B\overline{x}$

$$Y_i - \overline{Y} = B\left(x_i - \overline{x}\right) + E_i$$

$$y_i^* = B\,x_i^* + E_i$$

or, in vector notation,

$$\mathbf{y}^* = B\mathbf{x}^* + \mathbf{e}$$

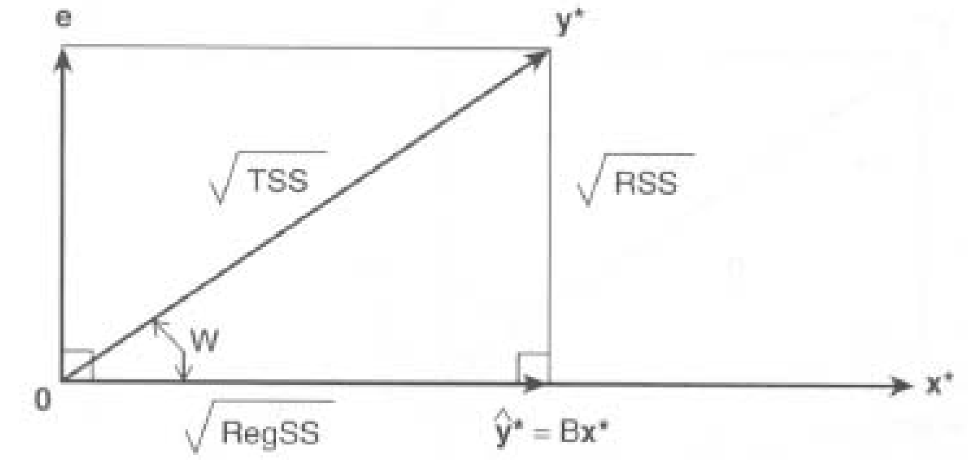


Figure 10.3 The vector geometry of least-squares fit in simple regression for variables in mean-deviation form. The analysis of variance for the regression follows from the Pythagorean theorem. The correlation between X and Y is the cosine of the angle W separating the $\mathbf{x}^*$ and $\mathbf{y}^*$ vectors.

5.2 Multiple regression

With the matrix notation, moving on to more than one explanatory variable is relatively trivial. The text (sect 5.2.1) writes

$\hat{Y} = A + B_2 X_1 + B_2 X_2$ or, indicating the observations with subscript i, $Y_i = A + B_1 X_{i1} + B_2 X_{i2} + E_i$

Then writes out the detail of minimizing the error sum of squares, deriving the normal equations and expressions for A, B_1, and B_2. Read through this to see how the error sum of squares, denoted by $S(A, B_1, B_2)$, is differentiated, but you need not try to remember the form of the least squares estimators in eqn (5.6) on p. 88, except for the fact that $A = \bar{Y} - B_1 \bar{X} - B_2 \bar{X}$. However, it is useful to recognize the form of the normal equations in eqn (5.5):

$$\begin{aligned}
&An + B_1 \sum X_{i1} + B_2 \sum X_{i2} = \sum Y_i \\
&A \sum X_{i1} + B_1 \sum X_{i1}^2 + B_2 \sum X_{i1} X_{i2} = \sum X_{i1} Y_i \\
&A \sum X_{i2} + B_1 \sum X_{i2} X_{i1} + B_2 \sum X_{i2}^2 = \sum X_{i2} Y_i
\end{aligned}$$

In matrix form (sect 9.1), let

$$\mathbf{y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & X_{11} & X_{12} \\ 1 & X_{21} & X_{22} \\ \vdots & \vdots & \vdots \\ 1 & X_{n1} & X_{n2} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} A \\ B_1 \\ B_2 \end{pmatrix}$$

Then (repeating the exact same notation we used for the simple linear regression in matrix form),

$$RSS = \|\boldsymbol{\varepsilon}\|^2 = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 = (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

So

$$\frac{\partial \mathrm{RSS}}{\partial \boldsymbol{\beta}} = 2\mathbf{X}^T (\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$$

and the normal equations are obtained by setting this to zero:

$$\mathbf{X}^T \mathbf{X}\boldsymbol{\beta} = \mathbf{X}^T \mathbf{y}$$

so $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$.

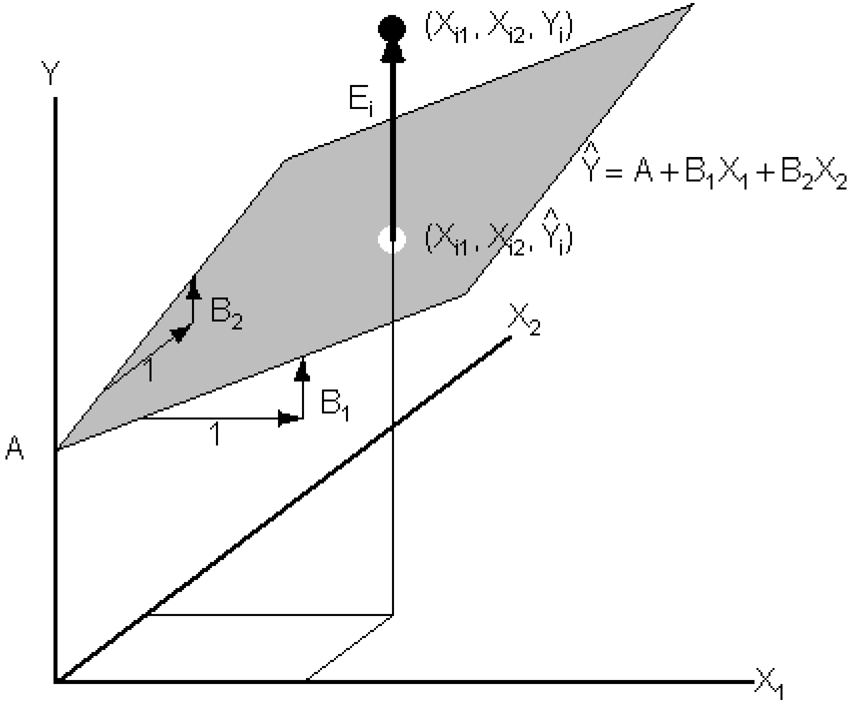
Note that the elements of $\mathbf{X}^T \mathbf{X}$ are the terms multiplying A, B_1, B_2 in the normal equations as written above:

$$\mathbf{X}^T \mathbf{X} = \begin{pmatrix} n & \sum X_{i1} & \sum X_{i2} \\ \sum X_{i1} & \sum X_{i1}^2 & \sum X_{i1} X_{i2} \\ \sum X_{i2} & \sum X_{i2} X_{i1} & \sum X_{i2}^2 \end{pmatrix}$$

And the right hand side of the normal equations is

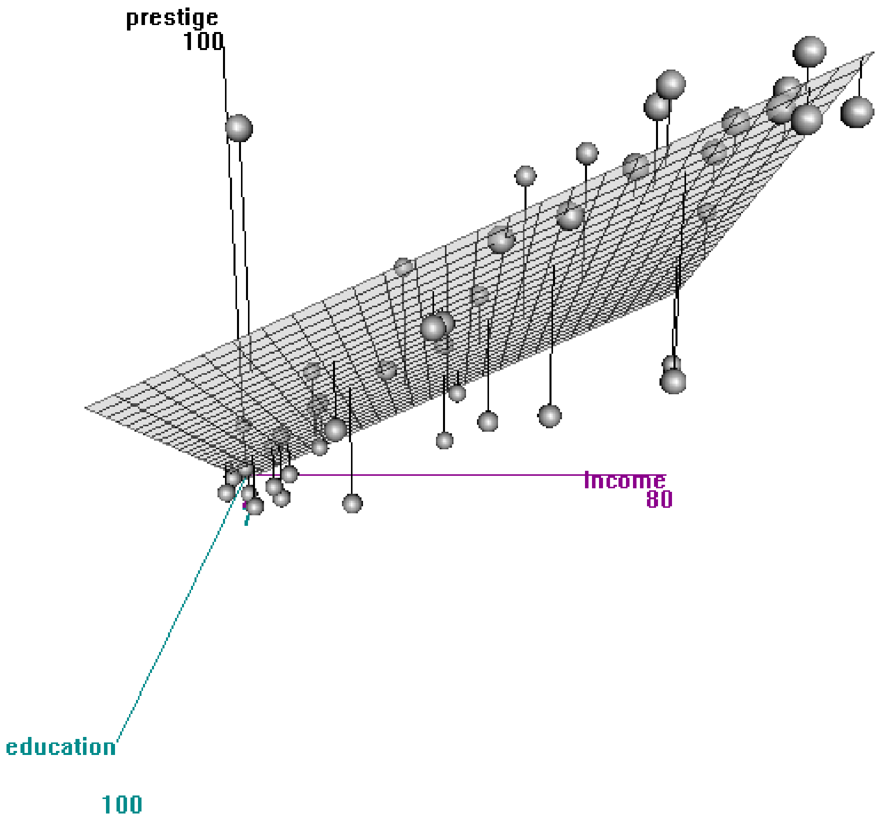
$$\mathbf{X}^T\mathbf{y} = \begin{pmatrix} \sum Y_i \\ \sum X_{i1} Y_i \\ \sum X_{i2} Y_i \end{pmatrix}$$

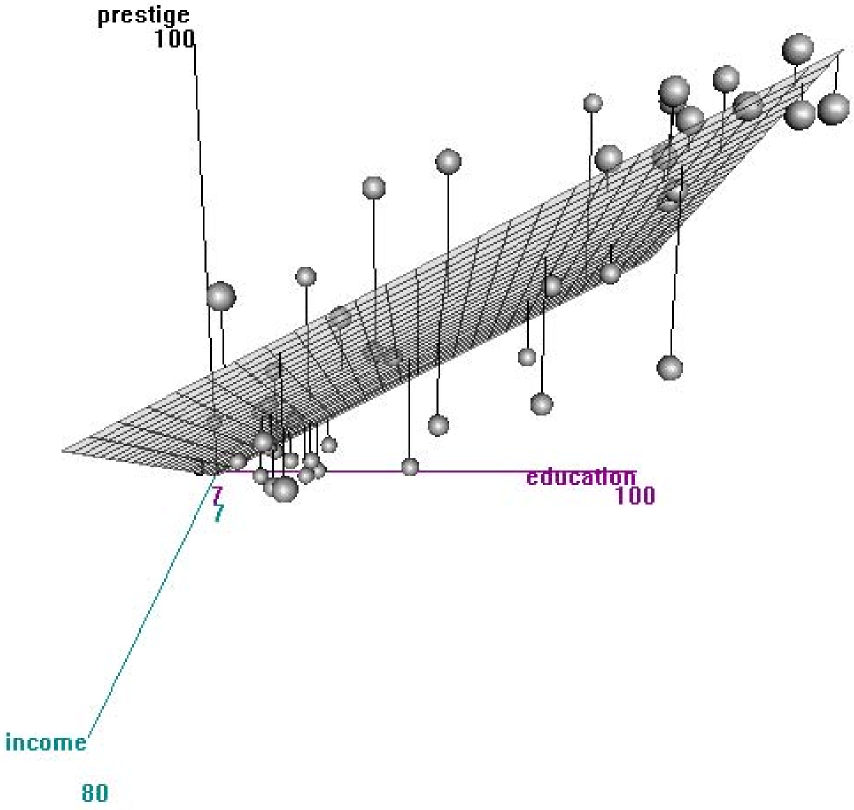
Figure 5.6 The multiple-regression plane, showing the partial slopes *B_1* and *B_2* and the residual *E_i* for the i$^{\text{th}}$ observation.



(X_{i1}, X_{i2}, Y_i)
Y
E_i
$\hat{Y} = A + B_1X_1 + B_2X_2$
$(X_{i1}, X_{i2}, \hat{Y}_i)$
B_2
X_2
1
B_1
1
A
X_1

Duncan occupation prestige example.

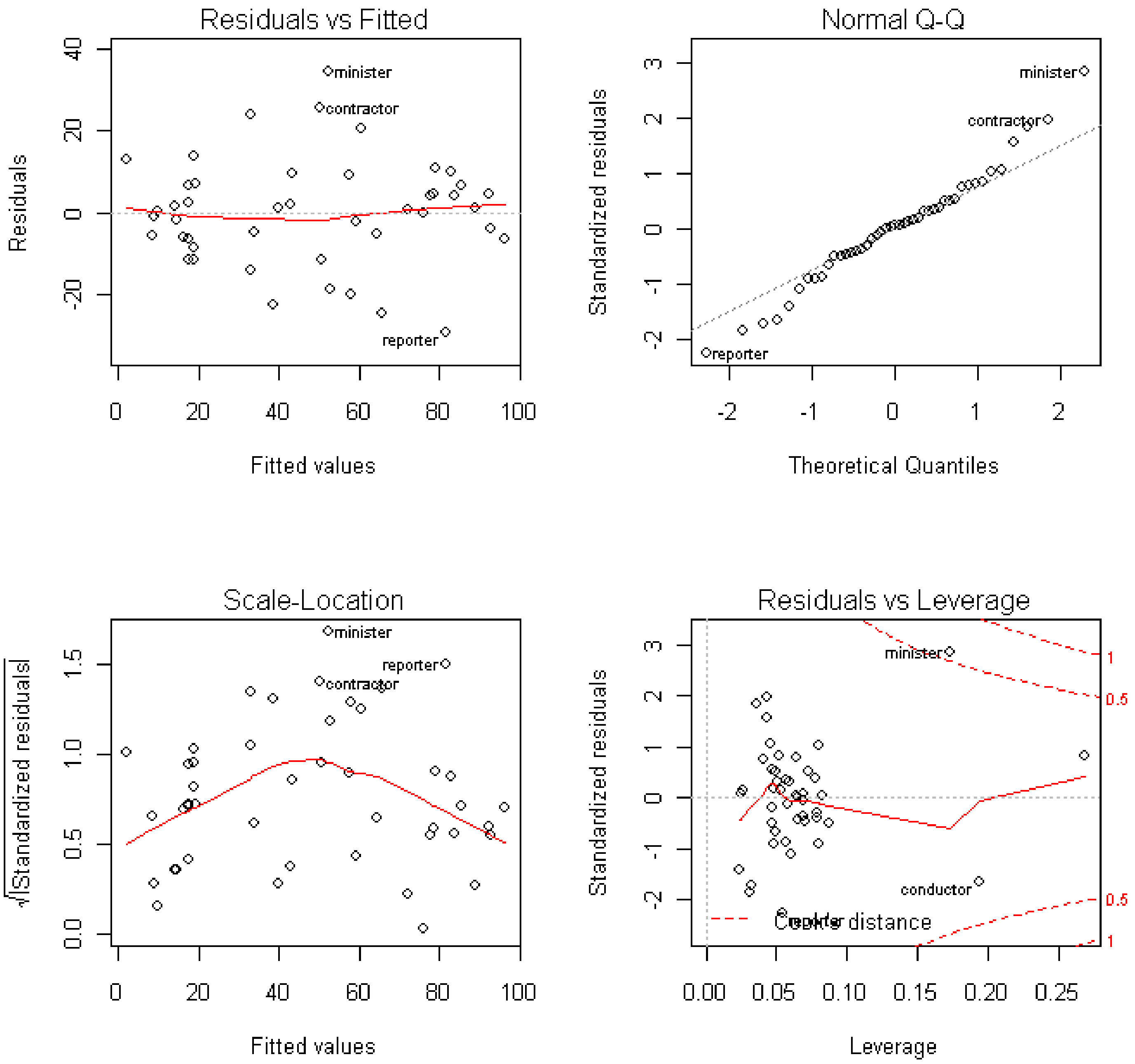




These are 2 different perspectives generated using the “`scatter3d`” function in the `Rcmdr` package. You can also “spin” the picture with dynamic rotations, but I can’t get it to work on my Mac (problem with the required `tcltk` package for rgl device driver).

```
> scatter3d(income, prestige, education, surface.col="gray", pos.res.col="black",
+     neg.res.col="black", point.col="gray", fogtype="none",revolutions=2)
> scatter3d(education, prestige, income, surface.col="gray", pos.res.col="black",
+     neg.res.col="black", point.col="gray", fogtype="none",revolutions=2)
>
> Tlmfit.fig5.7 <- lm( prestige ~ income + education, data=Duncan )
> summary(Tlmfit.fig5.7)

Call:
lm(formula = prestige ~ income + education, data = Duncan)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.06466    4.27194  -1.420    0.163
income       0.59873    0.11967   5.003 1.05e-05 ***
education    0.54583    0.09825   5.555 1.73e-06 ***
---

Residual standard error: 13.37 on 42 degrees of freedom
Multiple R-squared: 0.8282,     Adjusted R-squared:  0.82
F-statistic: 101.2 on 2 and 42 DF,  p-value: < 2.2e-16

> options(digits=2)
> anova(Tlmfit.fig5.7)

Analysis of Variance Table

Response: prestige
          Df Sum Sq Mean Sq F value  Pr(>F)
income     1  30665   30665   171.6 < 2e-16 ***
education  1   5516    5516    30.9 1.7e-06 ***
Residuals 42   7507     179
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
> windows()
> par(mfrow=c(2,2))
> plot(Tlmfit.fig5.7)
```



5.2.3 Standard error and multiple correlation

Standard error of the regression

$$s_E = \sqrt{\frac{\sum E_i^2}{n-k-1}}$$

where k is the number of independent variables in the regression; we lose $k+1$ df as there are $k+1$ coefficients, including the intercept. For the Duncan prestige data example above, $s_E = 13.37$. This is in the units of the response variable and it is directly interpretable. For this example Fox says:

"Recall that the response variable here is the percentage of raters classifying occupation as good or excellent in prestige; an average prediction error of 13 is substantial given Duncan's purpose, which was to use the regression equation to calculate substitute prestige scores for occupations for which direct ratings were unavailable."

The ANOVA decomposition is exactly as it was before

$$\text{TSS} = \text{RegSS} + \text{RSS}$$

where

$$\begin{aligned} \text{TSS} &= \sum \left(Y_i - \bar{Y}\right)^2 \\ \text{RegSS} &= \sum \left(\hat{Y}_i - \bar{Y}\right)^2 \\ \text{RSS} &= \sum \left(Y_i - \hat{Y}\right)^2 \end{aligned}$$

What we called just a squared correlation before, r^2, we now call the *squared multiple correlation*

$$R^2 \equiv \frac{\text{RegSS}}{\text{TSS}}$$

This is interpreted as "*the proportion of the variation in the response variable explained by the regression*". It can also be shown to be the square of the (usual) correlation coefficient computed between the $\{Y_i\}$ and the $\{\hat{Y}_i\}$, $r_{Y\hat{Y}}$, hence the name *squared multiple correlation*.

Note that we can also write

$$R^2 \equiv \frac{\text{TSS} - \text{RSS}}{\text{TSS}} = 1 - \frac{\text{RSS}}{\text{TSS}}$$

Because R^2 can only increase as you add more variables to a regression model, R^2 can be misleadingly large in a problem with k somewhat large compared to n. You will sometimes see reports of an "adjusted R^2" with a "correction" for degrees of freedom.

$$R^2 \equiv 1 - \frac{\text{s}_E^2}{s_Y^2} = 1 - \frac{\text{RSS}/(n-k-1)}{\text{TSS}/(n-1)}$$

5.2.4 Standardized regression coefficients

Regression coefficients are expressed in the units of the response variable relative to the units of the corresponding explanatory variable. Therefore we can't easily compare coefficients B_1 and B_2 when the corresponding explanatory variables are not measured in the same units.

- In the `Duncan` prestige dataset on US occupations (in 1950!), all the variables are expressed on percentage scales, so this isn't really an issue. (income: % of males earning > $3500; education: % of males who were high school graduates; prestige: % of raters who classified the occupation as having good or excellent prestige).

- In the `Prestige` dataset on prestige of Canadian occupations, the response prestige is on a points scale while education is measured in years and income in dollars. So the coefficient B_1 multiplying education has units of "points of prestige per year of education".

Social scientists often *standardize* the variables in order to directly compare regression coefficients.

$$Y_i - \bar{Y} = B_1(X_{i1} - \bar{X}_1) + \cdots + B_k(X_{ik} - \bar{X}_k) + E_i$$

Manipulate this equation by dividing both sides by the standard deviation of the response variable:

$$\frac{Y_i - \bar{Y}}{s_Y} = \left(B_1 \frac{s_1}{s_Y}\right)\frac{(X_{i1} - \bar{X}_1)}{s_1} + \cdots + \left(B_k \frac{s_k}{s_Y}\right)\frac{(X_{ik} - \bar{X}_k)}{s_k} + \frac{E_i}{s_Y}$$

or

$$Z_{iY} = B_1^* Z_{i1} + \cdots + B_k^* Z_{ik}$$

where $Z_{iY} = (Y_i - \overline{Y}) / s_Y$ is the *standardized response* variable and $Z_{ij} = (X_{ij} - \overline{X}_j)$ are *standardized explanatory* variables. The coefficients $B_j^* \equiv B_j(s_j / s_y)$ are called *standardized partial regression coefficients*.

Note that we have not changed the regression model in any important way.

- Has the multiple correlation coefficient changed?
- Has the standard error of the regression changed after this standardization?

Rather, as Fox says:

“By rescaling regression coefficients in relation to a meaure of variation---such as the IQR or standard deviation---standardized regression coefficients permit a *limited* comparison of the relative impact of incommensurable explanatory variables.”

Keep in mind that:

- The scaling is based on *sample* standard deviations, so the nature of the scaling, and hence the interpretation of these coefficients, is very much depending on the study design---that ranges of values of the variables that are represented in the sample. Samples for different populations can produce different results and interpretations.
- The language “comparison of the relative impact of incommensurable explanatory variables” suggests that (a) the explanatory variables have an “impact” (i.e., are causal) and (b) that their impact is accurately represented by this additive linear regression model.