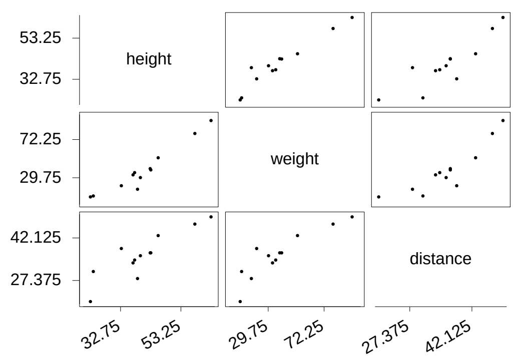
Read M&M Chapter 11 (skip part on logistic regression, pages 730–731).
Read M&M pages 681–686, for ANOVA tables. Multiple regression.

1. Least squares with more than one predictor

height (inch)	weight (lb.)	distance (cm)
42.8	40.0	37.0
63.5	93.5	49.5
37.5	35.5	34.5
39.5	30.0	36.0
45.5	52.0	43.0
38.5	17.0	28.0
43.0	38.5	37.0
22.5	8.5	20.0
37.0	33.0	33.5
23.5	9.5	30.5
33.0	21.0	38.5
58.0	79.0	47.0

The displayed data are borrowed from Chapter 14 of the text *Mathematical Statistics and Data Analysis* by John A. Rice (Duxbury 1995). They show the heights (in inches) and weights (in pounds) of twelve children, with the length (distance, in centimeter) of a path to the child's heart as response variable. As explained by Rice, the third variable was collected using an invasive procedure, which involved insertion of a catheter into a major vein or artery in the femoral region, with distances determined when the tip was seen (by means of a fluoroscope) to reach the pulmonary artery. The aim of the experiment was to see whether the distance could be estimate using only the child's height and weight. If it could, then doctors might be better able to estimate the length of the catheter needed for other children.

A Minitab MatrixPlot shows that the distance is well associated with both height and weight:



Following M&M notation, write y for the vector of distances, x_1 for the vector of heights, and x_2 for the vector of weights. An extra subscript i, will denote the child ($i = 1, \ldots, n$ with $n = 12$). The least squares procedure chooses constants b_0, b_1, b_2 to minimize

$$\sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i1} - b_2 x_{i2})^2$$

The formal interpretation of the output from a least squares fit assumes a model with normal errors: for some unknown constants $\beta_0, \beta_1, \beta_2$, and an unknown $\sigma > 0$,

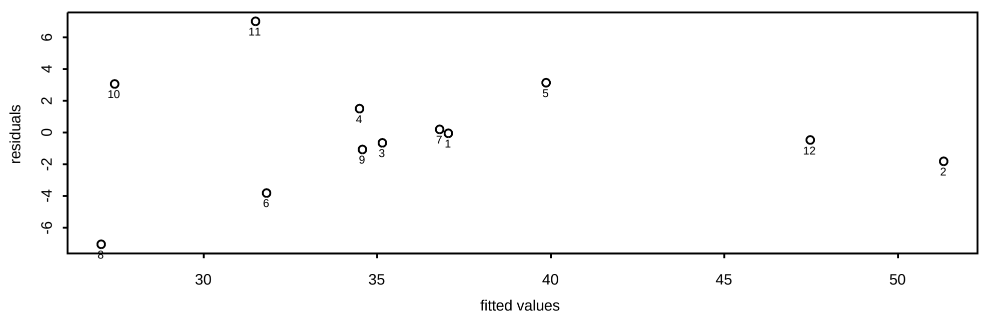
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

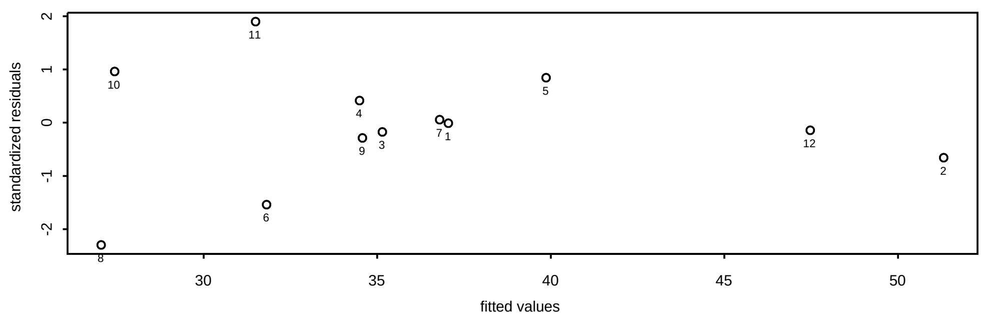
where the ϵ_i's are independent $N(0, \sigma)$ random variables, conditional on x_1 and x_2.

The least squares coefficients define ***fitted values*** and ***residuals*** much as before:

$$\widehat{y}_i = b_0 + b_1 x_{i1} + b_2 x_{i2} \qquad \text{and} \qquad e_i = y_i - \widehat{y}_i$$

If the model is correct, the residuals are independent of the fitted values.† Patterns in the plot of residuals versus fitted values sometimes suggest violations of the model. Some statisticians prefer first to standardize the residuals, dividing them by their estimated standard deviations, before plotting them against the fitted values. For the catheter data, the plots show a slightly curved pattern:





Mintab flags the residual for child 8 as unusually small (negative):

Unusual Observations

Obs	height	distance	Fit	StDev Fit	Residual	St Resid
8	22.5	20.00	27.05	2.48	-7.05	-2.30R

R denotes an observation with a large standardized residual

Perhaps child 8 is a newborn infant.

2. Coefficients and fitted values

As in the case of a single predictor (Lecture 9), the least squares procedure determines coefficients and fitted values as linear combinations of the observed y_i's, with the multipliers depending only on the predictor variables (the heights and weights). It is not necessary to calculate these multipliers explicitly, but they do help explain the behavior of the various quantities associated with the fit.

Here are the multipliers that I calculated for the fitted values:

† Because of the way least squares works, there can be no *linear* association between residuals and fitted values; the sample correlation between them must be zero. However, a zero *linear association* does not rule out other, non-linear relationships. Independence is a much stronger assertion.

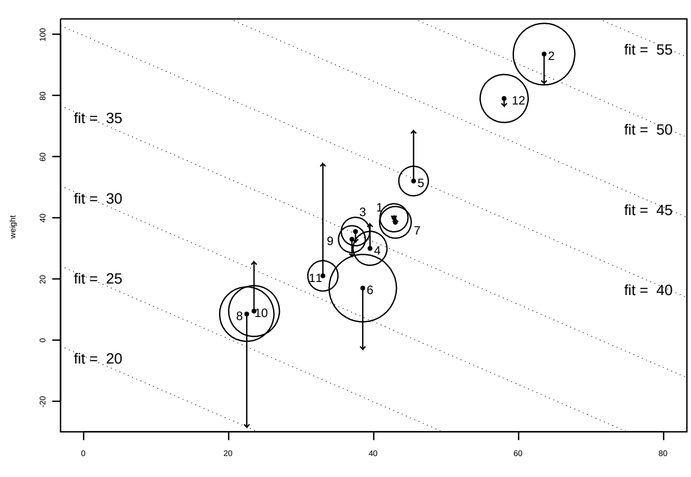
	y_1	y_2	y_3	y_4	y_5	y_6	y_7	y_8	y_9	y_{10}	y_{11}	y_{12}
fit1	0.11	0.08	0.06	0.12	0.07	0.18	0.12	0.01	0.07	0.02	0.08	0.09
fit2	0.08	0.51	0.08	-0.01	0.20	-0.15	0.06	-0.09	0.06	-0.08	-0.05	0.39
fit3	0.06	0.08	0.11	0.05	0.09	-0.02	0.05	0.16	0.10	0.15	0.09	0.07
fit4	0.12	-0.01	0.05	0.15	0.05	0.28	0.14	0.01	0.06	0.02	0.11	0.03
fit5	0.07	0.20	0.09	0.05	0.12	-0.02	0.06	0.07	0.08	0.06	0.05	0.16
fit6	0.18	-0.15	-0.02	0.28	-0.02	0.61	0.24	-0.13	0.03	-0.10	0.14	-0.05
fit7	0.12	0.06	0.05	0.14	0.06	0.24	0.13	-0.02	0.06	-0.01	0.09	0.08
fit8	0.01	-0.09	0.16	0.01	0.07	-0.13	-0.02	0.39	0.15	0.37	0.14	-0.06
fit9	0.07	0.06	0.10	0.06	0.08	0.03	0.06	0.15	0.10	0.14	0.09	0.06
fit10	0.02	-0.08	0.15	0.02	0.06	-0.10	-0.01	0.37	0.14	0.34	0.14	-0.06
fit11	0.08	-0.05	0.09	0.11	0.05	0.14	0.09	0.14	0.09	0.14	0.12	-0.01
fit12	0.09	0.39	0.07	0.03	0.16	-0.05	0.08	-0.06	0.06	-0.06	-0.01	0.31

For example,

$$\begin{aligned}\widehat{y}_6 &= 0.18y_1 - 0.15y_2 - 0.02y_3 + 0.28y_4 - 0.02y_5 + 0.61y_6 \\ &\quad + 0.24y_7 - 0.13y_8 + 0.03y_9 - 0.10y_{10} + 0.14y_{11} - 0.05y_{12}\end{aligned}$$

Notice the coefficient of y_6. The value of y_6 has quite a lot of influence on the fitted value $\widehat{y}_6$. In general it is worth taking special note of the numbers down the diagonal of the table of multipliers, which are called ***influence values***. Values close to 1 flag observations that have a lot of influence over the corresponding fitted values. An observation with infuence 1 is guaranteed to have a zero residual; plots of residuals would provide no clue to bad behavior for such a y_i.

In the next picture, the area of the the circle round each (height,weight) data point is proportional to the influence value for that point. Larger circles indicate points that have more influence on the fit. The dotted lines indicate the contours of the function $b_0 + b_i x_1 + b_2 x_2$; they indicate fitted values for the various height and weight combinations. The arrows indicate residuals from fitted values. The heads of the arrows correspond to the observed distances. For example, the arrow for child 8 reaches from the point height = 22.5 inches, weight = 8.5 pounds to the value that corresponds to distance = 20.0 cm on the dotted grid.



3. Normal distributions

Once again, the crucial consequence of the linearity in the y_i's **under the assumed model** is that

$$\begin{aligned}&\widehat{y}_i \text{ has a } N(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}, C_i\sigma) \text{ distribution} \\ &b_j \text{ has a } N(\beta_j, K_j\sigma) \text{ distribution,}\end{aligned}$$

where the C_i and K_j constants are determined by the height and weight variables. The residuals behave somewhat like the ϵ_i's, except for the fact that the fitting procedure imposes three linear constraints. The standard deviation σ is estimated by the quantity

$$S = \sqrt{\frac{\sum_i e_i^2}{n-3}} \qquad \text{with } n = 12 \text{ for the catheter data}$$

(M&M write s instead of S.) The next display shows where these values contribute to the Minitab output.

The regression equation is: distance $= 21.0 + 0.196\,\text{height} + 0.191\,\text{weight}$

Predictor	Coef	StDev	T	P
Constant	$21.008 = b_0$	$8.751 = K_0 S$	$2.40 = b_0/(K_0 S)$	$0.040 = \mathbb{P}\{\lvert t_9\rvert \geq 2.40\}$
height	$0.1964 = b_1$	$0.3606 = K_1 S$	$0.54 = b_1/(K_1 S)$	$0.599 = \mathbb{P}\{\lvert t_9\rvert \geq 0.54\}$
weight	$0.1908 = b_2$	$0.1652 = K_2 S$	$1.16 = b_2/(K_2 S)$	$0.278 = \mathbb{P}\{\lvert t_9\rvert \geq 1.16\}$

S = 3.943 R-Sq = 80.5% R-Sq(adj) = 76.2

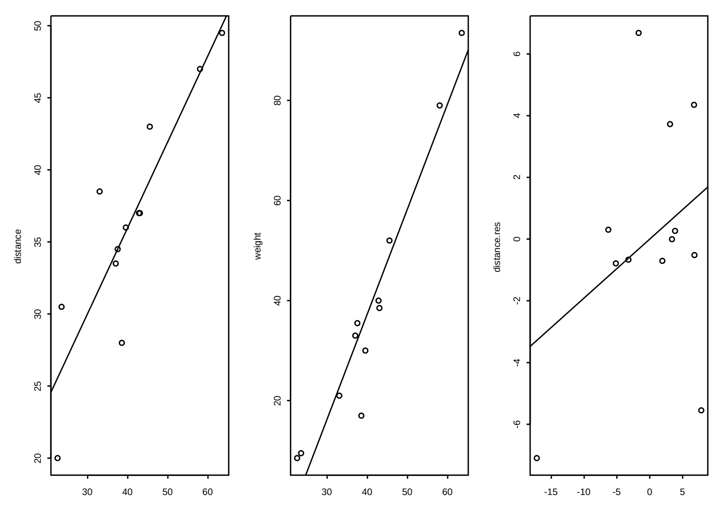
Be careful with the interpretation of the p-values. They can be used to test hypotheses of the form $\beta_j = 0$ within the context of the full model. For example, the value $T = 1.16$ in the weight row would be an observation from a t-distribution with 9 degrees of freedom if $y_i = \beta_0 + \beta_1 x_{i1} + \epsilon_i$, with the ϵ_i's independent $N(0, \sigma)$'s. In effect, the p-value is speaking to the question of whether, *after the distances have been adjusted for heights*, there is still a weight effect. A large value for $|T|$ (which would give a small value for P) would suggest that the weight term is accounting for a systematic effect beyond what is accounted for by the heights.

The p-value of 0.278 suggests that the weight term is doing little more than soaking up noise effects, once the distances are adjusted for heights.

Similarly, the p-value 0.599 suggests that the height term contributes little to the fit after the distances are adjusted for weights.

It would be a blunder to conclude that neither height nor weight is helpful for predicting distance. The t-statistic for each predictor is calculated assuming that the other predictor is already included in the fit.

4. Least squares one step at a time



Conceptually, and mathematically, regression of distance on height and weight is equivalent to a two step procedure:

(i) Regress both distance and weight on height, obtaining two sets of residuals, resid.dist and resid.weight, say.

(ii) Regress resid.distance on resid.weight.

The residuals from step (ii) are the same as the residuals that would be obtained from regressing distance on both height and weight simultaneously.

5. Interpretation of the Analysis of Variance

If we fit the predictors in the order "weight then height" rather than "height then weight" we get the same coefficients, the same residuals, and the same t-statistics. Minitab also produces tables giving an "analysis of variance". The main table decomposes the "variability" in the y_i's, as measured by $\sum_i (y_i - \overline{y})^2$, into a regression sum of squares $\sum_i (\widehat{y}_i - \overline{y})^2$ plus $\sum_i e_i^2$, the variablity that is left after the height and weight effects are taken out. The regression sum of squares represents the variability in the fitted values, which is due to both a noise effect and to the variabilty in the theoretical means $\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$.

If $\beta_1 = \beta_2 = 0$, the regression sum of squares would be derived from pure noise terms with no systematic parts due to differences in the theoretical means, in much the same way that the t-statistics represent noise when the b_j have zero expected values. With no systematic contribution, the regression sum of squares would be distributed as σ^2 times a $\chi^2(2)$ random variable. We estimate σ^2 by S^2. The ratio (regression sum of squares/2)$/S^2$ would have an ***F-distribution*** if $\beta_1 = \beta_2 = 0$. The p-value in the last column gives the probability that a random variable with such an F-distribution would be larger than the observed 18.62. A small p-value inclines us to doubt that the regression sum of squares was generated from pure noise, which leads us to reject the hypothesis that $\beta_1 = \beta_2 = 0$.

Mintab will also break the regression sum of squares into two components, one due to the height contribution to the variability in the distances, the other due to the weight contribution. Notice that order matters. Notice also that

$$578.82 \approx 558.06 + 20.75 = 574.21 + 4.61$$

The small discrepancy is due to round-off error.

Analysis of Variance (for taking out the height effect before the weight effect)

Source	DF	SS	MS	F	P
Regression	2	$578.82 = \sum_i (\widehat{y}_i - \overline{y})^2$	$289.41 = 578.82/2$	$18.62 = 289.41/15.55$	0.001
Residual Error	9	$139.91 = \sum_i e_i^2$	$15.55 = 139.91/9 = S^2$		
Total	11	$718.73 = \sum_i (y_i - \overline{y})^2$			

Source	DF	Seq SS
height	1	558.06
weight	1	20.75

Analysis of Variance (for taking out the weight effect before the height effect)

Source	DF	SS	MS	F	P
Regression	2	578.82	289.41	18.62	0.001
Residual Error	9	139.91	15.55		
Total	11	718.73			

Source	DF	Seq SS
weight	1	574.21
height	1	4.61