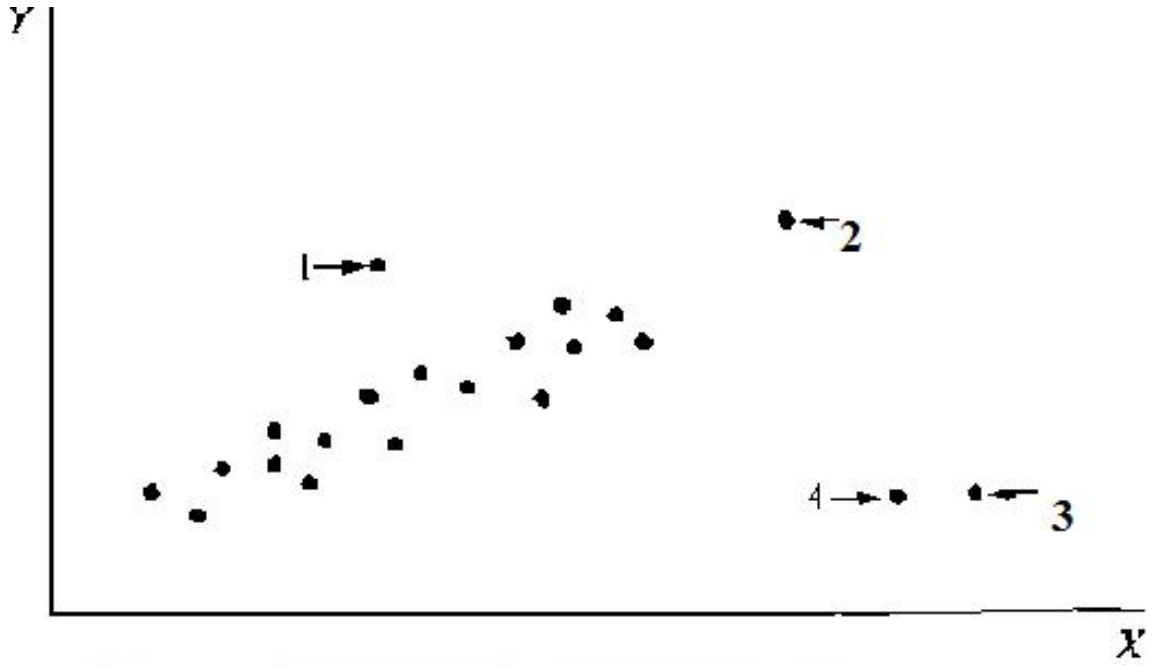
Chapter 10 Diagnostics

- Added variable (or partial regression) plots
- Testing for Y-outliers (Studentized deleted residuals)
- Identifying X-outliers (Hat matrix diagonals)
- Influential observations (Dffits, Cook's D, DFBETAS)
- Multicollinearity (VIF: Variance inflation factor)
- True residuals iid $N(0, \sigma^2)$

Figure 1: Example of outliers



Partial regression plots (Added variable plots)

- Evaluate the effect of x_i on y, given that other variables are in the model
- One plot for each x_i. To get the plot, run two regressions.

  — In the first, use the other x's to predict y .

  — In the second use the other x's to predict x_i.

  — Then plot the residuals from the first regression say $e\big(y|x_2, x_3\big)$ against the residuals from the second regression say $e\big(x_1|x_2, x_3\big)$. (Note: The correlation of these residuals was called the partial correlation coefficient.)

- $e(y|x_2, x_3)$ is the part of y that is orthogonal to (not explained by) x_2 and x_3.
  $e(x_1|x_2, x_3)$ is the part of x_1 that is orthogonal to (not explained by) x_2 and x_3.

- A linear pattern in this type of plot indicates that the variable would be useful in the model, and the slope is its regression coefficient.

  —The plots show the strength of a marginal relationship between y and x_i in the full model.

  — If the partial residual plot for x_i appears "flat", x_i may not need to be included in the model.

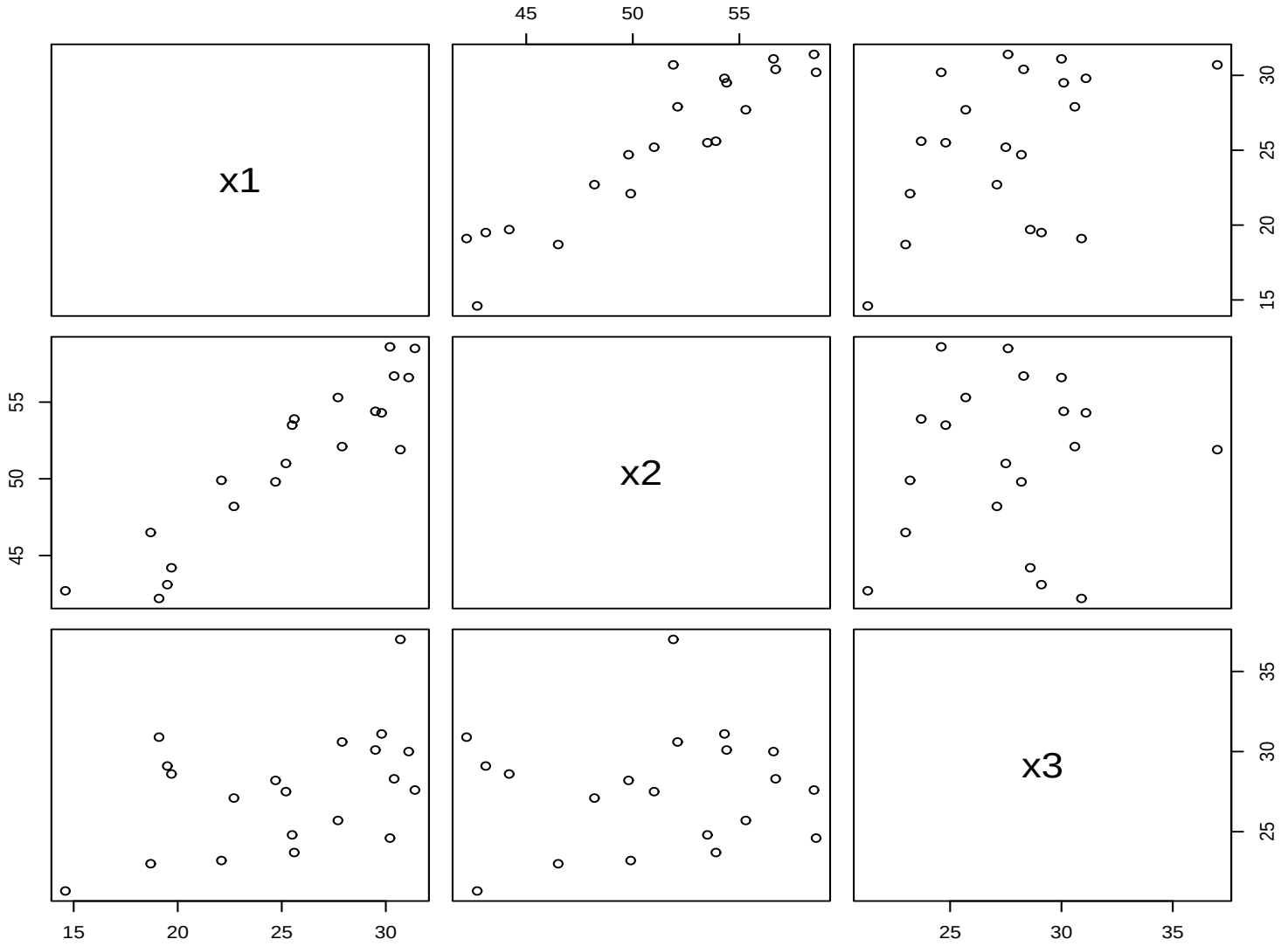
  — If it appears like a straight line (with non-zero slope), it suggests x_i should be included as a linear term.

- Nonlinear relationships, heterogeneous variances, and outliers may also be detected in these plots.

Example: Body fat data

- The data is a portion of data for a study of the relation of amount of body fat (y) to several possible predictor variables, based on a sample of 20 healthy females $25 - 34$ years old.
- Predictor variables are triceps skinfold thickness ($x1$), thigh circumference ($x2$), and midarm circumference ($x3$).
- Response variable is y. The amount of body fat for each of the 20 persons was obtained by a cumbersome and expensive procedure requiring the immersion of the person in water.

Figure 2: scatterplot of three predictor variables—body fat example



- Correlation matrix of $\mathbf{X}$ variables

$$r_{\mathbf{xx}} = \begin{bmatrix} 1.0 & .924 & .458 \\ .924 & 1.0 & .085 \\ .458 & .085 & 1.0 \end{bmatrix}$$

	Estimator	se	t-value	p-value
(Intercept)	117.085	99.782	1.173	0.258
x1	4.334	3.016	1.437	0.170
x2	-2.857	2.582	-1.106	0.285
x3	-2.186	1.595	-1.370	0.190

Notice that the overall F statistic is 21.52 on 3 and 16 DF, p-value: 7.343e-06. But none of the individual t's are significant. This indicates multicollinearity problem.

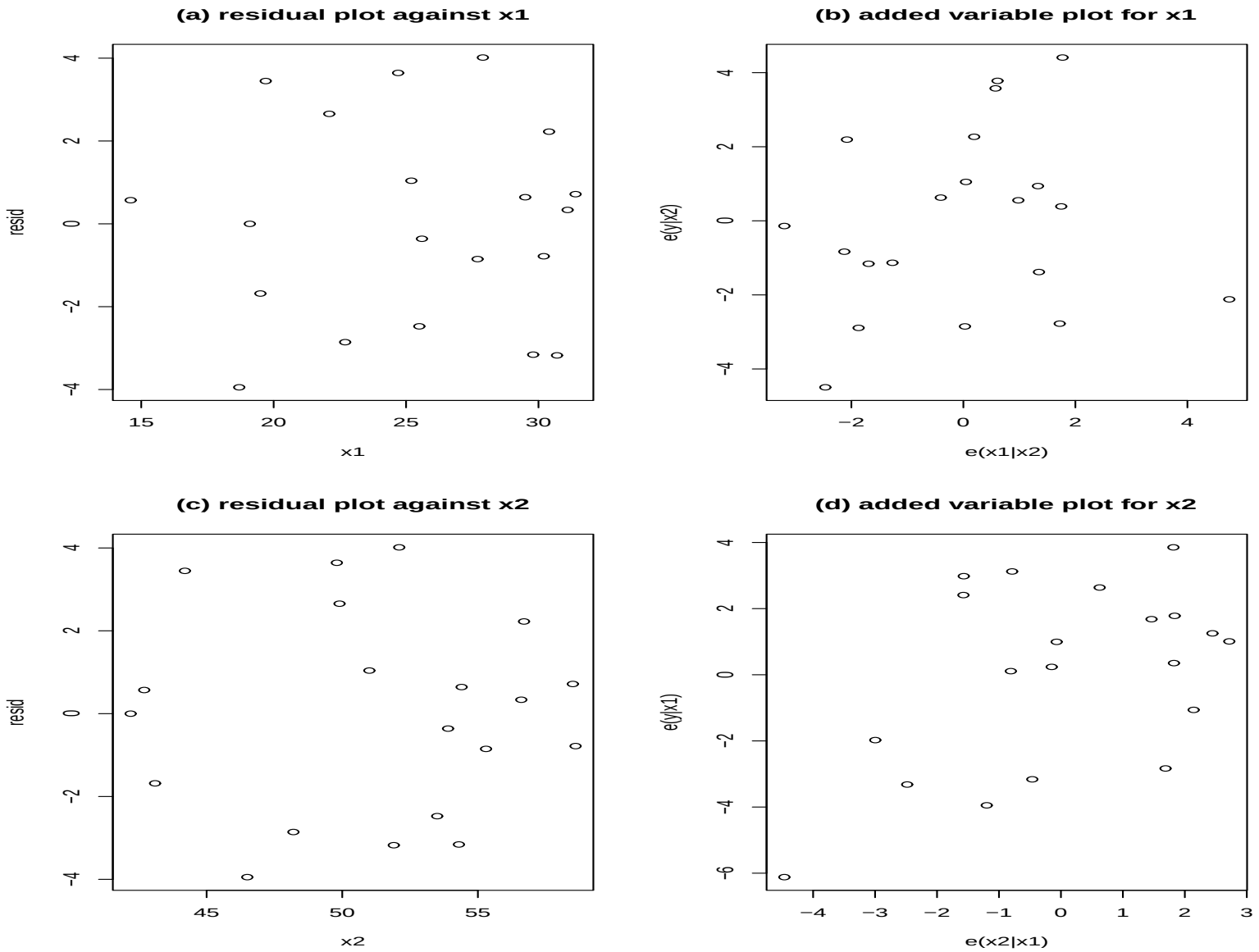
Example 2 (page 388)

Consider the regression of body fat (Y) only on triceps skinfold thickness (X_1) and thigh circumference (X_2).

- X_1 and X_2 are highly correlated ($r_{12} = 0.92$)
- Fitted regression function is

$$\hat{y} = -19.174 + 0.224x_1 + 0.6594x_2$$

Figure 3: Residual plots and added variable plots —body fat example with two predictor variables



Comments:

- Plots a and c do not indicate any lack of fit for the linear terms in the regression model or the existence of unequal variances of the error terms.

- Figures b and d contain the added-variable plots for x_1 and x_2 respectively. When the other predictor variable is already in the regression model.

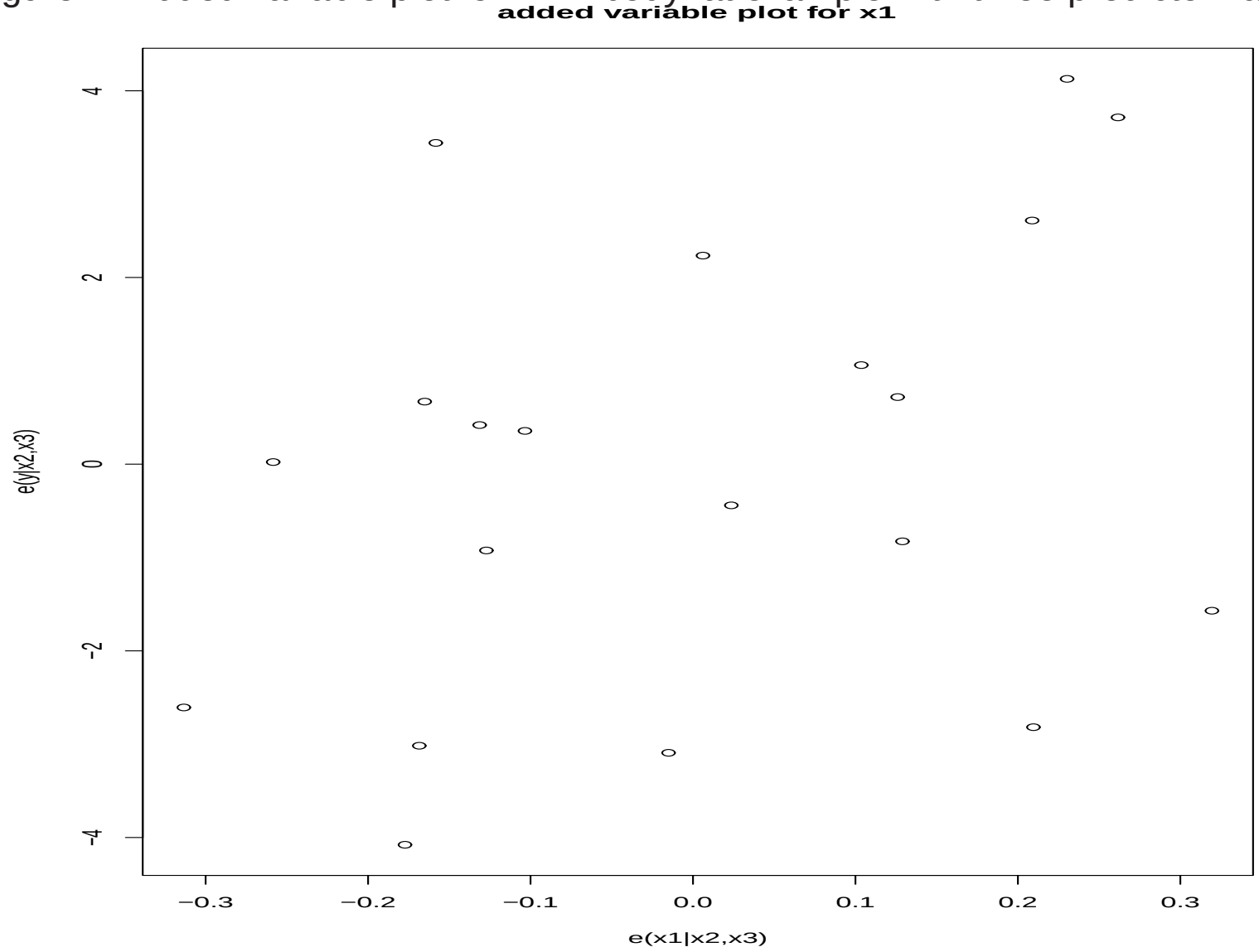
  —Both plots also show the line through the origin with slope equal to the regression coefficient for the predictor variable if it were added to the fitted model.

  —Figure b suggests that x_1 is of little additional help in the model when x_2 is already there. $R^2_{Y1|2} = 0.031$

  —Figure d suggests that x_2 is of little additional help in the model when x_2 is already there.

Now consider adding x_1 to the model that already contains x_2 and x_3

Figure 4: Added variable plot for x1 —body fat example with three predictor variables



Comments:

- Figure 3 suggests that x_1 is of little additional help in the model when x_2 and x_3 are already in the model.
- For Bodyfat example

$$\text{Body fat} = 117 + 4.33x_1 - 2.86x_2 - 2.19x_3 \qquad (1)$$

The regression equation for the residuals is

$$e(y|x_2, x_3) = -0.000 + 4.33e(x_1|x_2, x_3) \qquad (2)$$

- Notice that b_1 in (2) is the same as b_1 in the multiple regression (1). The t and p-values are essentially the same.

- The i-th slope b_i is the slope of the simple linear regression of the part of y that is orthogonal to all other predictors against the part of x_i that is orthogonal to all of the other predictors.
- If you are unsure about whether or not to include a particular variable, examine the added variable plot. It will tell you visually how strong the marginal relation between x_i and the response is.

Testing for outliers

- To do so we will need to figure out the sampling distribution of the estimated residuals.
- We know the distribution of the true residuals, if the assumptions are met. But what about the estimated residuals? Mean zero? Constant variance? Independent? Normal?

Recall

- $\hat{\mathbf{Y}} = \mathbf{HY}$
- $\mathbf{e} = \mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$

Variance and Covariance

$\text{Var}(e_i) = \sigma^2(1 - h_{ii})$

$\text{Cov}(e_i, e_j) = -h_{ij}\sigma^2, i \neq j$

Estimated variance and covariances

$s^2\{e_i\} = \text{MSE}(1 - h_{ii})$

$s\{e_i, e_j\} = -h_{ij}(\text{MSE}), i \neq j$

In general:

- Residuals have non-constant variance
- Residuals have mean zero
- Residuals are not independent
- Residuals are normally distributed

Types of Residuals

Raw residuals: $e_i = y_i - \hat{y}_i$

Semistudentized residuals: $e_i^* = \dfrac{e_i}{\sqrt{MSE}}$

Studentized residuals:

$$r_i = \frac{e_i}{s\{e_i\}} = \frac{e_i}{\sqrt{MSE \times (1 - h_{ii})}}$$

- the residuals e_i will have substantially different sampling variations if their standard deviations differ markedly
- the studentized residuals r_i have constant variance (when the model is appropriate)
- studentized residuals often are called internally studentized residuals

Deleted residuals: $d_i = y_i - \hat{y}_{i(i)}$

- If the point is an outlier, you don't want it messing up the fitted regression line, which will in turn mess up the calculation of the residual
- Better approach-use deleted residuals:
- Delete case i and refit the model using the remaining $n - 1$ cases. Compute the predicted value $\hat{y}_{i(i)}$ and residual d_i for case i using this model.
- An algebraically equivalent expression for d_i that does not require a recomputation of the fitted regression function omitting the ith case is

$$d_i = \frac{e_i}{1 - h_{ii}}$$

where e_i is the ordinary residual for the ith case and h_{ii} is the ith diagonal element in the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$.

Identifying Outlying y observations: Studentized deleted residual

$$t_i = \frac{d_i}{s\{d_i\}}$$

or equivalently,

$$t_i = \frac{e_i}{\sqrt{MSE_{(i)}(1 - h_{ii})}}$$

- Studentized deleted residual t_i is also called an externally studentized residual
- $$(n-p)MSE = (n-p-1)MSE_{(i)} + \frac{e_i^2}{1-h_{ii}}$$
- $$t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2}$$

Bonferroni Outlier Test

$$t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2}$$

$\alpha^* = \alpha/n$, because you are testing all residuals for outliers, not testing a specific case!

- Bonferroni critical value is $t(1-\alpha/2n; n-p-1)$
- If $t_i = e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2} \right]^{1/2} > t(1-\alpha/2n; n-p-1)$, we identify the case as a possible outlier based on this test.

Identifying Outlying x observations: Hat Matrix Leverage Values

Hat matrix Fun-Facts

- h_{ii} is sometimes called the leverage of the i-th observation
- $0 \leq h_{ii} \leq 1$ and $\sum_{i=1}^{n} h_{ii} = p$
- h_{ii} is a measure of the distance between the $\mathbf{x}$ values for the ith case and the means of the $\mathbf{x}$ values for all n cases
- $-1 \leq h_{ij} \leq 1, i \neq j$
- $\hat{Y}_i = \sum_{j=1}^{n} h_{ij} Y_j$

  —- $\hat{Y}_i$ is a weighted sum of the Y observations, where the weights are between -1 and 1 and they sum to one.

  —-h_{ii} is the specific weight assigned to Y_i in determining $\hat{Y}_i$

  —- h_{ii} is called the leverage of Y_i.

- A large value of h_{ii} suggests that the i-th case is distant from the center of all $\mathbf{x}$'s. The farther $\mathbf{x}_i$ is from $\mathbf{x}$, the greater the leverage! X-outliers have the biggest effect!

  The average value is

$$\bar{h} = \frac{\sum_{i=1}^{n} h_{ii}}{n} = \frac{p}{n}$$

- Leverage values that are greater than $2p/n$ cause concern
- Leverage values that are greater than $3p/n$ cause (at least mild) consternation
- Cases with leverages near 1 dominate any fitted regression
- Those outlying x values indicated by leverage should be examined carefully because they may have a substantial influence on the regression parameters.

  Note: n is the number of observations in the data and p is the number of regression parameters, including the intercept.

Example

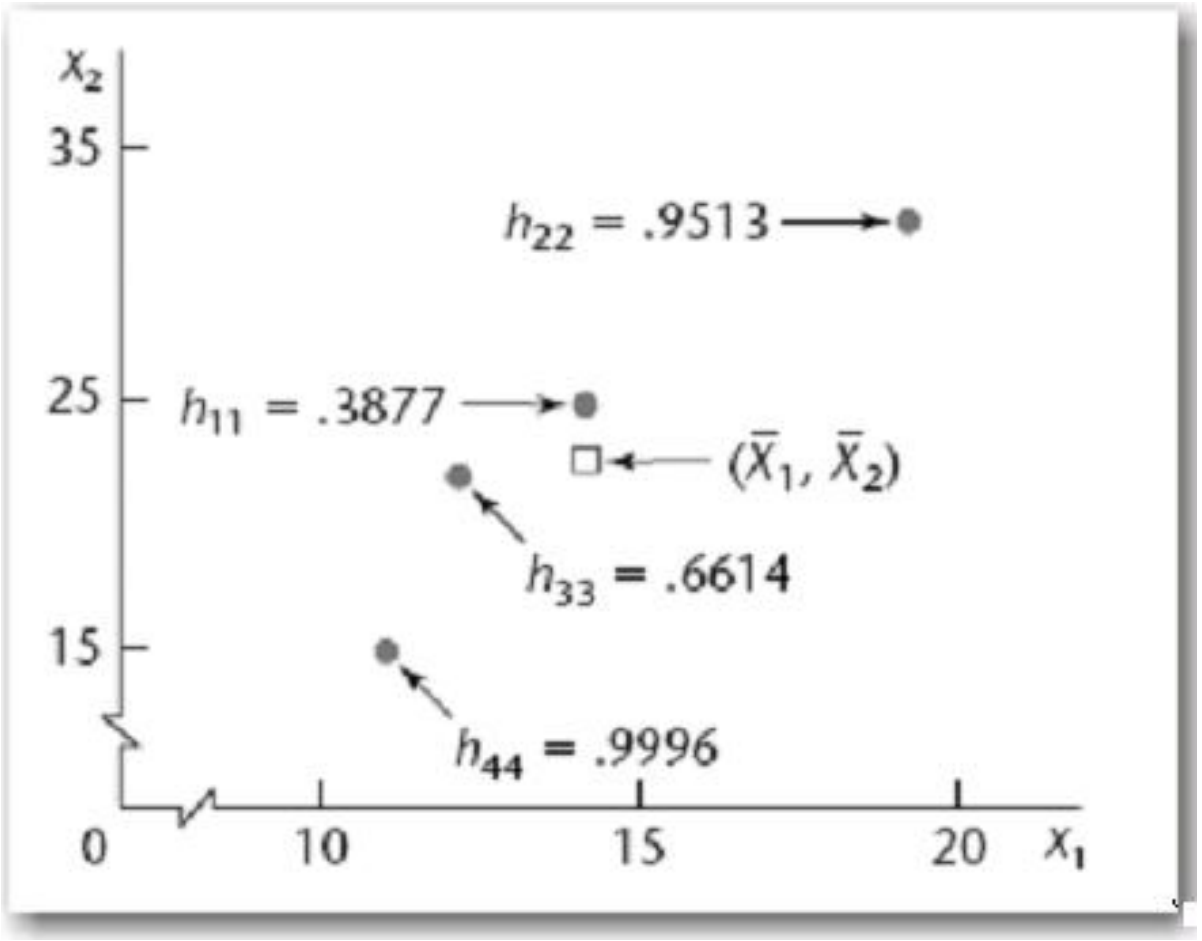
X_{i1}	X_{i2}	Y_i	$\hat{Y}_i$
14	25	301	282.2
19	32	327	332.3
12	22	246	260
11	15	187	186.5

Table 1: Hat Matrix

.3877	.1727	.4553	-.0157
.1727	.9513	-.1284	.0044
.4553	-.1284	.6614	.0117
-.0157	.0044	.0117	.9996

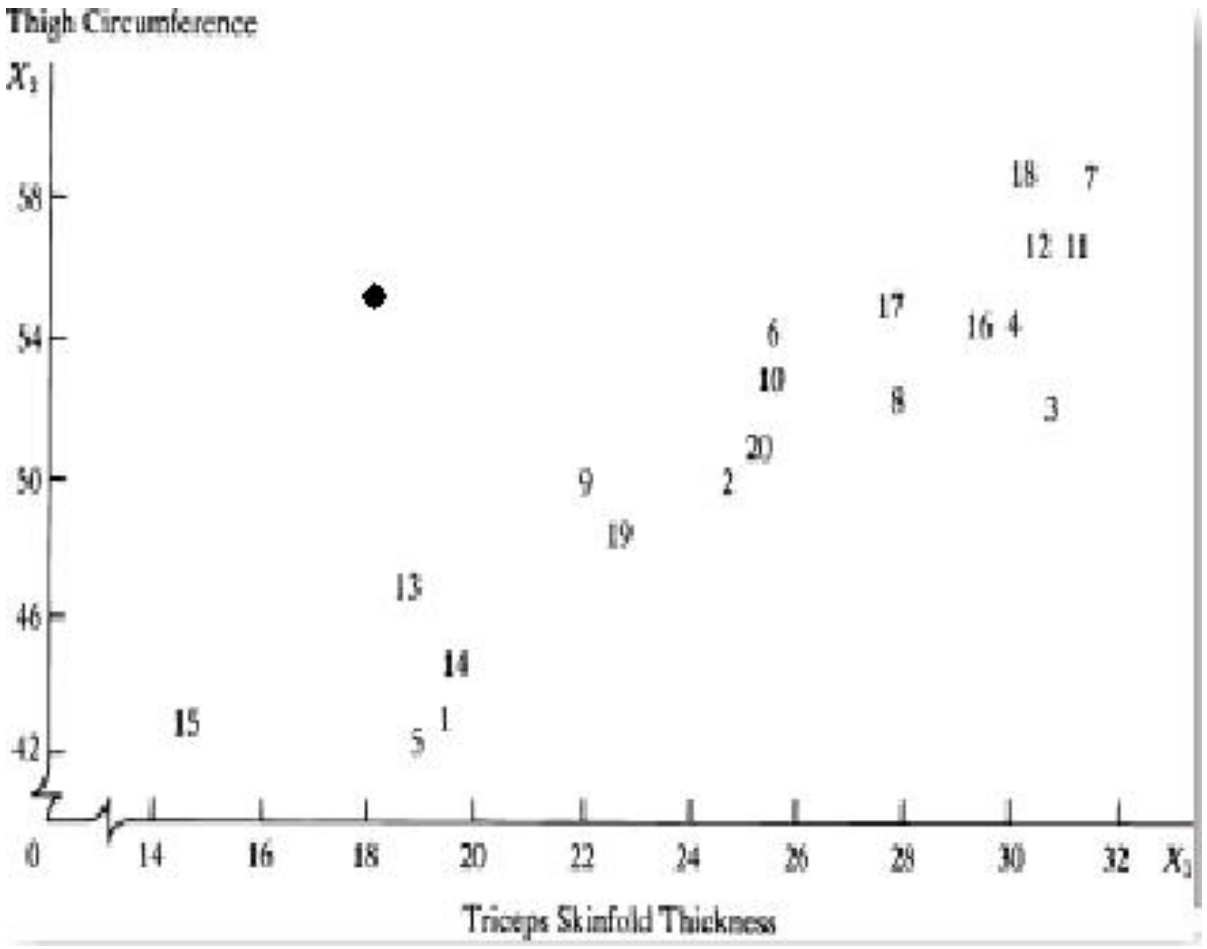
$$\begin{aligned}\hat{Y}_1 &= h_{11}Y_1 + h_{12}Y_2 + h_{13}Y_3 + h_{14}Y_4 \\ &= .3877 * 301 + .1727 * 327 + .4553 * 246 - .0157 * 187 \\ &= 282.2\end{aligned}$$

Figure 5: Leverages for Figure 1: example of outliers



Hidden Extrapolations

Suppose you want to predict at a point X_{new}, and you wonder if you are extrapolating. With more than three predictors you can't tell from a graph:


Thigh Circumference
X_3
58
54
50
46
42
0
14
16
18
20
22
24
26
28
30
32
X_2
Triceps Skinfold Thickness
18
7
12
11
17
16
4
6
10
8
3
20
9
2
19
13
14
15
1
5

Compute

$$h_{new,new} = \mathbf{X}'_{new}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_{new}$$

If this new leverage value is bigger than all of the leverages in the data, this is an extrapolation, because this point is farther from the center than any point in the data set.

Identifying Influential Cases: DFFITS, Cook's D, and DFBETAS Measures

- How can you really tell of a point is influential—that is if it is having an unduly large effect on the results?
- R. Dennis Cook, 1977: Delete the point, refit the regression, and see how much the predicted values change!

DFFITS

- A useful measure of the influence that case i has on the fitted value $\hat{y}_i$ is given by

$$(DFFITS)_i = \frac{\hat{y}_i - \hat{y}_{i(i)}}{\sqrt{MSE_{(i)}h_{ii}}}$$

$$= e_i \left[\frac{n-p-1}{SSE(1-h_{ii}) - e_i^2}\right]^{1/2} \left(\frac{h_{ii}}{1-h_{ii}}\right)^{1/2}$$

- The letters DF stands for the difference between the fitted value $\hat{y}_i$ for the ith case when all n cases are used in fitting the regression function and the predicted value $\hat{y}_{i(i)}$ for the ith case obtained when the ith case is omitted in fitting the regression function

- denominator is the estimated standard deviation of $\hat{y}_i$, but it uses the error mean square when the ith case is omitted in fitting the regression function for estimating the error variance σ^2

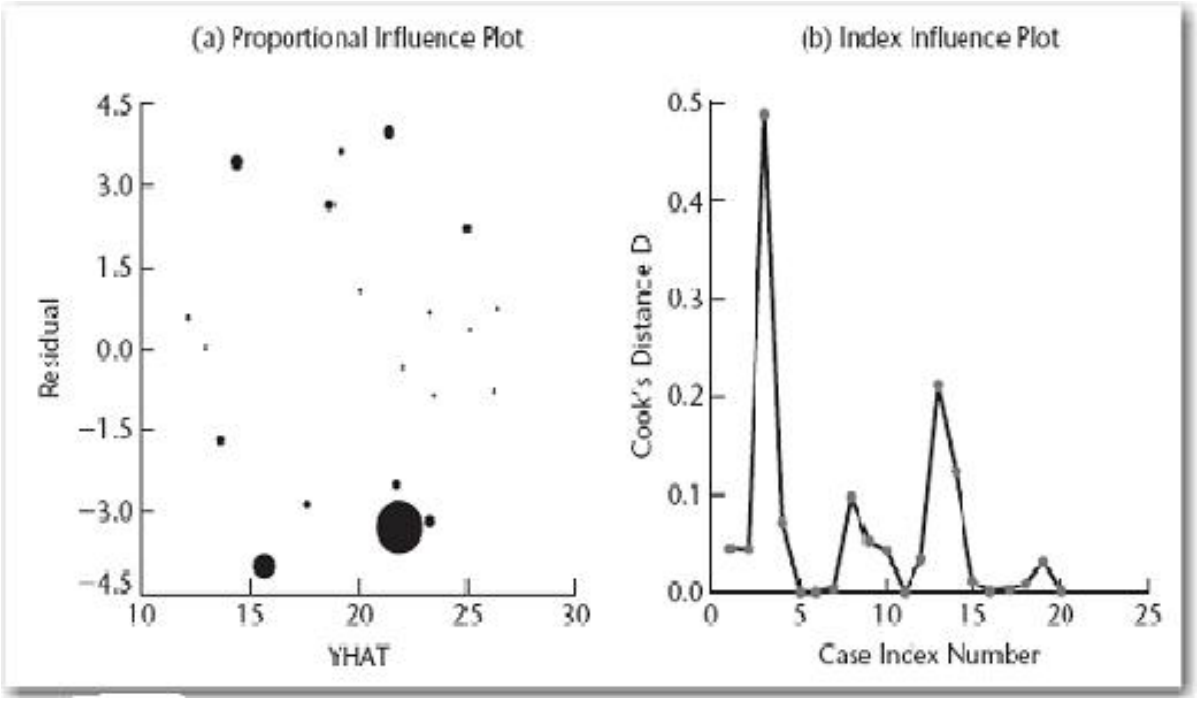
- Values larger than 1 (for small to medium size datasets) or $2\sqrt{p/n}$ (for large datasets) are considered influential.

Cook's distance

$$D_i = \frac{\sum_{j=1}^{n} (\hat{y}_j - \hat{y}_{j(i)})^2}{pMSE}$$

- This measures the influence of case i on all of the fitted $\hat{y}_i$'s
- It is a standardized version of the sum of squares of the differences between the predicted values computed with and without case i.
- Large values (larger than the 50-th percentile of the $F_{p;n-p}$ distribution) suggest an observation has a lot of influence.
- Use index plot to find big numbers.

Figure 6: Index plot: cook's distance



DFBETAS

$$(DFBETAS)_{k(i)} = \frac{b_k - b_{k(i)}}{\sqrt{MSE_{(i)}c_{kk}}}, k = 0, 1, \cdots, p-1$$

where c_{kk} is the kth diagonal element of $(\mathbf{X}'\mathbf{X})^{-1}$.

- A measure of the influence of case i on each of the regression coefficients $b_k (k = 0, 1, \cdots, p-1)$.
- It is a standardized version of the difference between the regression coefficient computed with and without case i.
- Values larger than 1 (for small-to-medium datasets) or $2/\sqrt{n}$ (for large datasets) are considered influential.

Measures of Multicollinearity

Informal Diagnostics

- large changes in the estimated regression coefficients when a predictor variable is added or deleted, or when an observation is altered or deleted
- regression coefficients change greatly when predictors are included/excluded from the model
- significant F-test but no significant t-tests for β's (ignoring β_0)
- type I and II SS are very different
- predictors have pairwise correlations

Formal Diagnostics

- Variance Inflation Factor (VIF)

$$VIF_k = 1/(1 - R_k^2)$$

where R_k^2 is the coefficient of multiple determination when x_k is regressed on the $p - 2$ other x variables in the model. We calculate it for each explanatory variable.

- If $R_k^2 = 0$, $(VIF)_k = 1$, i.e., when x_k is not linearly related to the other x variables. When $R_k^2 \neq 0$, $(VIF)_k$ is greater than 1, indicating an inflated variance for b_k' as a result of the intercorrelations among the x variables. When x_k has a per-

fect linear association with the other x variables in the model, $R_k^2 = 1, (VIF)_k \to \infty$

- If this R_k^2 is large, that means x_k is well predicted by the other x's. One suggested rule is that a value of 10 or more for VIF indicates excessive multicollinearity.

Regression Diagnostics Summary

- The ideas (especially with regard to the residuals) of Chapter 3 still apply, but we will also concern ourselves with the detection of outliers, influential data points and multicollinearity problem.
- Check normality of the residuals with a normal quantile plot. Plot the residuals versus predicted values, versus each of the X's and (when appropriate) versus time. Examine the partial regression plots for each X variable.
- Examine the studentized deleted residuals, The hat matrix diagonals, Dffits, Cook's D, and the DFBETAS. Check observations that are extreme on these measures relative to the other observations. Examine the VIF for each X

Life Insurance Example

- y: amount of insurance (in $1000)
- $x1$: Average Annual Income (in $1000)
- $x2$: Risk Aversion Score $(0 - 10)$
- $n = 18$ managers were surveyed.

Results

- rstudent: compare to $t_{14}(1 - .05/36) = t_{14}(.9986) = 3.621442$, none of them is y-outlier
- Leverage: comparing to $2 * 3/18 = .333$ or $4/18 = .22$, $No.6 = .35$, $No.7 = .62$ and $No.12 = .299$ are x outliers

- Cook's D: Compare to $F_{(3,15)}(.5) = .8256$, observation 7 with cook's d 2.889 has a lot of influence.
- Dffits: $No.7 = 3.5292$. Comparing to 1, it is an influential data point
- Dbeta's: $No.7$ with x_1 2.6598 and x_2 -2.8751, comparing to 1, it is an influential point

  According to all these measures, observation #7 appears to be influential. It has the smallest risk (1) and the highest income (79.380) among all the observations.