Chapter 11
Multiple Regression—Regression Diagnostics

In Chapter 9 we show how to set up and produce an initial analysis of a regression model with several predictors. In this chapter we discuss ways to investigate whether the model assumptions are met and, when the assumptions are not met, ways to revise the model to better conform with the assumptions. We also examine ways to assess the effect on model performance of individual predictors or individual cases (observations).

11.1 Example—Rent Data

11.1.1 Study Objectives

Alfalfa is a high-protein crop that is suitable as food for dairy cows. There are two research questions to ask the data in file `data(rent)` (from file (`alr162`) in Weisberg (1985)). It is thought that rent for land planted to alfalfa relative to rent for other agricultural purposes would be higher in areas with a high density of dairy cows and rents would be lower in counties where liming is required, since that would mean additional expense.

11.1.2 Data Description

The data displayed in the scatterplot matrices (`sploms`) in Figure 11.1 were collected to study the variation in rent paid in 1977 for agricultural land planted to alfalfa. The unit of analysis is a county in Minnesota; the 67 counties with appreciable rented

R.M. Heiberger, B. Holland, *Statistical Analysis and Data Display*,
Springer Texts in Statistics, DOI 10.1007/978-1-4939-2122-5_11

farmland are included. Note that we automatically conditioned the splom on the factor `lime`. The original data include:

`rnt.alf:` average rent per acre planted to alfalfa

`rnt.till:` average rent paid for all tillable land

`cow.dens:` density of dairy cows (number per square mile)

`prop.past:` proportion of farmland used as pasture

`lime:` "lime" if liming is required to grow alfalfa; "no.lime" otherwise
(Lime is a calcium oxide compound that is spread on a field as a fertilizer.)

We added one more variable

`alf.till:` the ratio of `rnt.alf` to `rnt.till`

to investigate the relative rent question.

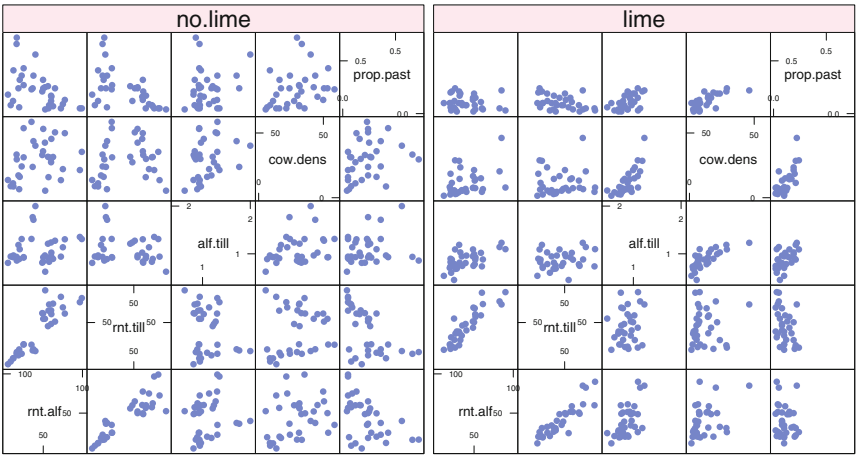

Fig. 11.1 Scatterplot matrices of all variables conditioned on `lime`.

11.1.3 Rent Levels

It is immediately clear from the sploms in Figure 11.1 that `lime` is very important in the distribution of `cow.dens` and `prop.past` as neither has any large values in the `lime` splom. The ratio `alf.till` is slightly higher in the `no.lime` splom.

`lime` does not seem to have an effect on either of the rent variables `rent.alf` or `rent.till`, as their panels have similar distributions in both sploms. The regression analysis of `rent.alf` in Table 11.1 supports that impression as `lime` has a very low t-value. `prop.past` also has a very low t-value.

Table 11.1 `rent.alf` regressed against all other observed variables.

```
> rent.lm3l <-
+     lm(rnt.alf ~ rnt.till + cow.dens + prop.past + lime,
+        data=rent)

> summary(rent.lm3l)

Call:
lm(formula = rnt.alf ~ rnt.till + cow.dens + prop.past + lime,
    data = rent)

Residuals:
    Min      1Q  Median      3Q     Max
-21.229  -4.869  -0.029   4.755  27.767

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -3.334      4.101   -0.81  0.41931
rnt.till       0.883      0.069   12.80  < 2e-16 ***
cow.dens       0.432      0.108    4.00  0.00017 ***
prop.past    -11.380     11.894   -0.96  0.34236
lime1         -0.506      1.425   -0.36  0.72371
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.31 on 62 degrees of freedom
Multiple R-squared:  0.84,Adjusted R-squared:  0.83
F-statistic: 81.6 on 4 and 62 DF,  p-value: <2e-16


> anova(rent.lm3l)
Analysis of Variance Table

Response: rnt.alf
          Df Sum Sq Mean Sq F value Pr(>F)
rnt.till   1  25824   25824  297.89 <2e-16 ***
cow.dens   1   2386    2386   27.53  2e-06 ***
prop.past  1     74      74    0.85   0.36
lime       1     11      11    0.13   0.72
Residuals 62   5375      87
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

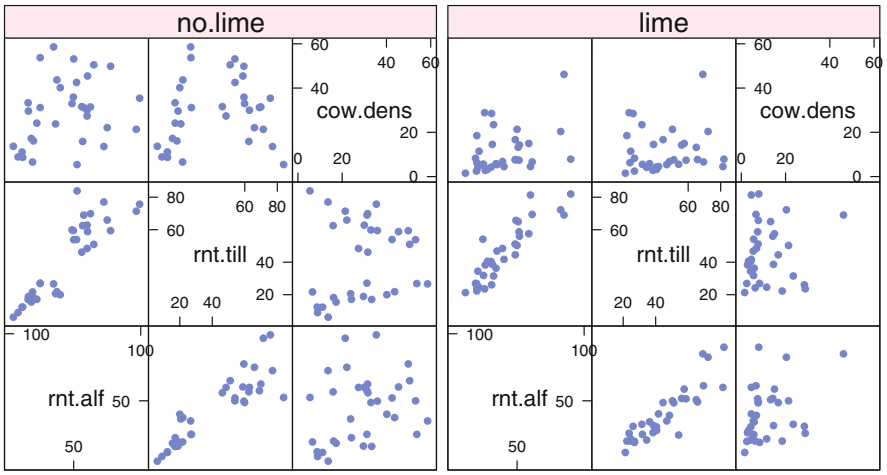

Fig. 11.2 Scatterplot matrices of `rnt.alf` with 2 X-variables, conditioned on `lime`.

We therefore look at a simpler model, without the `prop.past` predictor but with the `cow.dens:lime` interaction, in Figure 11.2 and Table 11.2. Although the regression analysis shows the `lime` coefficient as not significant, it shows the interaction of lime with cow density to be on the edge of significance ($p = .055$). We left both in the model because there appears to be much higher variability in the residuals for high values of `rnt.till` and lower variability in the residuals for low values of `cow.dens` in the `no.lime` counties as indicated in Figure 11.3.

Our conclusion from this portion of the analysis is that rent for alfalfa is related to rent for tillage and to cow density. The relationship with cow density may depend on the need for lime. We need to investigate the variability of the residuals.

Table 11.2 `rent.alf` regressed against all variables except `prop.past`, and including the interaction of cow density with lime.

```
> rent.lm4ln <- lm(rnt.alf ~ rnt.till + cow.dens +
+                  lime + cow.dens:lime, data=rent)

> summary(rent.lm4ln)

Call:
lm(formula = rnt.alf ~ rnt.till + cow.dens + lime + cow.dens:lime,
    data = rent)

Residuals:
    Min      1Q  Median      3Q     Max
-24.346  -4.251  -0.194   4.151  27.193

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -5.9584     3.0117   -1.98    0.052 .
rnt.till         0.9269     0.0536   17.28  < 2e-16 ***
cow.dens         0.4567     0.0991    4.61  2.1e-05 ***
lime1           -3.6034     2.1642   -1.66    0.101
cow.dens:lime1   0.1926     0.0986    1.95    0.055 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.1 on 62 degrees of freedom
Multiple R-squared:  0.847,Adjusted R-squared:  0.838
F-statistic: 86.1 on 4 and 62 DF,  p-value: <2e-16


> anova(rent.lm4ln)
Analysis of Variance Table

Response: rnt.alf
              Df Sum Sq Mean Sq F value  Pr(>F)
rnt.till       1  25824   25824  311.61 < 2e-16 ***
cow.dens       1   2386    2386   28.80 1.3e-06 ***
lime           1      5       5    0.07   0.799
cow.dens:lime  1    316     316    3.81   0.055 .
Residuals     62   5138      83
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

11.1.4 Alfalfa Rent Relative to Other Rent

Returning to the sploms in Figure 11.1, we see that that `lime` puts an upper bound on the `alf.till` ratio. The ratio does seem to go up with cow density and seems to have a variance relation with proportion in pasture. In Table 11.3, a regression of the `alf.till` ratio against the non-rent variables, we see that we can drop the `prop.past` variable.

We continue with Table 11.4 and Figure 11.4, which show an ordinary analysis of covariance with model

```
alf.till ~ cow.dens * lime
```
(11.1)

The ANOVA table in Table 11.4 shows the interaction is not quite significant.

We choose to investigate individual points by looking at plots of the residuals in Figure 11.5 (with the QQ-plot expanded in Figure 11.9) and the regression diagnostics in Figure 11.6. These show the three points (19, 33, 60) in the `no.lime` group and the single point (49) in the `lime` group as being potentially influential. Figure 11.6, produced with our functions `lm.case.s` and `plot.case.s`, includes boundaries for the standard recommended thresholds for the various diagnostic measures discussed in Section 11.3.

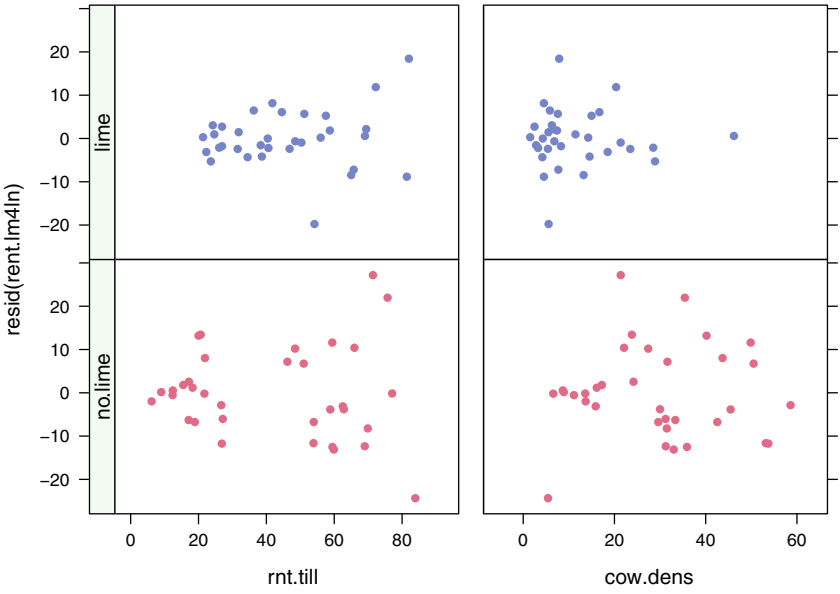

Fig. 11.3 Residuals from `rnt.alf ~ rnt.till + cow.dens*lime` (in Table 11.2 and Figure 11.2) plotted against the *X*-variables conditioned on `lime`.

Table 11.3 `alf.till` ratio regressed against cow density | `lime` and proportion in pasture.

```
> rent.lm12p <- lm(alf.till ~ lime * cow.dens + prop.past, data=rent)

> summary(rent.lm12p, corr=FALSE)

Call:
lm(formula = alf.till ~ lime * cow.dens + prop.past, data = rent)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3342 -0.1247 -0.0203  0.1045  0.7853

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    0.78957    0.05637   14.01  < 2e-16 ***
lime1         -0.09686    0.05333   -1.82  0.07419 .
cow.dens       0.00944    0.00259    3.64  0.00056 ***
prop.past      0.18989    0.22670    0.84  0.40546
lime1:cow.dens 0.00391    0.00242    1.62  0.11063
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.223 on 62 degrees of freedom
Multiple R-squared:  0.366,Adjusted R-squared:  0.325
F-statistic: 8.94 on 4 and 62 DF,  p-value: 9.17e-06


> anova(rent.lm12p)
Analysis of Variance Table

Response: alf.till
             Df Sum Sq Mean Sq F value  Pr(>F)
lime          1  0.846   0.846   17.03 0.00011 ***
cow.dens      1  0.754   0.754   15.19 0.00024 ***
prop.past     1  0.045   0.045    0.91 0.34503
lime:cow.dens 1  0.130   0.130    2.62 0.11063
Residuals    62  3.078   0.050
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We locate the potentially influential points in Figure 11.7 and see them as the three counties with the highest ratios and the one `lime` county with an unusually high cow density. In Section 11.3 we will discuss the statistics displayed in Figures 11.5 and 11.6 as well as their interpretation.

We redo the analysis without these four points in Table 11.6 and Figure 11.8. After isolating these four counties we see significantly different slopes in the `no.lime` and `lime` counties.

Table 11.4 `alf.till` ratio regressed against cow density | `lime`. See Figure 11.4.

```
> rent.lm12m <- aov(alf.till ~ lime * cow.dens, data=rent)

> anova(rent.lm12m)
Analysis of Variance Table

Response: alf.till
              Df Sum Sq Mean Sq F value  Pr(>F)
lime           1  0.846   0.846   17.11 0.00011 ***
cow.dens       1  0.754   0.754   15.26 0.00023 ***
lime:cow.dens  1  0.140   0.140    2.84 0.09708 .
Residuals     63  3.113   0.049
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> summary.lm(rent.lm12m)

Call:
aov(formula = alf.till ~ lime * cow.dens, data = rent)

Residuals:
    Min      1Q  Median      3Q     Max
-0.3296 -0.1362 -0.0139  0.0877  0.8408

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.80653    0.05248   15.37  < 2e-16 ***
lime1          -0.10424    0.05248   -1.99    0.051 .
cow.dens        0.01024    0.00241    4.25  7.1e-05 ***
lime1:cow.dens  0.00405    0.00241    1.68    0.097 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.222 on 63 degrees of freedom
Multiple R-squared:  0.359,Adjusted R-squared:  0.328
F-statistic: 11.7 on 3 and 63 DF,  p-value: 3.32e-06
```

Our conclusion at this step is that for most counties, there is a linear relationship of the rent ratio to the cow density, with the slope depending on the need for lime. The three `no.lime` counties and the one `lime` county need additional investigation.

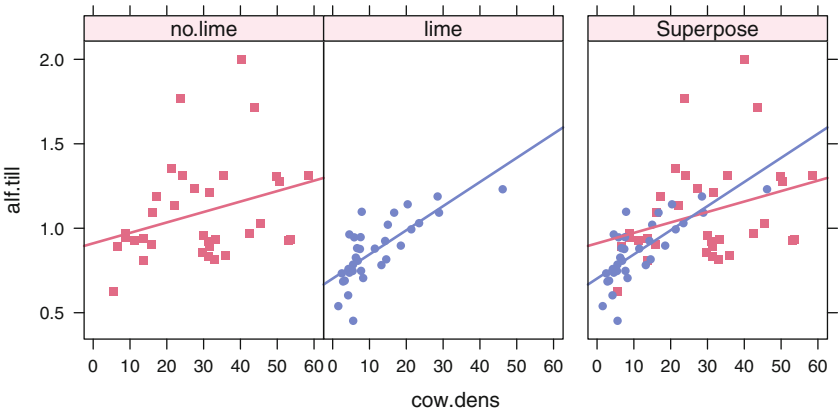

Fig. 11.4 ANCOVA `rnt.alf/rnt.till ~ cow.dens | lime`. See Table 11.4.

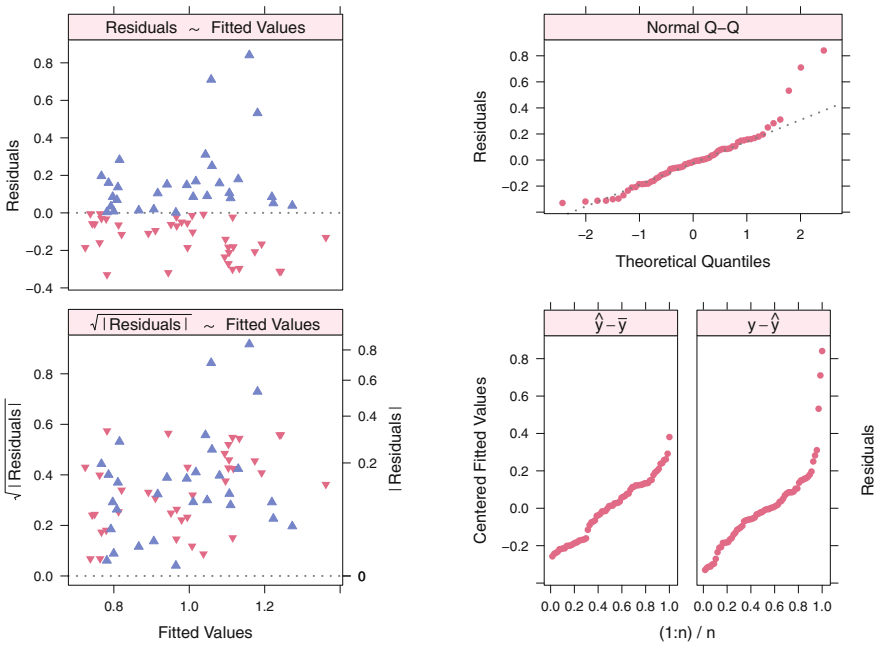

Fig. 11.5 Residuals from ANCOVA `(rnt.alf/rnt.till) ~ cow.dens | lime`. See Table 11.4 and Figure 11.4. The structure of the panels in this figure is discussed in Section 8.4. The figure itself is similar to Figure 8.6.

Table 11.5 Case diagnostics for model in Table 11.4. The diagnostics are plotted in Figure 11.6. The case numbers for the noteworthy cases are listed here.

```
> rent.case12m <- case(rent.lm12m)

> rent.case12m.trellis <-
+    plot(rent.case12m, rent.lm12m, par.strip.text=list(cex=1.2),
+         layout=c(3,3), main.cex=1.6, col=likertColor(2)[2], lwd=4)

> rent.case12m.trellis ## display both graph and list of noteworthy cases
                       Noteworthy Observations
Student del resid      19 33
deleted std dev        19 33
h                      13 40 49 56
Cook's distance
dffits                 5 19 32 33 49 60 66
DFBETAS (Intercept)    5 19
DFBETAS lime1          5 19
DFBETAS cow.dens       5 32 33 49 60
DFBETAS lime1:cow.dens 5 32 33 49 60
```

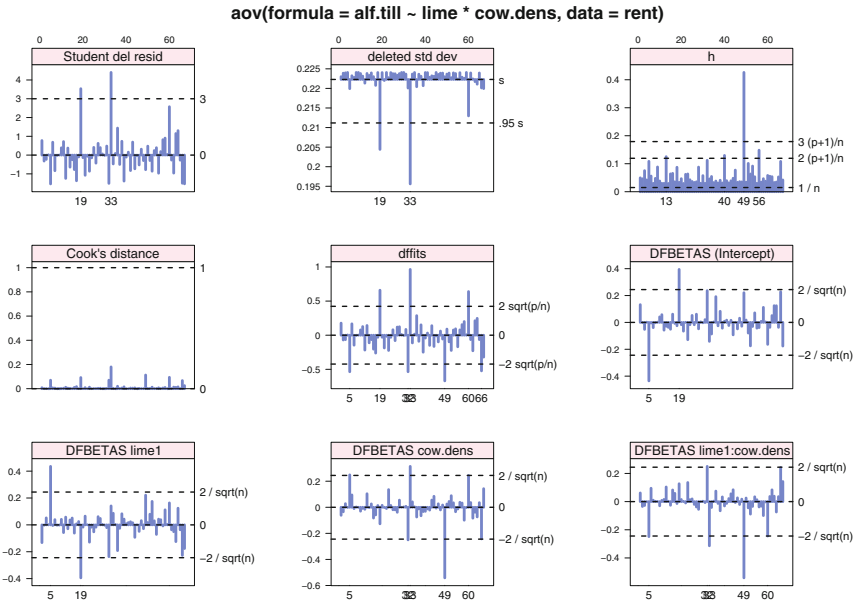

Fig. 11.6 Diagnostics from ANCOVA (`rnt.alf/rnt.till) ~ cow.dens | lime`. The model is displayed in Table 11.4 and Figure 11.4. Each of the statistics in these panels is discussed in Section 11.3 and shown enlarged in Figures 11.12–11.17. To work around the problem that identification in the graph's *x*-axis of noteworthy cases often suffers from overprinting, the `plot.case` function returns and prints a list of noteworthy cases. We show the list in Table 11.5.

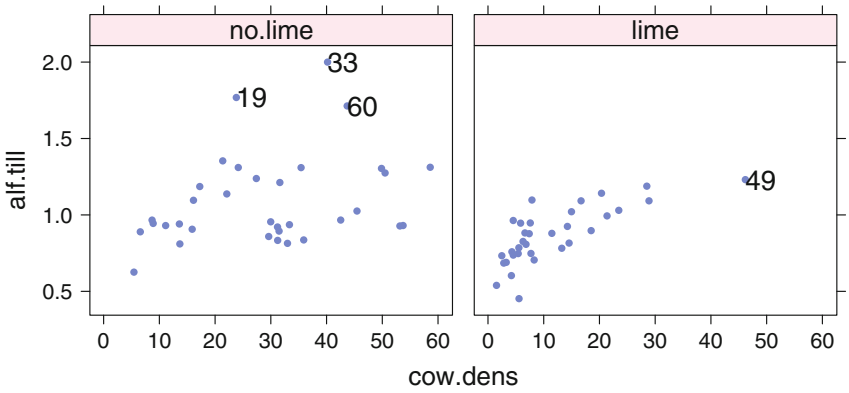

Fig. 11.7 Identified points in ANCOVA `(rnt.alf/rnt.till) ~ cow.dens | lime`.

Table 11.6 ANCOVA of `alf.till` ratio regressed against cow density and lime with four removed observations. See Figure 11.8. Compare to Table 11.4.

```
> rent.lm12ms.aov <- aov(alf.till ~ lime * cow.dens,
+                        data=rent[-c(19, 33, 60, 49),])

> anova(rent.lm12ms.aov)
Analysis of Variance Table

Response: alf.till
              Df Sum Sq Mean Sq F value  Pr(>F)
lime           1  0.428   0.428   17.81 8.5e-05 ***
cow.dens       1  0.395   0.395   16.43 0.00015 ***
lime:cow.dens  1  0.233   0.233    9.67 0.00288 **
Residuals     59  1.419   0.024
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

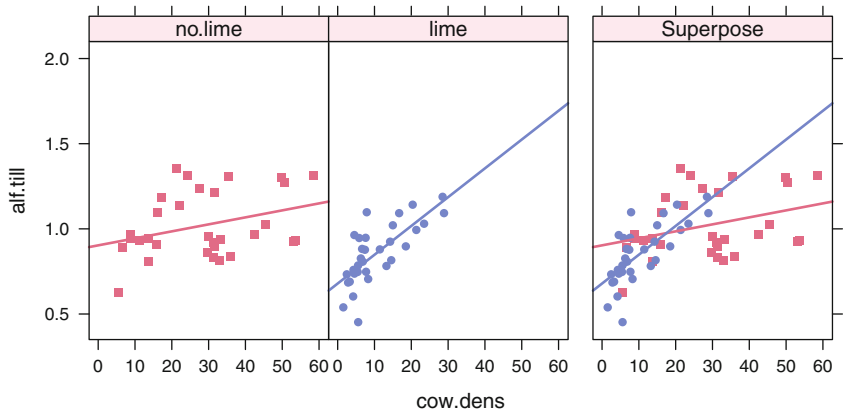

Fig. 11.8 Repeat of Figure 11.4 with four counties removed from ANCOVA `ancovaplot(alf.till ~ lime * cow.dens, data=rent[-c(19, 33, 60, 49),])`. See Table 11.6. Compare to Figure 11.4.

11.2 Checks on Model Assumptions

We assume in Section 9.3 that the model error terms $\epsilon_i \sim \text{NID}(0, \sigma^2)$ (Normal Independently Distributed), that is that they have the same variance σ^2 for all cases, are mutually uncorrelated or independent, and are normally distributed. In order for the conclusions from our analyses to be valid, these assumptions must be true. Therefore, we discuss ways to verify the assumptions and then suggest some remedies when assumptions are not met.

11.2.1 Scatterplot Matrix

We previously mentioned the importance of routinely producing scatterplot matrices as part of analyses involving several variables. We produced many such plots in our discussion in Section 11.1. Here we focus on the rows of the scatterplot matrix that correspond to the response variables. The panels in these rows, the plots of the response y vs each of the explanatory variables x_j, should each be approximately linear. In Section 11.1.3 the response is shown in the `rnt.alf` row in Figure 11.1 and in Figure 11.2. In Section 11.1.4 the response is the `alf.till` row in Figure 11.1 and in Figure 11.4. If the plot of y against any explanatory variable suggests curvature in the relationship, the analyst should consider transforming either the response variable or that explanatory variable so that following transformation the plot of y vs the transformed x_j is close to linear. A successful transformation suggests the use of this transformed predictor rather than the original in the regression model. Exercise 11.5 explores this idea.

11.2.2 Residual Plots

Before a model can be accepted for use in explanation or prediction, the analyst should produce and examine plots involving the residuals calculated from the fit of the model to the data. The residuals e_i should be plotted vs each of the following, one plot point per case:

- the fitted values of the response $\hat{y}_i$
- each of the model's explanatory variables x_j
- possibly other variables under consideration for the model but not yet a part of it
- time, if the data are time-ordered

In addition, the partial residuals (see Section 9.13.1) should be plotted against the corresponding predictors and against the residuals from regressing each predictor against the other predictors (added variable plots; see Section 9.13.4). Ideally, each

of these plots should exhibit no systematic character and have random scatter about the horizontal line at 0, the mean of the e_i.

In order to check for normality, the analyst should produce a normal probability plot of the residuals. If there is doubt that this plot confirms normality, the analyst can request the p-value from an all-purpose test of normality having good power against a variety of alternatives, such as the Shapiro–Wilk test mentioned in Section 5.7.

If a residual plot suggests that an assumption is not met, the analyst must seek a remedy following which the assumption is met.

We show in Figure 11.9 the normal probability plot for the rent ratio `alf.till` analysis in Table 11.4 and Figure 11.4. It does not look normal. Compare this plot to Figure 11.10, which shows probability plots of six normal and six non-normal variables.

From the `cow.dens` column, we again see similar behavior in Rows 1 and 3. We also note the higher variability in Y for the higher densities. We get a sense of why we see that difference in variability from the interaction `lime:cow.dens` column. Here we see, most clearly in the partial residuals plot in Row 3, that the high variability is observed when the interaction variable is negative, corresponding to the `no.lime` counties.

Figure 11.11 shows several plots of the residuals and partial residuals from the model in Table 11.4 and Figure 11.4. From the `lime` column, we see that the ratio `alf.till` is higher for `lime`=−1 (no lime) than for `lime`=1 (lime). The pattern is similar in the observed variable plots in Row 1 and the partial residuals plots in Row 3, suggesting that the `lime` effect is independent of the other variables.

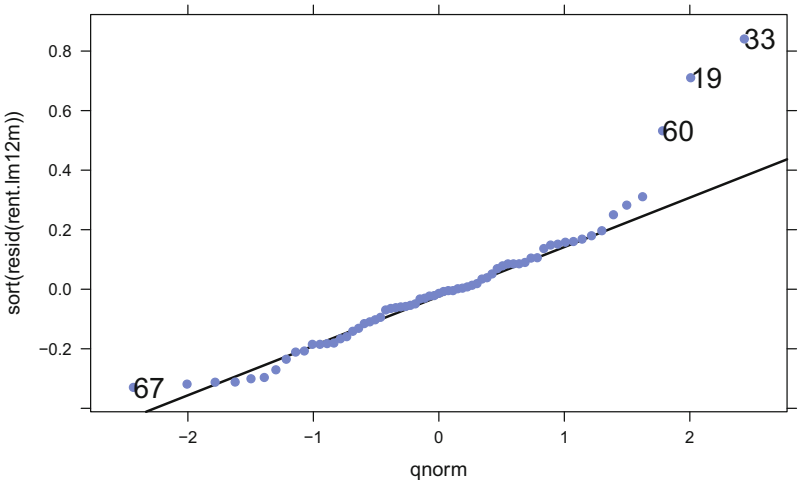

Fig. 11.9 Normal plot of residuals from ANCOVA `rnt.alf/rnt.till ~ cow.dens | lime`. See Table 11.4 and Figure 11.4. The results do not look normal. We ran the Shapiro-Wilk normality test with statistic W=0.8969 and $p = 4\,10^{-5}$. We identified the four most extreme points. Three of them are the three `no.lime` counties that we had previously identified.

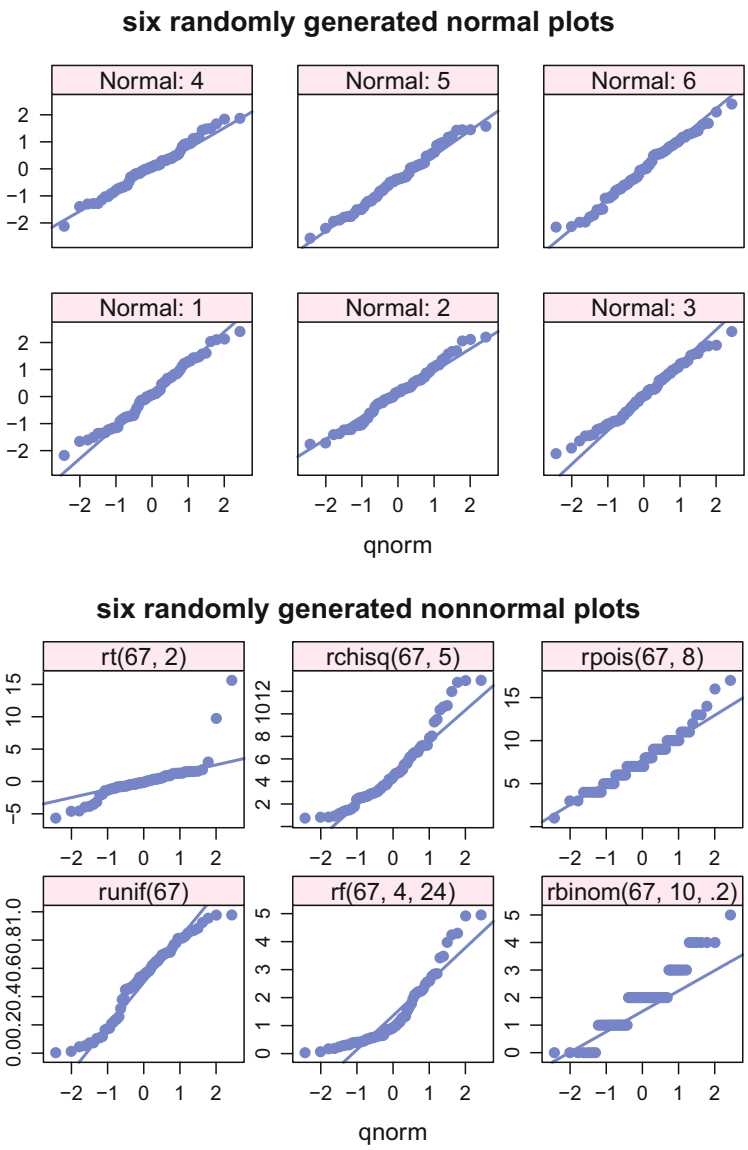

Fig. 11.10 Normal plot of six randomly generated normal variables and six randomly generated nonnormal variables. These plots are placed here to help you calibrate your eye to what normal and nonnormal distributions look like when plotted against the normal quantiles. *t*: long left and right tails as indicated by points below the diagonal on the left and above the diagonal on the right. Chi-square: short left and long right tails. Poisson: discrete appearance and long tail on the right. Uniform: short tails on left and right as indicated by points above the diagonal line on the left and below the diagonal line on the right. *F*: short tail on the left and long tail on the right. Binomial: discrete positions on the *y*-axis, with short tail on the left; this example with $p = .2$ is not symmetric and we see more points on the left.

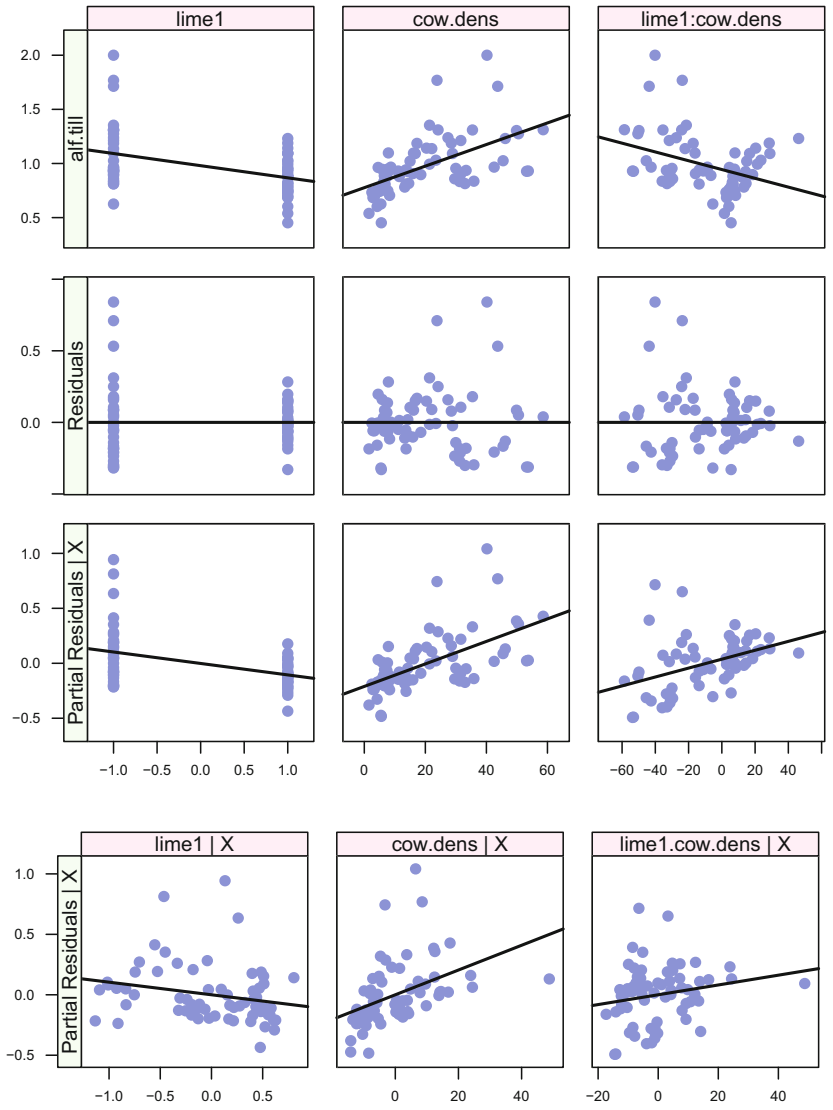

Fig. 11.11 Row 1 (at the top) shows the response variable `alf.till` against each of the three predictors. Row 2 shows the ordinary residuals $e = Y - \hat{Y}$ from the regression on all three variables against each of the three predictors. Row 3 shows the "partial residuals plots", the partial residuals for each predictor against that predictor. Row 4 shows the "added variable plots", the partial residuals against the residuals of X_j regressed on the other two predictors. The slope for both rows 3 and 4, the partial residuals and the added variables, is exactly the regression coefficient for that term.

Table 11.7 Regression Diagnostics Formulas

Name	Notation and definition	Sequenced calculation formulas	Description
Observed response variable	$\underset{n\times 1}{Y}$		
Observed predictor variables	$\underset{n\times(1+p)}{X}$	$X = [\mathbf{1}\, X_1 X_2 \ldots X_p]$	
Fitted value	$\hat{Y} = \left(\hat{Y}_i\right)$	$X\hat{\beta}$	
Residual	e_i	$Y_i - \hat{Y}_i$	
Standard deviation	$s = \sqrt{\mathsf{MSE}} = \sqrt{\text{var}(Y_i \vert X)}$	$\sqrt{\sum e_i^2/(n-p-1)}$	All n observations
Leverage	$h_i = h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i}$	$\text{diag}\left(X(X'X)^{-1}X'\right)$	
Variance of e_i	$\text{var}(e_i)$	$s^2(1-h_i)$	$Y_i = \hat{Y}_i + e_i$
Variance of $\hat{Y}_i$	$\text{var}(\hat{Y}_i)$	$s^2 h_i$	$\text{var}(Y_i) = \text{var}(\hat{Y}_i) + \text{var}(e_i)$
Standardized residual	e_i^*	$e_i/(s\sqrt{1-h_i})$	$e_i/(\sigma\sqrt{1-h_i}) \sim N(0,1)$ when H_0 is true
Data with i^{th} row deleted	$\underset{(n-1)\times(1+p)}{X_{(i)}}$	$X_{[(1,2,\ldots,i-1,i+1,\ldots,n),(0,1,\ldots p)]}$	
Deleted regression coefficients	$\hat{\beta}_{(i)} = (X'_{(i)}X_{(i)})^{-1}X'_{(i)}Y_{(i)}$	See description in Section 11.3.6.	Estimation of β based on $n-1$ observations, all except i. The definition isn't efficient. Use the algorithm in Section 11.3.6.
Deleted standard deviation	$s_{(i)} = \sqrt{\mathsf{MSE}_{(i)}}$	$\sqrt{\left((n-p)s^2 - e_i^2/(1-h_i)\right)/(n-p-1)}$	$n-1$ observations, all except i.
Deleted predicted value	$\hat{Y}_{(i)}$	$X_{[i,(0,\ldots p)]}\hat{\beta}_{(i)}$	Prediction Y_i based on the remaining $n-1$ observations

Studentized deleted residual	t_i	$e_i / \left(s_{(i)} \sqrt{1-h_i}\right)$	$t_i \sim t_{n-p-2}$ when H_0 is true
Cook's distance	$D_i = \left(\hat{Y} - \hat{Y}_{(i)}\right)' \left(\hat{Y} - \hat{Y}_{(i)}\right) / (p\, s^2)$ $= \left(\hat{\beta} - \hat{\beta}_{(i)}\right)' X'X \left(\hat{\beta} - \hat{\beta}_{(i)}\right) / (p\, s^2)$	$\frac{e_i^2}{p\,\mathsf{MSE}} \left(\frac{h_i}{(1-h_i)^2}\right)$	
Inverse of crossproduct of X	$C = (c_{ij})$	$(X'X)^{-1}$	
Covariance of coefficients	$s^2 C$	$s^2 (X'X)^{-1}$	Estimated covariance matrix of regression coefficients
DFBETAS	$\mathsf{DFBETAS}_{ik} = (\hat{\beta}_k - \hat{\beta}_{k(i)}) / (s_{(i)} \sqrt{c_{kk}})$		Standardized $\Delta\hat{\beta}_k$ when observation i is deleted
DFFITS	$\mathsf{DFFITS}_i = (\hat{Y}_i - \hat{Y}_{i(i)}) / (s_{(i)} \sqrt{h_i})$	$\left(\frac{n-p-1}{\mathsf{SSE}\,(1-h_i) - e_i^2}\right)^{\frac{1}{2}} \left(\frac{h_i}{1-h_i}\right)^{\frac{1}{2}}$	Standardized $\Delta\hat{Y}_i$ when observation i is deleted

11.3 Case Statistics

Many of the diagnostics discussed in this chapter fall under the heading *case statistics*, i.e., they have a value for each of the n cases in the data set. If a case statistic has a value that is unusual, based on thresholds we discuss, the analyst should *scrutinize* the case. One action the analyst might take is to delete the case. This is justified if the analyst determines the case is not a member of the same population as the other cases in the data set. But deletion is just one possibility. Another is to determine that the flagged case is unusual in ways apart from those available in its information in the present data set, and this may suggest a need to add one or more additional predictors to the model.

There are many case statistics used in regression diagnostics. The concepts are complex and the notation more so. We summarize the notation in Table 11.7. We discuss each of the formulas and illustrate them with the diagnostic plots for the rent data that we originally showed in Figure 11.6. We reproduce each of the panels in that figure as a standalone plot here as part of the discussion.

We focus on five distinct case statistics, each having a different function and interpretation. (One of these, DFBETAS, is a vector with a distinct value for each regression coefficient including the intercept coefficient.) For small data sets the analyst may choose to display each of these case statistics for all cases. For larger data sets we suggest that the analyst display only those values of the case statistics that exceed a threshold, or flag, indicating that the case is unusual in some way. Recommended thresholds are mentioned in the following sections.

Leverage measures how unusual a case is with respect to the values of its predictors, i.e., whether the values of a case's predictors are an outlying point in the p-dimensional space of predictors. Unlike the other case statistics, leverage does not involve the response variable.

Studentized deleted residuals suggest how unusual cases are with respect to the case's value of the response variable.

Cook's distance is a combined measure of the unusualness of a case's predictors and response. It sometimes happens that a case is flagged by Cook's distance but not quite flagged by leverage or Studentized deleted residuals.

DFFITS indicates the extent to which deletion of the case impacts predictions made by the model.

DFBETAS (one for each regression coefficient) show the extent to which deletion of a case would perturb that regression coefficient.

In the following sections we discuss these statistics in turn, presenting two formulas for each of them. The first, the definitional formula, is intended to be intuitive. It is used to explain to the reader what the formula measures and why it is helpful to view it in an analysis. It is also inefficient and should not be used as a computational

formula. The second formula, the computational formula, is an order of magnitude more efficient for computation. It is not intuitive. We leave for Exercise 11.8 the proofs that the two sets of formulas are equivalent.

11.3.1 Leverage

The calculation of leverages is briefly addressed in 9.3.1. Leverages measure how unusual a case is with respect to its set of predictors. Unlike other measures in this chapter, leverages do not involve the response variable. The leverage h_{ii} of case i, usually abbreviated to h_i, is the i^{th} diagonal entry of the *hat matrix* $H = X(X'X)^{-1}X'$. This matrix has come to be called the hat matrix because in matrix notation the predicted response is $\hat{Y} = X(X'X)^{-1}X'Y = HY$, i.e., H transforms Y to $\hat{Y}$ by placing a "hat" on the Y. It can be shown (see Exercise 11.9) that all leverages satisfy $\frac{1}{n} \leq h_i \leq 1$. If a model contains p predictors, an excessively large leverage is one for which

$$h_i > \frac{2(p+1)}{n} \quad \text{or} \quad h_i > \frac{3(p+1)}{n} \tag{11.2}$$

These suggested rules derive from the fact that the average of all n leverages is $\frac{p+1}{n}$, so they are based on exceeding 2 or 3 times this average. A case that is flagged because its leverage exceeds one or both of these thresholds has a value for at least one predictor that is unusual compared to values of such predictors for other cases. We can show that

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} \quad \text{and} \quad h_{ij} = \frac{\partial \hat{y}_i}{\partial y_j}$$

The leverage h_i of case i is geometrically interpreted as the generalized (Mahalanobis) distance of $X_{i\cdot}$ (the i^{th} row of X) from the $(p+1)$-dimensional centroid of all n rows of X.

More complicated forms of leverage have been devised to diagnose a group of cases that when considered together are unusual but when considered individually are not unusual.

Figure 11.12 displays the leverages for each case of the fit of the rent data using Model (11.1). This figure includes horizontal dotted lines demarking the two leverage thresholds given above. We observe that county 49 exceeds both thresholds, telling us that this county (requiring `lime`) has an unusually large `cow.dens`.

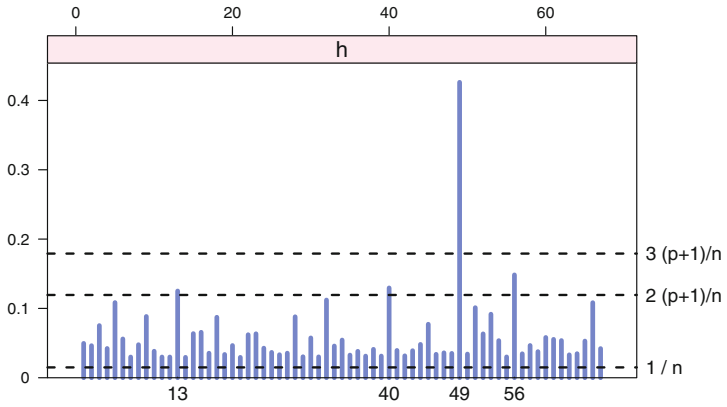

Fig. 11.12 Leverage for Model (11.1) for rent data.

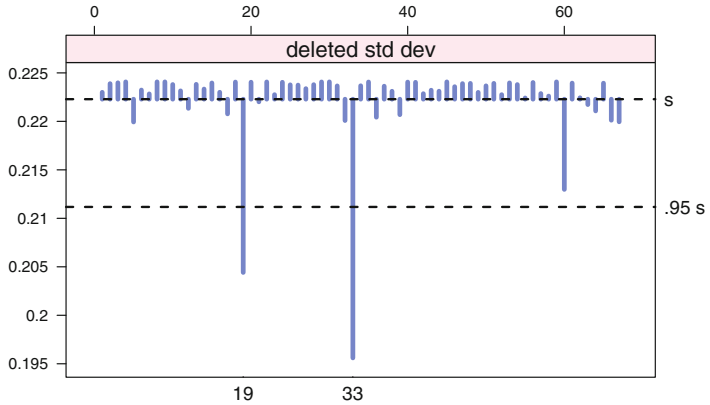

Fig. 11.13 Deleted standard deviations for Model (11.1) for rent data.

11.3.2 Deleted Standard Deviation

The deleted standard deviation $s_{(i)}$ is defined to be the value of s calculated from the same regression model using all cases *except* case i. Because the primary use of the $s_{(i)}$ is in the definition of the Studentized deleted residuals, there are no standard rules for interpreting these values themselves.

We compare the $s_{(i)}$ values to two thresholds, $.95s$ and $1.05s$. If deletion of an observation shifts the estimated standard deviation by 5% in either direction, we note it on the graph and choose to investigate the observation.

Figure 11.13 shows the deleted standard deviations for the rent data. We see two observations, 19 and 33, that are below our lower threshold.

11.3.3 Standardized and Studentized Deleted Residuals

The standardized and Studentized residuals help to assess the effect of each individual case on the calculated regression relationship. For case i the standardized residual

$$e_i^* = e_i / \sqrt{\widehat{\text{var}}(e_i)} \tag{11.3}$$

is the calculated residual, e_i, standardized by dividing by its estimated standard error

$$\sqrt{\widehat{\text{var}}(e_i)} = s\sqrt{1 - h_i} \tag{11.4}$$

Note that because this standard error depends on i, it differs slightly from case to case. The standardized residual is also called the *internally standardized residual* because the calculation of s includes case i.

The *Studentized deleted residual,* also called the *externally standardized residual,* for case i is calculated from the regular residuals, the deleted standard deviations, and the hat diagonals.

$$t_i = \frac{e_i}{s_{(i)}\sqrt{1 - h_i}} \tag{11.5}$$

As implied by this notation, t_i has a Student's t distribution with $n - p - 1$ degrees of freedom. Considering the t distribution with moderate degrees of freedom, we say that case i's response value is "unusual" (the actual response differs "appreciably" from the predicted response) if its absolute Studentized deleted residual exceeds 2 or 3. Such a case may be termed an *outlier*. We recommend a threshold of 2 for small data sets and 3 for large data. The reason for this recommendation is that for a large data set, 2 is the approximate 97.5$^{\text{th}}$ percentile of the t distribution so that when the model assumptions are satisfied for all cases, approximately 5% of these residuals will exceed 2 by chance alone.

We prefer the use of Studentized deleted residuals rather than standardized residuals because the former are interpretable as t statistics but the latter are not. A reason is that the numerator and denominator of t_i are statistically independent, but the numerator and denominator of the standardized residuals e_i^* are not independent.

It can be shown (see Exercise 11.8c) that the Studentized deleted residual defined intuitively in Equation (11.5) can be calculated more efficiently by the computational formula

$$t_i = e_i \left(\frac{n - p - 1}{\text{SSE}\,(1 - h_i) - e_i^2} \right)^{\frac{1}{2}} \tag{11.6}$$

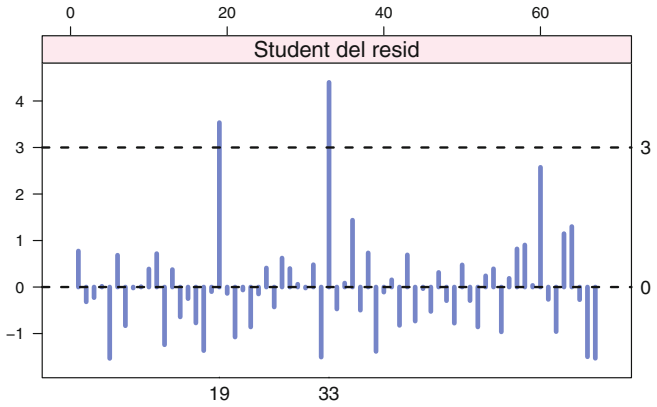

Fig. 11.14 Studentized deleted residuals for Model (11.1) for rent data.

where SSE is the error sum of squares under the full model having n cases. All terms in this expression are available from a single fitting with the n cases. Therefore, in calculating the n t_i's it is not necessary to refit the model n times corresponding to deleting each case in turn.

For our modeling of the rent data in Table 11.4, Figure 11.14 displays the Studentized (deleted) residuals for each case. We see that counties 19 and 33 both exceed the threshold 3, indicating that these counties have unusually large values of `alf.till`.

11.3.4 Cook's Distance

While leverage addresses the unusualness of a case's predictor variables, and Studentized deleted residuals address (primarily) the unusualness of a case's response variable, the Cook's distance D_i of a case assesses the unusualness of both its response and predictors. The Cook's distance D_i for case i can be interpreted in two ways.

Let $\hat{Y}$ be the n-vector of fitted values using all n cases and $\hat{Y}_{(i)}$ be the n-vector of fitted values when case i is not used in fitting. Then

$$D_i = \frac{\left(\hat{Y} - \hat{Y}_{(i)}\right)' \left(\hat{Y} - \hat{Y}_{(i)}\right)}{p\,\mathrm{MSE}} \tag{11.7}$$

This illustrates the interpretation that Cook's distance for case i measures the change in the vector of predicted values when case i is omitted.

Let $\hat{\beta}_{(i)}$ be the vector of estimated regression coefficients estimated without case i. Then

$$D_i = \frac{\left(\hat{\beta} - \hat{\beta}_{(i)}\right)' X'X \left(\hat{\beta} - \hat{\beta}_{(i)}\right)}{p\,\mathsf{MSE}} \tag{11.8}$$

This representation shows that D_i measures the change in the vector of estimated regression coefficients when case i is omitted.

As with the Studentized deleted residual, the n Cook's distances can be calculated without running n regressions omitting each case in turn. It can be shown that

$$D_i = \frac{e_i^2}{p\,\mathsf{MSE}} \left(\frac{h_i}{(1 - h_i)^2} \right) \tag{11.9}$$

From this formula it is apparent that a case with a large Cook's distance has either a large residual, a large leverage, or some combination of these two.

We recommend that a case be regarded as unusual if its Cook's distance exceeds 1. This threshold for what constitutes an unusually large value of Cook's distance D_i follows the recommendation of Weisberg (1985) (page 120).

> Since for most F distributions the 50% point is near 1, a value of $D_i = 1$ will move the estimate to the edge of about a 50% confidence region, a potentially important change. If the largest D_i is substantially less than 1, deletion of a case will not change the estimate of β by much. To investigate the influence of a case more closely, the analyst should delete the large D_i case and recompute the analysis to see exactly what aspects of it have changed.

There are also arguments, for example in Fox (1991), for a much smaller threshold $4/(n - p - 1)$ or $4/n$ that decreases with increasing sample size. We are unconvinced by these arguments.

Figure 11.15 displays the Cook's distances for the rent data. Counties 5, 19, 32, 33, 49, 60, and 66 have much larger Cook's distances than the other counties, but none of these 7 counties approaches the threshold of 1 that would flag a county as unusual. Therefore, Cook's distance flags no data points fitted by `alf.till ~ lime*cow.dens`.

11.3.5 DFFITS

DFFITS, shown in Figure 11.16, is an abbreviation for "difference in fits". DFFITS_i is a standardized measure of the amount by which predicted value $\hat{Y}_i$ for case i changes when the data on this case is deleted from the data set. A flag for a case with large DFFITS is one having absolute value greater than $2\sqrt{p/n}$.

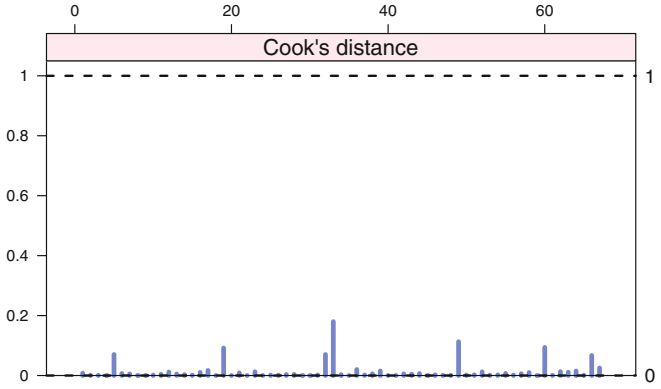

Fig. 11.15 Cook's distances for Model (11.1) for rent data.

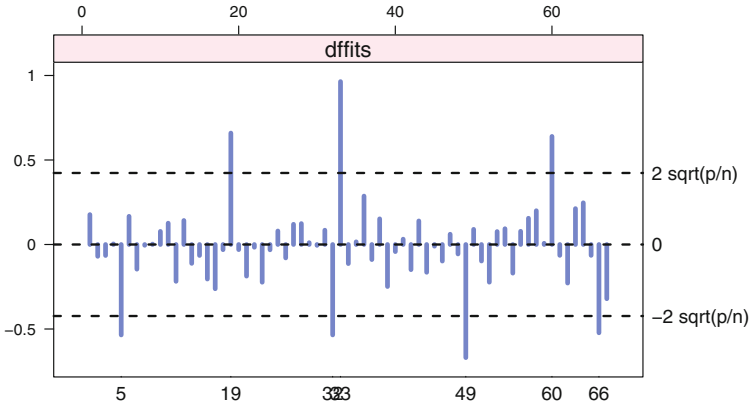

Fig. 11.16 DFFITS for Model (11.1) for rent data.

The interpretation of DFFITS_i is apparent from the formula

$$\mathsf{DFFITS}_i = \frac{\hat{Y}_i - \hat{Y}_{i(i)}}{\sqrt{\mathsf{MSE}_{(i)}\, h_i}} \tag{11.10}$$

where, as before, an (i) in a subscript means that the quantity is calculated with case i omitted from the data. As is seen from

$$\mathsf{DFFITS}_i = \left(\frac{n-p-1}{\mathsf{SSE}\,(1-h_i) - e_i^2}\right)^{\frac{1}{2}} \left(\frac{h_i}{1-h_i}\right)^{\frac{1}{2}} \tag{11.11}$$

DFFITS_i can be calculated from the output of the regression using all n cases.

11.3.6 DFBETAS

DFBETAS$_{ik}$ is a standardized measure of the amount by which the k^{th} regression coefficient changes if the i^{th} observation is omitted from the data set. A case is considered to have a large such measure if its absolute DFBETAS is greater than $2/\sqrt{n}$. Since a regression analysis has np DFBETAS in all, a request for DFBETAS in a large complicated regression analysis will generate a lot of output.

DFBETAS$_{ik}$ is defined by

$$\text{DFBETAS}_{ik} = \frac{\hat{\beta}_k - \hat{\beta}_{k(i)}}{\sqrt{\text{MSE}_{(i)}\, c_{kk}}}$$

for $k = 0, 1, \ldots, p$, where c_{kk} is the k^{th} diagonal entry in $(X'X)^{-1}$. The terms $\hat{\beta}_{k(i)}$ are called the deleted regression coefficients.

An efficient calculation algorithm is

1. Let $\hat{\beta}$ be the regression coefficients from regressing y on x.
2. Let X be the matrix of predictors including the column **1**.
3. Factor $X = QR$. See Section I.4.7 for details.
4. Multiply the i^{th} row of Q by $z_i = e_i/(1 - h_i)$. Call the result Q_z.
5. Solve $R\,\Delta b = Q_z'$ for Δb.
6. Then $\hat{\beta}_{(i)} = \hat{\beta} - \Delta b_i$, where Δb_i is the i^{th} column of Δb.

This algorithm is efficient because it does the hard work of solving a linear system only once, when it factors $X = QR$ to construct the orthogonal matrix Q and the triangular matrix R. The backsolve in step 5 is not hard work because it is working with a triangular system. All the remaining steps are simple linear adjustments to the original solution.

Another efficient algorithm, shown in Table 11.8, is essentially the same although with the steps in a different order. This is the algorithm used by R in function `stats:::dfbetas.lm`.

Figure 11.17 gives one DFBETAS plot for each predictor in the model in Table 11.4. We do not ordinarily interpret DFBETAS for the intercept term. Figure 11.6 shows that cases 5 and 19 impact the regression coefficient of `lime`, cases 33 and 49 impact the regression coefficient of `cow.dens`, and that these four counties plus county 32 are primarily responsible for the difference in slopes of the two regression lines in Figure 11.4.

Table 11.8 R's algorithm for dfbetas. The function `chol2inv` inverts a symmetric, positive definite square matrix from its Choleski decomposition. Equivalently, it computes $(X'X)^{-1}$ from the (R part) of the QR decomposition of X. The value `infl$sigma` is a vector whose i^{th} element contains the estimate of the residual standard deviation obtained when the i^{th} case is dropped from the regression. The value returned by the `stats:::dfbeta` function is the changes in the coefficients which result from dropping each case. Function `stats:::dfbeta` does the scaling.

```
> stats:::dfbetas.lm
function (model, infl = lm.influence(model, do.coef = TRUE),
    ...)
{
    qrm <- qr(model)
    xxi <- chol2inv(qrm$qr, qrm$rank)
    dfbeta(model, infl)/outer(infl$sigma, sqrt(diag(xxi)))
}
<bytecode: 0x10a0fa708>
<environment: namespace:stats>
```

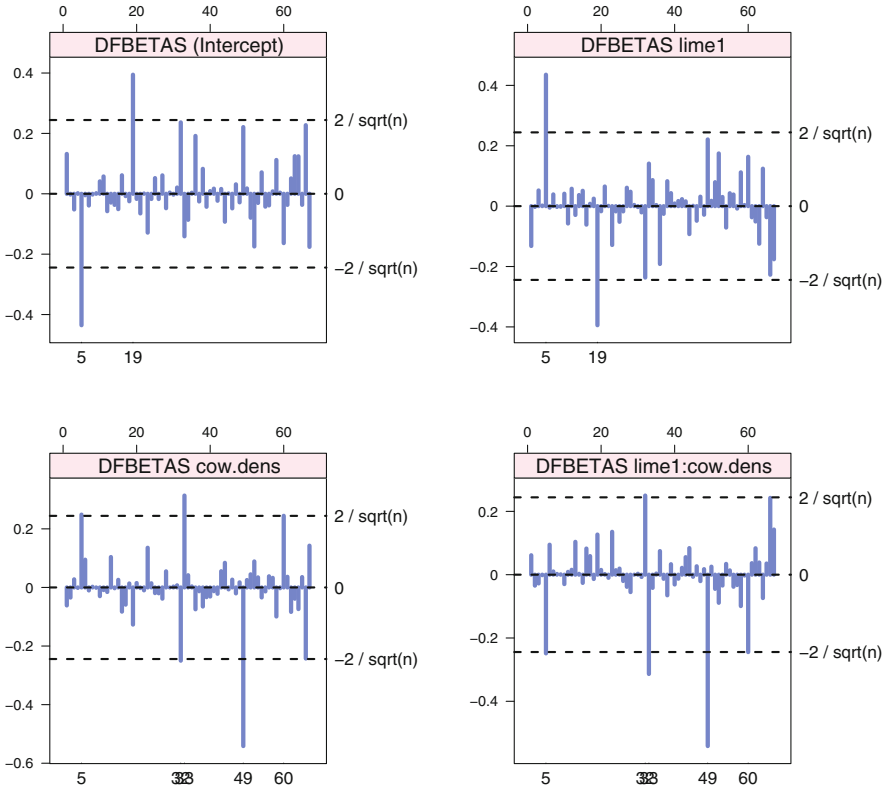

Fig. 11.17 DFBETAS for all four predictors in Model (11.1) for the rent data: the column of 1s for the intercept, the factor `lime`, the covariate `cow.dens`, and the interaction `lime:cow.dens`.

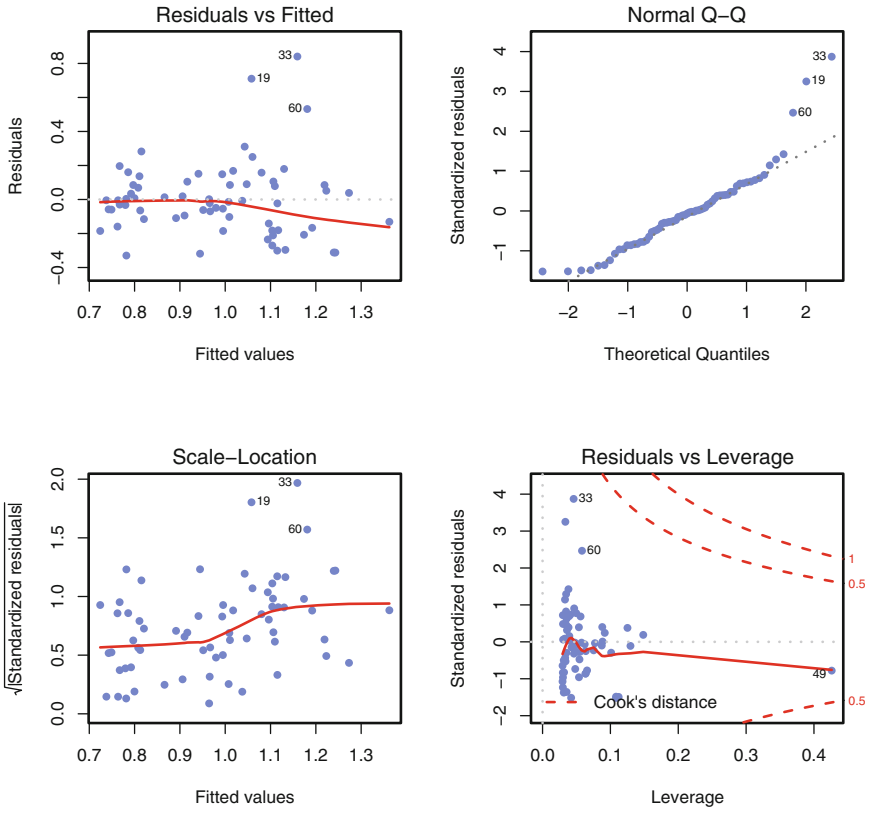

Fig. 11.18 Standard R plot of a linear model: `plot(rent.lm12m)`. The first three panels were discussed in Figure 11.5. The fourth is discussed in Figure 11.19. R by default fits a smooth curve to the points in these plots. The three largest residuals are indicated. "Largest" means larger than the others. There is no statistical significance associated with an identified point.

11.3.7 Residuals vs Leverage

We show R's standard set of regression diagnostic plots in Figure 11.18. The first three are essentially the same as the first three included in our Figure 11.5 constructed with `lmplot` from the **HH** package. R by default fits a smooth curve through both the plots of Residuals vs Fitted and $\sqrt{|Residuals|}$ vs Fitted. The fourth standard R plot, shown enlarged in Figure 11.19, shows the Residuals plotted against the leverage and includes contours of Cook's distance.

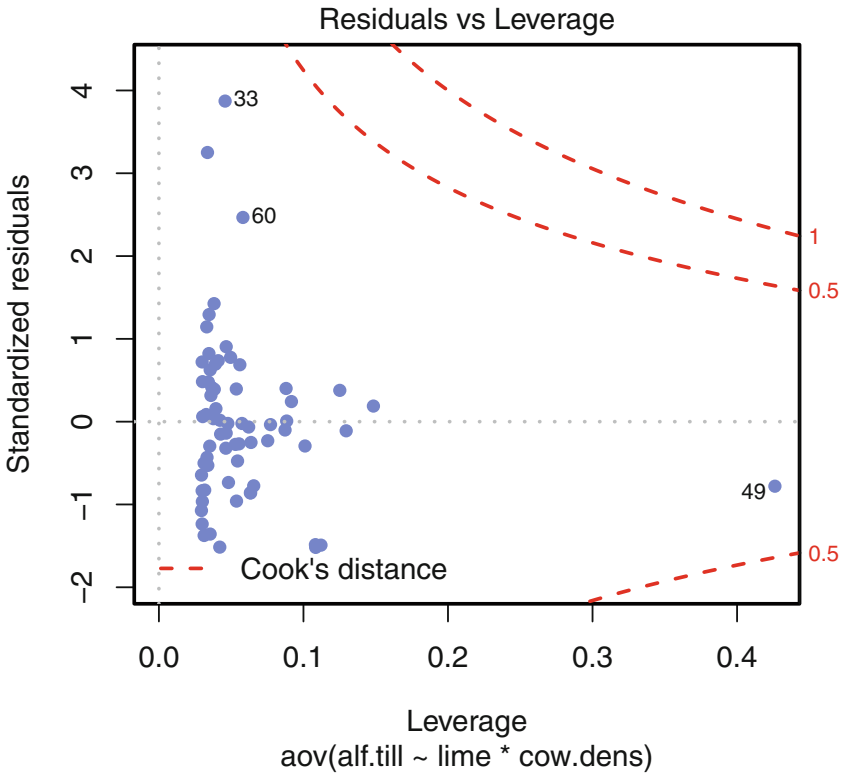

Fig. 11.19 This is a repeat of the fourth panel of Figure 11.18 with the smooth curve suppressed. The "Residuals vs Leverage" plot shows the standardized residuals e^* against the leverage h_i along with contours of Cook's distance. Cook's distance, a combined measure of the "unusualness" of a case's predictors and response, is discussed in Section 11.3.4. The contours of constant Cook's distance c are calculated as $\sqrt{c\,p\,(1-h_i)/h_i}$, where p is the number of estimated regression coefficients ($p = 2$ for simple linear regression). By default, contours are plotted for the two c-values 0.5 and 1. Note on the graph that the contour lines are closer to the 0-residual horizontal line for higher leverage values (corresponding to points farther away from $\bar{x}$) than for lower leverage values.

11.3.8 Calculation of Regression Diagnostics

Regression diagnostics are calculated from the matrix formulation of the equations in the "Sequenced calculation formulas" column of Table 11.7.

In R see the documentation for the functions `dfbetas`, `lm.influence`, and `plot.lm`. See also our functions `lm.case` and `plot.case` in the **HH** package.

Regression diagnostics in SAS are computed by adding the option `INFLUENCE` to the `MODEL` statement in `PROC REG`.

11.4 Exercises

We recommend that for all exercises involving a data set, you begin by examining a scatterplot matrix of the variables.

11.1. Data from Brooks et al. (1988), reprinted in Hand et al. (1994), relate the number of monthly man-`hours` associated with the anesthesiology service for 12 U.S. Naval hospitals to the number of surgical `cases`, the eligible `population` per thousand, and the number of operating `rooms`. The data appear in the file `data(hospital)`.

a. Construct and examine a scatterplot matrix of these data. Does it appear that multicollinearity will be a problem?

b. Fit the response to all three predictors, calculating the VIFs. Based on the analysis thus far, which predictor is the best candidate for removal? Why?

c. Fit the response with the predictor in part (b) removed.

d. Calculate the Studentized residuals, leverages, and Cook's distances for the model in part (c). Based on these calculations, what action would you recommend?

11.2. We previously encountered the dataset `data(hardness)` in Section 9.7 and Exercise 4.5. Since density is easily measured but hardness is not, it is desired to model hardness as a function of density.

a. Construct a histogram of `hardness` and confirm that a transformation is required in order to use this chapter's regression modeling procedures.

b. Regress the transformation of `hardness` you chose based on either part (a) or Exercise 4.5. For this regression, produce a scatterplot of the residuals vs the fitted values and of the residuals vs `density`. Conclude from these plots that a quadratic regression is appropriate.

c. We illustrate a linear and a quadratic fit of the hardness data in Figure 9.5 and Table 9.4. Produce residual plots and regression diagnostics for both models.

11.3. The dataset `data(concord)` is described in Exercise 4.6. Use multiple regression analysis to model `water81` as a function of a subset of the five candidate predictors. Consider transforming variables to assure that the assumption of regression analysis are well satisfied. Carefully interpret, in terms of the original model variables, all regression coefficients in your final model.

11.4. Creatine clearance is an important but difficult to measure indicator of kidney function. It is desired to estimate `clearance` from more readily measured variables. Neter et al. (1996) discuss data, originally from Shih and Weisberg (1986), relating

clearance to serum clearance concentration, age, and weight. The datafile is data(kidney).

a. Regress clearance on each of the three individual predictors. Investigate the adequacy of this model.

b. Improve on the model in part (a) by adding to the set of candidate predictors the squares and pairwise products of the three original predictors. Conclude that the addition of one of these six new candidates improves the original model.

c. Investigate the adequacy of this model.

d. Carefully interpret each of the four estimated regression coefficients in terms of the model variables.

11.5. Heavenrich et al. (1991) provide data on the gasoline mileage (MPG) of 82 makes and models of automobiles as well as 4 potential predictors of MPG. The data appear in data(mileage). The potential predictors are

WT: vehicle weight in 100 lbs

HP: engine horsepower

SP: top speed in mph

VOL: cubic feet of cab space

We wish to use them to model MPG.

a. Produce a scatterplot matrix and comment on the plots of MPG vs HP and of HP vs SP.

b. Regress MPG on WT, HP, and SP. Are the signs of the estimated regression coefficients as expected? Explain what is causing the anomaly.

c. First regress MPG on WT and SP and then regress MPG on WT and HP. Which of these two regressions is preferred?

d. For the model you prefer in part (c), produce a normal plot of the residuals and a plot of the residuals vs the fitted values. What do you conclude?

e. Regress the log of MPG on WT and SP and also the log of MPG on the log of WT and SP. Produce residual plots and normal probability plots from both of these runs. Based on the numerical output and plots, explain which model is preferred.

f. For the preferred model, produce case diagnostics. For each flagged case, indicate what is unusual about it.

11.6. Neter et al. (1996) discuss a dataset relating the amount of life insurance carried in thousands of dollars (lifeins) to average annual income in thousands of dollars (anninc) and risk aversion score (riskaver), for 18 managers, where higher scores connote greater risk aversion. The data are contained in the file data(lifeins).

a. Produce a scatterplot matrix. Which of `anninc` and `riskaver` appears to be more closely related to `lifeins`?

b. Regress `lifeins` on `anninc` and `riskaver`, storing the residuals.

c. From a scatterplot of these residuals vs `anninc`, conclude that the relationship between `lifeins` and `anninc` is nonlinear. Define the square of average annual income, `annincsq` = `anninc`^{2}. Regress `lifeins` on the three predictors `anninc`, `annincsq`, and `riskaver`. Plot the residuals from this run against `anninc`. Based on this plot, discuss whether addition of the curvature term seems worthwhile.

d. Identify cases (managers) whose values indicate either high influence or high leverage. Also note whether these cases have high values of any of the measures Cook's distance, DFFITS, or DFBETAS. If so, interpret such high values in terms of the model variables.

11.7. Refer to `data(hpErie)`, previously considered in Exercise 9.3.

a. Rerun the regression for the final model you found in Exercise 9.3b, this time requesting a complete set of regression diagnostics.

b. Closely examine the values of the diagnostics for the two high-priced houses that are the focus of Exercise 9.3c. Would you recommend both of these houses or just one of them for special scrutiny?

11.8. Prove the equivalence of the intuitive and computational formulas for the following case statistics:

a. DFFITS in Equations (11.10) and (11.11)

b. Cook's distance in either intuitive Equation (11.7) or (11.8), and computational Equation (11.9)

c. Studentized deleted residual in Equations (11.5) and (11.6)

11.9. Explore the diagonals of the hat matrix $H = X(X'X)^{-1}X'$.

a. Prove that all leverages satisfy $\frac{1}{n} \le h_i \le 1$. Since H is a projection matrix, show that the upper bound on the diagonals is 1. Since the column $X_0 = 1$ is included in the X matrix, show that the lower bound on the diagonals is $\frac{1}{n}$.

b. Show that the average leverage

$$\frac{\sum_i h_i}{n} \equiv (p+1)/n$$