Appendix D
R Code
Jon Baldock and Kimberly Garland Campbell

Figures and Captions

Fig. 2.1

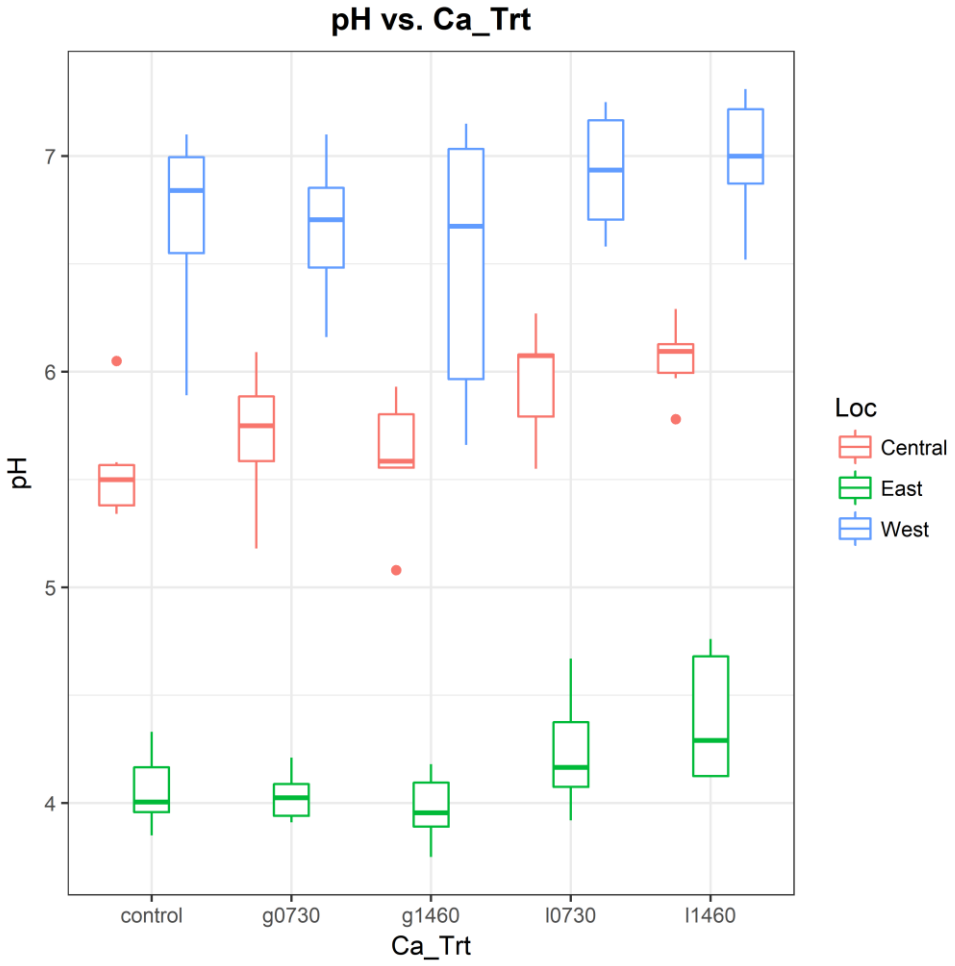


Fig. 2.1. This boxplot is the R equivalent of Fig. 5 in Chapter 2. The R version identifies three observations as being potential outliers, but Fig. 5 in Chapter 2 did not. This discrepancy is due to SAS and R using different algorithms for determining quantiles, which can result in sizable differences for the interquartile range and outlier fence, especially in small samples such as these (use the R help for the quantile() function for more information on the nine options for quantiles).

Fig. 2.2

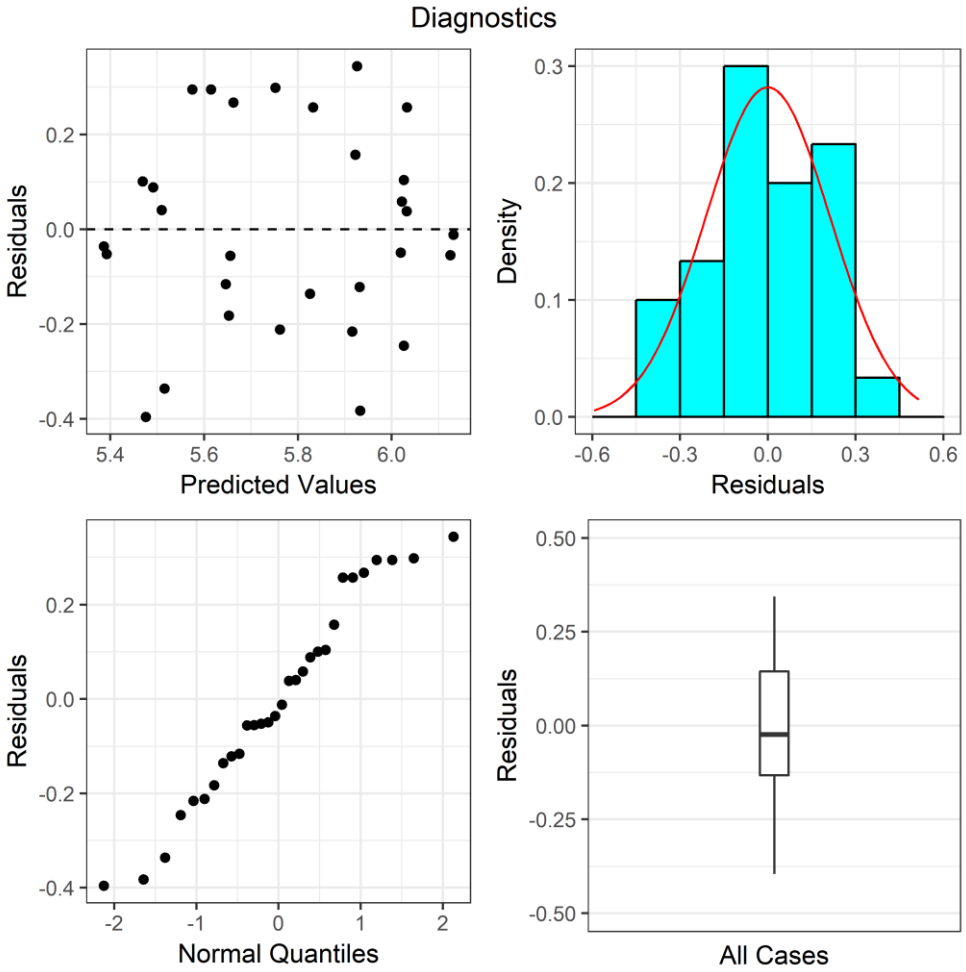


Fig. 2.2. These are the diagnostic plots for the ANOVA of the Statbean Study pH data from the Central Location. This panel is comparable to the top section of Fig. 6 in Chapter 2. Both show there are only minor deviations from normality and no evidence of systematic departures from the model. The R packages used to make these graphs do not provide a method to overlay a straight line on the Q-Q plot (lower left). Also, SAS uses the lower right section to print a table with four statistics describing the distribution of the residuals and four fit statistics. Because the R's summary() function gave the five-number summary for the residuals and we are not comparing models, I found it easier and more informative to put a box and whisker plot in this position.

Fig. 2.3

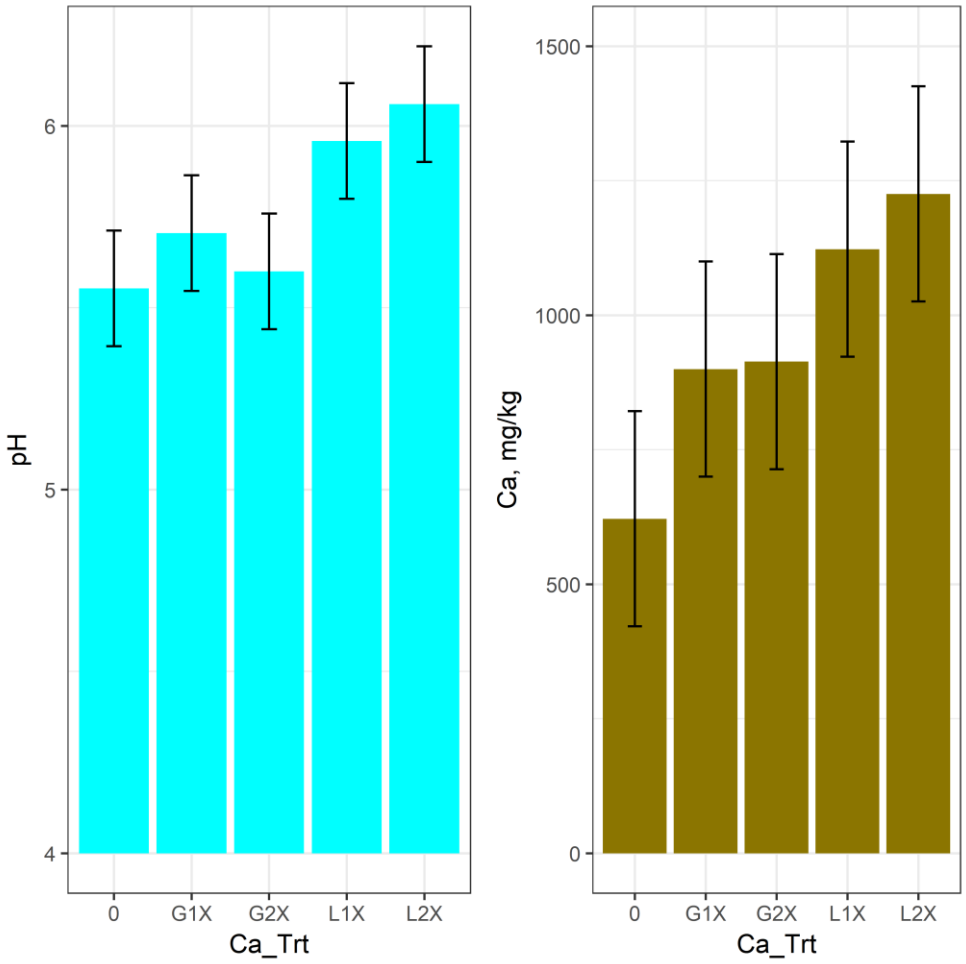


Fig. 2.3. These bar charts of the calcium treatment means for soil pH and Ca with LSD error bars (averaged over the mulch levels so n=6) at the Central Location correspond well with Fig. 7 in Chapter 2. Observe the pH chart above and the one in Chapter 2 move the baseline to pH=4 to emphasize the differences among treatments. In both graphs above, the LSD error bars are centered so half the LSD is above the mean and half below. Thus, if these error bars overlap, the treatments are not significantly different at the 5% level (Ca LSD = 400 mg/kg, pH LSD = 0.318).

Fig. 2.4.

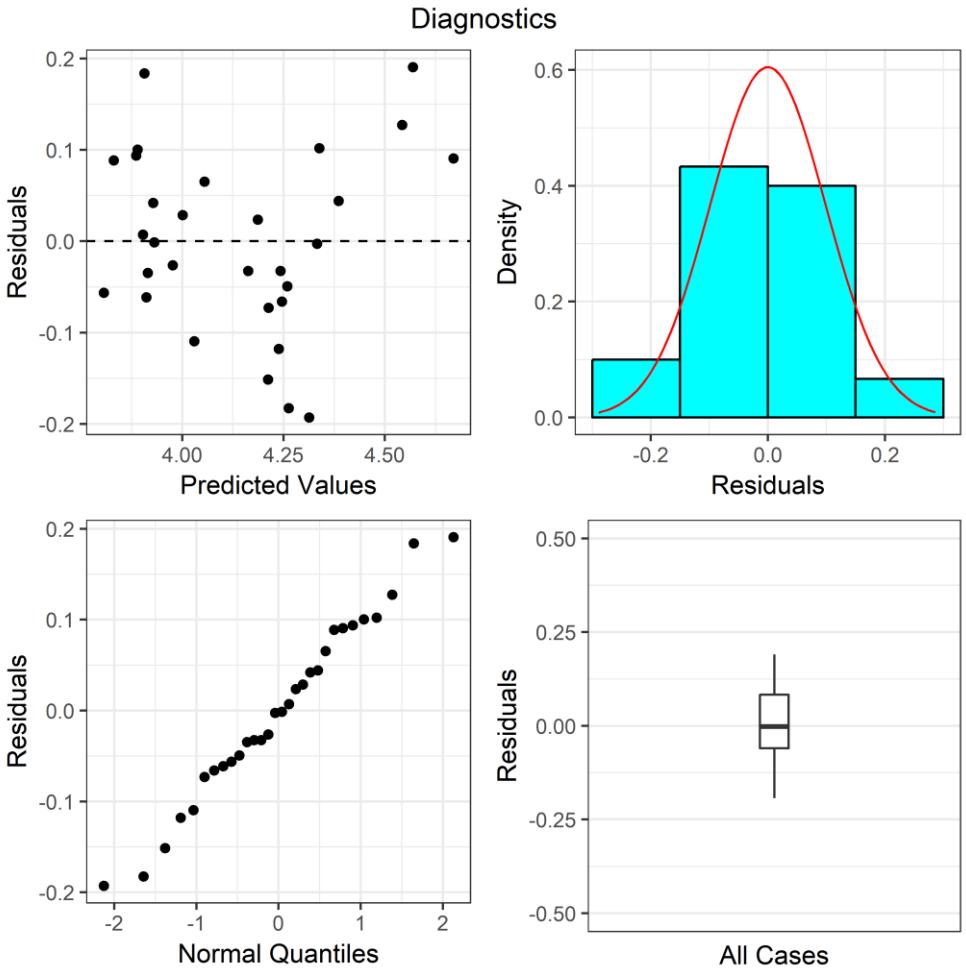


Fig. 2.4. These are the diagnostic plots for the ANOVA of the Statbean Study pH data from the East Location. This panel is comparable to the middle section of Fig. 6 in Chapter 2. Both show there are only minor deviations from normality and no evidence of systematic departures from the model. The structural differences between these R graphs and those in the SAS panel are discussed in the caption for Fig. 2.2.

Fig. 2.5.

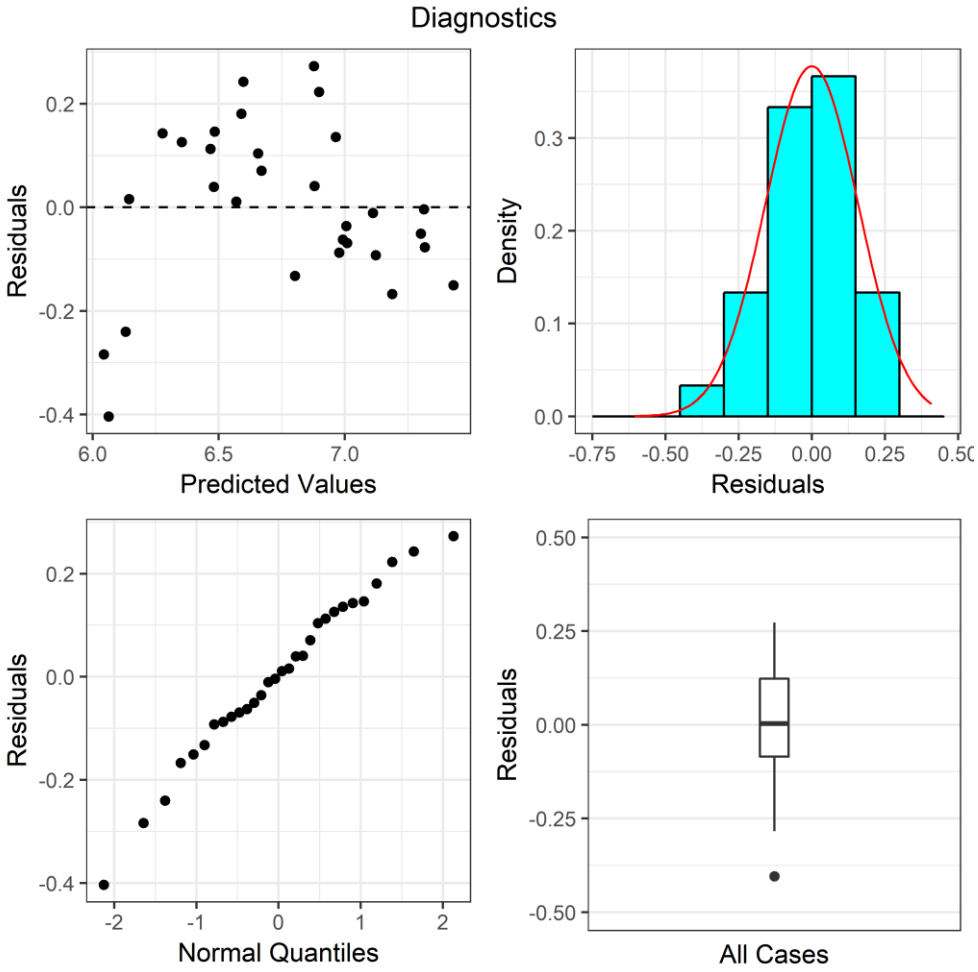


Fig. 2.5. These are the diagnostic plots for the ANOVA of the Statbean Study pH data from the West Location. This panel is comparable to the bottom section of Fig. 6 in Chapter 2. Both show there are only minor deviations from normality and no evidence of systematic departures from the model. One potential outlier shows on the box and whisker plot, but the scaled residuals place at about -2 on a standardized scale so it is not likely an outlier (see R output text). The structural differences between these R graphs and those in the SAS panel are discussed in the caption for Fig. 2.2.

Fig. 2.6

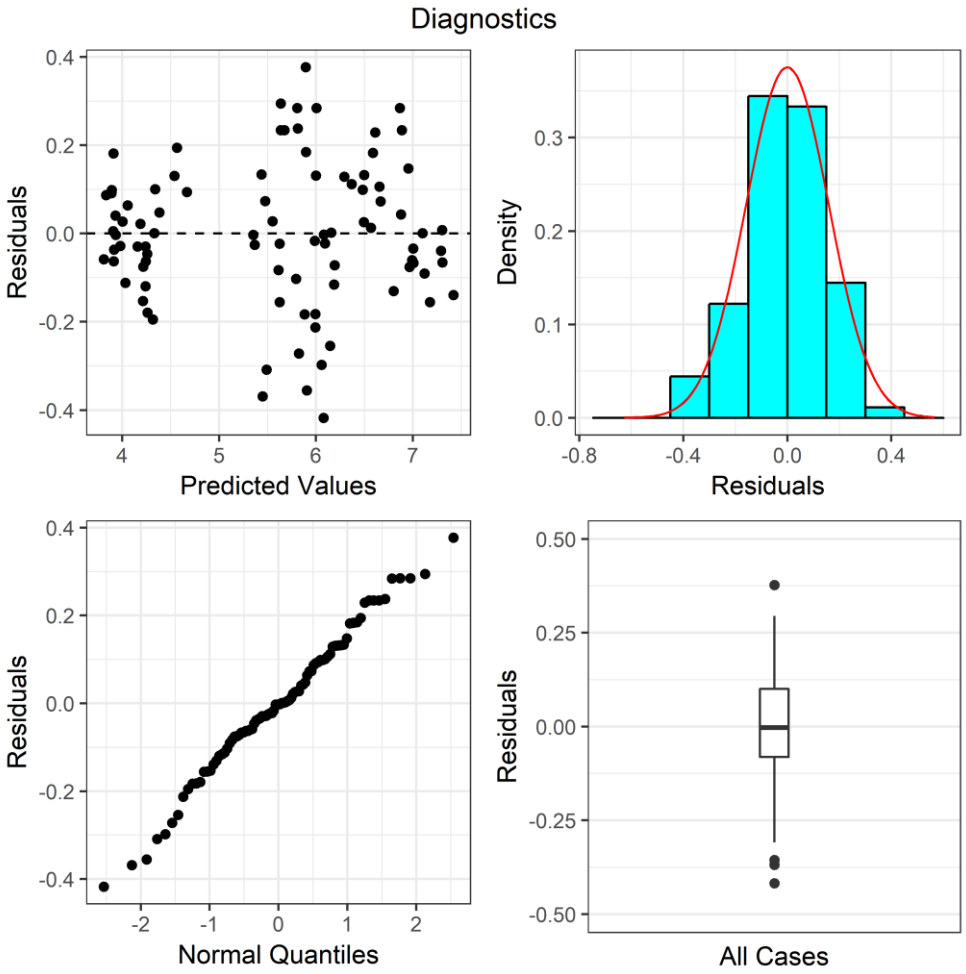


Fig. 2.6. These are the diagnostic plots for the ANOVA of the Statbean Study pH data combined over all three locations. This panel is comparable to Fig. 8 in Chapter 2. Both show there are only minor deviations from normality and no evidence of systematic departures from the model. Four potential outliers show on the box and whisker plot, but the minimum and maximum scaled residuals were about -2 and +1.8 on a standardized scale so it is unlikely that these are outliers (see R Output 2.4). The structural differences between these R graphs and those in the SAS panel are discussed in the caption for Fig. 2.2.

Fig. 2.7

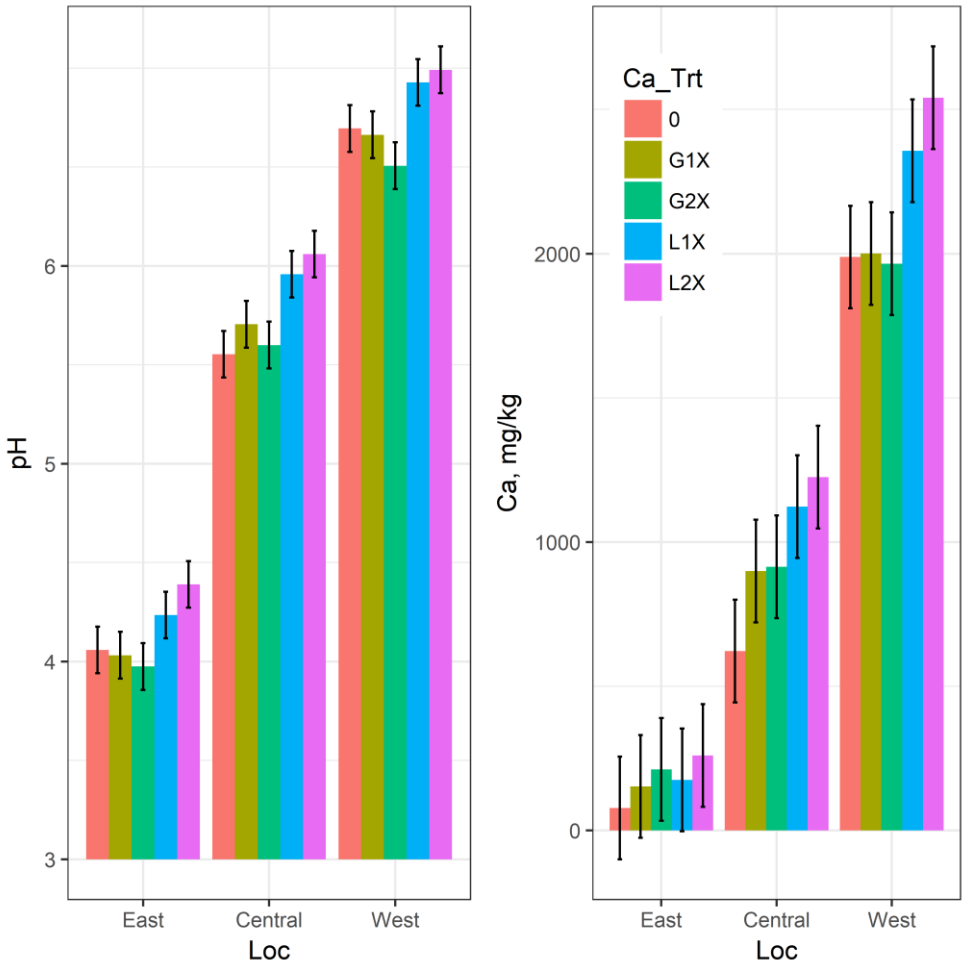


Fig. 2.7. These bar graphs plot the soil pH and Ca means for the Loc-by-Ca_Trt combinations with error bars equal to the LSD (n=6 per bar). They are the equivalent of Fig. 9 in Chapter 2. The LSD error bars are centered so half the LSD is above the mean and half below. Thus, if these error bars overlap, the treatments are not significantly different at the 5% level (pH LSD = 0.236; Ca LSD = 341).

Fig. 3.1

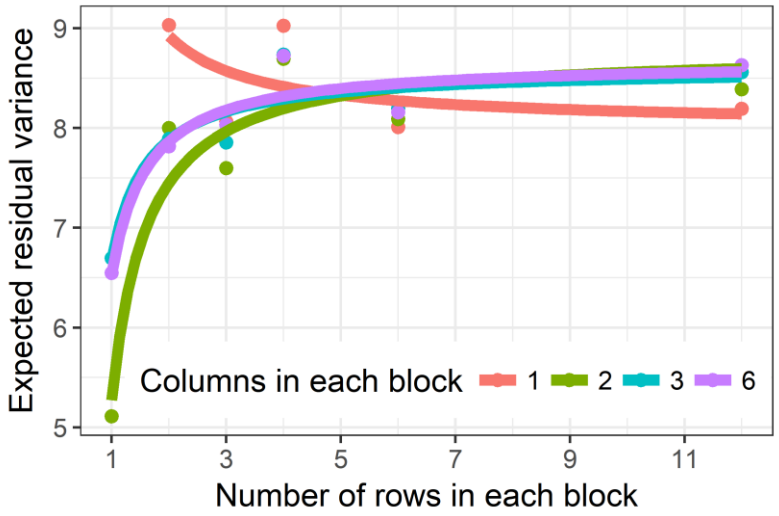


Fig. 3.1. Within block mean squares for 23 block arrangements of row and columns. This figure compares well with the figure in Box 3 of Chapter 3. Both show the minimum variation within blocks occurs with one row per block and two columns per block and that one row per block with three or six columns per block are viable alternatives.

Fig. 4.1

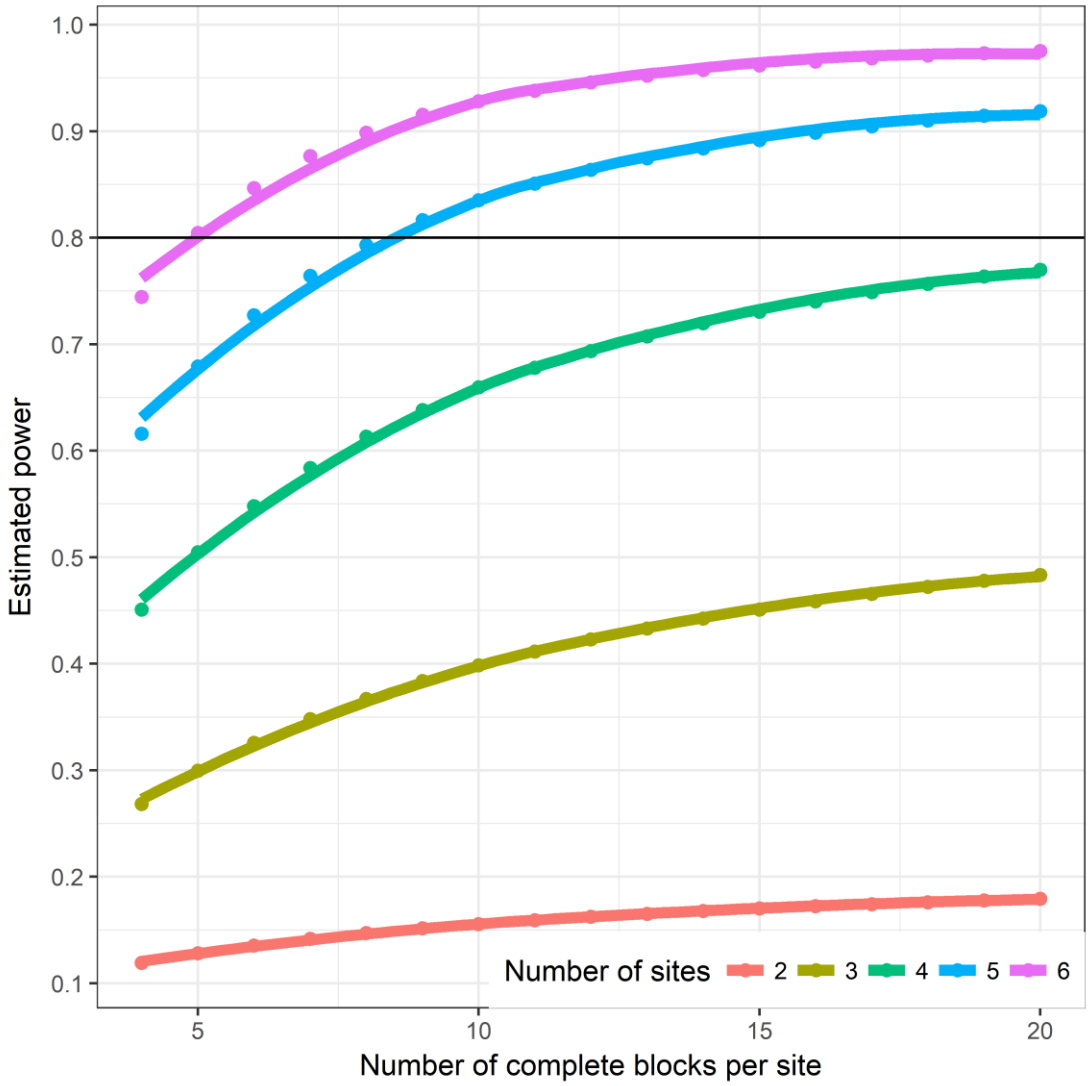


Fig. 4.1. This graph emulates Fig. 3 in Chapter 4 and shows the relationship between power and replications (complete blocks per site) for a range of number of locations (sites) in the study.

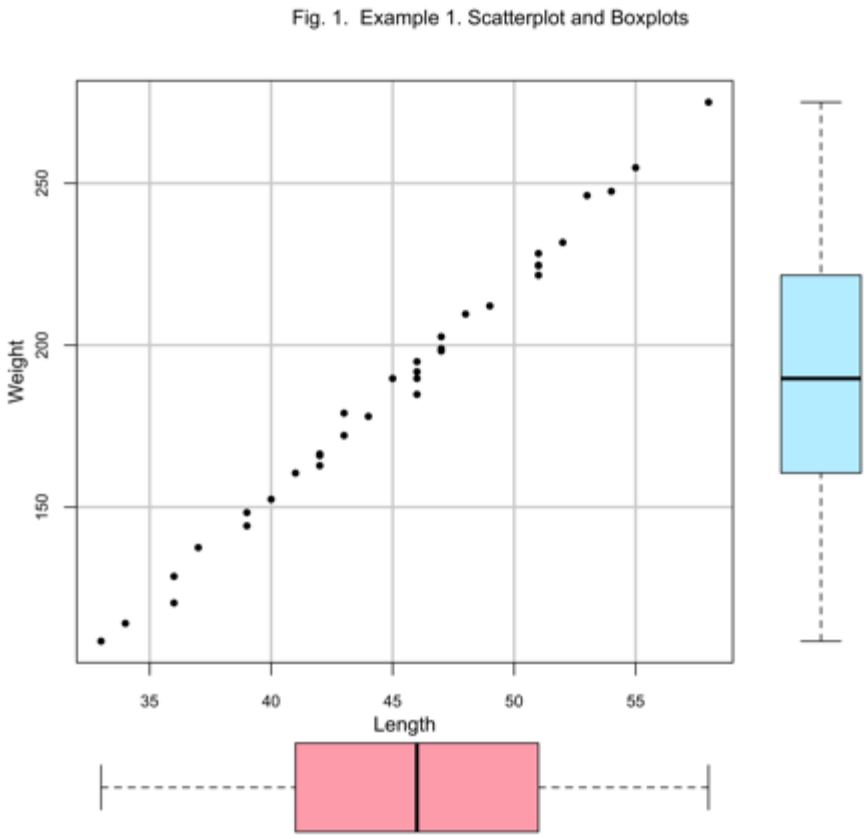


Figure 6.1 Scatter plot of data from Chapter 6, Example 1, eel.dat with marginal box plots.

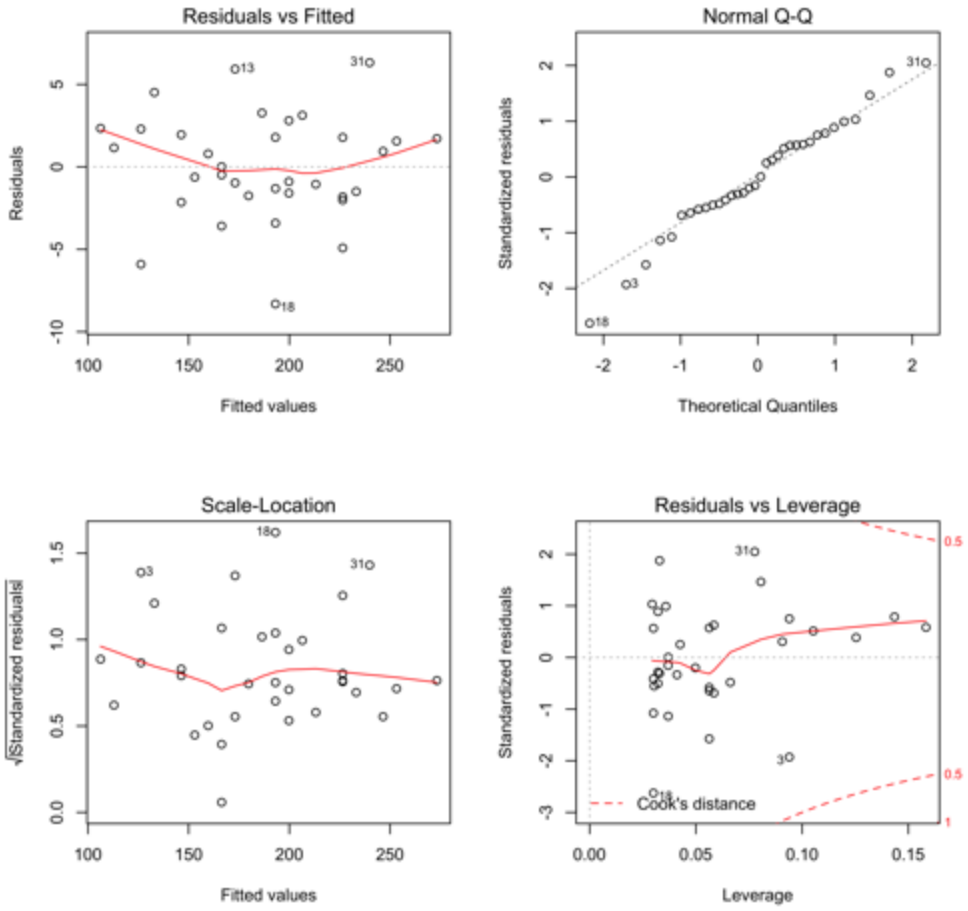


Figure 6.2. Resduals plots from simple linear regression of eel wieight on length for Chapter 6, Example 1.

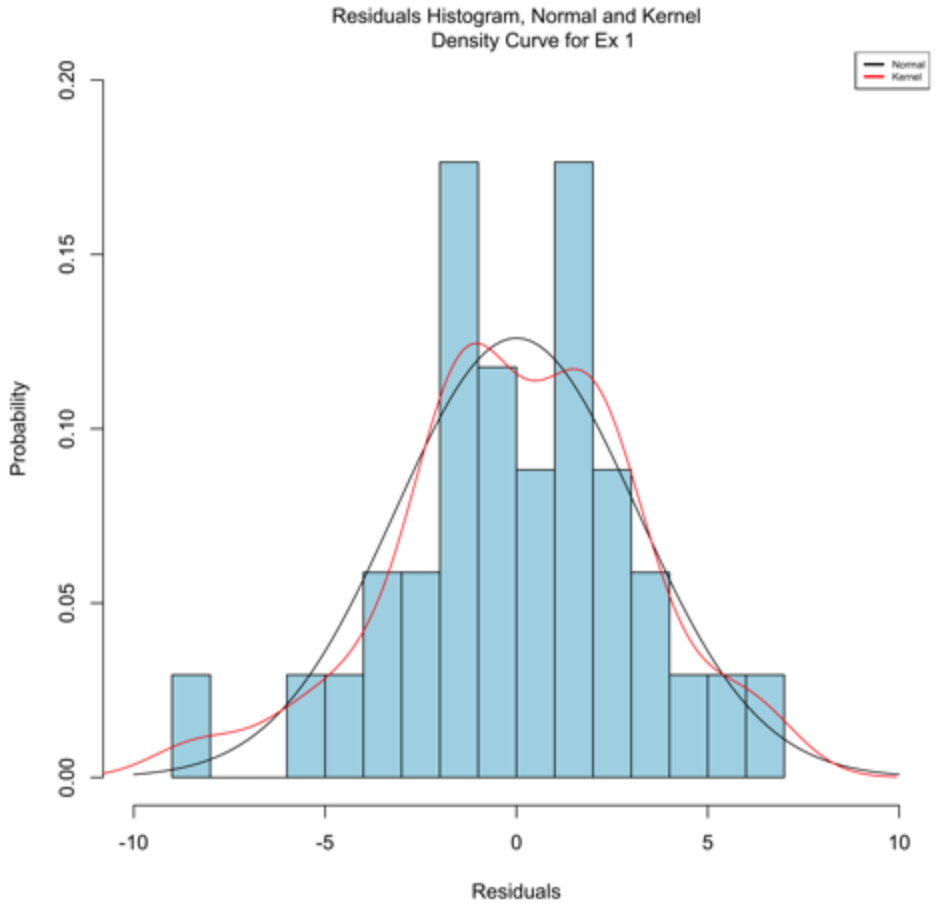


Fig. 6. 3a. Frequency histogram of residuals from model weight ~ length for eel data from Chapter 6, Example 1.

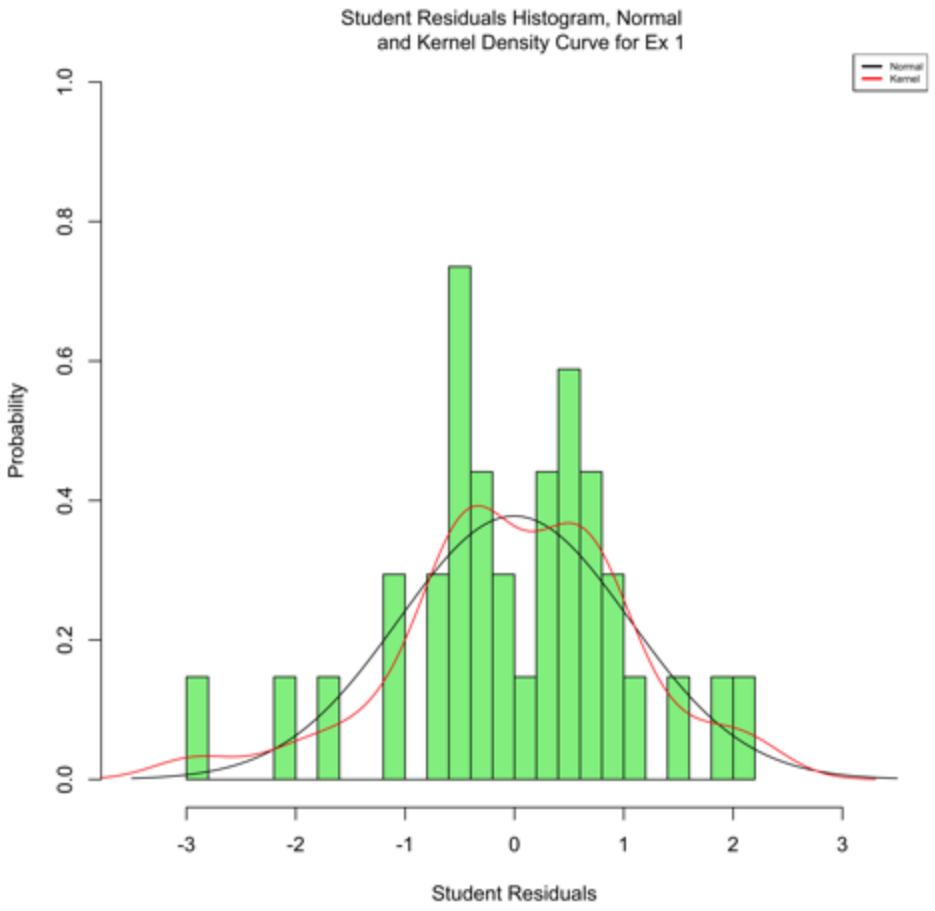


Figure 6.3b. Frequency histogram of standardized (also known as student) residuals from linear model weight ~ length for the eel data from Chapter 6, Ex. 1.

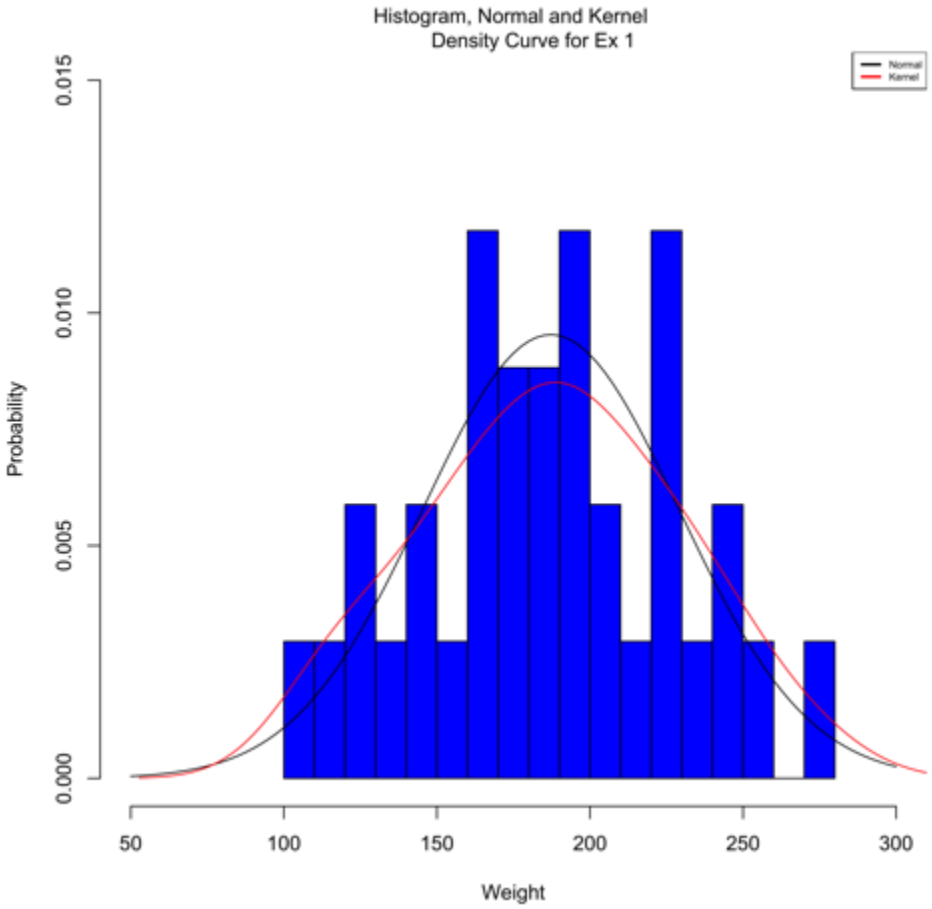


Figure 6.3c. Frequency histogram of Eel weights for Chapter 6, Example 1.

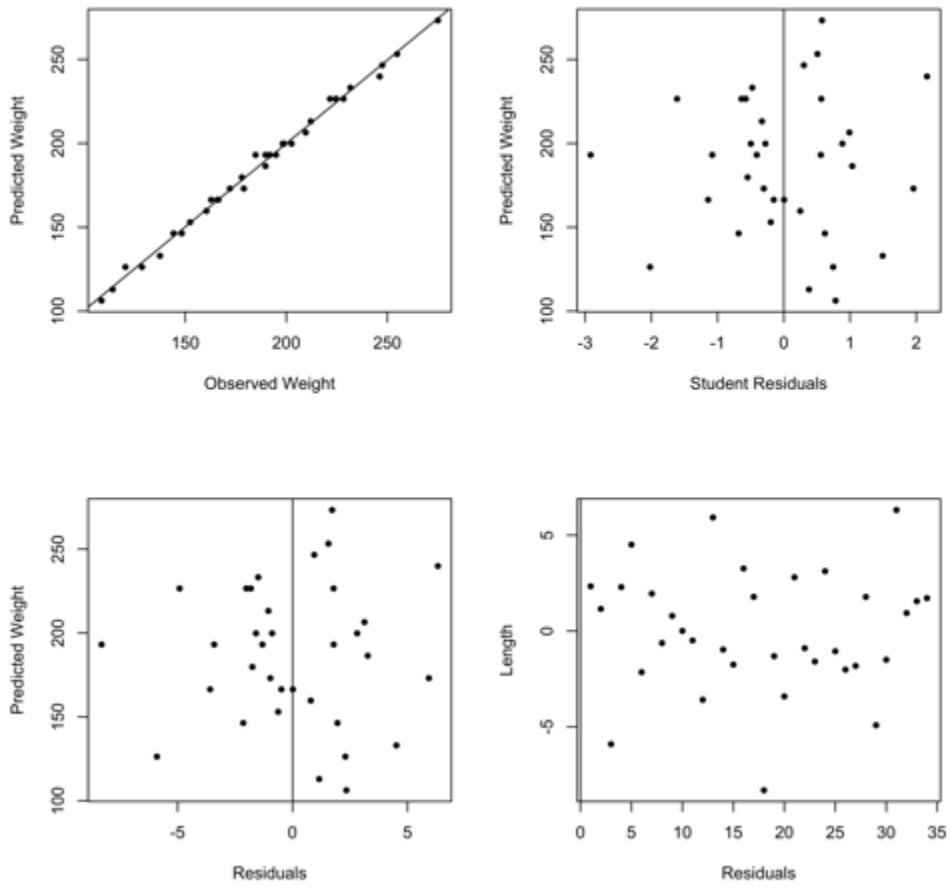


Figure 6. 4. Several diagnostic plots for regression of weight on length for eel data from Chapter 6, Example 1. The residuals do not exhibit a pattern but there are a few outliers.

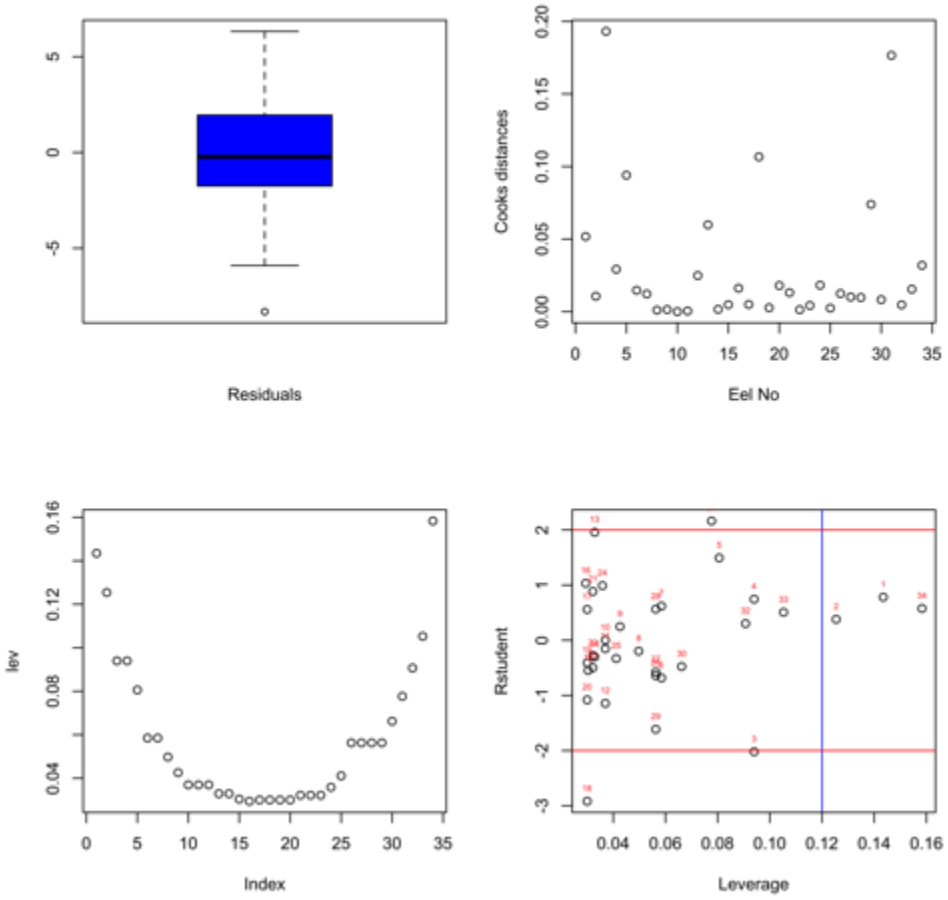


Figure 6.5. Additional diagnostic plots for linear regression of weight ~ length for Eel data from Chapter 6, Example 1. These plots include a residuals boxplot and several measures of influence.

Fig

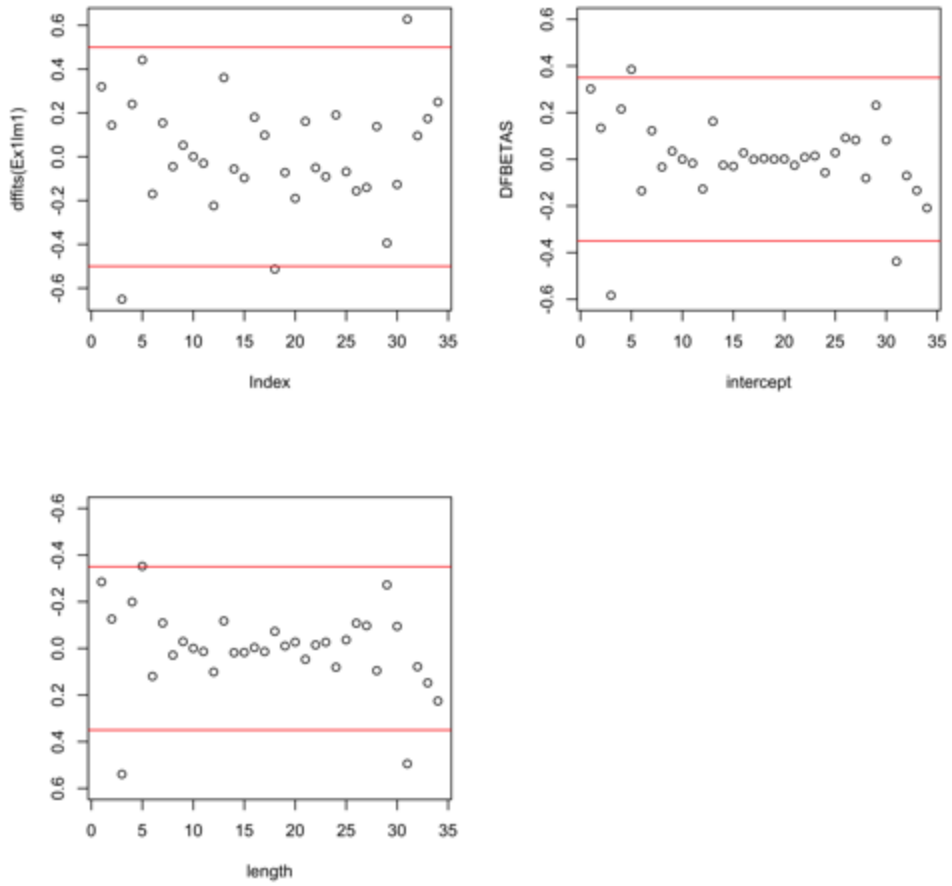


Figure 6.6 Additional influence plots for regression of weight ~ length for eel data from Chapter 6, Example 1.

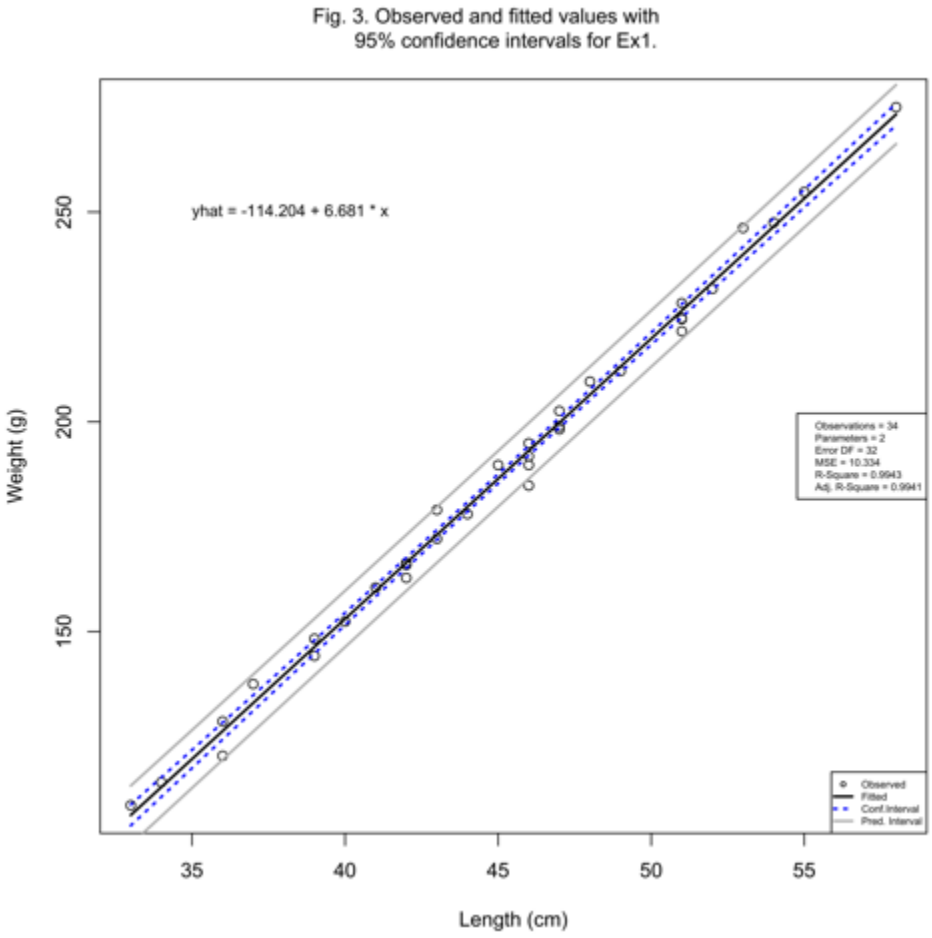


Figure 6.7. Plot of eel weight on length with confidence and prediction intervals for Chapter 6, Example 1.

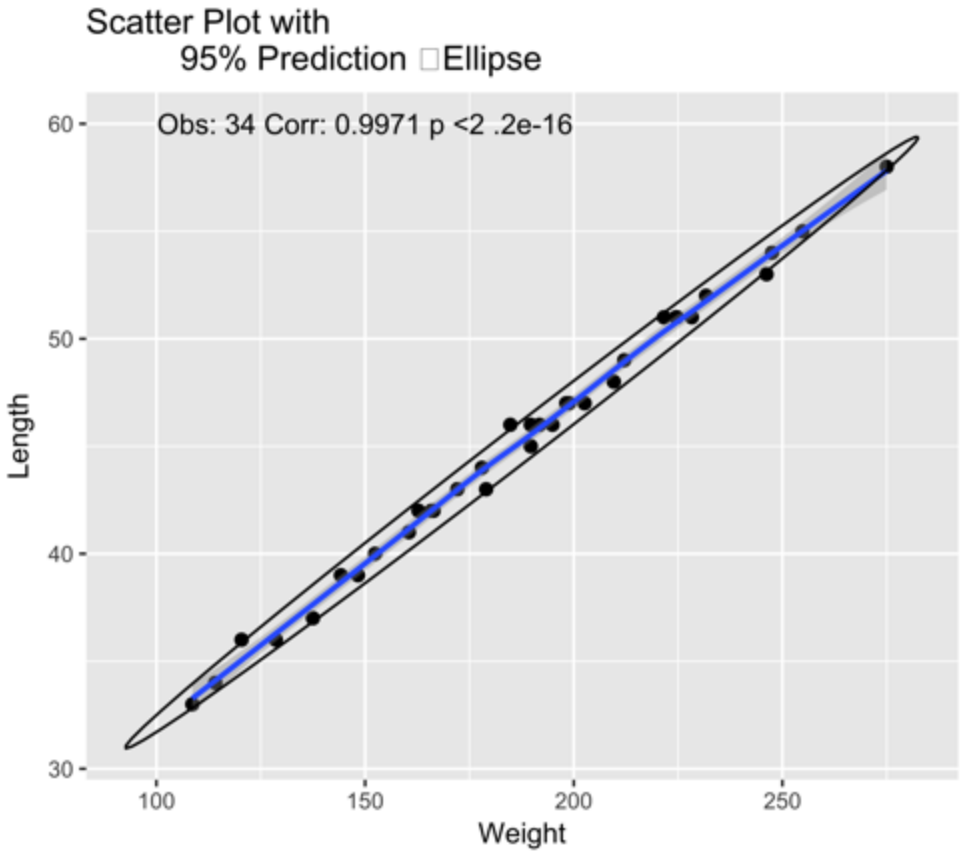


Figure 6.8. Plot of prediction ellipse for regression of weight on length for eel data in Chapter 6, Example 1.

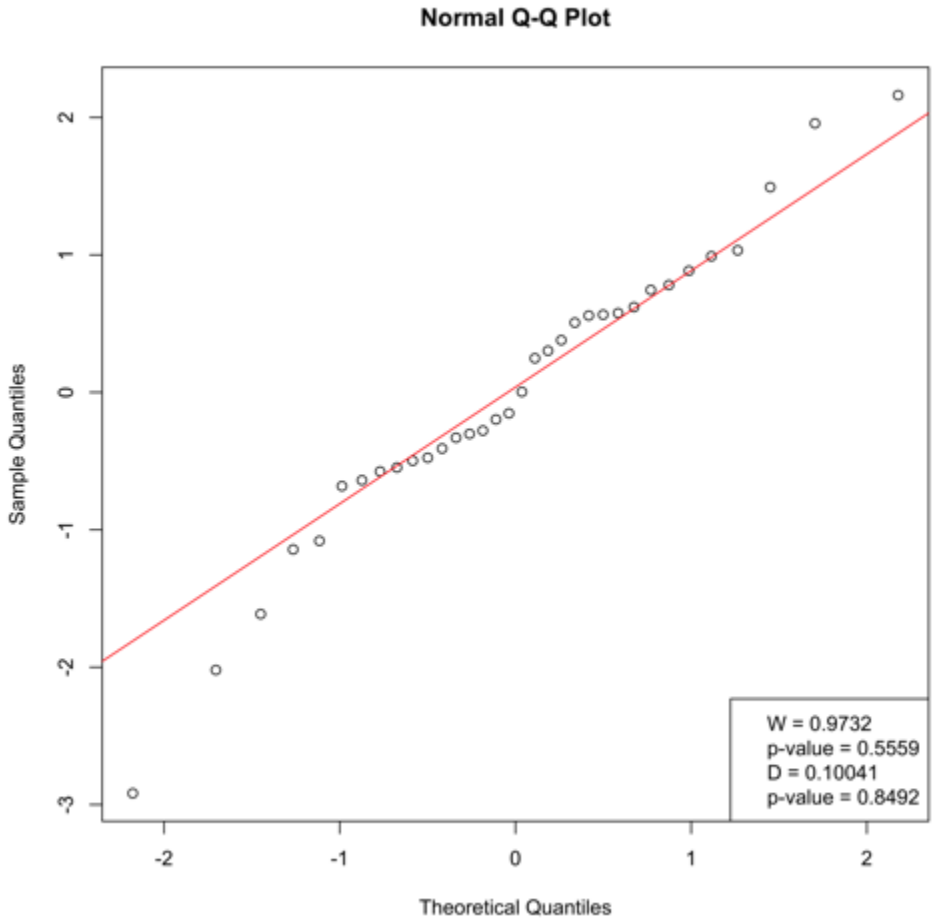


Figure. 6.9. The QQ plot for the regression of weight on length for the eel data from Chapter 6, Example 1. Some outliers are evident at extremes of plot.

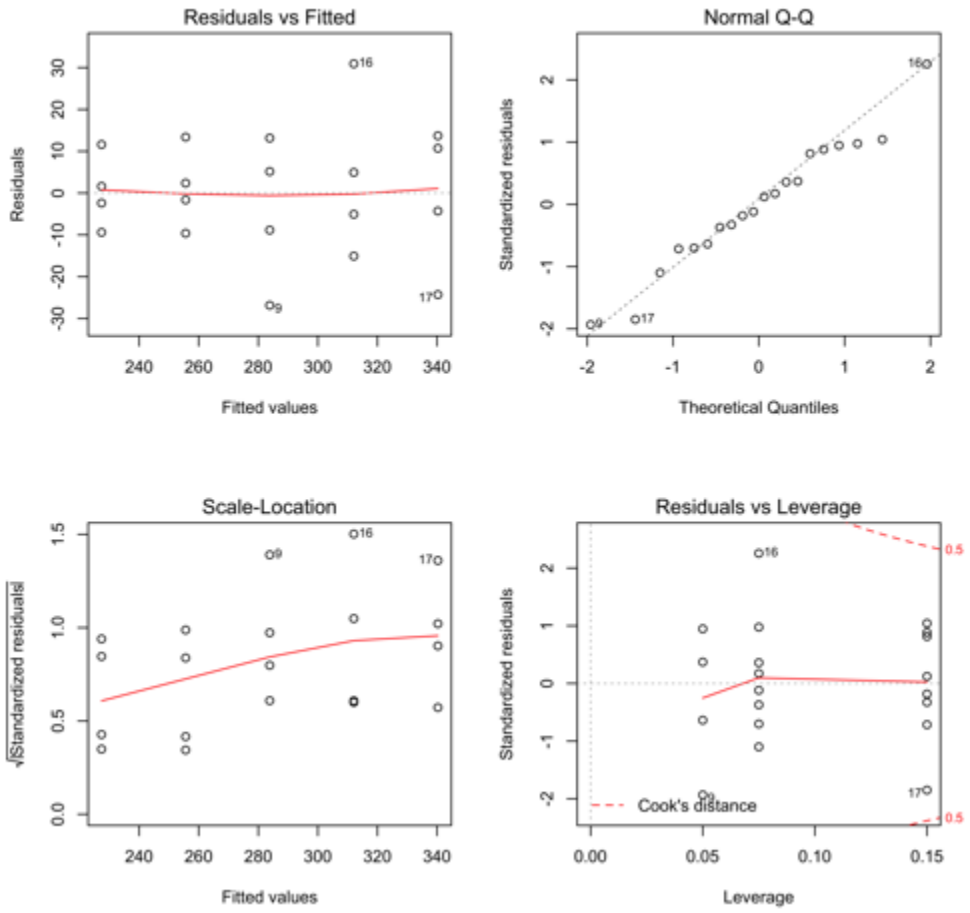


Figure 6.10. Diagnostic plots from simple linear regression for fibre content on day for Chapter 6, Example 3.

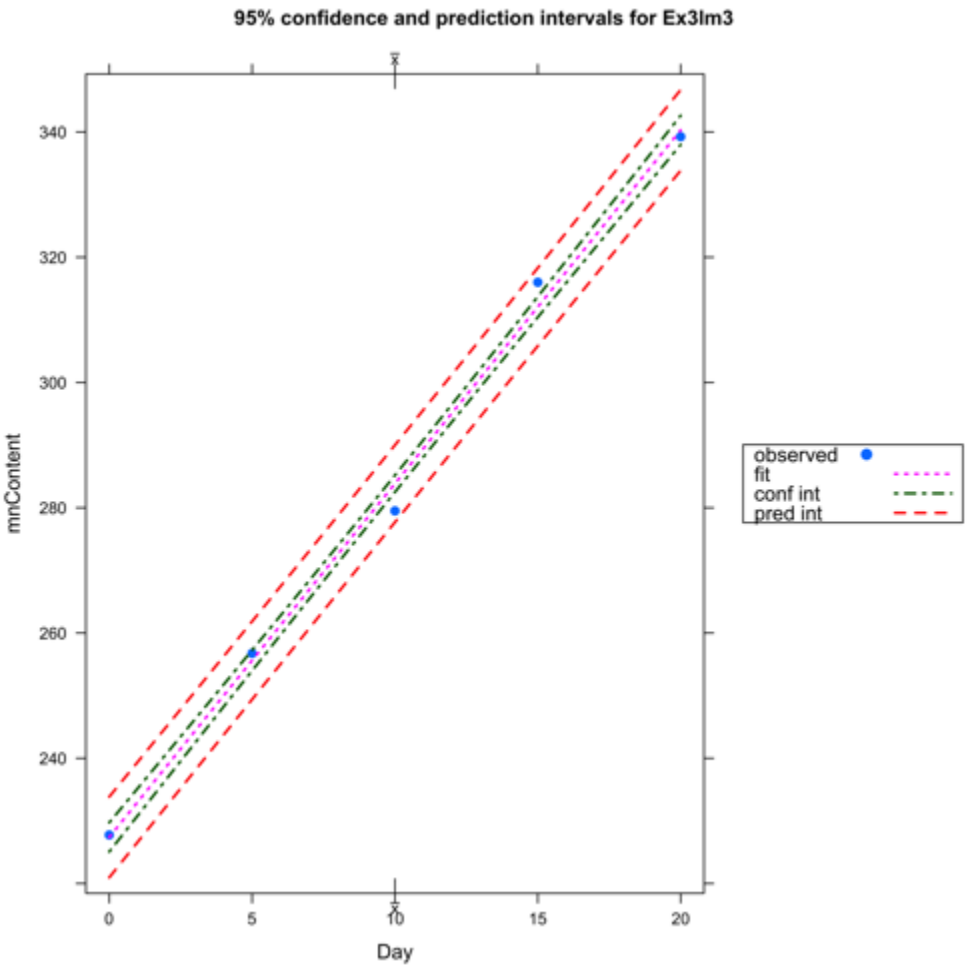


Figure 6.11. Confidence and predication interval plot for the mean of content regressed on day for the fibre data from Chapter 6, Example 3.

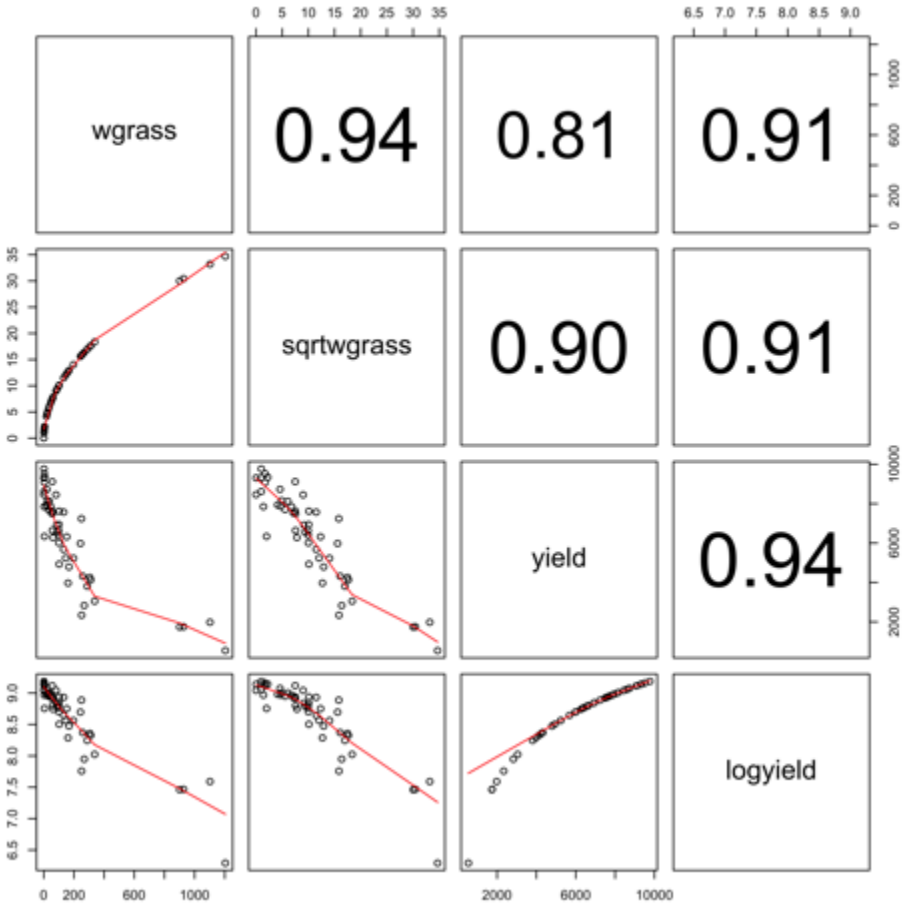


Figure 6.12. Correlations among variables in the grass dataset where the plot yields are regressed on counts of windgrass. Several variables were calculated from the initial count and yield data. Plots of the relationships among the variables are on the lower diagonal and the pearson correlation coefficients are above the diagonal. The data are from Chapter 6, Example 4.

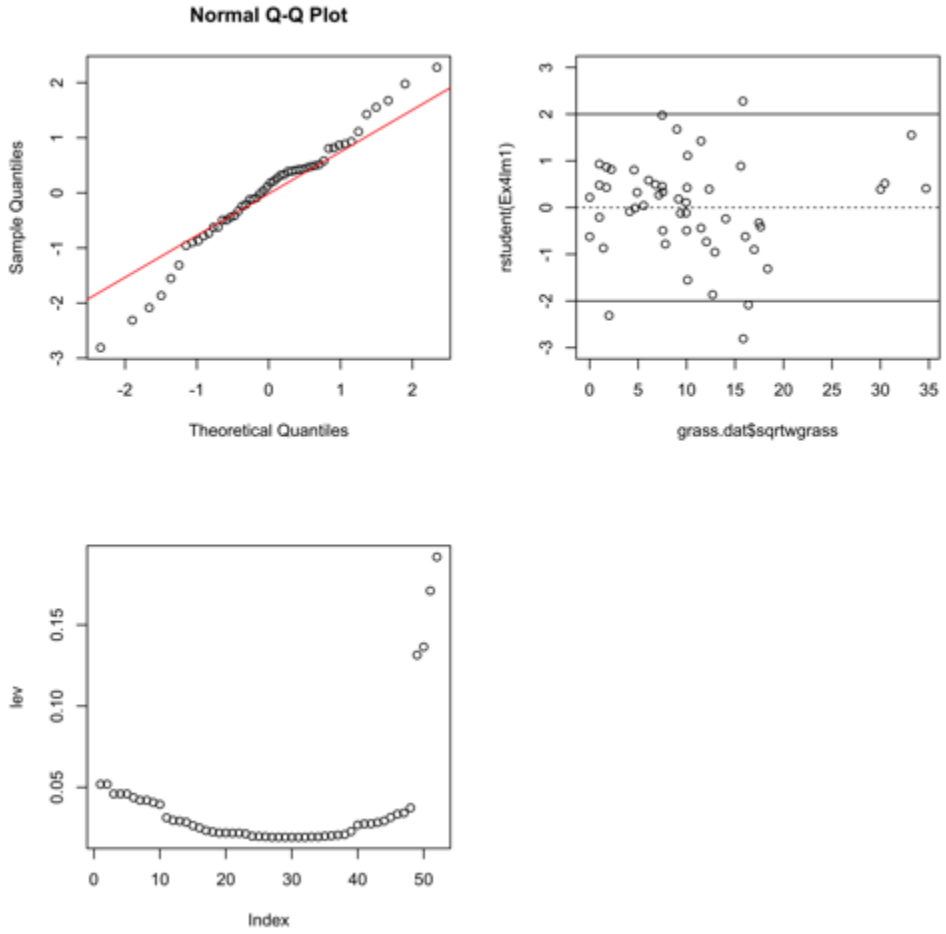


Figure 6.13. Diagnostic plots from linear regression of the plot yield on the square root of the windgrass counts for the grass data from Chapter 6, Example 4. The regression model is not an adequate fit to the data.

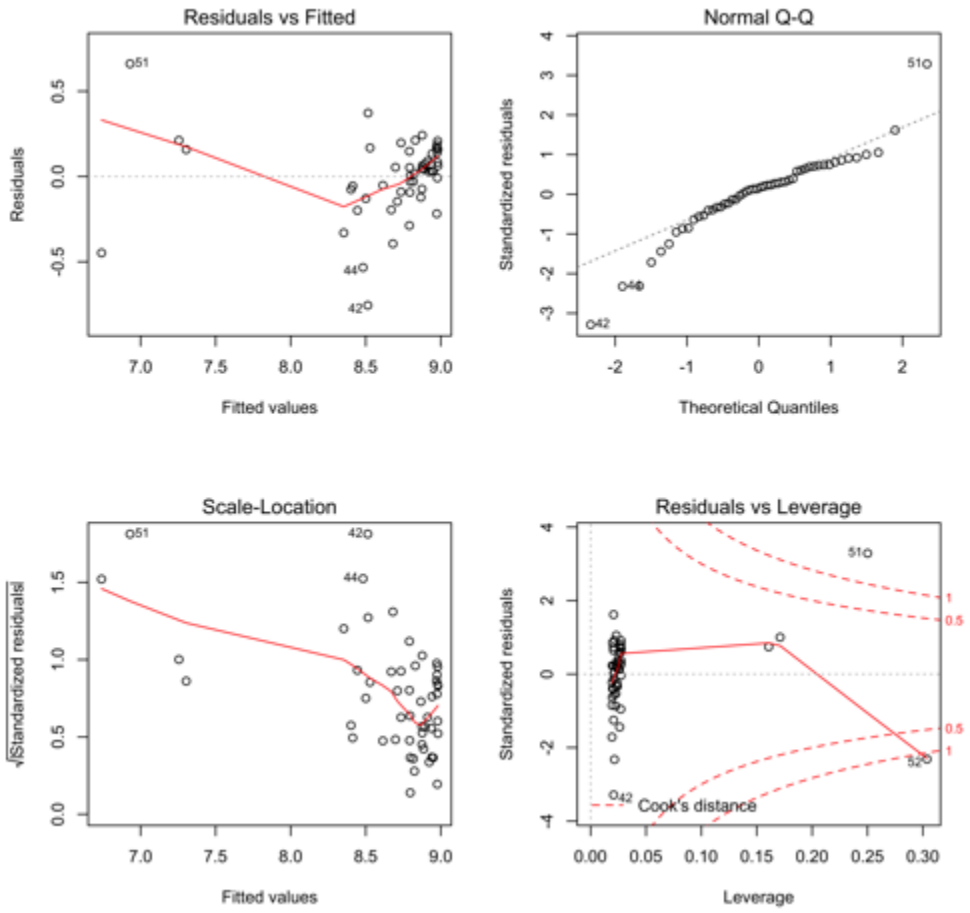


Figure 6.14. Diagnostic plots from regression of the log-transformed plot yield on the windgrass counts for the grass data in Chapter 6, Example 4. These data illustrate that this model is not an adequate fit to the data.

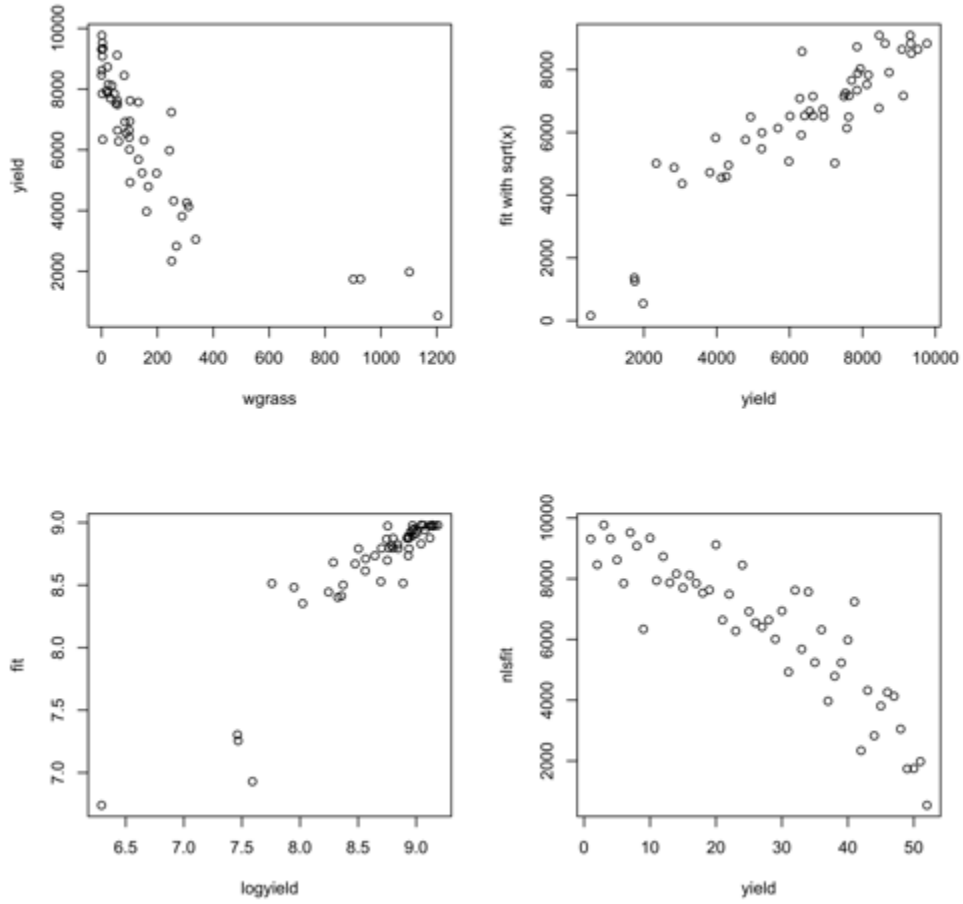


Figure 6.15. Plots of the various regressioin models that have been tested for the relationship between the plot yield and the windgrass counts in Chapter 6, Example 4. All of the plots illustrate some departure from the fitted values and the measured values. The plot in the upper left of the regression of yield on the square root of the windgrass counts has the best match of data to the fitted values but there are still problems.

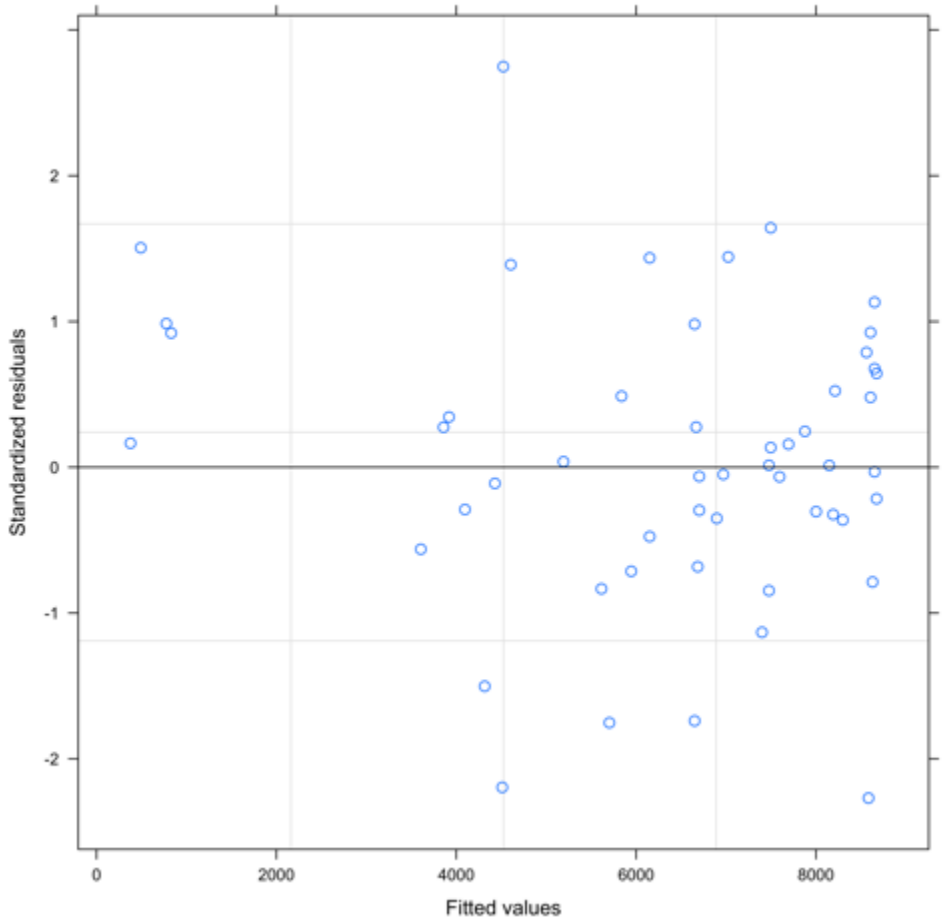


Figure 6.16. Plot of studentized residuals against fitted values for the nonlinear regression of plot yield on windgrass counts, using the formula: Yield ~ (a*exp(b*wgrass)), where a= 8000, and b = -0.01) for the grass data in Chapter 6, Example 4.

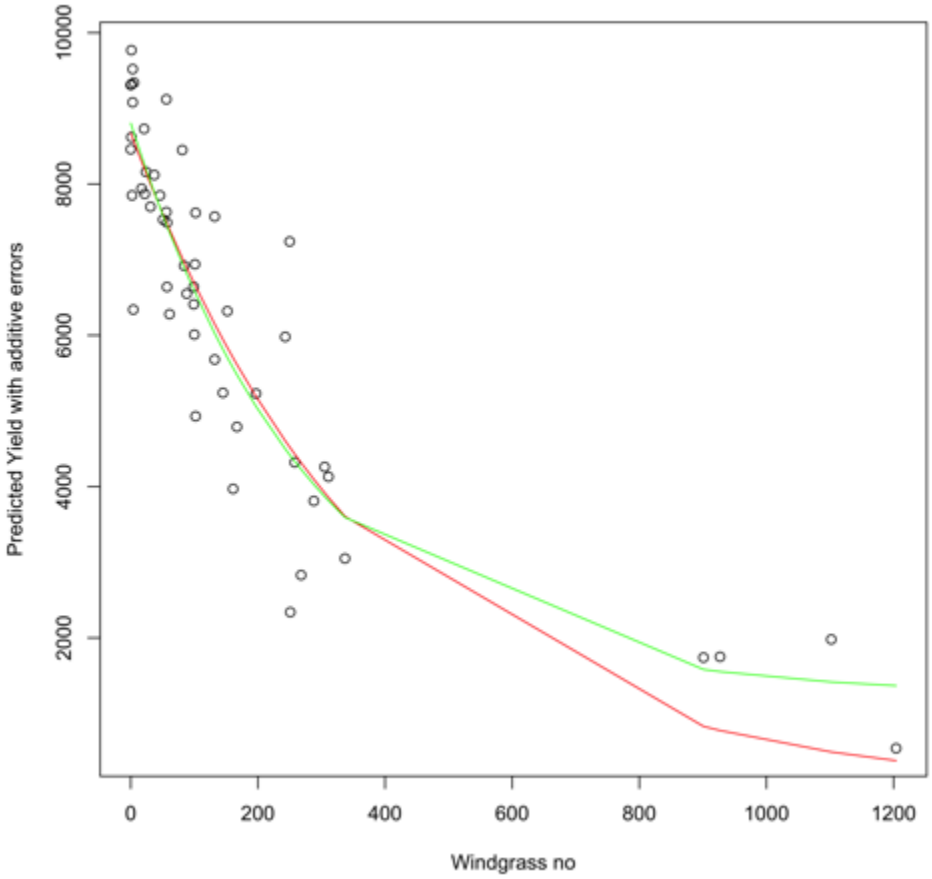


Figure 6.17. Comparison of two nonlinear regressions, both with the same formula, Yield ~ (a*exp(b*wgrass)). The red line in the fit when a= 8000, and b = -0.01 and the green line is the fit when a = 7000, b = -0.01, and c = 1200. The second model (green line) seems to be a better fit. These data are from the grass data in Chapter 6, Example 4.

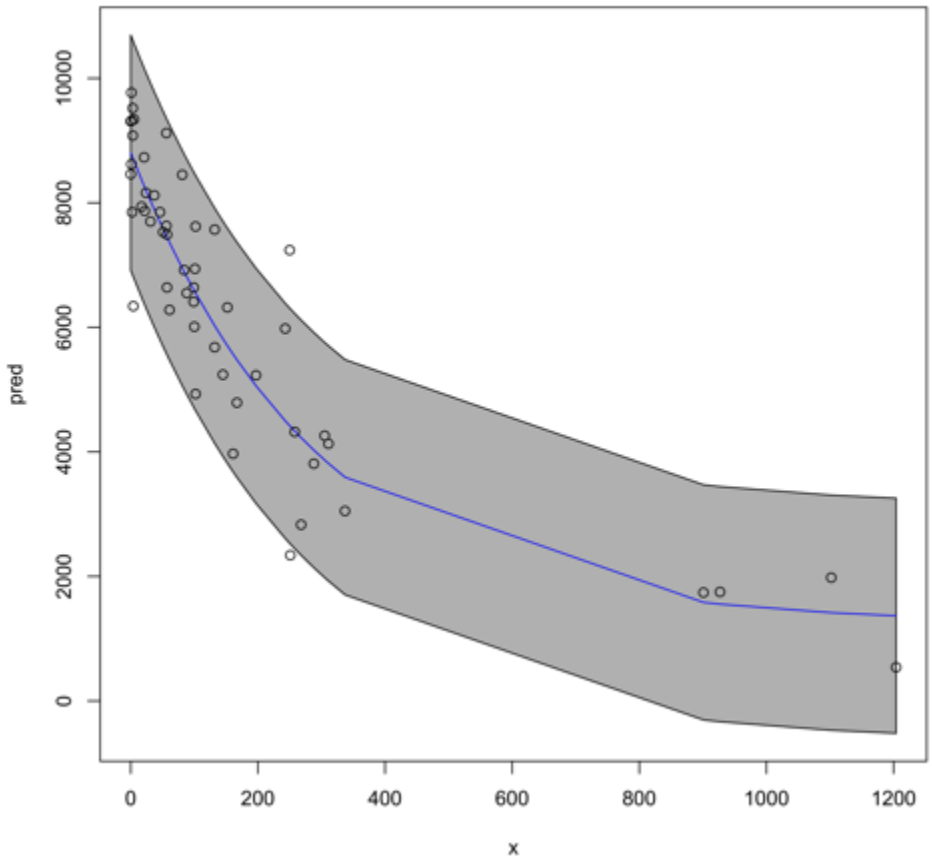


Figure 6.18. Plot of fit to data for the nonlinear regression for yield on windgrass counts where the formula was Yield ~ (a*exp(b*wgrass)) where a = 7000, b = -0.01, and c = 1200. The 95% confidence intervals and included and illustrate variance heterogeneity as windgrass counts increase.

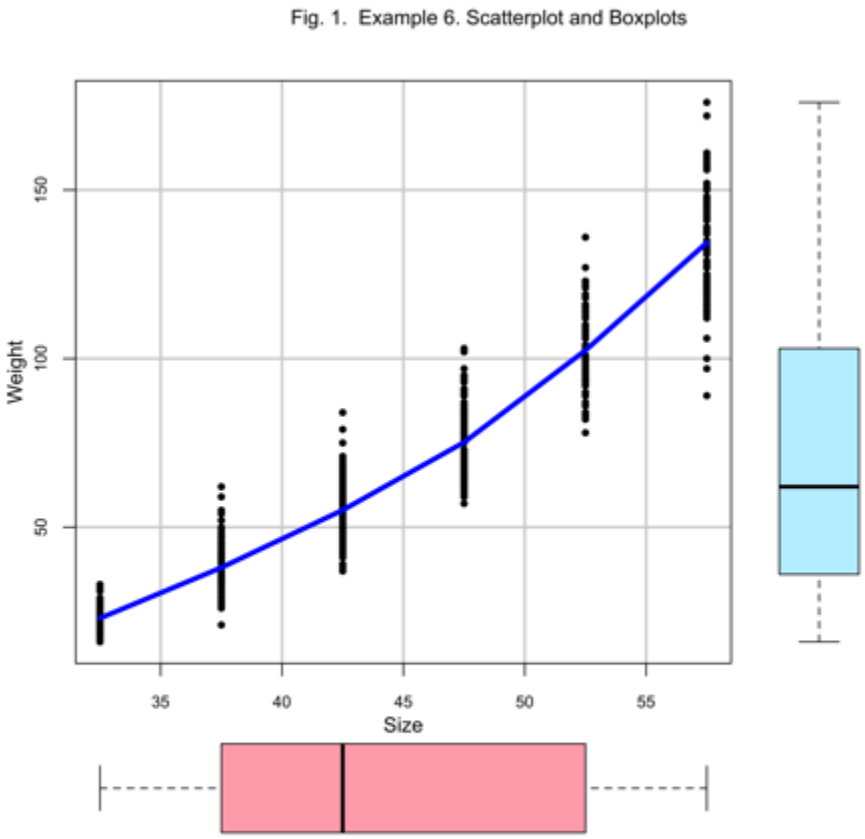


Figure 6.19. Scatter plot of potato weight by size with box plots of the data for the potato dataset in Chapter 6, Example 6. The data are skewed and variance heterogeneity is present.

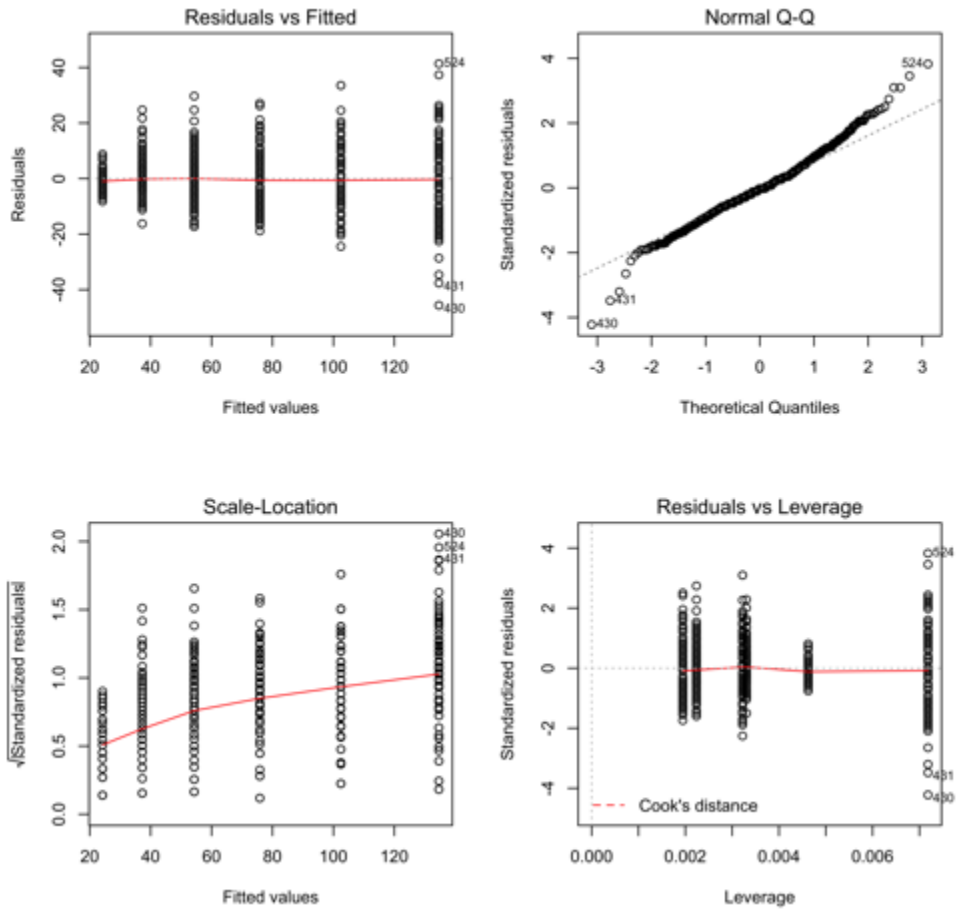


Figure 6.20. Diagnostic plots for the best regression model for the potato data in Chapter 6, Example 6. The best model regressed potato weight on the size3 variable and did not include an intercept. Even though this model was the best based on all of the models tested, the residuals still show some inadequacy in the model, specifically for variance heterogeneity.

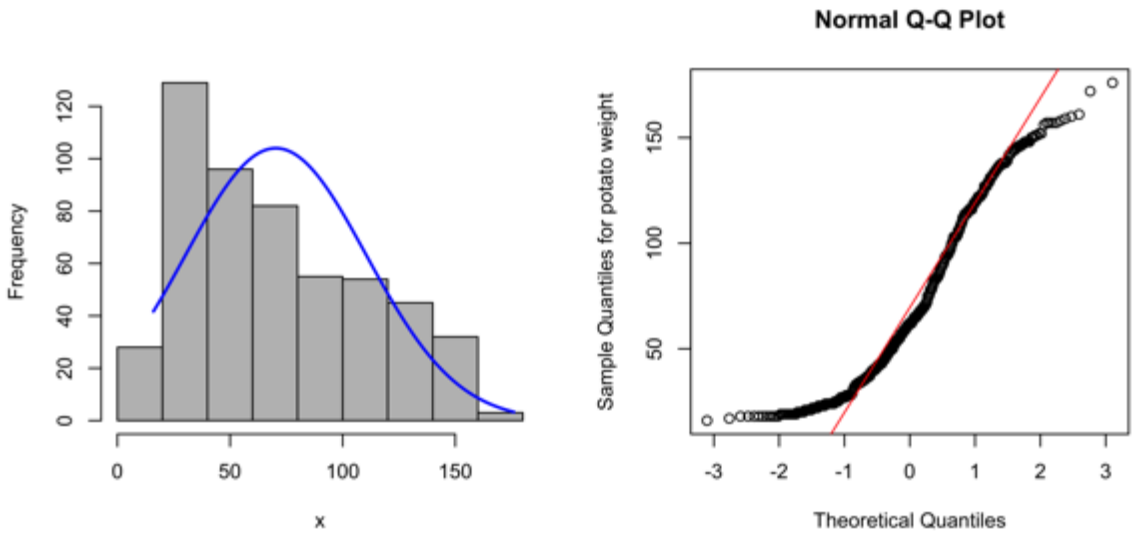
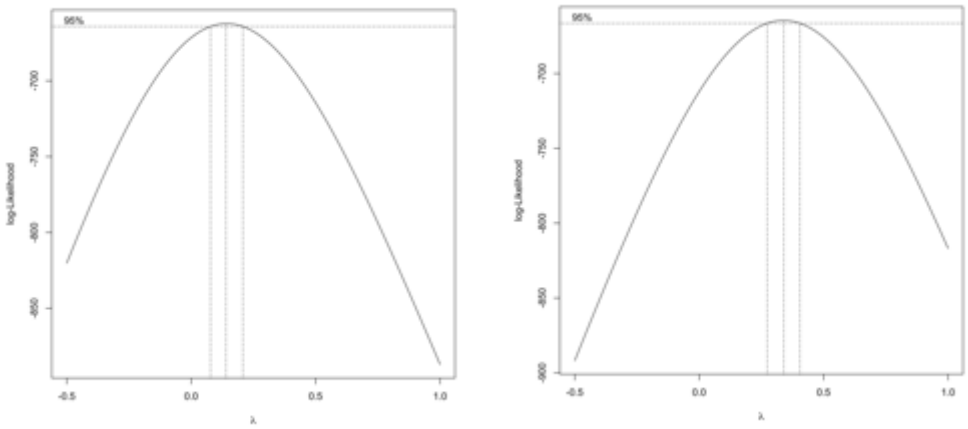


Figure 6.21. The histogram for the residuals and the QQ plot for the untransformed weight data in the potato data set from Chapter 6, Example 6. These plots illustrate the skewed nature of the data.



Figures 6.22 (left) and 6.23 (right). The plots from box cox transformations of the size data. Plot 6.22 is the box cox results when weight is regressed on log(size) while 6.23 is the box cox results when weight is regressed on size. These data are from Chapter 6, Example 6.

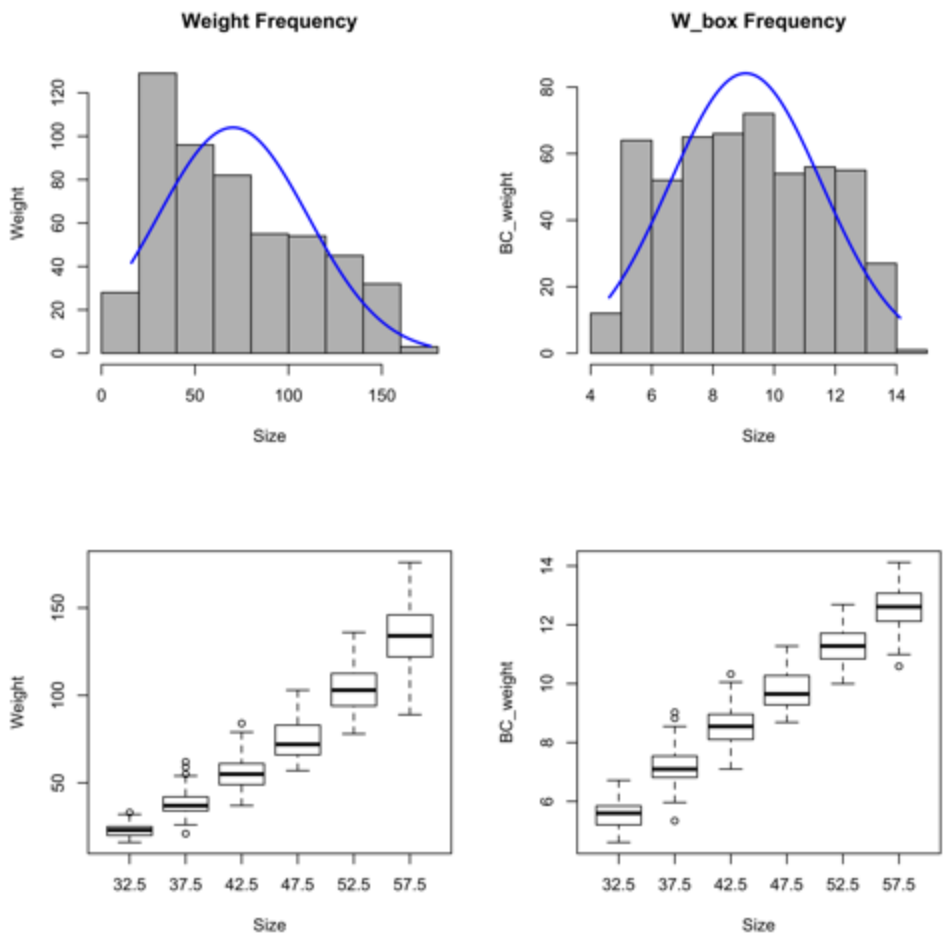


Figure 6.24. Residuals histograms and box plots for the original weights and the box cox transformed weights with lambda = 0.34. The transformation does help to normalize the data and remove some variance heterogeneity. These data are from the potato dataset in Chapter 6, Example 6.

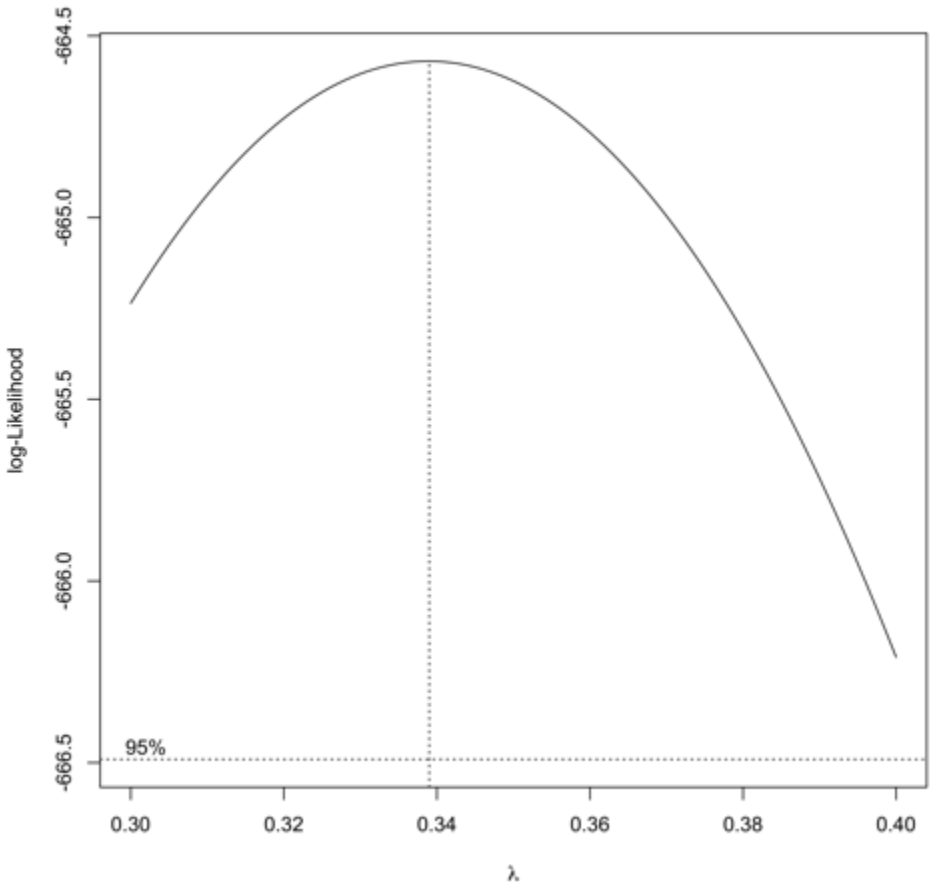


Figure 6.25. The final box cox transformation plot showing that 0.34 has the highest log likelihood. The transformation is for the weight data in the potato dataset from Chapter 6, Example 6.

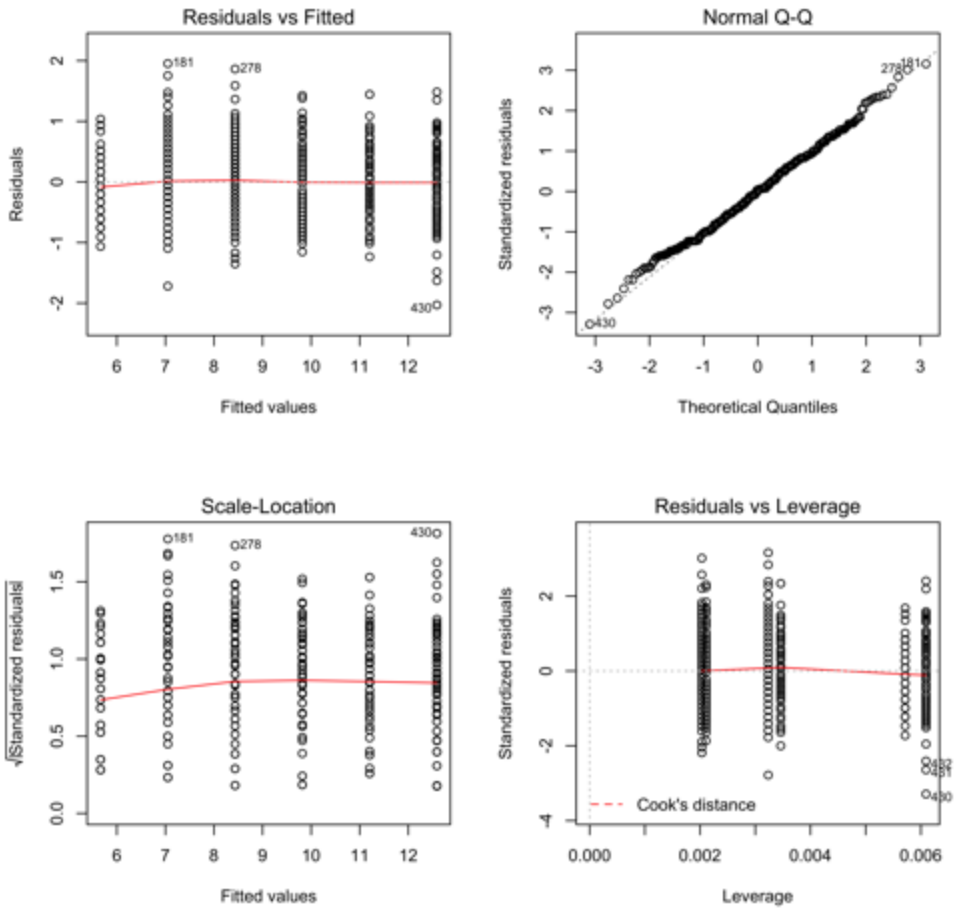


Fig. 6.26. Diagnostic plots from the regression model for the box cox transformed potato weight on size. These plots look better but there are still some problems with the fit of the model. These data are from Chapter 6, Example 6.

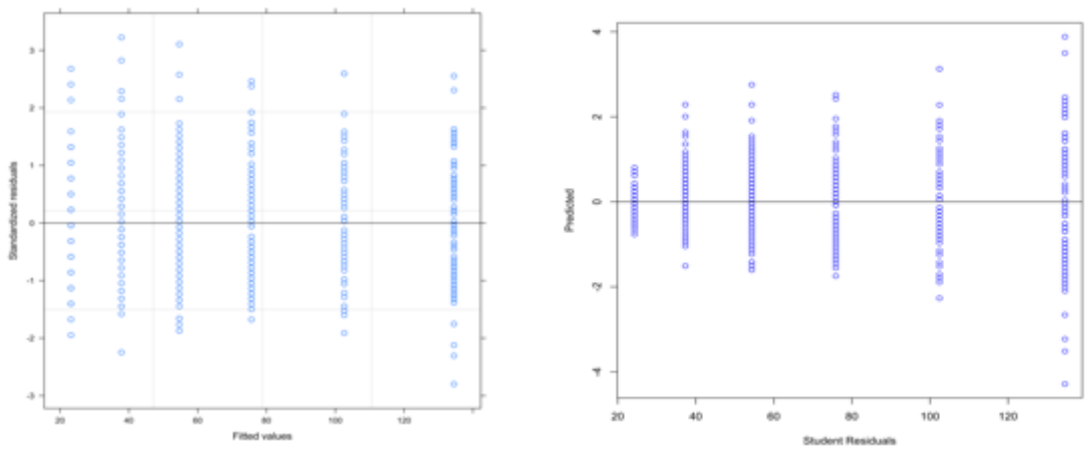


Fig. 6.27a and b. Residuals plots comparing the regression of weight on size, with individual weights (a) with the regression of weight on size 3 with no weighting (b) . The variance heterogeneity is not as great when the weighted analysis is run. These data are from the potato dataset in Chapter 6, Example 6.

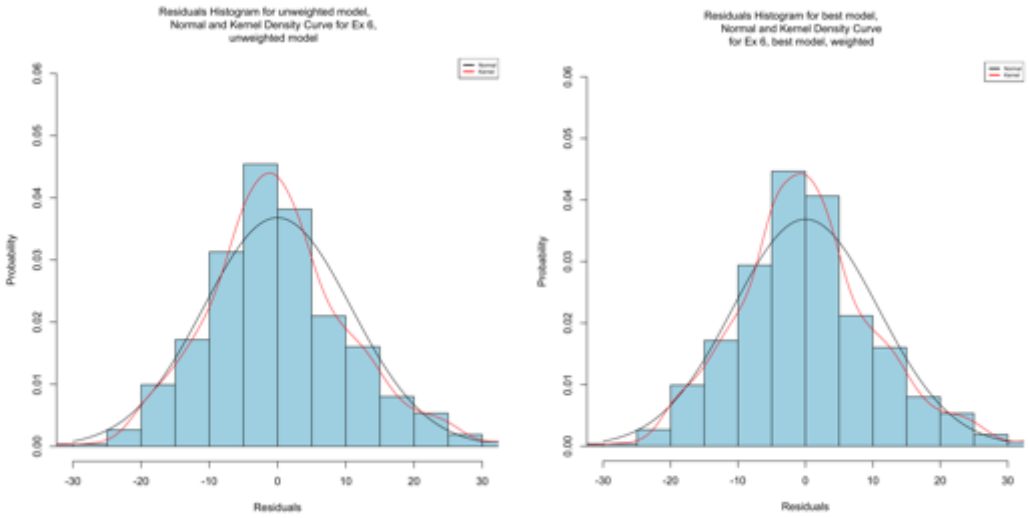


Fig. 6.28 a and b. The residuals plots for the two models are similar but not identical. These data are from Chapter 6, Example 6.

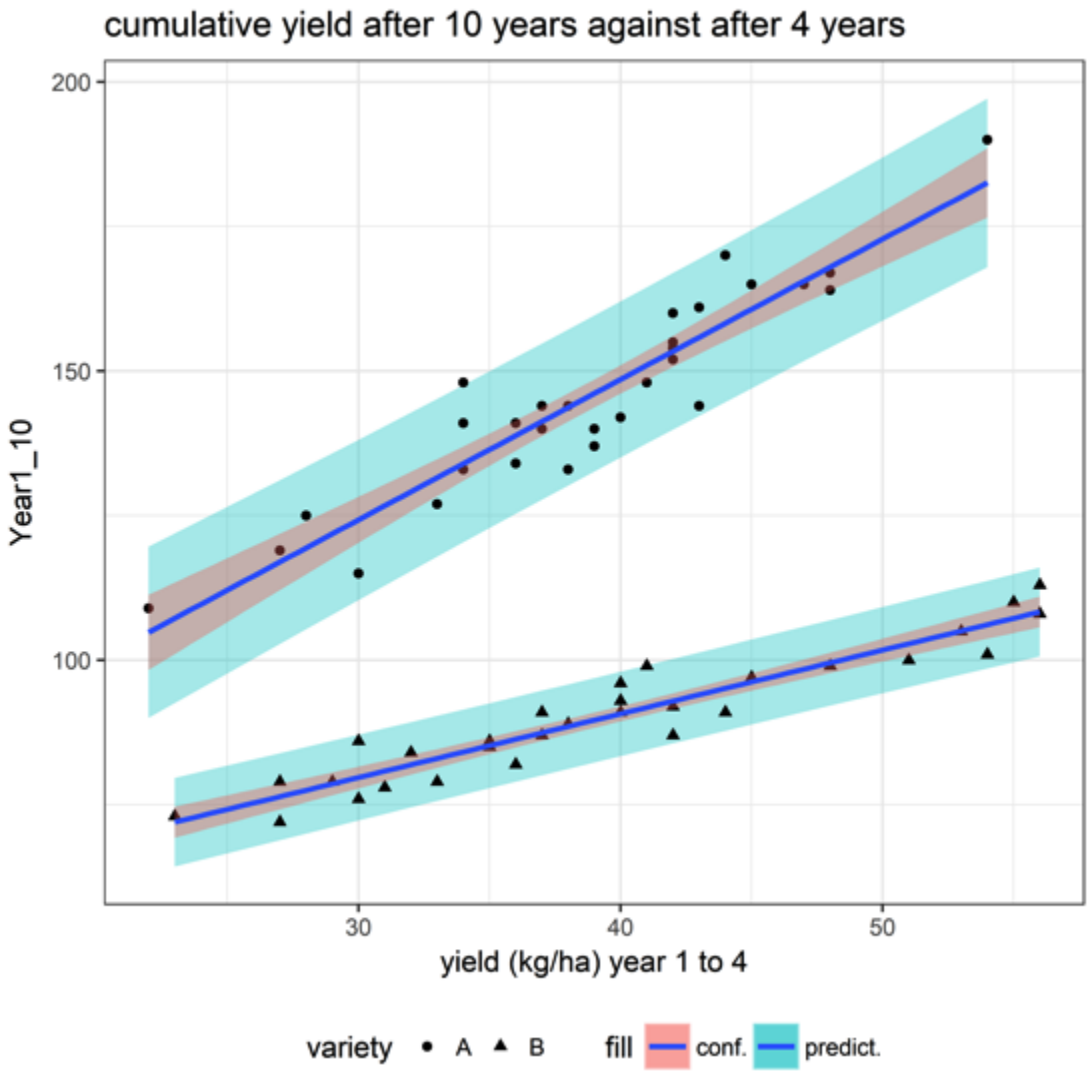


Fig. 6.29. The observed data and separate regression lines with 95% confidence and prediction intervals for the apple data in Example 7 of Chapter 6, which reconstructs Fig. 14 in that chapter.

Fig. 6.30a

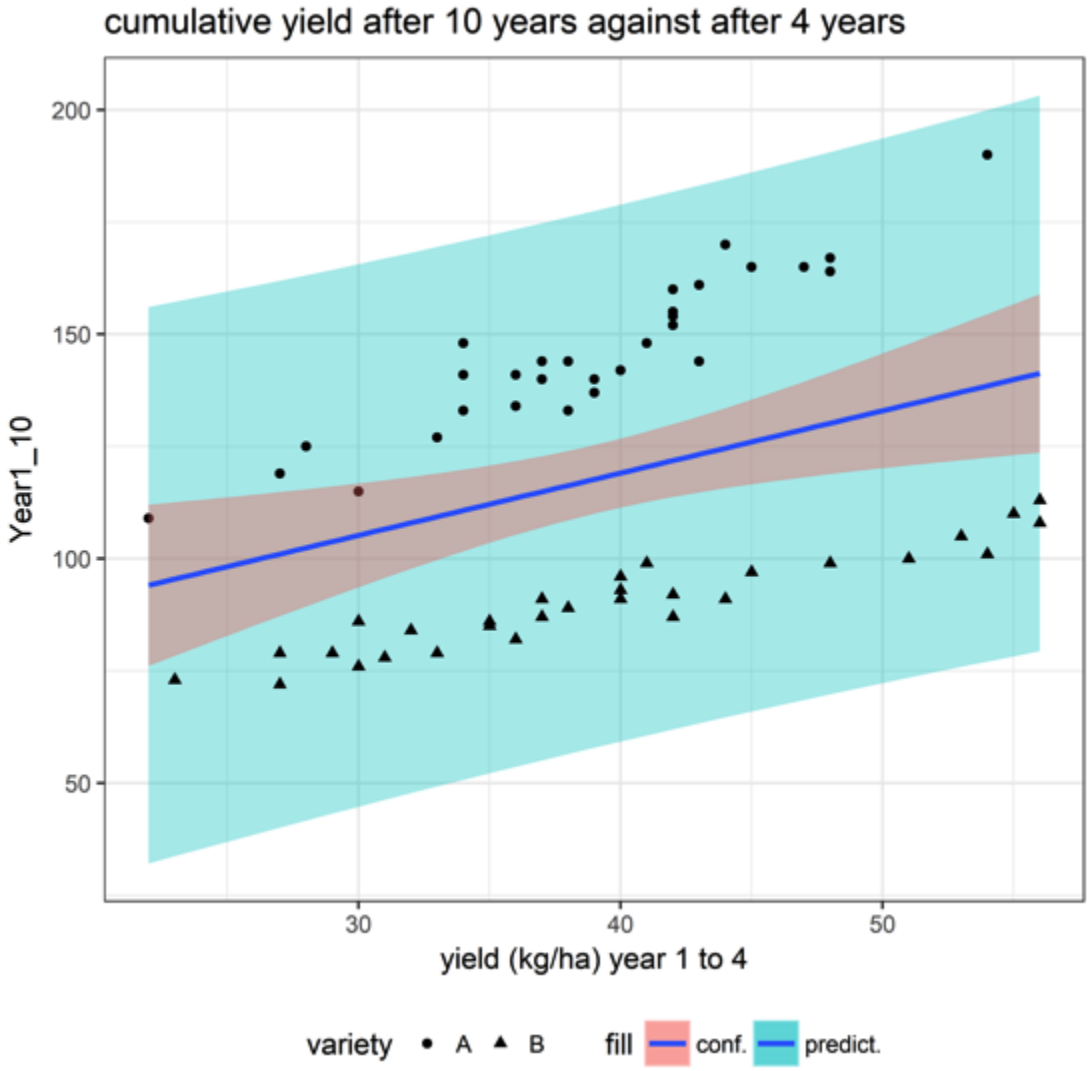


Fig. 6.30a. The observed data and common regression line with 95% confidence and prediction intervals for the apple data in Example 7 of Chapter 6, which reconstructs the upper graph for Approach a in Fig. 15 in that chapter.

Fig. 6.30a

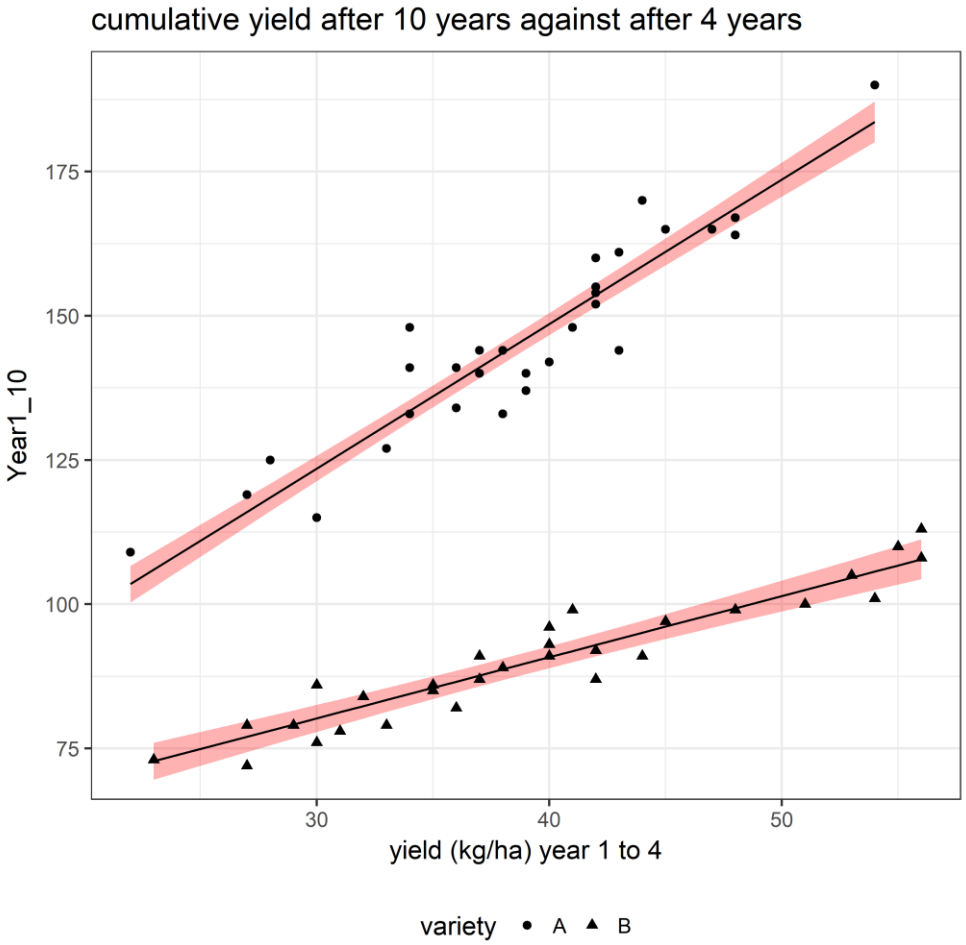


Fig. 6.30a. The observed data and combined analysis regression lines with 95% confidence intervals for the apple data in Example 7 of Chapter 6. This graph reconstructs the lower left graph for Approach c (equal variances) in Fig. 15 in that chapter.

Fig. 6.31

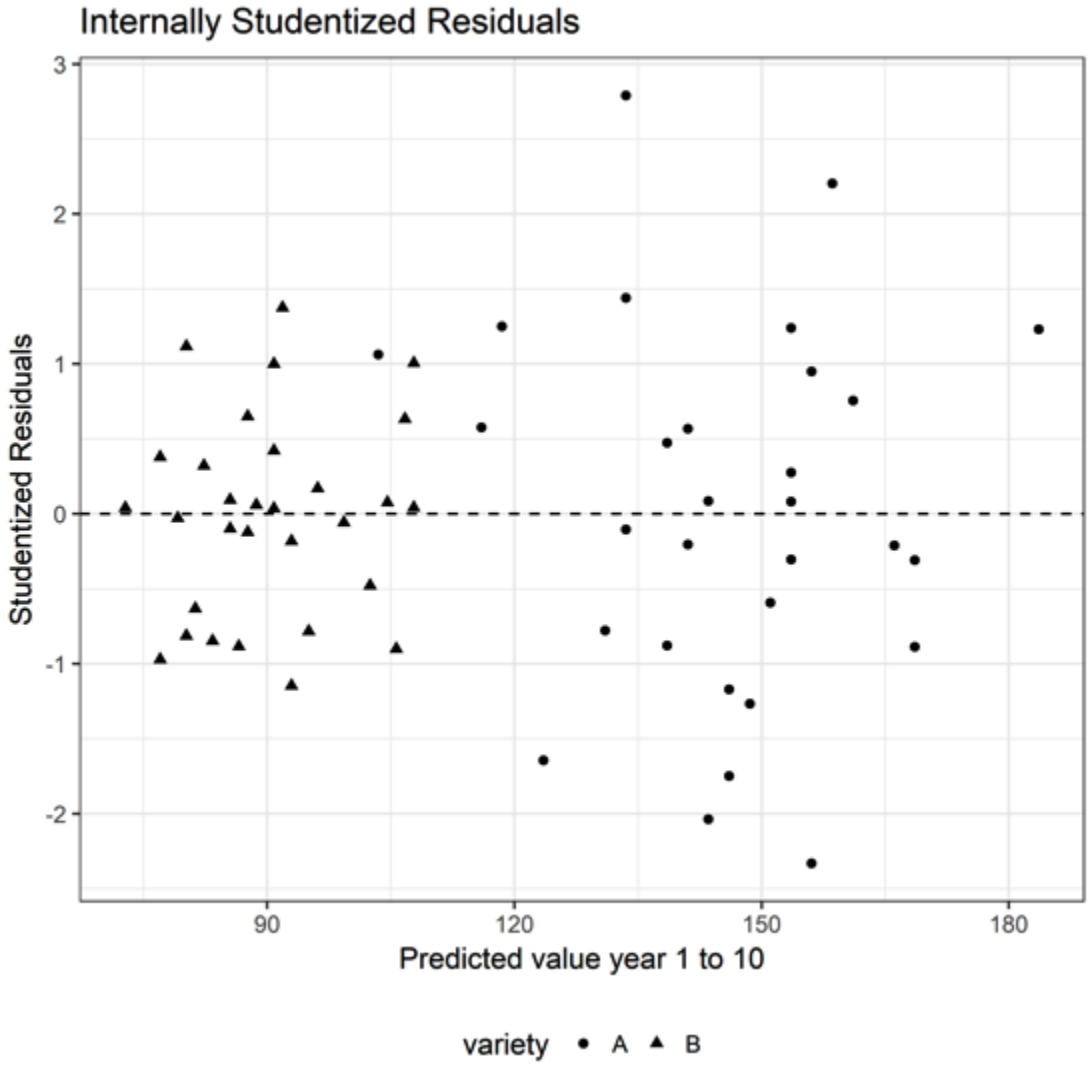


Fig. 6.31. The internally Studentized residuals plotted against the predicted values for Approach c (equal variances) for the apple data in Example 7 of Chapter 6. This graph reconstructs the lower right graph in Fig. 15 in that chapter, except that there isn't any function to calculate the externally Studentized (jackknifed) residuals for a gls object in R. However, there is a function for the internally Studentized residuals, which should not make a big difference when the number of cases per group is large as it is in this example (30 per variety).

Fig. 6.32a

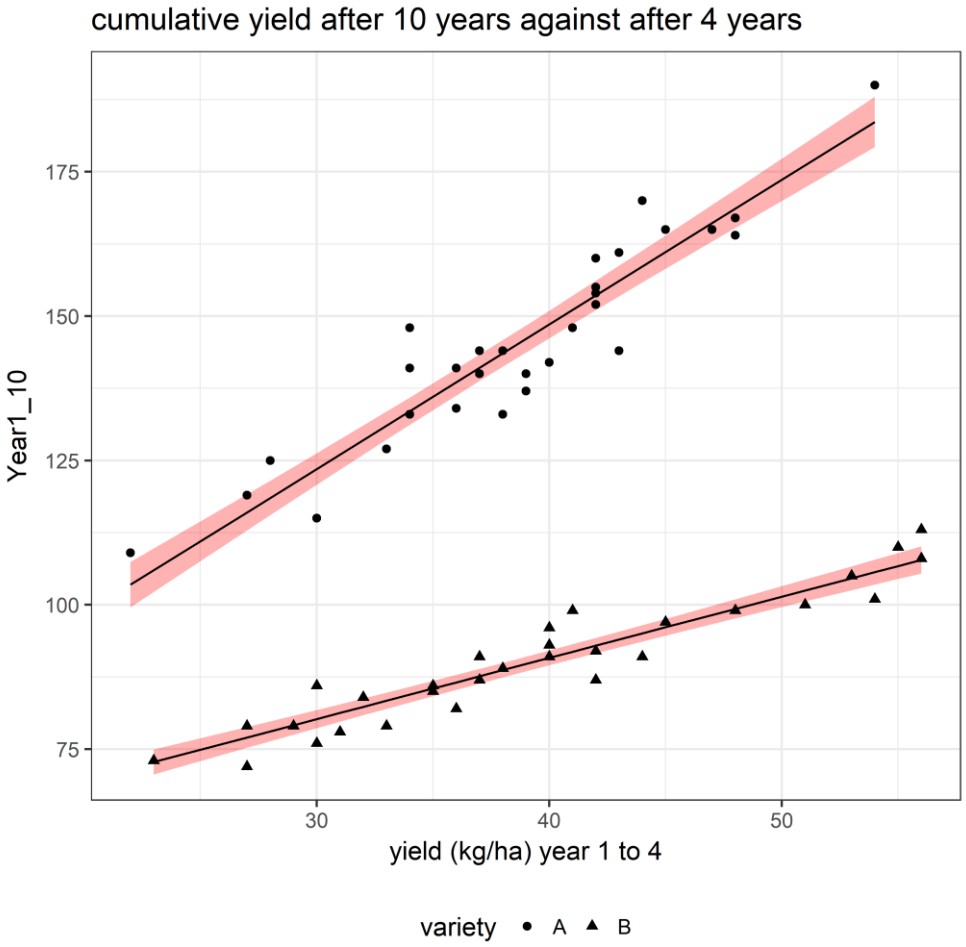


Fig. 6.32a. The observed data and combined analysis regression lines with 95% confidence intervals for the apple data in Example 7 of Chapter 6. This graph reconstructs the lower left graph for Approach c (unequal variances) in Fig. 16 in that chapter.

Fig. 6.32b

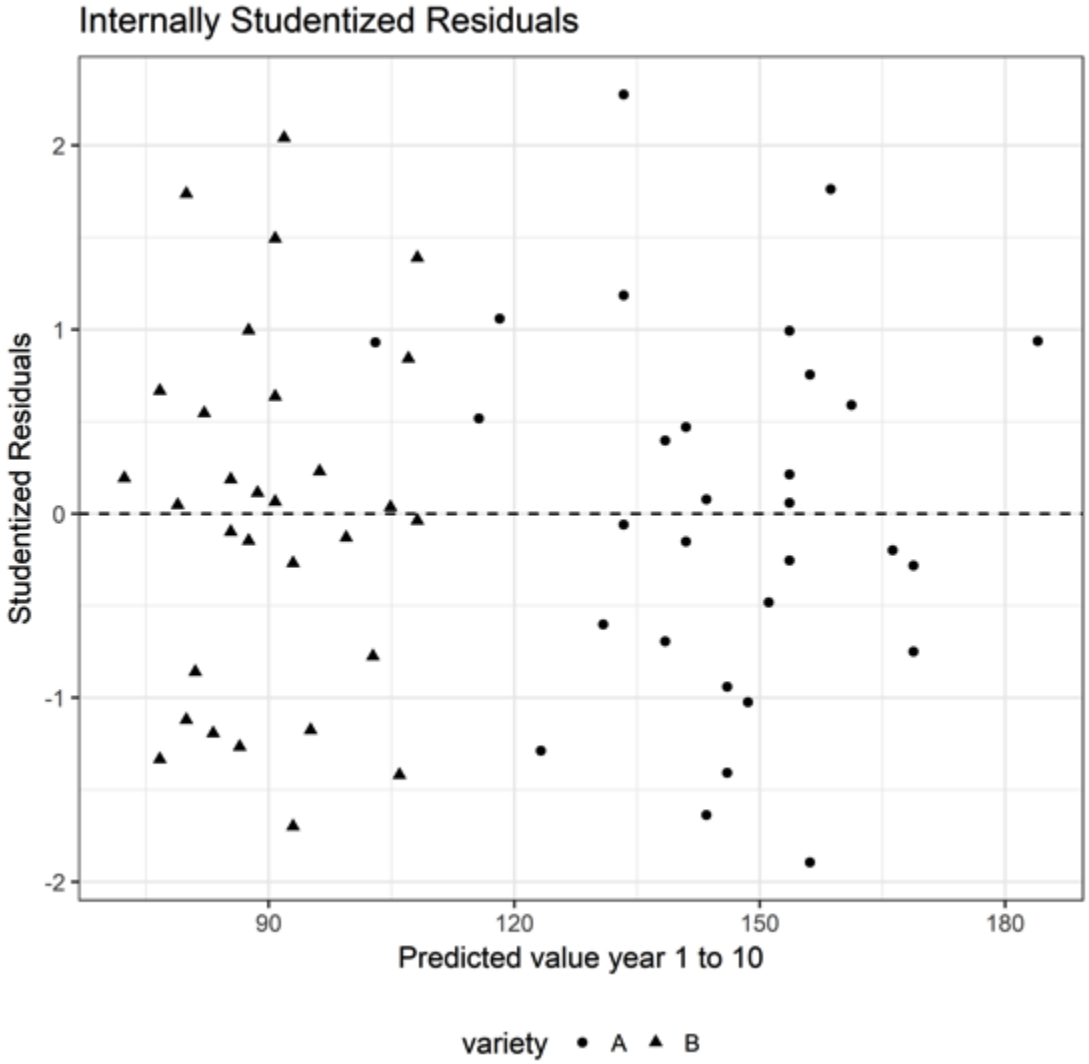


Fig. 6.32b. The internally Studentized residuals plotted against the predicted values for Approach c (unequal variances) for the apple data in Example 7 of Chapter 6. This graph reconstructs the lower right graph for in Fig. 16 in that chapter, except that the internally instead of the externally Studentized (jackknifed) residuals are used. See the caption for Fig. 6.31 for an explanation.

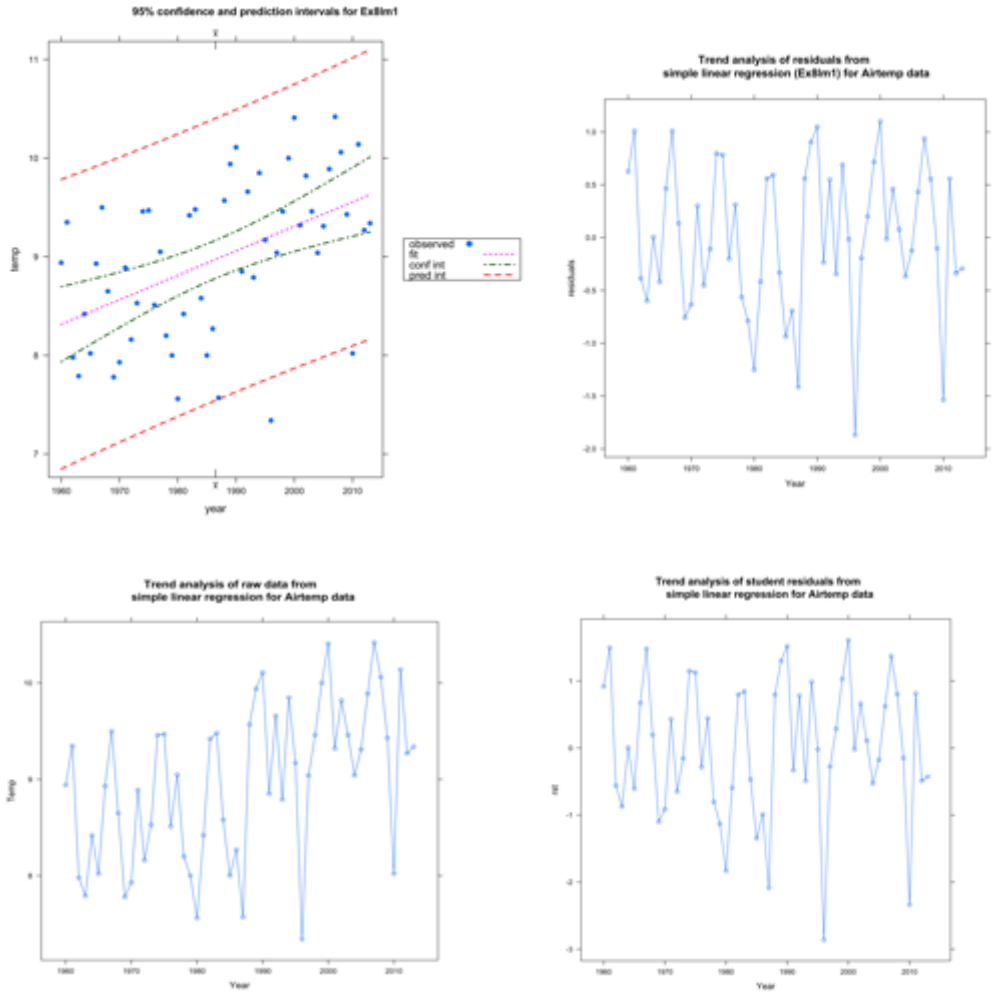


Figure 6.33 a, b, c and d. Data from Example 8 in Chapter 6 are examined. In 6.33a (upper right), a residuals plot is fit to a simple linear regression of air temperature on year. Confidence and prediction intervals are included. In 6.33b, (upper left), the residuals from the model indicate that variance is increasing with time. In 6.33c (lower left), the temperature data is plotted against year and the student residuals are plotted against year in 6.33d. These also show increased variation with time.

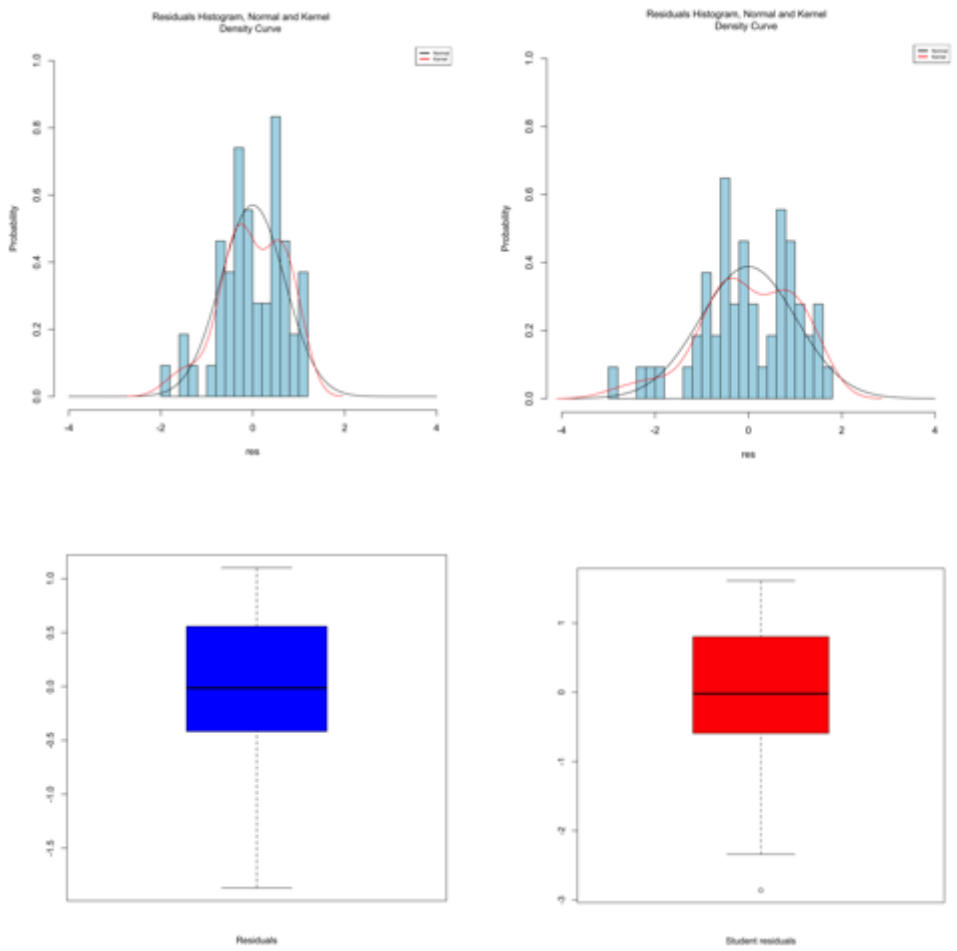


Figure 6.34 a, b, c, and d. Residuals, histograms and boxplots from simple linear regression of the air temperature on year for the airtemp dataset in Chapter 6, Example 8. The residuals are in plots 6.34a and 6.34c (left side) and the studentized residuals are in 6.34b and 6.34d.

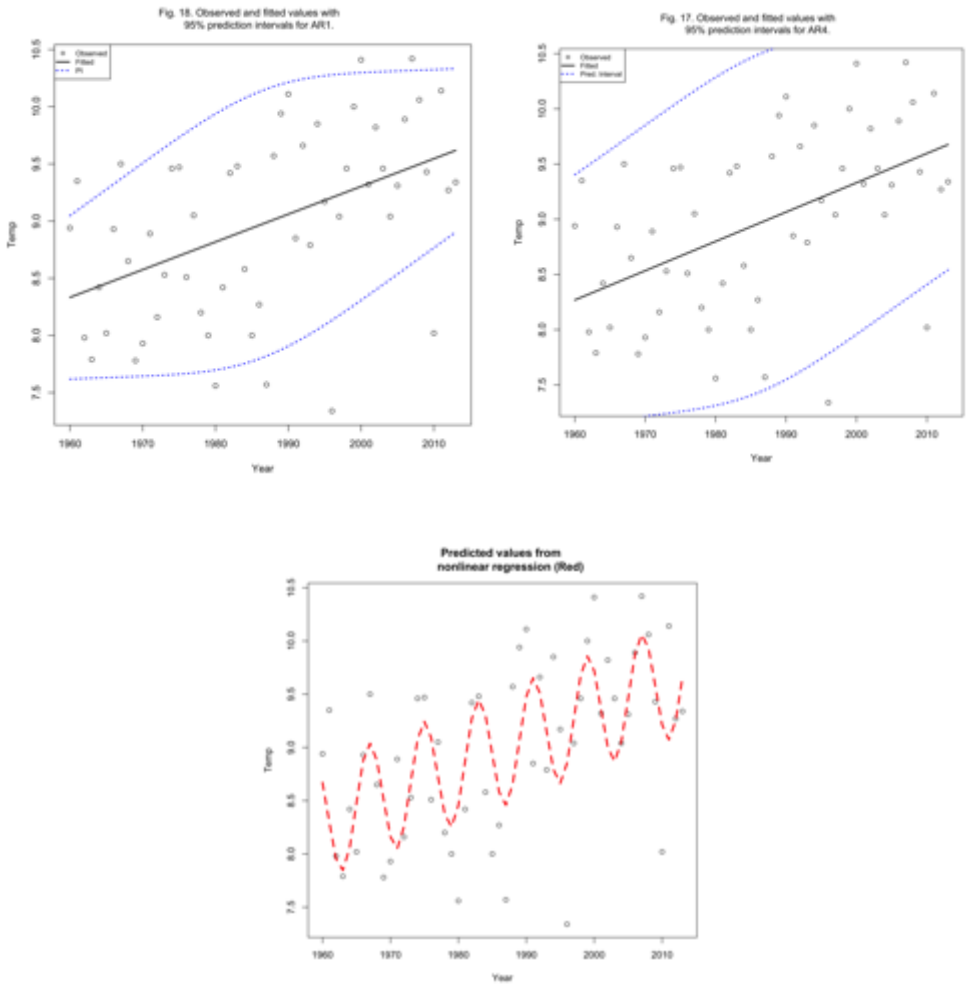


Fig. 6.35a, b, and c. Plots of temperature against year for two autoregressive models, with prediction intervals, (AR1 (a) and AR4 (b). In both cases, some of the data points fall outside of the intervals. These plots match Figures 18 and 17 in Chapter 6. Plot 6.35c is the raw data with the predicted values for the nonlinear regression model marked in red. These data are from the airtemp dataset in Chapter 6, Example 8.

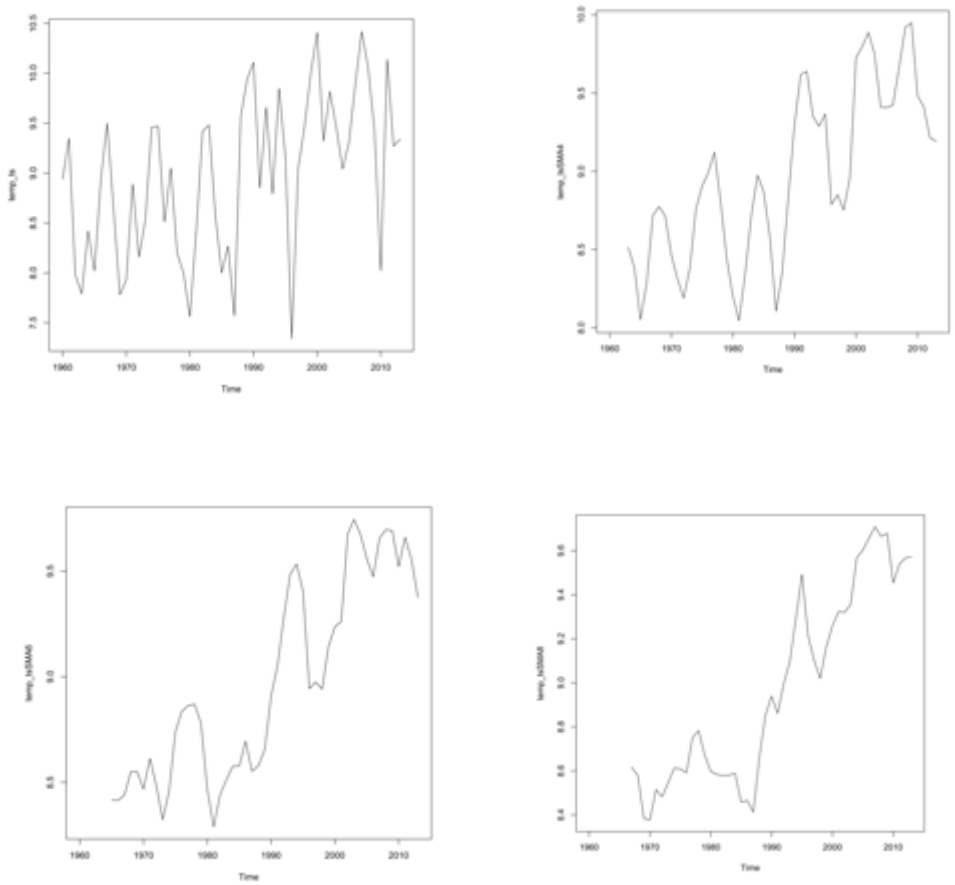


Fig. 6.36 a, b, c, and d. Moving average plots of air temperature against year for 1(a), 4(b), 6(c) and 8(d) years. The trend towards more variation and greater temperatures after 1985 is illustrated. These data are from the air temp dataset in Chapter 6, Example 8.

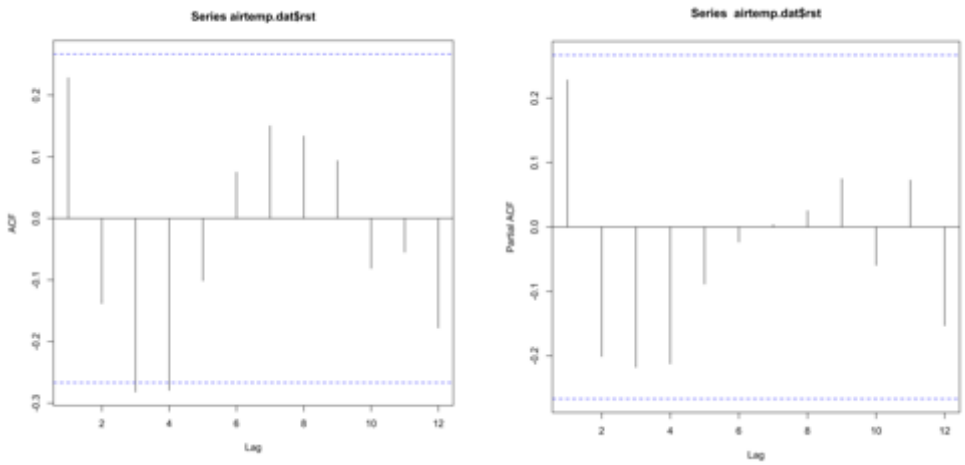


Figure 6.37 a and b. The autocorrelation function (ACF)(a) and the partial autocorrelation function (PACF) (b) plots for the studentized residuals from the simple linear model of air temperature on year indicates that it the autocorrelation alternates sign and the first 4 lags are the largest.

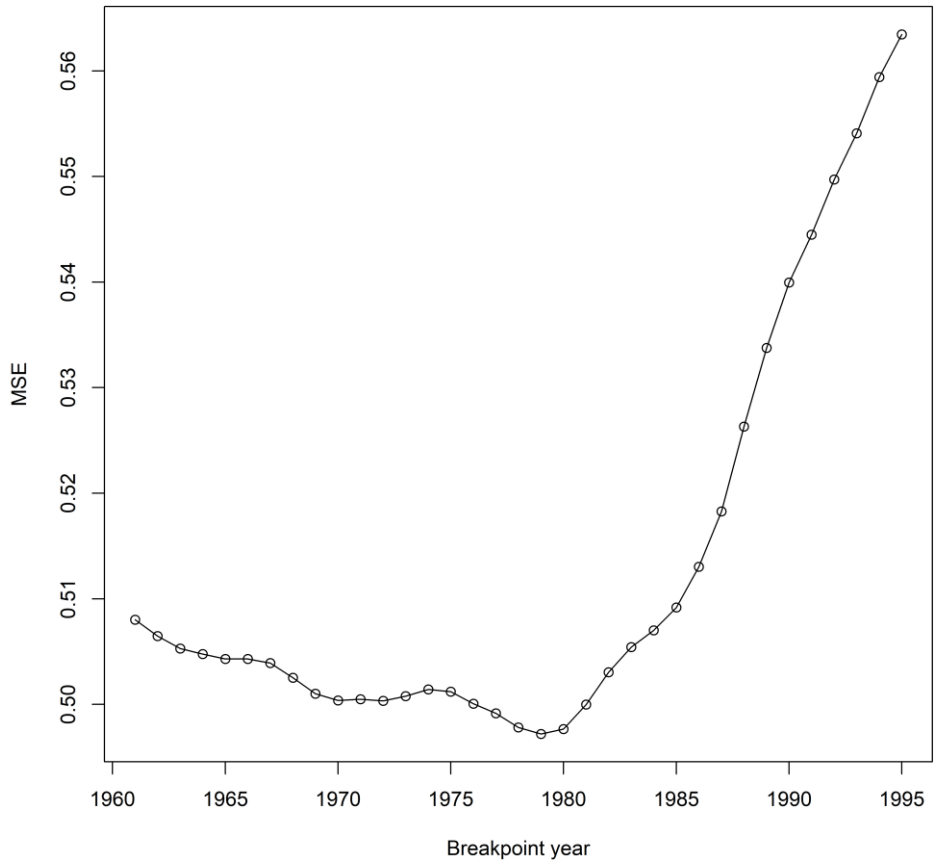


Fig. 6.38. A plot of the MSE against breakpoint year for the airtemp data. The minimum MSE occurred when the breakpoint was 1979. These data are from Chapter 6, Example 8.

Fig. 6.39

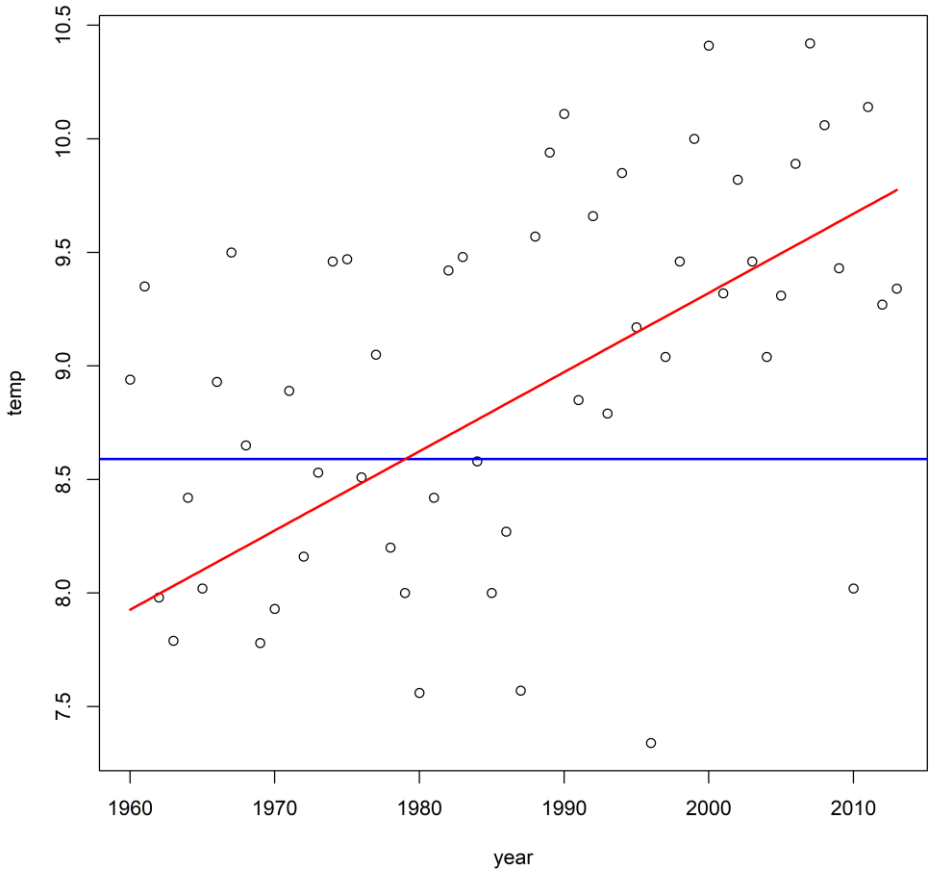


Figure 6.39. A graph of the two line segments identified in the piecewise regression of air temperature on year for the airtemp dataset from Chapter 6, Example 8. The first line (blue) lasts through 1978 and the second (red) from 1979 to the present. The equation for the blue line is temp = 8.59; and for the red line, temp = 0.0349(year – 1979) + 8.59.

Fig. 6.40

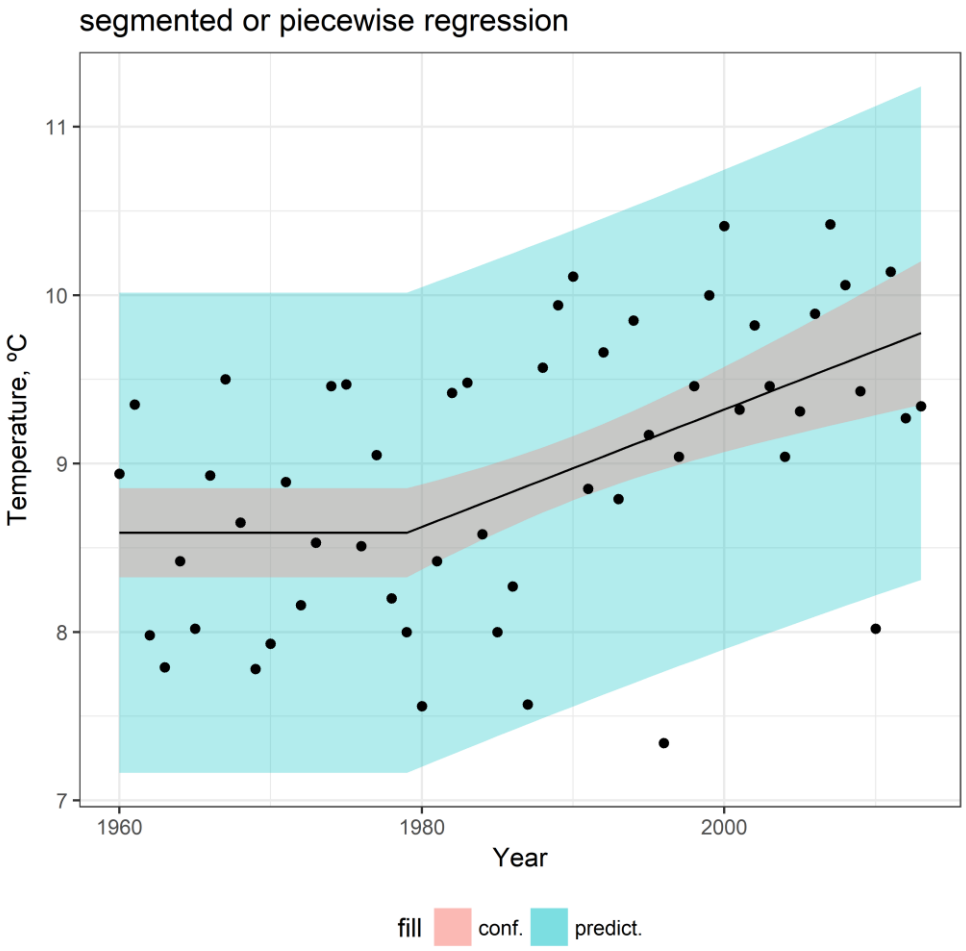


Fig. 6.40. The predicted regression line with confidence and prediction intervals for a piecewise regression of temperature on year. The breakpoint was set to 1979 after the iterative search found that breakpoint to minimize the MSE. These data are from the airtemp dataset in Chapter 6, Example 8. This graph corresponds to Fig. 19 in that chapter and the model used forces the initial line segment to be horizontal and the lines to meet at the breakpoint, which is the same as the authors did for their figure.

Fig. 7.1

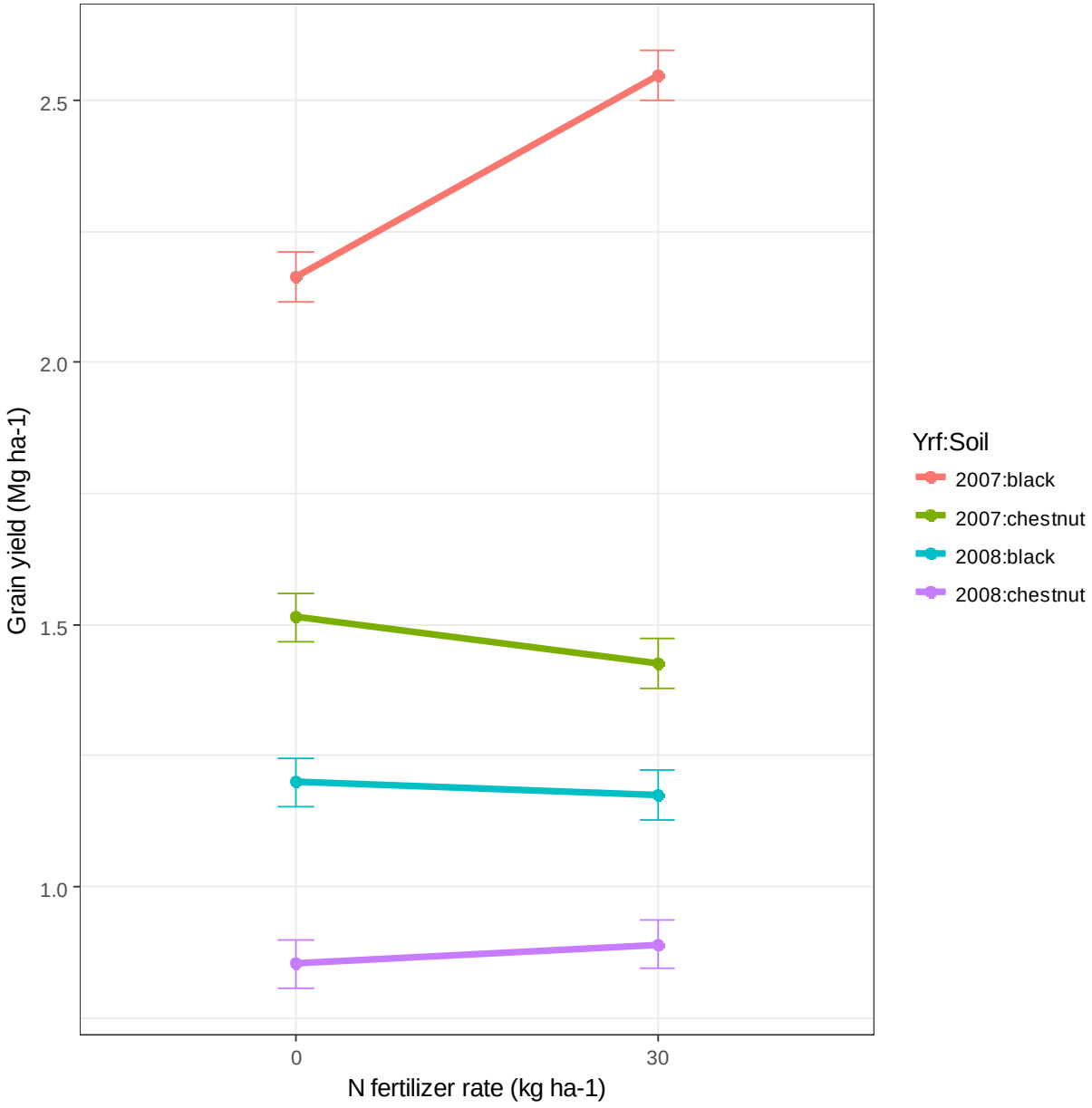


Fig. 7.1. Mean grain yield for Year-by-Soil combinations at two levels of N. This figure does not have all the detail of Fig. 1.1 in Chapter 7, but it contains the main points. It does include half the LSD above and below each mean so whether or not the error bars overlap agrees with the compact letter display of significance. The LSD calculated in R was 0.09419 compared to the 0.0983 listed in Fig. 1.1 of Chapter 7.

Fig. 7.2

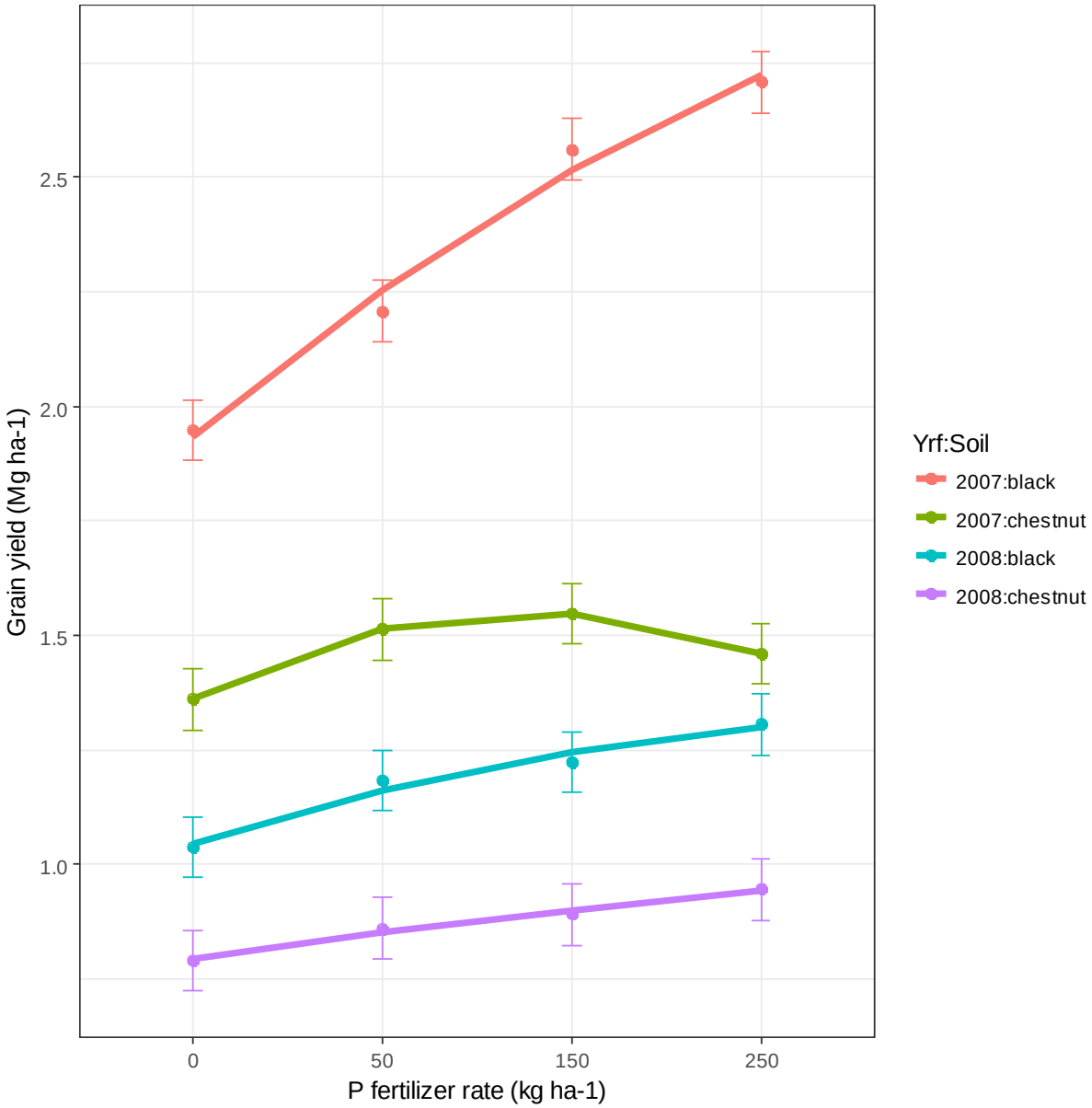


Fig. 7.2. Mean grain yield for Year-by-Soil combinations at four levels of P. This figure does not have all the detail of Fig. 1.2 in Chapter 7, but it contains the main points as in the graphs on Page 13 of App. 2 in Chapter 7. Note that a quadratic fit was applied to all four responses instead of the best fit polynomial. It does include half the LSD above and below each mean so whether or not the error bars overlap agrees with the compact letter display of significance. The LSD calculated in R was 0.1332 compared to the 0.1359 given in Fig. 1.2 of Chapter 7.

Fig. 7.3

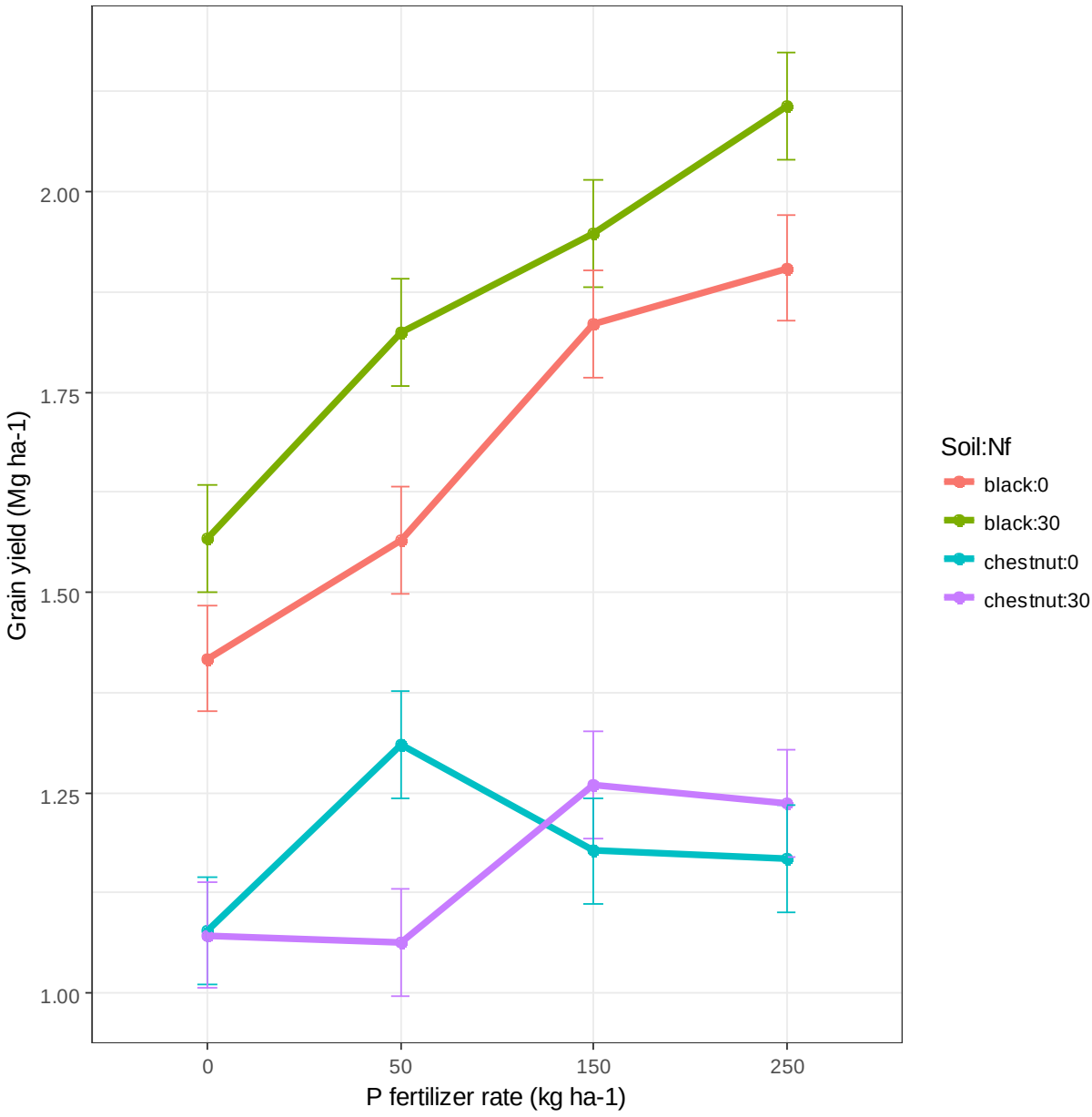


Figure 7.3. Mean grain yield for Soil-by-N combinations at four levels of P. This figure does not have all the detail of Fig. 1.3 in Chapter 7, but it contains the main points as in the graph on Page 18 of App. 2 in Chapter 7. Note that a cubic fit was applied to all four responses. It does include half the LSD above and below each mean so whether or not the error bars overlap agrees with the compact letter display of significance. The LSD calculated in R was 0.1332 compared to the 0.1341 given in Fig. 1.2 of Chapter 7

Fig. 9.1

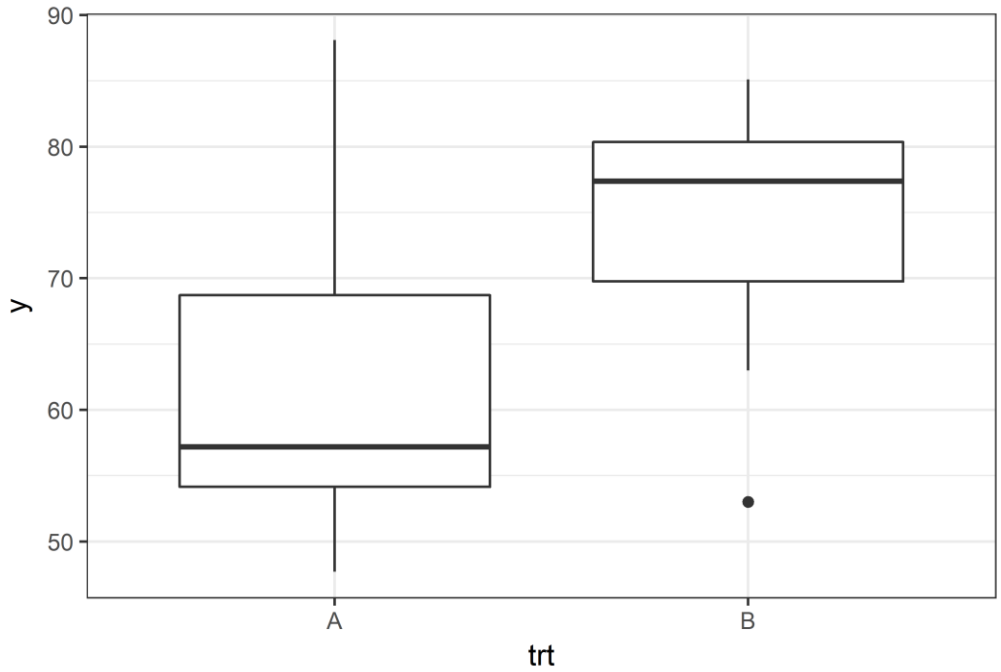


Fig. 9.1. This box and whisker plot does not exactly replicate Fig. 3 of Chapter 9 in that the boxes are narrower and the lower whisker for Trt B is shorter. The latter allows the minimum value to be identified as a possible outlier. These disparities are due to SAS and R using different algorithms for determining quantiles, which can result in sizable differences for small samples such as this (use the R help for the quantile() function for more information on the nine options for quantiles).

Fig. 9.2

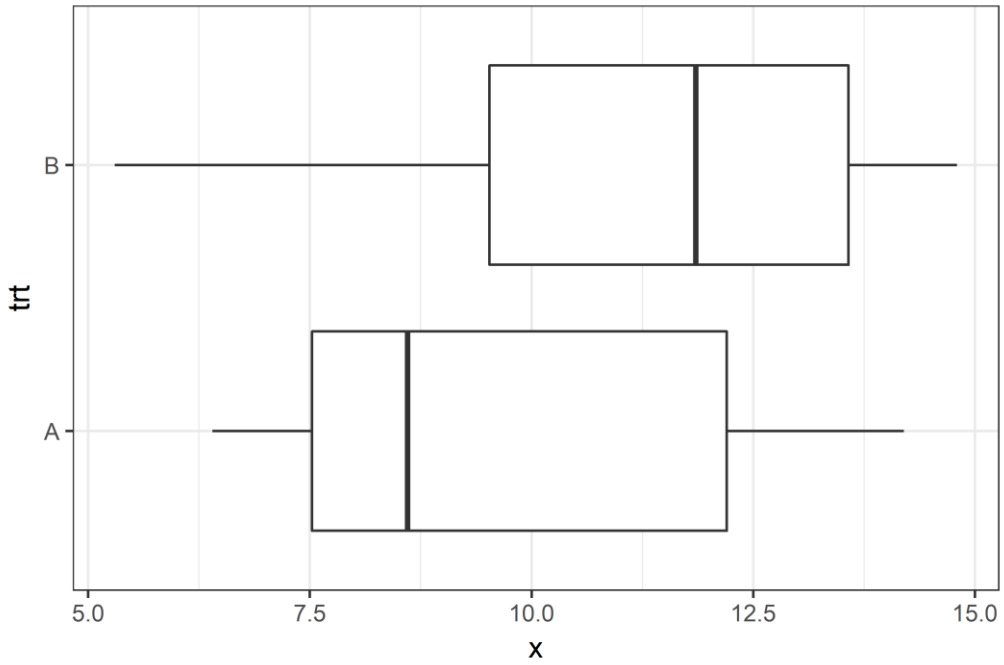


Fig. 9.2. This box and whisker plot varies slightly from Fig. 4 of Ch. 9 for the same reason Fig. 9.1 differs from Fig. 3 of Chapter 9.

Fig. 9.3

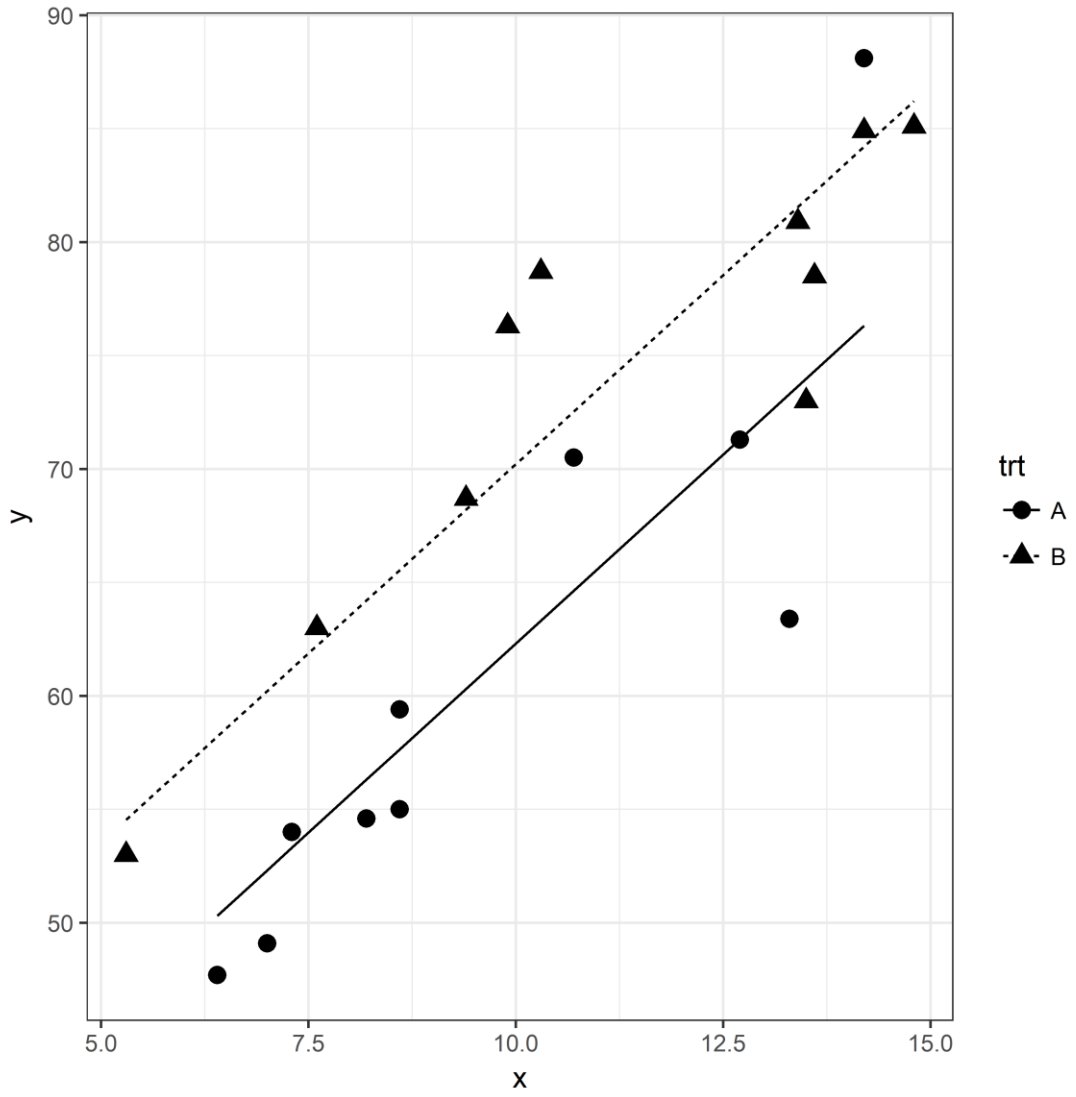


Fig. 9.3. This figure reliably follows Fig. 6 of Chapter 9, in which the ANCOVA quantifies the linear response of y to x for Treatments A and B.

Fig. 9.4

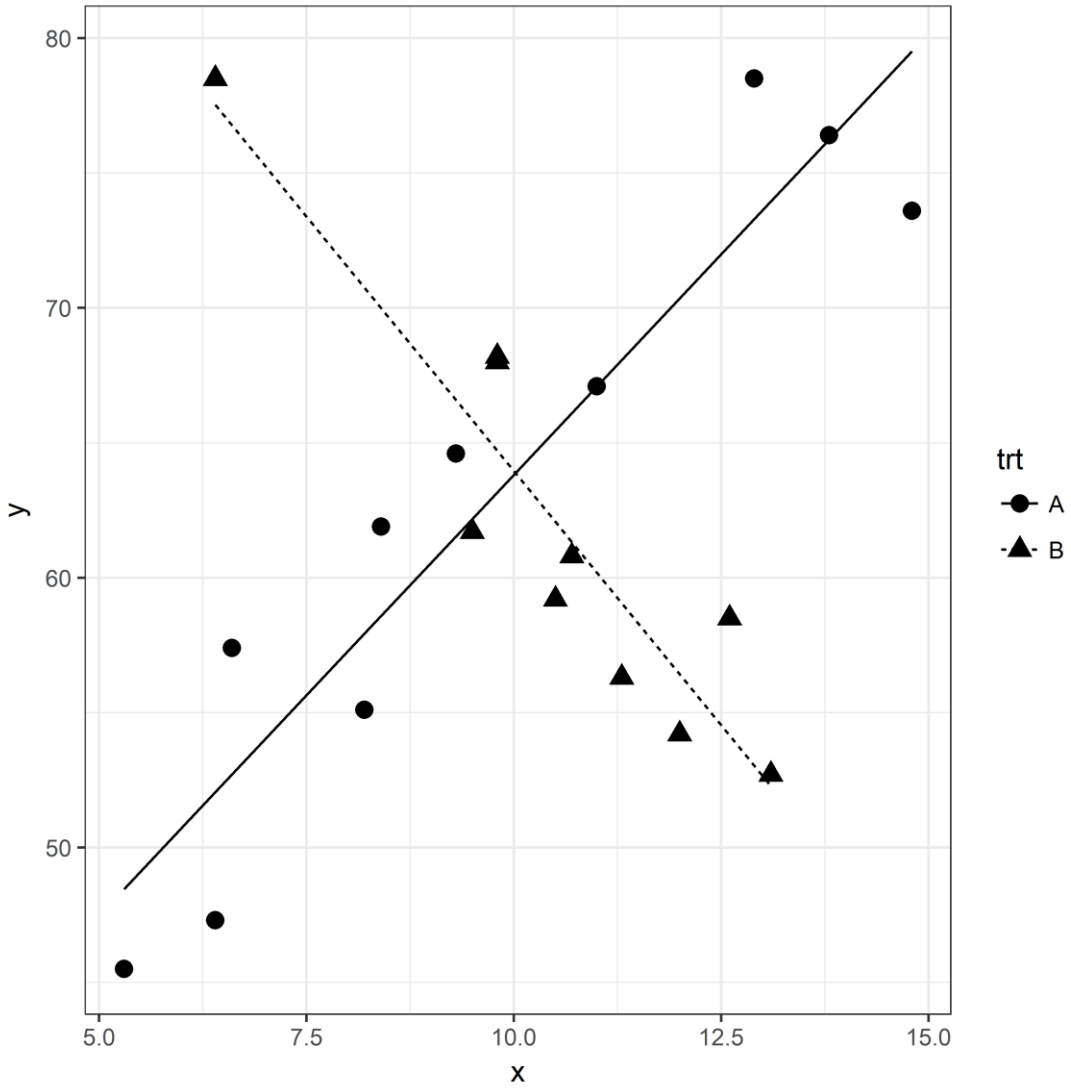


Fig. 9.4. This figure reliably emulates Fig. 18 of Chapter 9 showing the large difference in slopes for Treatments A and B.

Fig. 14.1

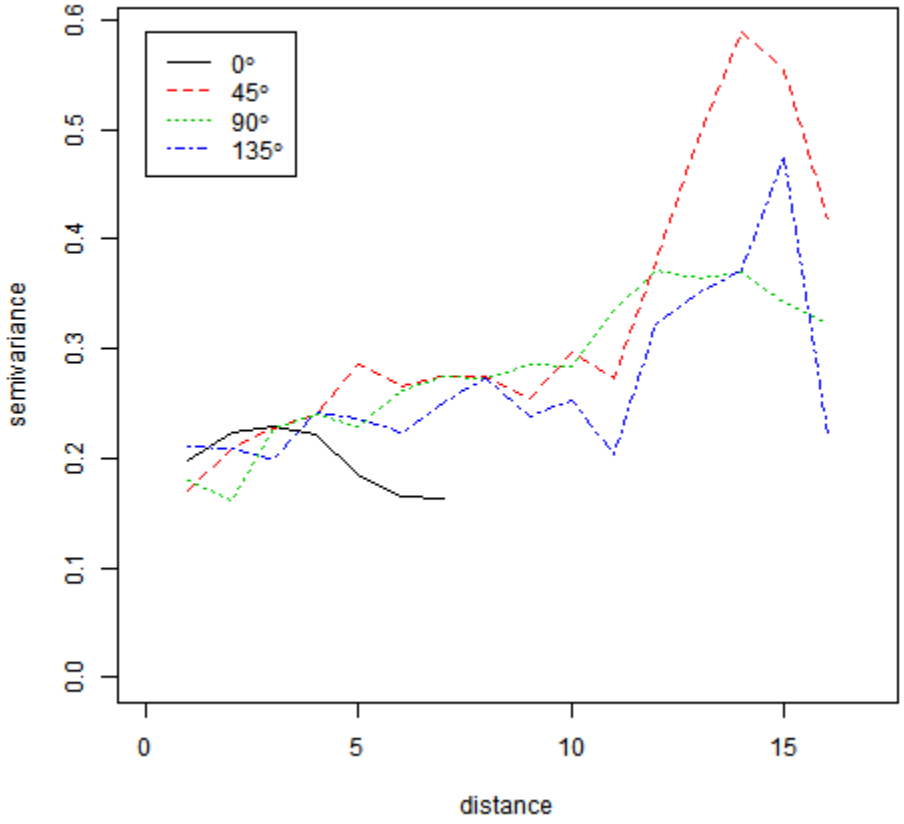


Fig. 14.1. This graph combines four of the panels in Fig. 1 in Chapter 14 into one. Note the relatively large increase in variance for points 10 or more units apart suggesting the residuals are spatially correlated. The different patterns for the different angles indicate the correlations are not isotropic, but the departures are not substantial.

Fig. 14.2

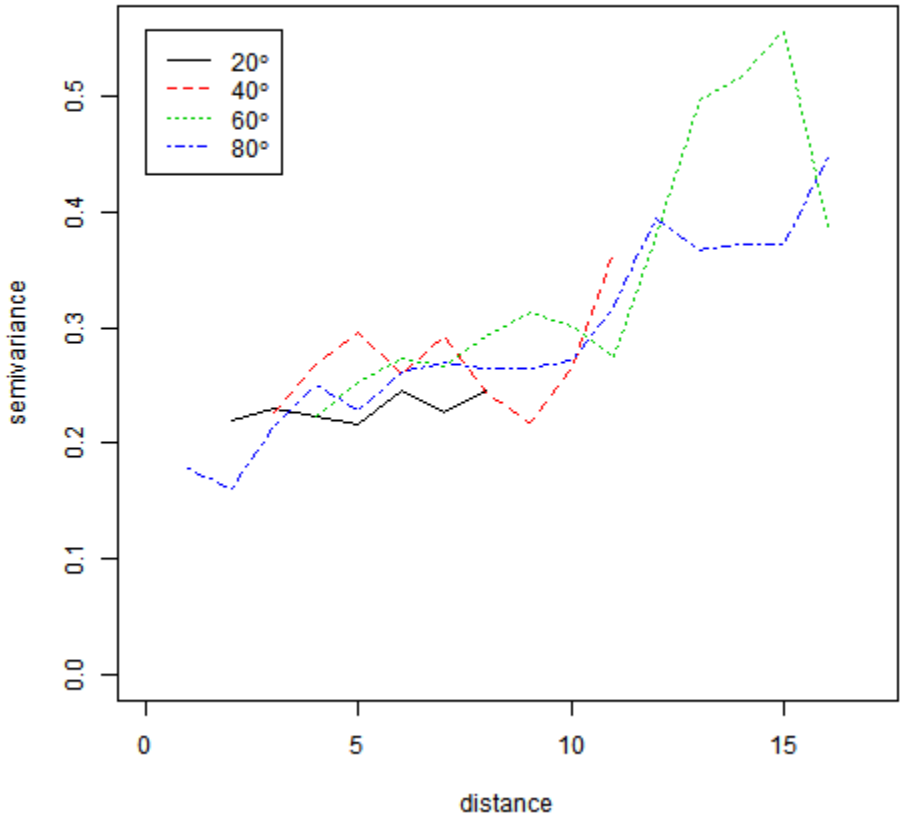


Fig. 14.2. This figure narrows the range of angles to show a little of the clustering that Fig. 2 in Chapter 14 shows; specifically, there are areas of similarity and areas of substantial differences in variance.

Fig. 14.3

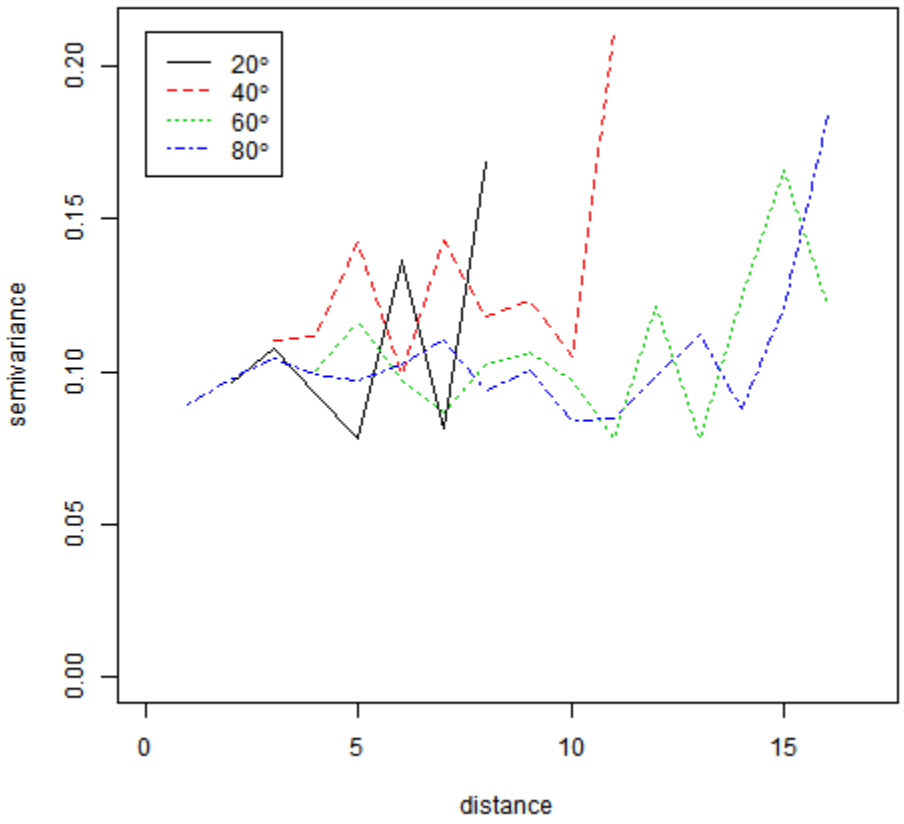


Fig. 14.3. The narrow angle variograms for the residuals from the quadratic row-column model without a spatial covariance structure show reasonable control of the spatial correlation, except at the outer edges.

Fig. 14.4

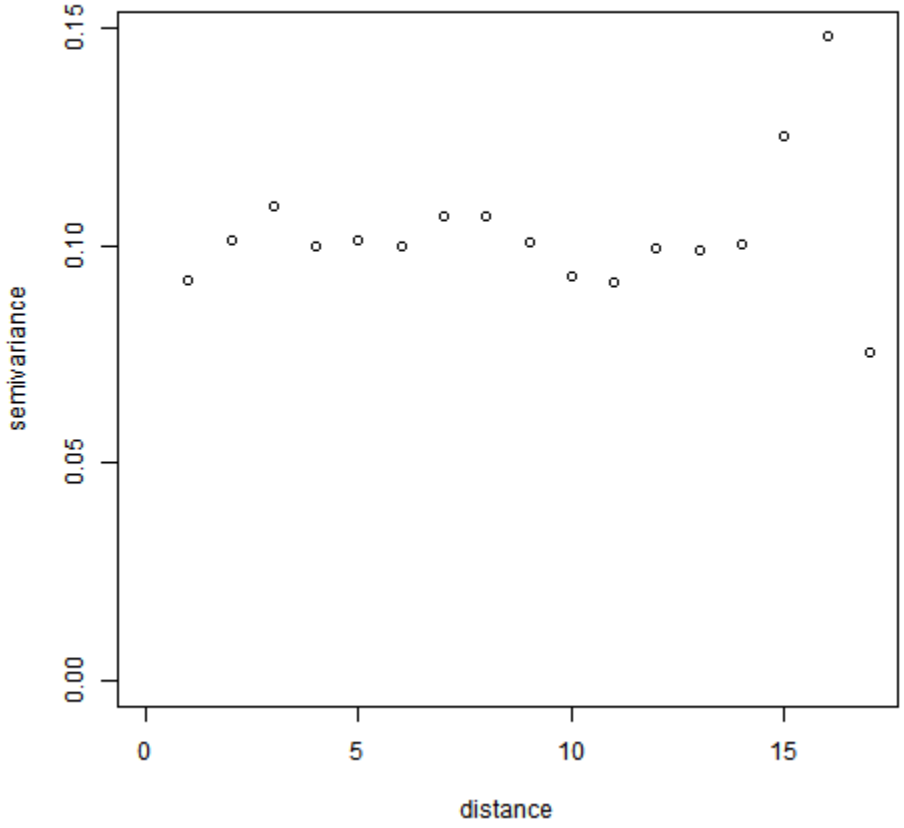


Fig. 14.4. The isotropic (or omnidirectional) variogram for the residuals from the quadratic row-column model without a spatial covariance structure also displays reasonable control of the spatial correlation, except at the outer edges. These results are similar to those illustrated in Fig. 4 and 5 in Chapter 14.

Fig. 14.5

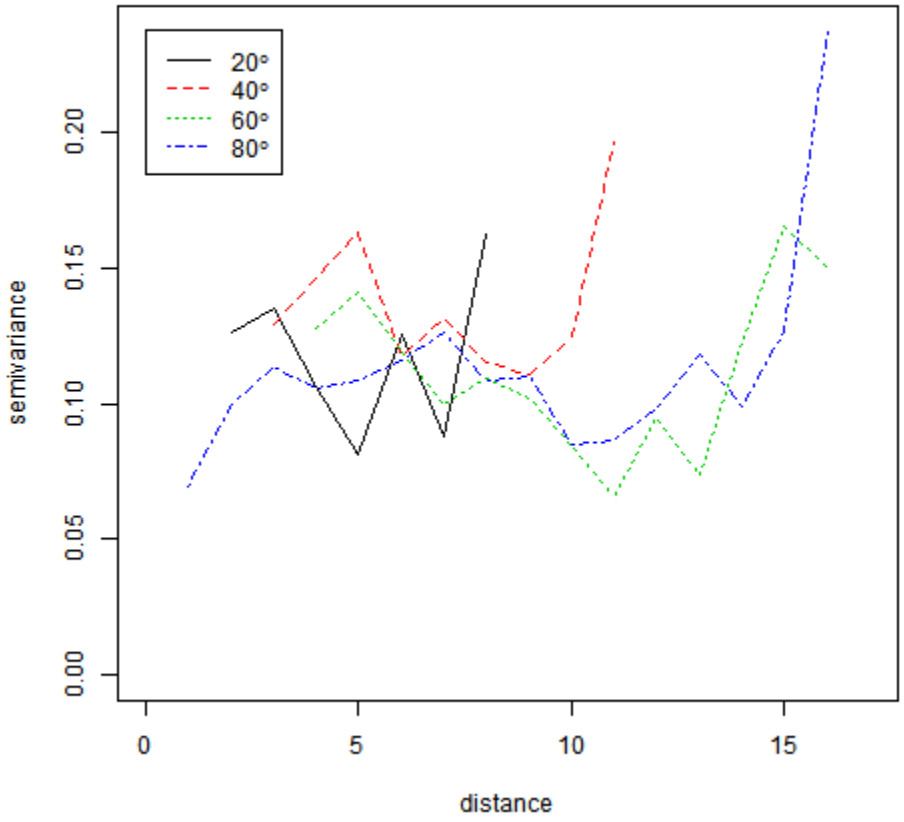


Fig. 14.5. Except for the longest distance at 80º, the quadratic row-column model plus an exponential spatial covariance structure showed excellent reduction in the observable spatial correlation; but it did not appear to be superior to the quadratic row-column model without a covariance structure.

Fig. 14.6

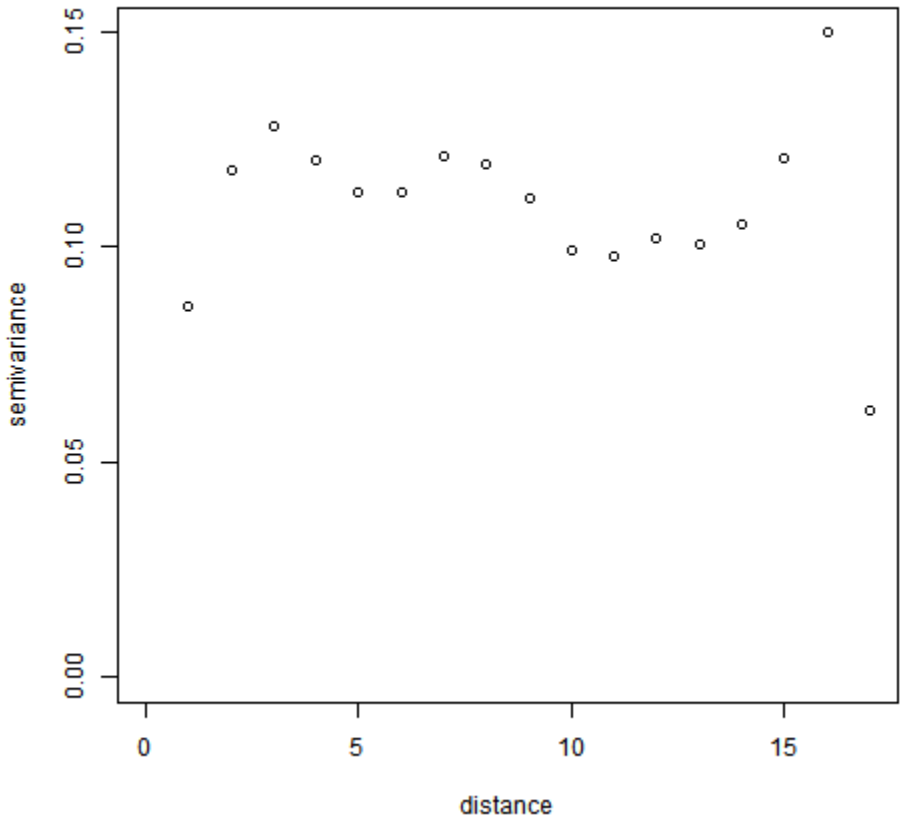


Fig. 14.6. The isotropic (or omnidirectional) variogram for the residuals from the quadratic row-column model plus a spatial covariance structure also showed a substantial reduction in the observable spatial correlation. However, it was not visibly better than the quadratic row-column model without a covariance structure.