ALSM: Chapter 2

Diagnostics and Remedial Measures

Darshan Patel

5/19/2020

```
library(tidyverse)
library(latex2exp)
library(gridExtra)
library(wesanderson)
```

Problem 1:

Distinguish between (1) residual and semistudentized residual, (2) $\mathrm{E}[\varepsilon_i] = 0$ and $\overline{e} = 0$ and (3) error term and residual.

Answer:

Problem 2:

Prepare a prototype residual plot for each of the following cases: (1) error variance decreases with X; (2) true regression function is $\bigcup$ shaped, but a linear regression function is fitted.

Answer:

Problem 3:

Refer to Grade point average Problem 1.19.

(a) Prepare a box plot for the ACT scores X_i. Are there any noteworthy features in this plot?

Answer:

(b) Prepare a dot plot of the residuals. What information does this plot provide?

Answer:

(c) Plot the residual e_i against the fitted values $\hat{Y}_i$. What departures from regression model (2.1) can be studied from this plot? What are your findings?

Answer:

(d) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlationbetween the ordered residuals and their expected values under normality. Test the reasonableness of the normality assumption here using Table B.6 and $\alpha = 0.05$. What do you conclude?

Answer:

(e) Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into two groups, $X < 26$, $X \geq 26$, and use $\alpha = 0.01$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

Answer:

(f) Information is given below for each student on two variables not included in the model, namely, intelligence test score (X_2) and high school class rank percentile (X_3). (Note that larger class rank percentiles indicate higher standing in the class, e.g., 1% is near the bottom of the class and 99% is near the top of the class.) Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either of these variables. What do you conclude?

Answer:

Problem 4:

Refer to Copier maintenance Problem 1.20.

(a) Prepare a dot plot for the number of copiers serviced X_i. What information is provided by this plot? Are there any outlying cases with respect to this variable?

Answer:

(b) The cases are given in time order. Prepare a time plot for the number of copiers serviced. What does your plot show?

Answer:

(c) Prepare a stem-and-leaf plot of the residuals. Are there any noteworthy features in this plot?

Answer:

(d) Prepare residual plots of e_i versus $\hat{Y}_i$ and e_i versus X_i on separate graphs. Do these plots provide the same information? What departures from regression model (2.1) can be studied from these plots? State your findings.

Answer:

(e) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Does the normality assumption appear to be tenable here? Use Table B.6 and $\alpha = .10$.

Answer:

(f) Prepare a time plot of the residuals to ascertain whether the error terms are correlated over time. What is your conclusion?

Answer:

(g) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .05$. State the alternatives, decision rule and conclusion.

Answer:

(h) Information is given below on two variables not included in the regression model, namely, mean operational age of copiers serviced on the call (X_2, in months) and years of experience of the service person making the call (X_3). Plot the residuals against X_2 and X_3 on separate graphs to ascertain whether the model can be improved by including either or both of these variables. What do you conclude?

Answer:

Problem 5:

Refer to Airfreight breakage Problem 1.21.

(a) Prepare a dot plot for the number of transfers X_i. Does the distribution of number of transfers appear to be asymmetrical?

Answer:

(b) The cases are given in time order. Prepare a time plot for the number of transfers. Is any systematic pattern evident in your plot? Discuss.

Answer:

(c) Obtain the residuals e_i and prepare a stem-and-leaf plot of the residuals. What information is provided by your plot?

Answer:

(d) Plot the residuals e_i against X_i to ascertain whether any departures from regression model (2.1) are evident. What is your conclusion?

Answer:

(e) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is reasonable here. Use Table B.6 and $\alpha = .01$. What do you conclude?

Answer:

(f) Prepare a time plot of the residuals. What information is provided by your plot?

Answer:

(g) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .10$. State the alternatives, decision rule and conclusion. Does your conclusion support your preliminary findings in part (d)?

Answer:

Problem 6:

Refer to Plastic hardness Problem 1.22.

(a) Obtain the residuals e_i and prepare a box plot of the residuals. What information is provided by your plot?

Answer:

(b) Plot the residuals e_i against the fitted values $\hat{Y}_i$ to ascertain whether any departures from regression model (2.1) are evident. State your findings.

Answer:

(c) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation betwewn the ordered residuals and their expected values under normality. Does the normality assumption appear to be reasonable here? Use Table B.6 and $\alpha = .05$.

Answer:

(d) Compare the frequencies of the residuals against the expected frequencies under normality, using the 25th, 50th and 75th percentiles of the relevant t distribution. Is the information provided by these comparisons consistent with the findings from the normal probability plot in part (c)?

Answer:

(d) Use the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into the two groups, $X \leq 24$ and $X > 24$ and use $\alpha = .05$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (b)?

Answer:

Problem 7:

Refer to Muscle mass Problem 1.27.

(a) Prepare a stem-and-leaf plot for the ages X_i. Is this plot consistent with the random selection of women from each 10-year age group? Explain.

Answer:

(b) Obtain the residuals e_i and prepare a dot plot of the residuals. What does your plot show?

Answer:

(c) Plot the residuasl e_i against $\hat{Y}_i$ and also against X_i on separate graphs to ascertain whether any departures from regression model (2.1) are evident. Do the two plots provide the same information? State your conclusions.

Answer:

(d) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality to ascertain whether the normality assumption is tenable here. Use Table B.6 and $\alpha = .10$. What do you conclude?

Answer:

(e) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with the level of X. Use $\alpha = .01$. State the alternatives, decision rule and conclusion. Is your conclusion consistent with your preliminary findings in part (c)?

Answer:

Problem 8:

Refer to Crime rate Problem 1.28.

(a) Prepare a stem-and-leaf plot for the percentage of individuals in the county having at least a high school diplima X_i. What information does your plot provide?

Answer:

(b) Obtain the residuals e_i and prepare a box plot of the residuals. Does the distribution of the residuals appear to be symmetrical?

Answer:

(c) Make a residual plot of e_i against $\hat{Y}_i$. What does the plot show?

Answer:

(d) Prepare a normal probability plot of the residuals. Also obtain the coefficient of correlation between the ordered residuals and their expected values under normality. Test the reasonableness of the normality using Table B.6 and $\alpha = .05$. What do you conclude?

Answer:

(e) Conduct the Brown-Forsythe test to determine whether or not the error variance varies with the level of X. Divide the data into two groups, $X \geq 69$ and $X > 69$ and use $\alpha = .05$. State the decision rule and conclusion. Does your conclusion support your preliminary findings in part (c)?

Answer:

Problem 9:

Electricity consumption. An economist studying the relation between household electricity (Y) and number of rooms in the home (X) employed linear regression model (2.1) and obtained the residuals. Plot the residuals against X_i. What problem appears to be present here? Might a transformation alleviate the problem?

Answer:

Problem 10:

Per capita earnings. A sociologist employed linear regression model (2.1) to relate per capita earnings (Y) to average number of years of schooling (X) for 12 cities. The fitted values $\hat{Y}_i$ and the semistudentized residuals e_i^* are given.

(a) Plot the semistudentized residuals against the fitted values. What does the plot suggest?

Answer:

(b) How many semistudentized residuals are outside ± 1 standard deviation? Approximately how many would you expect to see if the normal error model is appropriate?

Answer:

Problem 11:

Drug concentration. A pharmacologist employed linear regression model (2.1) to study the relation between the concentration of a drug in plasma (Y) and the log-dose of the drug (X). The residuals and log-dose levels are given.

(a) Plot the residuals e_i against X_i. What conclusions do you draw from the plot?

Answer:

(b) Assume that (3.10) is applicable and conduct the Breusch-Pagan test to determine whether or not the error variance varies with log-dose of the drug (X). Use $\alpha = .05$. State the alternatives, decision rule and conclusion. Does your conclusion support your preliminary findings in part (a)?

Answer:

Problem 12:

A student does not understand why the sum of squares defined in (3.16) is called a pure error sum of squares "since the formula looks like one for an ordinary sum of squares" Explain.

Answer:

Problem 13:

Refer to Copier maintenance Problem 1.20.

(a) What are the alternative conclusions when testing for lack of fit of a linear regression function?

Answer:

(b) Perform thetest indicated in part (a). Control the risk of Type I error at .05. State the decision rule and conclusion.

Answer:

(c) Does your test in part (b) detect other departures from regression model (2.1), such as lack of constant variance or lack of normality in the error terms? Could the results of the test of lack of fit be affected by such departures? Discuss.

Answer:

Problem 14:

Refer to Plastic hardness Problem 1.22.

(a) Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .01$. State the alternatives, decision rule and conclusion.

Answer:

(b) Is there any advantage of having an equal number of replications at each of the X levels? Is there any disadvantage?

Answer:

(c) Does the test in part (a) indicate what regression function is appropriate when it leads to the conclusion that the regression function is not linear? How would you proceed?

Answer:

Problem 15:

Solution concentration. A chemist studied the concentration of a solution (Y) over time (X). Fifteen identical solutions were prepared. The 15 solutions were randomly divided into five sets of three, and the five sets were measured, respectively, after 1, 3, 5, 7 and 9 hours.

(a) Fit a linear regression function.

Answer:

(b) Perform the F test to determine whether or not there is lack of fit of a linear regression function; use $\alpha = .025$. State the alternatives, decision rule and conclusion.

Answer:

(c) Does the test in part (b) indicate what regression function is appropriate when it leads to the conclusion that lack of fit of a linear regression function exists? Explain.

Answer:

Problem 16:

Refer to Solution concentration Problem 3.15.

(a) Prepare a scatter plot of the data. What transformation of Y might you try, using the prototype patterns in Figure 3.15 to achieve constant variance and linearity?

Answer:

(b) Use the Box-Cos procedure and standardization (3.36) to find an appropriate power transformation. Evaluate SSE for $\lambda = -.2, -.1, 0, .1, .2$. What transformation of Y is suggested?

Answer:

(c) Use the transformation $Y' = \log_{10} Y$ and obtain the estimated linear regression funcion for the transformed data.

Answer:

(d) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

Answer:

(e) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

Answer:

(f) Express the estimated regression function in the original units.

Answer:

Problem 17:

Sales growth. A marketing researcher studied annual sales of a product that had been introduced 10 years ago. The data is given as follows, where X is the year (coded) and Y is sales in thousands of units.

(a) Prepare a scatter plot of the data. Does a linear relation appear adequate here?

Answer:

(b) Use the Box-Cox procedure and standarization (3.36) to find an appropriate power transformation of Y. Evaluate SSE for $\lambda = .3, .4, .5, .6, .7$. What transformation of Y is suggested?

Answer:

(c) Use the transformation $Y' = \sqrt{Y}$ and obtain the estimated linear regression function for the transformed data.

Answer:

(d) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

Answer:

(e) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

Answer:

(f) Express the estimated regression function in the original units.

Answer:

Problem 18:

Production time. In a manufacturing study, the production times for 111 recent production runs were obtained. The table lists for each run the production time in hours (Y) and the production lot size (X).

(a) Prepare a scatter plot of the data. Does a linear relation appear adequate here? Would a transformation on X or Y be more appropriate here? Why?

Answer:

(b) Use the transformation $X' = \sqrt{X}$ and obtain the estimated linear regression function for the transformed data.

Answer:

(c) Plot the estimated regression line and the transformed data. Does the regression line appear to be a good fit to the transformed data?

Answer:

(d) Obtain the residuals and plot them against the fitted values. Also prepare a normal probability plot. What do your plots show?

Answer:

(e) Express the estimated regression function in the original units.

Answer:

Problem 19:

A student fitted a linear regression function for a class assignment. The student plotted the residuals e_i against Y_i and found a positive relation. When the residuals were plotted against the fitted values $\hat{Y}_i$, the student found no relation. How could this difference arise? Which is the more meaningful plot?

Answer:

Problem 20:

If the error terms in a regression model are independent $N(0, \sigma^2)$, what can be said about the error terms after transformation $X' = \frac{1}{X}$ is used? Is the situation the same after transformation $Y' = \frac{1}{Y}$ is used?

Answer:

Problem 21:

Derive the result in (3.29).

Answer:

Problem 22:

Using (A.70), (A.41) and (A.42), show that E[MSPE] $= \sigma^2$ for normal error regression model (2.1).

Answer:

Problem 23:

A linear regression model with intercept $\beta_0 = 0$ is under consideration. Data have been obtained that contain replications. State the full and reduced models for testing the appropriateness of the regression function under consideration. What are the degrees of freedom associated with the full and reduced models if $n = 20$ and $c = 10$?

Answer:

Problem 24:

Blood pressure. Data was obtained in a study of the relation between diastolic blood pressure (Y) and age (X) for boys 5 to 13 years old.

(a) Assuming normal error regression model (2.1) is appropriate, obtain the estimated regression function and plot the residuals e_i against X_i. What does your residual plot show?

Answer:

(b) Omit case 7 from the data and obtain the estimated regression function based on the remaining seven cases. Compare this estimated regression function to that obtained in part (a). What can you conclude about the effect of case 7?

Answer:

(c) Using your fitted regression function in part (b), obtain a 99 percent prediction interval for a new Y observation at $X = 12$. Does observation Y_7 fall outside this prediction interval? What is the significance of this?

Answer:

Problem 25:

Refer to the CDI data set in Appendix C.2 and Project 1.43. For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more appropriate in one case than in the others?

Answer:

Problem 26:

Refer to the CDI data set in Appendix C.2 and Project 1.44. For each geographic region, obtain the residuals and prepare a residual plot against X and a normal probability plot. Do the four regions appear to have similar error variances? What other conclusions do you draw from your plots?

Answer:

Problem 27:

Refer to the SENIC data set in Appendix C.1 and Project 1.45.

(a) For each of the three fitted regression models, obtain the residuals and prepare a residual plot against X and a normal probability plot. Summarize your conclusions. Is linear regression model (2.1) more apt in one case than in the others?

Answer:

(b) Obtain the fitted regression function for the relation between length of stay and infection risk affter deleting cases 47 ($X_{47} = 6.5$, $Y_{47} = 19.56$) and 112 ($X_{112} = 5.9$ and $Y_{112} = 17.94$). From this fitted regression function obtain separate 95 percent prediction intervals for new Y observations at $X = 6.5$ and $X = 5.9$, respectively. Do observations Y_{47} and Y_{112} fall outside these prediction intervals? Discuss the significance of this.

Answer:

Problem 28:

Refer to the SENIC data set in Appendix C.1 and Project 1.46. For each geographic region, obtain the residuals and prepare a residual plot against X and a normal probability plot. Do the four regions appear to have similar error variance? What other conclusions do you draw from your plots?

Answer:

Problem 29:

Refer to Copier maintenance Problem 1.20.

(a) Divide the data into four bands according to the number of copiers serviced (X). Band 1 ranges from $X = .5$ to $X = 2.5$; band 2 ranges from $X = 2.5$ to $X = 4.5$; and so forth. Determine the median value of X and the median value of Y in each of the bands and develop the band smooth by connecting the four pairs of medians by straight lines on a scatter plot of the data. Does the band smooth suggest the regression relation is linear? Discuss.

Answer:

(b) Obtain the 90 percent confidence band for the true regression line and plot it on the scatter plot prepared in part (a). Does the band smooth fall entirely inside the confidence band? What does this tell you about the appropriateness of the linear regression function?

Answer:

(c) Create a series of six overlapping neighborhoods of width 3.0 beginning at $X = .5$. The first neighborhood will range from $X = .5$ to $X = 3.5$; the second neighborhood will range from $X = 1.5$ to $X = 4.5$; and so on. For each of the six overlapping neighborhoods, fit a linear regression function and obtain the fittedvalue $\hat{Y}_c$ at the center X_c of the neighborhood. Develop a simplified version of the lowess smooth by connecting the six$(X_c, \hat{Y}_c)$ pairs by straight lines on a scatter plot of the data. In what ways does your simplified lowess smooth differ from the band smooth obtained in part (a)?

Answer:

Problem 30:

Refer to Sales growth Problem 3.17.

(a) Divide the range of the predictor variable (coded years) into five bands of width 2.0, as follows: Band 1 ranges from $X = -.5$ to $X = 1.5$; band 2 ranges from $X = 1.5$ to $X = 3.5$; and so on. Determine the median value of X and the median value of Y in each band and develop the band smooth by connecting the five pairs of medians by straight lines on a scatter plot of the data. Does the band smooth suggest that the regression relation is linear? Discuss.

Answer:

(b) Create a series of seven overlapping neighborhoods of width 3.0 beginning at $X = -.5$. The first neighborhood will range from $X = -.5$ to $X = 2.5$; the second neighborhood will range from $X = .5$ to $X = 3.5$; and so on. For each of the seven overlapping neighborhoods, fit a linear regression function and obtain the fitted value $\hat{Y}_c$ at the center X_c of the neighborhood. Develop a simplified version of the lowess smooth by connecting the seven $(X_c, \hat{Y}_c)$ pairs by straight lines on a scatter plot of the data.

Answer:

(c) Obtain the 95 percent confidence band for the true regression line and plot it on the plot prepared in part (b). Does the simplified lowess smooth fall entirely within the confidence band for the regression line? What does this tell you about the appropriateness of the linear regression function?

Answer:

Problem 31:

Refer to the Real estate sales data set in Appendix C.7. Obtain a random sample of 200 cases from the 522 cases in this data set. Using the random sample, build a regression model to predict sales price (Y) as a function of finished square feet (X). The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If the regression assumptions are not met, include and justify appropriate remeidal measures. Use the final model to predict sales price for two houses that are about tocome on the market: the first has $X = 1100$ finished square feet and the second has $X = 4900$ finished square feet. Assess the strengths and weaknesses of the final model.

Answer:

Problem 32:

Refer to the Prostate cancer data set in Appendix C.5. Build a regression model to predict PSA lebel (Y) as a function of cancer volume (X). The analysis should include an assessment of the degree to which the key regression assumptions are satisfied. If the regression assumptions are not met, include and justify appropriate remedial measures. Use the final model to estimate mean PSA level for a patient whose cancer volume is 20 cc. Assess the strengths and weaknesses of the final model.

Answer: