1 Regression Residuals and Transformations

The file transforms.csv on the course website contains 4 pairs of Xs and Ys. For each pair:

(a) Fit the linear regression model $Y = \beta_0 + \beta_1 X + \varepsilon, \varepsilon \sim N(0, \sigma^2)$. Plot the data and fitted line.
(b) Provide a scatterplot, normal Q-Q plot, and histogram for the studentized regression residuals.
(c) Using the residual scatterplots, state how the SLR model assumptions are violated.
(d) Determine the data transformation to correct the problems in (c), fit the corresponding regression model, and plot the transformed data with new fitted line.
(e) Provide plots to show that your transformations have (mostly) fixed the model violations.

2 Infant Nutrition

Revisit the infant nutrition data from homework 3.

(Aside: One reason we revisit data sets so often in this course is to get a sense of the progression of an analysis. Trying different models, diagnostics, etc, comparing what we learn. The course builds over time, and this way you can see all the tools being brought to bear on the same problem.)

In that homework, we fit three different models:

$$E[\text{woh}|\text{age}] = \beta_0 + \beta_1 \text{age}$$
$$E[\text{woh}|\text{age}] = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2$$
$$E[\text{woh}|\text{age}] = \beta_0 + \beta_1 \text{age} + \beta_2 1\{\text{age} > 7\} + \beta_3 \text{age} \times 1\{\text{age} > 7\}$$

(a) Use our main three diagnostic tools for the three models above. Discuss what you conclude from comparison of all plots.
(b) In homework 3, residual plots motivated us to consider $\text{age}^2$. Present a statistical test that justifies adding $\text{age}^2$. Repeat the same test to see if adding $\text{age}^3$ is appropriate. What is the highest-order polynomial fit you can justify?
(c) Using our suite of diagnostic tools, compare the simple linear model (just $\text{age}$) to the highest-order polynomial model you tested in (b) and the third model displayed above, that has a dummy for the first 7 observations. Based on the residual plots, what do you conclude about each model?
(d) Upon a scatterplot of the data, overlap predicted regression lines and predictive intervals for the third model and the model you selected in (a) and (b).
(e) Overall, what model do you prefer and why?

(f) In part (b) we tested adding higher order polynomials one at a time to see how far we could go. Now, using a partial $F$-test evaluate adding entire batches of higher order polynomials and see what order is justified. That is, instead of testing

$$E[\text{woh}|\text{age}] = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \cdots + \beta_k \text{age}^k$$

vs.

$$E[\text{woh}|\text{age}] = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \cdots + \beta_k \text{age}^k + \beta_{k+1} \text{age}^{k+1},$$

use the partial $F$-test for

$$E[\text{woh}|\text{age}] = \beta_0 + \beta_1 \text{age}$$

vs.

$$E[\text{woh}|\text{age}] = \beta_0 + \beta_1 \text{age} + \beta_2 \text{age}^2 + \cdots + \beta_k \text{age}^k.$$

for $k = 2, 3, \ldots$. What is the highest-order polynomial fit you can justify?

(g) Using our suite of diagnostic tools, compare the simple linear model (just $\text{age}$) to the highest-order polynomial model you tested in (f). What do you conclude about which model is best? What do you conclude about the diagnostic tools themselves?

3 Newspaper Circulation

Data were collected on the average Sunday and daily (i.e., weekday) circulations (in thousands) for 48 of the top 50 newspapers in the United States for the period March–September, 1993. See the newspaper.csv file on the course web site.

(a) Construct a scatter plot of Sunday circulation versus daily circulation. Does the plot suggest a linear relationship between the variables? Do you think this is a plausible relationship?

(b) Fit a regression line predicting Sunday circulation from daily circulation.

(c) What do $\beta_0$ and $\beta_1$ represent in this model? Be precise.

(i) Is there any (statistically) significant relationship between Sunday circulation and daily circulation? Justify your answer by a statistical test. Fully describe the test you are using, include null and alternative hypothesis, test statistic, and critical value.

(ii) What is special in this context about the value $\beta_1 = 1$? Given this data, is $\beta_1 = 1$ plausible? Justify your answer statistically.

(iii) Test the null hypothesis of $\beta_1 = 1$ against a two-sided alternative, with robust standard errors. What is the conclusion you reach? What is the $p$-value associated with this test?

(d) Suppose that you are proposing to add a Sunday edition of a newspaper with a weekday circulation of 225,000 copies. What would you tell advertisers is the expected Sunday circulation? What is the standard deviation of this expectation? What would you say when they ask you to predict a likely range of possible Sunday circulation numbers?

(e) Argue that working with the logarithm of the circulation(s) might be better than using the raw numbers. Fit the corresponding log-log regression model. Compare and contrast the fit and the predictive interval obtained for the Sunday edition of a newspaper with a weekly circulation of 225,000 copies.
4 Price Elasticity and Cheese

This question considers sales volume as well as price and display activity for packages of Borden Sliced Cheese. The data, available as cheese.csv on the course site, are taken from Rossi, Allenby, and McCulloch’s Bayesian Statistics and Marketing. For each of 88 stores (store) in different US cities, we have repeated observations of the sales volume (vol, in terms of packages sold), unit price (price), and whether the product was advertised with an in-store display (disp = 1 for display).

Answer the following questions in a clear and concise manner. Include the appropriate plots and hypothesis tests to illustrate and support your conclusions. Present your solutions as though you are a consultant seeking to inform and convince a (very statistics savvy) client of your results.

(a) Ignoring price, do the in-store displays have an effect on log sales? Is there reason to suspect that your result is confounded by pricing strategies?

(b) A better question: is price elasticity for Borden cheese effected by the presence of in-store advertisement?

(i) Test this by running two separate regressions. Note that testing if one value is equal to another is the same as testing if the difference is equal to zero. Also, if $b$ and $b^{*}$ are least squares coefficients from independent regression fits, then $sd(b - b^{*}) = \sqrt{s_{b}^{2} + s_{b^{*}}^{2}}$ (and $n$ is big enough that the $b$’s are all normal).

(ii) How can you test this with only one regression? Is the result the same?

(c) Do you have a possible economic explanation for your results in (b)?

5 Furniture Data

Return to the furniture.csv data from Homework 5. See the description and information in that file. Create the time variable that counts the months starting with the first, and regress sales on time, as we did in Homework 1:

$$sales_{i} = \beta_{0} + \beta_{1} \times time_{i} + \varepsilon_{i}.$$ 

We want to check our regression assumptions for this data, the best we can.

(a) Use plots to check if the residuals are normally distributed. What do you think?

Importantly, these checks do not tell us if the residuals are independent of each other, which we also required. Here they are probably not independent, that’s the whole point of time series data. All the variance formulas we used in class relied on independence (see the derivation handout). Therefore, we can’t really trust our inference (i.e. standard errors, tests, confidence intervals) at this point. Doing this stuff with time series data can be subtle, and in this class we will not get into the details.

(b) Plot the residuals against fitted values (remember that this is equivalent to plotting the residuals against the $X$ variable). Does it appear that our assumption of constant variance is satisfied?

The second thing we check with the plot of residuals against $X$, or against fitted values, is if there is any information/pattern left in the $X$ variable that we should be extracting. In class so far, this has been in terms of functional form, usually looking to add polynomial terms or dummy variables. Now, we are looking for patterns over time, and this is a little different. By definition, the linear regression
above extracts all the *linear* information contained $X$ variable *time*, but this does *not* mean that we have captured the entire time dependence pattern.

(c) Again plot the residuals against the fitted values, but this time make all the points for December a different color. Also, plot the data and the regression line (as you did for Homework 1), but this time make all the points for December a different color. What do these two plots tell you? What would you change about the regression after seeing these plots?