Chapter 11: r.m.s. error for regression

Context

- In Chapter 10 we looked at predicting the value for y from a given value of x.
- How precise are these estimates?
- To determine that, we will look at the r.m.s. error of the regression line
- Other new things: residual, residual plot, homoscedasticity, heteroscedasticity, normal approximation for points within narrow strip

2 / 19

Prediction error

3 / 19

r.m.s. error for the regression line

- See overhead
- prediction error = actual y value - predicted y value
- To determine the typical size of these errors, we take the r.m.s. size of them:

$$\sqrt{\frac{(error\#1)^2 + (error\#2)^2 + \cdots + (error\#n)^2}{n}}$$

- This is the r.m.s. error of the regression line, which says how far typical points are above or below the regression line

4 / 19

$68\% - 95\%$ rule

- $68\% - 95\%$ rule:
  - About 68% of the points on a scatter diagram are within one r.m.s. error from the regression line
  - About 95% of the points on a scatter diagram are within two r.m.s. errors from the regression line
  - See overhead

5 / 19

Computing the r.m.s. error

- We know how to compute the r.m.s. error:
  - For each data point, predict y using the regression method
  - Compute the prediction errors:
    actual value - predicted value
  - Take the r.m.s. size of the prediction errors
- That is a lot of work! You'll learn a shortcut soon.

6 / 19

r.m.s. error vs. SD of y

- Compare the following two lines for predicting y
  - a horizontal line through the average of y
  - the regression line
- What are the differences?
  - The horizontal line ignores the value of x
  - The regression line uses the value of x
  - The prediction errors for the horizontal line are expected to be larger. Why? Because we don't use information on x.
  - The r.m.s. size of the prediction errors for the horizontal line is just the SD of y. Why? Because now the prediction errors are the deviations from the average, and the r.m.s. size of those is the SD of y.

7 / 19

When to use which one?

- So when to use the r.m.s. error and when to use the SD of y?
  - If there is no information on x, then:
    - The best prediction for y is its average
    - In other words: we use the horizontal line
    - This prediction is likely to be off by the e r.m.s. error of this line, which is the SD of y
  - If we know the value of x, and the scatter diagram is football shaped, then:
    - The best prediction for y is the regression estimate
    - In other words, we use the regression line
    - This prediction is likely to be off by the r.m.s. error of the regression line

8 / 19

Computing the r.m.s. error

- So the r.m.s. error for the regression line is smaller than the SD of y. By how much?
- By a factor $\sqrt{1-r^2}$:

$$\text{r.m.s. error of the regression line} = \sqrt{1-r^2} \times \text{SD of } y$$

- That is a nice shortcut!
- How does the r.m.s. error of the regression line compare to the r.m.s. error of other straight lines? For football shaped diagrams, the regression line is the line with the smallest r.m.s. error (see Chapter 12).

9 / 19

r.m.s. error vs. r

- Both the r.m.s. error of the regression line and r tell us about the spread/clustering of the points around the regression line
  - r measures clustering. If r close to -1 or 1, then the points are tightly clustered.
  - r.m.s. error measures distance of points to the regression line. If the r.m.s. error is small, the points are close to the line.
- What are the units of the r.m.s. error?
  - The same as the original units of y
- Why need r.m.s. error when we already had r, which tells us about the clustering of the points?
  - r measures clustering relative to the SD
  - the r.m.s. error measures the spread in the original units of y

10 / 19

Residual plots

11 / 19

Residual plot

- Prediction errors are also called residuals
- In the scatter diagram we usually plot (x-value, y-value)
- We can make a new plot, by plotting (x-value, residual)
- This is called a residual plot
- See overhead
- The average of the residuals is always equal to zero. Hence, the SD of the residuals is the same as the r.m.s. error of the regression line. Why?
- Why do we make residual plots?
  - To check if we can use the regression method:
    The residual plot should have no pattern, and should look like a horizontal football

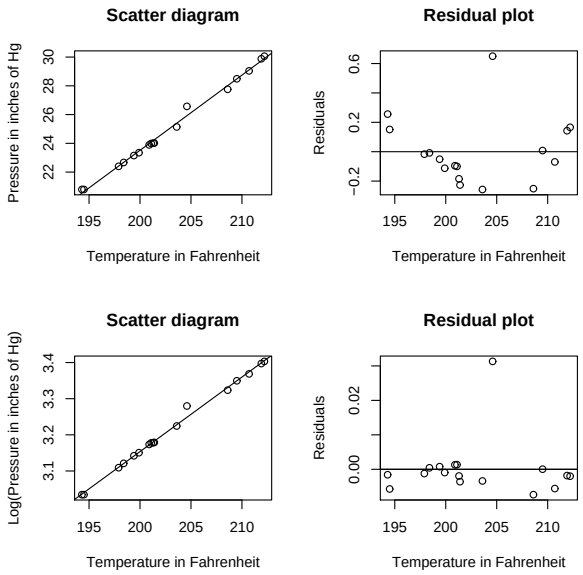
12 / 19

Example

- James D. Forbes, Scottish scientist in 1840s - 1850s.
- He wanted to estimate the altitude above see level.
- He knew that altitude could be determined from atmospheric pressure, measured by a barometer.
- In the 1840s it was difficult to transport the fragile barometers.
- Therefore he wanted to use the boiling temperature of water to estimate the atmospheric pressure, and that would then tell him the altitude.

13 / 19

Example



14 / 19

Looking at vertical strips

15 / 19

r.m.s. error

- See overhead (Exercise 3 on page 190)
- Take the points in a narrow vertical strip
- The SD of their y-coordinates is given by the r.m.s. error of the regression line, if the scatter diagram is homoskedastic (football shaped).

16 / 19

Homoscedastic vs heteroscedastic

- Some Greek:
  - homos = same, heteros = different, skedastos = scatter
- homeskedastic (football shaped):
  - Vertical strips in scatter diagram have similar spread
  - Note that the range of the points is larger in the middle of the football, due to larger number of points there. But the SDs of the y-values in each strip are similar.
  - This SD is given by the r.m.s. error of the regression line
- heteroskedastic (not football shaped):
  - The vertical strips have different amounts of spread
  - The r.m.s. error of the regression line gives a sort of average spread across all vertical strips, and cannot be used for any one vertical strip

17 / 19

Use normal curve inside strip

- Suppose we have a football shaped scatter diagram
- Consider points in a narrow vertical strip as new data set
- The new average of this data set is the regression estimate
- The new SD of this data set is the r.m.s. error of regression line
- The football shape means that we can use the normal approximation
- Hence, we can use the normal approximation as before, with the new average and the new SD.

18 / 19

Example

- Data on height and forearm length of 1000 men:
  - height: average = 68 inches, SD = 2.7 inches
  - forearm length: average = 18 inches, SD = 1 inch
  - $r = 0.80$
  - the scatter diagram is football shaped
- What percentage of men have forearms longer than 20 inches? See overhead. Straightforward normal approximation
- Of the men who are 71 inches tall, what percentage have forearms longer than 20 inches? See overhead:
  - Step 1: Draw picture
  - Step 2: Find new average of strip: regression estimate
  - Step 3: Find new SD of strip: r.m.s. error
  - Step 4: Then do normal approximation in the strip

19 / 19