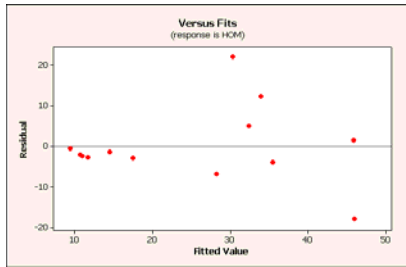


Chapter 8: Linear Regression

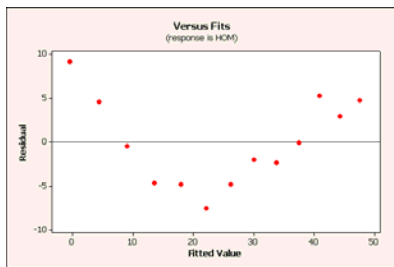
- Provide an example (draw it or describe it) of residuals where
 - there is non-constant variance

THE RESIDUAL PLOT WILL CHANGE SPREAD IN Y AS YOU MOVE ON THE X-AXIS



- the underlying model is not linear

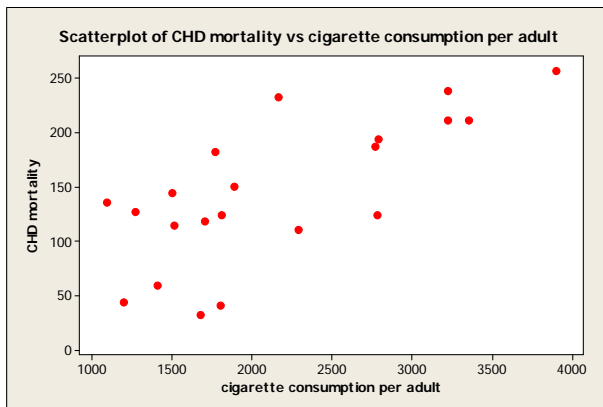
THE RESIDUAL PLOT WILL HAVE A CURVE TO IT



Data Set:

- Smoking
 - Data shows the average number of cigarettes consumed per adult per year and the rate of deaths from coronary heart disease (in deaths per 100,000) for several countries.
- <http://www.canyons.edu/faculty/morrowa/140/datasets/>

- Create an appropriate graph and find appropriate statistics for the data.



BECAUSE WE ARE DEALING WITH TWO QUANTITATIVE VARIABLES, A SCATTER PLOT IS APPROPRIATE.

BECAUSE THE RELATIONSHIP APPEARS LINEAR AND THERE ARE NO APPARENT OUTLIERS, WE REPORT THE CORRELATION: $r = 0.730$

IN THE 21 COUNTRIES SURVEYED, THE MEAN CIGARETTE CONSUMPTION WAS APPROXIMATELY 2148 CIGARETTES PER ADULT PER YEAR, WITH A STANDARD DEVIATION OF 809 CIGARETTES PER ADULT PER YEAR.

THE MEAN CORONARY HEART DISEASE RATE WAS 144.9 DEATHS PER 100,000 CITIZENS, WITH A STANDARD DEVIATION OF 66.5 DEATHS PER 100,000 RESIDENTS.

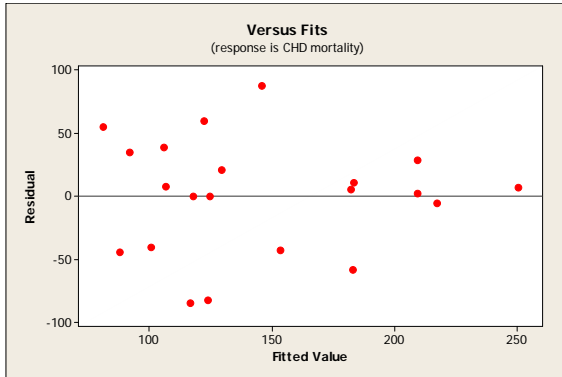
- Describe the association between cigarette smoking and coronary heart disease.

THERE IS A MODERATE, POSITIVE LINEAR ASSOCIATION.

- Create a linear model for the data.

$$\text{CHD mortality} = 15.8 + 0.0601 \text{ cigarette consumption per adult}$$

5. Evaluate the strength and appropriateness of the model (hint: use the original scatterplot along with residual plots and R^2).



THE RESIDUAL PLOT DOES NOT HAVE ANY OBVIOUS PATTERN. THEREFORE, THE ORIGINAL DATA WAS LINEAR.

R^2 FOR THE MODEL IS 53.2%. THIS IS SLIGHTLY HIGHER THAN OUR CUTOFF OF 50%. THEREFORE, THE LINEAR PATTERN IS MODERATE TO WEAK.

FINAL CONCLUSION: THE LINEAR MODEL IS GOOD.

NOTE: THERE IS MILD CAUSE FOR CONCERN AS THE VARIATION DECREASES AS WE INCREASE THE FITTED VALUE, BUT THIS COULD BE BECAUSE THERE ARE NOT MANY POINTS FOR LARGE FITTED VALUES.

6. Explain the meaning of R^2 in the context of this problem.

53.2% OF THE VARIATION IN CORONARY HEART DISEASE MORTALITY RATE IS EXPLAINED BY THE LINEAR MODEL.

7. Interpret the slope and intercept of the model.

THE SLOPE TELLS US THAT FOR EACH INCREASE IN CIGARETTE CONSUMPTION BY ONE PER ADULT, THE CORONARY HEART DISEASE MORTALITY RATE WILL INCREASE BY 0.0601.

TO INTERPRET THE INTERCEPT, WE NEED TO CONSIDER IF A VALUE OF 0 IS PLAUSIBLE FOR THE NUMBER OF CIGARETTES CONSUMED PER ADULT FOR A COUNTRY. THIS IS NOT PLAUSIBLE. THEREFORE, THE Y-INTERCEPT HAS NO INTERPRETATION.

8. Would it be better for a country to have a positive or a negative residual from this model? Why?

NEGATIVE—THIS WOULD MEAN THAT THE OBSERVED CORONARY HEART DISEASE MORTALITY RATE IS BELOW THE EXPECTED VALUE.

9. What if the United States were to cut cigarette consumption in half over the next decade? What would the model suggest about the rate of deaths from coronary heart disease?

NOTE: TO SOLVE THIS PROBLEM, WE FIRST LOOK AT THE DATA TO FIND THE NUMBER OF CIGARETTES CONSUMED IN THE U.S., 3900. HALF OF THIS IS 1950. WE THEN USE MINITAB TO MAKE A PREDICTION FOR THE CORRESPONDING CORONARY HEART DISEASE MORTALITY RATE.

THIS WOULD CUT THE CORONARY HEART DISEASE MORTALITY RATE IN HALF. THE EXPECTED CDH MORTALITY RATE WOULD BE 133.

10. Did you remember to “think” about the data? I didn’t disclose everything you might want to know. Take a look at the source of the data: <http://www.pubmedcentral.nih.gov/picrender.fcgi?artid=1349138&blobtype=pdf> and try to find the “When” for the data.

THE DATA IS FROM THE EARLY 1960'S

11. Why might this data not be relevant today? What other factors today might contribute more to coronary heart disease?

MANY OTHER FACTORS ARE OF A CONCERN WITH RESPECT TO CDH MORTALITY RATE (SUCH AS OBESITY).