3.4 Redux

1. extrapolation : be wary of making predictions too far in the future.
2. outliers : a regression outlier is far from the trend and may not nesecarily be an outlier in either individual sample. always remember to investigate the reasons an observation is an outlier before removing it from the data sample.
3. example 16 : correlation $\neq$ causation
   (a) smoking example survey of 1,314 women in the uk during 1972 - 1974

	Survival Status		
Smoker	No	Yes	Total
Yes	138	443	582
No	230	502	732
Total	369	945	1314

   conditional proportions.

	Survival Status		
Smoker	No	Yes	Total
Yes	0.24	0.76	582
No	0.31	0.69	732

   percentages of deaths for the explanatory variable 'age group'.

	Age Group			
Smoker	18-34	35-54	55-64	65+
Yes	0.028	0.172	0.443	0.857
No	0.027	0.095	0.331	0.855

4.1 Gathering Data

1. experiment vs. observe
2. experiment : assigning subjects to certain experimental conditions (treatments) and then observing outcomes on response variables
   - treatment : experimental conditions, assigned values of the explanatory variables
3. observational study : researcher observes values of response and explanatory variables for sampled subjects w/o assigning treatments.
4. example
   (a) German study : compared 118 cancer patients w/ 475 healthy patients. patient cell phone use measured w/ a survey. cancer patients used cell phones more often.

(b) US study : compared 469 patients w/ brain cancer to 422 patients w/o. cell phone use measured w/questionnaire. similar cell phone use in both

(c) Australian study : 200 mice bred to be susceptible to cancer of immune system. 100 mice exposed for 2 half hour periods a day to same type of radiation from cell phones. 100 mice not exposed. after 18 months brain cancer rate for radiation mice twice as high.

(d) how are (a) and (b) different from (c)?

- (c) imposed conditions on mice. it is an experiment.

5. ex 3

(a) study of 76,000 students in 497 high schools and 225 middle schools. each student filled out a survey. students questioned about drug use. the conditional proportions are given below

	Drug Use		
Drug Tests	Yes	No	N
Yes	0.37	0.63	5,653
No	0.36	0.64	17,473

i. which is the explanatory variable?

ii. what are we studying?

iii. what kind of study is this?

6. advantages of an experiment

(a) elimination of lurking variables. all conditions controlled

(b) achieved by random sampling (?)

(c) experiments can determine cause and effect due to controlled conditions

(d) causality is best determined using an experiment

7. when can we not use an experiment?

8. why would we use an observational study?

(a) ethics : cannot assign human subjects in a study to expose themselves to harmful experimental conditions.

(b) collection problems : making sure people due the assigned treatments is often not possible.

(c) time frame : experimental time frames may be too long to wait for an answer.

(d) often the point of a study is not to establish causality, but simply determine association. in this case observational studies are the best method.

9. anecdotal evidence not a good source of data since it is probably not representative of a population.

10. data widely available on many topics. remember to rely on reputable sources.

11. sample survey : select a sample from a population and interviews them to collect data.

4.2 Sampling

1. sampling frame : list of subjects in the population from which the sample is taken
2. sampling design (or just design) : method of selecting subjects from the sampling frame
3. simple random sample : a sample where each possible sample of n subjects has an equal chance of being selected

   usually done with either a random number list or a computer
4. methods of collecting data in a sample survey

   (a) personal interview (face-to-face) :
   pros : subjects more likely to participate
   cons : cost, some subjects less likely to answer sensitive questions when a real person is involved

   (b) telephone interview :
   pros : the pros of the personal interview, lower cost
   cons : interview may be shortened because people are impatient on the phone

   (c) self-administered interview (mail/email) :
   pros : cheaper still than the phone interview
   cons : smaller completion rate
5. be wary of bias : over/under representation of certain elements of the population. happens when sample not taken completely randomly.

   sampling bias occurs from using nonrandom sampling (telephone example)

   nonresponse bias when some subjects can't be reached or refuse to participate (mindshare)

   response bias when subjects give incorrect responses. subjects could lie, or they could misunderstand the questions.
6. the problem with bias is that there is no way to determine whether the bias exists. there is also no way to correct for possible bias once the survey is taken. bias needs to be removed before the survey is taken by designing the sample properly.
7. convenience samples : samples that are taken with the convenience of the investigator in mind

   (a) the internet survey : volunteer sample. some segments of the population more likely to respond. surveys done at the mall or union, certain segments of the population are naturally under/over represented

   (b) may be necessary for some kinds of studies. medical studies sometimes need to use this method surveying patients at a certain kind of office to test the effects of a treatment or drug.

8. the short story about surveying
    (a) identify the population
    (b) construct a sampling frame listing all subjects in the population (if possible)
    (c) use a random sampling design
    (d) be cautious of sampling bias

4.3 - 4.4

skipping the rest of the chapter. if you need to perform an experiment, contact a statistician before the experiment has been performed. the statistician will take you through the non-trivial process of designing the experiment and analyzing the results.