STAT 350: Lecture 18

Reading:

Summary of Distribution theory conclusions

1. $\epsilon^T Q\epsilon/\sigma^2$ has the same distribution as $\sum \lambda_1 Z_i^2$ where the Z_i are iid $N(0,1)$ random variables (so the Z_i^2 are iid χ_1^2) and the λ_i are the eigenvalues of Q.

2. $Q^2 = Q$ (Q is idempotent) implies that all the eigenvalues of Q are either 0 or 1.

3. Points 1 and 2 prove that $Q^2 = Q$ implies that $\epsilon^T Q\epsilon/\sigma^2 \sim \chi^2_{\text{trace}(Q)}$.

4. A special case is
$$\frac{\hat\epsilon^T\hat\epsilon}{\sigma^2} \sim \chi^2_{n-p}$$

5. t statistics have t distributions.

6. If $H_o : \beta = 0$ is true then
$$F = \frac{(\hat\mu^T\hat\mu)/p}{\hat\epsilon^T\hat\epsilon/(n-p)} \sim F_{p,n-p}$$

Many extensions of this theory are possible. The most important of these are:

1. If a "reduced" model is obtained from a "full" model by imposing k linearly independent linear restrictions on β (like $\beta_1 = \beta_2$, $\beta_1 + \beta_2 = 2\beta_3$) then
$$\text{Extra SS} = \frac{\text{ESS}_R - \text{ESS}_F}{\sigma^2} \sim \chi_k^2$$
assuming that the null hypothesis (the restricted model) is true. So the Extra Sum of Squares F test has an F-distribution.

2. In ANOVA tables which add up the various rows (not including the total) are independent.

3. When the null hypothesis H_o is **not true** the distribution of the Regression SS is **Non-central** χ^2. This is used in power and sample size calculations.

Experimental Designs leading to multiple regression analysis

1. (Randomized) designed experiments:

   - want to study the effect of variables $x_1, x_2, \ldots, x_p$ on a response variable Y.
   - Experimenter **chooses** n sets of values of $x_1, x_2, \ldots, x_p$ and measures the response Y on n **experimental units**.
   - Experimental Units are assigned **at random** to levels (that is to the particular combinations of x values). (This is a much better method that other methods for deciding which experimental units get which x values.

   Example:

   - Experimental Unit is a batch of plaster
   - $n = 18$ batches made.
   - x_1 is the sand content and x_2 is the fibre content. We tried 3 settings of x_1, 3 of x_2 and tried each of the $3 \times 4 = 9$ combinations twice.

2. Randomized Block Designs

   - want to study the effect of variables $x_1, x_2, \ldots, x_p$ on a response variable Y of an experimental unit.
   - BUT Y is probably influenced by variable B which the experimenter cannot control.

   Example

   - x_1 is log(Dose) of some drug.
   - B = sex of patient (the patient is the experimental unit.

- experimenter can assign patient to level of x_1 but NOT to the level of B.
- B is called a blocking factor.

Example

- Y is lung capacity
- B_1 is cigarettes smoked per day
- B_2 is age
- B_3 is sex
- x_1 is daily vitamin C intake
- x_2 is daily Echinacea dose
- Key point is that x_1 and x_2 are under control of the experimenter but the other factors are not.

3. Observational Studies

   - values of Y and variables $x_1, x_2, \ldots, x_p$ are determined by sampling from a population.
   - covariates $x_1, x_2, \ldots, x_p$ are not controlled by the experimenter.

   Example:

   - As in the previous example but suppose viatmin C and echinacea intakes are not controlled, just measured.

Vital Distinction

- Cause and effect relations are convincingly deduced only for controlled variables.
- Interpretation of regression coefficients is difficult in observational studies.

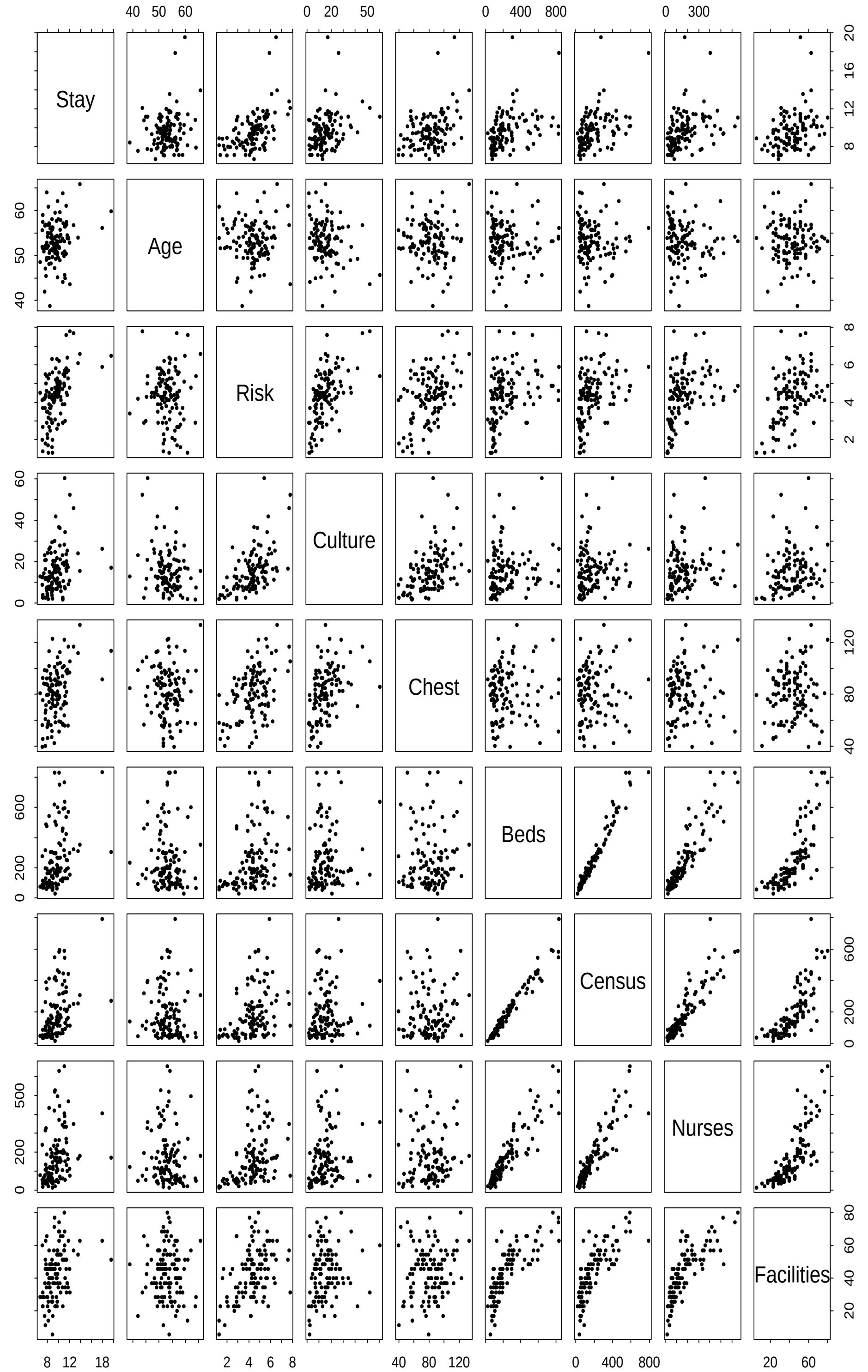
I am now going to illustrate many of the techniques we are developing in this course with an extended example. I will be using a data set from the textbook. The example will last for several lectures.

The SCENIC data set

The data set consists of a sample of 113 hospitals selected by some means which we are not told. We appear to have a purely observational study. For each hospital we have the values of the following variables:

- Average length of stay of patients in days
- Average age of patients.
- Probability of acquiring an infection in the hospital. (I don't know how this is measured.)
- Culturing ratio: 100 times the ratio (Cultures performed) divided by (number of patients with no infection).
- Chest X-ray ratio defined similarly.
- Number of beds.
- Medical school affiliation (A dichotomous, Yes or no, variable).
- Geographic region (in the US) – NE, NC, S or W.
- Number of patients.
- Number of nurses.
- Available facilities (available at the given hospital).

The data set is described in the Appendix of the text. Here I reproduce a page of pair-wise scatter plots for all variables except the categorical variables Region and School.


Stay
Age
Risk
Culture
Chest
Beds
Census
Nurses
Facilities

It is evident from the plot that, as expected, several of the variables are quite highly correlated. Here is the correlation matrix:

	Stay	Age	Risk	Culture	Chest	Beds	Census	Nurses	Facilities
Stay	1.00	0.19	0.53	0.33	0.38	-0.49	0.47	0.34	0.36
Age	0.19	1.00	0.00	-0.23	-0.02	-0.02	-0.05	-0.08	-0.04
Risk	0.53	0.00	1.00	0.56	0.45	-0.19	0.38	0.39	0.41
Culture	0.33	-0.23	0.56	1.00	0.42	-0.31	0.14	0.20	0.19
Chest	0.38	-0.02	0.45	0.42	1.00	-0.30	0.06	0.08	0.11
Beds	0.41	-0.06	0.36	0.14	0.05	-0.11	0.98	0.92	0.79
Census	0.47	-0.05	0.38	0.14	0.06	-0.15	1.00	0.91	0.78
Nurses	0.34	-0.08	0.39	0.20	0.08	-0.11	0.91	1.00	0.78
Facilities	0.36	-0.04	0.41	0.19	0.11	-0.21	0.78	0.78	1.00