Bayesian Statistics, Assignment for Friday, Sept. 20

From *Statistical Treatment of Experimental Data* **and Looking Ahead**

Finish Chapter 3 of Young (to the end of Section 11, p. 86).

If you didn't know any more statistics than what we covered in Young these four weeks, you would actually already be able to think critically about most scientific results that are popularly presented, especially in medicine and the social sciences.

In these four weeks, we covered descriptive statistics and we covered the single most important concept from frequentist statistics, which is linear regression. Linear regression is used in countless studies. In medicine it is part of the analysis for every claim of a dose-response relationship.

After the exam, we will begin Bayesian statistics. Make sure you have our Bayes textbook at hand.

For Problem Set 6

Best Fit, Linear Regression, and χ^2

The long recap is for the benefit of the people who were absent, and as a possibly-helpful review for the people who were present.

1. In Tuesday's class, we had a water pressure example. The assumption was that the water pressure is rising linearly, perhaps because a tank is being steadily filled. There were three measurements at three different times of the water pressure. We assume that water pressure actually obeys:

$p(t) = a \cdot t + b$

The water pressure was measured at three different times t_1, t_2, and t_3. The measurements were p_1 to within a range Δp, p_2 also to within a range Δp, and p_3 to within a range Δp.

If the errors in the measurements are Gaussian-distributed around the true values, then the probability of getting p_1 to within Δp is

$Ae^{-(p_1-(at_1+b))^2/2\sigma^2}\ \Delta p$

the probability of getting p_2 to within Δp is

$$A e^{-(p_2-(a t_2+b))^2/2\sigma^2}\ \Delta p$$

and the probability of getting p_3 to within Δp is

$$A e^{-(p_2-(a t_3+b))^2/2\sigma^2}\ \Delta p$$

As usual, $A = \frac{1}{\sqrt{2\pi}\ \sigma}$, but that isn't going to matter much in what follows.

The joint probability is the product of the three probabilities which is:

$$A^3\,(\Delta p)^3\ e^{-\left[(p_1-(a t_1+b))^2+(p_2-(a t_2+b))^2+(p_3-(a t_3+b))^2\right]/2\sigma^2}$$

Now if you think of what is up in the exponent as a function of b, $-\left[(p_1-(a t_1+b))^2+(p_2-(a t_2+b))^2+(p_3-(a t_3+b))^2\right]/2\sigma^2$ is a downward opening parabola.

You could also think of this whole mess as a function of a, in which case it is some other downward opening parabola.

If you multiple through by $-2\,\sigma^2$, then either of these parabolas is now upward opening. We are very interested in finding the minimum of both of these upward opening parabolas.

(a) Multiply out

$$(p_1-(a t_1+b))^2+(p_2-(a t_2+b))^2+(p_3-(a t_3+b))^2$$

and group the terms involving a^2, $a\cdot b$, b^2, a, b, and neither a nor b separately. In other words, you will have six groupings. For example, the grouping involving neither a nor b is $p_1{}^2+p_2{}^2+p_3{}^2$.

(b) What you have written is of the form

$$M_{\mathrm{aa}}\,a^2+2\,M_{\mathrm{ab}}\,a\cdot b+M_{\mathrm{bb}}\,b^2-2\,M_a\,a-2\,M_b\,b+M$$

What are M_{aa}, M_{ab}, M_{bb}, M_a, M_b, and M? Please note where I have put the 2's and the minus signs. I have put them in to make the algebra tidier. For example $M=p_1{}^2+p_2{}^2+p_3{}^2$.

(c) Using the Σ summation notation, rewrite your answers for all 6 of M_{aa}, M_{ab}, etc. For example $M=\Sigma_{i=1}^{3}\,p_i{}^2$. *We will be making great use of these six answers on Friday to finish linear regression.*

2. The notation $P(S)$ means "the probability that S is true." The notation $P(F)$ means "the probability that F is true." The notation $P(S \mid F)$ means "knowing that F is true, what is the probability that S is true." It is often read as "the probability of *S given F*." Let F be "person caused a fatal accident on a certain day," and for sake of the discussion the day will be Sept. 20. Let S be "person was drunk or stoned on a certain day," also Sept. 20.

Use these numbers:

There are 40,000,000 Californians.
1,000,000 were drunk or stoned on Sept. 20.
4 were the cause of a fatal accident on Sept 20.
Of the 4 that caused fatal accidents on Sept. 20, 2 were also drunk or stoned.

What is

$P(S)$, $P(F)$, $P(F \mid S)$, $P(S \mid F)$?

I'll get you started so you know what I am looking for: $P(S) = 1,000,000/40,000,000 = 1/4$. Finding $P(F \mid S)$ is the tricky one, but think about it, and you'll get it.

The entire point of this exercise is to make an example where $P(F \mid S) \neq P(S \mid F)$. Do you see that not only are they different, but they can be very different? Anybody that writes $P(F \mid S) = P(S \mid F)$, hasn't done many examples and hasn't met Thomas Bayes.

3. Problem 27 on p. 91. You will need to have studied pp. 78-79. The tables being referred to are in the back of the book, but you don't need to refer to them, because he has already told you in the reading the values that you need to know for the data at hand.

4. Problem 28 on p. 91. You will need to have studied p. 82, and you are just being asked to calculate Eq. 11.1. We can discuss what calculating 11.1 means after you turn in an answer on Friday.