MATH 2B RECITATION 3/1/12

ALDEN WALKER

1. Regression!

Again, read the notes on the course website!

Linear regression is very interesting. I like it because it's actually linear algebra. The fact that it's "linear" regression doesn't mean that you can only use it to fit lines—in fact, you can find a best fit of any degree polynomial, or in general the best fit of a linear combination of whatever functions you want. (For example, you could find the a in $a\log(x)$ but not a in x^a, although perhaps you could actually find the latter with the former, if you see what I mean).

You may have done this in linear algebra. Basically, you stick the outputs of the various functions you will take the linear combination of into a matrix, so for the linear combination $\beta_1 x + \beta_2 1$, that is, a best fit line, suppose we have n data points $(x_1, y_1), \ldots, (x_n, y_n)$. Then we are searching for β such that the following equation holds:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_1 & 1 \\ x_2 & 1 \\ \vdots & \vdots \\ x_n & 1 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}$$

(Denote the matrix on the right by M).

Well, of course you can't do that perfectly because the system is very overdetermined (note the image has at most dimension 2, but the range probably has way more than that, so picking everything at random, we would have a probability of 0 of picking a system which can be solved).

However, you can find β such that the right hand side is as close as possible to the $\mathbf{y}$ vector. Note that this closest vector will clearly (if you think about it) be the projection of $\mathbf{y}$ onto the column space of M. And, it turns out there's a nice formula for this projection! It's pretty cool that such a thing exists (well, that such a thing exists and it's not very complicated), when you think about what you would have to do if you projected it by hand (Gram-Schmidt probably, etc).

There are two relevant formulas here; the formula for the *projection* $\hat{\mathbf{y}}$ is $\hat{\mathbf{y}} = M(M^TM)^{-1}M^T\mathbf{y}$. The formula for the *coefficients* of the projection in terms of the basis given by the columns of M (this is what you will want in this senario) is then $\hat{\beta} = (M^TM)^{-1}M^T\mathbf{y}$. Note that if you were doing this over $\mathbb{C}$ rather than $\mathbb{R}$, you would have adjoints instead of transposes (the matrix is $M(M^*M)^{-1}M^*$).

The difference between the vector $\mathbf{y}$ and its projection $\hat{\mathbf{y}}$ (i.e. the length) is the SSR (sum of the squares of the residuals). You could write it as $\|\mathbf{y} - \hat{\mathbf{y}}\|^2$.

1.1. Distributions.

Ok, so the right way to think about the situation is to imagine that there is a real (as in, it exists) vector β which would solve the matrix equation above exactly, except that $N(0, \sigma^2)$ errors have been inserted into the $\mathbf{y}$ observation vector. Therefore, it makes sense to ask the question: how close did my coefficients $\hat{\beta}$ come to the real ones β?

- Suppose that you have n observations, and r regression coefficients (the length of β). Then it turns out that
$$\frac{c\beta - c\hat{\beta}}{s\sqrt{c(M^TM)^{-1}c^T}} \sim t_{n-r}$$
where: "$\sim$" means "is distributed as," $s = \sqrt{\frac{SSR}{n-r}}$, and c is any vector of length r (like, if you're asking about β_0, you could use $c = (1, 0, \ldots, 0)$).
- And $E(c\hat{\beta}) = c\beta$, which means $c\hat{\beta} \sim N(c\beta, c(M^TM)^{-1}c^T\sigma^2)$.

- The SSR is also called the RSS, and we can use it to produce an unbiased estimator of the error variance σ^2, i.e. $\widehat{\sigma^2} = s^2 = \frac{\|\mathbf{y}-\hat{\mathbf{y}}\|^2}{n-r}$
- Another useful fact is that $Var(c\hat{\beta}) = c(M^TM)^{-1}c^T\sigma^2$, and we use this to make the definition of the *standard error* of an estimate $c\hat{\beta}$ as $c(M^TM)^{-1}c^Ts^2$. The standard error of one of the coefficients β_i is just the standard error of $c\hat{\beta}$ where c has a 1 in the ith spot and zeros elsewhere.
- (This item is obselete with our new book, but I'm leaving it here because it might come in handy). The book tends to use the notation s_x^2 for the variance of x, or $s_{\hat{\beta}_i}^2$ for the standard error of $\hat{\beta}_i$. When the book says that the coefficient of multiple determination is:

$$R^2 = \frac{s_y^2 - s_{\hat{\epsilon}}^2}{s_y^2}$$

  This means that you should take the variance of the original given $\mathbf{y}$, subtract the variance of the observed errors, and divide by the variance of $\mathbf{y}$.
- You can also guess a new observation. Suppose you'd like to know $\mathbf{y}_{n+1} = c\beta$. The obvious guess is $c\hat{\beta}$, and the question is: how good a guess is this? Well, it turns out that

$$\frac{\mathbf{y}_{n+1} - c\hat{\beta}}{s\sqrt{1 + c(M^TM)^{-1}c^T}} \sim t_{n-r}$$

  From which you can get confidence intervals or test hypotheses. The reason there is a 1 added in is that we aren't totally sure about the variance, hence the $c(M^TM)^{-1}c^T$, but there will also be experimental error in our new observation!

This essentially tells you all you need to know to give confidence intervals and deal with hypotheses involving β.

1.2. **Example.** Let's fit a line $\beta_1 x + \beta_2$ to the following data and give 95% confidence intervals for the β_i:

y	x
0.801033	0
4.23744	1
5.48164	2
3.20569	3
6.24067	4
7.45512	5
8.21995	6
10.3746	7
9.10986	8
11.861	9
12.25	10

The matrix M here is:

$$M = \begin{bmatrix} 0 & 1 \\ 1 & 1 \\ 2 & 1 \\ 3 & 1 \\ 4 & 1 \\ 5 & 1 \\ 6 & 1 \\ 7 & 1 \\ 8 & 1 \\ 9 & 1 \\ 10 & 1 \end{bmatrix}$$

Therefore we compute

$$\hat{\beta} = (M^TM)^{-1}M^T\mathbf{y} = \begin{bmatrix} 1.04492 \\ 1.97878 \end{bmatrix}$$

and

$$\hat{\mathbf{y}} = M\hat{\beta} = \begin{bmatrix} 1.97878 \\ 3.0237 \\ 4.06862 \\ 5.11353 \\ 6.15845 \\ 7.20337 \\ 8.24828 \\ 9.2932 \\ 10.3381 \\ 11.383 \\ 12.428 \end{bmatrix}$$

I did this in Mathematica (I will put the notebook on my website).

The $SSR = \|\mathbf{y} - \hat{\mathbf{y}}\|^2 = 11.506$

Let's get confidence intervals: first, let $c = (1, 0)$. We know that $\frac{c\beta - c\hat{\beta}}{s\sqrt{c(M^TM)^{-1}c^T}} \sim t_{n-r}$, where r here is 2, and $s = \sqrt{SSR/(n-r)} = \sqrt{11.506/(11-2)} = 1.13$. Using Mathematica (or noting that $c(M^TM)^{-1}c^T$ is just the upper left element of $(M^TM)^{-1}$), we find that $\sqrt{c(M^TM)^{-1}c^T} = 0.0953$, and thus

$$\begin{aligned} \frac{c\beta - c\hat{\beta}}{1.13 \times 0.0953} \sim t_{n-r} \quad &\Rightarrow \quad P\left(t_{9,0.025} \leq \frac{\beta_1 - \hat{\beta}_1}{0.107806} \leq t_{9,0.975}\right) = 0.95 \\ &\Rightarrow \quad P\left(\hat{\beta}_1 + (0.107806)t_{9,0.025} \leq \beta_1 \leq \hat{\beta}_1 + (0.107806)t_{9,0.975}\right) = 0.95 \end{aligned}$$

And thus a 95% confidence interval for β_1 is $\hat{\beta}_1 \pm (0.107806)t_{9,0.975}$. Plugging in values, this is $1.04492 \pm (0.107806)t_{9,0.975} = [0.801042, 1.28879]$.

Following an identical calculation, we find a 95% confidence interval for β_2 of $[0.6219, 3.336]$. Thinking about this for a moment in the context of this being an example problem, you can guess that $\beta_1 = 1$, and β_2 is 1, 2, or 3, and in fact, checking my Mathematica notebook shows how I generated the data and that this is true.

We can also get the standard error of, for example, $\hat{\beta}_1 + \hat{\beta}_2$. We know

$$s^2_{\hat{\beta}_1 + \hat{\beta}_2} = c(M^TM)^{-1}c^T s^2 = 0.2363 \times 1.13 = 0.267$$

Where $c = (1, 1)$.

Let's calculate how good a fit this is, generally. We calculate the coefficient of multiple determination:

$$R^2 = \frac{s_y^2 - s_\epsilon^2}{s_y^2} = \frac{13.161 - 1.15}{13.161} = 0.91258$$

So that's a pretty good fit.

Now suppose that we make another observation $\mathbf{y}_{n+1}$ at 11, so set $c = (11, 1)$. Now we know

$$\frac{\mathbf{y}_{n+1} - c\hat{\beta}}{s\sqrt{1 + c(M^TM)^{-1}c^T}} \sim t_{n-r}$$

So we compute $c\hat{\beta} = 13.47$ and $s\sqrt{1 + c(M^TM)^{-1}c^T} = 1.3457$, so we know

$$P\left(t_{9,0.025} \leq \frac{\mathbf{y}_{n+1} - 13.47}{1.3457} \leq t_{9,0.975}\right) = 0.95$$

And thus a 95% confidence interval for the new observation is:

$$13.47 \pm 1.3457 t_{9,0.975} = [10.426, 16.514]$$

2. χ^2 TESTS

2.1. **Goodness-Of-Fit Tests.** You can use a χ^2 test to see if data is distributed according to a given distribution. The assumption here are that your data is in bins. Then you do the following.

- Let the null hypothesis be that the data is normal (or whatever distribution)
- Compute the expected number E_i for each bin, given that the distribution is normal (or whatever) with the sample mean and stddev (or the given mean and stddev, in the case of question 5)
- Find the $\sum_{i=0}^{n}(O_i - E_i)^2/E_i$, where O_i is the observed number in the bin i.
- That sum is the chi-square statistic X^2, and we assume/it is distributed as a chi-square distribution with $n - m - 1$ degrees of freedom, where m is the number of parameters being estimated (in the normal case, 2)
- Find the probability that a chi-square random variable with $n - m - 1$ degrees of freedom is larger than X^2. This is the p-value, and thus for any $\alpha > p$, we reject the null hypothesis that the data is normal.

2.1.1. *Example.* This is copied off p.344 of the textbook we used to use, because it's a good example that's not the normal distribution. Suppose you have a bunch of bacteria clumps, and you think that the number of clumps per square is Poisson. Well, let's use a χ^2 test. Here is the data:

Number per square	0	1	2	3	4	5	6	7	8	9	10	19
Frequency	56	104	80	62	42	27	9	9	5	3	2	1

The sample mean here is just the average $(0 \times 56 + 1 \times 104 + \cdots + 19 \times 1)/400 = 2.44$. Under this mean, we would expect the following counts:

Number per square	0	1	2	3	4	5	6	≥ 7
Observed	56	104	80	62	42	27	9	20
Expected	34.9	85.1	103.8	84.4	51.5	25.1	10.2	5.0
$\frac{(O-E)^2}{E}$	12.8	4.2	5.5	5.9	1.8	0.14	0.14	45.0

The total statistic is $X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 75.4$. We have 6 degrees of freedom because there are 8 bins and we are estimating one parameter (the mean of the Poisson), so $8 - 1 - 1 = 6$. Then we find that if $Y \sim \chi^2_6$, we have $P(Y > 75.4) \approx 3.1 \times 10^{-14}$, so pretty much there's no chance that the data is Poisson. This is interesting because looking at a plot of the data, you might think it was. You can see how it fails the test by looking that the bottom row of the table—notice that the main failure is that there were too many observed counts that were high and low.

2.2. **Test of Independence.** Section 13.4 in your book is good reading.

The idea here is to take a table of data and decide whether or not the two axis labels (like, eye color and hair color) are independent. The null hypothesis is that they are indpendent. See p.522 for step by step instructions here. If you let n_{ij} be the matrix of observations with I rows and J columns, then the chi-square statistic is:

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n}$$

where $n_{i\cdot} = \sum_{j=1}^{J} n_{ij}$, and likewise for $n_{\cdot j}$, and n is the total number of observations.

The degrees of freedom of this is $(I-1)(J-1)$. Given a table of data, you just compute all that stuff, and see what the probability that a chi-square statistic with $(I-1)(J-1)$ degrees of freedom is larger than X^2, and this gives you a p-level. You then reject the null hypothesis that the two axes are independent at significance level α for any $\alpha > p$.

2.2.1. *Example.* Here is a data chart that I completely made up from a study in which researchers told subjects that gullible was written on the ceiling and then observed whether or not they looked up (and a control group in which they didn't tell them anything):

	Gullible on the ceiling	control (not on ceiling)
Looking up	12	20
Not looking up	38	30

Ok now we compute a whole bunch of stuff:

$$\begin{aligned}
n_{1\cdot} &= \sum_{j=1}^{2} n_{1j} = 32 \\
n_{2\cdot} &= 68 \\
n_{\cdot 1} &= 50 \\
n_{\cdot 2} &= 50 \\
X^2 &= \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n)^2}{n_{i\cdot}n_{\cdot j}/n} \\
&= \frac{(12 - (32\times 50)/100)^2}{(32\times 50)/100} + \frac{(20 - (32\times 50)/100)^2}{(32\times 50)/100} + \frac{(38 - (68\times 50)/100)^2}{(68\times 50)/100} + \frac{(30 - (68\times 50)/100)^2}{(68\times 50)/100} \\
&= 2.94118
\end{aligned}$$

We have $(2-1)(2-1) = 1$ degree of freedom, so we have a p-level of 0.0863, meaning that for any level of significance greater than that, we will reject the null. For instance, at an α of 0.1, we would reject.

Note that rejecting means that we reject the hypothesis of independence, but we do not conclude anything about the structure of the dependence. For instance, I made up this example with the idea that telling someone that gullible is on the ceiling probably causes them to not look up more often, since they don't want to look gullible, so it's sort of a negative dependence.