MA 2B RECITATION 03/01/12

1. Linear Regression

Linear regression is probably one of the most important things to come out of the study of linear algebra. However, it is deceptively named. You can use "linear" regression to fit not just lines, but even polynomials or functions or even generalized functions, if you want to get fancy.

Nothing here is too complicated. You could have done all of this last year in Math 1b. Here's the setup.

Suppose that we have n data points $(x_1, y_1), \ldots, (x_n, y_n)$. We want to find a line $\beta_1 x + \beta_0 1$ that best fits our data. Namely, we want to find β_i such that

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

Let M denote the left-most matrix.

However, in most cases, we can't solve this precisely, because the system is overdetermined. Namely, the *image* of M has dimension at most 2, but the *range* may have arbitrarily many dimensions.

So the world is not perfect and life is not easy. But we want to find a β vector that is as close as possible to the y-vector. However, linear algebra has already provided given us the answer to this question. Namely, the closest vector is the *projection of* $\boldsymbol{y}$ *onto the column space of* M. This is essentially the definition of projection onto a vector, but if you've forgotten what it means, look it up! We even have a nice formula for the projection, which is great, because it would be tedious to try and compute it by hand (e.g. by an explicit Gram-Schmidt computation).

The formula for the projection $\hat{\mathbf{y}}$ is

$$\hat{\mathbf{y}} = M(M^T M)^{-1} M^T \mathbf{y}.$$

Therefore, the formula for the coefficients of the projection with respect to the basis given by the columns of M is

$$\hat{\beta} = (M^T M)^{-1} M^T \mathbf{y}.$$

Remark 1.1. It is important to state that all of these formulas are in the case of vector spaces over $\mathbf{R}$. There are slightly different formulas over other base fields. For instance, if the vector spaces were over the complex numbers $\mathbf{C}$, we would want to use the adjoint matrix M^* instead of the transpose M^T.

OK. That's great, Brian. But why are we learning this now instead of as a homework exercise last year? It's because the right way to view this situation is by interpreting the situation probabilistically, with distributions.

Date: March 1, 2011.

1.1. **The Probabilistic Viewpoint.** The way you want to think about this situation above is to imagine that these some vector β that will solve our equation exactly, but that some trickster (like *Mother Nature* or *entropy*) has inserted $N(0, \sigma^2)$ errors into the observation vector $\mathbf{y}$. So you figured this out, but you still want to know: How close did the observed coefficients $\hat{\beta}$ come to the true coefficients?

Suppose that we have n observations and r regression coefficients (i.e. the length of the β-vector). Then

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

where $s = \sqrt{\frac{RSS}{n-r}}$ and c is any vector of length r. (Recall RSS = "residual sum of squares" $= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$.) For instance, if we wanted to find out about the first entry β_0, we might want to use $c = (1, 0, \dots 0)$.

Remark 1.2. Sometimes residual sum of squares (RSS) is also called the sum of squared residuals (SSR). They are the same thing.

I'm going to give a rundown of some of the more important facts and definitions. Refer to the book or to the posted lecture notes for details. (I especially recommend the notes on the class website.)

We can use the SSR to produce an unbiased estimator of the error variance σ^2:

$$\hat{\sigma^2} = s^2 = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{n-r}.$$

We have

$$E(c\hat{\beta}) = c\beta$$

and so

$$c\beta \sim N(c\beta, c(M^TM)^{-1}c^T\sigma^2).$$

Another good thing to know is that

$$\mathrm{Var}(c\hat{\beta}) = c(M^TM)^{-1}c^T\sigma^2.$$

It is so useful that we make it a definition.

Definition 1.3. The **standard error** of an estimate $c\hat{\beta}$ is $c(M^TM)^{-1}c^Ts^2$. (Note the s instead of σ.)

The standard error of one the coefficients β_i is just the standard error of $c\hat{\beta}$ where $c = e_i$, that is, where c has a 1 in the ith coordinate and zeroes elsewhere.

Following the convention in the book, we will usually write s_x^2 for the variance of x and $s_{\hat{\beta}_i}^2$ for the standard error or $\hat{\beta}_i$. Therefore, the coefficient of multiple determination in the book is

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

In other words, we take the variance of y, subtract the variance of the observed errors, and then divide this difference by the variance of y.

We often refer to something called a **standard statistical model**, and by this, we just mean a model that has the following two properties: (1) the equation is linear and (2) the errors are of mean zero and have the same variance.

Given our information, we can also guess a new observation. Suppose that we want to know $\mathbf{y}_{n+1} = c\beta$. The obvious guess for this is just $c\hat{\beta}$ itself, but how good is this guess? We have

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

and so we can use this to determine confidence intervals or run hypothesis tests.

Example 1. We want to fit a polynomial of degree 2 to the data points:

x	y
1	3.92296
2	9.107
3	20.1724
4	37.0409
5	60.0193
6	89.0312
7	124.228
8	165.182
9	211.967
10	264.492

Our model here is $y = \beta_1 x^2 + \beta_2 x + \beta_3$. Our matrix here is

$$M = \begin{bmatrix} 1 & 1 & 1 \\ 4 & 2 & 1 \\ 9 & 3 & 1 \\ 16 & 4 & 1 \\ 25 & 5 & 1 \\ 36 & 6 & 1 \\ 49 & 7 & 1 \\ 64 & 8 & 1 \\ 81 & 9 & 1 \\ 100 & 10 & 1 \end{bmatrix}$$

with the first column corresponding to β_1, the second to β_2, and the third to β_3. However, note that linear regression doesn't "know" about the difference between x^2 and x columns, even though the pattern is obvious to us. We then calculate our guesses

$$\hat{\beta} = (M^T M)^{-1} M^T \mathbf{y} = \begin{bmatrix} 2.98209 \\ -3.82872 \\ 4.76363 \end{bmatrix} \quad \hat{\mathbf{y}} = \begin{bmatrix} 3.917 \\ 9.03456 \\ 20.1163 \\ 37.1622 \\ 60.1728 \\ 89.1466 \\ 124.085 \\ 164.988 \\ 211.855 \\ 264.686 \end{bmatrix}$$

Therefore, the RSS is

$$RSS = \frac{||\mathbf{y} - \hat{\mathbf{y}}||^2}{10 - 3} = 0.155078.$$

Setting $c = [1, 0, 0]$, we get $s\sqrt{c(M^TM)^{-1}c^T} = 0.00674889$, so a 90% confidence interval for β_1 is

$$\hat{\beta}_1 \pm 0.00674889 t_{7,0.95} = [2.96931, 2.99488]$$

and similarly, we calculate the one for β_3 as

$$\hat{\beta}_3 \pm 0.182395 t_{7,0.95} = [4.41807, 5.10919].$$

Note that the interval for β_3 is larger, because β_1 affects the output value by much more, so we guess it more accurately.

The true polynomial for this data is actually $3x^2 - 4x + 5$. Thus, we see that our 90% confidence interval is wrong! This does indeed give a 90% interval, but such an interval has no guarantee of getting the right answer.

2. Chi-Squared (χ^2) Test

We use a χ^2-test to test to see if the data does indeed follow a given distribution. There are two main uses for this: to test for goodness of fit, and to test for independence.

2.1. **Goodness of Fit.**

You first want to divide your data into bins, following whatever scheme is most appropriate. We then apply the following method.

(a) Set H_0 as "data is normal" (or Poisson or exponential, etc.) We'll assume normal here, but you can replace all subsequent instances of "normal" with any other distribution.

(b) Given that the distribution is normal, compute the expected number E_i for for each bin with sample mean and standard deviation.

(c) Find the value

$$X^2 = \sum_{i=0}^{n} \frac{(O_i - E_i)^2}{E_i}$$

where O_i is the observed number in the bin i. The value X^2 is called the **chi-square statistic** and we assume that it is distributed as a chi-square distribution with $n - m - 1$ degrees of freedom, where m is the number of parameters being estimated. For instance, with a normal distribution, $m = 2$.

(d) Find the probability that a chi-squared 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.

Example 2. This is the example from p. 344 in the book, chosen because it's a good example of something that doesn't follow the normal distribution. (You have questions on the homework for the normal case.)

Suppose we have clumps of bacteria, and believe that the number of clumps per square is Poisson. We'll try and use a χ^2 test to verify this.

No. 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 in this case is the average

$$\frac{0 \cdot 56 + 1 \cdot 104 + \cdots + 19 \cdot 1}{400} = 2.44$$

Under this mean, we expect the following counts

No. per square	0	1	2	3	4	5	6	≥ 7
Observed	56	104	80	62	42	27	9	20
Expected	39.4	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 test statistic is

$$X^2 = \sum \frac{(O_i - E_i)^2}{E_i} = 75.4.$$

Since there are 8 bins and we are estimating one parameter (the mean of the Poisson), we have $8 - 1 - 1 = 6$ degrees of freedom. If $Y \sim \chi^2_6$, then

$$P(Y > 75.4) \approx 3.1 \times 10^{-14}$$

so it is extremely unlikely that this data is Poisson. However, if you just look at the table, you might believe that the data is Poisson, because the numbers aren't particularly far apart especially near the mean. Nevertheless, there were too many observed counts at the higher and lower ends that deviated from a Poisson-like distribution, and so affected our final conclusion.

2.2. **Test of Independence.** What we want to do here is to take a table of data and decide whether the two axis labels (like eye color and hair color on your homework set) are independent. The null hypothesis is that they are independent. There are step-by-step instructions on page 522, but here is the gist of the argument. If we let n_{ij} be the matrix of observations with I rows and J columns, then the chi-square statistic in this case is

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

where $n_{i*} = \sum_{j=1}^{J} n_{ij}$ sums over all the j or a fixed i, and similarly for n_{*j}, and where n is the total number of observations.

Here, we have $(I-1)(J-1)$ degrees of freedom. Given a table of data, we just compute the above things and see what the probability that a chi-squared 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 $\alpha > p$.

Example 3. We have the following table of data of heights of father and child, split into three categories: short, average, and tall.

	child short	average	tall
father short	14	11	8
average	11	11	9
tall	6	10	12

We want to know: are the heights of fathers and their children independent? We calculate

$$X^2 = 3.81436.$$

Since $I = J = 3$, we have 4 degrees of freedom. Our p-value is $1 - F(3.81436)$, where F is the CDF of the χ^2 distribution. It turns out the p-value is 0.431712. This is not small, so we probably should not reject the hypothesis that the father's height and the child's height are independent.