An introduction to biostatistics: part 2

Cavan Reilly

September 4, 2019

Table of contents

Statistical models

Statistical models can be powerful tools for understanding complex relationships among variables.

We will first suppose that we observe 2 continuous variables.

Typically we would start out by looking at a scatterplot: so let's look at an example.

Let's read in some data from the genetic association study we looked at in the previous lecture:

```
fms=read.delim("http://www.biostat.umn.edu/~cavanr/FMS_data.txt")
```

Statistical models

Then we can just use the `plot` command as follows.

```
> pdf("wght-hght.pdf")
> plot(fms$Pre.height,fms$Pre.weight,xlab="Height",ylab="Weight")
> dev.off()
```

It looks like the 2 variables increase together, but we clearly don't have an equation like:

$$\text{Weight} = \beta_0 + \beta_1 \text{Height},$$

for 2 constants β_0 and β_1.

Note: we look at the distribution of the response variable conditional on the explanatory variable.

Correlation

A commonly used measure of the extent to which 2 continuous variables have a linear association is the correlation coefficient.

The `cor` function in R allows one to compute this summary.

If it is positive then large values of one variable are generally associated with large values of the other.

If it is negative then large values of one variables are associated with small values of the other.

If the absolute value of the correlation coefficient exceeds 0.7 then there is a strong association, if less than 0.3 then weak, otherwise moderate.

Statistical models

A statistical model adds a random variable to this equation so that it can be a true equality

$$\text{Weight} = \beta_0 + \beta_1 \text{Height} + \text{Error}.$$

To make this more precise, let y_i represent the weight of subject i and x_i that subject's height, then our model is

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

for $i = 1, \ldots, n$.

We observe y_i and x_i, so given any values for β_0 and β_1, ϵ_i is something we can compute.

β_0 and β_1 are called *regression coefficients*.

Statistical models

Some values of β_0 and β_1 will result in larger ϵ_i than others: in ordinary least squares regression we choose β_0 and β_1 to minimize

$$\sum_i (y_i - \beta_0 - \beta_1 x_i)^2$$

Once the problem is formulated in this fashion we can use calculus to get expressions for β_0 and β_1 that just depend on the data.

In R we can get these estimates by typing

```
> m1=lm(fms$Pre.weight~fms$Pre.height)
```

Statistical models

```
> m1

Call:
lm(formula = fms$Pre.weight ~ fms$Pre.height)

Coefficients:
    (Intercept)  fms$Pre.height
       -181.377           5.039
```

Statistical models

We can add these to our figure to see how well this works

```
> pdf("wght-hght.pdf")
> plot(fms$Pre.height,fms$Pre.weight,xlab="Height",ylab="Weight")
> abline(m1$coef)
> dev.off()
```

So that looks reasonable, but raises the question: if I have a statistical model, how do I "get values for the unknowns"?

A *parameter* is an unknown number in a statistical model.

Statisticians would formulate the "get a value for" a parameter, as "how to estimate" a parameter.

Statistical models

The statistical models we used in the previous lecture were so simple that we didn't even dwell on their existence.

But we still had models: for example, if I measure a dichotomous random variable I might be assuming that everyone I measure has the same chance of a success.

If I didn't think that was the case then why would we summarize those data with sample proportions from all subjects?

So in fact we had simple models in mind and these were very much tied to the way I estimate the parameters in those models.

Maximum likelihood estimation

A number of general approaches to the problem of how to estimate parameters in statistical models have been developed.

We will mostly discuss and use the *method of maximum likelihood*.

The *likelihood* is the probability of observing the data that was observed: one will typically have a model for this.

One then expresses the likelihood in terms of the parameters of the model and maximizes the likelihood with respect to the parameters.

If the sample size is sufficiently large the parameter estimates one gets using these techniques have a number of desirable features.

Maximum likelihood estimation

Estimators depend on the data, so they are also random variables.

As such we can talk about their expectation and their variance.

If the expectation of a parameter estimate is equal to the actual parameter estimate then we say that the estimator is *unbiased*.

One can mathematically demonstrate that there is a lower bound to the variance that one can obtain using an unbiased estimator.

In fact we have mathematical expressions for the lower bound.

Maximum likelihood estimation

One can demonstrate that as the sample size goes to infinity, maximum likelihood estimators (MLEs) are unbiased and their variance obtains this lower bound.

Moreover they are approximately normal random variables and there are formulae we can use for their variance.

This makes it easy to test if a parameter is equal to some hypothesized value, say θ_0, since, if $\hat{\theta}$ is the MLE and $\text{se}(\theta)$ is its estimated standard deviation we just compute

$$T = \frac{\hat{\theta} - \theta_0}{\text{se}(\theta)}$$

and use that to get a p-value.

We can also get 95% confidence intervals with: $\hat{\theta} \pm 1.96\text{se}(\theta)$.

Maximum likelihood estimation

So for large sample sizes MLEs are about as good estimators as one can expect.

It transpires that many estimators that are quite natural are in fact MLEs for certain models.

For example, if I suppose that I have independent measurements on a Bernoulli random variable, then the sample proportion is the maximum likelihood estimate of the probability of success.

If I assume I have independent measurements from a normal distribution then the MLE of the expectation of that distribution is given by the sample mean.

Maximum likelihood estimation: an example

If I observe x_i where each x_i is zero or one all with the same probability, then I have a collection of independent Bernoulli observations.

Since the observations are independent I can compute the likelihood by taking the product over all subjects, so I just need the likelihood for each subject.

If the probability of success is π this is given by $\pi^{x_i}(1-\pi)^{1-x_i}$, why?

Now maximize the likelihood over all subjects. What do you get?

Maximum likelihood estimation: linear model

It also turns out that if I specify the following model:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

where ϵ_i are independent observations from a normal distribution with mean zero, the maximum likelihood estimates of β_0 and β_1 are the same as what we obtain by ordinary least squares regression.

While this is interesting, the real value of MLEs is that we have standardized tools for statistical inference which can be adapted to new models.

Linear models

Suppose we observe more than one explanatory variable for each subject: then we might entertain models like

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

where ϵ_i is normally distributed as before.

Such models are called *multiple regression models* and have many uses.

For example, suppose I observe race for each subject and in my data set it takes 3 levels (say Caucasian, African American and Asian).

Linear models

I could create an x_i variable that has values 1, 2, and 3 and fit a model like

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

but such a model makes a very strong assumption about the difference in the mean of y_i across the levels of race-what is that assumption?

Linear models

A better approach is to create 2 *indicator variables*, for example, let x_{1i} be 1 if subject i is African American and 0 otherwise and let x_{2i} be 1 if subject i is Asian and 0 otherwise, then

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i$$

is a more general model that allows for distinct differences between the means of the racial groups. For example, restricting attention to these 3 groups:

```
> x1=c(fms$Race=="African Am")
> x2=c(fms$Race=="Asian")
> m1=lm(fms$Pre.height~x1+x2, subset=c(fms$Race=="African Am"
+ | fms$Race=="Asian" | fms$Race=="Caucasian"))
```

Linear models

```
> summary(m1)

Call:
lm(formula = fms$Pre.height ~ x1 + x2, subset = c(fms$Race ==
    "African Am" | fms$Race == "Asian" | fms$Race == "Caucasian"))

Residuals:
     Min       1Q   Median       3Q      Max
-11.0526  -2.7276  -0.0526   2.4474  10.9474

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  67.0526     0.1289 520.349  < 2e-16 ***
x1TRUE       -1.2958     0.5607  -2.311  0.02105 *
x2TRUE       -1.2742     0.3895  -3.272  0.00111 **
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.62 on 927 degrees of freedom
  (365 observations deleted due to missingness)
Multiple R-squared:  0.01591, Adjusted R-squared:  0.01378
F-statistic: 7.492 on 2 and 927 DF,  p-value: 0.0005922
```

Linear models

But couldn't one do that with ANOVA, yes and we get exactly the same result:

```
> a1=aov(fms$Pre.height~fms$Race, subset=c(fms$Race=="African Am"
+ | fms$Race=="Asian" | fms$Race=="Caucasian"))
> summary(a1)
             Df Sum Sq Mean Sq F value   Pr(>F)
fms$Race      2    196   98.15   7.492 0.000592 ***
Residuals   927  12145   13.10
```

So multiple regression generalizes ANOVA (and 2 sample t-tests) by allowing us to consider categorical and continuous predictors simultaneously.

How would you do a t-test using this sort of approach.

Confounding

Such models also allow us to potentially understand complex relationships between more than 2 variables simultaneously.

When trying to deduce the nature of the association between 2 variables, a confounding variable is one that is related to both.

Such variables can make it difficult to interpret pairwise associations: in particular they can make what appears to be an association go away, or even change direction.

We try to *statistically control* for the effect of such variables by incorporating them in regression models.

Randomization is the best way to avoid confounders.

Statistical adjustments

In observational studies we often want to make comparisons between groups that differ in some way.

Ebola survivor study: we have survivors and contacts, but the contacts are younger.

It appears eyesight is worse in survivors, but this could be due to age.

I can fit regression models with survivorship status and age as covariates and eyesight as the response variable.

Statistical adjustments

Let y_i represent visual acuity, x_{1i} represent age and let x_{2i} represent an indicator for survivorship status.

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \epsilon_i.$$

If $\hat{\beta}_i$ are the estimated regression coefficients then we could compare eyesight for someone with the median age in the study (28)

$$\hat{\beta}_0 + \hat{\beta}_1 \times 28$$

to

$$\hat{\beta}_0 + \hat{\beta}_1 \times 28 + \hat{\beta}_2$$

With such a model I could make predictions about the mean sight of a survivor and a contact who are the same age.

I can also construct confidence intervals: contrasts.

Contrasts

As age adjustments make clear, we are frequently interest in *linear combinations* of the regression coefficients.

If we use β to represent the vector of regression coefficients, then the last application had us examine the inner product of this vector with the vectors

$$c_1 = (1, 28, 1)$$

$$c_2 = (1, 28, 0).$$

The regression coefficients have a *covariance matrix*: this is a table with as many rows and columns as there are regression coefficients.

Contrasts

Each element in the table either has the variance of a regression coefficient (on the diagonal) or the covariance between 2 distinct regression coefficients.

The covariance between 2 variables is the correlation between them times the standard deviations of the 2 variables.

Represent this table (i.e. matrix) with $\text{Var}\beta = \Sigma_\beta$.

Contrasts

We had computed $T = c_1\beta$: it turns out that the variance of this expression is given by

$$\text{Var}\, T = c_1 \Sigma_\beta c_1^T,$$

where c^T represents the transpose of a vector or matrix.

Contrasts arise in other settings: suppose we wanted to test if the heights of Asians are different from the heights of African Americans using our previous regression model.

Contrasts

I can compute the difference in the means as follows

```
> c(0,1,-1)%*%m1$coef
              [,1]
[1,] -0.02153233
```

To get at the variance of this expression note that this gives us the standard errors of the regression coefficients

```
> summary(m1)$sigma*sqrt(diag(summary(m1)$cov.unscaled))
(Intercept)      x1TRUE      x2TRUE
  0.1288608   0.5606824   0.3894500
```

Contrasts

So the covariance matrix is given by

```
> (summary(m1)$sigma^2)*summary(m1)$cov.unscaled
```

and we can get a test statistic with the following

```
> c(0,1,-1)%*%m1$coef/sqrt((summary(m1)$sigma^2)*c(0,1,-1)%*%
+ summary(m1)$cov.unscaled%*%c(0,1,-1))
             [,1]
[1,] -0.03272913
```

and so get a 2 sided p-value by

```
> 2*(1-pt(0.03272913,df=927))
[1] 0.9738976
```

Statistical interaction

Again consider the regression model that motivated the discussion of contrasts: visual acuity is the response variable and age and survivor status are predictors.

That model assumes that the association between age and visual acuity is the same in both survivor groups, in contrast

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i} + \epsilon_i$$

is a model that allows the slopes that describe the relationship between age and visual acuity to differ between survivors and contacts.

Such a model has a statistical interaction between age and survivor status.

Generalized linear models

Linear models have been generalized in a number of ways.

One way to think about a linear model is in terms of relating the mean structure of a random variable to a collection of other variables.

$$y_i \sim \mathrm{N}(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi}, \sigma^2).$$

This form makes clear how we could allow for something other than normally distributed data.

Logistic regression

For example, suppose I observe a binary response variable for each subject y_i and so

$$y_i \sim \text{Ber}(\pi_i),$$

where π_i is the probability of success for subject i.

If I then assume that for some function g,

$$g(\pi_i) = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{pi},$$

then I have a model where the success probability is impacted by a collection of other variables.

Logistic regression

If I suppose that $g(x)$ has the form

$$g(x) = \log\left(\frac{x}{1-x}\right)$$

then we are using a technique called *logistic regression*.

We typically estimate the parameters in this model using the method of maximum likelihood.

Logistic regression

The likelihood has the form

$$\prod_i \pi_i^{y_i}(1-\pi_i)^{1-y_i}$$

and if we invert g we find that

$$\pi_i = \frac{e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi}}}{1+e^{\beta_0+\beta_1 x_{1i}+\cdots+\beta_p x_{pi}}}.$$

We then use numerical techniques to maximize this expression with respect to the set of regression coefficients.

Logistic regression

These numerical techniques can sometimes fail and there is no guarantee that a unique collection of estimated regression coefficients will in general exist.

Suppose there is only 1 predictor variable and it is binary: then we can show that regression coefficient is the log of the odds ratio across the 2 levels of the binary predictor.

Logistic regression: an example

Here is the syntax that is necessary to conduct a logistic regression analysis.

```
> m1=glm(fms$Met_syn~fms$pre.BMI>30,family=binomial)
```

Previously we found that the odds ratio was 13.68.

Logistic regression: an example

```
> summary(m1)

Call:
glm(formula = fms$Met_syn ~ fms$pre.BMI > 30, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.0502  -0.3237  -0.3237  -0.3237   2.4394

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)           -2.9229     0.1688 -17.320   <2e-16 ***
fms$pre.BMI > 30TRUE   2.6161     0.2702   9.684   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Logistic regression: an example

and we again find a huge odds ratio

```
> exp(2.6161)
[1] 13.68226
```

We can also see if there are SNPs associated with metabolic disorder:

```
> table(fms$akt2_969531)

 AA  GA  GG
 98 495 614
```

```
> m1=glm(fms$Met_syn~c(fms$pre.BMI>30)+fms$akt2_969531+
+ fms$Gender,family=binomial)
```

Logistic regression: an example

```
> summary(m1)

Call:
glm(formula = fms$Met_syn ~ c(fms$pre.BMI > 30) + fms$akt2_969531 +
    fms$Gender, family = binomial)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.5128  -0.4557  -0.2373  -0.2049   3.1427

Coefficients:
                              Estimate Std. Error z value Pr(>|z|)
(Intercept)                    -4.9309     0.7291  -6.763 1.35e-11 ***
c(fms$pre.BMI > 30)TRUE         2.6768     0.2960   9.043  < 2e-16 ***
factor(fms$akt2_969531)GA       1.3751     0.6928   1.985   0.0472 *
factor(fms$akt2_969531)GG       1.0784     0.6890   1.565   0.1176
fms$GenderMale                  1.6399     0.2959   5.542 3.00e-08 ***
```

So hard to tell if this SNP is associated: we can fit a model without the SNPs and compare fits

Logistic regression: an example

```
> m0=glm(fms$Met_syn~c(fms$pre.BMI>30)+fms$Gender,family=binomial,
+ subset=!is.na(fms$akt2_969531))
> anova(m0,m1)
Analysis of Deviance Table

Model 1: fms$Met_syn ~ c(fms$pre.BMI > 30) + fms$Gender
Model 2: fms$Met_syn ~ c(fms$pre.BMI > 30) + fms$akt2_969531 +
    fms$Gender
  Resid. Df Resid. Dev Df Deviance
1       785     378.70
2       783     373.55  2   5.1505
> 1-pchisq(5.1505,df=2)
[1] 0.07613479
```

So not quite significant.

Logistic regression: an example

We had included 2 covariates that are strongly related to our outcome variable.

What happens if we had excluded these?

Do those results make sense?

Model selection

Choosing which variables should be in a model and how those variables should enter are difficult problems.

Automated model fitting algorithms are popular but they are frequently not very useful.

When analyzing studies that don't employ randomized treatment assignment, you can never really be sure which are the relevant variables.

Trying to find the "best model" is probably not a good use of one's time: better to fix a data analysis plan prior to conducting the data analysis.

I devise such plans when the experiment is being designed: I strongly suggest you do too.