Stratification

Patrick Breheny

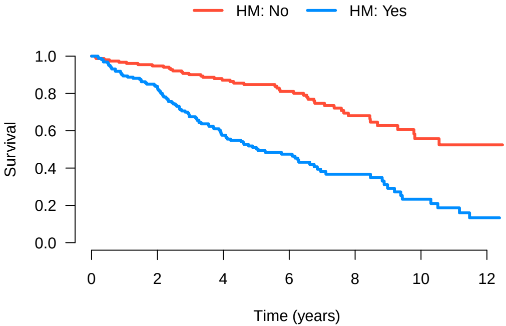
September 19

Introduction

- This is a short lecture on the idea of *stratified* analyses
- A stratified analysis is one in which the data set is broken down into multiple, more homogeneous subsets and the analysis repeated in each subset
- This is often done because one is worried about confounding factors, or because a treatment might be more effective in one group than another

Hepatomegaly in the PBC data

- As an example, let's examine whether an enlarged liver (hepatomegaly) is associated with survival using the `pbc` cholangitis data set
- In a marginal analysis, the two are clearly associated ($p = 4 \times 10^{-11}$):



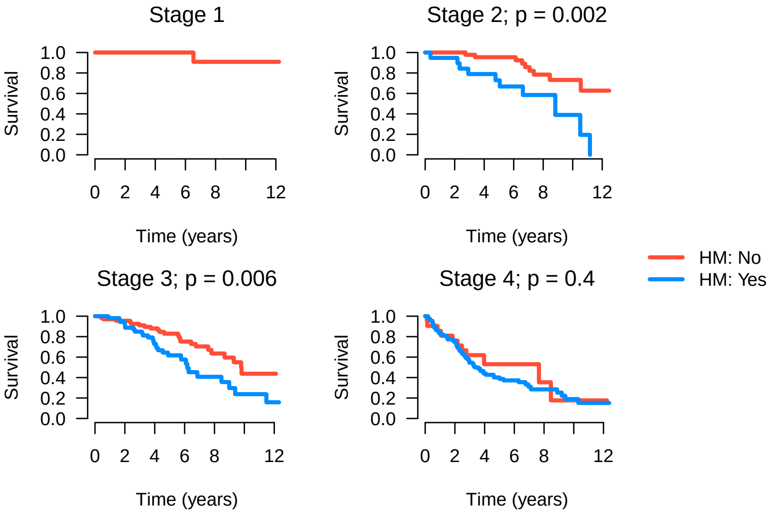
Possible confounding?

Hepatomegaly, however, is strongly correlated with stage:

	1	2	3	4
	Stage			
Hepatomegaly: No	16	48	67	21
Hepatomegaly: Yes	0	19	53	88

Perhaps this could be driving the association?

Stratified results



Remarks

- Thus, we have a significant association for stages 2 and 3, but not 1 and 4
- Still, the direction of association is consistent: the survival of patients with hepatomegaly is consistently worse than those without it
- It is desirable, then, to come up with a way of pooling these test results across disease stages
- This would yield an overall test of hepatomegaly's association with survival, but would account for the possibly confounding influence of disease stage

Stratified log-rank tests

- Fortunately, this is very straightforward to accomplish with the log-rank test
- Our test statistic already consists of sums across failure times; we can simply add across strata as well:

$$\frac{W^2}{V} = \frac{\left(\sum_k \sum_j w_{jk}\right)^2}{\sum_k \sum_j v_{jk}} \dot{\sim} \chi^2_1,$$

  where w_{jk} denotes the observed minus expected number of failures at the jth failure time within the kth stratum, and v_{jk} is its variance (both of which we have previously derived)
- The same extension applies to the multi-sample case as well

Hepatomegaly results

Stratified log-rank test: Observed vs. expected failures for patients with hepatomegaly

	Observed	Expected	Difference
Stage 1	0	0.0	0.0
Stage 2	10	4.4	5.6
Stage 3	28	18.7	9.3
Stage 4	62	59.4	2.6
Total	100	82.4	17.6

Hepatomegaly results (cont'd)

- Thus, we see an extra 17.6 failures in the hepatomegaly group, compared with a standard error of 5.1
- Therefore, our test of association is still significant, with $p = 0.001$, but the association is far less dramatic once we adjust for the effect of stage

The strata() function

The `survdiff` function accommodates stratified log-rank tests through the use of a `strata()` function:

```
> survdiff(S ~ hepato + strata(stage), pbc)

           N Observed Expected (O-E)^2/E (O-E)^2/V
hepato=0 152       44     61.6      5.03        12
hepato=1 160      100     82.4      3.76        12

 Chisq= 12  on 1 degrees of freedom, p= 0.000541
```

The power of stratified tests

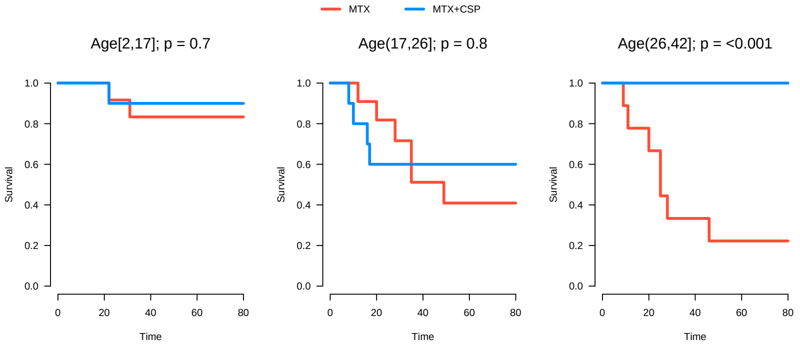
- In the previous example, the p-value for stratified log-rank test was much less significant than the one from the original analysis
- However, this is due to confounding, not the fact that stratified log-rank tests are inherently low-powered
- For example, let's consider our GVHD study
- Here, the MTX/MTX+CSP assignment was random, so confounding shouldn't be an issue

LAF example

- Nevertheless, the data set contains information on whether or not the patient was assigned to a laminar airflow isolation room
- Restricting laminar airflow helps to maintain a sterile environment, which may help reduce the risk of GVHD
- Stratifying on LAF, however, produces a p-value of 0.02; this is exactly the same result we obtained earlier from the non-stratified test

GVHD: Age example

As a final example, consider stratifying the GVHD analysis on age:



Stratified test?

- We can combine these results with a stratified test to get $p = 0.01$, again very close to the original result of 0.02
- However, this approach may not be appropriate here: should we report a single overall effect when the stratified analysis suggests a large benefit for older patients and little to no benefit for children?
- On the other hand, maybe we're reading too much into small samples. . .

Conclusions

- Stratified tests are useful ways of carrying out tests of an overall effect while allowing for the possibly confounding effects of other variables
- The log-rank test is easily extended to allow strata
- The obvious limitations of stratified analyses are that they do not accommodate continuous factors, and do not allow the simultaneous analysis of multiple factors
- For that, we need regression models, which is what we will focus on for the rest of the course