Case study: leukemia data

The four eras of data

From Jeff Leek 2016/12/16 simply statistics post:

https://simplystatistics.org/2016/12/16/the-four-eras-of-data/

The three eras of statistics from Efron (2012) book:

- The age of huge census-level data sets were brought to bear on simple but important questions: Are there more male than female births? Is the rate of insanity rising?
- The classical period of Pearson, Fisher, Neyman, Hotelling, and their successors, intellectual giants who developed a theory of optimal inference capable of wringing every drop of information out of a scientific experiment. The questions dealt with still tended to be simple — *Is treatment A better than treatment B?* — but the new methods were suited to the kinds of small data sets individual scientists might collect.
- The era of scientific mass production, in which new technologies typified by the *microarray* allow a single team of scientists to produce *high-dimensional data*. But now the flood of data is accompanied by a deluge of questions, perhaps thousands of estimates or hypothesis tests that the statistician is charged with answering together.

Jeff Leek breakdown goes like this:

1. **The era of not much data** Prior to about 1995, usually we could collect a few measurements at a time. The whole point of statistics was to try to optimaly squeeze information out of a small number of samples - so you see methods like maximum likelihood and minimum variance unbiased estimators being developed.
2. **The era of lots of measurements on a few samples** This one hit hard in biology with the development of the microarray and the ability to measure thousands of genes simultaneously. This is the same statistical problem as in the previous era but with a lot more noise added. Here you see the development of methods for **multiple testing** and **regularized regression** to separate signals from piles of noise.
3. **The era of a few measurements on lots of samples** This era is overlapping to some extent with the previous one. Large scale collections of data from EMRs and Medicare are examples where you have a huge number of people (samples) but a relatively modest number of variables measured. Here there is a big focus on statistical methods for knowing how to model different parts of the data with hierarchical models and separating signals of varying strength with model calibration.
4. **The era of all the data on everything** This is an era that currently we as civilians don't get to participate in. But Facebook, Google, Amazon, the NSA and other organizations have thousands or millions of measurements on hundreds of millions of people. Other than just sheer computing I'm speculating that a lot of the problem is in segmentation (like in era 3) coupled with avoiding crazy overfitting (like in era 2).

Leukemia data

Let's consider a data set from era 3.

Leukemia data concerns the gene expression measurements on 72 leukemia patients, 47 ALL (acute lymphoblastic leukemia), 25 AML (acute myeloid leukemia, a worse prognosis). These data arise from the landmark Golub et al (1999) Science paper.

Research question

The research question is whether there is any effect of the gene expressions on the clinical outcome (ALL or AML).

Data

Download the data:

```
leukemia <- read.csv("http://web.stanford.edu/~hastie/CASI_files/DATA/leukemia_big.csv")
```

We have normalized gene expression measurements of n subjects for p genes, where

- $y_i \in \{0, 1\}$ is the clinical outcome (0=ALL or 1=AML) for the i-th subject;
- $x_{ij} \in \mathbb{R}$ is the gene expression of the jth gene for the i-th subject.

Prepare the data

```
# y = 0 if ALL, = 1 if AML
y <- (substr(names(leukemia),1,3)=="AML")*1
# gene expression matrix X
X <- t(leukemia)
# n and p
n <- nrow(X)
p <- ncol(X)
colnames(X)  = paste0("X",1:p)
```

Model

Modelling the way in which Y depends on X, we adopt the framework of generalized linear models, which includes logistic regression as a special case.

In the logistic model we assume that

$$\text{logit}\{\mathbb{E}(Y_i)\} = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j$$

where β_0 is the intercept and β_j is the regression coefficient for gene j, $j = 1, \ldots, p$.

The research question translates into testing the null hypothesis that all the coefficients of the logistic model are zero

$$H_0 : \beta_1 = \ldots = \beta_p = 0$$

against the alternative $H_1 : \bigcup_{j=1}^{p} \{\beta_j \neq 0\}$ that at least one coefficient is non-zero.

Analysis

We will use the global test for testing the null hypothesis H_0. This requires to install the R package `globaltest`

```
source("https://bioconductor.org/biocLite.R")
biocLite("globaltest")
```

1. To perform the global test of H_0

```
library(globaltest)
gt(y~X, null=~1, model = "logistic")
```

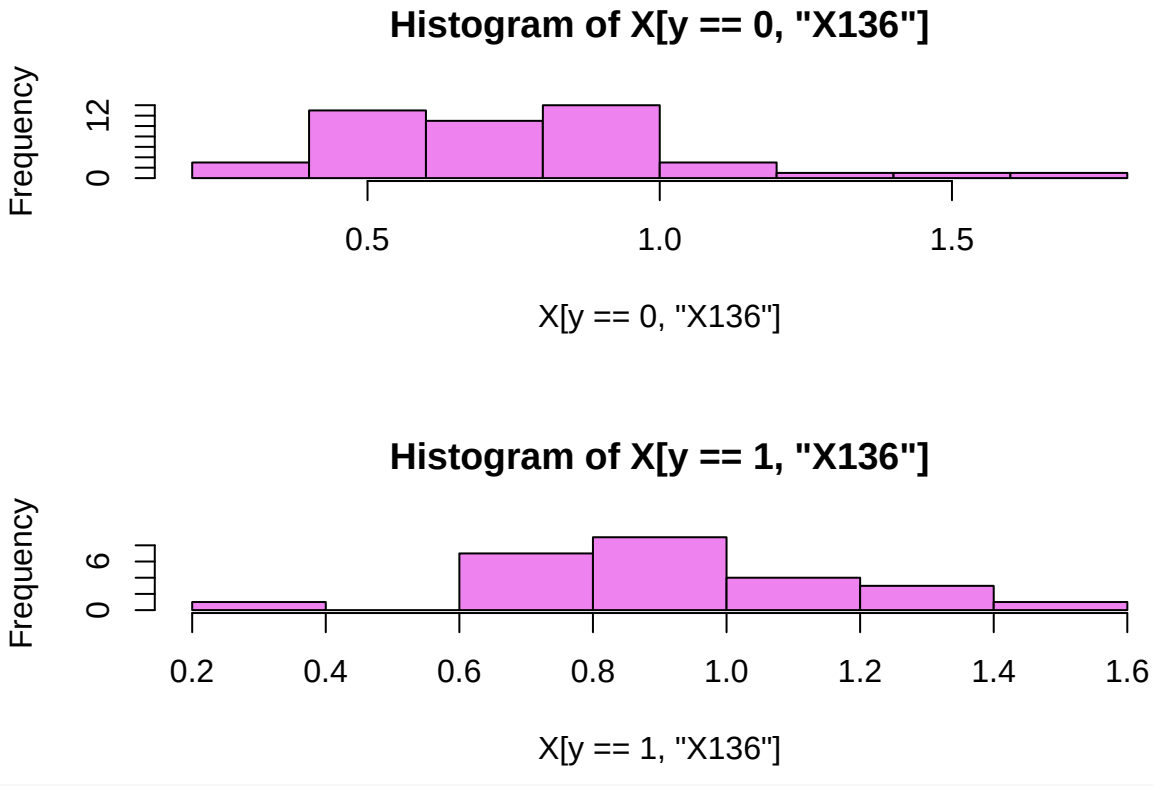
```
##    p-value Statistic Expected Std.dev #Cov
## 1 1.48e-19      6.91     1.41   0.243 7128
```

The result shows that AML and ALL patients do indeed differ with respect the overall gene expression profile.

In the case that the test is not significant, patients with the same clinical outcome do not have very similar gene expression patterns. In this case is unlikely that that a good classification rule to discriminate AML and ALL can be found on the basis of these data.

2. To reproduce Figure 1.4 of Efron and Hastie (2015): scores for gene 136, leukemia data. A two-sample t-statistic $= 3.01$ with p-value $= 0.0036$

```
op <- par(mfrow = c(2, 1))
hist(X[y==0,"X136"], col="violet")
hist(X[y==1,"X136"], col="violet")
```



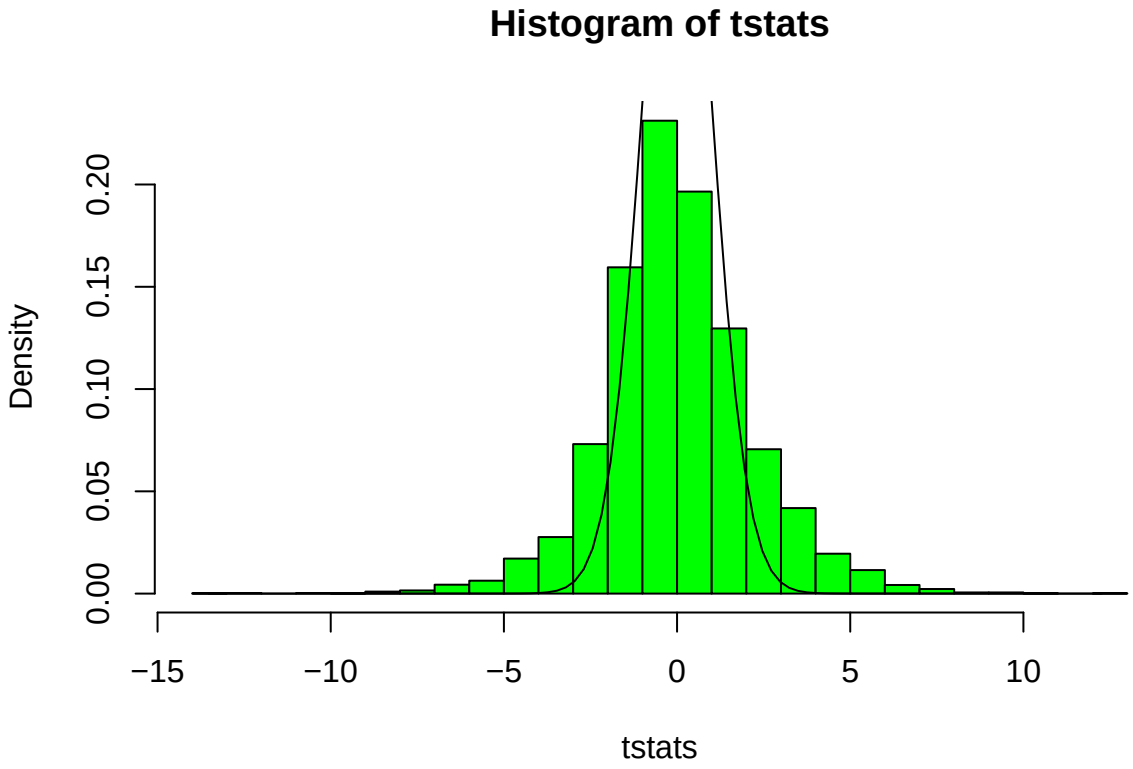

```
par(op)
t.test(X[y==0,"X136"],X[y==1,"X136"], var.equal =T)
```

```
##
```

```
##  Two Sample t-test
##
## data:  X[y == 0, "X136"] and X[y == 1, "X136"]
## t = -3.014, df = 70, p-value = 0.003589
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.32817995 -0.06680742
## sample estimates:
## mean of x mean of y
## 0.7524794 0.9499731
```

3. To reproduce Figure 1.5 of Efron and Hastie (2015): an histogram of the 7128 two-sample t-test statistics on the 7128 genes and the null distribution for the t-statistic.

```
tstats = apply(X,2, function(x)
  t.test(x[y==0],x[y==1], var.equal =T )$stat
)
hist(tstats, 20, freq=FALSE, col="green")
curve(dt(x,df=n-2), from=min(tstats), to=max(tstats), add=TRUE)
```



References

- Efron (2012) Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. IMS Monographs, Cambridge University Press.
- Efron and Hastie (2016) Computer-Age Statistical Inference: Algorithms, Evidence, and Data Science, Cambridge University Press