Statistics and Linguistic Applications
Hale
March 1, 2010

Motivation

second language How well does performance on the Cambridge English Proficiency Test predict error rates in spoken English?

psycho- How well does token frequency predict response latency in lexical decision?

socio- How well does age predict vowel height in Michiganders?

corpus How well do all my annotators agree on what constitutes a Verb Phrase?

Academic example

Vasishth uses Faraway's `stat500` grading example.

```
> library(faraway)
> data(stat500)
> attach(stat500)
> head(stat500, n = 3)
```

```
  midterm final   hw total
1    24.5  26.0 28.5  79.0
2    22.5  24.5 28.2  75.2
3    23.5  26.5 28.3  78.3
```

As scores on the midterm vary, how do scores on the final vary? Is it the case that they change in lockstep? If they did, instructors could promise students that improving one's score by 1 point on the midterm would necessarily yield a similar 1 point improvement on the final. In such an obviously hypothetical situation (figure 1) all pairs (midterm, final) of scores would lie on a straight line of slope 1.

```
> plot(cbind(final, final))
> abline(0, 1)
```

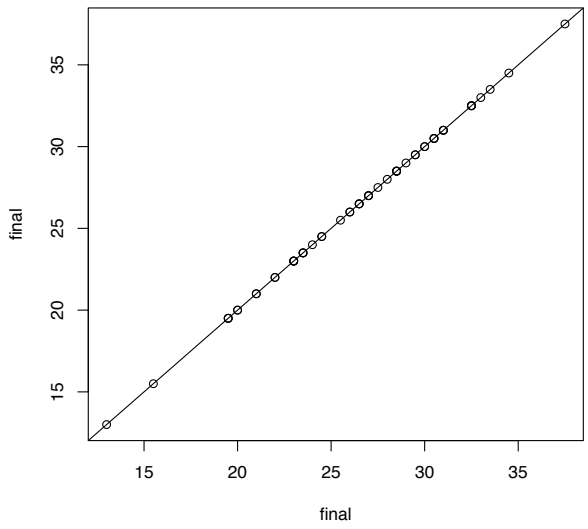


Figure 1: Ridiculous but true: performance on the final perfectly predicts performance on the final

The whole point of using statistics, though is that real life is far from this ideal. The actual relationship between midterm and final scores in Faraway's data are somewhat messier (figure 2). If there were a relationship between midterm and final, then more dots representing (midterm, final) pairs would inhabit the *first* and *third* quadrants than second and fourth.

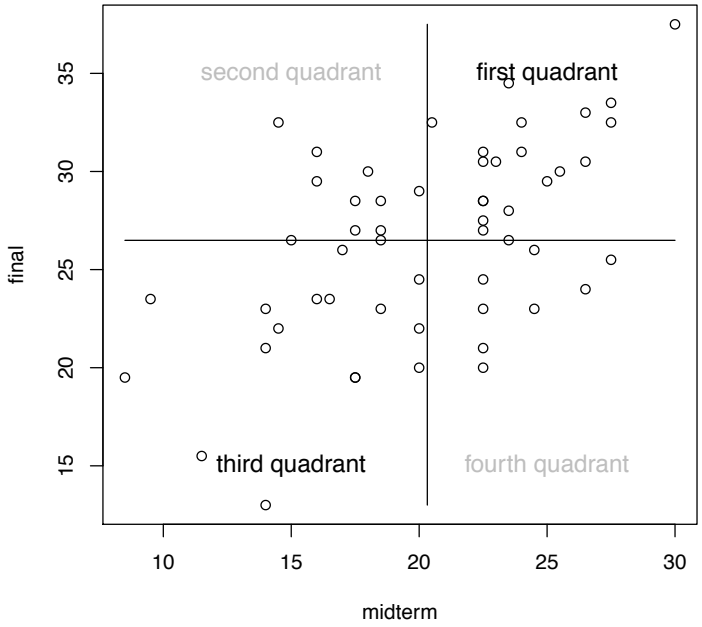


Figure 2: Relationship between midterm and final score in Faraway's `stat500`

Writing out the graphical, quadrant-based concept (figure 2) as words,

first quadrant I am above average on the midterm AND above average on the final.

third quadrant I am below average on the midterm AND below average on the final.

second quadrant I am below average on the midterm, but above average on the final.

fourth quadrant I am above average on the midterm, but below average on the final.

invokes the *difference from the mean*. Consider paired data points x_i and y_i. If $x_i - \bar{x}$ is positive, then in a sample where the two are related, $y_i - \bar{y}$ ought to be positive as well. Corners matter a lot — if both of these differences are big, then the (x, y) association is strong. Of course, since multiplying negative numbers gives positive numbers, credit is given for alignment in the third quadrant as well as the first. Combining all these ideas leads to the sum of product of deviations:

$$\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad \text{sum of product of deviations} \tag{1}$$

Sadly, quantity 1 will give different results in large vs. small samples. Dividing by something close to the sample size yields something that might be called the the average product of deviations.

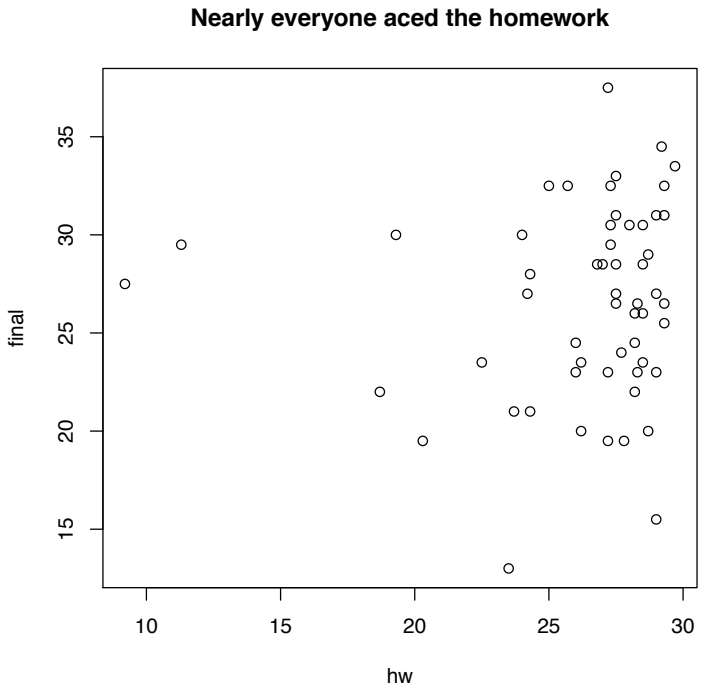
$$\frac{\sum_{i=0}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} \qquad \text{Covariance of } x \text{ and } y \tag{2}$$

Quantity 2 is the Covariance, a measure of the relationship between two quantitative variables. The Covariance of two random variables X and Y is a well-defined notion of *lack* of independence: $Cov(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$. Conversely, one can also use Covariance to ask about the relative strength of association between different sample-pairings.

```
> deviations <- function(s) {
+     (s - rep(mean(s), length(s)))
+ }
> sum(deviations(midterm) * deviations(final))/(length(final) - 1)

[1] 12.95202

> sum(deviations(hw) * deviations(final))/(length(final) - 1)

[1] 1.730960
```

Homework grades seem to be less-strongly related to performance on the final. Probably this is because nearly everyone did well on the homework.

```
> plot(final ~ hw, main = "Nearly everyone aced the homework")
```



Equation 2 has its own shortcoming, which can be understood by reference to the difference between centimeters and meters. If we are linearly predicting peoples' weight from their height, using centimeters rather than meters will inflate things a hundredfold. It would be nice to standardize-away this sort of problem to obtain a scale-independent measure of relatedness between two variables, no matter what their units of measurement. Z-scores do this

$$Z_y = \frac{y - \bar{y}}{s_y} \qquad \text{standardized scores}$$

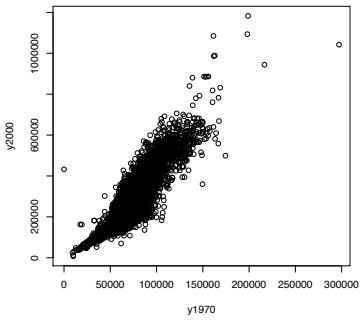
where s_y is just the standard deviation of the sample y.

$$\frac{1}{n-1}\sum_{i=0}^{n}\left(\frac{x_i - \bar{x}}{s_x}\right)\left(\frac{y_i - \bar{y}}{s_y}\right) = \frac{1}{n-1}\sum_{i=0}^{n}\left(Z_x\right)\left(Z_y\right) = \frac{Cov(X,Y)}{s_x s_y} \tag{3}$$

Equation 3 is the **correlation** r. It is scale-independent, trapped within $[-1.00, 1.00]$ and indicative of how strongly two variables are related in a linear way. If $r = 0$ then no relationship exists. If r is positive, the cloud slopes up and "the more the more" i.e. acceleration and gas-pedal depression in a properly-maintained car. If r is negative, things are inverted, i.e. the more education women get the fewer children they tend to have (in the USA).

Home value

One is the assessed value of homes in Maplewood, NJ. Are homes that were valuable in back in '70 still valuable in 2000?



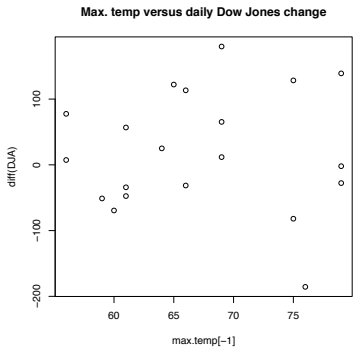
How strong is relationship between 1970 assessed value and 2000 assessment?

```
> homemade.cov <- sum(deviations(y1970) * deviations(y2000))/(6841 - 1)
> c(homemade.cov, cov(y1970, y2000))

[1] 2610880277 2610880277

> homemade.cor <- homemade.cov/(sd(y1970) * sd(y2000))
> c(homemade.cor, cor(y1970, y2000))

[1] 0.8962155 0.8962155

> detach(homedata)
```

Weather vs. Dow Jones Industrial Average

Perhaps the maximum temperature in Central Park is reliably correlated with the net daily change in the stock market?

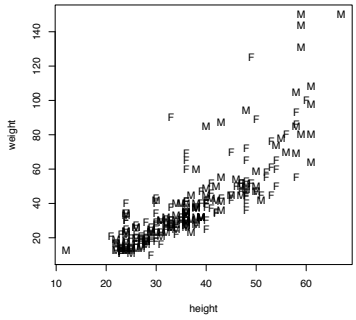
```
> attach(maydow)
> names(maydow)

[1] "Day"      "DJA"      "max.temp"

> plot(max.temp[-1], diff(DJA), main = "Max. temp versus daily Dow Jones change")
> cor(max.temp[-1], diff(DJA))

[1] 0.01028846

> detach(maydow)
```



Boys & Girls Height vs. Weight

Taller people are heavier. But is this relationship a linear one?

```
> attach(kid.weights)
> plot(height, weight, pch = as.character(gender))
> cor(height, weight)

[1] 0.8237564

> detach(kid.weights)
```



Despite the high correlation, the true relationship between weight and height is a quadratic one.

Outliers

The 2000 election for President of the USA was very close. Both Pat Buchanan and George Bush were conservative candidates, and we might expect a strong relationship between the number cast for one versus the other. Consider the Florida voter records county-by-county. The best-fitting line (figure 3) suggests that in most counties, every additional vote for Bush is corresponds to about 0.005 of a vote for Buchanan. Except for two places!

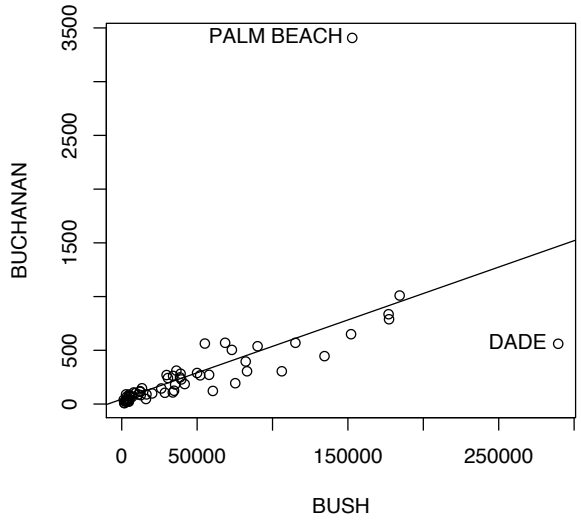


Figure 3: Votes for BUSH and BUCHANAN in Florida counties

Interpretation

- there could be systematic nonlinear relationships in your data — correlation is blind to these
- If $r_{xy} = 0$ then X cannot be causing Y. However, even if r_{xy} is large, do not conclude that X causes Y. The causal story may go through some third factor, or there may in fact be no causal story.
- the **coefficient of determination** r^2 is interpretable as the proportion of variance explained

Hypothesis test

If you are willing to assume that the data X and Y are both Normally distributed, you can do a t-test against the hypothesis that the best-fitting line is actually flat.

$\mathcal{H}_0$ true correlation is zero
$\mathcal{H}_1$ true correlation is nonzero

The R command `cor.test` carries out such a test.

```
> attach(maydow)
> cor.test(max.temp[-1], diff(DJA))

        Pearson's product-moment correlation

data:  max.temp[-1] and diff(DJA)
t = 0.0437, df = 18, p-value = 0.9657
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 -0.4342092  0.4507570
sample estimates:
       cor
0.01028846

> detach(maydow)
```

The confidence interval that R reports is based on Fisher's Z-transformation. This transformation converts a correlation r into something approximately normal.

$$Z = \frac{1}{2} \ln \left(\frac{1+r}{1-r} \right)$$

It can be used to test a hypothesis about *pairs* of correlation coefficients.

$\mathcal{H}_0$ true correlations are identical
$\mathcal{H}_1$ true correlation are different

The difference z in 4 is approximately $\mathcal{N}(0,1)$.

$$z = (Z_1 - Z_2) / \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}} \tag{4}$$

The graph produced by `qqnorm` will be straight if Normality holds.

Rank correlation

When Normality fails, or precise numerical quantities are not available (e.g. rate this item on a 1-7 scale), it is still possible to order the data by ranks. The Spearman ranked correlation r_{rank} between ranked data R and S is

$$r_{\text{rank}} = 1 - \frac{6 \sum_{i=0}^{n} (R_i - S_i)^2}{n(n^2 - 1)} \tag{5}$$

Is the k-tallest kid also the k-fattest?

```
> attach(kid.weights)
> cor(height, weight, method = "spearman")

[1] 0.8822136

> detach(kid.weights)
```

Now You Try

1. Match up the r values to their scatterplots in figure 4.

2. File `englishgreek.txt`, posted on the class website, presents data on the "gravity" of English-language errors made by Greek-Cypriot learners. The `english` column has the error gravity scores assigned to the learners by native English speaking teachers. The `greek` column gives the corresponding observations from teachers whose native language is Greek. An important political question is whether or not the two nationalities converge in their assessment of second-language learners.

   (a) Plot a scattergram and visually examine this bivariate data.

   (b) Calculate the covariance of the two sets of scores.

   (c) Calculate the correlation coefficient. What does this value say about the linear relationship between the scores of hte Greek-speaker teachers and the native English-speaking teachers?

   (d) Is the correlation coefficient you obtained statistically significant?

   (e) Oops, we just realized that two data points were left off. It turns out that there were some student work that the Greek and English teachers completely agreed on.

sentence number	Greek teachers	English teachers
33	3	3
34	0	0

   Recalculate your correlation cofficient on this new, enlarged dataset. What difference does the addition of these two points for each group make?

The figure below has six scatter diagrams for hypothetical data. The correlation coefficients, in scrambled order, are:

$$-0.85 \quad -0.38 \quad -1.00 \quad 0.06 \quad 0.97 \quad 0.62$$

Match the scatter diagrams with the correlation coefficients.

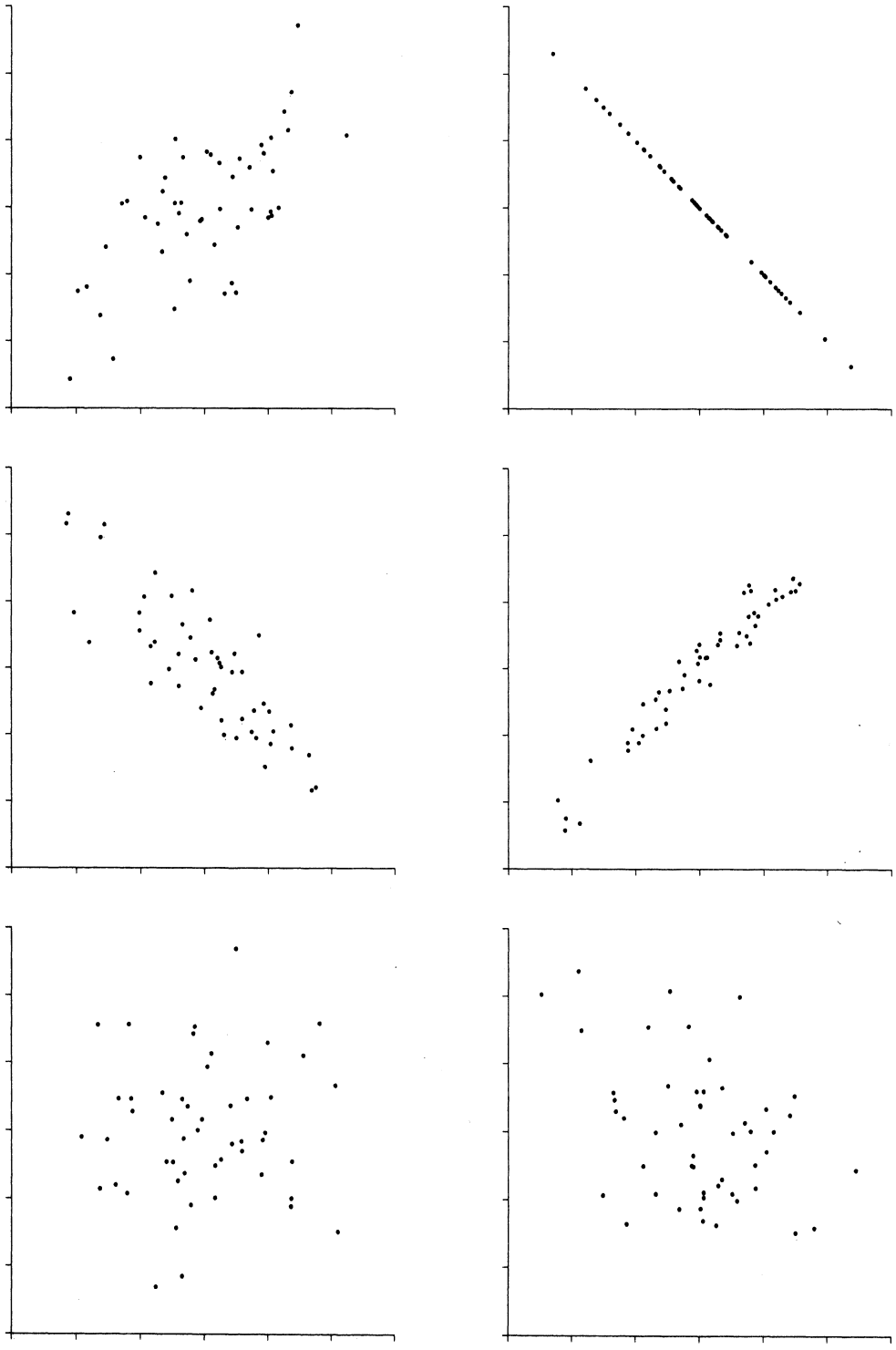


Figure 4: For homework problem 1