Lecture 2: Vulnerable

Call and response

To get started on this topic, I sent out a handful of emails to try to acquire some of the base vulnerability data from researchers publishing indices -- Here's a typical note, with identifying references stripped

This is "the call"...

hi

i'm a professor in the statistics department at ucla... i'm writing to see if i might be able to get a copy of the data you assembled for your XXXXXX article on vulnerability to extreme climate events. i was specifically looking for the data referred to for your first regression in your XXXXXX section.

is that possible?

thanks!

M.

Mark Hansen | www.stat.ucla.edu/~cocteau

Call and response

When I teach 202a, we often discuss how much of computational science is not particularly reproducible -- Two of the emails I received in response to my “call” tell the story nicely...

Mark,

Sure. It has been a little while, and I am not 100% sure which files I used, but I am 99% sure that it was the sheet "late with population" in the attached excel spreadsheet, which ought to match what is in the STATA file. If this doesn't seem right, let me know, and I can spend a few more minutes hunting for the right data.

Cheers,

XXXXX

Dear Mark

This will take some digging through old files and backup disks as this was some time ago and my filing system is probably not what it should be. However, I'll try and find some time to do the necessary excavation. The data we used were all publicly available, and you should be able to reproduce the analysis based on the description of the methodology in the paper, if that's your interest. In any case feel free to give me a nudge in a week or so. When do you need to have the data, given your teaching schedule?

It would be good to have a professional statistician cast an eye over this, and the data we used were up to 2000, and could do with being updated. In any case I'm sure you'll find much to criticise! As someone who finds themselves having to deal with vulnerability indices I'm very sceptical of them, even the ones I've produced myself.

All the best

XXXXXX

Call and response

What's beautiful here is that you find two very different ways of working -- In one, the researcher is able to produce data almost immediately that (while not exactly right) got us really close to the published results

In the other case, there's a certain amount of hunting that has to take place -- I don't mean to fault this researcher in any way, I simply wanted to make the case that we should strive to be more like the first researcher (and better)

It's also worth noting the reception that statisticians get if we're not careful -- Writing to researchers outside of statistics often elicits a kind of fear that we're going to check up on them or criticize their work in some way

My (unwanted and highly biased) advice to you is to be the kind of statistician that tries to help!

PNAS PNAS PNAS

Estimating least-developed countries' vulnerability to climate-related extreme events over the next 50 years

Anthony G. Patt[a,1], Mark Tadross[b], Patrick Nussbaumer[c], Kwabena Asante[d], Marc Metzger[e,f], Jose Rafael[g], Anne Goujon[a,h], and Geoff Brundrit[i]

[a]International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]Climate Systems Analysis Group, University of Cape Town, Rondebosch 7701, South Africa; [c]Institute of Environmental Science and Technology, Autonomous University of Barcelona, 08193 Bellaterra, Spain; [d]Climatus LLC, Mountain View, CA 94041; [e]Centre for the Study of Environmental Change and Sustainability, University of Edinburgh, EH8 9XP, Scotland; [f]Alterra, Wageningen University and Research Centre, 6700 AA Wageningen, The Netherlands; [g]Department of Geography, University of Eduardo Mondlane, Maputo, Mozambique; [h]Vienna Institute of Demography, Austrian Academy of Sciences, 1040 Vienna, Austria; and [i]Department of Oceanography, University of Cape Town, Rondebosch 7701, South Africa

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approved December 4, 2009 (received for review September 10, 2009)

When will least developed countries be most vulnerable to climate change, given the influence of projected socio-economic development? The question is important, not least because current levels of international assistance to support adaptation lag more than an order of magnitude below what analysts estimate to be needed, and scaling up support could take many years. In this paper, we examine this question using an empirically derived model of human losses to climate-related extreme events, as an indicator of vulnerability and the need for adaptation assistance. We develop a set of 50-year scenarios for these losses in one country, Mozambique, using high-resolution climate projections, and then extend the results to a sample of 23 least-developed countries. Our approach takes into account both potential changes in countries' exposure to climatic extreme events, and socio-economic development trends that influence countries' own adaptive capacities. Our results suggest that the effects of socio-economic development trends may

sensitivity to those stressors, which in turn is determined by a complex set of social, economic, and institutional factors collectively described as determining its adaptive capacity (5, 6). As the UNFCCC secretariat suggested in its needs assessment, "one of the key limitations in estimating the costs of adaptation is the uncertainty about adaptive capacity. Adaptive capacity is essentially the ability to adapt to stresses such as climate change. It does not predict what adaptations will happen, but gives an indication of differing capacities of societies to adapt *on their own* to climate change or other stresses" (1, p. 97).

Human losses to extreme weather events can serve as a reliable indicator for this vulnerability, and with it the need for financial assistance, for two reasons. First, measures to reduce vulnerability to extreme weather events account for a particularly large share of estimated adaptation financial needs (1). Second, in the context of efforts to achieve a wide range of development goals, it is only

Vulnerability

The underlying question here is interesting and relevant (they usually are, for what it's worth) -- Here we are interested in understanding how climate change (and the accompanying increase in extreme weather events) will affect different parts of the world

Specifically, the researchers produce a model that relates variables capturing some notion of vulnerability to the impacts that weather-related natural disasters have had, country by country

Estimating least-developed to climate-related extreme 50 years

Anthony G. Patt[a,1], Mark Tadross[b], Patrick Nussbaumer[c], Kwa
Anne Goujon[a,h], and Geoff Brundrit[i]

[a]International Institute for Applied Systems Analysis, 2361 Laxenburg, Austria; [b]
South Africa; [c]Institute of Environmental Science and Technology, Autonomous
View, CA 94041; [d]Centre for the Study of Environmental Change and Sustainab
University and Research Centre, 6700 AA Wageningen, The Netherlands; [g]Depa
Mozambique; [h]Vienna Institute of Demography, Austrian Academy of Sciences
Cape Town, Rondebosch 7701, South Africa

Edited by Stephen H. Schneider, Stanford University, Stanford, CA, and approv

When will least developed countries be most vulnerable to climate change, given the influence of projected socio-economic development? The question is important, not least because current levels of international assistance to support adaptation lag more than an order of magnitude below what analysts estimate to be needed, and scaling up support could take many years. In this paper, we examine this question using an empirically derived model of human losses to climate-related extreme events, as an indicator of vulnerability and the need for adaptation assistance. We develop a set of 50-year scenarios for these losses in one country, Mozambique, using high-resolution climate projections, and then extend the results to a sample of 23 least-developed countries. Our approach takes into account both potential changes in countries' exposure to climatic extreme events, and socio-economic development trends that influence countries' own adaptive capacities. Our results suggest that the effects of socio-economic development trends may begin to offset rising climate exposure in the second quarter of the century, and that it is in the period between now and then that vulnerability will rise most quickly. This implies an urgency to the need for international assistance to finance adaptation.

vulnerability | adaptive capacity | development | natural disasters | natural hazards

Results

The first stage of our analysis was to estimate statistical models of losses from climate-related disasters, based on a set of climatic and socio-economic variables that will likely change over time, which appear in Table 1. The dependent variables are logged values of the number of people per million of national population killed or affected, respectively, by droughts, floods, or storms over the period 1990–2007. The variable number of disasters is the logged value of numbers reported by each country over the same period, and accounts for climate exposure; estimated coefficient values greater than 1 in both models indicate that average losses per disaster are higher in more disaster-prone countries. We expected that larger countries are likely to experience disasters over a smaller proportion of their territory or population, and also benefit from potential economies of scale in their disaster management infrastructure, both resulting in lower average per capita losses; the negative coefficient estimates for the variable national population in both models are consistent with this expectation. The variable HDI represents the Human Development Index, a United Nations (UN) indicator comprised of per capita income, average education and literacy rates, and average life expectancy at birth. Recent studies of disaster losses —not limited to climate-related events—have shown that countries with medium HDI values experience the highest average losses, whereas countries with high HDI values experience the lowest (14, 15). We therefore included the logged HDI values in quadratic form. Negative coefficient estimates for both HDI and HDI^2 in both models are thus consistent with these expectations, given that logged HDI values are always negative, and the square of the logged values are in turn positive. Finally, we considered several additional socio-economic variables not directly captured by HDI, and found only two that improved model fit. For the model of the number of people killed, the positive coefficient estimate for female fertility indicates that countries with higher birth rates experience greater average numbers of deaths. We do not take this to mean that there is a direct connection between fertility and natural hazard deaths, but rather that higher birth rates are associated with lower female empowerment, and lower female empowerment is associated with higher disaster vulnerability, as has been shown previously (16, 17). For the model of the number of people affected, the negative coefficient estimate for the proportion urban population is consistent with urban residents being less likely to require post-disaster assistance than rural residents, also observed previously (18, 19). Both models yield an R^2 statistic slightly greater than 0.5, indicating that variance in the independent variables explains just over half of the variance in the numbers killed and affected. This is consistent with results from past analyses based on similar data and methods (8–10).

Vulnerability

In the end, a great deal of attention is paid to a regression table (below), the form of which we should be fairly familiar with

In each row they present the regression of the logarithm of the number of people killed by weather-related natural disasters from 1990 to 2007 as a function of several predictors, one of which is slightly special...

Table 1. Ordinary least-squares regression results

Independent variables	Killed	Affected
Number of disasters	1.36* (0.15)	1.88* (0.19)
National population	−0.56* (0.09)	−0.79* (0.11)
HDI	−5.97* (1.95)	−13.55* (2.16)
HDI^2	−6.26* (1.52)	−9.82* (1.86)
Female fertility	1.45* (0.43)	
Proportion urban population		−0.41 (0.37)
Constant	−3.86* (0.49)	5.33* (1.71)
Number of observations	150	154
R^2	0.52	0.55

The dependent variable in the Killed model is the logged value of the number of people reported by CRED as killed by the three types of disasters considered (droughts, floods, and storms) divided by population. The dependent variable in the *Affected* model is the same for the number of people reported affected, but not killed, by the same disasters. All independent variables are logged values. Because HDI occupies the range of 0–1, all logged HDI values were negative, whereas the squares of these values were positive. *Values significant (*two-tailed student's t* test) at the 99% confidence level. Values in parentheses are SEs.

HDI

The HDI or Human Development Index, a United Nations (UN) is an “indicator comprised of per capita income, average education and literacy rates, and average life expectancy at birth”

From the table on the previous page, we see that the variable and its square are both included in the final model and are given the following interpretation

“Of particular importance to rapidly developing countries is the observed nonlinear relationship between HDI and disaster losses. Fig. 1 illustrates the magnitude of this effect in both models, compared with the background variance, and taking into account the effects of the other variables. The estimated regression curve in Fig. 1A suggests that the risk of being affected by a climate disaster is highest in countries with HDI values of ~0.5, whereas the curve in Fig. 1B suggests that the highest risk level is for countries with HDI values somewhat higher, ~0.6. This suggests that for countries with HDI values of less than 0.5, the transition to higher levels of development could potentially, in the absence of targeted intervention, exacerbate vulnerability.”

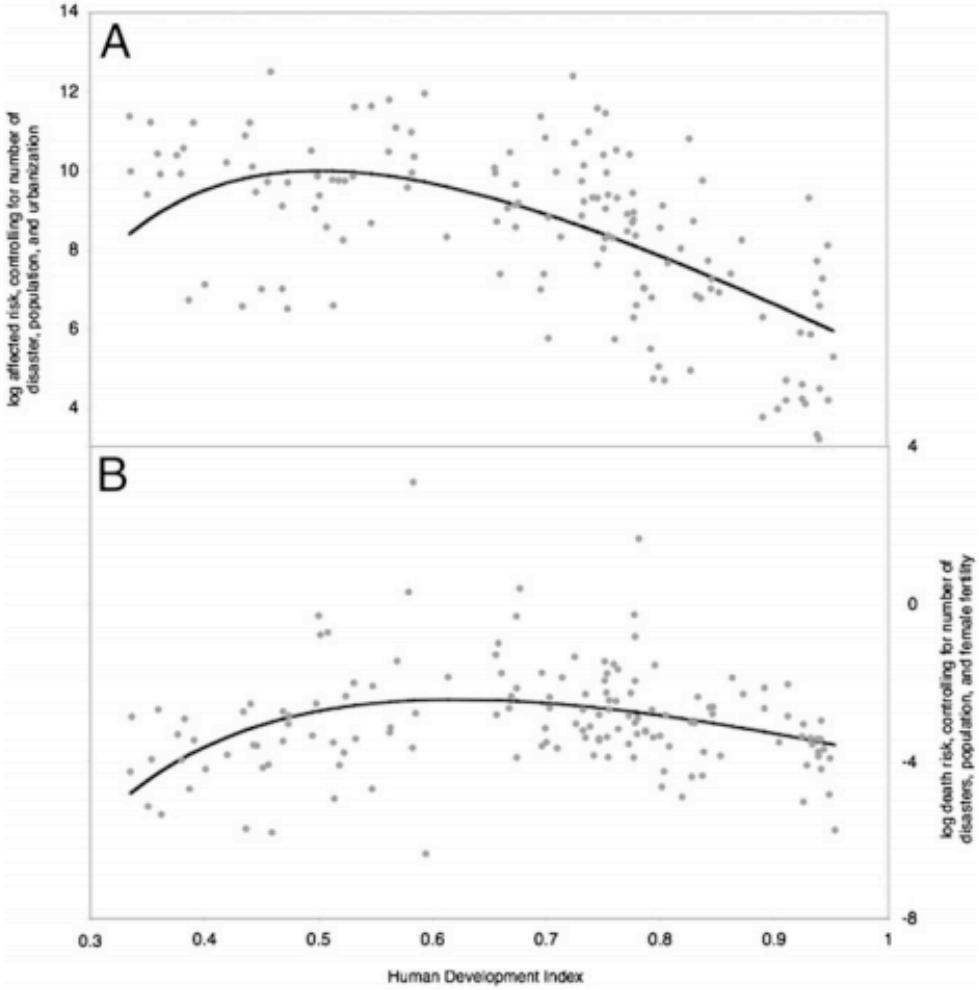


Fig. 1. Relationship between risk and HDI for (*A*) the number of people affected, i.e., needing emergency or recovery assistance, by a flood, drought, or cyclone, per million of population, and (*B*) the number of people killed. Each dot represents a country in the CRED database during the period 1990–2007, with its position on the vertical scale being the logarithm of the annual value per million population, after subtracting the predicted influence of other risk factors. Regression line in each figure shows predicted values including the influence of HDI.

Looking at the data

The data we were given consist of measurements associated with 144 different countries -- For each we have the following variables

`country_name` the name of the country

`ln_events` the natural logarithm of the number of droughts, floods and storms occurring in the country from 1990-2007

`ln_pop` the natural logarithm of the country's population

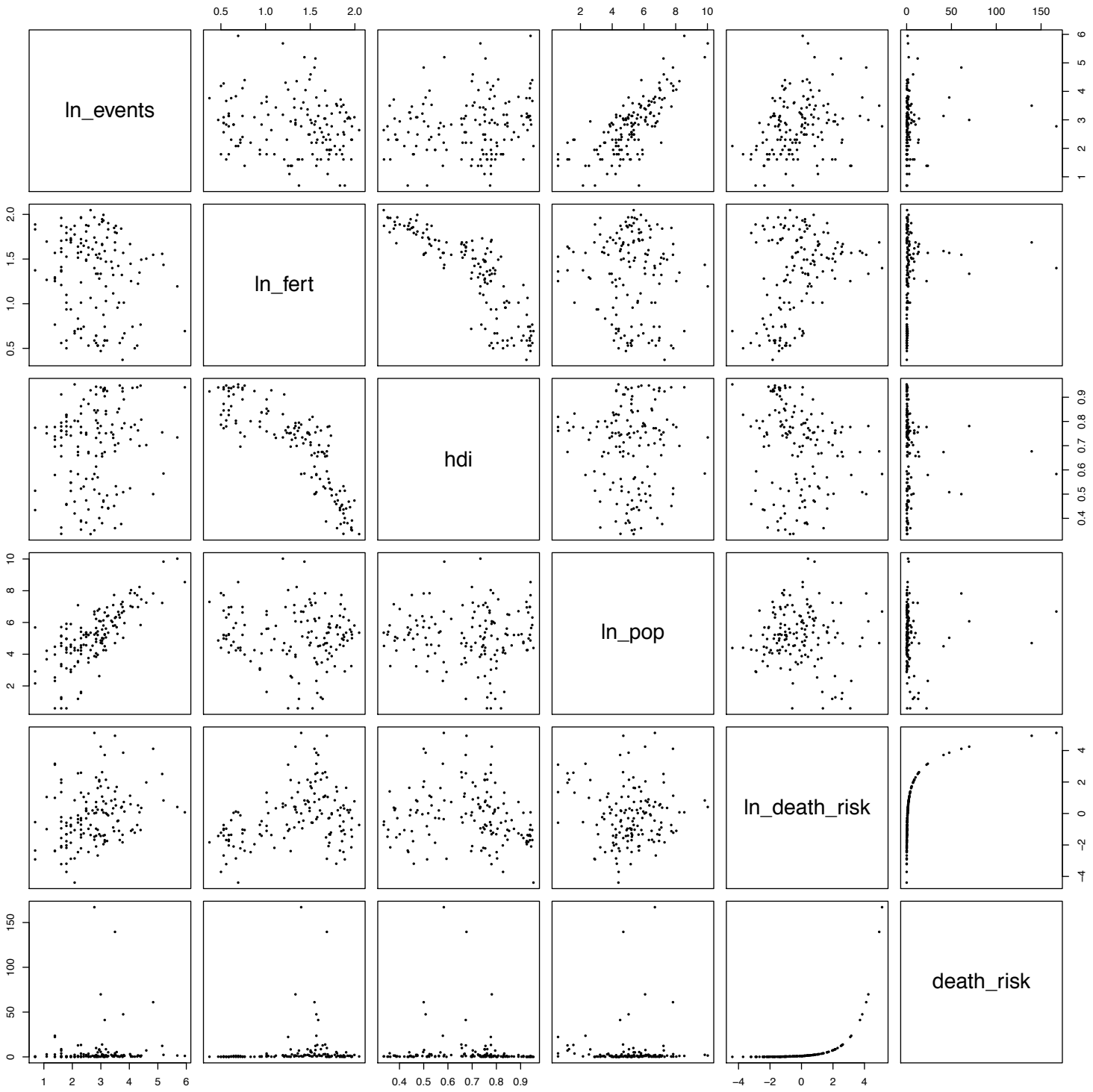
`ln_fert` the natural logarithm of an estimate of the country's female fertility

`hdi` the Human Development Index for the country

`death_risk` the proportion of people out of 1M in population killed in droughts, floods and storms

There are four predictor variables (if you count HDI and its square as one) which, while not big by any stretch of the imagination, is complex enough to keep us from "seeing" the whole data set

Instead, we might opt for partial views...


ln_events
ln_fert
hdi
ln_pop
ln_death_risk
death_risk

```
# first, load the vulnerability data

> load(url("http://www.stat.ucla.edu/~cocteau/stat201b/vulnerability.RData"))
> names(vul)

# [1] "country_name"  "ln_urb"        "ln_events"     "ln_fert"
# [5] "hdi"           "ln_pop"        "ln_death_risk" "death_risk"

# fit without quadratic on hdi for the moment

> fit <- lm(ln_death_risk~ln_events+ln_fert+ln_pop+hdi,data=vul)
> summary(fit)

# Call:
# lm(formula = ln_death_risk ~ ln_events + ln_fert + ln_pop + hdi,
#    data = vul)
#
# Residuals:
#     Min      1Q  Median      3Q     Max
# -3.4518 -0.7673 -0.1513  0.5669  6.2271
#
# Coefficients:
#             Estimate Std. Error t value Pr(>|t|)
# (Intercept)  -5.3485     1.5175  -3.524 0.000575 ***
# ln_events     1.3708     0.1792   7.649 3.04e-12 ***
# ln_fert       2.1961     0.4614   4.760 4.81e-06 ***
# ln_pop       -0.5672     0.1026  -5.529 1.54e-07 ***
# hdi           1.9922     1.2628   1.578 0.116928
# ---
# Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 #
# Residual standard error: 1.35 on 139 degrees of freedom
# Multiple R-squared: 0.4221,!Adjusted R-squared: 0.4055
# F-statistic: 25.38 on 4 and 139 DF,  p-value: 8.522e-16
```

The table

Shortly we'll talk about the variance-covariance structure of the coefficients (or rather their sampling distribution) which will get us closer to understanding some of the stochastic results in this table

The table of t-statistics come (ultimately) from a distributional result about the residual sum of squares and independence between RSS and both the fitted values as well as the regression coefficients -- We'll cover that material little by little

Finally, the (multiple) R-squared statistic is just the proportion of the variance explained by the model -- In R code, this is simply

```
> tss = (nrow(vul)-1)*var(vul$ln_death_risk)
> rss = sum(residuals(fit)^2)
> 1-rss/tss
[1] 0.4221038

> summary(fit)$r.squared
[1] 0.4221038

> cor(fitted(fit),vul$ln_death_risk)^2
[1] 0.4221038
```

The table

R’s tabular summary is not unlike output produced by large-scale computer packages in the late 1960s -- SAS output still bears some resemblance to this approach

At about this time, however, there was a growth in graphical diagnostics, strategies that depended on the visual inspection of plots to identify a model’s shortcomings

In the next lecture, we will take on residual diagnostics in earnest -- Before that, however, we will beat a hasty retreat and build up some notation

The linear model

We have n observations or pairs of data points where each pair consisted of a "response" y_i, and one or more "predictors" $x_{i1}, \ldots, x_{ip}$

We assume that our data were generated by **an underlying probability model** in which the responses y_i are observations from a random variable Y_i -- Collecting these variables into a vector $Y = (Y_1, \ldots, Y_n)^t$, our model assumes

$$EY = \mu = (\mu_1, \ldots, \mu_n)^t \quad \text{and} \quad \text{var } Y = \sigma^2 I_{n \times n}$$

In a linear model, we link the mean vector μ of our responses to our set of predictors via a linear equation

$$EY_i = \mu_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$$

The linear model

If we assemble our predictors into a matrix M (for Model matrix, also known as the design matrix for those of you who took 201a)

$$M = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

and our parameters into a vector $\beta = (\beta_1, \ldots, \beta_p)^t$ we can write this dependence in the form $EY = M\beta$

The normal linear model

The **normal linear model** adds the assumption that the responses are independent and normally distributed -- Therefore, each Y_i has a normal distribution with mean μ_i and variance σ^2

Put another way, our vector of responses can be written as

$$Y = M\beta + \epsilon$$

where $\epsilon = (\epsilon_1, \ldots, \epsilon_n)^t$ are independent normal random variables, each with mean 0 and variance σ^2

Notice that in this formulation, we think of the **responses as being random variables while our predictors are fixed** -- The predictors can come from a designed experiment in which they are intentionally chosen, otherwise we conduct our analysis conditional on their values

Estimation

With this model, we have partitioned our responses into a systematic component Mβ and a random component ϵ

In terms of estimation, therefore, given M and observations y from this model, we would like to identify a value for the parameter vector β that explains as much of the response as possible

To formalize this idea a bit, let's step back and review a small amount of linear algebra (a very small amount)

Review: Independence, orthogonality, subspaces

A set of vectors $m_1, \ldots, m_p$ in $\mathbb{R}^n$ is linearly independent if

$$\sum_{j=1}^{p} a_i m_i = 0 \quad \text{if and only if} \quad a_1 = \cdots = a_p = 0$$

Otherwise, a nontrivial combination of the $m_1, \ldots, m_p$ is zero and the collection is said to be linearly dependent

Review: Independence, orthogonality, subspaces

A subspace of $\mathbb{R}^n$ is a subset that is also a vector space -- The set of all linear combinations of $m_1, \ldots, m_p$ in $\mathbb{R}^n$

$$\{a_1 m_1 + \cdots + a_p m_p | a_1, \ldots, a_p \in \mathbb{R}\}$$

is referred to as the span of the set $\{m_1, \ldots, m_p\}$

Review: Independence, orthogonality, subspaces

A subset $\{m'_1, \ldots, m'_r\}$ is a maximal linearly independent subset of $\{m_1, \ldots, m_p\}$ if its not properly contained in any linearly independent subset of $\{m_1, \ldots, m_p\}$

If $\{m'_1, \ldots, m'_r\}$ is maximal, then

$$\mathsf{span}\{m'_1, \ldots, m'_r\} = \mathsf{span}\{m_1, \ldots, m_p\}$$

and $\{m'_1, \ldots, m'_r\}$ is a basis for $S = \mathsf{span}\{m_1, \ldots, m_p\}$

All bases have the same number of elements and this number is called the dimension of S

Review: Independence, orthogonality, subspaces

Two vectors m_1 and m_2 in $\mathbb{R}^n$ are said to be orthogonal if

$$m_1^t m_2 = m_2^t m_1 = \sum_{i=1}^{n} m_{i1} m_{i2} = 0$$

We say a set of vectors $\{m_1, \ldots, m_p\}$ is orthogonal if $m_i^t m_j = 0$ whenever $i \neq j$; and orthonormal if $m_i^t m_j = \delta_{ij}$

The orthogonal complement of a subspace S is defined by

$$S^{\perp} = \{v \in \mathbb{R}^n \mid v^t m \text{ for all } m \in S\}$$

Review: Independence, orthogonality, subspaces

Finally, there are two important subspaces associated with a matrix M in
-- The range of an n x p matrix M is defined to be

$$R(M) = \{v \in \mathbb{R}^n \mid Mb \text{ for some } b \in \mathbb{R}^p\}$$

and the null space of M is given by

$$N(M) = \{b \in \mathbb{R}^p \mid Mb = 0\}$$

If we write $M = [m_1, \ldots, m_p]$ then the range of M is just the span of $\{m_1, \ldots, m_p\}$
and we define the rank of M to be the dimension of the span of its columns

Review: Length

We begin with standard Euclidean distance -- Let z be an n-dimensional vector with components z_i, then we can define the length of z

$$\|z\| = \sqrt{\sum_{i=1}^{n} z_i^2}$$

using slightly more compact notation $\|z\| = \sqrt{z^t z}$

Estimation

Returning to our estimation problem -- Ideally we would like to find that element in the span of our predictors, the range of M, that is as close to our data as possible

Formally, we are looking for a value of β such that the length of the error or residual vector is as small as possible -- That is, we want to minimize

$$\|y - M\beta\|^2 = (y - M\beta)^t(y - M\beta) = \sum_{i=1}^{n}(y - \beta_1 x_{i1} - \ldots - \beta_p x_{ip})^2$$

Estimation

As we saw last time (albeit with different notation), we can differentiate this expression to yield the normal equations

$$\mathrm{M}^{\mathrm{t}}\mathrm{M}\beta = \mathrm{M}^{\mathrm{t}}\mathrm{y}$$

If M has full rank p, then we can invert $\mathrm{M}^{\mathrm{t}}\mathrm{M}$ to form the (now infamous) solution

$$\widehat{\beta} = (\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{y}$$

Sampling distributions

Recall that we are viewing the responses y as observations of a random vector Y, the distribution of which was governed by the (normal) linear model -- We are also conditioning on the values of the predictors in M

Therefore, using the expression on the previous slide, we can show that the sampling distribution of $\widehat{\beta}$ is normal with mean

$$\begin{aligned} \mathrm{E}\widehat{\beta} &= (\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{E}\mathrm{Y} \\ &= (\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{M}\beta \\ &= \beta \end{aligned}$$

and variance-covariance matrix

$$\begin{aligned} \mathrm{var}\ \widehat{\beta} &= (\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{var}\ \mathrm{Y}\ \mathrm{M}(\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1} \\ &= \sigma^2(\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{M}(\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1} \\ &= \sigma^2(\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1} \end{aligned}$$

Sampling distributions

We can go a bit farther consider the estimated means $\widehat{\mu} = \mathsf{M}\widehat{\beta}$ which are normal with mean μ and variance-covariance matrix $\sigma^2\mathsf{M}(\mathsf{M}^\mathsf{t}\mathsf{M})^{-1}\mathsf{M}^\mathsf{t} = \sigma^2\mathsf{H}$ where we recall the hat matrix $\mathsf{H} = \mathsf{M}(\mathsf{M}^\mathsf{t}\mathsf{M})^{-1}\mathsf{M}^\mathsf{t}$

Similarly, the residuals $\mathsf{y} - \mathsf{M}\widehat{\beta} = (\mathsf{I} - \mathsf{H})\mathsf{y}$ should be normally distributed with mean 0 and variance-covariance matrix $\sigma^2(\mathsf{I} - \mathsf{H})$

A geometric view

The least squares estimate for μ is the closest point to the observed data y in the range of M, the set of possible mean vectors -- In the parlance of linear algebra, we say that $\widehat{\mu}$ is the projection of y onto R(M)

This closest point (in n-dimensional space) is unique -- Depending on whether M is invertible or not, we might have different representations for it, but the point itself is unique

A geometric view

Recall that the hat matrix H is symmetric ($\mathsf{H}^t = \mathsf{H}$) and idempotent ($\mathsf{H}^2 = \mathsf{H}$) -- We can use these two facts to show that the residual vector

$$\mathsf{y} - \widehat{\mu} = \mathsf{y} - \mathsf{H}\mathsf{y} = (\mathsf{I} - \mathsf{H})\mathsf{y}$$

is orthogonal to the estimated means

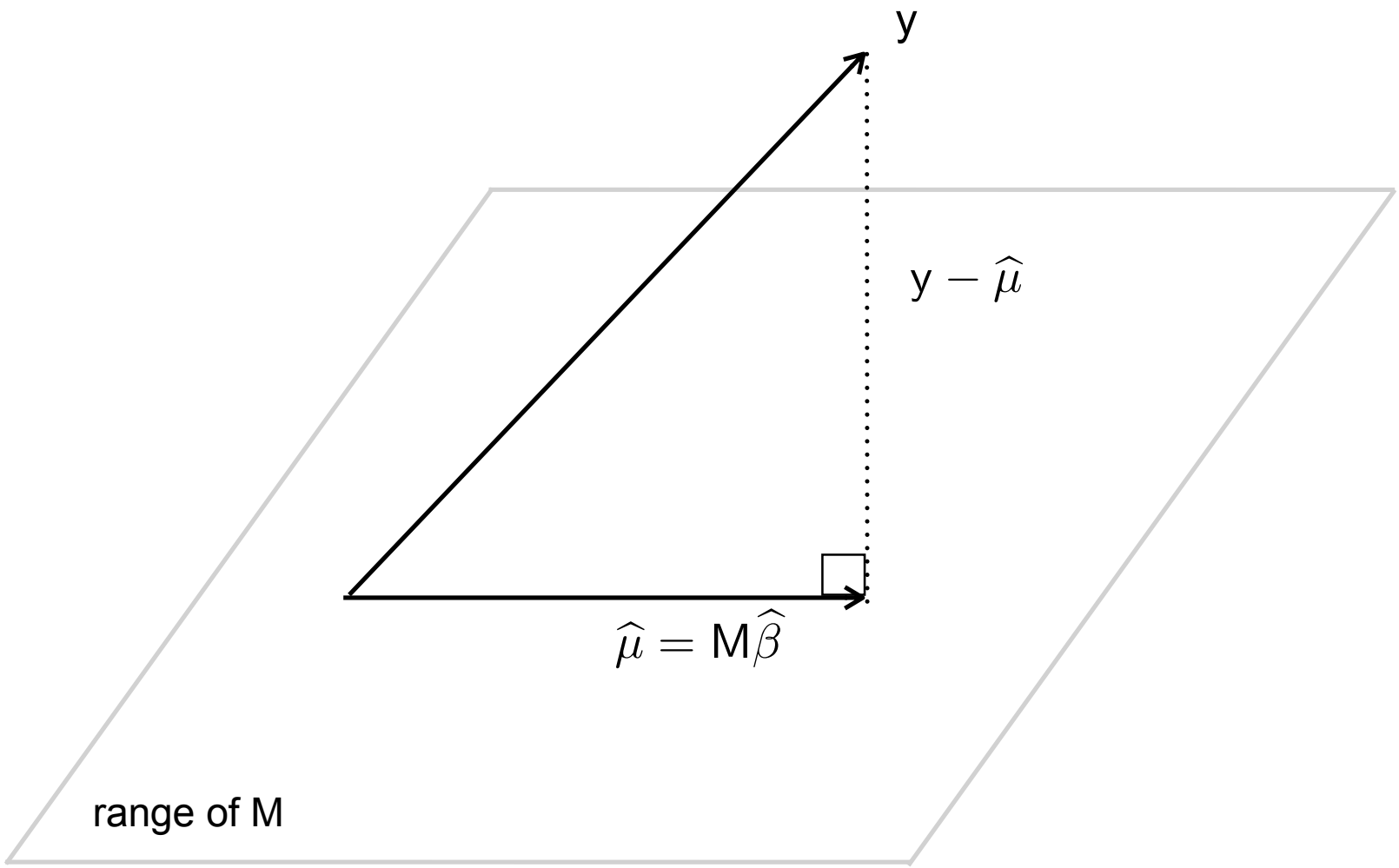
$$\begin{aligned} \widehat{\mu}^t(\mathsf{y} - \widehat{\mu}) &= \mathsf{y}^t\mathsf{H}^t(\mathsf{I} - \mathsf{H})\mathsf{y} \\ &= \mathsf{y}^t(\mathsf{H} - \mathsf{H}\mathsf{H})\mathsf{y} \\ &= \mathsf{y}^t(\mathsf{H} - \mathsf{H})\mathsf{y} \\ &= 0 \end{aligned}$$

A geometric view

This also means that we have partitioned the length of the vector y into two components, one for the (estimated) systematic part and one for the (estimated) random component

$$\begin{aligned} \|\mathsf{y}\|^2 &= \|\mathsf{y} - \widehat{\mu} + \widehat{\mu}\|^2 \\ &= \|\mathsf{y} - \widehat{\mu}\|^2 + \|\widehat{\mu}\|^2 + 2\widehat{\mu}^{\mathsf{t}}(\mathsf{y} - \widehat{\mu}) \\ &= \|\mathsf{y} - \widehat{\mu}\|^2 + \|\widehat{\mu}\|^2 \end{aligned}$$

This is essentially the Pythagorean theorem and we can see it graphically on the following (somewhat lame) slide


y
$y - \widehat{\mu}$
$\widehat{\mu} = M\widehat{\beta}$
range of M

Projection matrices

A square n-by-n matrix H is known as a projection matrix if Hy is the point in the subspace

$$S = \{Hb \mid b \in \mathbb{R}^n\}$$

that is closest to y

From linear algebra we might recall that **a square matrix H is a projection matrix if and only if it is symmetric and idempotent**

Projection matrices

We've already established one direction of that statement by deriving exhibiting the hat matrix H associated with a least squares fit -- To go the other way, assume we have a symmetric and idempotent matrix H that's n-by-n

1. If H is symmetric and idempotent, so is I-H

2. Hy and (I-H)y are orthogonal

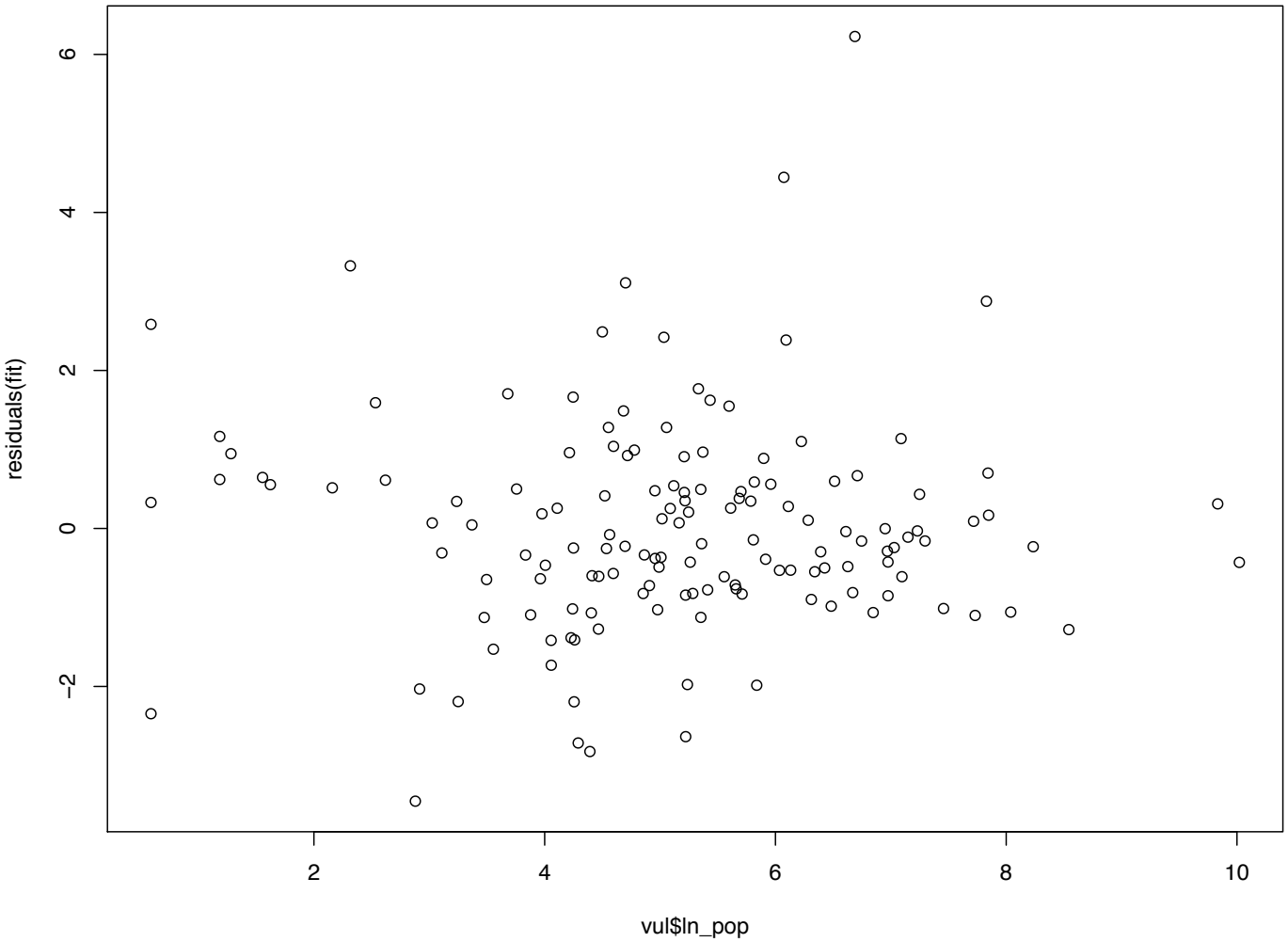
3. Finally, for any vector $b \in \mathbb{R}^n$ we have

$$\|y - Hb\|^2 = \|y - Hy\|^2 + \|Hy - Hb\|^2$$

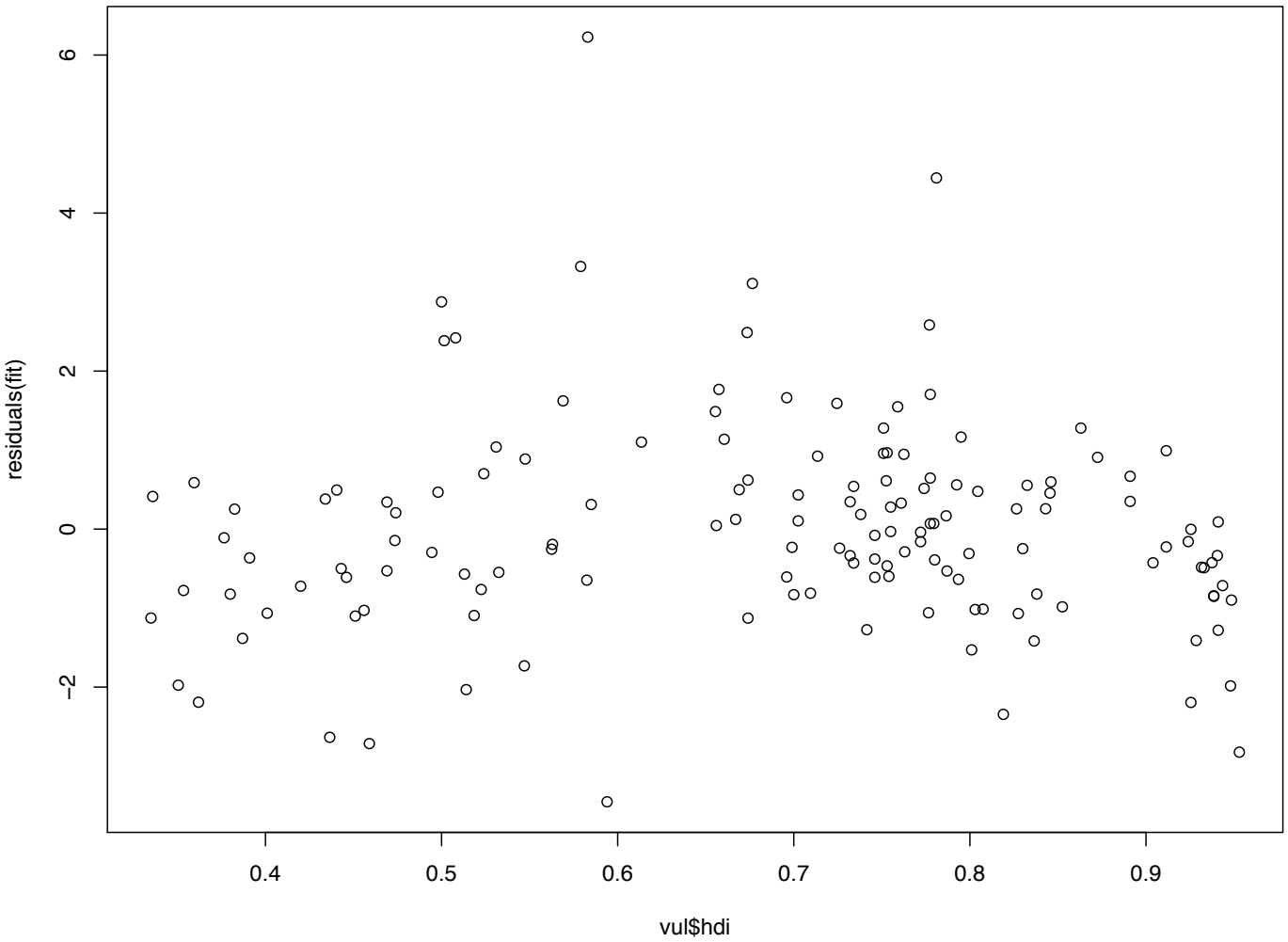
Diagnostic plots

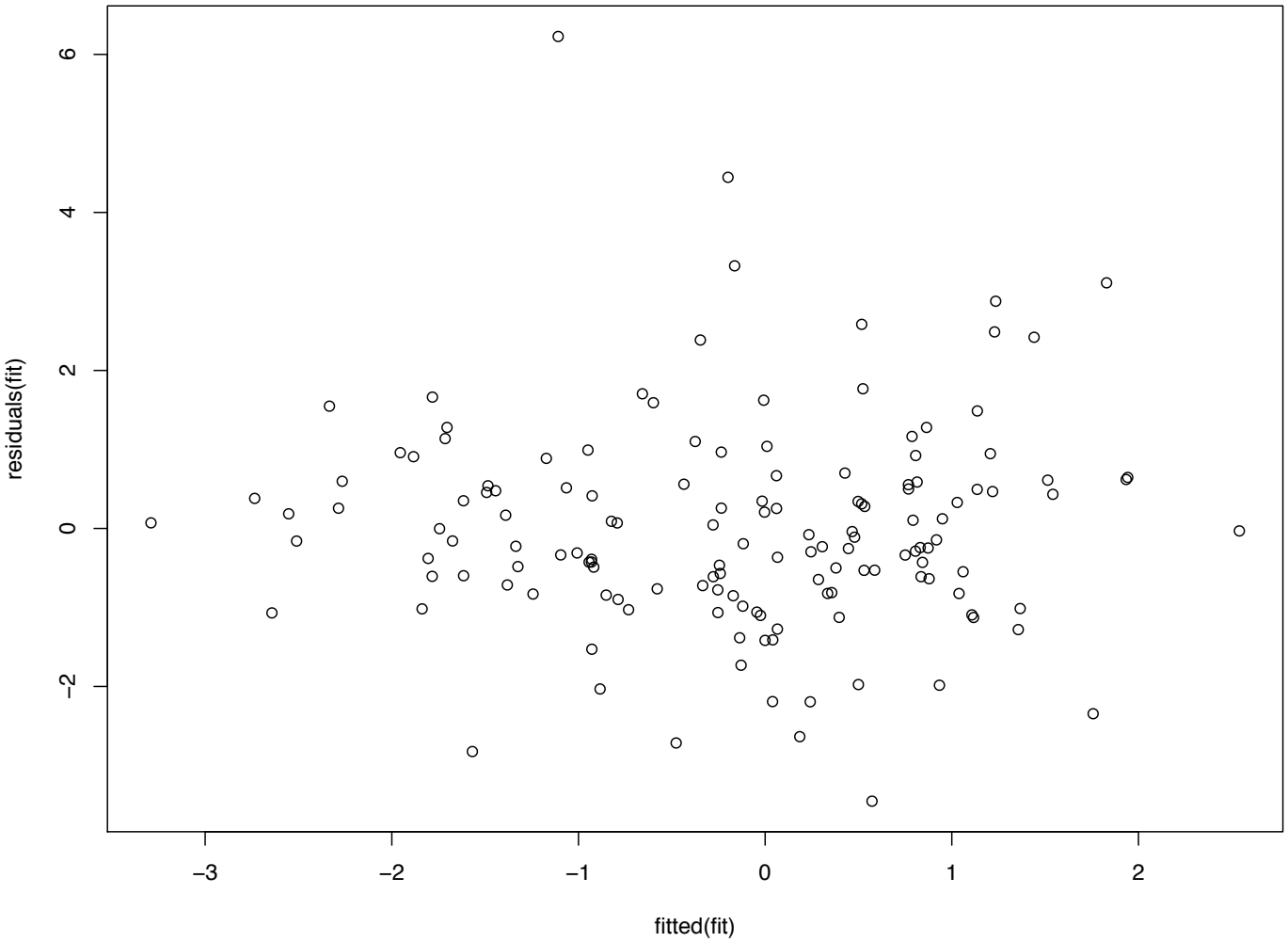
These orthogonality relationships happen thanks to the least squares procedure -- We should examine various plots of these quantities to see if there is any structure missing in the current fit

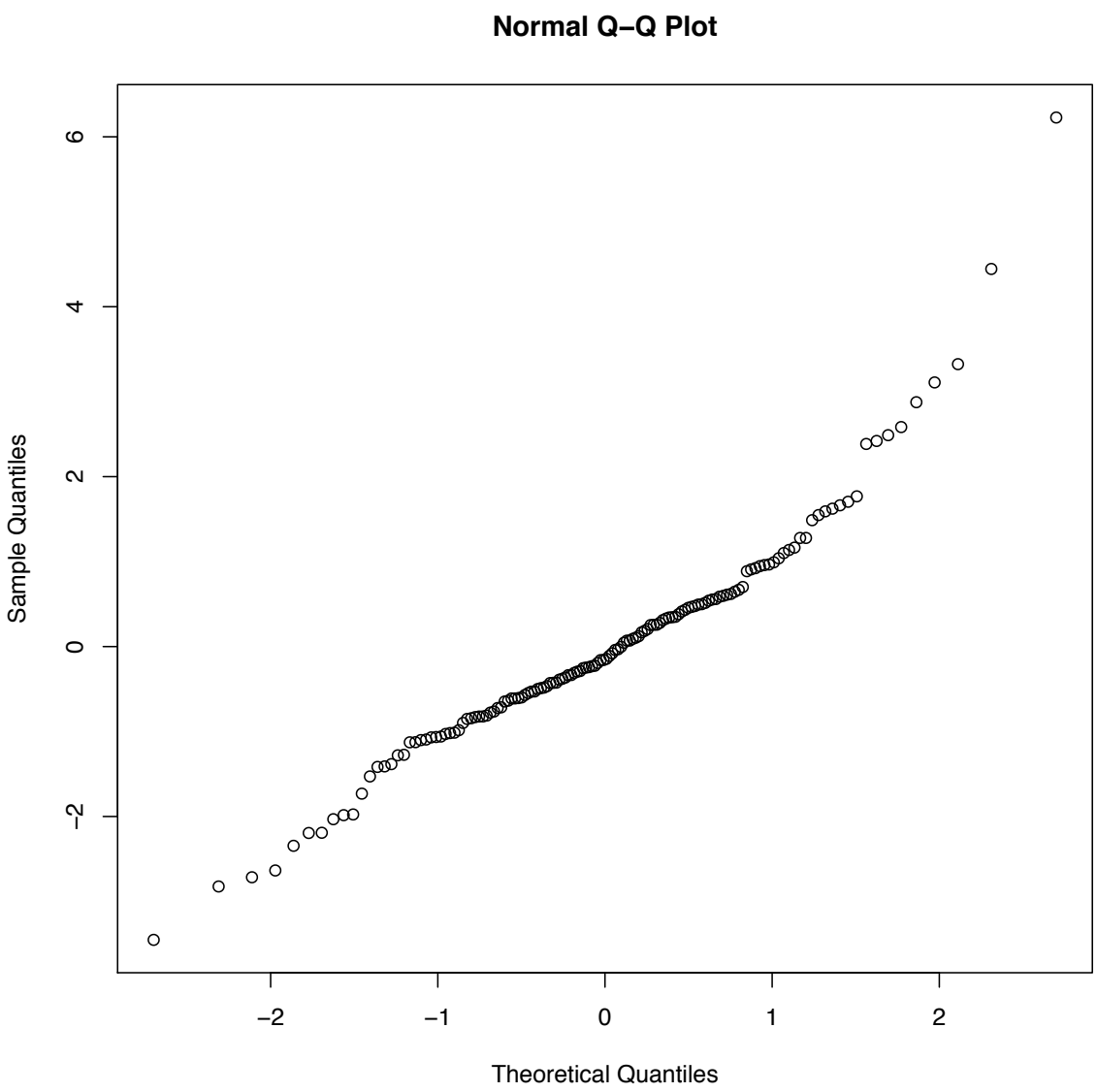
```
# now, some diagnostic plots...

# residuals against predictors
plot(vul$ln_pop,residuals(fit))
plot(vul$ln_hdi,residuals(fit))

# residuals against and against the fitted values
plot(fitted(fit),residuals(fit))

# a normal quantile-quantile plot
qqnorm(residuals(fit))
```


residuals(fit)
6
4
2
0
−2
2
4
6
8
10
vul$ln_pop

Do we need a quadratic term?




residuals(fit)
6
4
2
0
−2
−3
−2
−1
0
1
2
fitted(fit)


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
6
4
2
0
−2
−2
−1
0
1
2

Aside: Other notions of length

Squared error loss is an extremely popular choice for estimation -- In the normal linear model it corresponds to Maximum Likelihood (although it predates the MLE by a century or so) and even without the assumption of normality, the least squares fit is the best linear unbiased estimate (Gauss-Markov, later)

There are, however, many other measures for the length of a vector -- You might recall from your linear algebra course that a vector norm is any real-valued function on $\mathbb{R}^n$ such that

1. $f(x) \geq 0$ for all $x \in \mathbb{R}^n$ with equality if and only if x = 0

2. $f(x + y) \leq f(x) + f(y)$ for all $x, y \in \mathbb{R}^n$

3. $f(ax) = |a|f(x)$ for all $a \in \mathbb{R}$ and $x \in \mathbb{R}^n$

Aside

The Euclidean norm we've been using for length is part of the so-called Holder family or p-norms are defined by

$$\|z\|_p = (|z_1|^p + \cdots + |z_n|^p)^{1/p}$$

of which

$$\|z\|_1 = (|z_1| + \cdots + |z_n|)$$

$$\|z\|_2 = (|z_1|^2 + \cdots + |z_n|^2)^{1/2} = \sqrt{z^t z}$$

Least absolute deviation

Rather than measuring the length of our residual vector with a 2-norm or squared error loss, we might instead choose our parameter vector so as to minimize the quantity

$$\|y - M\beta\|_1 = \sum_{i=1}^{n} |y - \beta_1 x_{i1} - \cdots \beta_p x_{ip}|$$

As a minimization problem, what's different here?

Least absolute deviation

While we can’t depend on differentiation any longer, we can reason about the fit itself and what it might look like...

Least absolute deviation

In the end, the solution will be a line that connects two points in the data set -- Which one is the subject of search or rather more intelligent optimization algorithms

Later in the quarter we'll return to this idea when we discuss quantile regression -- You can think of this as a way to estimate a conditional median (as opposed to least squares and the conditional mean)

One historical note...

A bit of history (again)

While we date least squares to Legendre's publication on the topic in 1805 or maybe Gauss's claim of 1795, Boscovich was working nearly a half century prior with similar data

Boscovich was interested in the ellipticity of the earth -- Newton and others had suggested that the earth's rotation could be expected to make it bulge at the equator with a corresponding flattening at the poles

To estimate this effect, five measurements had been made -- Each was a ("rather arduous") direct measurement of the arc-length of 1 degree of latitude at five quite dispersed points -- From Quito on the equator to a site in Lapland

	sin^2(lat)	arclen
Quito	0.0000	56751
Cape Hope	0.2987	57037
Rome	0.4648	56979
Paris	0.5762	57074
Lapland	0.8386	57422

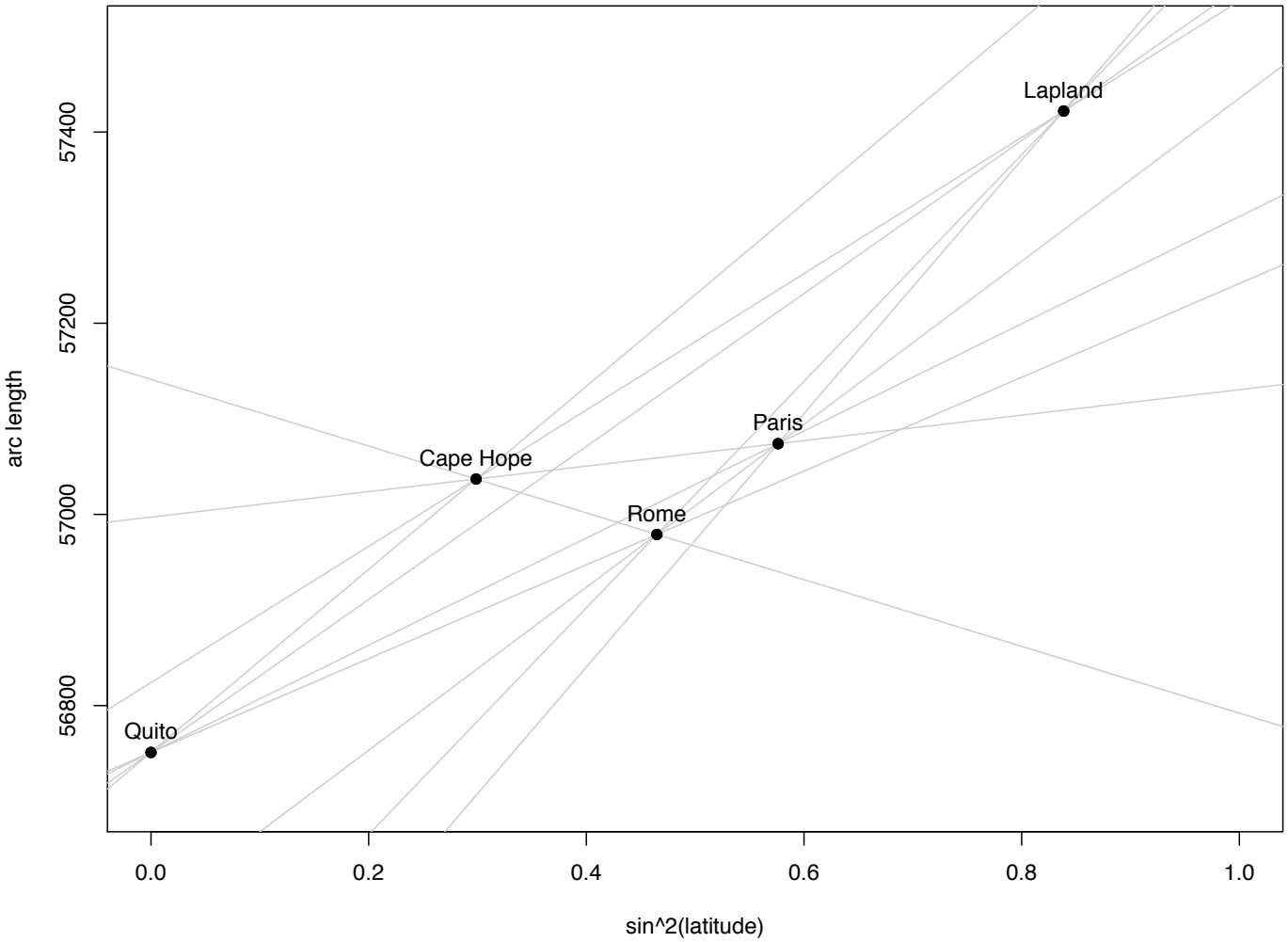
From these data, it's clear that arc length increased as you moved toward the pole from the equator, but how should you combine the five measurements to produce one estimate of the earth's ellipticity?

Aside

Boscovich used the same linearization that Stigler attempted -- For small arcs, we have that $y = a + b \sin^2\lambda$ where y is arc length and λ is latitude

The parameter a could be thought of as the length of a degree of arc at the equator and b as the exceedence of a degree of arc at the pole over its value at the equator -- In the end ellipticity could be estimated as b/(3a)

Boscovich noted that any pair of lines could be used to compute an estimate of a and b and so computed all 5-choose-2 or 10 ellipticity values -- From these he took an average (and later an average leaving out a couple ill-fitting lines)


Lapland
Paris
Cape Hope
Rome
Quito
57400
57200
57000
56800
arc length
0.0
0.2
0.4
0.6
0.8
1.0
sin^2(latitude)

Aside

Actually, it turns out that you can write the least squares estimate as a weighted average of all the pairwise fits -- For the moment assume we have an intercept and just one predictor and let h = (i,j), a pair of (distinct) row numbers

Then, define

$$M(h) = \begin{bmatrix} 1 & x_i \\ 1 & x_j \end{bmatrix} \quad \text{and} \quad y(h) = \begin{bmatrix} y_i \\ y_j \end{bmatrix}$$

It can be shown that the least squares estimate $\widehat{\beta}$ is given by

$$\widehat{\beta} = \sum_h w(h)\widehat{\beta}(h)$$

where $w(h) = |M(h)|^2 / \sum_h |M(h)|^2$ -- What can you tell me about the weights?

Aside

Interestingly, the weights are proportional to the distance between each pair of points in the data set -- What does this suggest to you about least squares fits and “outlying” data points?

And we're back...

Returning to least squares -- We've had enough tedious calculations for one class and instead we'll spend some time talking about the different matrix components we've seen so far and interpreting them in statistical terms

We'll start with the hat matrix H -- It takes us from our observed data to an estimate of the vector of conditional means

$$\widehat{\mu}_j = \sum_{i=1}^{n} h_{ji} y_i$$

if we let

$$H = \begin{bmatrix} h_{11} & h_{12} & \cdots & h_{1n} \\ h_{21} & h_{22} & \cdots & h_{2n} \\ \vdots & \vdots & & \vdots \\ h_{n1} & h_{n2} & \cdots & h_{nn} \end{bmatrix}$$

The hat matrix

Given this expression, we see that the elements h_{ij} express the degree of "leverage" that an observation y_i has on a fitted value $\widehat{\mu}_j$

In general, the greatest impact that y_i has on the fit is through $\widehat{\mu}_i$, and hence people often focus their attention on the diagonal elements h_{ii} of the hat matrix

Because H is symmetric and idempotent, we can write

$$h_{ii} = \sum_j h_{ij}^2 = h_{ii}^2 + \sum_{j \neq i} h_{ij}^2 \geq 0$$

Therefore we see that the diagonal elements satisfy $0 \leq h_{ii} \leq 1$

The hat matrix

The bound helps, but it's hard to know when a particular value is "big" -- We can show that the eigenvalues of a projection matrix are either 0 or 1

This sounds bad but it's pretty easy to derive once we have an orthogonal basis for the column space of M -- We'll do this next time with the so-called Gram-Schmidt procedure

If $q_1, \ldots, q_p$ form an orthonormal basis for the columnspace of M, and we let $q_{p+1}, \ldots, q_n$ denote an orthonormal basis for the orthogonal complement of this space, we can create a new matrix $Q = [q_1, \ldots, q_n]$

The hat matrix

Then multiplying by H gives (remember $q_1, \ldots, q_p$ will project onto themselves)

$$HQ = H\left[q_1, \ldots, q_p, q_{p+1}, \ldots, q_n\right] = \left[q_1, \ldots, q_p, 0, \ldots, 0\right]$$

so that $Q^t HQ = \mathrm{diag}(1, \ldots, 1, 0, \ldots, 0)$, with p one's

The hat matrix

So, the eigenvalues of the hat matrix are either 0 or 1 and the total number of 1's matches the rank of M -- that means that the trace of H (the sum of its diagonal elements) which is also the sum of its eigenvalues is just p

Long story short, we have

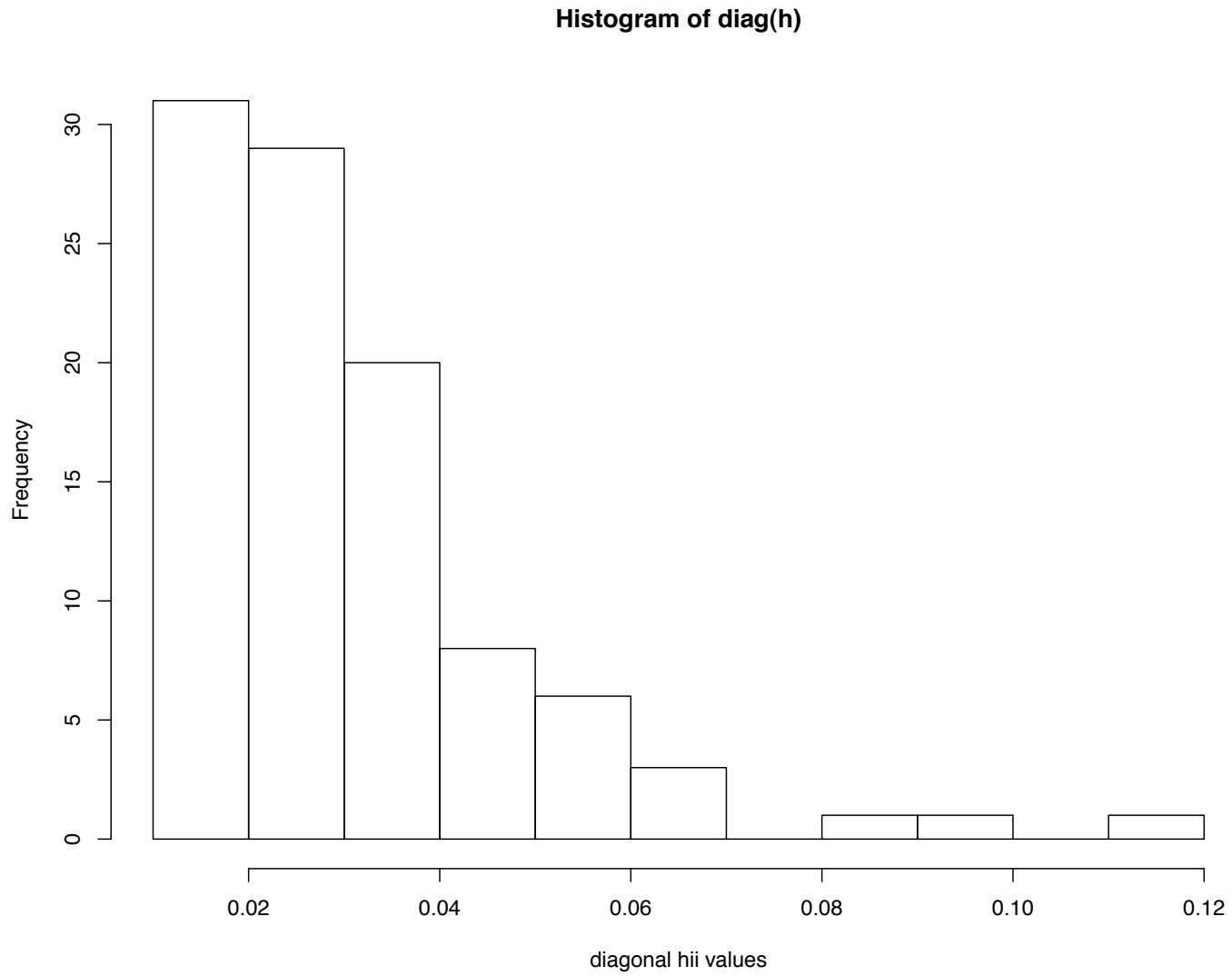
$$\sum_{i=1}^{n} h_{ii} = p \quad \text{or} \quad \frac{1}{n}\sum_{i=1}^{n} h_{ii} = p/n$$

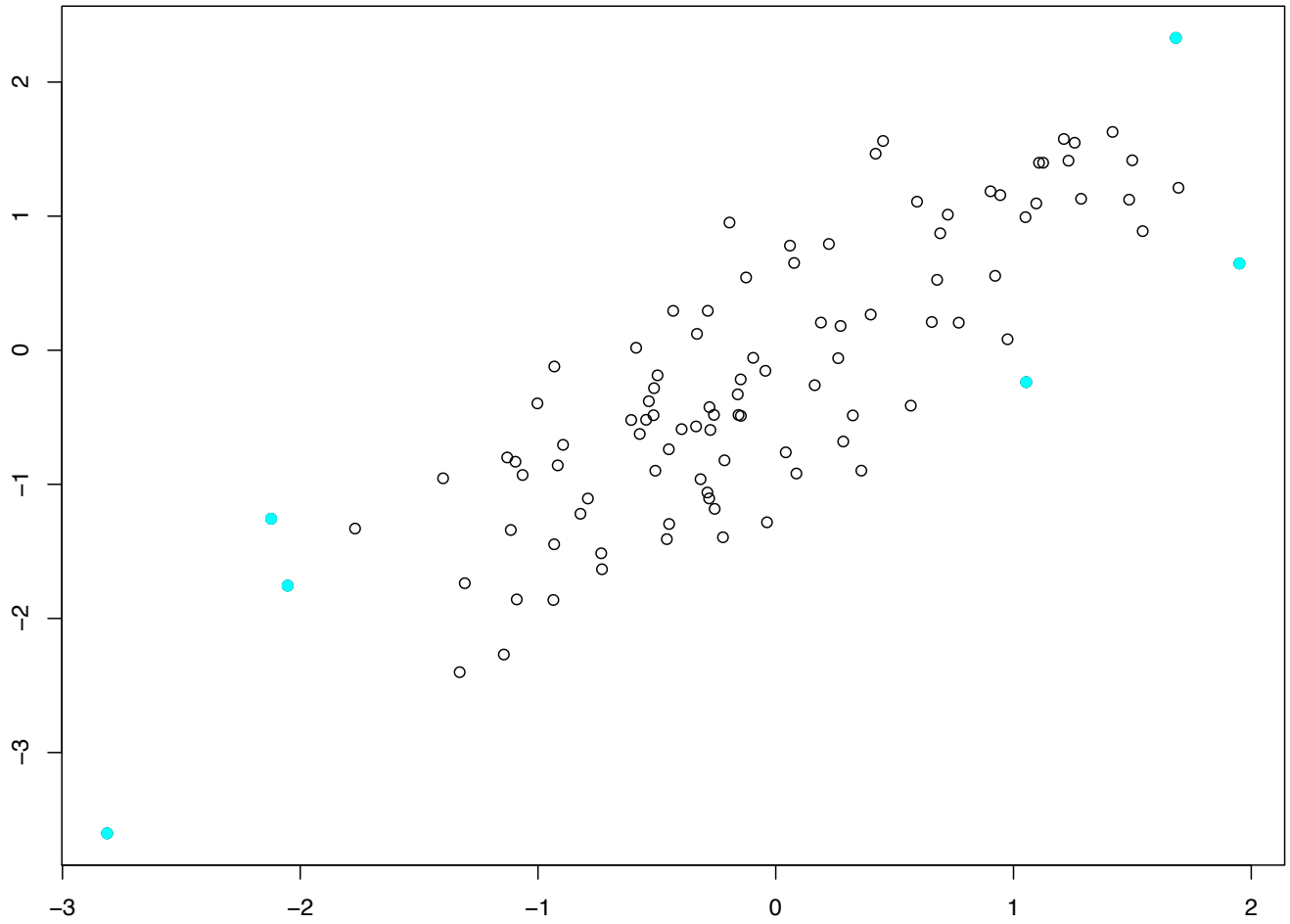
The hat matrix

In searching for high-leverage observations, then, we often look for “big” values along the diagonal of the hat matrix, where big is larger than 2p/n

On the next slide, we have simulated predictors (bivariate normal) -- 100 rows and p=3 (an intercept and two predictors) translates to a “cutoff” of 0.06

The histogram shows the 100 leverage values, six of which are above the cutoff... the points are colored cyan on the following plot -- Do they match your intuition?

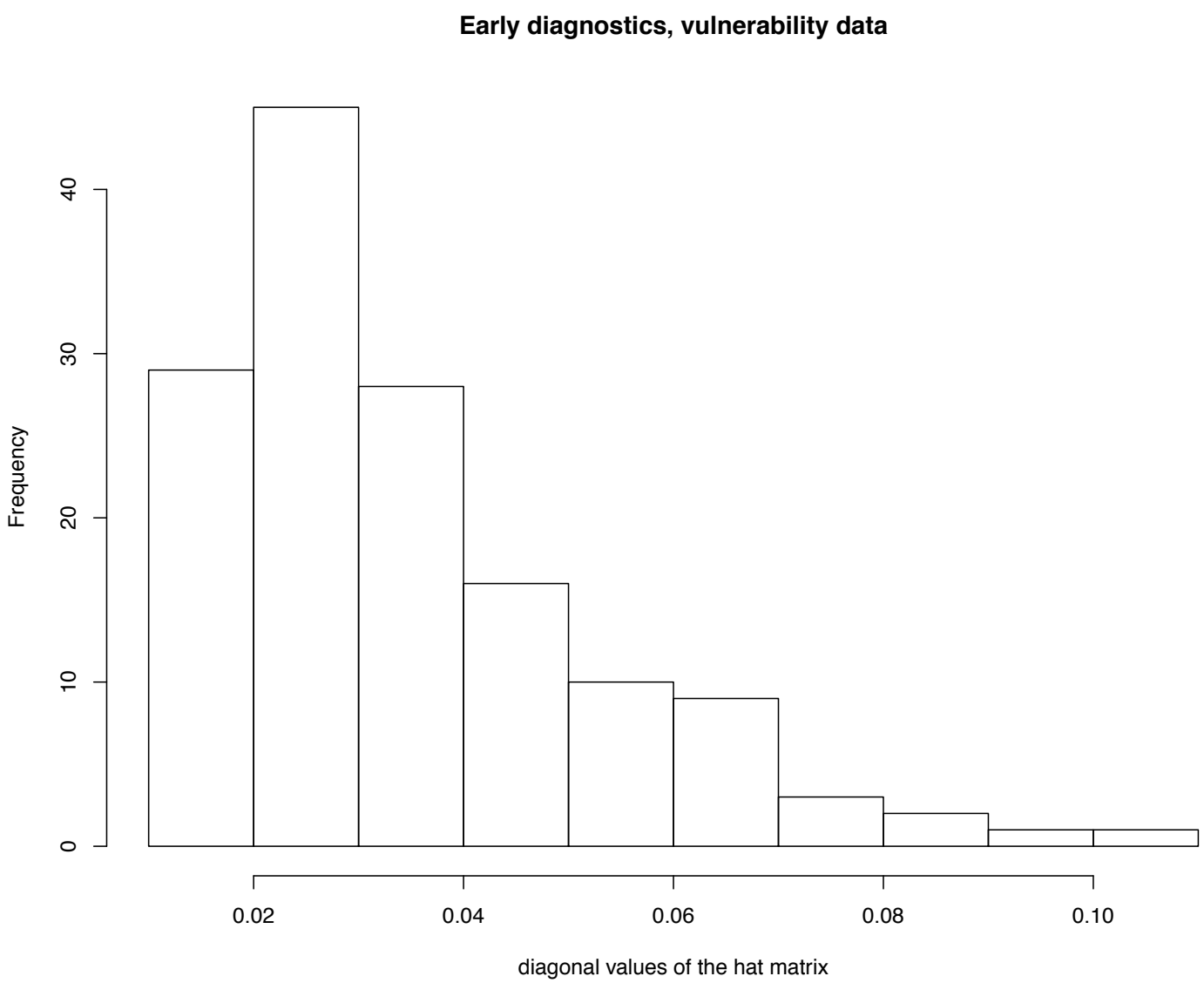

Histogram of diag(h)
Frequency
0
5
10
15
20
25
30
0.02
0.04
0.06
0.08
0.10
0.12
diagonal hii values

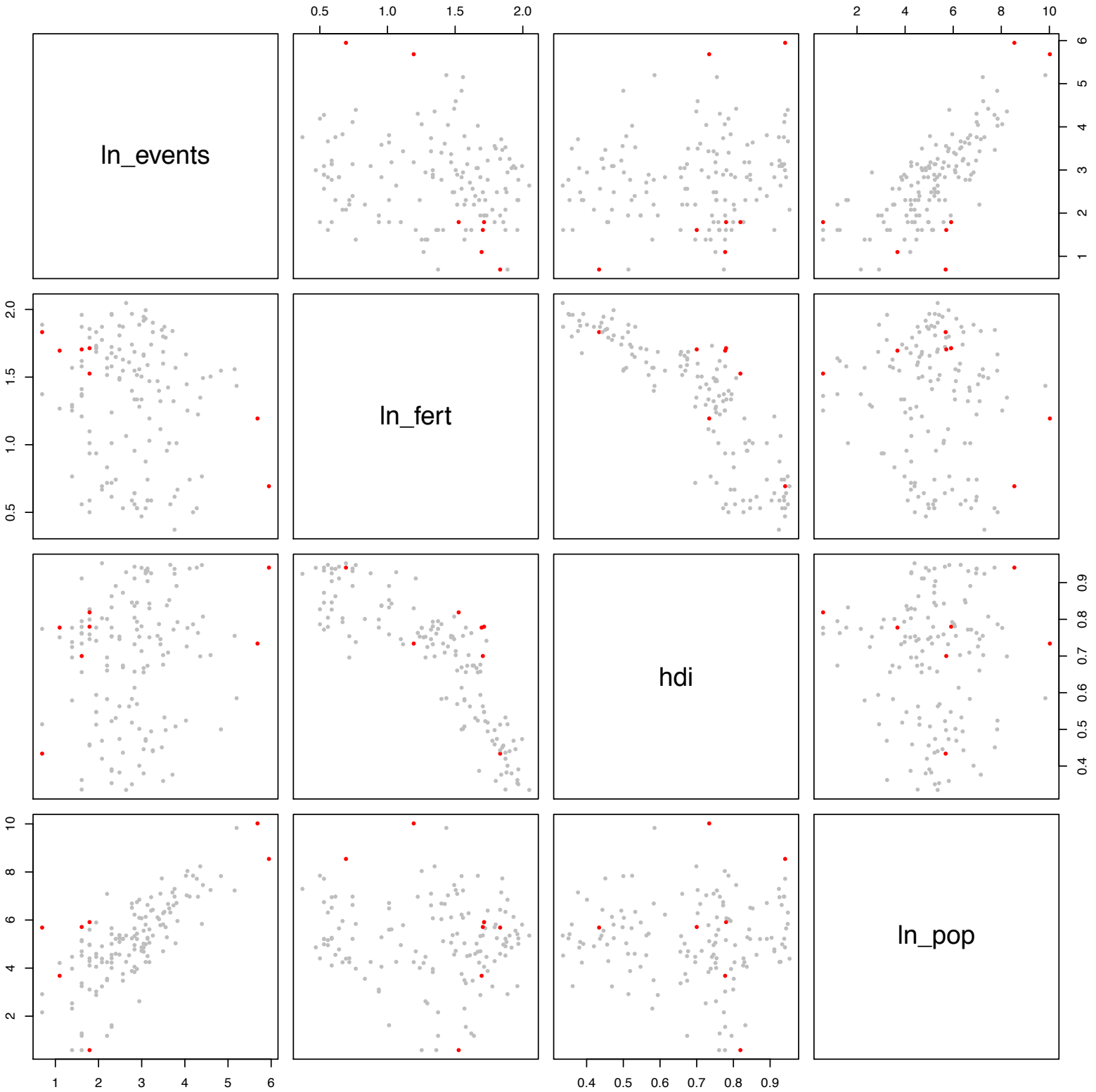

2
1
0
−1
−2
−3
−3
−2
−1
0
1
2

Applying the metric

Applying this to our simple model, we find that there are a handful of points that seem to have “big” leverage on the fits -- In R we can easily compute the hat matrix values and have a look

```
> inf = lm.influence(fit)
> names(inf)
[1] "hat"          "coefficients" "sigma"        "wt.res"

# the cutoff
> 10/nrow(vul)
[1] 0.06944444

> vul[inf$hat>0.069,1]
[1] China P Rep     Cote d'Ivoire   Oman            Saudi Arabia
[5] Syrian Arab Rep Tonga           United States
144 Levels: Albania Algeria Angola Argentina Armenia Australia ... Zimbabwe
```


Early diagnostics, vulnerability data
Frequency
0
10
20
30
40
0.02
0.04
0.06
0.08
0.10
diagonal values of the hat matrix


ln_events
ln_fert
hdi
ln_pop
0.5
1.0
1.5
2.0
2
4
6
8
10
1
2
3
4
5
6
0.5
1.0
1.5
2.0
0.4
0.5
0.6
0.7
0.8
0.9
2
4
6
8
10
1
2
3
4
5
6
0.4
0.5
0.6
0.7
0.8
0.9

The hat matrix

The goal of going down this path is simply to show that certain constructions from linear algebra have interpretations in statistical terms -- The elements of a projection matrix providing us insight into points that are exerting undo “influence” on the fit

Using the hat matrix we can also assess the effect of removing, say, a single point -- Using the Sherman-Morrison-Woodbury formula, for example, we can derive the effect on the variance-covariance matrix by dropping a point