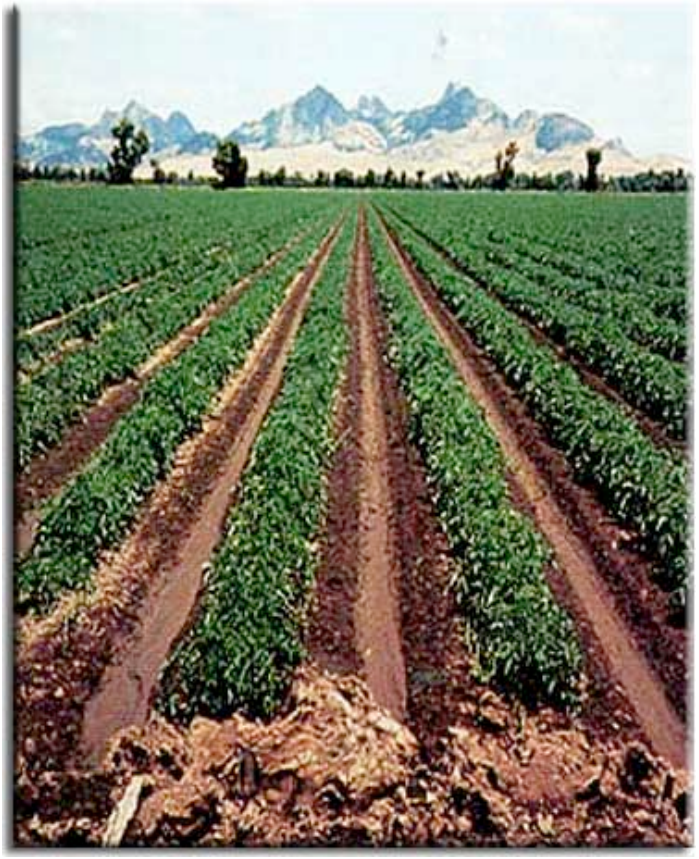


Lecture 14: Vulnerable

Last time

We developed further the idea behind regression as a tool for assessing relationships between variables -- We approached it from several directions

We started with some historical context, discussing how modern regression really began with the work of one person, Francis Galton, and his pursuit in the context of the bivariate normal distribution

Galton and the bivariate normal

Galton established that if Y_1 and Y_2 had a bivariate normal distribution, then the conditional distribution of Y_2 given that $Y_1 = y_1$ is univariate normal with mean

$$E(Y_2|Y_1 = y_1) = \mu_2 + \rho\frac{\sigma_2}{\sigma_1}(y_1 - \mu_1)$$

and variance $(1 - \rho^2)\sigma_2^2$

The regression line emerged as describing the conditional mean of Y_2 given different values of $Y_1 = y_1$

DIAGRAM BASED ON TABLE I.

(all female heights are multiplied by 1·08)

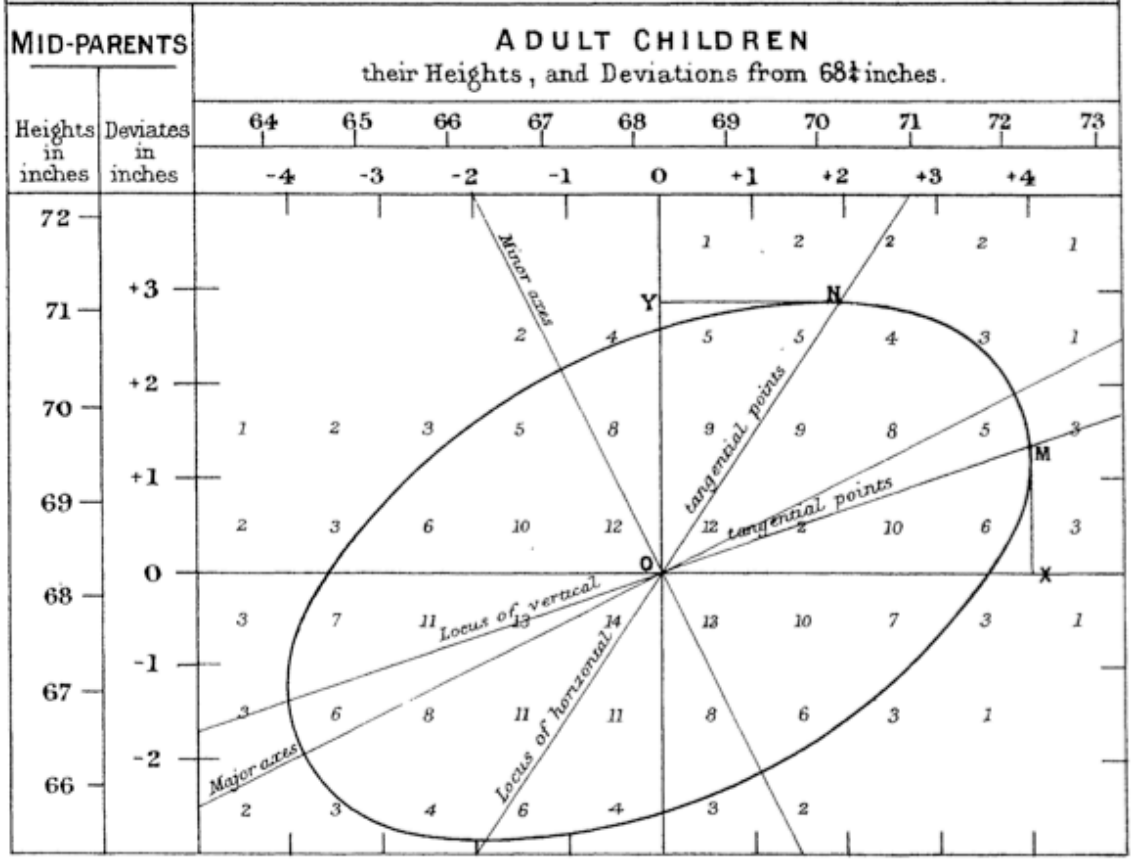
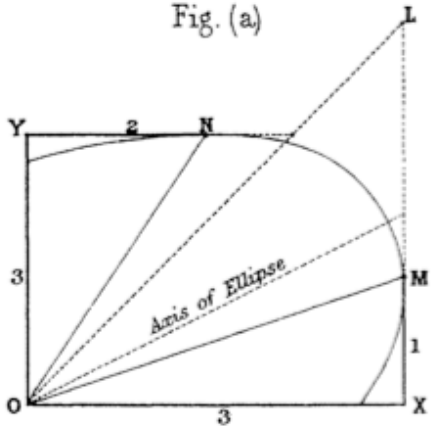


Fig. (a)



J.P. & W.R. Emslie, lith.

Journ. Anthropolog. Inst., Vol. XV, Pl. X.

Regression and least squares

We then saw that the same line emerges when we attempt to fit a model of the form

$$Y = \beta_0 + \beta_1 x + \epsilon$$

where ϵ is assumed to have mean 0 and standard deviation σ^2

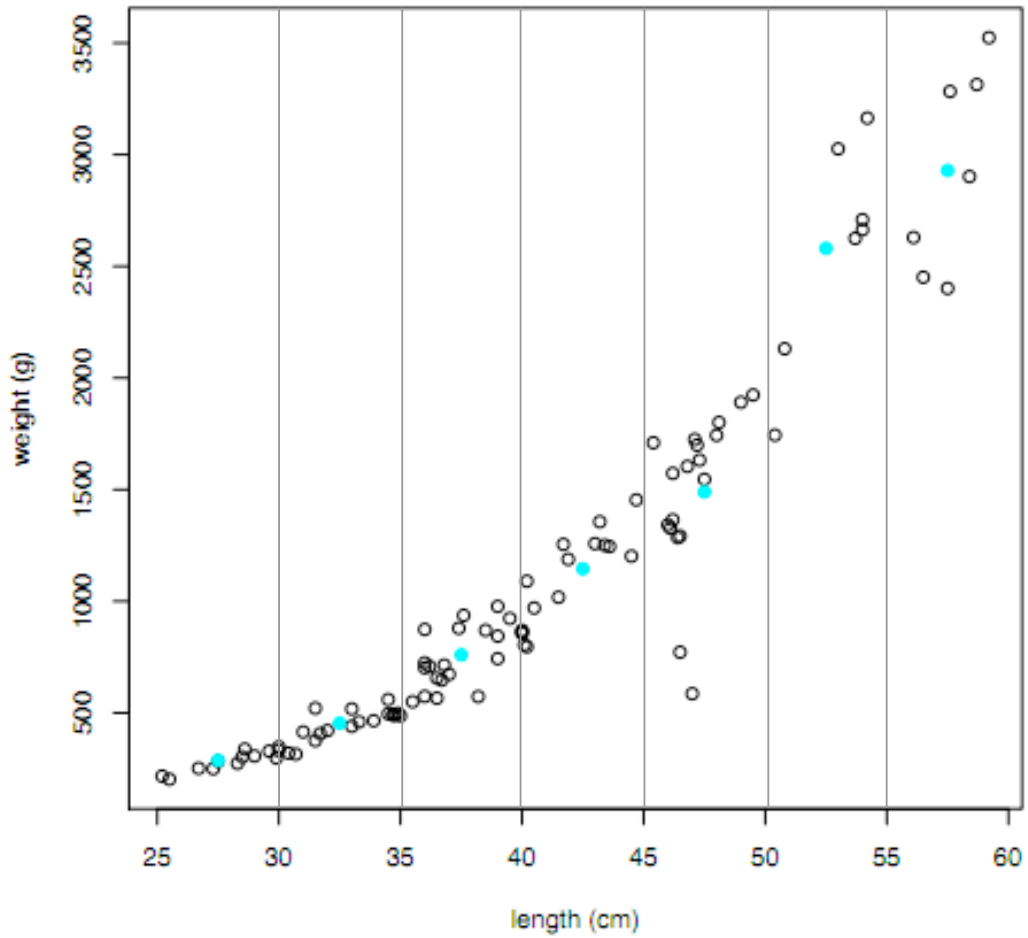
Given data $(x_1, Y_1), \ldots, (x_n, Y_n)$ we fit this model via ordinary least squares

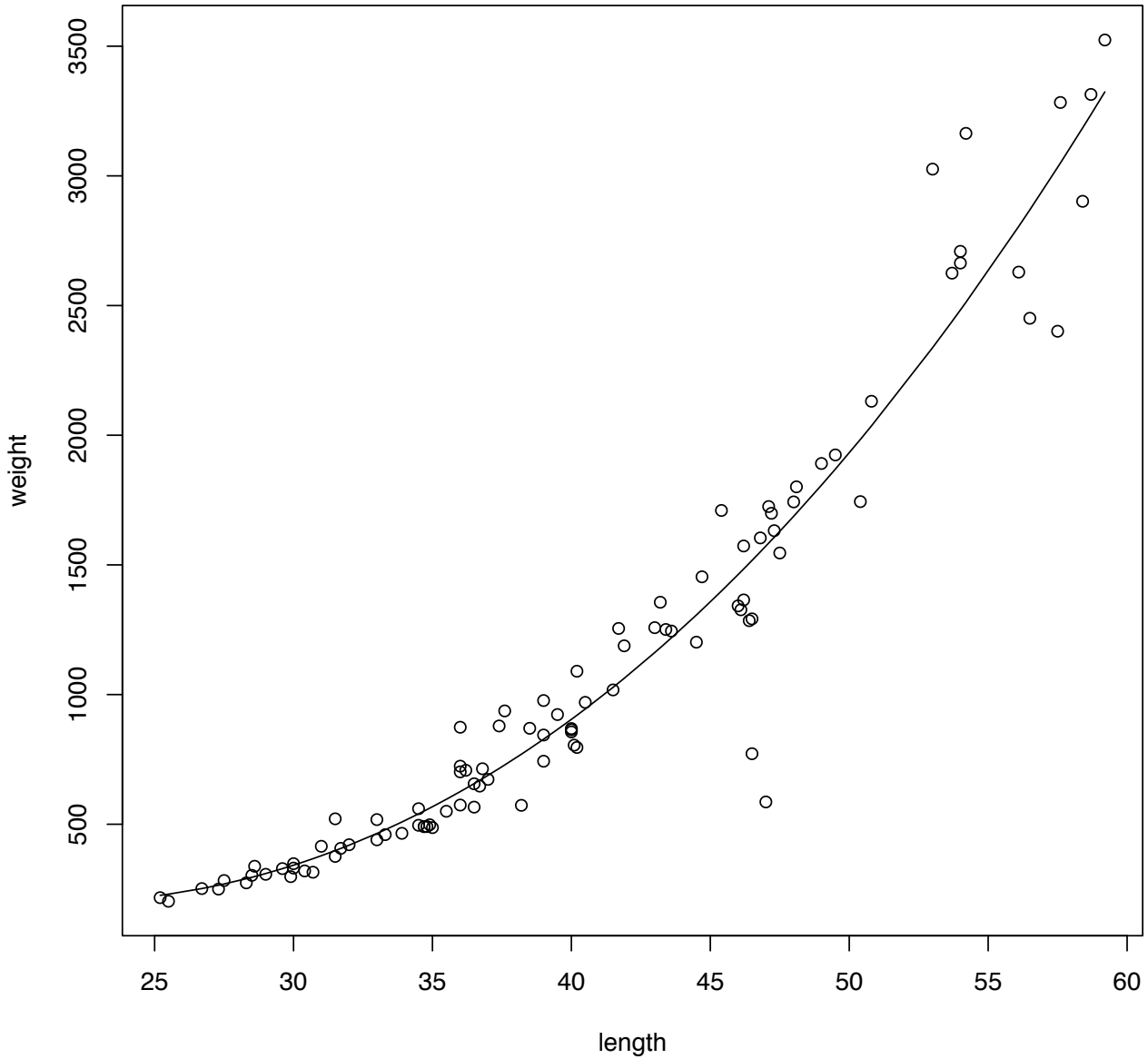
$$\sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 x_i)^2$$

which could be motivated either heuristically or via maximum likelihood if we assumed our errors were normally distributed

Last time

We then considered simple extensions to the model if we didn't really have a linear dependence on the predictor -- We saw one case, for example, in which a quadratic rather than a linear fit seemed more reasonable...

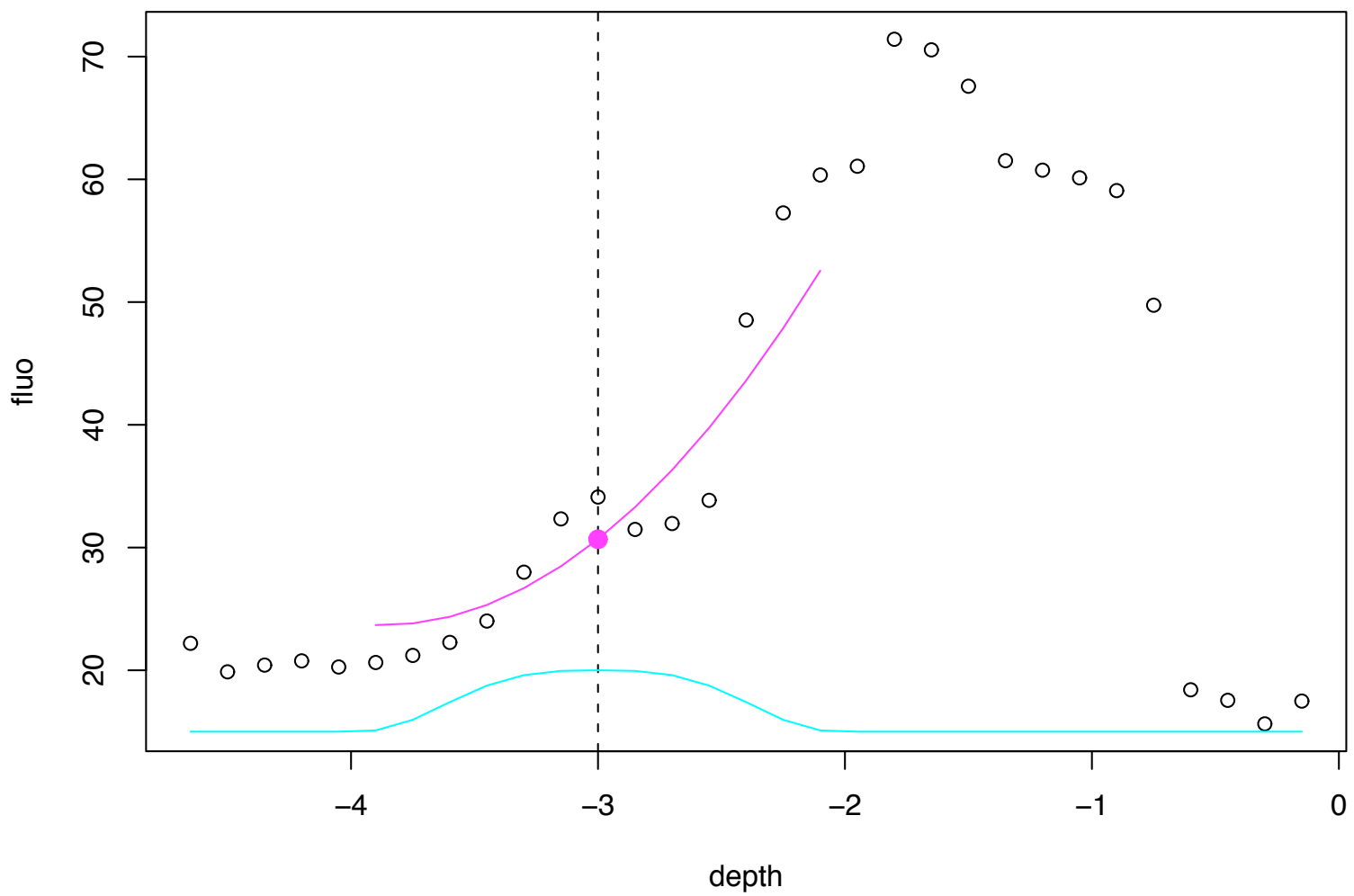

weight (g)
500
1000
1500
2000
2500
3000
3500
25
30
35
40
45
50
55
60
length (cm)

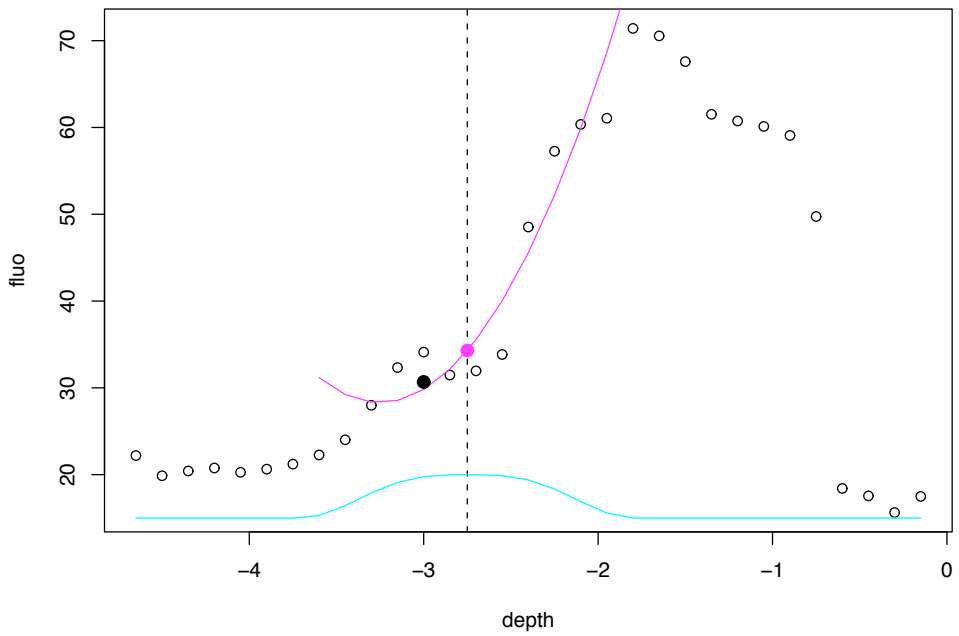

weight
500
1000
1500
2000
2500
3000
3500
25
30
35
40
45
50
55
60
length

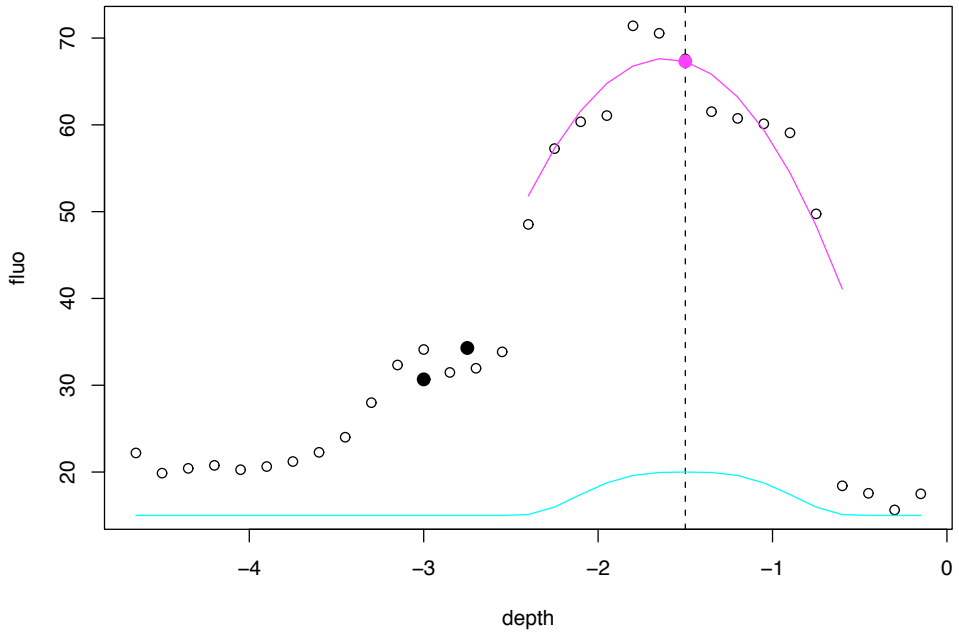
Last time

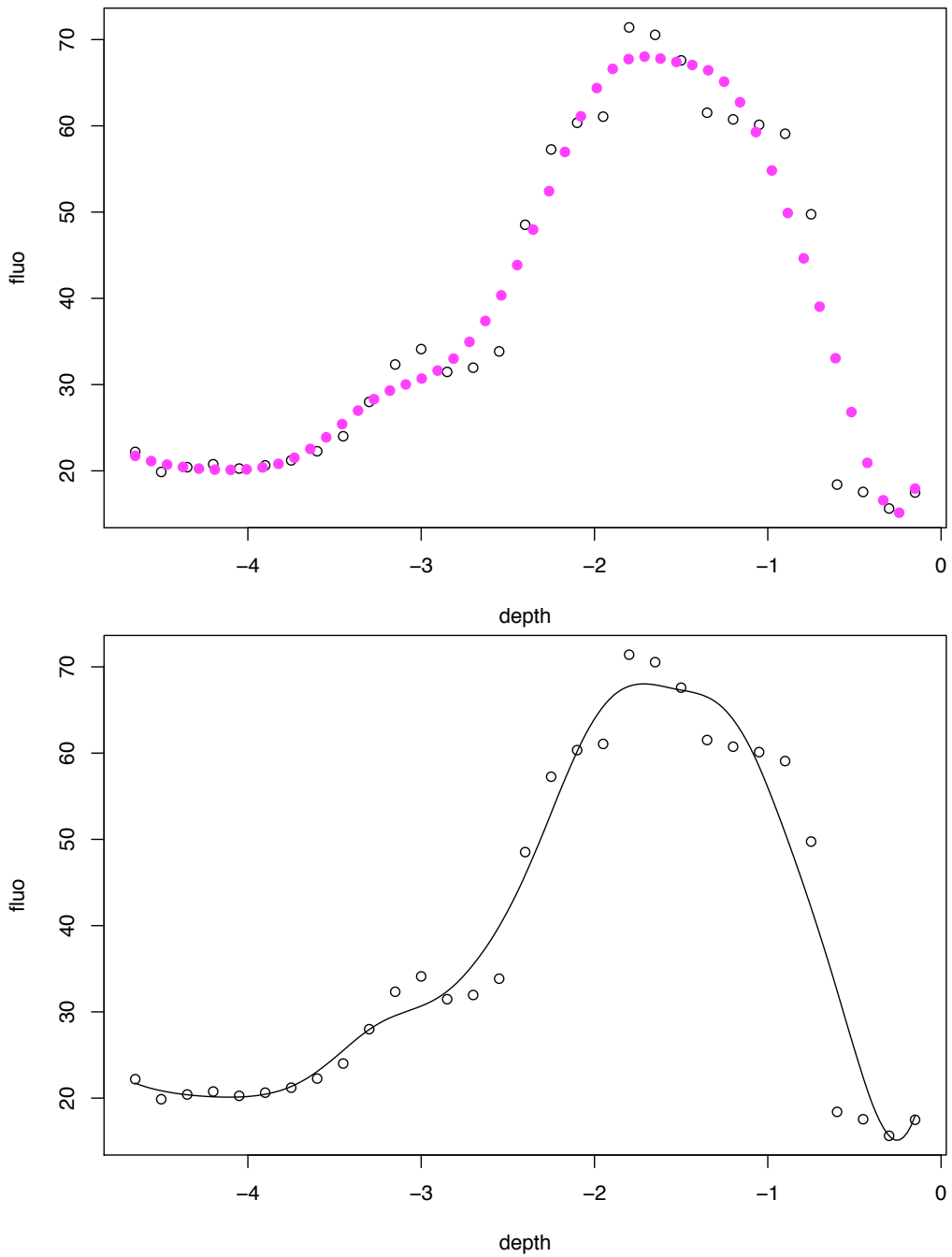
And we ended by considering what to do when a polynomial doesn't even seem reasonable -- When the shape of the data is sufficiently complex that the degree of the polynomial would have to be incredibly large

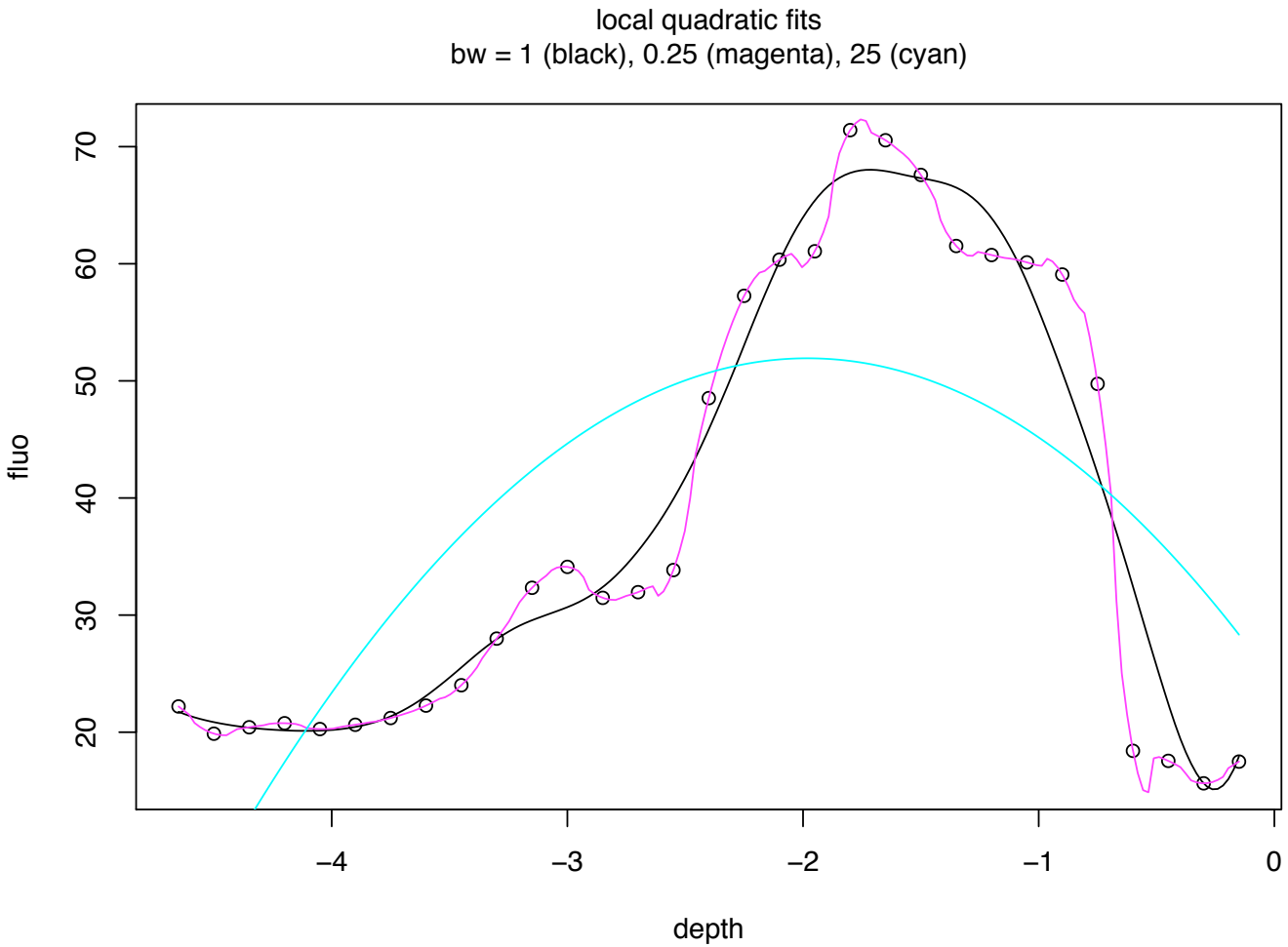
In this case, we were motivated by the idea behind a Taylor expansion (all smooth functions look, locally, like polynomials) and considered so-called local regression models...

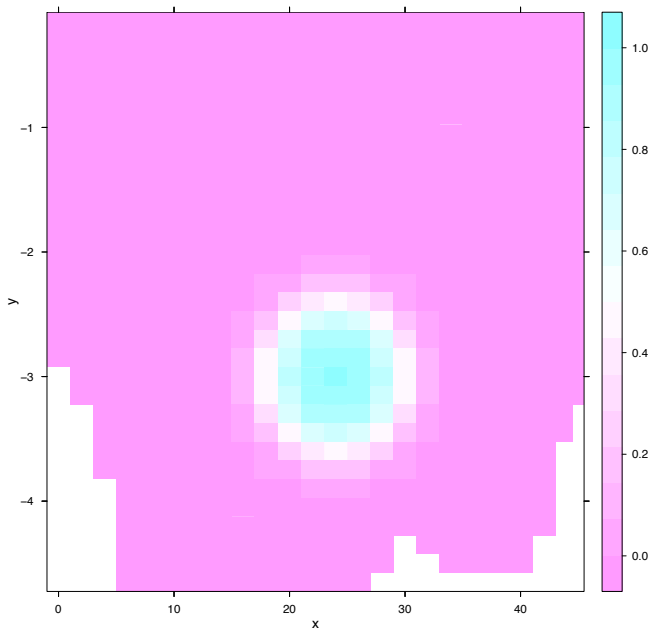

fluo
20
30
40
50
60
70
−4
−3
−2
−1
0
depth

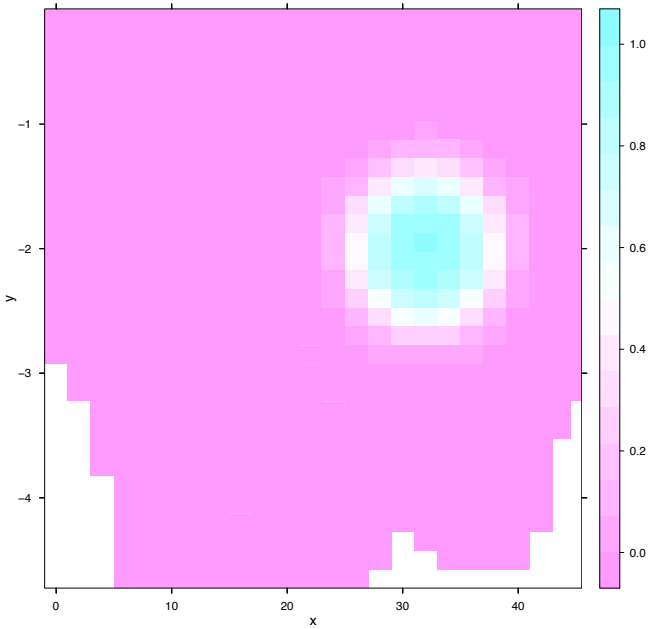

fluo
depth
20
30
40
50
60
70
−4
−3
−2
−1
0

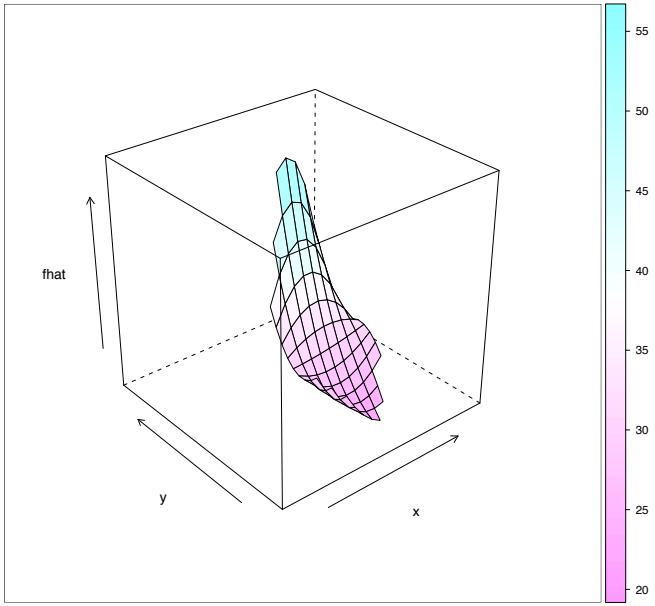

fluo
depth
20
30
40
50
60
70
−4
−3
−2
−1
0

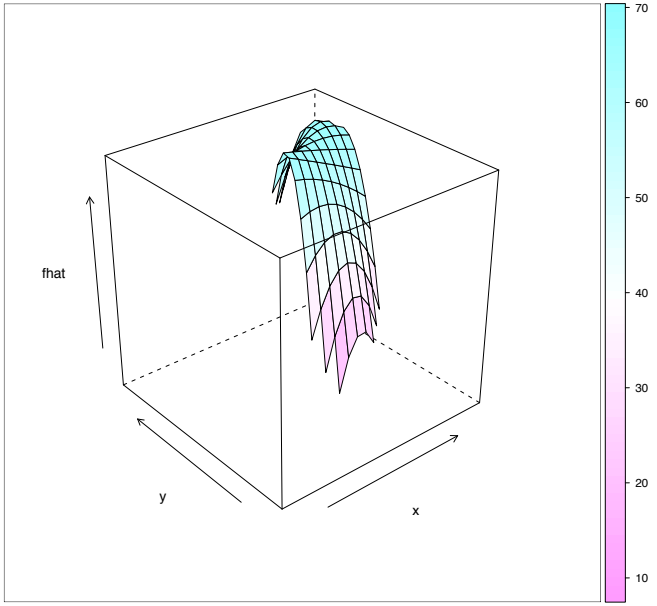

fluo
depth
20
30
40
50
60
70
−4
−3
−2
−1
0

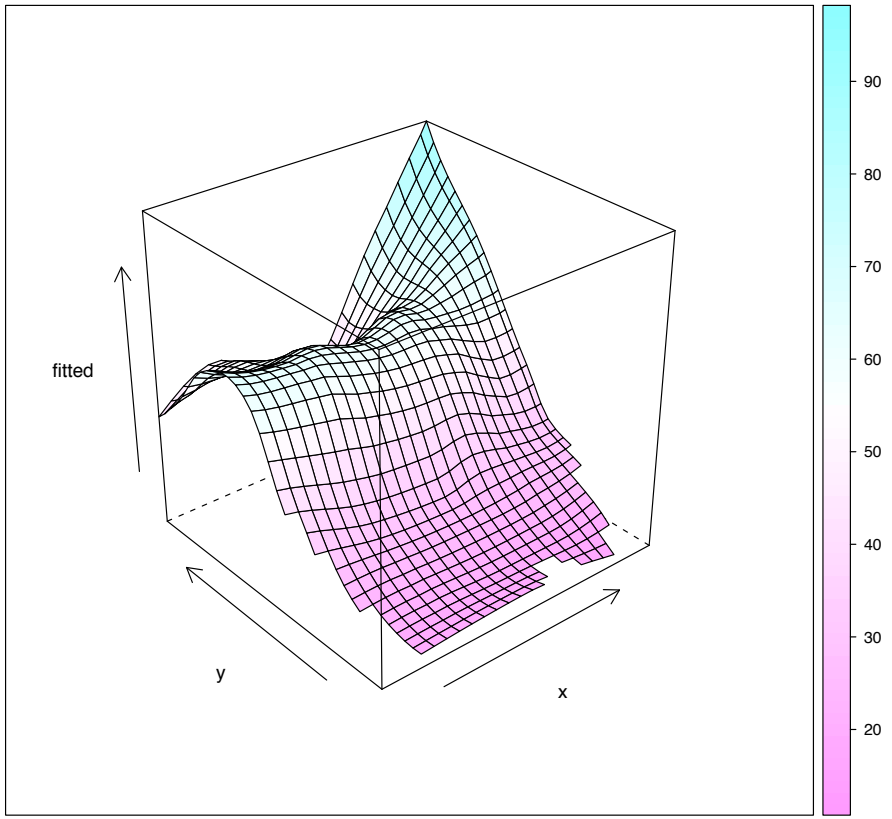

local quadratic fits
bw = 1 (black), 0.25 (magenta), 25 (cyan)
fluo
20
30
40
50
60
70
−4
−3
−2
−1
0
depth


y
x
−1
−2
−3
−4
0
10
20
30
40
1.0
0.8
0.6
0.4
0.2
0.0

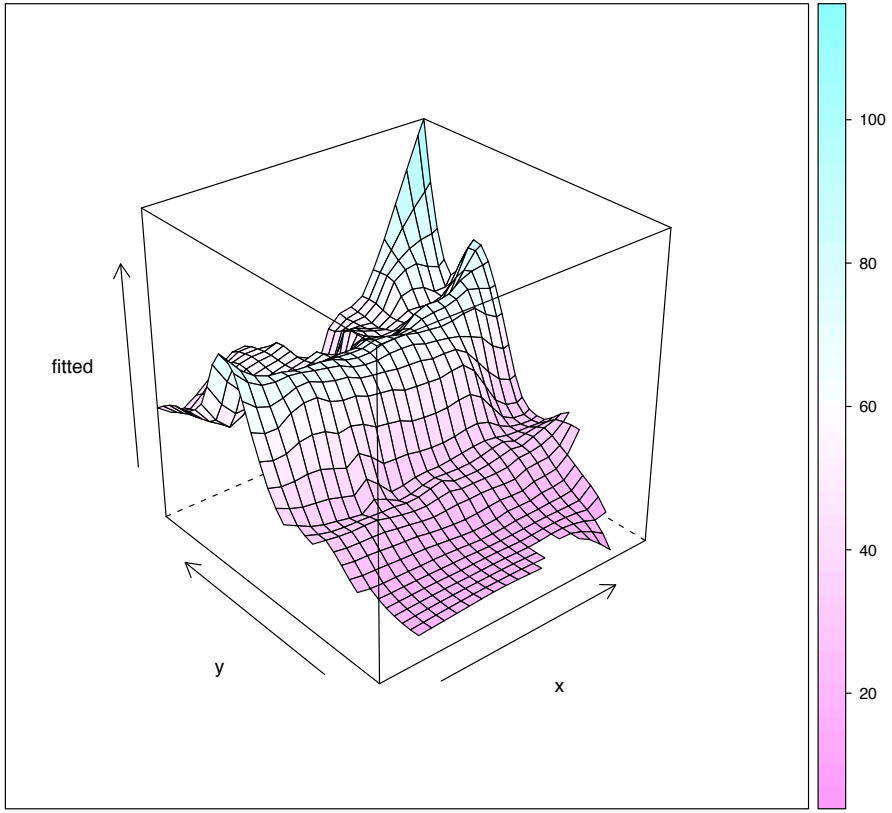

y
x
−1
−2
−3
−4
0
10
20
30
40
1.0
0.8
0.6
0.4
0.2
0.0


fhat
y
x
55
50
45
40
35
30
25
20


fhat
y
x
70
60
50
40
30
20
10




```
# load up graphics library to make lovely 3d plots

library(lattice)

# fit a local polynomial in two dimensions using the full
# lake data set

fit = loess(fluo~x+y,data=lake,enp.target=25)

# create a new data frame to plot with and plot the fit

newlake = data.frame(x=lake$x,y=lake$y,fitted=predict(fit))
wireframe(fitted~x+y,data=newlake,drape=T)
```



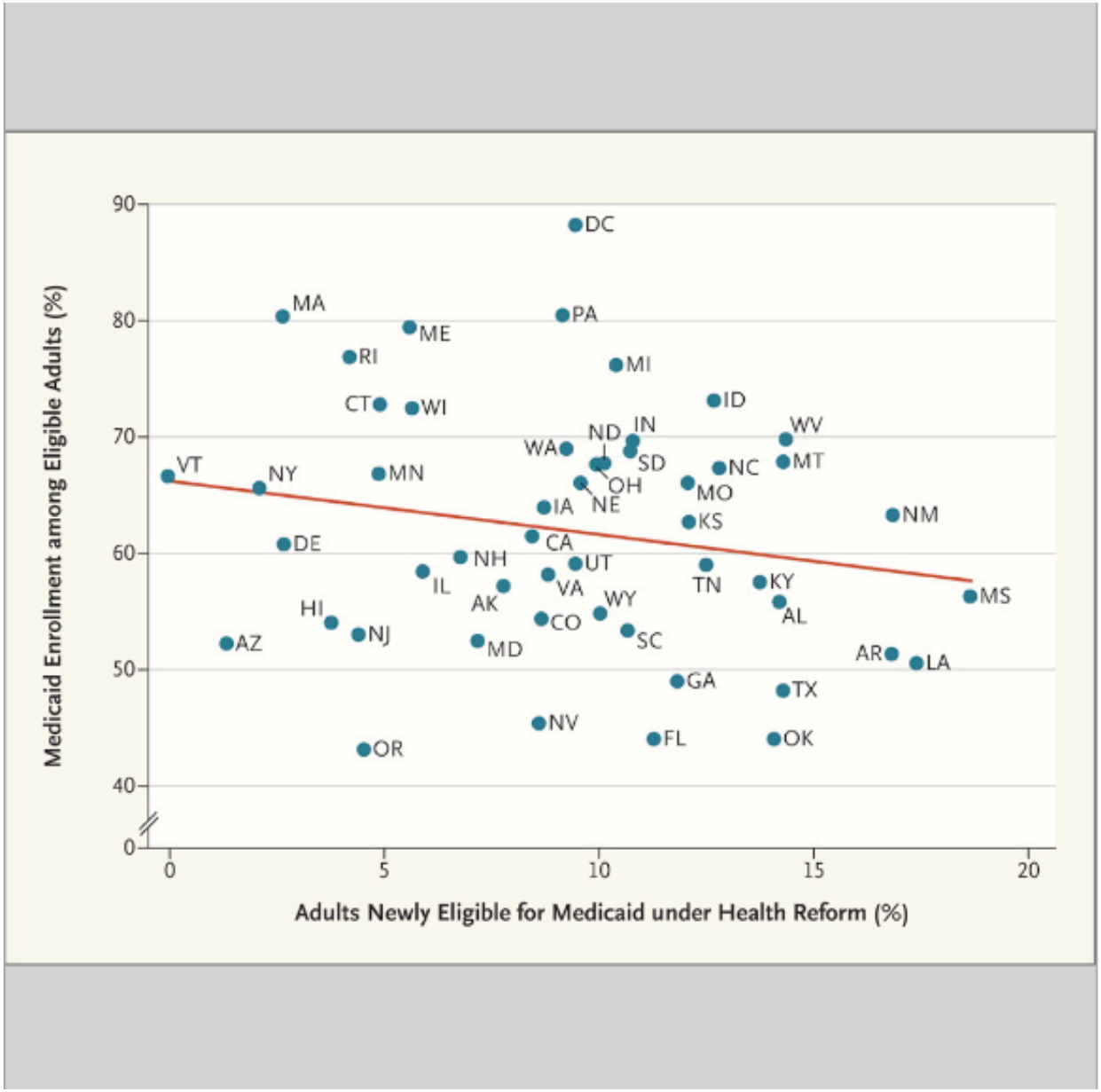

```
# load up graphics library to make lovely 3d plots

library(lattice)

# fit a local polynomial in two dimensions using the full
# lake data set; this time with 100 degrees of freedom

fit = loess(fluo~x+y,data=lake,enp.target=100)

# create a new data frame to plot with and plot the fit

newlake = data.frame(x=lake$x,y=lake$y,fitted=predict(fit))
wireframe(fitted~x+y,data=newlake,drape=T)
```

Regression today

Of course univariate (simple linear) regression or curve and surface fitting is only the start of the story

Regression has become a powerful tool in a number of quantitative disciplines -- In many cases, regression models act as a kind of **social probe**, providing researchers with a glimpse into the workings of some larger phenomenon

How we reason with these models is the subject of a big part of today's lecture -- Let's start by looking at a few examples of modern regression analysis



Medicaid Enrollment among Currently Eligible Adults (2007 through 2009) and Percentage of Adults Who Will Become Eligible in 2014 under Health Care Reform, by State.

The population sample was restricted to eligible adults with no other form of health insurance; noncitizens were excluded from the analysis. Results are based on an analysis of data from the Current Population Survey of 2007 through 2009. The red line shows the regression equation: Enrollment=0.660.46×Newly Eligible (P=0.17).

Table 2. Percent Approving Same-Sex Marriage Ban; OLS Estimates, U.S. Counties

Independent Variable	1	2	3	4
Percent Women Not Working in Labor Force	.152**	.170**	.075*	.090*
Occupational Sex Segregation	16.310***	18.048***	8.216***	9.559***
Percent Households Married with Children	.219*	.322***	.156**	.217***
Percent Same-Sex Households	−2.038	−2.237	−5.428***	−5.956***
Percent Unmarried Opposite-Sex Households	−1.682***	−1.720***	−1.314***	−1.210***
Residential Instability	−.033	−.024	−.083**	−.066*
Percent Homes Not Owner Occupied	−.076*	−.084*	−.079***	−.109***
Crime Rate		.099***		.022*
Percent Production or Construction Occupations	.194***	.234***	.067**	.063*
Percent Professional Occupations	−.178*	−.100	−.118*	−.119*
Percent Self-Employed	−.100	−.093	−.198***	−.234***
Median Family Income ($1,000s)	−.138***	−.045	−.166***	−.148***
Percent Receiving Public Assistance	−.054	.037	.125	.207
Mean Years of Education	−1.328*	−2.611***	−2.040***	−2.606***
Percent Enrolled in College	−.060	.020	−.185**	−.139
Population Density (logged)	−.111	−.332	.001	.040
Percent Urban	.029**	.011	.015**	.011
Republican Voting (percent Bush 2000)	.350***	.381***	.287***	.317***
Percent Evangelical	.180***	.179***	.031***	.020
Percent Catholic	.063***	.064***	−.025**	−.035**
Median Age	−.155	−.022	−.314***	−.247**
Percent African American	.092***	.085***	−.006	.009
Percent Latino	.043*	.024	−.108***	−.118***
LGBT Organizations	−.114	.974	−2.096***	−1.576**
Civil Rights Organizations	−.007	−.004	.073	.113
Antidiscrimination Legislation	−3.283***	−2.797***	−2.656***	−1.918*
Alabama			−6.679***	−6.132***
Arizona			−23.814***	−23.878***
Arkansas			−9.454***	−9.085***
California			−2.079*	−1.867*
Colorado			−13.112***	−12.478***
Florida			−7.906***	−8.263***
Georgia			−2.787***	−3.307***
Idaho			−17.638***	−17.591***
Kansas			−5.857***	−6.051***
Kentucky			−7.623***	−9.111***
Louisiana			−4.771***	−5.492***
Michigan			−15.107***	−15.211***
Mississippi			−2.865***	−1.619
Missouri			−5.502***	−5.583***
Montana			−8.701***	−8.893***
Nebraska			−5.916***	−5.754***
Nevada			−3.886***	−3.964***

Table 3. Regression Equations Predicting Sales, Number of Customers, Market Share, and Relative Profitability with Racial and Gender Diversity and Other Characteristics of Establishments

	Model 1	Model 2	Model 3	Model 4
Independent Variables	Sales	Customers	Market Share	Profitability
Constant	4.998***	61545.4	3.403***	3.363***
Racial diversity	.093***	433.86***	.007**	.006*
Gender diversity	.028**	195.642**	.001	.005**
Proprietorship	−.821	−370.78	−.232*	−.161
Partnership	.663	−6454.6	−.017	.256
Public corporation	−.109	7376.29	.214*	.202
Private corporation	−1.484**	−8748.7*	.008	.019
Company size	.00001*	.352*	.000**	.000**
Establishment size	.00001**	.119	.000	.000
Organization age	.013**	44.813	.001	.001
Agriculture	−1.942	4188.66	−.206	−.033
Mining	.739	−28856*	−.168	−.264
Construction	−.967	−875.7	−.152	−.036
Transportation/communications	−.052	1498.75	.119	.226
Wholesale trade	.008	−16383**	.136	−.064
Retail trade	−1.183**	7209.83*	.08	−.087
F. I. R. E.	−.683	−8335.4	−.212	.085
Business services	−1.49*	1552.28	.204*	.112
Personal services	−1.566*	1480.03**	.423**	−.001
Entertainment	−4.708**	−1504.5	−.191	−.076
Professional services	−.615	−13539**	.138	−.023
North	2.196***	23143.9***	−.06	−.039
Midwest	2.616***	14968.3***	−.023	−.073
South	1.82***	21152.8***	−.055	.059
R^2	.165***	.155***	.075**	.064**
N	506	506	469	484

Notes: Coefficients are unstandardized. For the dummy (binary) variable coefficients, significance levels refer to the difference between the omitted dummy variable category and the coefficient for the given category.
* $p < .1$; ** $p < .05$; *** $p < .01$.

Table 2. OLS Estimates of Effect of Selected Measures of Residential Segregation on Log of Total Foreclosures

	Dissimilarity Index		Isolation Index	
Variables	*B*	*SE*	*B*	*SE*
Index of Segregation				
African Americans	3.718**	.725	2.122**	.619
Hispanics	−.773	.596	.080	.656
Asians	−2.080*	.920	−2.161	1.636
Control Variables				
Housing Starts Ratio	2.980**	.960	3.067**	1.077
Wharton Land Use Index	.250**	.082	.272**	.096
Change in Housing Price Index	.082**	.024	.092**	.029
CRA-Covered Lending Share	−1.295	.912	−.810	1.061
Subprime Loan Share	3.022*	1.353	4.310**	1.581
MSA Credit Score Index	−.015*	.007	−.016*	.007
Log of Population	1.008**	.089	1.013**	.093
Percent with College Degree	−1.341	1.315	−.997	1.459
Log Median Household Income	.253	.509	.340	.515
Percent with Second Mortgage	.751	3.687	.225	4.350
Percent Workforce Unionized	−.025**	.011	−.022*	.011
Unemployment Rate	−.010	.064	.012	.071
Change in Unemployment Rate	.245**	.052	.213**	.063
Age of Housing Stock	.004	.012	.014	.013
Region				
Midwest	.434*	.200	.631**	.200
South	.042	.257	.081	.296
West	.463	.384	.679	.436
Coastal MSA	−.053	.123	.070	.133
Borders Rio Grande	−1.030**	.370	−1.054**	.380
Constant	1.960	7.557	.979	8.150
R^2	.91		.90	
Joint *F*-Test for Region	3.35*		7.97**	
Joint *F*-Test for Segregation	10.48**		6.28**	

Note: $N = 99$. Robust standard errors. Model also includes percent black, percent Hispanic, and percent Asian.

*$p < .05$; **$p < .01$ (two-tailed tests).

Regression today

In each of these cases, the model being examined is of the form

$$Y = \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

where each response Y is taken to be some linear function of a set of input variables $x_1, \ldots, x_p$ and the error ϵ is assumed to be independent of the inputs and to have mean zero

The applications on the previous page take Y to be everything from segregation measures to indices of profitability and sales

Regression today

We'll see today how these relationships are estimated (or "fit", although it's just least squares again) -- But researchers are rarely satisfied with simply producing coefficient estimates and instead they want to judge the size of the coefficients both in terms of **statistical as well as practical significance**

The coefficient β_2 in a model of the form

$$Y = \beta_1 x_1 + \cdots + \beta_p x_p + \epsilon$$

tells us how a unit change in a predictor (unemployment rate in a county or the age of its housing stock, say) affects the response (some measure of racial segregation, say), holding all other conditions the same -- **The absolute size of the coefficient, then, provides us with a sense of the practical importance of the predictor**

Regression today

We won't have access to the true coefficient β_2 , however, and instead need to reason from the estimate $\widehat{\beta}_2$ (again, we'll talk about how this is formed, but think least squares)

The statistical significance of a coefficient refers to its size relative to its standard error -- If the ratio $\widehat{\beta}_2 / \mathsf{se}(\widehat{\beta}_2)$ is less than two, for example, then a 95% confidence interval, $\widehat{\beta}_2 \pm 2\,\mathsf{se}(\widehat{\beta}_2)$, will include zero

This means zero is a "plausible" value for the population parameter β_2 and, in turn, that it's plausible that the associated predictor does not belong in the model -- That is, in an expression of the form

$$\mathsf{Y} = \beta_1 \mathsf{x}_1 + \cdots + \beta_\mathsf{p} \mathsf{x}_\mathsf{p} + \epsilon$$

if $\beta_2 = 0$, then making changes to x_2 will have no effect on the response Y

Table 2. OLS Estimates of Effect of Selected Measures of Residential Segregation on Log of Total Foreclosures

	Dissimilarity Index		Isolation Index	
Variables	*B*	*SE*	*B*	*SE*
Index of Segregation				
African Americans	3.718**	.725	2.122**	.619
Hispanics	−.773	.596	.080	.656
Asians	−2.080*	.920	−2.161	1.636
Control Variables				
Housing Starts Ratio	2.980**	.960	3.067**	1.077
Wharton Land Use Index	.250**	.082	.272**	.096
Change in Housing Price Index	.082**	.024	.092**	.029
CRA-Covered Lending Share	−1.295	.912	−.810	1.061
Subprime Loan Share	3.022*	1.353	4.310**	1.581
MSA Credit Score Index	−.015*	.007	−.016*	.007
Log of Population	1.008**	.089	1.013**	.093
Percent with College Degree	−1.341	1.315	−.997	1.459
Log Median Household Income	.253	.509	.340	.515
Percent with Second Mortgage	.751	3.687	.225	4.350
Percent Workforce Unionized	−.025**	.011	−.022*	.011
Unemployment Rate	−.010	.064	.012	.071
Change in Unemployment Rate	.245**	.052	.213**	.063
Age of Housing Stock	.004	.012	.014	.013
Region				
Midwest	.434*	.200	.631**	.200
South	.042	.257	.081	.296
West	.463	.384	.679	.436
Coastal MSA	−.053	.123	.070	.133
Borders Rio Grande	−1.030**	.370	−1.054**	.380
Constant	1.960	7.557	.979	8.150
R^2	.91		.90	
Joint *F*-Test for Region	3.35*		7.97**	
Joint *F*-Test for Segregation	10.48**		6.28**	

Note: $N = 99$. Robust standard errors. Model also includes percent black, percent Hispanic, and percent Asian.

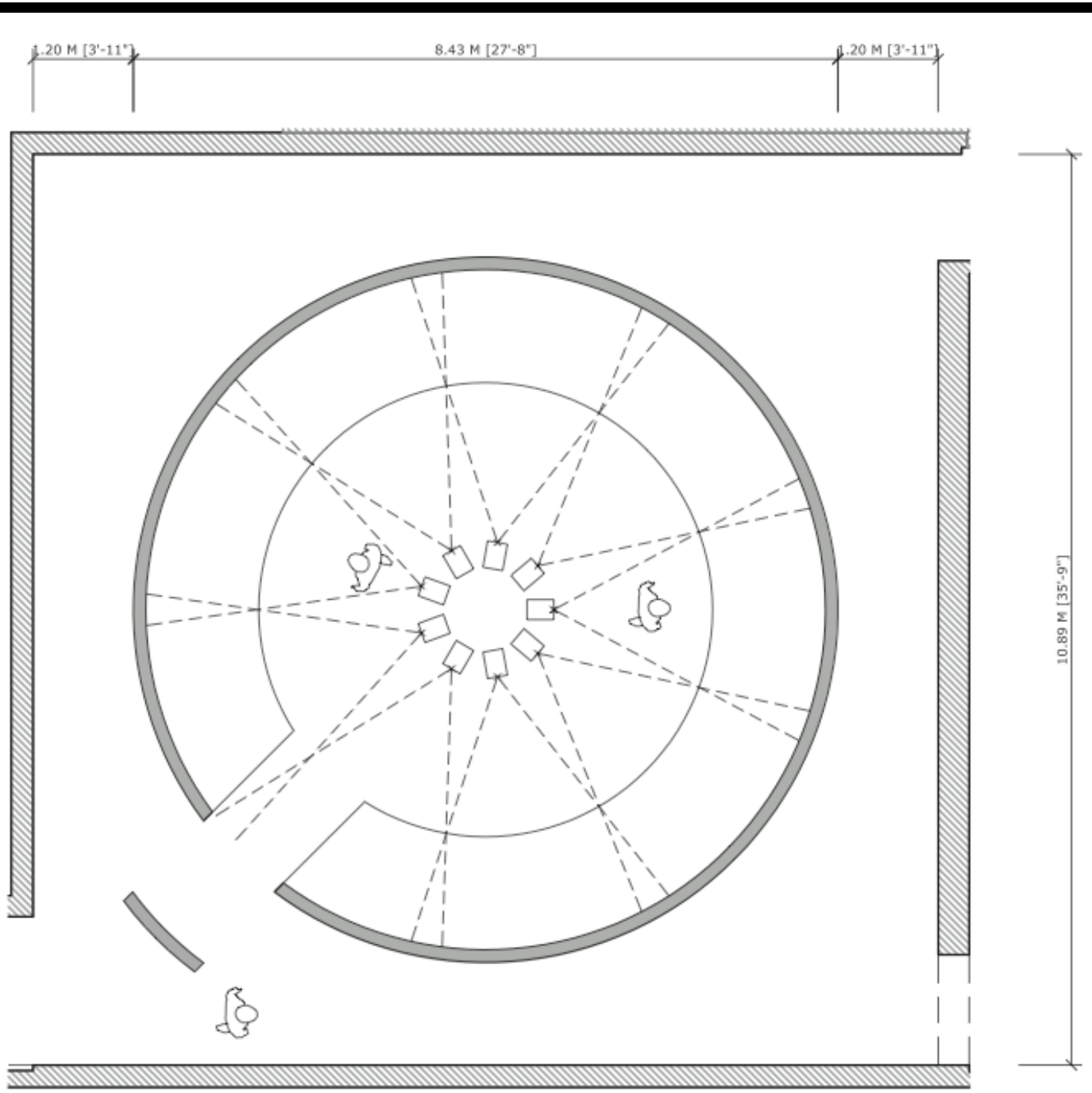
*$p < .05$; **$p < .01$ (two-tailed tests).

Digging in

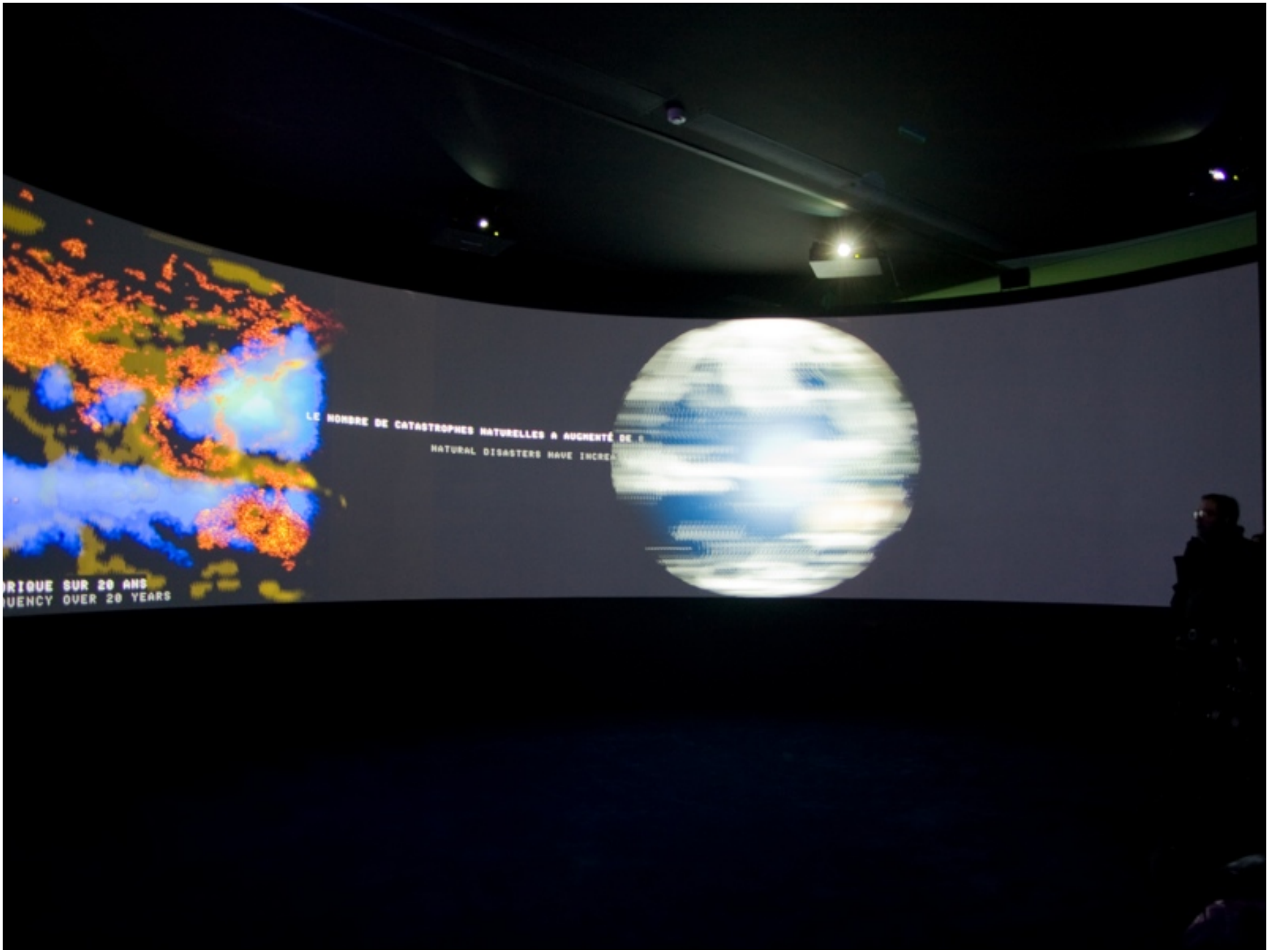
Today we'll take this idea on a bit more completely and illustrate how estimates are formed for the multiple linear regression model -- We will examine inference procedures as well

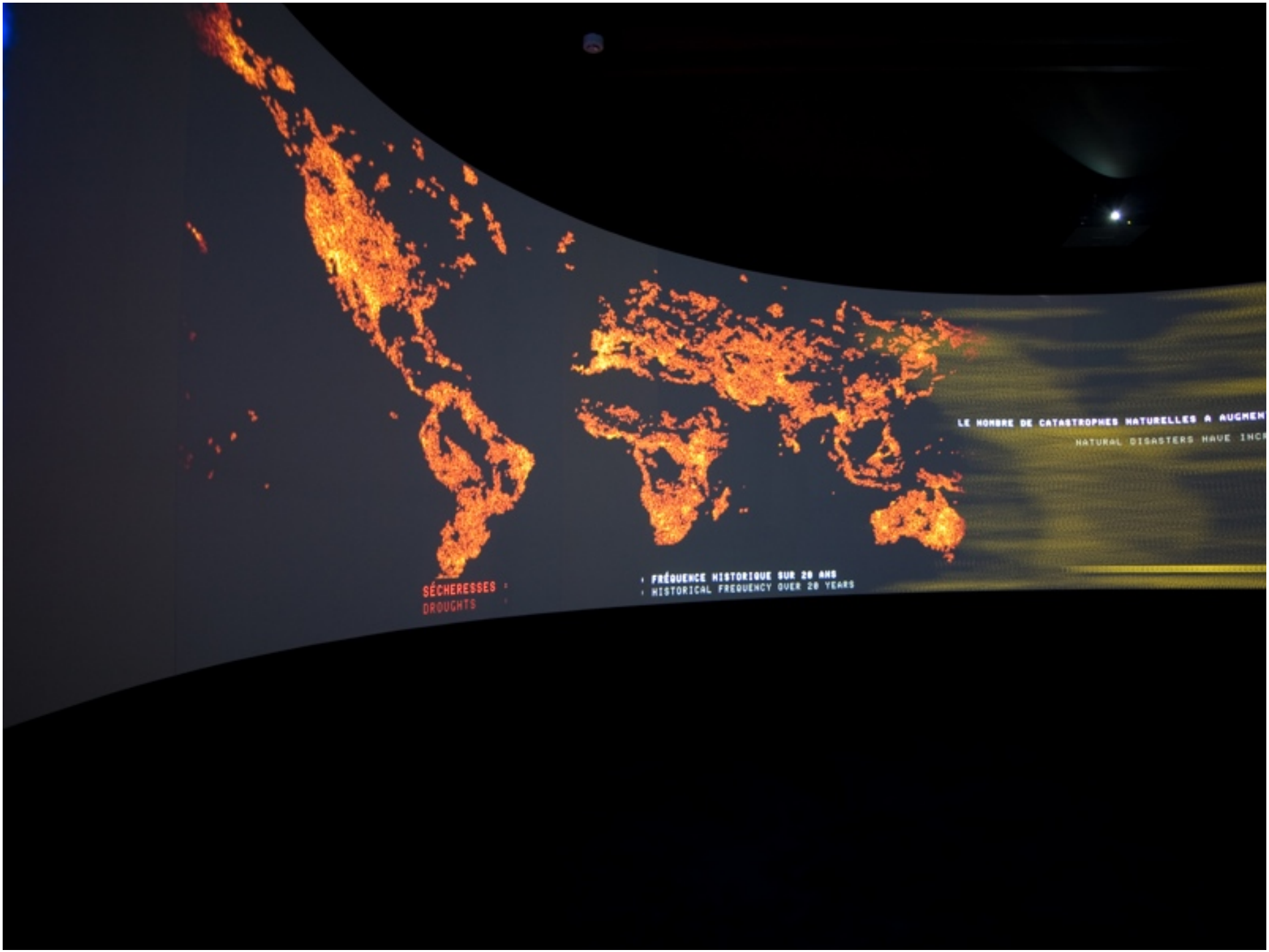
Our test case will be a relatively recent class of applications relating to the vulnerability of different countries to the effects of climate change, and specifically their ability to cope with extreme climatic events

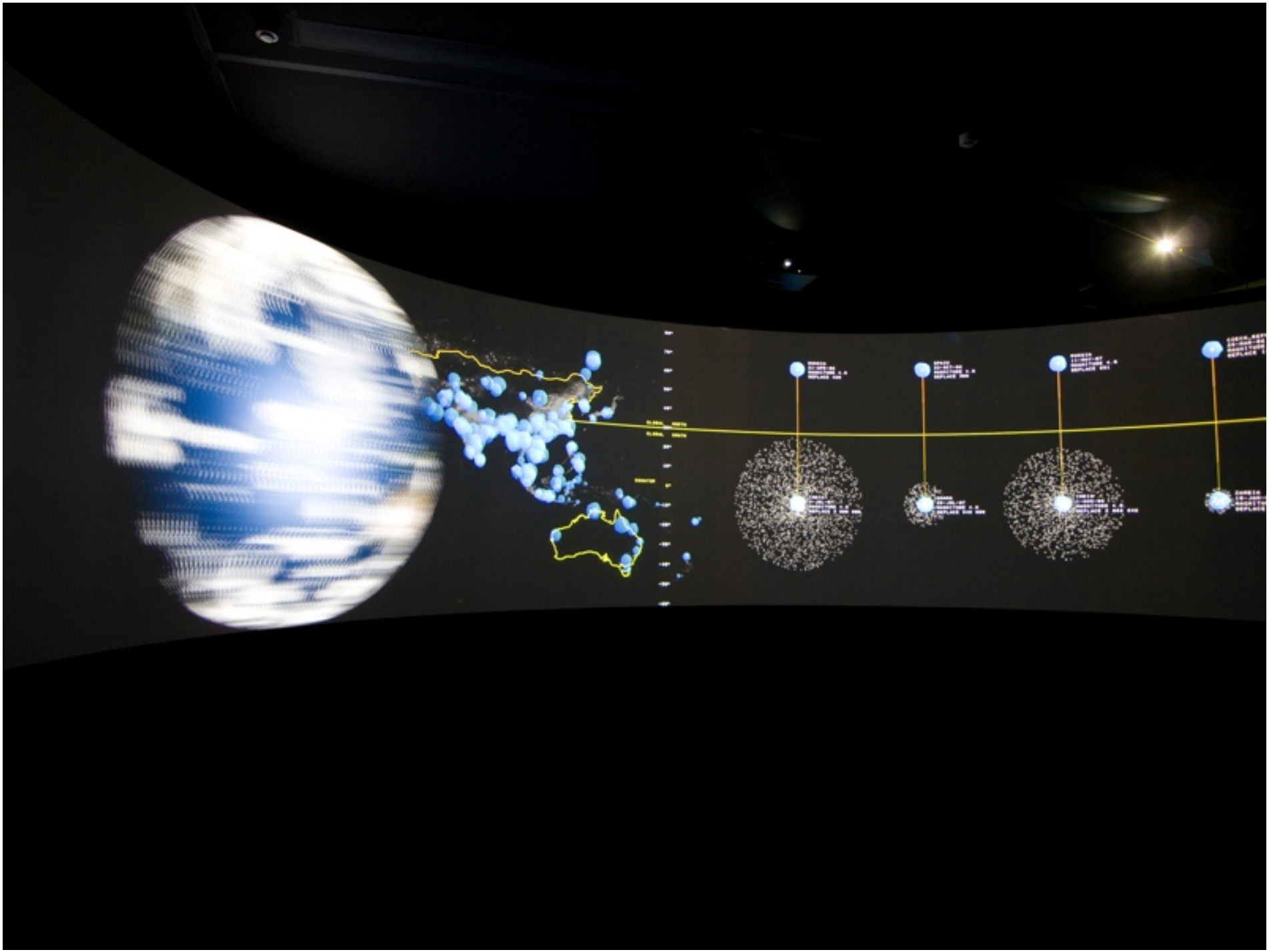
In terms of background....

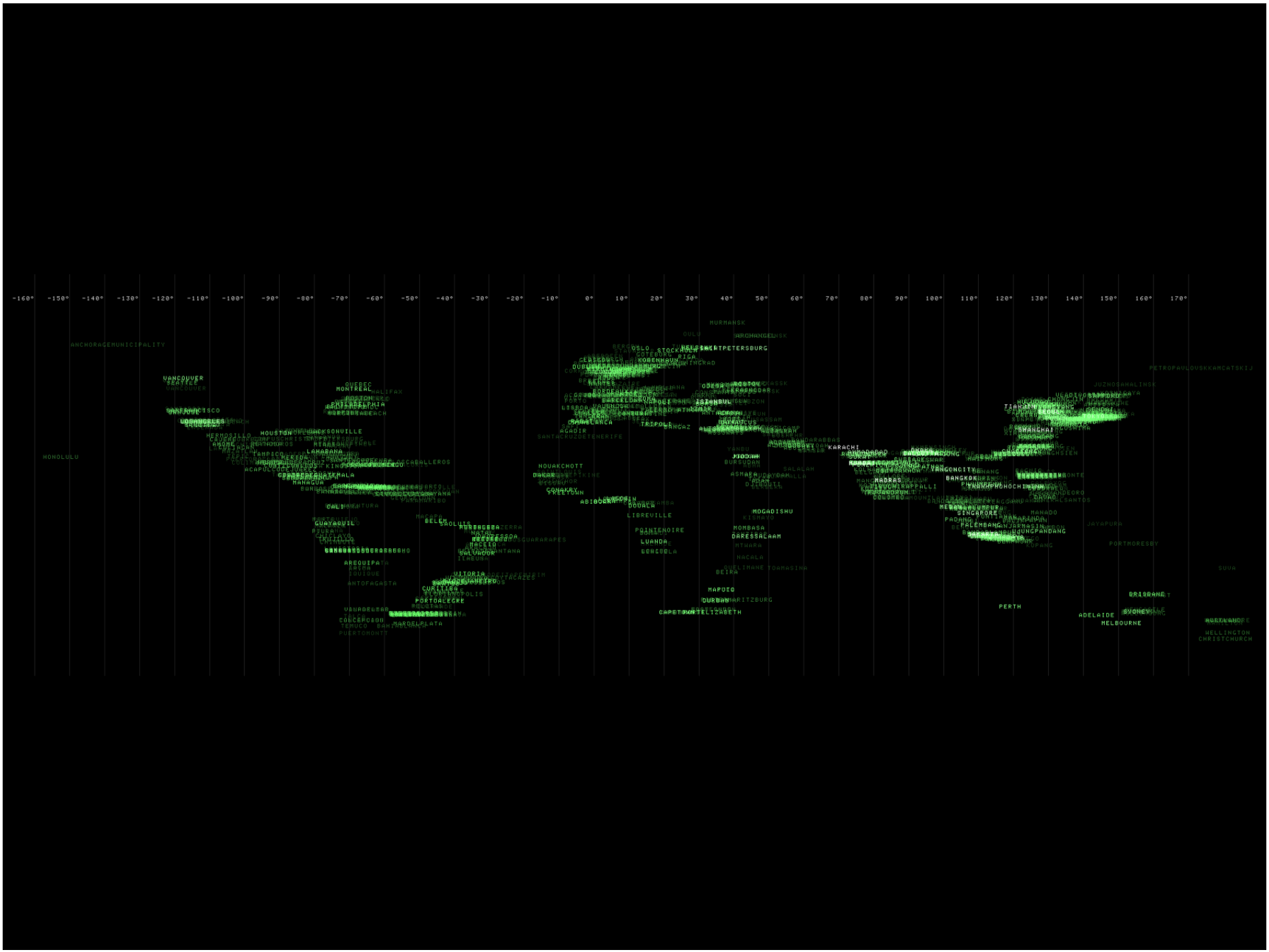

1.20 M [3'-11"]
8.43 M [27'-8"]
1.20 M [3'-11"]
10.89 M [35'-9"]

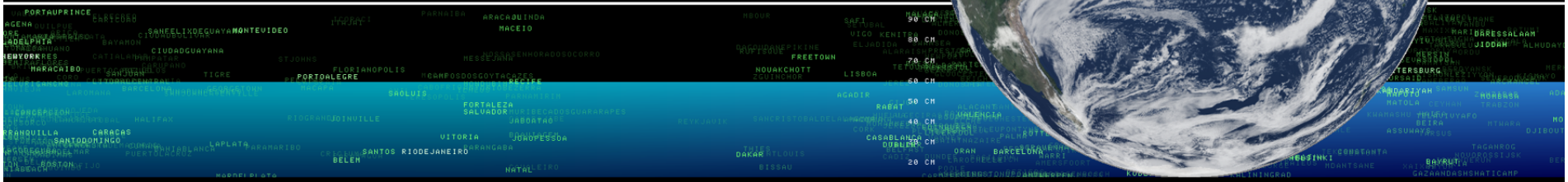

UGANDA
KENYA
RWANDA
BURUNDI
UNITED REP. OF TANZANIA
OCT 1996


LE NOMBRE DE CATASTROPHES NATURELLES A AUGMENTÉ DE
NATURAL DISASTERS HAVE INCRE
RIQUE SUR 20 ANS
UENCY OVER 20 YEARS


LE NOMBRE DE CATASTROPHES NATURELLES A AUGMEN
NATURAL DISASTERS HAVE INC
SÉCHERESSES
DROUGHTS
FRÉQUENCE HISTORIQUE SUR 20 ANS
HISTORICAL FREQUENCY OVER 20 YEARS




-160° -150° -140° -130° -120° -110° -100° -90° -80° -70° -60° -50° -40° -30° -20° -10° 0° 10° 20° 30° 40° 50° 60° 70° 80° 90° 100° 110° 120° 130° 140° 150° 160° 170°
MURMANSK
ANCHORAGEMUNICIPALITY
VANCOUVER
SEATTLE
MONTREAL
QUEBEC
HALIFAX
BOSTON
PHILADELPHIA
SANFRANCISCO
LOSANGELES
HOUSTON
JACKSONVILLE
HERMOSILLO
MONTERREY
LAHABANA
MERIDA
ACAPULCO
MANAGUA
HONOLULU
CALI
GUAYAQUIL
BELEM
SALVADOR
AREQUIPA
ANTOFAGASTA
CURITIBA
PORTOALEGRE
VITORIA
OSLO
STOCKHOLM
HELSINKI
ST.PETERSBURG
ARKHANGELSK
COPENHAGEN
RIGA
DUBLIN
BERGEN
NANTES
BARCELONA
LISBON
MADRID
ISTANBUL
ODESSA
ROSTOV
KRASNODARSK
TRIPOLI
BENGAZ
ALGIERS
CASABLANCA
AGADIR
SANTACRUZDETENERIFE
HOUAKCHOTT
DAKAR
CONAKRY
FREETOWN
ABIDJAN
DOUALA
LIBREVILLE
POINTENOIRE
LUANDA
BENGUELA
NACALA
BEIRA
MAPUTO
DURBAN
PIETERMARITZBURG
CAPETOWN
PORTELIZABETH
KARACHI
MUMBAI
MADRAS
BANGALORE
COCHIN
COLOMBO
CALCUTTA
DHAKA
RANGOON
BANGKOK
SINGAPORE
KUALALUMPUR
PENANG
PALEMBANG
JAKARTA
SURABAYA
DENPASAR
MANADO
JAYAPURA
PORTMORESBY
DARWIN
PERTH
ADELAIDE
MELBOURNE
SYDNEY
BRISBANE
SUVA
AUCKLAND
WELLINGTON
CHRISTCHURCH
PETROPAVLOVSKKAMCATSKIJ
JUZNOSAHALINSK
VLADIVOSTOK
KHABAROVSK
SAPPORO
HUANGDAO
TIANJIN
SHANGHAI
HONGKONG
TAIPEI
MANILA
SEOUL
PUSAN
OSAKA
NAGOYA
TOKYO
YOKOHAMA


90 CM
80 CM
70 CM
60 CM
50 CM
40 CM
30 CM
20 CM
PORTAUPRINCE
SANFELIXDEGUAYANA
MONTEVIDEO
MACEIO
ARACAJUINDA
CIUDADGUAYANA
PORTOALEGRE
FLORIANOPOLIS
MARACAIBO
SAOLUIS
FORTALEZA
SALVADOR
RECIFE
JABOATAO
JOINVILLE
VITORIA
JUNDIAI
SANTOSRIODEJANEIRO
BELEM
NATAL
CARACAS
SANTODOMINGO
LAPLATA
BOSTON
FREETOWN
NOUAKCHOTT
LISBOA
DAKAR
BISSAU
RABAT
AGADIR
CASABLANCA
DUBLIN
ORAN
BARCELONA
VALENCIA
BARESSALAAM
JIDDAH
MATOLA
BEIRA
BAYKUT

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, Autonomou
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.”

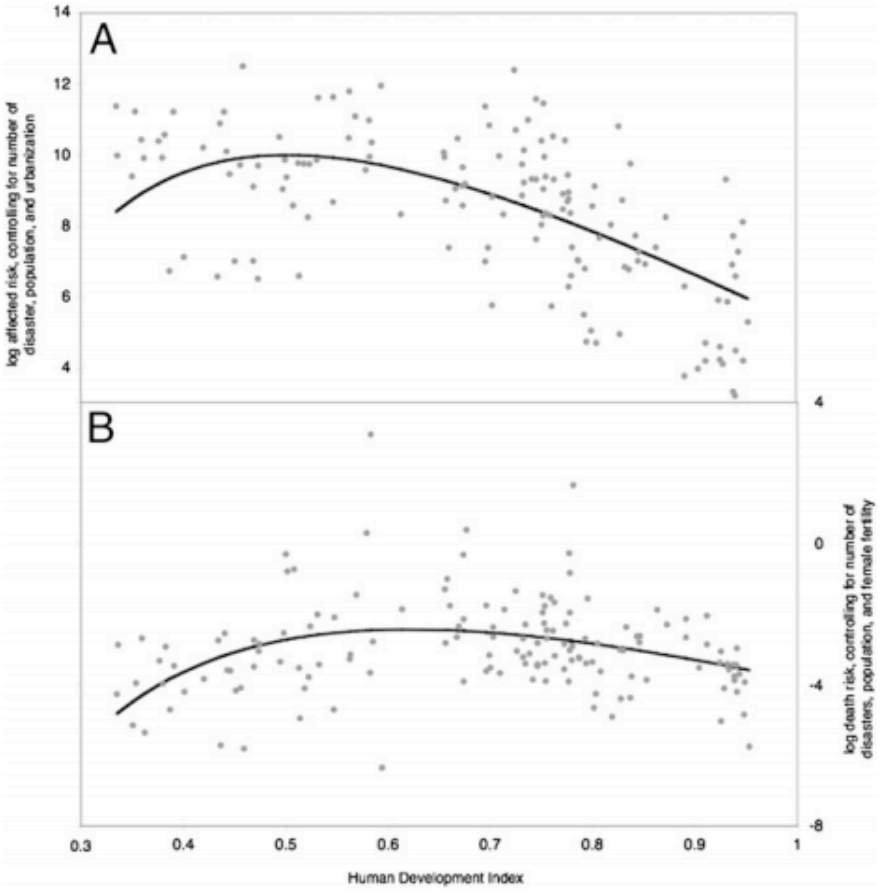


Fig. 1. Relationship between risk and HDI for (*A*) the number of people affected, i.e., needing emergency or recovery assistance, by a flood, drought, or cyclone, per million of population, and (*B*) the number of people killed. Each dot represents a country in the CRED database during the period 1990–2007, 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

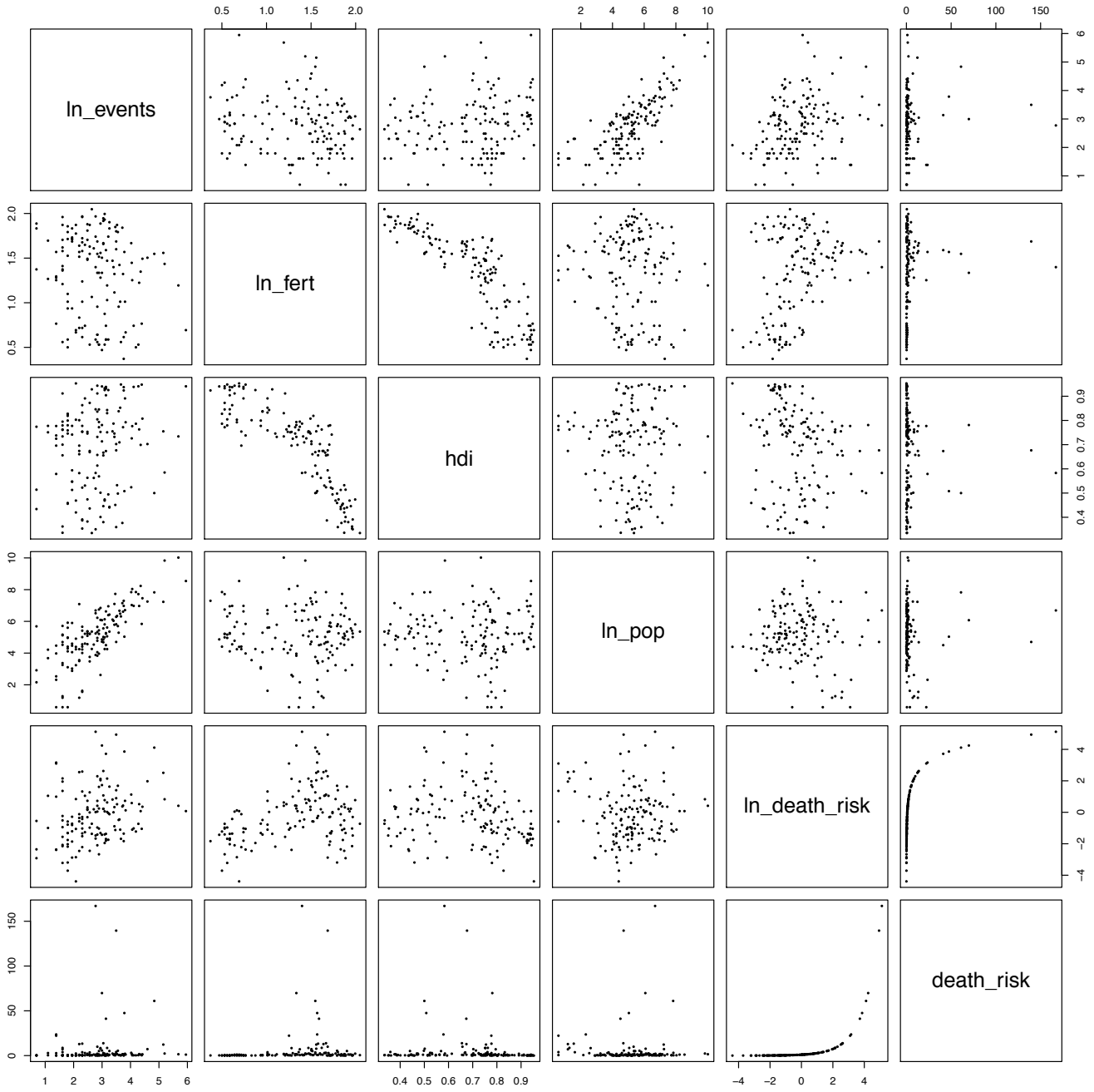
`ln_fert` the natural logarithm of an estimate of the country's female fertility

`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

```
> vul$country[order(vul$hdi)]

  [1] Niger                          Sierra Leone
  [3] Mali                           Burkina Faso
  [5] Mozambique                     Guinea Bissau
  [7] Ethiopia                       Burundi
  [9] Chad                           Central African Rep
 [11] Rwanda                         Zaire/Congo Dem Rep
 [13] Benin                          Cote d'Ivoire
 [15] Zambia                         Malawi
 [17] Tanzania Uni Rep               Angola
 ...
[125] Portugal                       Greece
[127] Hong Kong (China)              Israel
[129] Germany                        Italy
[131] New Zealand                    Ireland
[133] Spain                          Austria
[135] United Kingdom                 Belgium
[137] France                         Switzerland
[139] Japan                          United States
[141] Netherlands                    Australia
[143] Canada                         Norway
144 Levels: Albania Algeria Angola Argentina Armenia Australia ... Zimbabwe
```

Comment

This kind of study is not atypical for a modern regression analysis -- A researcher has a **socially important research** question and attempts to probe some hypothesis about its underlying structure by assembling **collections of variables representing inputs and a response**

A regression is then fit and the table of coefficients (or various regression diagnostics) are consulted to see which of the various factors are both **practically important and statistically significant**

I'll refrain from commenting on the wisdom of this general strategy except to say that there are a lot of well-meaning, hard working investigators "out there" that are trying to reason from data about complicated political, social, financial questions -- Regression can be a valuable tool if used correctly

```
# first, load the vulnerability data

> load(url("http://www.stat.ucla.edu/~cocteau/stat105/data/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

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

# 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

We'll talk about how the estimates are derived and their standard errors computed both via the bootstrap as well as the classical formula-based approach (quoted in the table here)

The column of t-statistics (ratios of coefficient estimates to their standard errors) is then interpreted and *'s assigned according to whether zero is in a 95%, 99% or 99.9% confidence interval

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

```
> tss = sum((vul$ln_death_risk-mean(vul$ln_death_risk)^2)
> 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 linear model

We suppose our data are given by pairs Y_i and $m_i = (x_{i1}, \ldots, x_{ip})$ for i ranging from 1 to the sample size n -- We then assume that for each i

$$Y_i = x_{i1}\beta_1^* + \cdots + \beta_p^* x_{ip} + \epsilon_i$$

where the errors $\epsilon_1, \ldots, \epsilon_n$ are independent, identically distributed random variables with mean zero and common standard deviation σ^*

The vulnerability data cover 144 countries and so our sample size n=144 -- In the model two slides back, are estimating p=5 coefficients (four of them for predictors and one for the intercept or constant term)

Note

In this case the researchers worked with data from as many countries as they could -- **There was no random sampling involved in the selection of the countries**

In general, there are two kinds of regression settings -- In one the predictor variables are considered to be **random variables** (think sampling fish from the rivers in South Carolina) and in the other they are **fixed** (here, with countries or perhaps via a designed experiment)

How we choose to treat the predictors will affect our analysis as we will see -- For the moment, I'll use X if a predictor is random and x if it is fixed (or we choose to make all of our inferences conditional on its value)

This sounds cryptic but the difference will be clear shortly...

The normal linear model

The normal linear model adds the assumption that the errors have a normal distribution with mean 0 and common standard deviation σ^*

If we assemble our predictors into a matrix M (for Model matrix -- also known as the design matrix)

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

and introduce the vectors $\beta^* = (\beta_1^*, \ldots, \beta_p^*)$, $\epsilon = (\epsilon_1, \ldots, \epsilon_n)$ and $Y = (Y_1, \ldots, Y_n)$; then we can write our model in the form

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

(Make sure the dimensions of the matrices match!)

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

Estimation

We operationalize this notion via the least squares criterion -- That is we will determine estimates for the unknown parameters $\beta^* = (\beta_1^*, \ldots, \beta_p^*)$ and σ^* via ordinary least squares

Whether we motivate this heuristically or via an MLE for the normal linear model, we want to choose $\beta = (\beta_1, \ldots, \beta_p)$, say, so as to minimize

$$\sum_{i=1}^{n} (Y_i - \beta_1 x_{i1} - \cdots - \beta_p x_{ip})^2$$

Taking partial derivatives, we can derive the so-called normal equations

$$\begin{aligned}
\beta_1 \sum x_{i1}^2 + \beta_2 \sum x_{i1}x_{i2} + \cdots + \beta_p \sum x_{i1}x_{ip} &= \sum Y_i x_{i1} \\
\beta_1 \sum x_{i1}x_{i2} + \beta_2 \sum x_{i2}^2 + \beta_3 \sum x_{i2}x_{i3} + \cdots + \beta_p \sum x_{i2}x_{ip} &= \sum Y_i x_{i2} \\
&\vdots \\
\beta_1 \sum x_{i1}x_{ip} + \cdots + \beta_{p-1} \sum x_{i,p-1}x_{ip} + \beta_p \sum x_{ip}^2 &= \sum Y_i x_{ip}
\end{aligned}$$

Estimation

We can rewrite these equations more compactly as follows

$$M^t M\beta = M^t Y$$

Now, assuming the matrix $M^t M$ is invertible (oh, boy, here comes the linear algebra!) we can form an estimate of the regression coefficients using the (symbolic!) manipulation

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

Similarly, the conditional mean (the fitted value) for the ith data point is

$$\widehat{Y}_i = \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip}$$

which we can write in matrix notation as

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

A hat

The matrix H is known as the hat matrix (for the obvious reason that it carries our observed data into an estimate of the associated conditional means, the fitted value, in effect placing a “hat” on Y)

We can derive some simple properties of H easily -- For example H is symmetric (check!) and it’s idempotent

$$H^2 = HH = M(M^tM)^{-1}M^tM(M^tM)^{-1}M^t = M(M^tM)^{-1}M^t = H$$

We can compute the residuals from our fit $\widehat{\epsilon}_i = Y_i - \widehat{Y}_i$ as $\widehat{\epsilon} = (I - H)Y$, where we have set $\widehat{\epsilon} = (\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n)$ -- The residual sum of squares can be written as

$$\sum_{i=1}^{n} \widehat{\epsilon}_i^2 = \widehat{\epsilon}^t\epsilon = Y^t(I - H)(I - H)Y = Y^t(I - H)Y$$

A hat

We can also do a bit of geometry while we're here -- Because the hat matrix is symmetric and idempotent, we can show that the residual vector

$$\widehat{\epsilon} = (\mathsf{I} - \mathsf{H})\mathsf{Y}$$

is orthogonal to the estimated means (the fitted values)

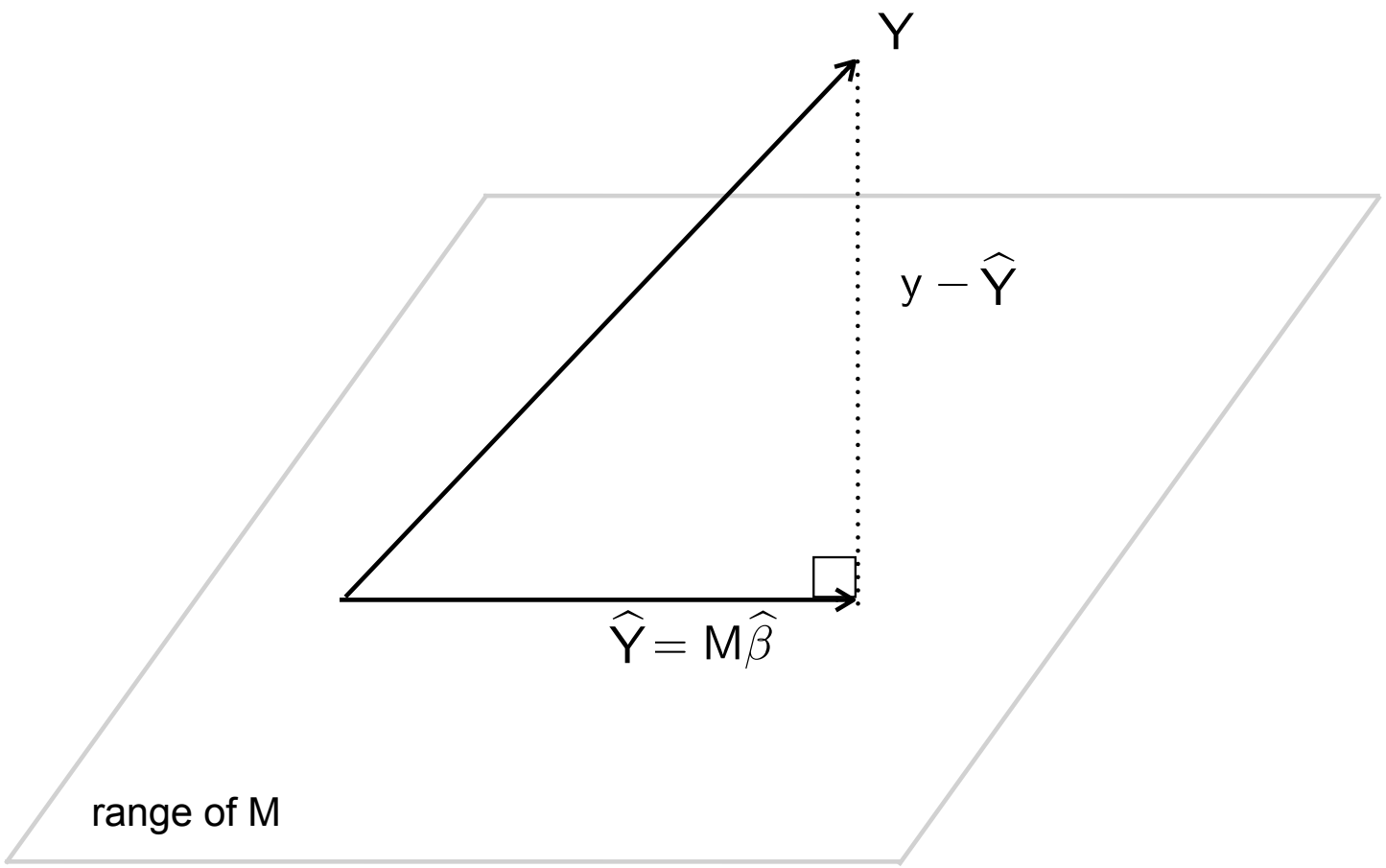
$$\begin{aligned}
\widehat{\mathsf{Y}}^{\mathsf{t}}\widehat{\epsilon} &= \widehat{\mathsf{Y}}^{\mathsf{t}}(\mathsf{Y} - \widehat{\mathsf{Y}}) \\
&= \mathsf{Y}^{\mathsf{t}}\mathsf{H}^{\mathsf{t}}(\mathsf{I} - \mathsf{H})\mathsf{Y} \\
&= \mathsf{Y}^{\mathsf{t}}(\mathsf{H} - \mathsf{H}\mathsf{H})\mathsf{Y}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

A geometric view

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

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

In pictures this becomes...


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

Sampling distributions

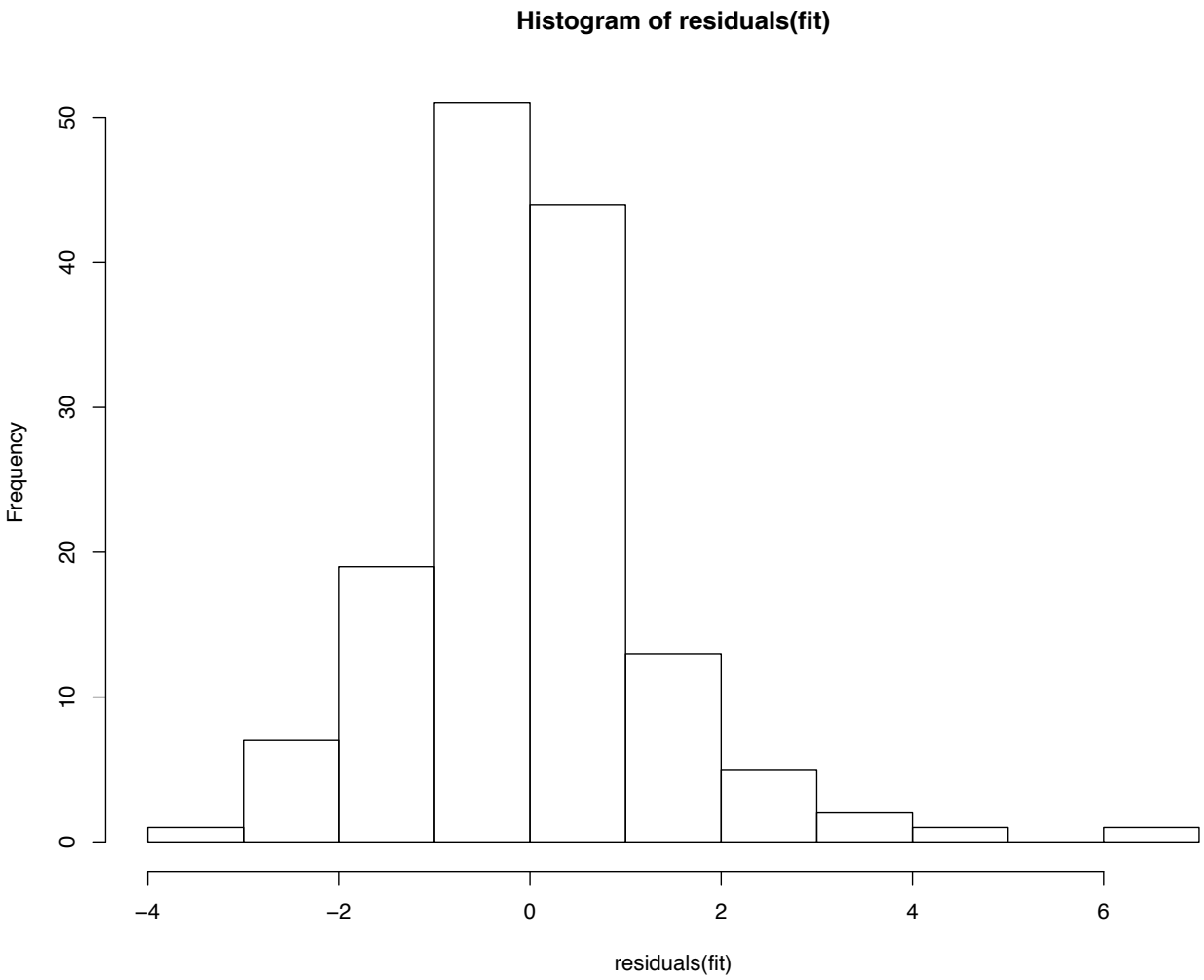
We can work out some things about the sampling distribution of our estimated regression coefficients -- If we assume our errors are independent, identically distributed and with a common standard deviation then we can write out their mean

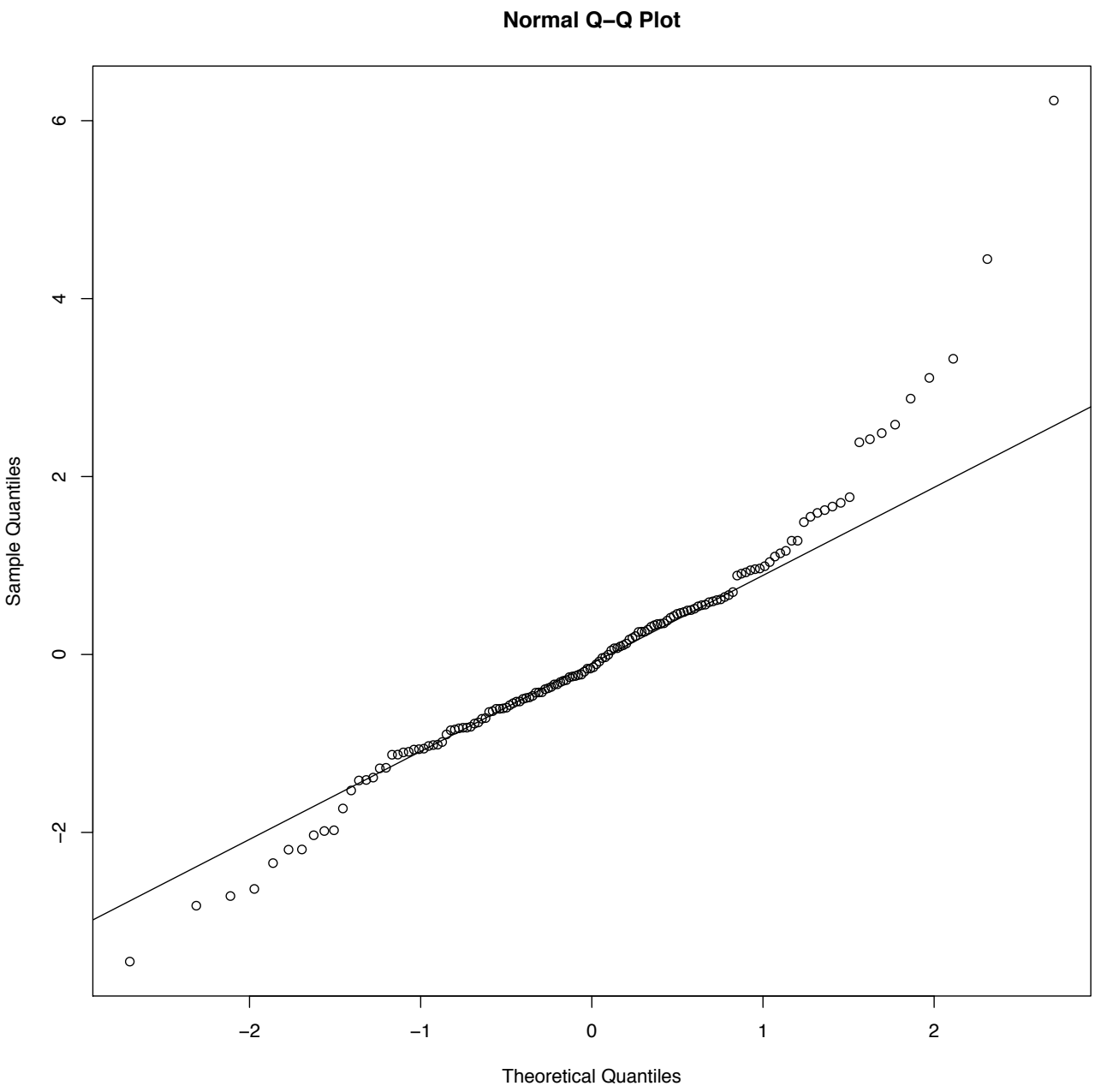
$$\begin{aligned} \mathrm{E}\widehat{\beta} &= (\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{EY} \\ &= (\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{M}\beta \\ &= \beta \end{aligned}$$

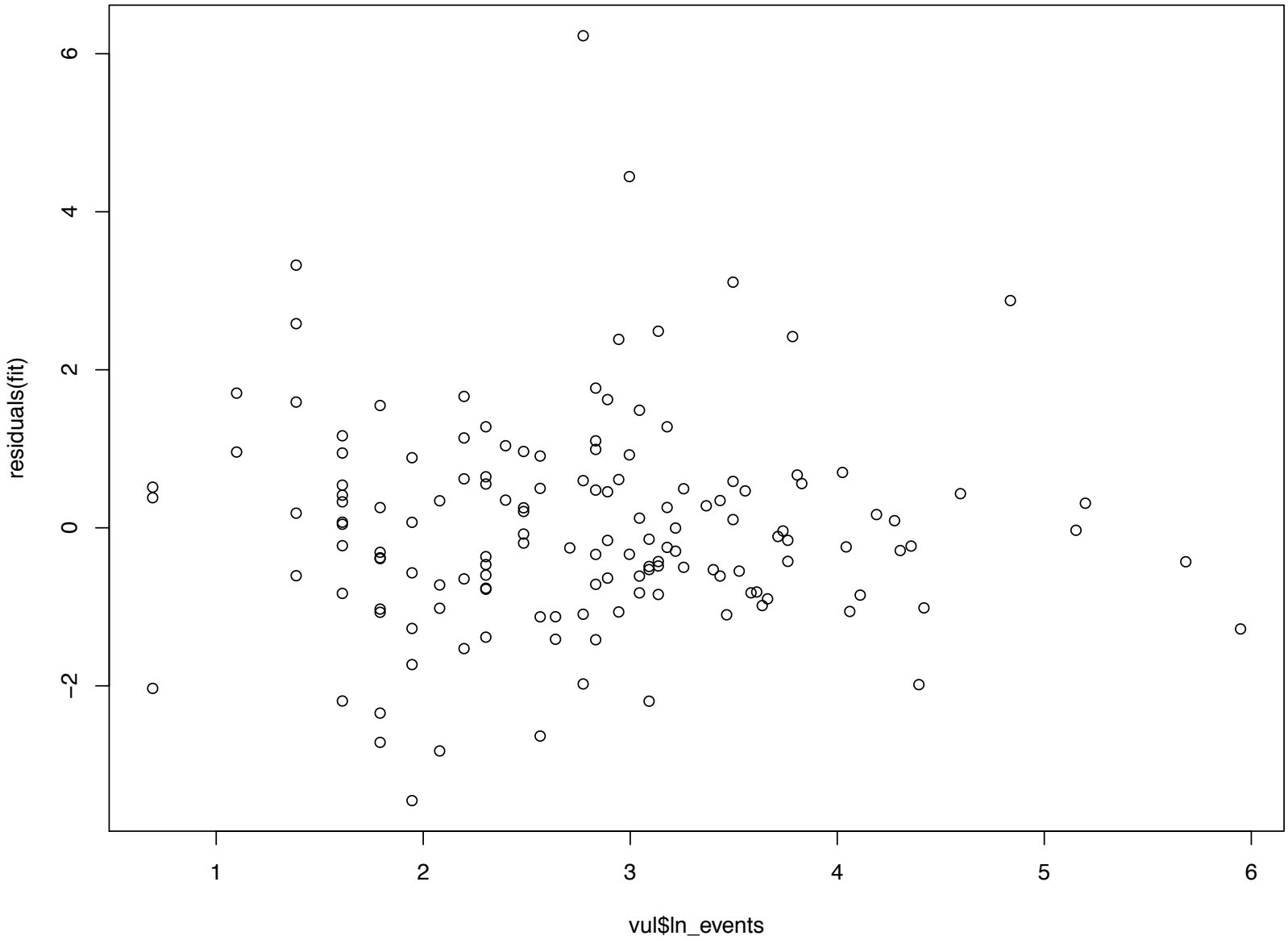
and variance-covariance matrix

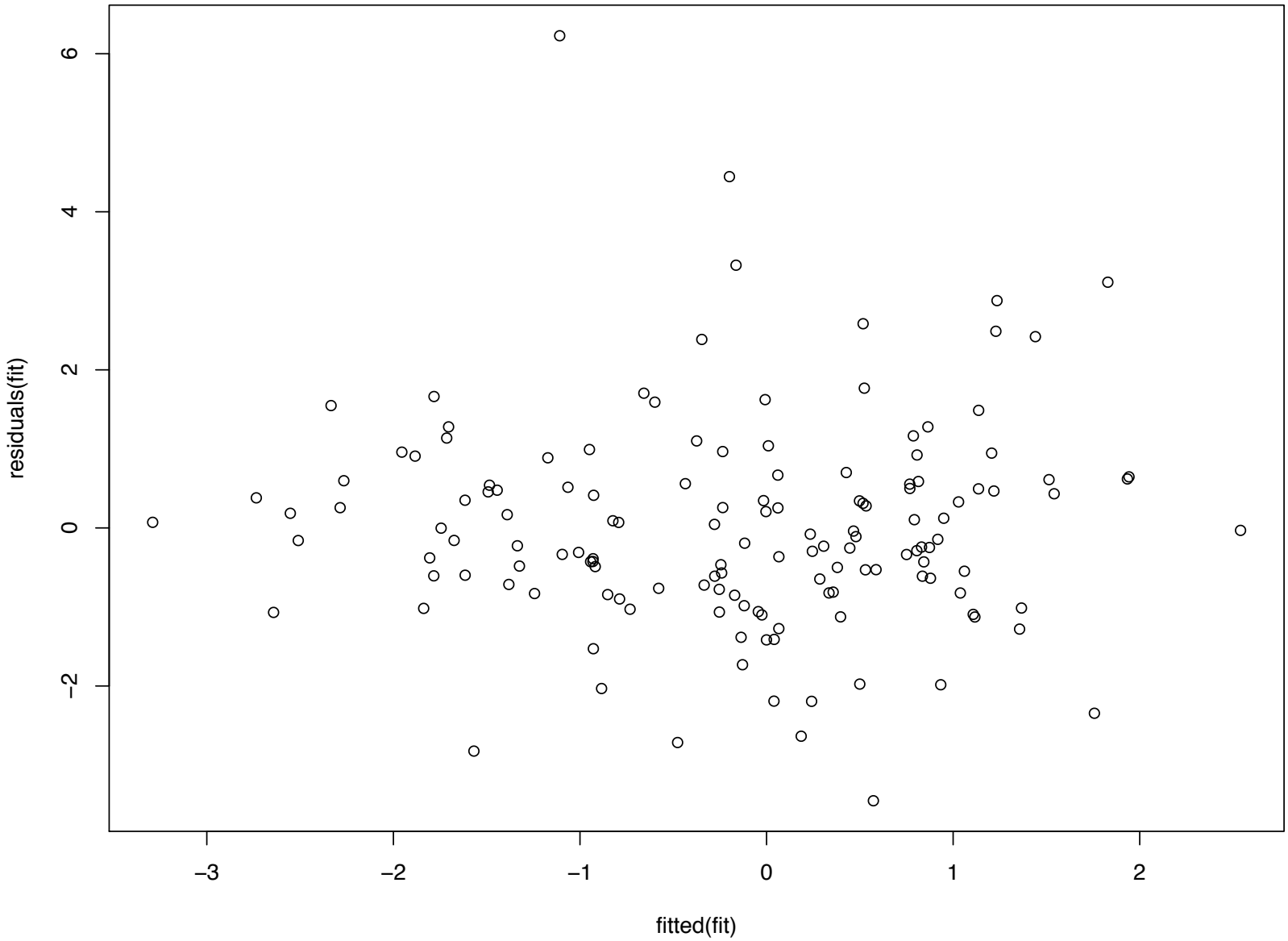
$$\begin{aligned} \mathrm{var}\ \widehat{\beta} &= (\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{var}\ \mathrm{Y}\ \mathrm{M}(\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1} \\ &= \sigma^2(\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}\mathrm{M}^{\mathrm{t}}\mathrm{M}(\mathrm{M}^{\mathrm{t}}\mathrm{M})^{-1}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

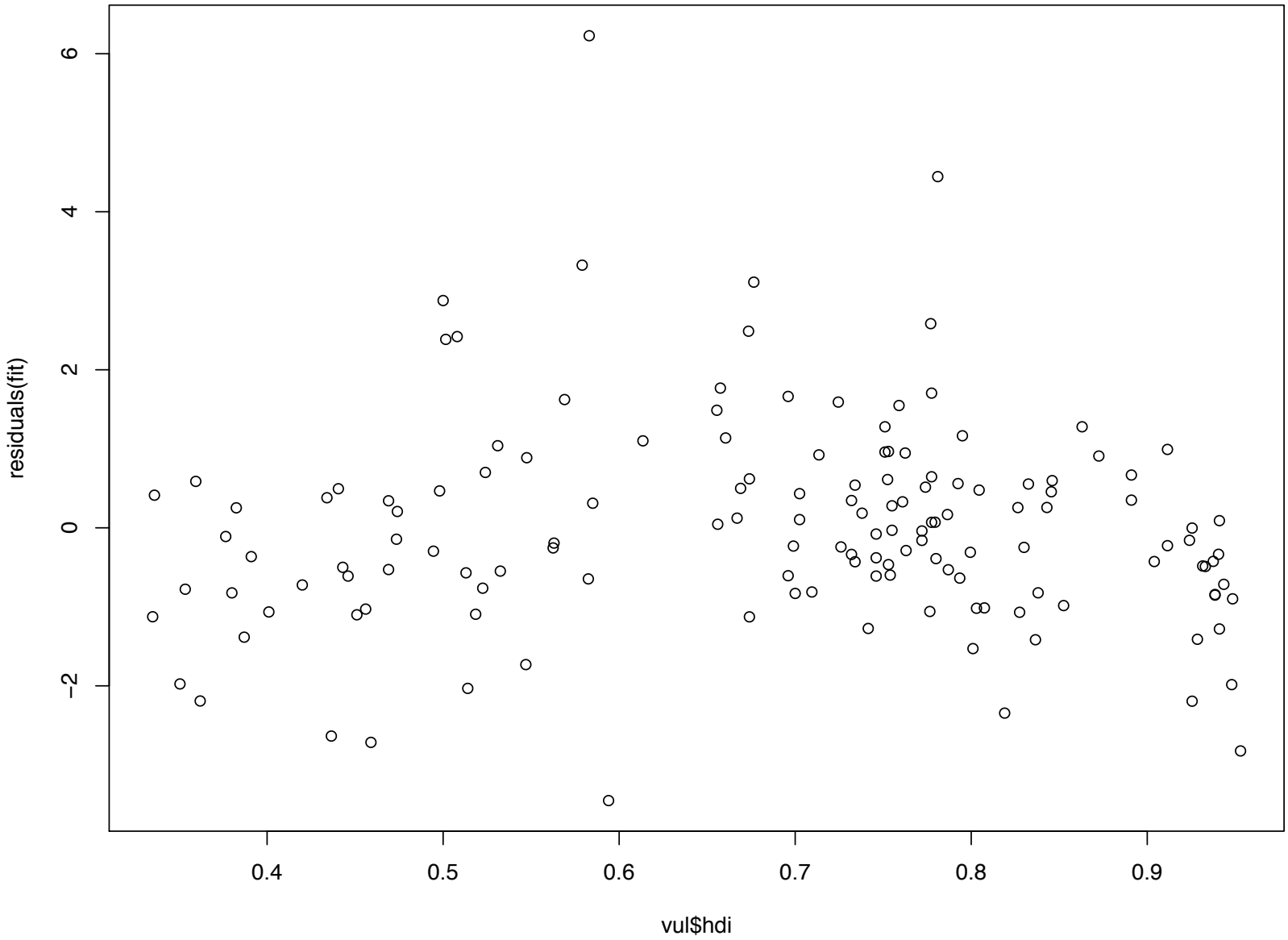
```
> model <- lm(ln_death_risk~ln_events+ln_fert+ln_pop+hdi,data=vul)

# the Y-hats or fitted values are obtained with, well fitted() and the
# residuals are computed with, well, residuals()

> sum(residuals(model)*fitted(model))
[1] 2.959438e-15

# zero! the residuals are orthogonal to the fit
# have a look at the residuals

> hist(residuals(model))

> qqnorm(residuals(model))
> qqline(residuals(model))

> plot(fitted(modelmodel),residuals(model))
> plot(vul$hdi,residuals(model))
> plot(vul$ln_events,residuals(model))

# add a “smooth” curve to see if something’s missing!
> plot(vul$hdi,residuals(model))
> lo <- loess(residuals(model)~vul$hdi)
> points(vul$hdi,predict(lo),pch=19,col=5)
```

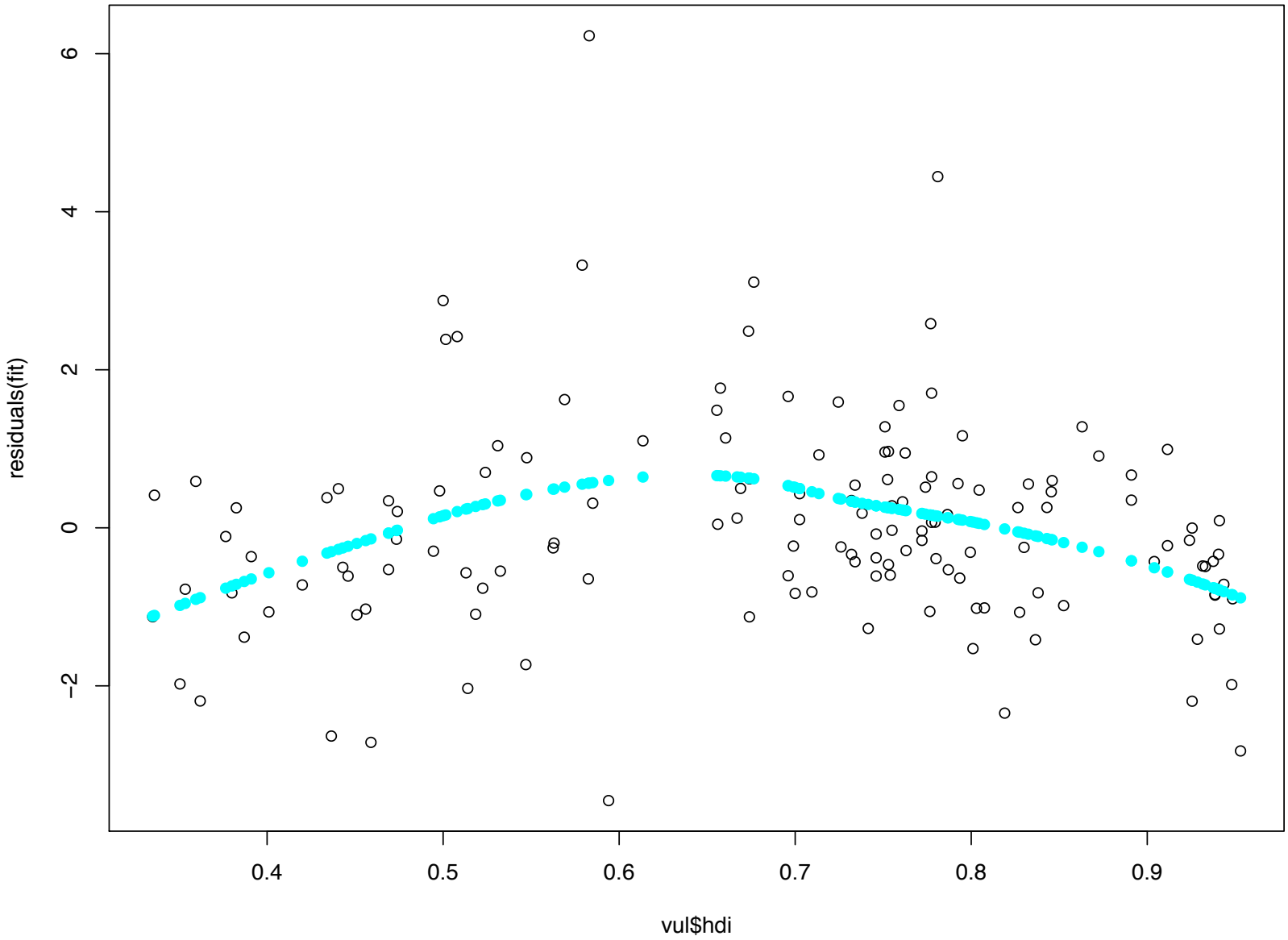

Histogram of residuals(fit)
Frequency
0
10
20
30
40
50
−4
−2
0
2
4
6
residuals(fit)


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


residuals(fit)
6
4
2
0
−2
1
2
3
4
5
6
vul$ln_events


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


residuals(fit)
6
4
2
0
−2
0.4
0.5
0.6
0.7
0.8
0.9
vul$hdi


residuals(fit)
6
4
2
0
−2
0.4
0.5
0.6
0.7
0.8
0.9
vul$hdi

Inference

To assess questions like what variables have a significant impact on the response, we are led to wonder what our data have to say more broadly -- What can we say in the face of the uncertainty that's present? Which coefficients are statistically significant (different from zero)?

To address this we first have to consider the random mechanism at play -- We already noticed that it doesn't come from sampling because we have essentially all the countries (or, rather, the researchers took those for which data were available)

Where is the randomness?

A model

In this case, we assume that the countries were not sampled in some way but that instead the errors were drawn from a fixed distribution -- For the normal linear model this would be, well, a Gaussian distribution, but in general we may not be willing to make that assumption

To apply the bootstrap here, we analyze as we randomized -- If the stochastic mechanism comes from the errors

$$Y_i = x_{i1}\beta_1^* + \cdots + \beta_p^* x_{ip} + \epsilon_i$$

then we need to estimate the error distribution!

The bootstrap

Here, we propose sampling not pairs (x_i, Y_i), but instead sample from the residuals from our fitted model

$$\widehat{\epsilon}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_1 x_{i1} - \cdots - \widehat{\beta}_{ip} x_{ip}$$

To be precise, we fit our regression model and create the residuals , then we do the following

For b = 1,..., B

Form a new error vector $\widetilde{\epsilon} = (\widetilde{\epsilon}_1, \ldots, \widetilde{\epsilon}_n)$ by sampling n times with replacement from the $\widehat{\epsilon}_1, \ldots, \widehat{\epsilon}_n$

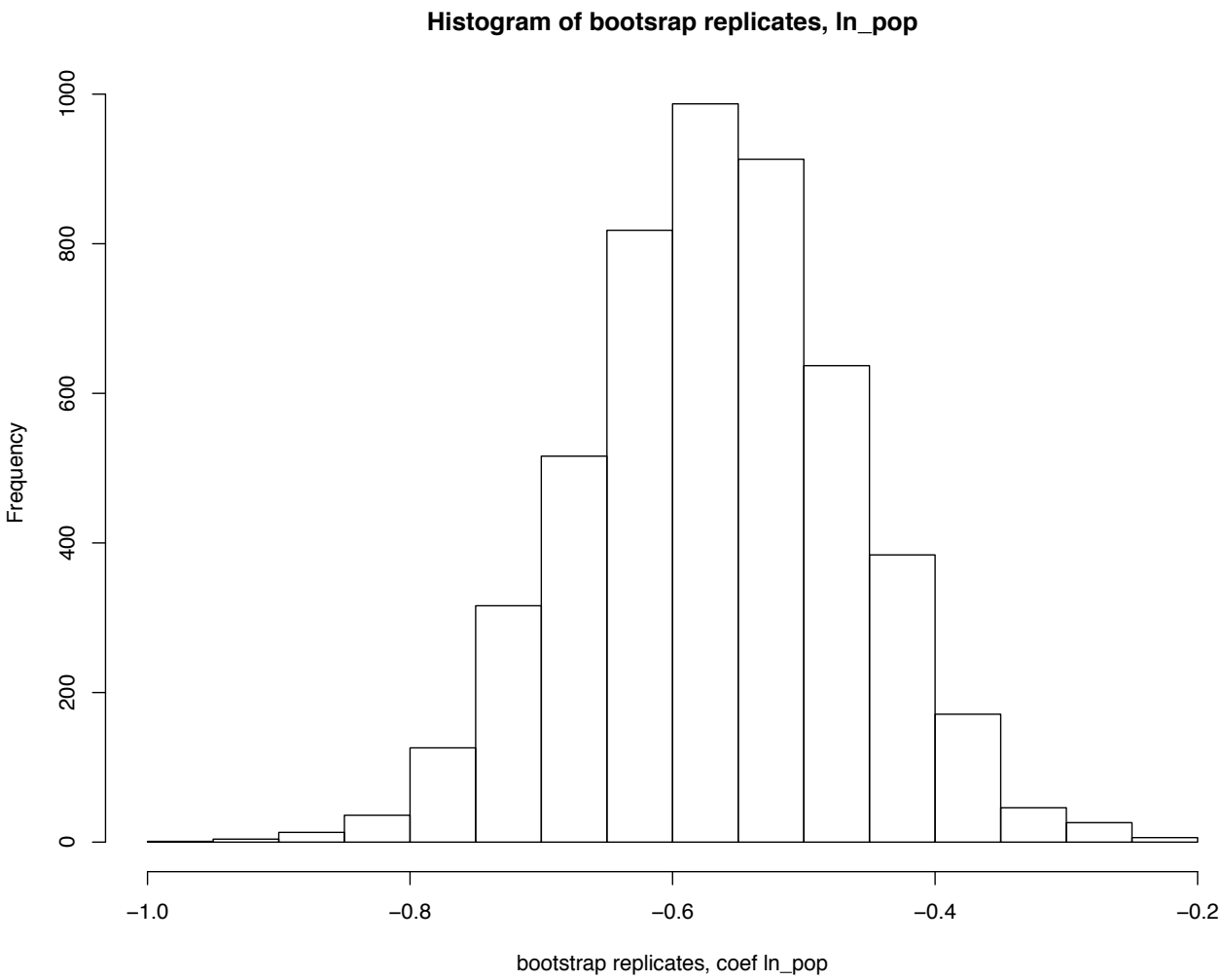
Then, create new observations

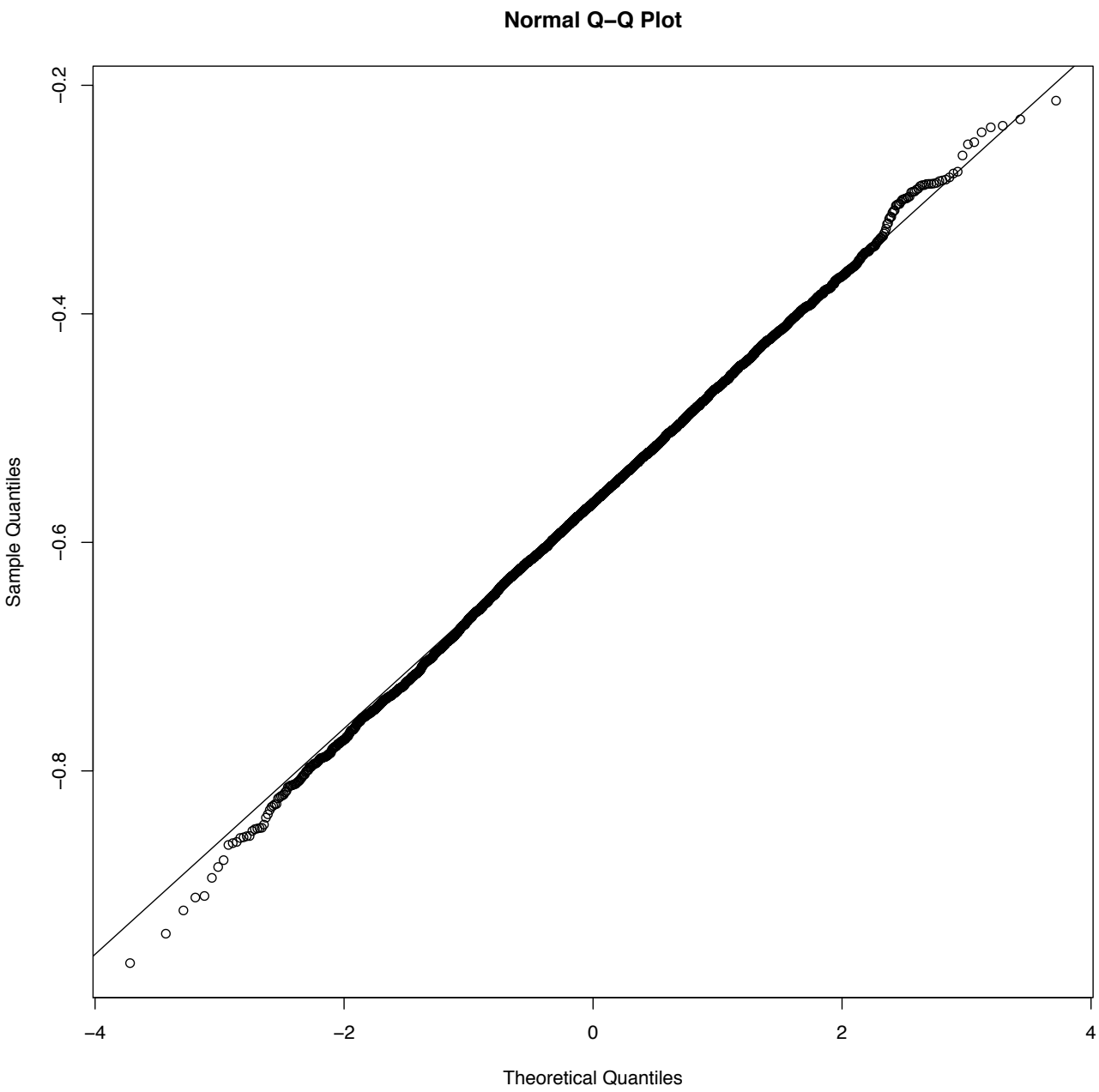
$$\widetilde{Y}_i = \widehat{Y}_i + \widetilde{\epsilon}_i = \widehat{\beta}_1 x_{i1} + \cdots + \widehat{\beta}_p x_{ip} + \widetilde{\epsilon}_i$$

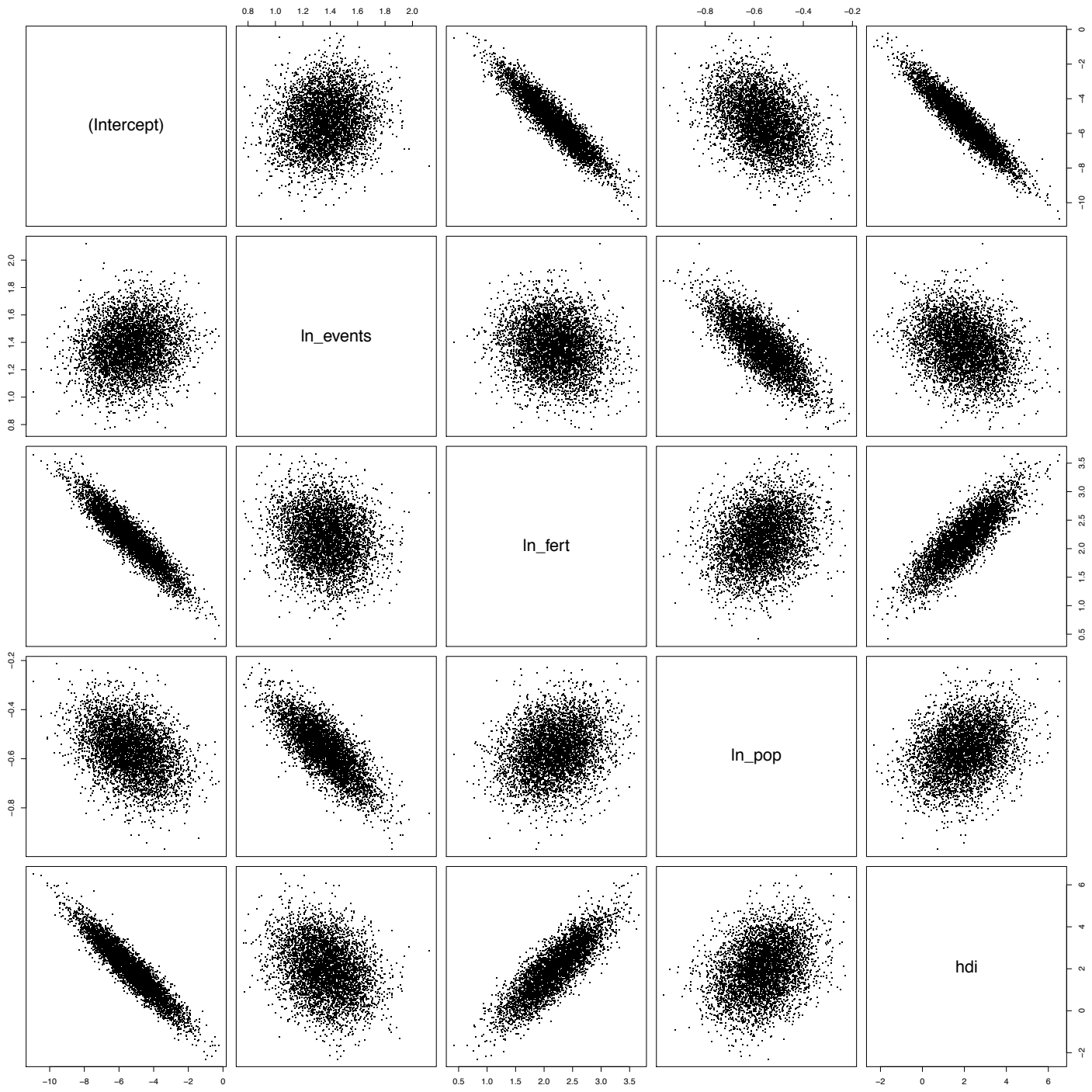
and fit a new model using OLS on the pairs $(x_i, \widetilde{Y}_i)$

The resulting coefficients $\widetilde{\beta} = (\widetilde{\beta}_1, \ldots, \widetilde{\beta}_p)$ is the bth bootstrap replicate

```
 model <- lm(ln_death_risk~ln_events+ln_fert+ln_pop+hdi,data=vul)

 res <- residuals(model)
 fit <- fitted(model)

 bootreps <- matrix(0,ncol=5,nrow=5000)

 for(i in 1:5000){

   # new data from sampled residuals
   vul$boot_y <- fit+sample(res,replace=T)

   #fit to the new data
   bootmodel <- lm(boot_y~ln_events+ln_fert+ln_pop+hdi,data=vul)

   # save bootstrap replicates of the coefficients
   bootreps[i,] <- coefficients(bootmodel)
}

# match these to the table!
> sd(bootreps[,1])
[1] 1.483672

> sd(bootreps[,2])
[1] 0.1780579

> sd(bootreps[,3])
[1] 0.4536994

> hist(bootreps[,4])
> qqnorm(bootreps[,4])
> pairs(bootreps)

# what can you say about the joint (sampling) distribution of the betas"
```


Histogram of bootsrap replicates, ln_pop
Frequency
0
200
400
600
800
1000
−1.0
−0.8
−0.6
−0.4
−0.2
bootstrap replicates, coef ln_pop


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
−0.2
−0.4
−0.6
−0.8
−4
−2
0
2
4


(Intercept)
ln_events
ln_fert
ln_pop
hdi

A more complicated statistic

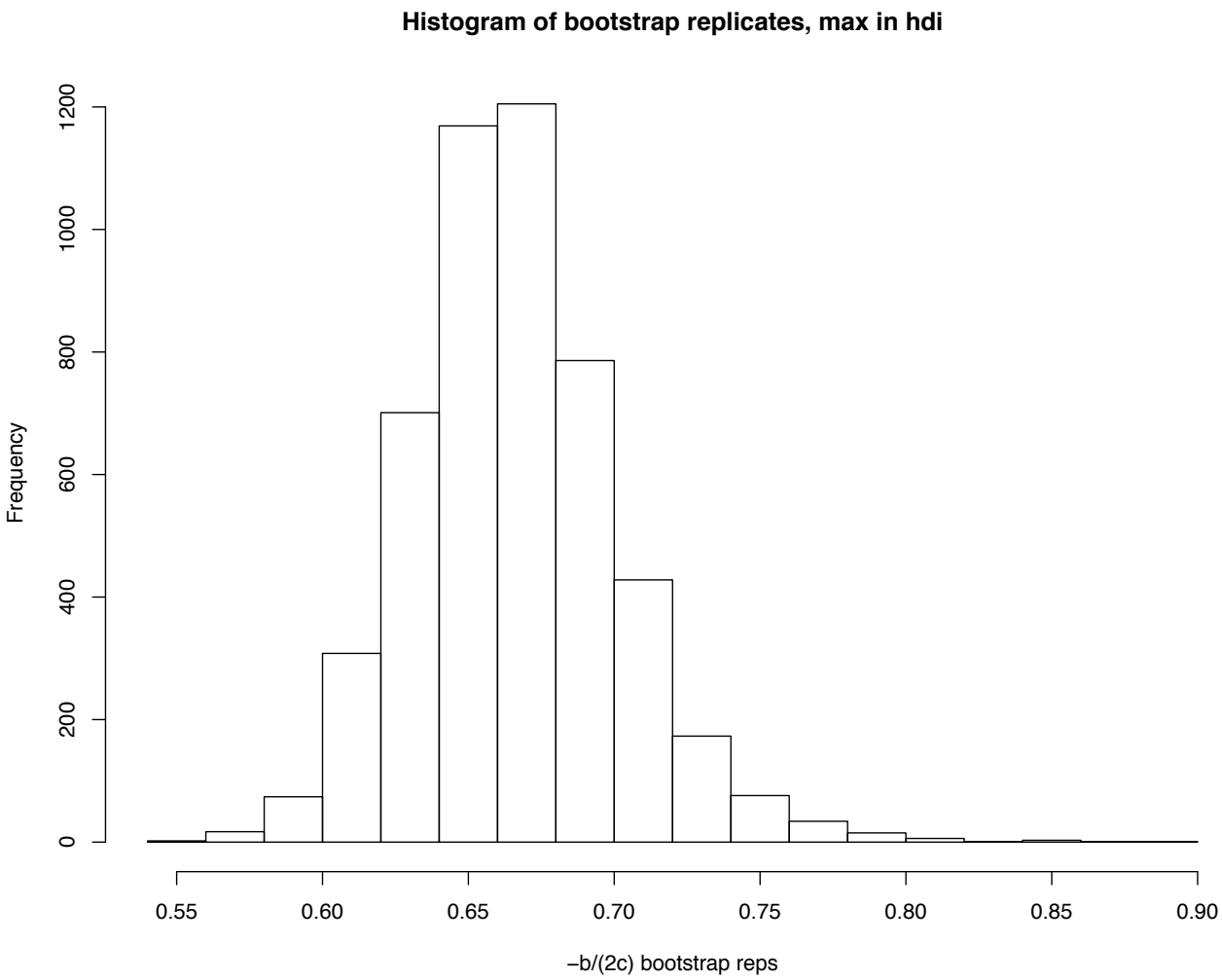
Suppose we were interested in understanding where the maximum was in HDI -- We can fit the author's original quadratic and estimate its maximum

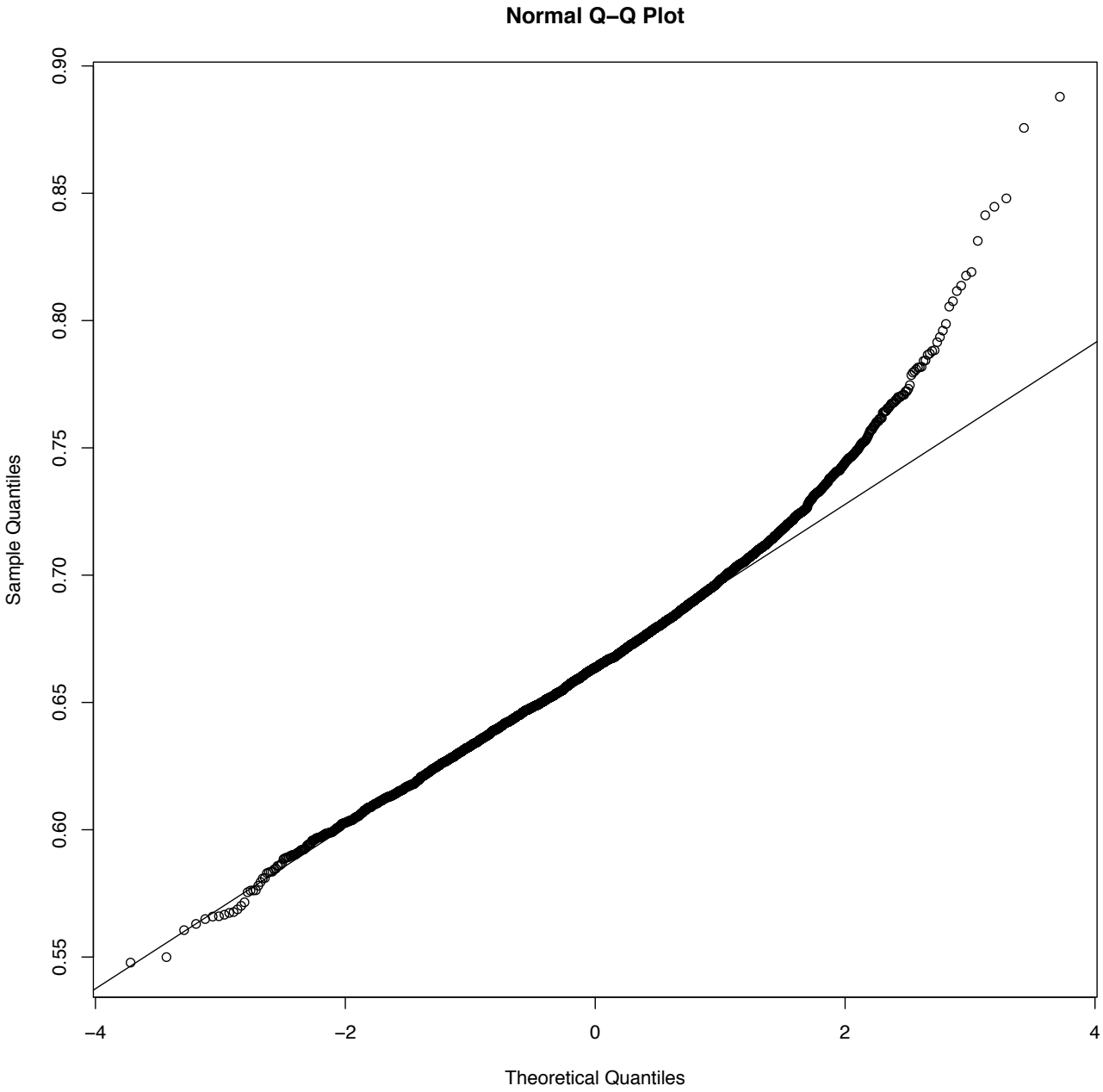
Recall that a quadratic of the form $a + b\,z + c\,z^2$ has a maximum (well, extremum which in this case is a maximum) at -b/(2c)

We can then use the bootstrap replicates to estimate the sampling distribution for the location of the maximum in HDI and construct a confidence interval! This is on the next page

(We probably should make sure that we are in fact estimating a maximum and that some of the curves aren't flipped around more like a smile than a frown -- Re-implement the code on the next page to do this!)

```
# now, let's consider the max in hdi -- fit with the quadratic
# model here...

model <- lm(ln_death_risk~ln_events+ln_fert+ln_pop+hdi+I(hdi^2),data=vul)

res <- residuals(model)
fit <- fitted(model)

bootreps <- rep(0,5000)

for(i in 1:5000){

   # new data from sampled residuals
   vul$boot_y <- fit+sample(res,replace=T)

   #fit to the new data
   bootmodel <- lm(boot_y~ln_events+ln_fert+ln_pop+hdi+I(hdi^2),data=vul)

   # save bootstrap replicates of ratio -b/(2c)
   bootcoef <- coefficients(bootmodel)
   bootreps[i] <- -bootcoef[5]/(2*bootcoef[6])
}

> hist(bootreps)

# you often see heavy tails with a ratio!
> qqnorm(bootreps)

> quantile(bootreps,c(0.025,0.975))
     2.5%     97.5%
0.6036005 0.7412837
```


Histogram of bootstrap replicates, max in hdi
Frequency
0
200
400
600
800
1000
1200
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
−b/(2c) bootstrap reps


Normal Q–Q Plot
Sample Quantiles
Theoretical Quantiles
0.55
0.60
0.65
0.70
0.75
0.80
0.85
0.90
−4
−2
0
2
4

Pivot again

If our error distribution is normal (or if we have a large sample size and the CLT kicks in), then we expect the distribution of our $\widehat{\beta}$ to be multivariate normal with mean β^* and variance-covariance matrix (given some slides back)

$$\sigma^2(\mathsf{M}^\mathsf{t}\mathsf{M})^{-1}$$

And so

$$\frac{\widehat{\beta}_\mathsf{j} - \beta^*_\mathsf{j}}{\mathsf{se}(\widehat{\beta}_\mathsf{j})} = \frac{\widehat{\beta}_\mathsf{j} - \beta^*_\mathsf{j}}{\sigma\sqrt{[(\mathsf{M}^\mathsf{t}\mathsf{M})^{-1}]_{\mathsf{jj}}}} = \frac{\widehat{\beta}_\mathsf{j} - \beta^*_\mathsf{j}}{\mathsf{c}\sigma}$$

has a standard normal distribution and we can use it to construct confidence intervals much in the same way we did when we discussed Gosset

Ah, but in the same way, we don't know the error variance and we have to estimate it, leading you to expect that (and you can prove this rigorously)

$$\frac{\widehat{\beta}_\mathsf{j} - \beta^*_\mathsf{j}}{\mathsf{c}\,\widehat{\sigma}}$$

has a t-distribution, here with n-p degrees of freedom where $\widehat{\sigma}^2 = \mathsf{RSS}/(\mathsf{n}-\mathsf{p})$

Confidence intervals

This classical result (again, assuming the errors are either exactly normal or that we have enough data to make an application of the CLT realistic) lets us move easily between hypothesis testing and confidence intervals

Again, if we think of the pivot notion, we know that

$$\frac{\widehat{\beta}_j - \beta_j^*}{c\,\widehat{\sigma}}$$

has a t-distribution so that

$$P(-t \leq \frac{\widehat{\beta}_j - \beta_j^*}{c\,\widehat{\sigma}} \leq t) = 0.95$$

which we inverted to give us confidence intervals

Hypothesis testing

For hypothesis testing, we instead start a hypothesized value, in this case, it might be that $\beta_j^* = 0$ -- It would be interesting to be able to reject this hypothesis and declare that an input variable does have an effect on the response

Under the null $\beta_j^* = 0$, then the ratio

$$\frac{\widehat{\beta}_j}{c\,\widehat{\sigma}}$$

has a t-distribution again with n-p degrees of freedom -- We can then use the t-distribution as our null sampling distribution just like we did when we were doing rerandomizations for A/B testing

Hypothesis testing

The P-value quoted by R in the table a few slides back for each coefficient asks how likely is it under the null to see a value of $\widehat{\beta}_j$ as large as we have -- Here the notion of large is measured in the absolute sense

That is, a priori we often don't have strong feelings whether the coefficient should be positive or negative and so we compare the chance of seeing

$$\frac{|\widehat{\beta}_j|}{c\,\widehat{\sigma}}$$

as large as we have

(In the Neyman-Pearson framework, this is equivalent to specifying a two-sided alternative, namely that $\beta_j^* \neq 0$)

Bootstrap hypothesis testing

To mimic this situation using our standard analysis tools, we recall that in the Bootstrap World, our estimate, in this case $\widehat{\beta}_j$, plays the role of β_j^* -- Therefore, we can create a bootstrap sampling distribution for

$$\frac{\widehat{\beta}_j - \beta_j^*}{c\,\widehat{\sigma}}$$

from a collection of bootstrap replicates of

$$\frac{\widetilde{\beta}_j - \widehat{\beta}_j^*}{c\,\widetilde{\sigma}}$$

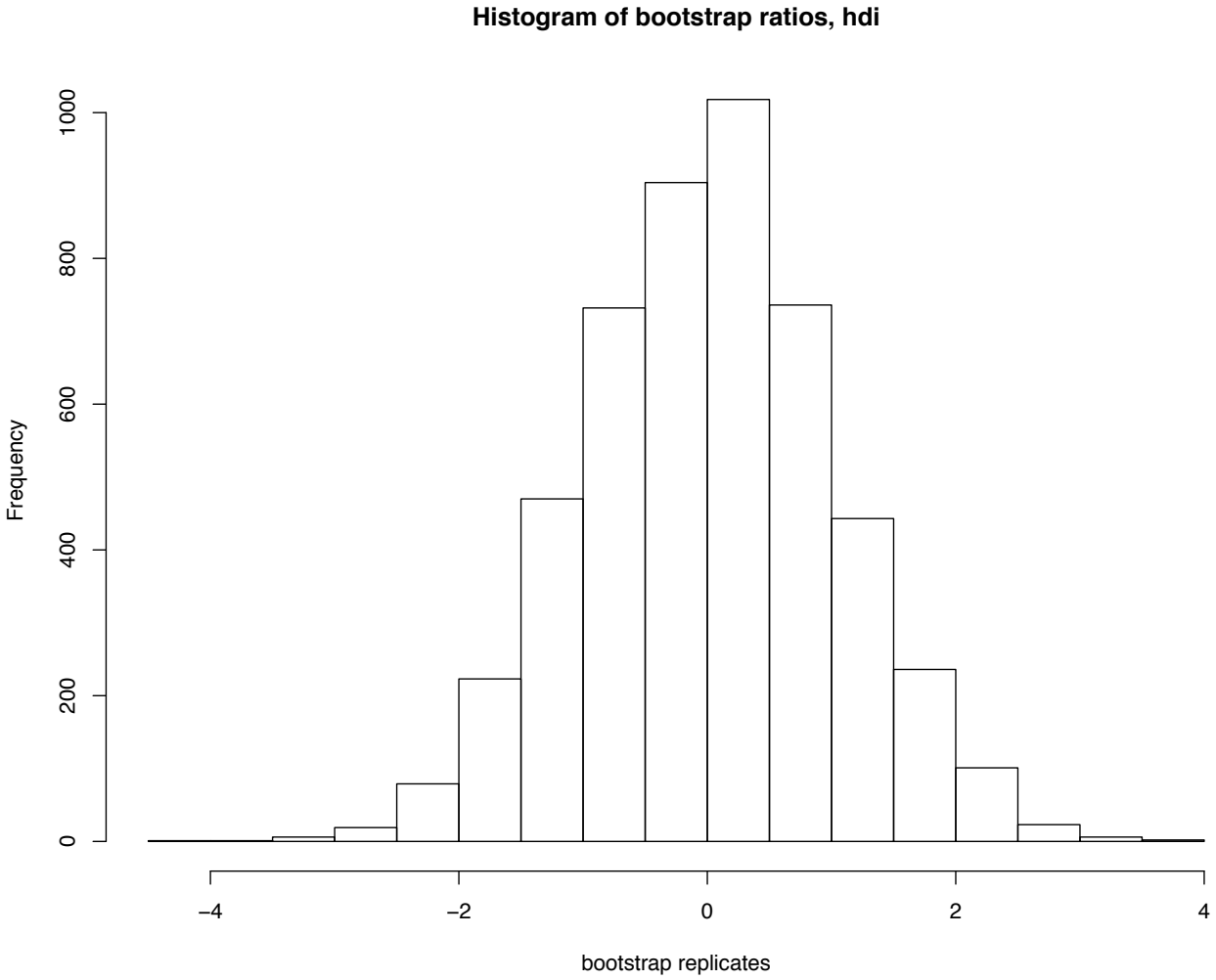
Bootstrap hypothesis testing

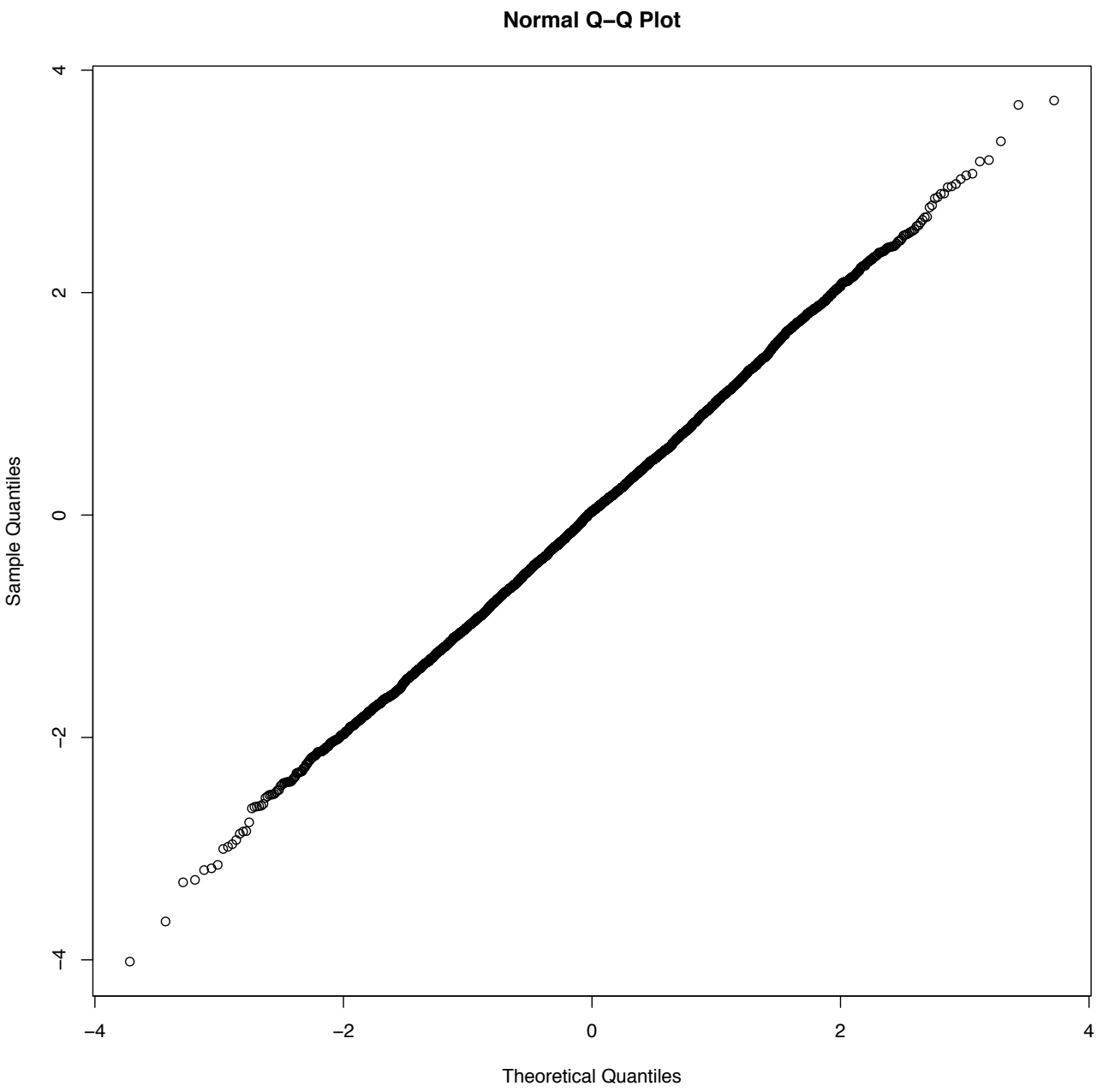
Under the null hypothesis that $\beta_j^* = 0$, we can then compare our observed

$$\frac{|\widehat{\beta}_j|}{c\,\widehat{\sigma}}$$

to the distribution of bootstrap replicates and compute a P-value!

```
# now, let's consider the max in hdi -- fit with the quadratic
# model here...

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

 res <- residuals(model)
 fit <- fitted(model)

 bootreps <- matrix(0,ncol=5,nrow=5000)

 for(i in 1:5000){

   # new data from sampled residuals
   vul$boot_y <- fit+sample(res,replace=T)

   #fit to the new data
   bootmodel <- lm(boot_y~ln_events+ln_fert+ln_pop+hdi,data=vul)

   # save bootstrap replicates of the ratios
   bootsum <- summary(bootmodel)$coefficients
   bootreps[i,] <- (bootsum[,1]-coefficients(model))/bootsum[,2]
}

> hist(bootreps[,5])

# the observed ratio is given in the table for the regression on slide 48
# estimate the p-value
> mean(abs(bootreps[,5])> 1.578)
[1] 0.1236
```


Histogram of bootstrap ratios, hdi
Frequency
0
200
400
600
800
1000
−4
−2
0
2
4
bootstrap replicates


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