STA465: *Theory and Methods for Complex Spatial Data*

Instructor: Dr. Vianey Leos Barajas

COURSE ADMIN

HELLO AND WELCOME!

- Welcome to STA465: Theory and Methods for Complex Spatial Data
- In this course we are going to extend the methods you learnt in STA302 to cover data where observations are not independent.
- We will specifically focus on dependence structures that are specified by the location at which the measurement occurs.
- We will learn the theory and practice of spatial data analysis through a number of case studies.
- **There will be R. (and INLA and Stan)**

WHO AM I?

- Dr. Vianey Leos Barajas (she/her)
- First name: Vianey
- Last names: Leos Barajas (born in México!)

- PhD in Statistics from Iowa State University

- Office Hours: Friday 1:30-2:30 pm

  Other times **by appointment only**

- Email: vianey.leosbarajas@utoronto.ca

TEACHING ASSISTANT + OTHER HELPERS

- Dayi (David) Li
- PhD student in statistics — research focus on topics in astrostatistics and has worked on spatial point process modelling of stellar objects in the M33 galaxy
- Will answer questions posted on discussion section in Quercus
- Office hours — any preference?
- Additional assistants:

EMAIL POLICY

- Email is for questions that aren't appropriate for the discussion forum on Quercus or office hours
- I'll answer as soon as feasible. I typically do not check emails on nights or weekends, so allow for 2 **working days after** you send the email before you start getting annoyed.
- I will occasionally miss something so if I don't answer, please re-send and (politely!) remind me to answer.
- **If you don't include your name somewhere in the message I will not know who you are!**

ASSESSMENT

- 5 Homeworks worth 20% each
  - Data analysis. Some modelling or explanation component. R will be required.
- 1 **optional** Final Exam worth 20%.
- Due dates are in the syllabus.

ASSESSMENT

- Lateness policy:
  - Homeworks are due **sharply** at the appointed time. No late assignments will be accepted without documentation of a valid reason. Remember, you can take the final if you can't complete a homework assignment for any reason!

- Re-grading policy:
  - Regrading requests should only be made for **genuine grading errors**, and should be initiated by writing or typing a complete explanation of your concern (together with your full name, student number, and e-mail address) on a **separate piece of paper**, and giving this together with your original **unaltered** homework/test paper to the instructor within one week of when the graded item was first available. **Warning:** your mark may end up going down rather than up.

ACADEMIC HONESTY

- Don’t Cheat.
- Don’t pay someone else to do your homework for you.
- The assignments are designed to test your practical skills with data. They should require your to synthesize your new skills.

THE FINAL EXAM

- **Optional.**
- If you have turned in all of your homework assignments and are happy with your grade, you do not need to take the final exam.
- If, for any reason, you were not able to complete one of the homework assignments OR you are unhappy with your lowest homework score, you have the option of taking the final exam to makeup for a missing homework or low score.

- I will drop the lowest score of the homeworks and final exam to compute the final grade.

TEXTBOOK AND SLIDES

- This course does not have a single fixed text.
- Four books will be used occasionally (available electronically through the library or for free online):
  - *Spatio-temporal statistics with R* by Christopher K. Wikle, Andrew Zasmmit-Mangion and Noel Cressie
  - *Geospatial Health Data: Modeling and Visualization with R-INLA and Shiny* by Paula Moraga
  - *Spatial and spatio-temporal Bayesian models with R-INLA* by Marta Blangiardo and Michela Cameletti
  - *Statistical Analysis and Modelling of Spatial Point Patterns* by Janine Illian, Antti Penttinen, Helga Stoyan, and Dietrich Stoyan (maybe!)
- Further information will be contained in slides, handouts, and specific references that will be available on Quercus before classes.

COMPUTING

- The course will be run using the **R** computing environment.
- You are **strongly encouraged** to use **RStudio** (https://www.rstudio.com), which is a **free** IDE for **R.**
- All instructions in the course will assume that you have the **latest version** of **both** RStudio and R installed. We **will not answer** any R related questions unless both of these things are true.

- The best resource for R help is always **google**.
- This course will use the R package **INLA**. This is not available from CRAN but can be installed into R using the command

```
install.packages("INLA", repos=c(getOption("repos"), INLA="https://inla.r-inla-download.org/R/testing"), dep=TRUE)
```

THE CONTENT OF THE COURSE

- Linear regression as a Bayesian model
- Multivariate Gaussian distributions and conditional independence
- Bayesian multilevel models
- Models for areal data
- Model checking, validation, and workflow
- Gaussian random fields in theory and practice
- Modelling non-Gaussian spatial data
- Simulation

GLOBAL AIR POLLUTION

AIR POLLUTION IS BAD

- The World Health Organization estimates that 7 million deaths each year may be directly attributable to air pollution

- Poor air quality in major cities has been a problem for hundreds of years

- A wide range of pollutants have been implicated in adverse effects on human health, but particular attention has tended to focus on particulate matter

MORE SPECIFICALLY, AIRBORNE PARTICULAR MATTER IS BAD

- Complex mixture of extremely small particles and liquid droplets.

- “Inhalable coarse particles” such as those found near roadways and dusty industries (PM10)

- “Fine particles” such as those found in smoke and haze (PM2.5) can be directly emitted from sources such as forest fires, or they can form when gases emitted from power plants, industries and automobiles react in the air.

THIS IS A GLOBAL PUBLIC HEALTH PROBLEM

- Many international institutions, including the World Health Organisation and Global Burden of Disease (GBD) collaboration (Institute for Health Metrics and Evaluation) require population exposures to air pollution in order to estimate the associated burden of disease.

- Accurate estimates of PM2.5 concentrations are required on a global scale.

- These estimates need to be linked with population data to estimate population-level exposures.

- We need an accurate, global, multi-resolution PM2.5 data product that acknowledges the inherent uncertainty in measurements.

SUSTAINABLE DEVELOPMENT GOALS

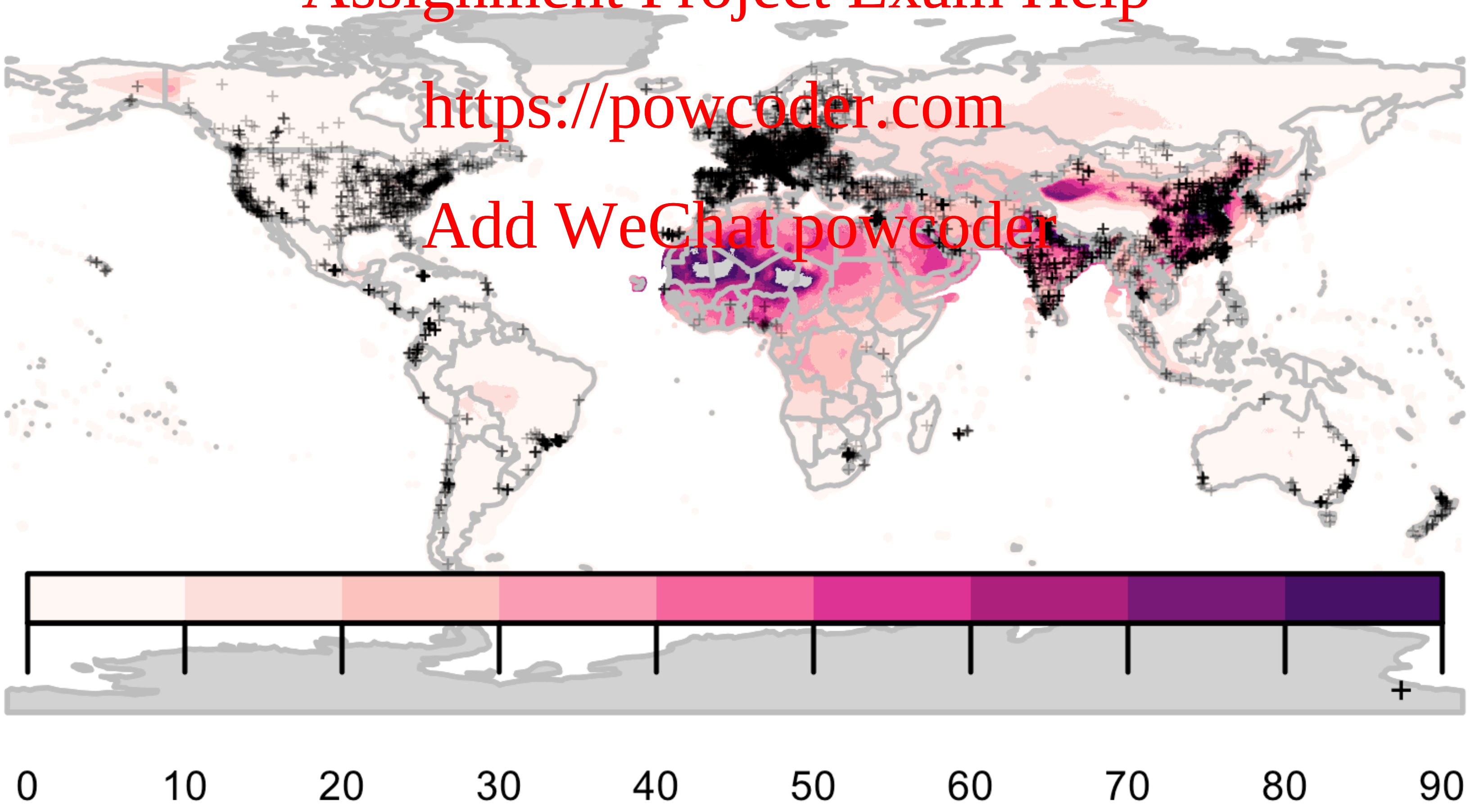
- The Sustainable Development Goals are a set of 17 goals proposed by the United Nations in 2015.
- Air pollution is a part of SDGs 3 (Health), 7 (Energy), and 11 (Cities)

- There are also two relevant SDG indicators:
  - 11.6.2: Annual mean levels of fine particulate matter (PM2.5) (population-weighted)
  - 3.9.1: Mortality rate attributed to household and ambient air pollution.

BREATHE

Goal Estimate global PM2.5 concentration

Problem Most data from noisy satellite measurements (6003 ground monitors provide sparse, heterogeneous coverage)



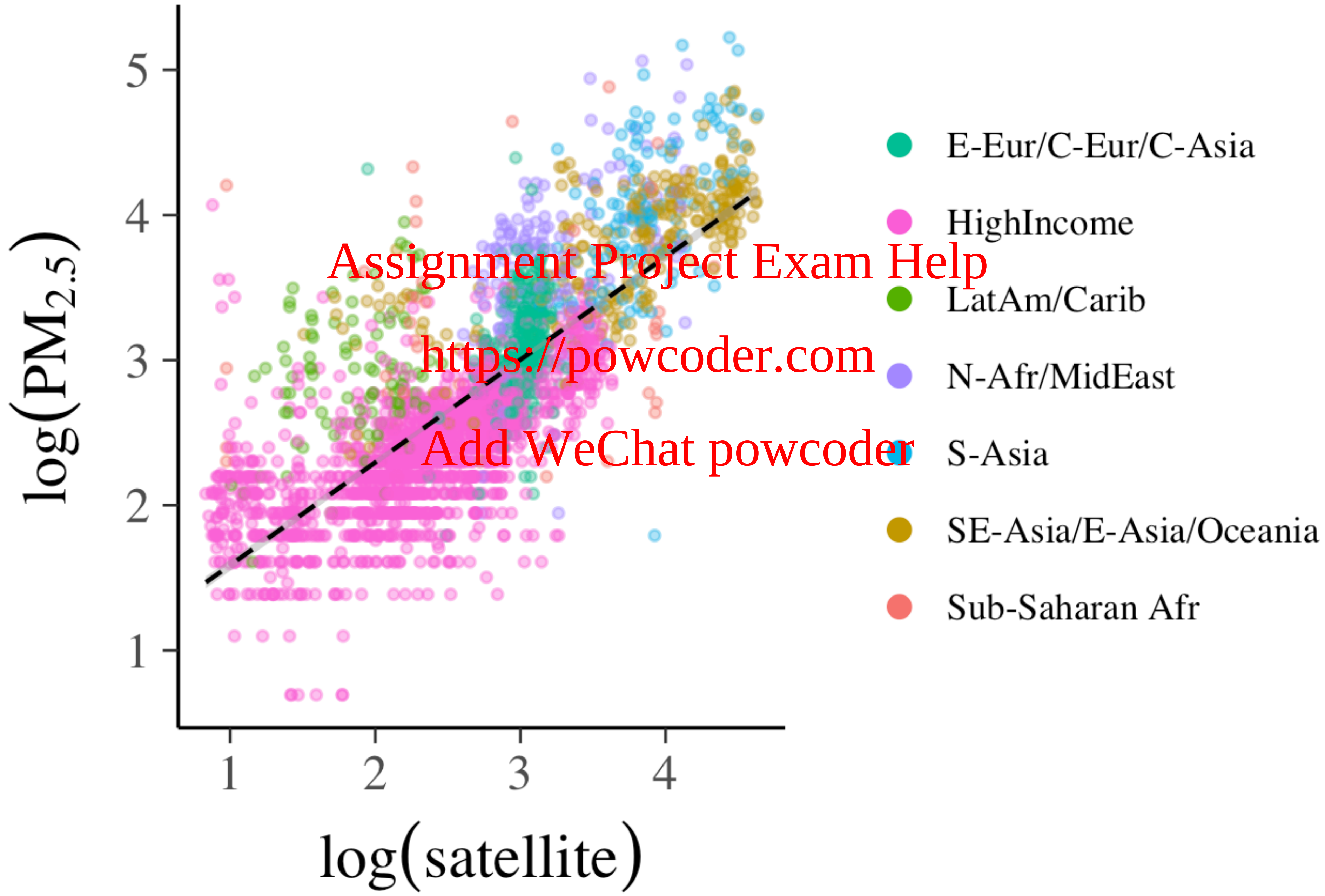
black points indicate ground monitor locations

Satellite estimates of PM2.5 and ground monitor locations

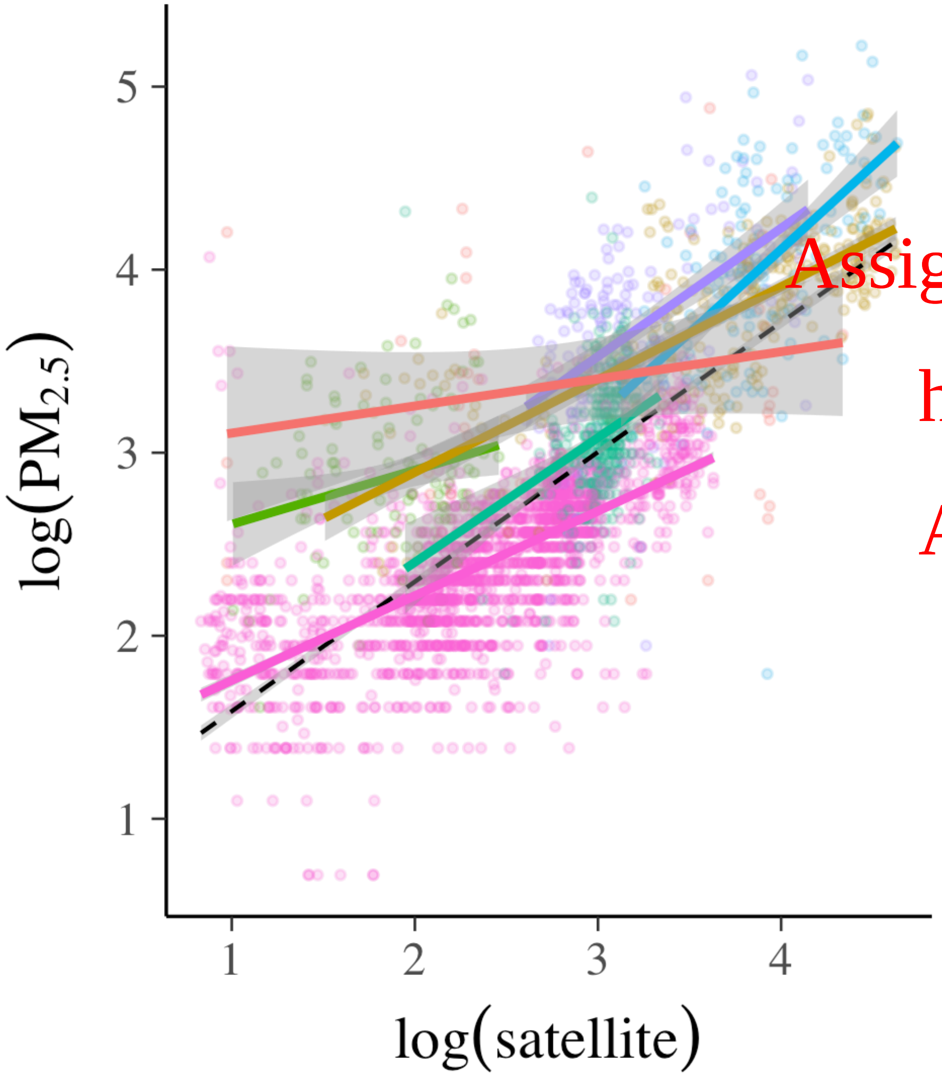
WHAT SORT OF INDIRECT INFORMATION DO WE HAVE?

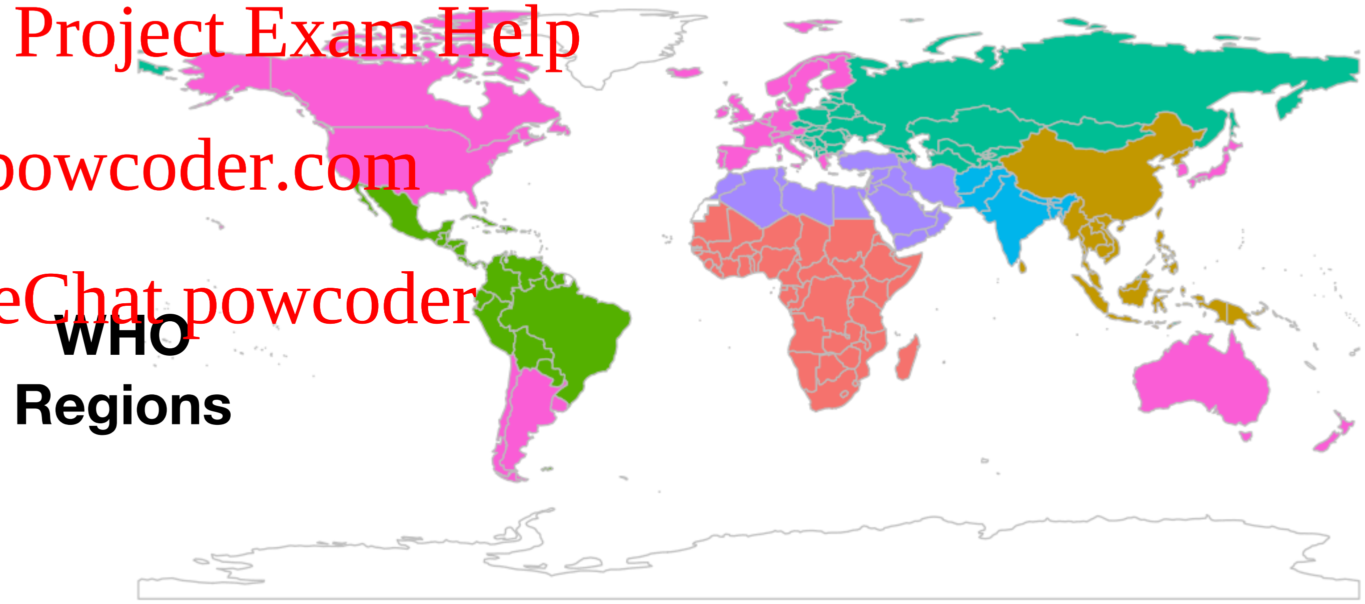
- Gridded satellite measurements of Aerosol Optical Depth converted to PM2.5 estimates
- Gridded general circulation / chemical transport models (TM5-FASST, GEOS-Chem)
- Gridded estimates of the sum of sulphate, nitrate, ammonium and organic carbon (SNAOC) and the compositional concentrations of mineral dust (DUST) based on simulations from the GEOS-Chem chemical transport model
- Gridded estimates of the global population from the Gridded Population of the World project

HOW CAN WE USE THIS INDIRECT INFORMATION?



BUT IS ONE STRAIGHT LINE ENOUGH?





WHAT WE'RE GOING TO DO

- Re-interpret linear regression in a way that makes it more amenable to the type of extensions we need.
- Find methods for **pooling information** between different countries so that data from similar countries can be used to improve estimation in countries with little data.
- Find methods for accounting for the geographic closeness of countries
- Look at ways to do sub-national estimation.

REINTERPRETING LINEAR REGRESSION

LINEAR REGRESSION

- Today we're talking about linear regression. The fine art of putting a straight line through data.

- There is an **entire** course (STA302) dedicated to this topic, which we are going to use as a launchpad into spatial modelling.

- Our aim today is to re-interpret linear regression in a new framework that is easier to extend to non-iid data

THE SETUP

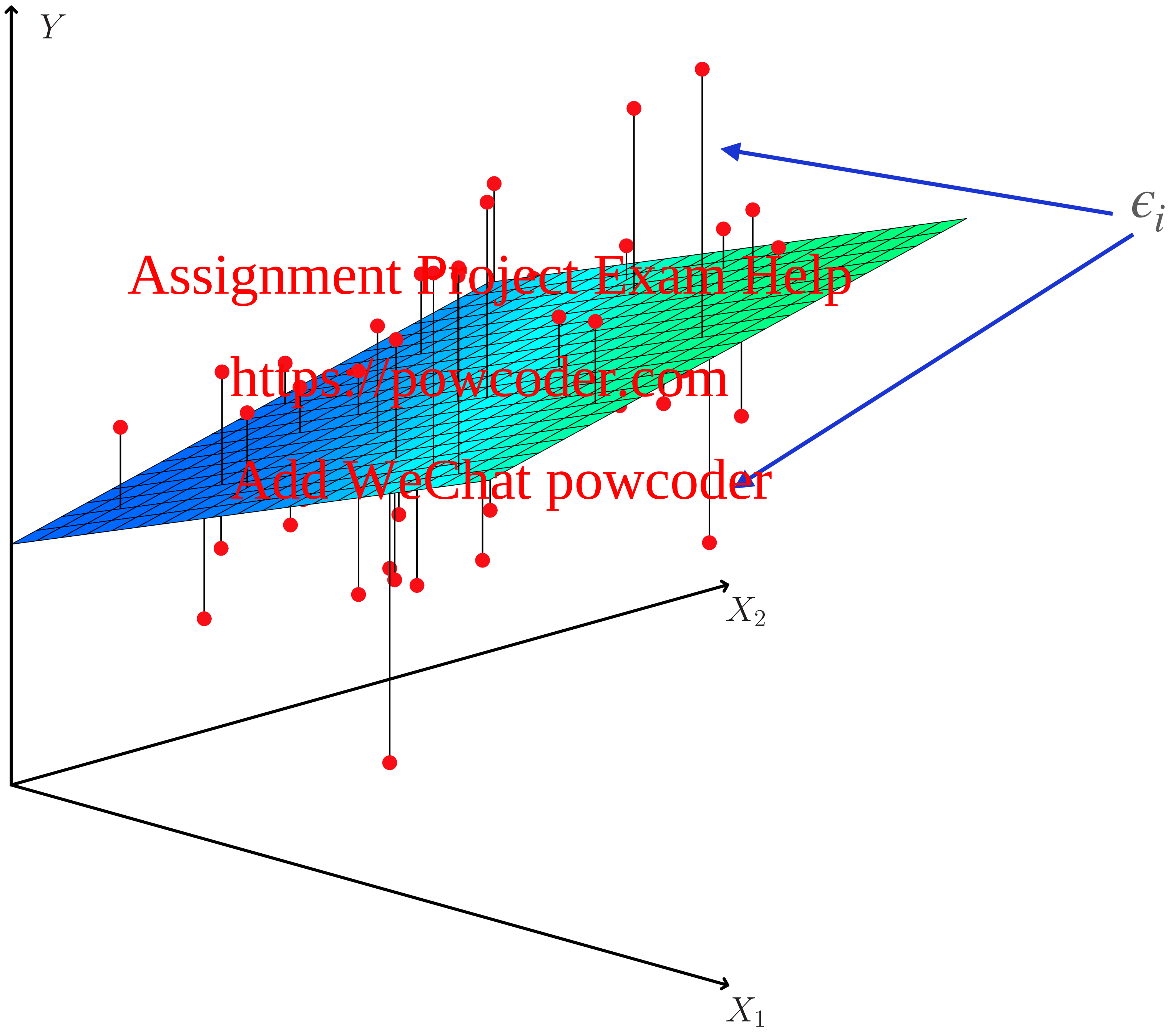
- Our linear regression model is as follows: We have independent data points y_i which come with a **vector** of features x_i

- We **model** the relationship between the data and the features as a linear regression model

$$y_i = x_i^T \beta + \epsilon_i$$

- Our main assumptions are that the noise (or residual) ϵ_i is **independent, identically distributed, and Gaussian** (The Gaussian bit can be relaxed somewhat, the others are critical)

REGRESSION IN A PICTURE



Some figures taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

FREQUENTIST INFERENCE

LINEAR REGRESSION MINIMIZES THE SUM OF SQUARES

- The empirical risk is

$$R_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2 = \frac{1}{n} \parallel y - X\beta \parallel^2$$

- Here X is a matrix and the rows of X are the feature vectors.

- Similarly y is the data as an n-dimensional vector.

MINIMIZING THE EMPIRICAL RISK

- We can get an estimate for β by minimizing the empirical risk.
- Taking derivates we get

$$\begin{aligned}\nabla R_n(\beta) &= \frac{1}{n}\nabla(y^Ty - 2y^TX\beta + \beta^TX^TX\beta) \\ &= \frac{1}{n}(-2X^Ty + 2X^TX\beta)\end{aligned}$$

- If we set the to zero (to find the minimum), we find that β solves

$$X^TX\beta = X^Ty$$

- These are commonly called the **normal equations** and their solution is called the **least squares** estimate.

SO WHAT ARE WE DOING?

- This type of classical inference (sometimes known as **frequentist** inference) has an underlying probabilistic framework:
  - The data y is random.
  - The estimator $\hat{\beta}(y)$ is a **deterministic** function of the data
  - We can then make probability statements about how often the **true value is within some interval around the estimator.**

INTERPRETING LEAST SQUARES

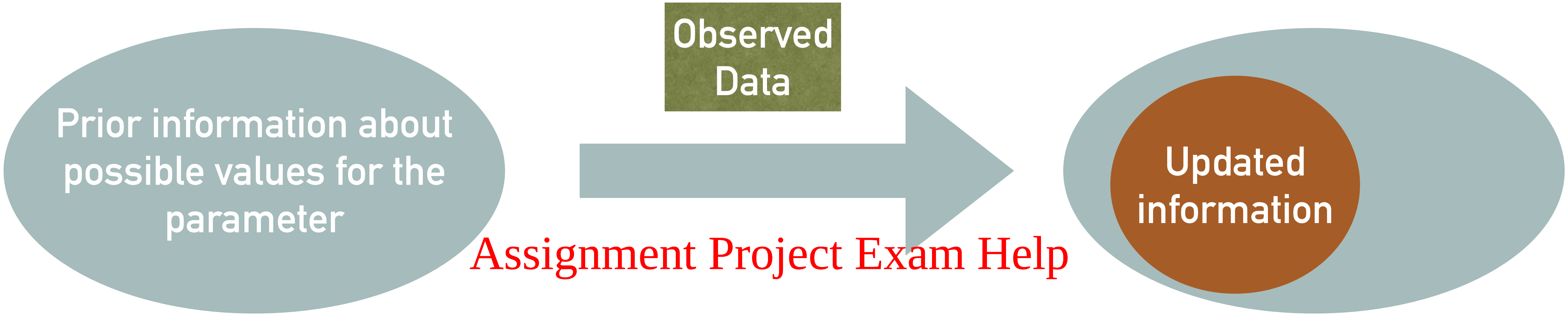
- This means that we actually construct a **box** of values for β and say that with some high probability the **true value** is in that box.
- So we are always making probabilistic statements about the true value of the regression line and how uncertain we are as a function of data

BUT THERE'S ANOTHER WAY

A DIFFERENT QUESTION

- What if, instead of asking questions about the true value that we would know perfectly if we had infinite data, we asked a slightly different question?
- ***Which values of β are consistent with the data we have observed?***
- This is a different question, but in a lot of cases it can give similar answers
- It does not rely on getting infinite amounts of data, but instead focuses on the data in hand.

HOW DO WE DO THIS?



Prior information about possible values for the parameter

Observed Data

Updated information

$$\Pr(\beta)$$

Prior

$$\Pr(y \mid \beta)$$

Likelihood

$$\Pr(\beta \mid y) = \frac{\Pr(y, \beta)}{\Pr(y)}$$

Posterior

THE LIKELIHOOD

- For linear regression, we model our observations as being the underlying regression surface + iid Gaussian noise, ie

$$y_i \mid \beta, \sigma \sim N(x_i^T \beta, \sigma^2)$$

- Because these are independent, the joint probability density is just the product

$$p(y \mid \beta, \sigma) = \prod_{i=1}^{n} p(y_i \mid \beta, \sigma)$$

$$= (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right]$$

UP TO A CONSTANT

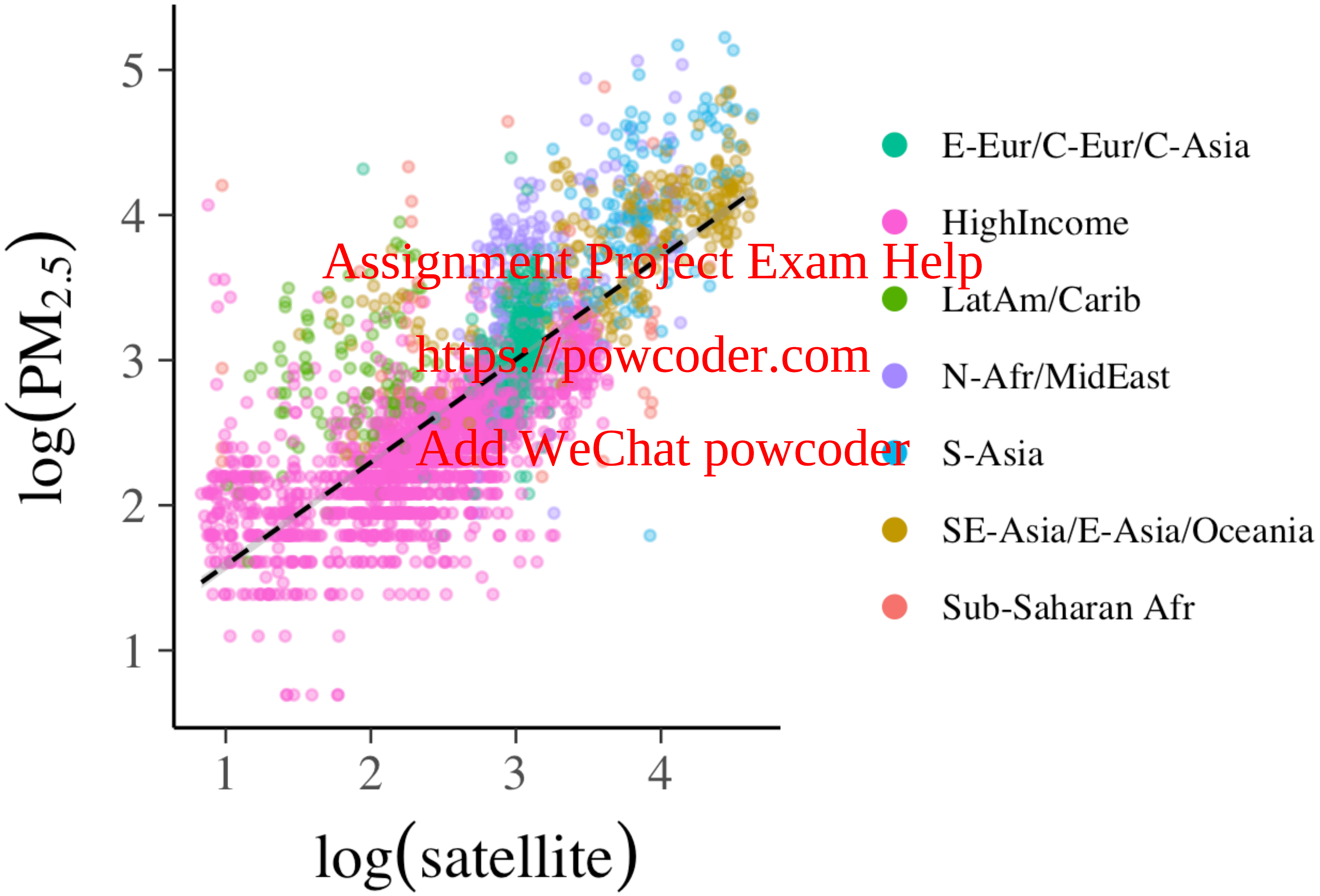
- Because we know that all probability densities integrate to one, we actually don't need to explicitly state the constant term, so we will usually write the likelihood as

$$p(y \mid \beta, \sigma) \propto \sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2\right]$$

WHAT ABOUT THE PRIOR?

- We need to define our set of reasonable states for β
- In some sense, this is hard if we don't know **anything** about the problem.
- But in reality, we do know some things!!

WHAT DO WE KNOW ABOUT PM2.5?

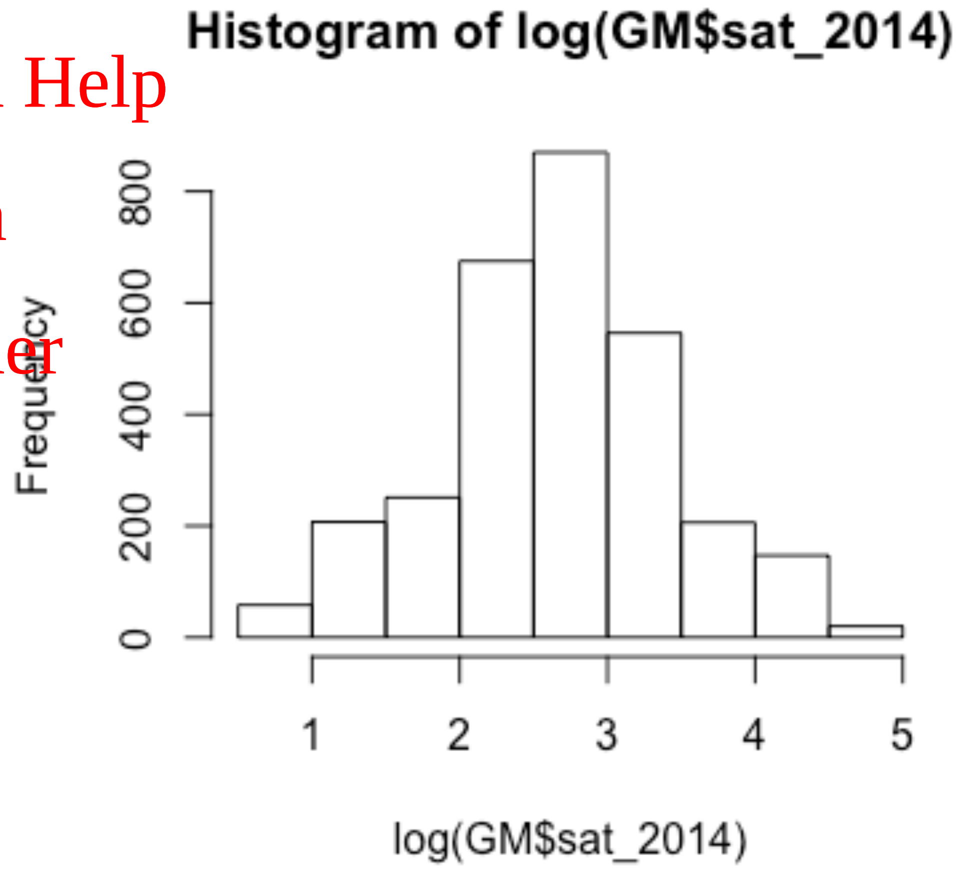


A PRIOR FOR PM2.5

- So we know that, on the log-scale, PM2.5 is probably not too small, and also probably not super-large.
- It would be surprising for it to be outside the range of [-10,10].
- (This corresponds to an upper limit of $\sim$2200 $\mu g m^{-3}$)
- For contrast, anything bigger than 28 on the log-scale is denser than concrete.
- **Should we make these hard limits?**

HOW DOES THAT HELP US MAKE A PRIOR?

- It doesn’t directly. It instead helps us tell if a prior is sensible.
- For instance, we know the log-satellite is in [0,5]
  so beta shouldn’t be much larger than 2
- Then the **data generated from the prior model** fits our constraints.



- One way to do this is to make

$$\beta \sim \text{Uniform}(0,2)$$

- Another option is

$$\beta \sim N(0,1) \text{ or } \beta \sim N(0,4)$$

- **Which is better?**

AN AWKWARDNESS

- There is a degree of arbitrariness to the specification of prior distributions
- And it can make a difference to inference!
- There are two things we must do to guard against this arbitrariness:
  - Make the prior justifiable *a priori*
  - Do model checking *a posteriori*
- **We will talk a lot about this later on**

THE POSTERIOR

COMPUTING THE POSTERIOR

- For simplicity, we’re going to assume that

$$\beta \sim N(0, \Sigma_\beta)$$

- This is a **multivariate Gaussian distribution** with pdf

$$p(\beta) \propto |\Sigma_\beta|^{-n/2} \exp\left(-\frac{1}{2}\beta^T \Sigma_\beta \beta\right)$$

- Most of the time, the **covariance matrix** Σ_β will be **diagonal** with the same value on the diagonal.
- This corresponds to each $\beta_j \sim N(0, \sigma_\beta^2)$

BAYES THEOREM

- We now have a prior and a likelihood, so we can compute a posterior using Bayes Theorem for densities

$$p(\beta|y,\sigma) = \frac{p(y|\beta,\sigma)p(\beta)}{p(y)}$$
$$\propto p(y|\beta,\sigma)p(\beta)$$

- The denominator on the first line is inconvenient to compute in a lot of cases, so we typically use the second expression.

BAYES FOR LINEAR REGRESSION

- Manipulating the densities and dropping all the terms that don't depend on β we get

$$
\begin{aligned}
p(\beta \mid y, \sigma) &\propto p(y|\beta, \sigma)p(\beta) \\
&\propto \exp\left(-\frac{1}{2\sigma}(y - X\beta)^T(y - X\beta) - \frac{1}{2}\beta^T\Sigma_p^{-1}\beta\right) \\
&= \exp\left[-\frac{1}{2}\left(\sigma^{-1}y^Ty - \frac{2}{\sigma}y^TX\beta + \sigma^{-1}\beta^TX^TX\beta + \beta^T\Sigma_\beta^{-1}\beta\right)\right] \\
&\propto \exp\left[-\frac{1}{2}\beta^T(\sigma^{-1}X^TX + \Sigma_\beta^{-1})\beta + \sigma^{-1}y^TX\beta\right]
\end{aligned}
$$

- YIKES!!!!

COMPARE IT TO A NORMAL DISTRIBUTION

- For a multivariate normal distribution, the density is

$$N(\beta;\mu,\Sigma) \propto |\Sigma|^{-n/2}\exp\left[\frac{-1/2}{(}\beta-\mu)^T\Sigma^{-1}(\beta-\mu)\right]$$

$$\propto |\Sigma|^{-n/2}\exp\left[\frac{-1}{2}\beta^T\Sigma^{-1}\beta+\mu^T\Sigma^{-1}\beta-\frac{1}{2}\mu^T\Sigma^{-1}\mu\right]$$

$$\propto |\Sigma|^{-n/2}\exp\left[\frac{-1}{2}\beta^T\Sigma^{-1}\beta+\mu^T\Sigma^{-1}\beta\right]$$

- Hence

$$\beta \mid y,\sigma \sim N\left[(X^TX)^{-1}X^Ty,(\sigma^{-1}X^TX+\Sigma_\beta^{-1})^{-1}\right]$$

BUT WHAT ABOUT SIGMA?