WEEK 1: SEARCHING FOR SIGMA

STA465: Theory and Methods for Complex Spatial Data

Instructor: Dr. Vianey Leos Barajas

WHAT ABOUT SIGMA?

A REMEMBRANCE OF THINGS PAST

- Consider the slightly simplified model

$$y_i \mid \beta, \sigma \sim N(X\beta, \sigma^2 I)$$
$$\beta \sim N(0, \tau^2 I)$$

- The posterior is

$$\beta \mid y, \sigma \sim N\left[\left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1} X^T y, \left(X^T X + \frac{\sigma^2}{\tau^2} I\right)^{-1}\right]$$

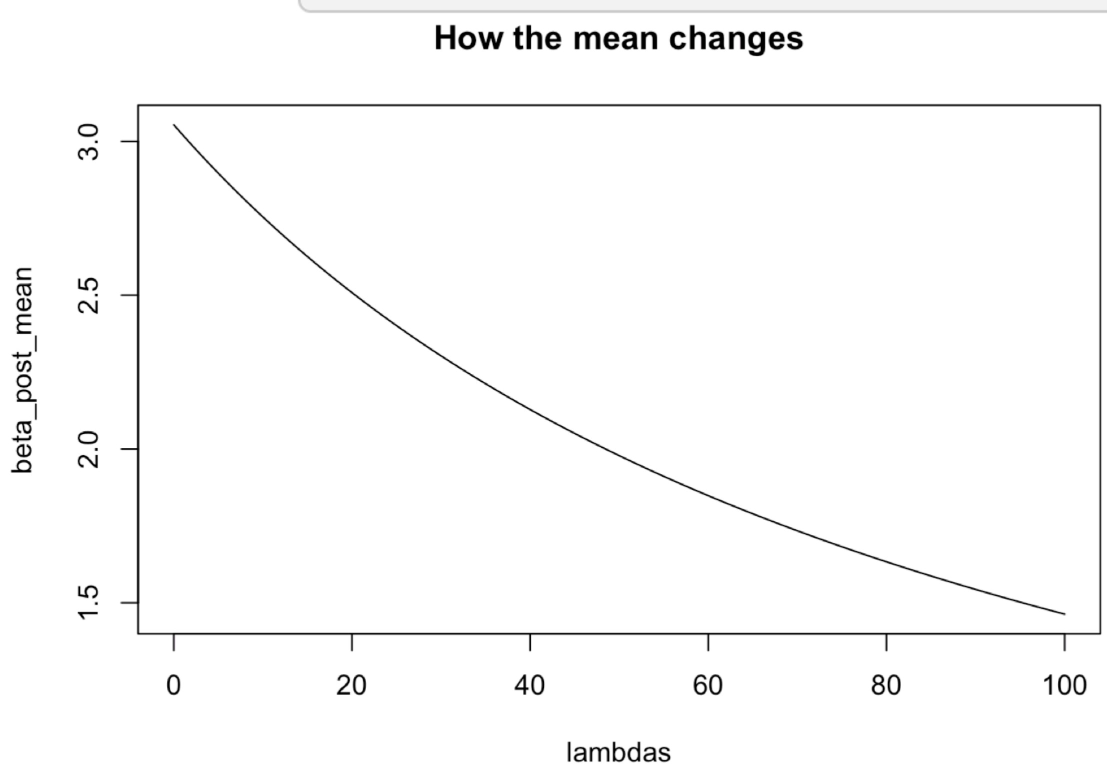
- The ratio $\frac{\sigma^2}{\tau^2}$ controls the amount of information that comes from the data vs the amount that comes from the prior

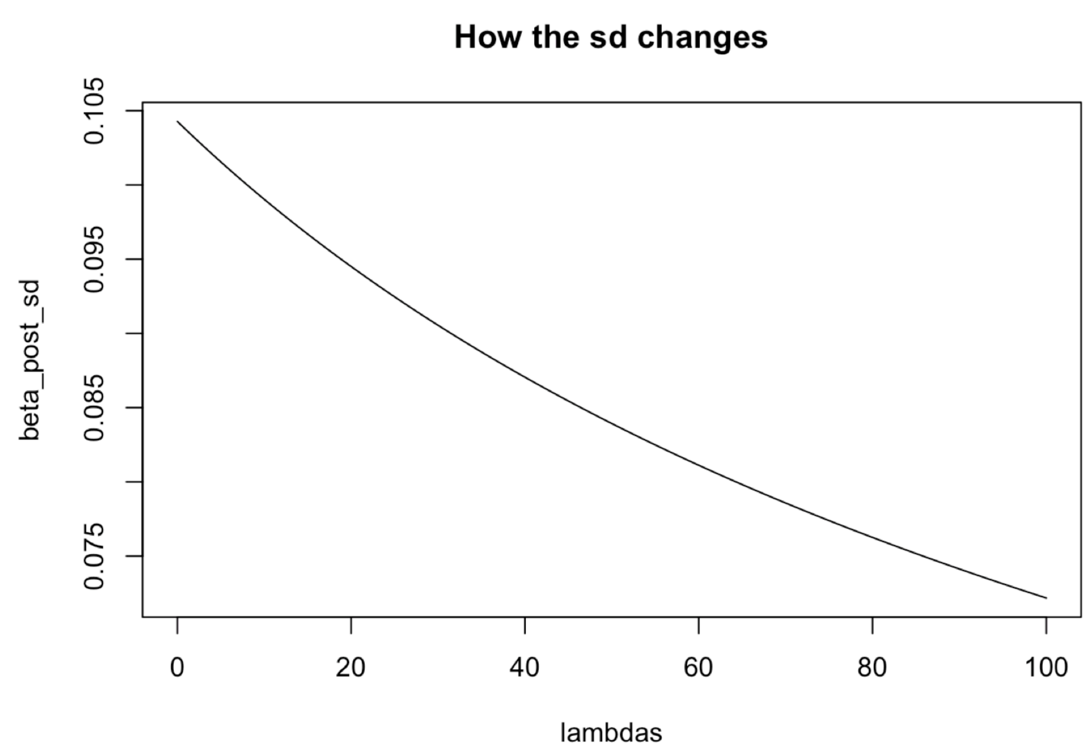
AN EXPERIMENT

```r
N = 100 #size of problem
x = rnorm(N) #make a covariate
y = 0.2 + 3*x + -.5*rnorm(N)

# a set of lambda = sigma^2/tau^2
lambdas = seq(0,100,length.out=100)
beta_post_mean = rep(NA,100)
beta_post_sd = rep(NA,100)
for (i in 1:100){
  beta_post_mean[i] = (t(x)%*%x + lambdas[i])^(-1)*t(x)%*%y
  beta_post_sd[i] = (t(x)%*%x + lambdas[i])^(-1/2)
}

plot(lambdas, beta_post_mean,type="l",main="How the mean changes")
```



SO THE PRIOR HAS AN EFFECT

- So how do we choose this value of $\frac{\sigma^2}{\tau^2}$?
- Answer: We don't!

- The prior variance is selected as we showed last week, but we actually don't need to choose a specific value of σ

- We can instead run the Bayesian machinery again!

AVERAGE IT OUT

- If we can find the **joint posterior** $p(\beta, \sigma^2 \mid y)$ then we can **marginalize out** the standard deviation

$$p(\beta \mid y) = \int p(\beta, \sigma^2 \mid y)\, d\sigma^2 = \int p(\beta \mid y, \sigma^2) p(\sigma^2 \mid y)\, d\sigma^2$$

- If we do this, instead of choosing the **best** value of $\dfrac{\sigma^2}{\tau^2}$ we can average over the values that are most consistent with the data

- This allows us to reflect the uncertainty we have about this parameter

BUT HOW DO WE GET THE POSTERIOR FOR SIGMA?

- There are lots of ways to do this, but here's one cute trick:

$$p(A, B \mid y) = \frac{p(A, B, y)}{p(y)}$$

$$p(A \mid B, y)p(B \mid y) = \frac{p(A, B, y)}{p(y)}$$

$$p(A \mid B, y) = \frac{p(A, B, y)}{p(y)p(B \mid y)}$$

$$p(A \mid B, y) \propto \frac{p(A, B, y)}{p(B \mid y)}$$

- So we can get the posterior for σ^2

$$p(\sigma^2 \mid y) \propto \frac{p(y \mid \beta, \sigma^2)p(\beta)p(\sigma^2)}{p(\beta \mid \sigma^2 y)}$$

NB: Left hand side does not depend on beta, so we can put any value of beta on the right!

HOW DO WE CHOOSE THE PRIOR FOR SIGMA

- The prior on the observation variance can come from two places:
  - Structural knowledge about the measurement process (accuracy of measurement equipment)
  - Knowledge of the data (you can't breathe concrete)
- One important thing: **never specify a prior directly on the variance!**
- Recall: if data is Gaussian, we are always within 3 **standard deviations** of the mean, which makes the standard deviation the natural parameter to put a prior on.

ASIDE: CHANGING VARIABLES

- Recall that if we have a prior $p(u)$ and we want a prior for the parameter $v = g(u)$ we need to transform it **carefully!**

$$\Pr(u < t) = \int_{-\infty}^{t} p(u)\, du = \int_{g^{-1}(\infty)}^{g^{-1}(t)} p\left[g^{-1}(v)\right] \left| \frac{d}{dv} g^{-1}(v) \right| \, dv = \Pr\left[v < g^{-1}(t)\right]$$

- **So a prior for standard deviation** $p_\sigma(\sigma)$ **becomes a prior for variance** $v = \sigma^2$

$$p(v) = \frac{1}{2} v^{-1/2} p_\sigma\left(\sqrt{v}\right)$$

THE POSTERIOR FOR SIGMA

- Putting this in we get

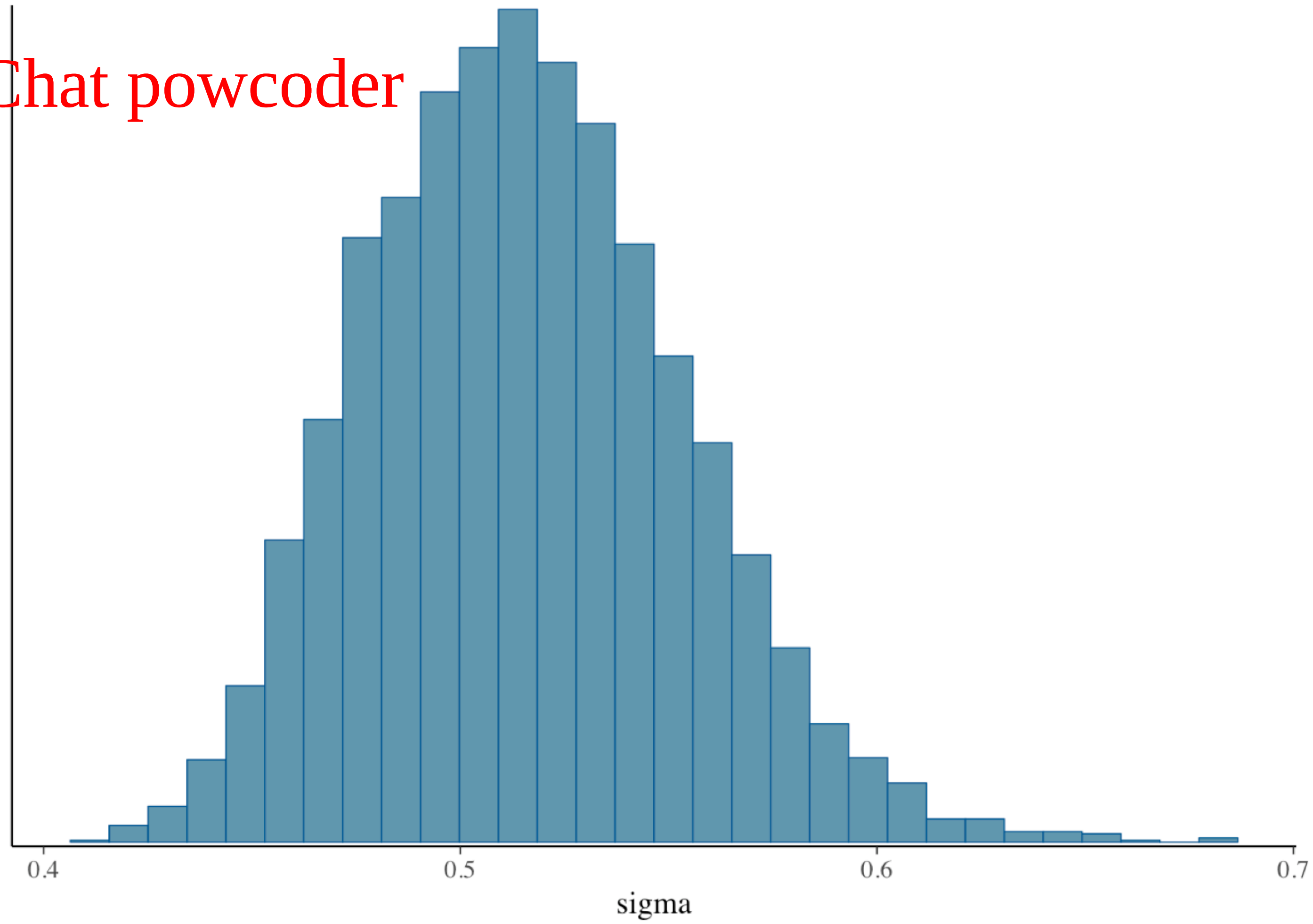
$$p(\sigma^2 \mid y) \propto \frac{\sigma^{-n} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)^T(y - X\beta) - \frac{1}{2\tau^2}\beta^T\beta\right] p(\sigma^2)}{|\Sigma_{post}|^{-1/2} \exp\left[-\frac{1}{2}(\beta - \mu_{post})^T \Sigma_{post}^{-1}(\beta - \mu_{post})\right]}$$

$$\propto \sigma^{-n} |\Sigma_{post}|^{1/2} \exp\left[-\frac{1}{2\sigma^2} y^T y + \frac{1}{2}\mu_{post}^T \Sigma_{post}^{-1} \mu_{post}\right] p(\sigma^2)$$

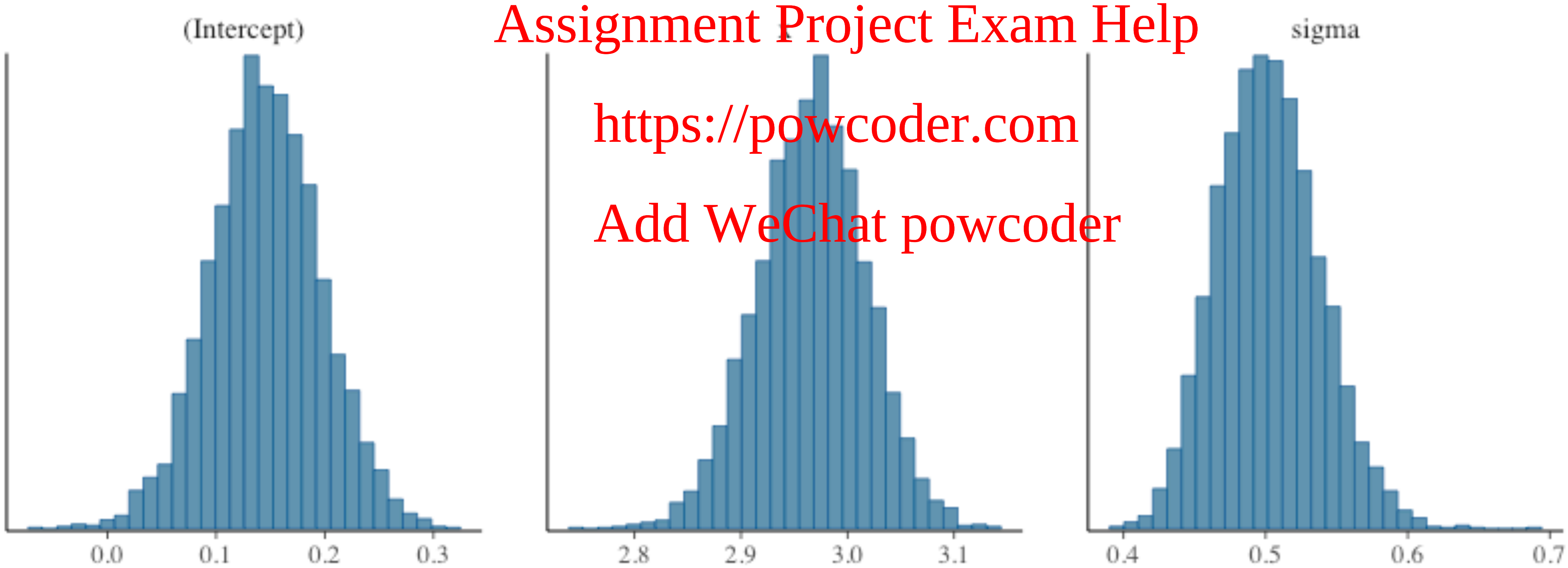
- Yuck!

WE DO IT COMPUTATIONALLY

```
library(rstanarm)
fit <- stan_glm(formula= y~1+ x, data = data.frame(y=y,x=x), family=gaussian(),
               prior_intercept = normal(location=0,scale=5,autoscale=FALSE),
               prior =normal(location=0,scale=5,autoscale=FALSE))
plot(fit,pars="sigma","hist")
```

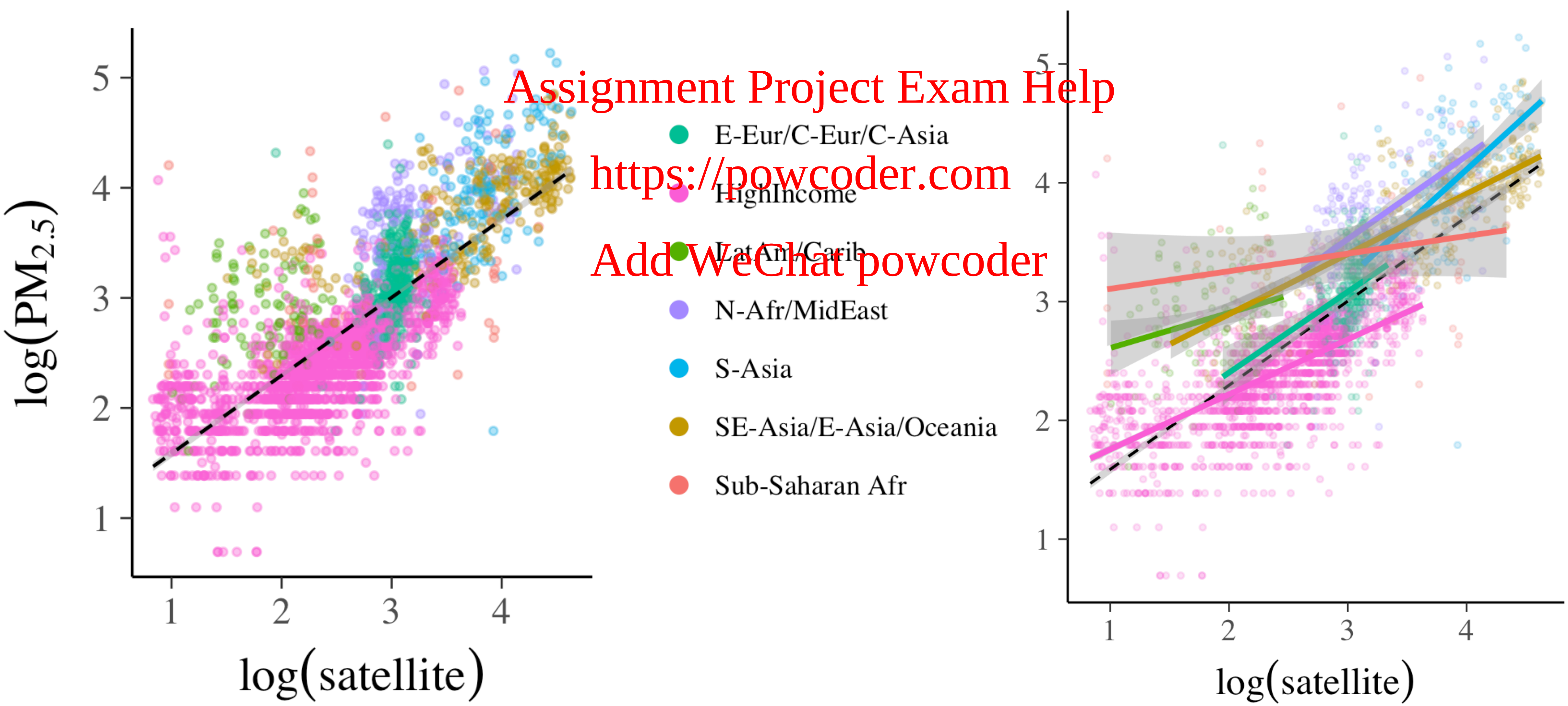



HISTOGRAMS



BACK TO THE AIR POLLUTION

- Last time, we looked at the global PM2.5 data



FITTING ONE LINE TO ALL THE DATA

➤ The (log) PM2.5 measurement is well-predicted by the (log) satellite data

```
load("~/Documents/bayes-vis-paper/bayes-vis.RData")
fit_total = lm(log(pm25) ~ log(sat_2014), data = GM)
summary(fit_total)
```

```
##
## Call:
## lm(formula = log(pm25) ~ log(sat_2014), data = GM)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.86688 -0.26833 -0.06241  0.23273  2.63544
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    0.87899    0.02938   29.92   <2e-16 ***
## log(sat_2014)  0.70828    0.01058   66.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4485 on 2978 degrees of freedom
## Multiple R-squared:  0.6008, Adjusted R-squared:  0.6006
## F-statistic:  4481 on 1 and 2978 DF,  p-value: < 2.2e-16
```

FITTING ONE LINE TO EACH REGION

- What if we instead fit an individual line to each region?

```
output = data.frame(super_region = super_region_list, R2 = rep(NA,7))
dat = data.frame(pm25 = GM$pm25, sat_2014 = GM$sat_2014, super_region = GM$super_region)
for(i in 1:7) {
  dat_local = dat %>% filter(super_region ==i)
  output$R2[i] = summary(lm(log(pm25)~log(sat_2014), data=dat_local))$r.squared
}
print(output,digits=2)
```



```
##             super_region    R2
## 1             HighIncome 0.453
## 2           N-Afr/MidEast 0.322
## 3                 S-Asia 0.345
## 4       E-Eur/C-Eur/C-Asia 0.087
## 5             LatAm/Carib 0.055
## 6 SE-Asia/E-Asia/Oceania 0.561
## 7          Sub-Saharan Afr 0.046
```

TINY R-Squared!

TWO BAD SOLUTIONS

- We are doing pretty badly if we use all the data
  - The estimate is dominated by all the data in high-income countries.

- We are doing pretty badly if we separate into regions

  - A lot of regions don't have enough information to get a good regression line.

- Is there a compromise?

A COMPROMISE

TWO ENDS OF A CONTINUUM

- What we saw were two ends of a continuum for fitting regressions over multiple, linked, data sets.

- **Complete pooling:** When we fitted the global regression, we pooled all of our data together to get a single estimate

- **No pooling:** When we fitted individual regressions to each region, we did not share any information between the regions

- **Partial pooling: ????**

LET'S START WITH A SIMPLER CASE

- Instead of trying to estimate a mean and an intercept, let's just focus on the intercept (ie we have no covariate).

- The complete pooling estimate is

$$\mu = \frac{1}{n}\sum_{i=1}^{n} y_i$$

- The no pooling estimate is

$$\mu_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}$$

here y_{ij} is the ith measurement of region j

WHAT IF WE MADE A BAYESIAN MODEL

- Consider the following model

$$y_{ij} \mid \mu, \sigma \sim N(\mu_j, \sigma^2)$$
$$\mu_j \sim N(\mu, \tau^2)$$

- This is different from complete pooling because it estimates a different intercept for each region.

- This is different from no pooling because the intercepts are no longer independent: they're *a priori* assumed to be draws from the same normal distribution

- This means that we don't expect the region means to be more than 3τ apart.

THE POSTERIOR MEAN

- The partially pooled posterior mean is

$$\mathbb{E}(\mu_j \mid y) = \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\mu}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

$$\approx \frac{\frac{n_j}{\sigma^2}\bar{y}_j + \frac{1}{\tau^2}\bar{y}_{all}}{\frac{n_j}{\sigma^2} + \frac{1}{\tau^2}}$$

- So when a region has a lot of data, we are close to the region average, while when there is little data we are close to the global average.

EXTENDING TO REGRESSION

- We can extend this model to cover regression in a straightforward way

$$y_{ij} \mid \mu, \beta, \sigma \sim N(\mu_j + \beta_j x_{ij}, \sigma^2)$$

$$\mu_j \sim N(\mu, \sigma_\mu^2)$$

$$\beta_j \sim N(\beta, \sigma_\beta^2)$$

- This allows the slope to change in each region as well.

- Note that the intercept and slope will generally have different amounts of pooling

- These models are known as **multilevel models**

HOW DO WE CHOOSE THE AMOUNT OF POOLING?

- We don't!!!!!

- Once again, we let Bayes do the hard work for us.

- Set a (sensible) prior on τ, compute the posterior, and average over all of the partially pooled models that are consistent with the observed data!

- (Once again, the posterior for τ is ugly, but we get it the same way as yesterday)

THIS IS BASICALLY WHAT WE WANT

- We want the local data in the region to be represented, but we also want to use the global data when needed.

- We're assuming the regions are **exchangeable**, which may not be the best assumption.

- We're going to work on breaking this as we go further in the course.