Announcements

Make sure to sign in on the google form (I send a list of which section questions are useful for which pset questions afterwards)

Final project proposals due 11/21

Introductions

- One question or thought related to lecture last week (Weighted least squares, quantile regression, mixed effects models)

Many ways to peel an orange

Recall the following linear model extensions:

- Heteroscedasticity-consistent standard errors: $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$, $\epsilon \sim \text{MVN}(\vec{0}, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is a diagonal matrix with non-identical entries.
- Weighted least squares: $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$, $\epsilon \sim \text{MVN}(\vec{0}, \mathbf{W}^{-1}\sigma^2\mathbf{I}_{n\times n})$ where $\mathbf{W}$ is a diagonal matrix of weights.
- Huber's method: Minimize loss $= \sum_i^n g(\hat{\epsilon}_i)$ with

$$g(x) = \begin{cases} x^2/2 & |x| < c \\ c|x| - c^2/2 & |x| \geq c \end{cases}$$

- Block correlations: $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$, $\epsilon \sim \text{MVN}(\vec{0}, \mathbf{\Sigma})$ with a covariance matrix like

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho_1 & \rho_1 & 0 & 0 \\ \rho_1 & 1 & \rho_1 & 0 & 0 \\ \rho_1 & \rho_1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & \rho_2 \\ 0 & 0 & 0 & \rho_2 & 1 \end{bmatrix}$$

- Autoregressive: $\vec{Y} = \mathbf{X}\vec{\beta} + \vec{\epsilon}$, $\epsilon \sim \text{MVN}(\vec{0}, \mathbf{\Sigma})$ with a covariance matrix like

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{bmatrix}$$

- Random intercepts: $Y_{ij} \sim \alpha + \alpha_i + \beta X_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_Y^2)$ and $\alpha_i \sim \mathcal{N}(0, \sigma_\alpha^2)$. (Note that this version explicitly includes a fixed intercept and a random intercept with mean 0. Random effects as you'll see them elsewhere (and in `lmer`) are usually specified to have mean 0, so including the fixed effect separately is a good practice.)
- Random slopes: $Y_{ij} \sim \alpha + (\beta + \beta_i)X_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_Y^2)$ and $\beta_i \sim \mathcal{N}(0, \sigma_\beta^2)$.

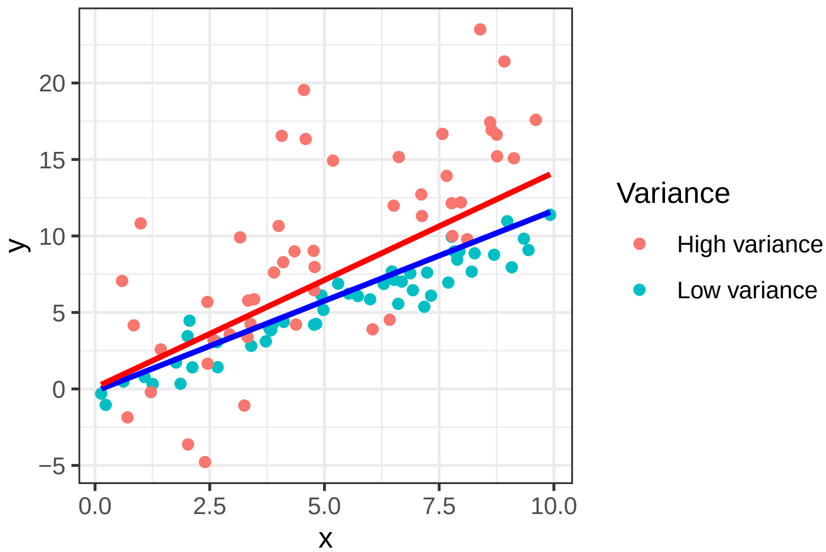
1. For each of the scenarios below, which of the above linear model extensions would you use? Why?

- Stock price over a week: Autoregressive (a high residual on one day might indicate a high residual on the next day)
- Wheat yield vs phosphorus fertilizer use in various Kansas counties: Block correlation (residuals are correlated within a county), random intercepts (different counties have different baseline wheat yields), or random slopes (phosphorus use changes wheat yield differently in each county)

- Soil nitrate concentration vs fertilizer use where the nitrate measurements are taken with instruments of varying precision: Heteroscedasticity-consistent standard errors or weighted regression (less precise instruments have higher standard errors or lower weights)
- Spotify listens vs release year: Huber's method (trying to mitigate the outlier effect of songs that are very popular)
- The day in the year leaves start to fall vs latitude for various tree types: Block correlation (trees of the same type lose their leaves around the same time), random intercepts (trees of the same type have different baseline leaf shedding times), or random slopes (different trees respond to different latitudes to lose their leaves at different times)

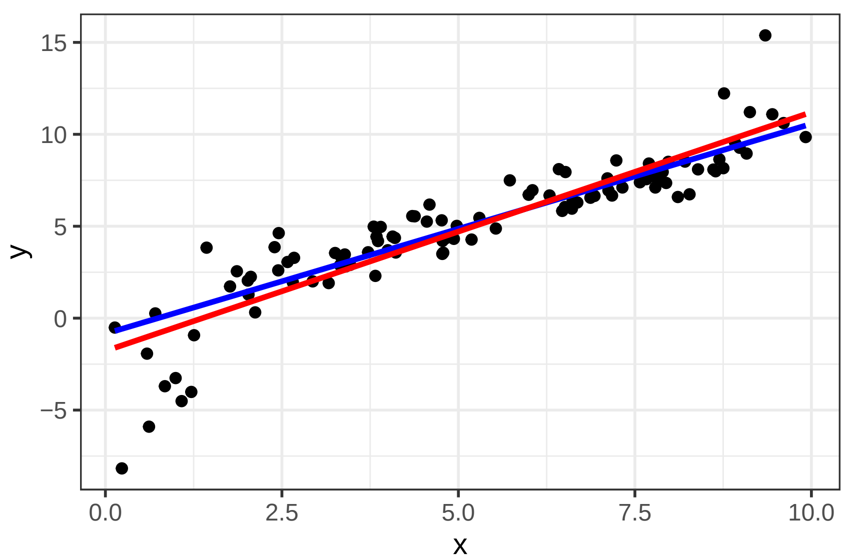
2. Which line is which in the following models?

- One line is a standard OLS; the other uses weighted regression.



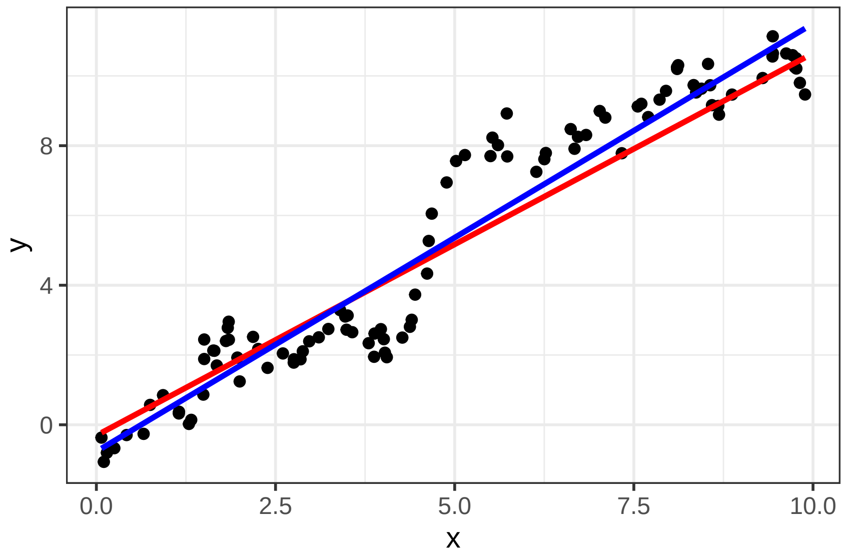
Blue is the weighted regression. It weights low variance observations more heavily as expected.

- One line is a standard OLS; the other uses Huber's method.



Blue is Huber's method. It is not as sensitive to outliers.

- One line is a standard OLS; the other uses autoregression.

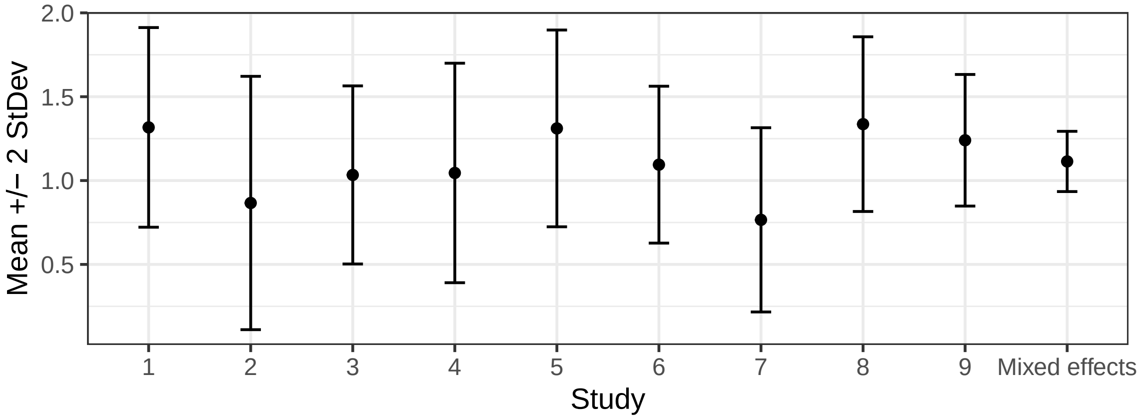


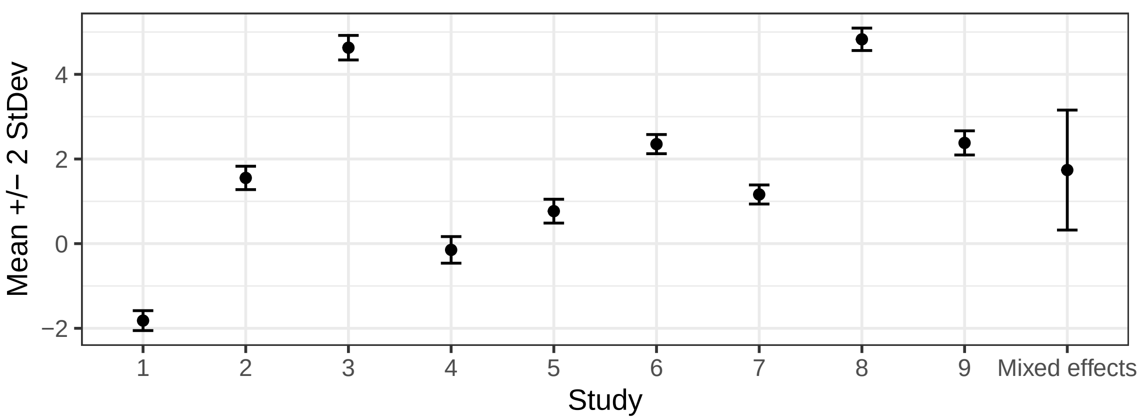
Red is the autoregression. It's less influenced by tightly correlated nearby points.

Meta-analysis analysis

One common use of mixed effects models is in meta-analysis to aggregate the results of multiple studies. This question will explore various methods of meta-analysis aggregation.

1. A common plot in meta-analysis is a forest plot, showing the mean and standard error of some quantity of interest as determined by various studies as well as an aggregated mean and standard error. Given the following forest plots, plot the approximate mean and standard error for the aggregated effect from a linear mixed effects model.





2. Sometimes, studies will not make available all their individual data points. In these cases, we might only have a point estimate and standard error for a particular estimand of interest. Assume that for study i we have our statistic of interest

$$\hat{\beta}_i \sim \mathcal{N}(\beta, \sigma^2_{\hat{\beta}_i})$$

This model specifies that there is some underlying effect β and that each study will find some $\hat{\beta}_i$ with a standard error based on the study's sample size etc. If we have n_{studies} each with a $\hat{\beta}_i$ and a $\sigma^2_{\hat{\beta}_i}$, use maximum likelihood estimation to find a point estimate for β.

Since the standard errors are known, we can keep only the pieces that depend on β,

$$L \propto \prod_{i=1}^{n_{\text{studies}}} e^{\frac{1}{2}\left(\frac{\hat{\beta}_i - \beta}{\sigma_{\hat{\beta}_i}}\right)^2} \implies \log(L) = \sum_{i=1}^{n_{\text{studies}}} \frac{1}{2}\left(\frac{\hat{\beta}_i - \beta}{\sigma_{\hat{\beta}_i}}\right)^2$$

Taking the derivative and setting it equal to 0 shows

$$0 = \sum_{i=1}^{n_{\text{studies}}} \left(\frac{\hat{\beta}_i - \hat{\beta}}{\sigma_{\hat{\beta}_i}}\right) = \sum_{i=1}^{n_{\text{studies}}} \left(\frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}\right) - \hat{\beta} \sum_{i=1}^{n_{\text{studies}}} \left(\frac{1}{\sigma_{\hat{\beta}_i}}\right) \implies \hat{\beta} = \frac{\sum_{i=1}^{n_{\text{studies}}} \left(\frac{\hat{\beta}_i}{\sigma_{\hat{\beta}_i}}\right)}{\sum_{i=1}^{n_{\text{studies}}} \left(\frac{1}{\sigma_{\hat{\beta}_i}}\right)}$$

3. Find the distribution of this $\hat{\beta}$ assuming the studies are independent.

Each of the $\hat{\beta}_i$s are Normal, so $\hat{\beta}$ has a Normal distribution. Its mean is

$$E(\hat{\beta}) = \frac{\sum_{i=1}^{n_{\text{studies}}}\left(\frac{E(\hat{\beta}_i)}{\sigma_{\hat{\beta}_i}}\right)}{\sum_{i=1}^{n_{\text{studies}}}\left(\frac{1}{\sigma_{\hat{\beta}_i}}\right)} = \beta$$

Its variance is

$$\text{Var}(\hat{\beta}) = \frac{\sum_{i=1}^{n_{\text{studies}}}\left(\frac{\text{Var}(\hat{\beta}_i)}{\sigma^2_{\hat{\beta}_i}}\right)}{\left(\sum_{i=1}^{n_{\text{studies}}}\left(\frac{1}{\sigma_{\hat{\beta}_i}}\right)\right)^2} = \frac{n_{\text{studies}}}{\left(\sum_{i=1}^{n_{\text{studies}}}\left(\frac{1}{\sigma_{\hat{\beta}_i}}\right)\right)^2}$$

Therefore,

$$\hat{\beta} \sim \mathcal{N}\left(\beta, \frac{n_{\text{studies}}}{\left(\sum_{i=1}^{n_{\text{studies}}}\left(\frac{1}{\sigma_{\hat{\beta}_i}}\right)\right)^2}\right)$$

4. The code below simulates data from the following scenarios:

- In the first scenario, there is a true effect β that holds in all the studies. The only thing that differs by study is the sample size. Therefore, for study i, the j^{th} observation is given by $Y_{ij} = \beta_0 + \beta X_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_Y^2)$.
- In the second scenario, the effect in study i is $\beta_i \sim \mathcal{N}(\beta, \sigma_\beta^2)$. Then, for study i, the j^{th} observation is given by $Y_{ij} = \beta_0 + \beta_i X_{ij} + \epsilon_{ij}$ with $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_Y^2)$.

We will consider 5 methods:

- Use the likelihood method above
- Find $\hat{\beta}_i$ for each study and take the average to estimate β.
- Use a mixed effects model `ys ~ xs + (xs - 1 | study)` (random slopes, no random intercepts)
- Combine all the data points and run a linear model with `ys ~ xs + xs:as.factor(study_id)`
- Combine all the data points and run a linear model with `ys ~ xs`

Interpret the outputs: how do the methods perform on the two scenarios?

```
##                            Method   Bias    SE   MSE
## 1                      Likelihood -0.003 0.502 0.252
## 2                       Raw means -0.005 0.624 0.389
## 3             Linear mixed effects  0.000 0.464 0.215
## 4 Fixed effects with interaction -0.015 1.869 3.495
## 5   Fixed effects no interaction  0.001 0.455 0.207
```

```
##                            Method   Bias    SE   MSE
## 1                      Likelihood -0.001 0.874 0.763
## 2                       Raw means -0.008 0.908 0.825
## 3             Linear mixed effects -0.009 0.862 0.742
## 4 Fixed effects with interaction  0.006 2.719 7.395
## 5   Fixed effects no interaction  0.006 0.954 0.909
```

The fixed effects with interactions is clearly the worst model: it actually just takes study 1 as the estimated effect and fits adjustments for the rest of the studies. The raw means also does rather poorly with a higher standard error than the other methods. Fixed effects with no interactions does the best when all the participants in all the studies are from the same distribution, but its standard error is higher when there are by-study differences. Linear mixed effects performs very well in both scenarios, and the likelihood method also does quite well.

5. Under the random slopes model specified above but assuming an intercept of 0, given a new X from a new study, find the probability that its associated Y value will be less than τ. You can leave your answer as a double integral.

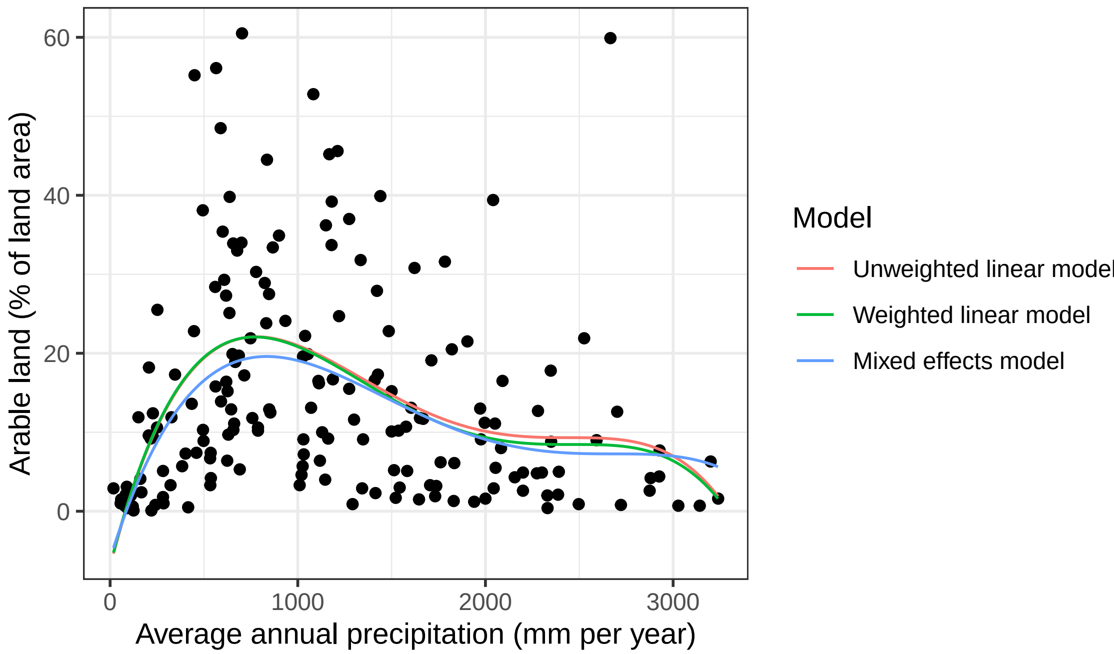
Conditioning on the group β_i and applying the Law of Total Probability,

$$\begin{aligned} P(Y < \tau | X) &= \int_{-\infty}^{\infty} P(Y < \tau | \beta_i, X) P(\beta_i) d\beta_i \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\tau} \frac{1}{\sqrt{2\pi}\sigma_Y} e^{-\frac{1}{2}\left(\frac{y - X\beta_i}{\sigma_Y}\right)^2} \frac{1}{\sqrt{2\pi}\sigma_\beta} e^{-\frac{1}{2}\left(\frac{\beta_i - \beta}{\sigma_\beta}\right)^2} dy d\beta_i \end{aligned}$$

Crops continued

This problem will deal with a dataset of country-level statistics from UNdata, Varieties of Democracy, and the World Bank.

1. The following shows a quartic regression model to predict the percent of arable land from the annual precipitation. The weighted model is weighted by the log GDP per capita (the idea being that wealthier countries might have better measurement and reporting). The mixed effects model uses random intercepts for each world region (e.g., because temperature can differ by region). Interpret the differences.



The unweighted and weighted linear models are very similar. The random intercepts model predicts a lower proportion of arable land across all precipitations. All models look roughly similar with similar precipitations at which arable land is maximized.