Effect of chemical exposure on preterm risk

Longnecker data and the Collaborative Perinatal Project

- Data from the National Collaborative Perinatal Project (CPP)
- Relates gestational age to chemical exposure (DDE and PCBs) and other factors in 2380 pregnant women.

Goal:

- Assess how exposure to DDE and PCBs impact the risk of preterm birth, defined as delivery before 37 weeks.

This talk:

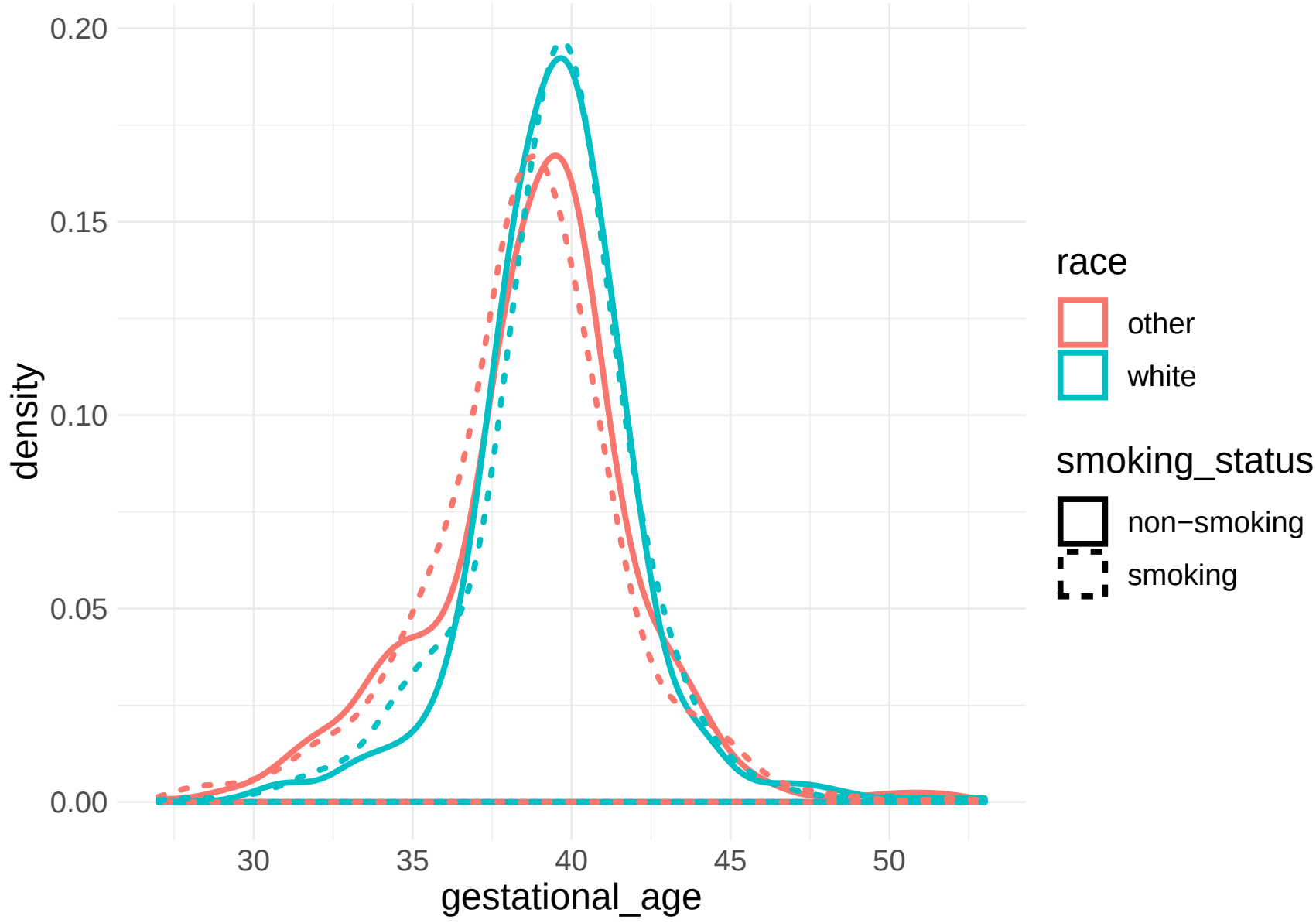
- Walk through some of our thought process addressing this.

Data cleaning and transformations

- Removed pregnancies over 55 weeks
- Dropped albumin variable (93% missing values)
- Combined "other" (n=136) and "black" in a single class of size 1336.
- Summarized PCBs by summation.
- Observation # 1857 has missing values; we removed it.

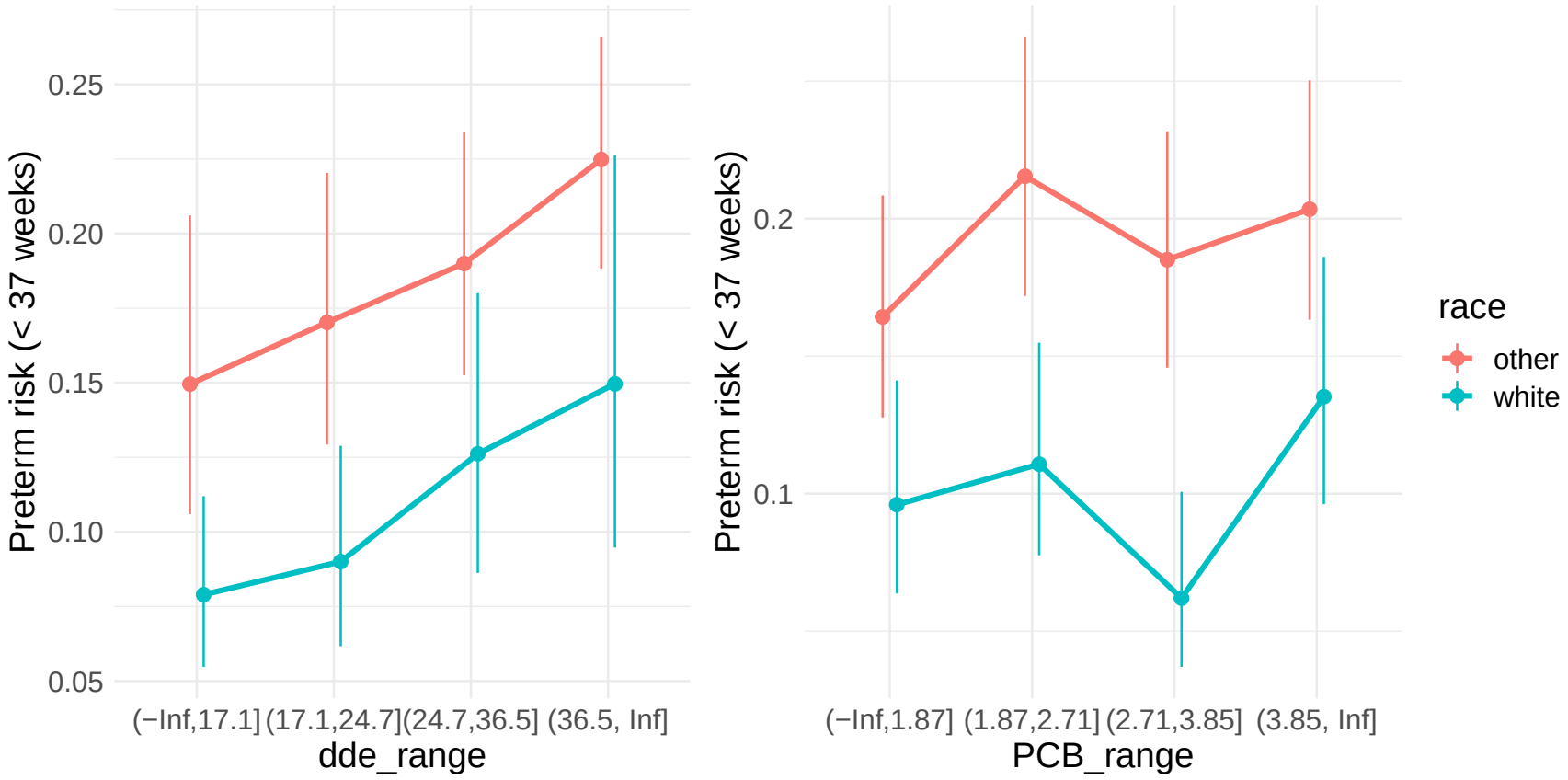
Note: Still 22% missing values in the score_* variables.

A look at the data



−>

A look at the data



Imputation

- Approximately 79% of the observations were complete cases. Most incomplete cases came from the "score" variables.
- A standard Bayesian approach to data imputation was taken for each score variable:
  - The observed score variables were regressed onto the other predictors.
  - Missing values were treated as model parameters and were estimated using their respective posterior mean.
- A few potential problems with this approch:
  - Treating each score variable as a linear function of the other predictors is hard to justify.
  - Multicolinearity amongst the predictors (e.g., *cholesterol* and *triglycerides*).
- Possible improvements:
  - Propogate uncertainty associated with using imputation methods.
  - Apply a non-linear model to estimate missing values.

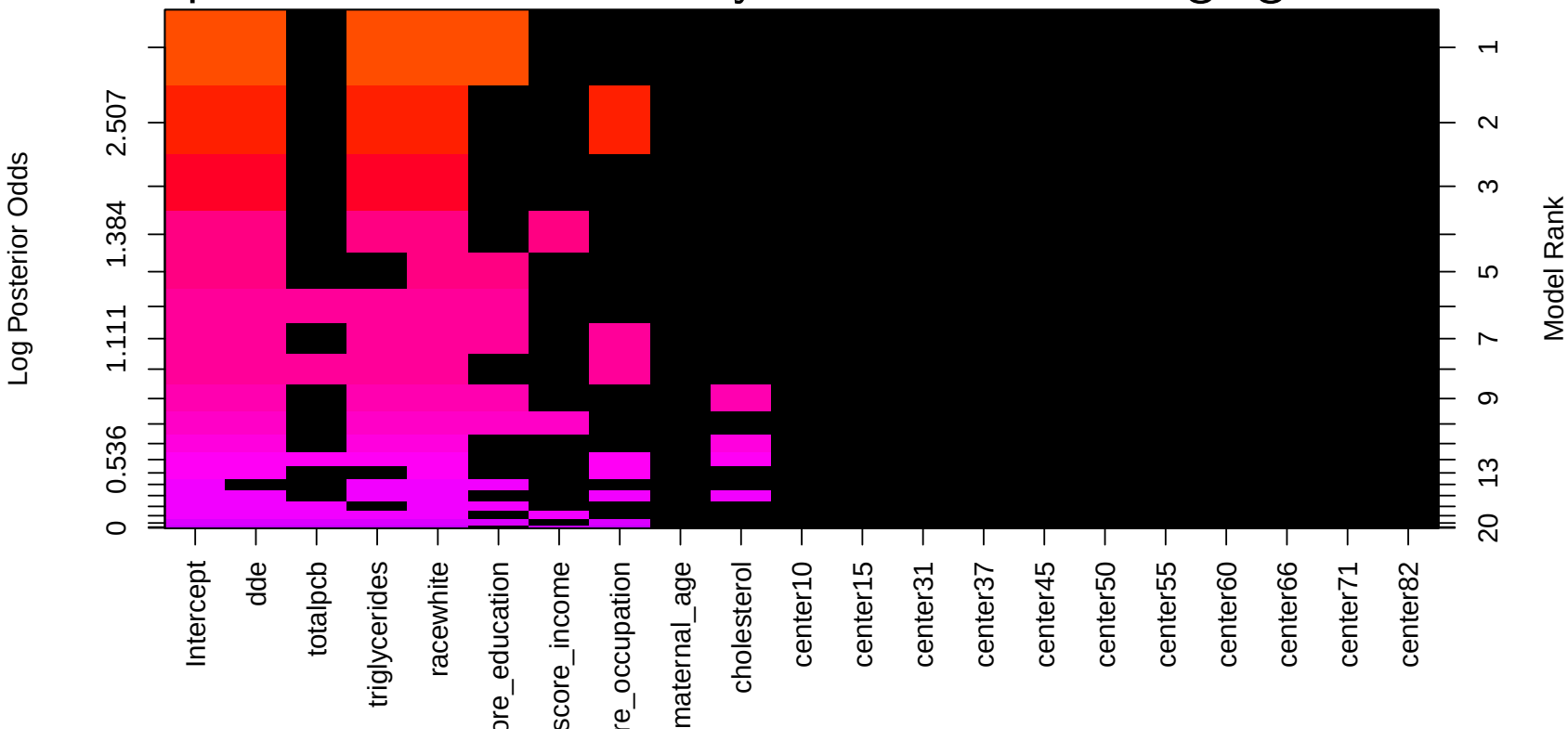
Model

We first define "preterm" as gestational age less than 37 months, and compare it to "early preterm," defined as gestational age less than 34 months.

Since the response is a binary, We use logistic regression to regress preterm against all other variables. We find that dde, totalpcb, triglycerides, race, and center 37 are statistically significant

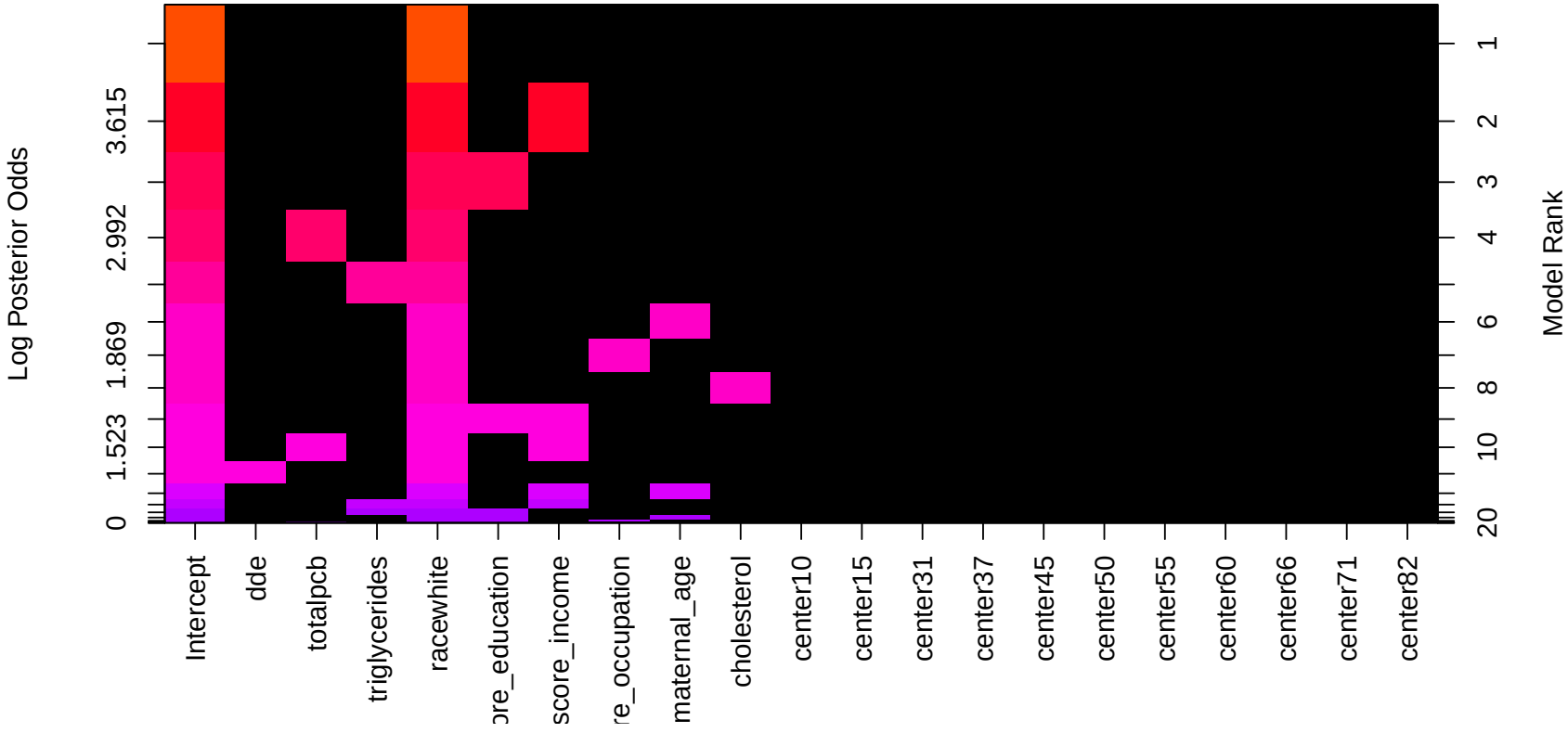
Model

We compare this result with Bayesian Model Averaging.



Model

When we regress "early preterm" through logistic regression, we find only totalbcp and race significant. We compare this with BAS, where we
find that only race shows up substantially often in simulated data sets.



Model

In comparing results based on different definitions of preterm birth, we find that it is likely the case that DDE (and to a lesser extent, PCB) are associated with early birth. However, in the more extreme cases, where gestational age is less than 34 years, there are likely other factors not represented in the data that account for this difference.

BAS and GLM produced different results for PCB in each case. We believe this is due to low group size. (74/1853 women had births before 34 months)

Imputation

- Approximately 78% of the observations were complete cases. Most incomplete cases came from the "score" variables.
- A standard Bayesian approach to data imputation was taken for each score variable:
- The observed score variables were regressed onto the other predictors.
- Missing values were treated as model parameters and were estimated using their respective posterior mean.
- A few potential problems with this approach:
- The missing at random (MAR) assumption may be inappropriate.
- Treating each score variable as a linear function of the other predictors is hard to justify.
- Multicollinearity amongst the predictors (e.g., *cholesterol* and *triglycerides*).

Conclusion

- Both models indicated *dde* may play a role in preterm risk. In addition, *pcb* appears to play at most a minor role in preterm risk.
- Partial correlation amongst *dde* and *pcb* were analyzed through the following procedure:
- *dde* and *pcb* were regressed onto the other predictors first to remove any linear dependence.
- The residuals from the model were then extracted and yielded a significant correlation $\rho \approx 0.30$.
- This limits any causal interpretation of the effects of *dde* and *pcb* on preterm births.
- Non-linear dependence was also explored using a support vector machine and produced similar results.

Improvements

- Account for the heterogeneity amongst the different centers. Some centers differed significantly with respect to race, sample size, and occurrence of preterm births.
- Control for non-linear relationships between the socio-economic variables and preterm risk through a non-linear model.
- Improve the data imputation method. A couple examples include propagating uncertainty associated with using imputation methods or applying a non-linear model to estimate missing values.