Assessing the Influence of DDE and PCBs on the Likelihood and Severity of Preterm Births

Phuc Nguyen, Joseph Lawson, Emily Gentles

1/23/2020

Abstract

We examine a data set from the Collaborative Perinatal Project to assess how the risk of pre-term births are influenced by the mother's exposure to DDE and PCBs. We find positive evidence of an effect, using an ordinal logistic regression model to assess risk of different severities of pre-term births accounting for exposure to DDE, PCBs as well as potential socioeconomic and health-status confounders. Specifically, we find that a 10% increase in exposure to DDE is associated with a 1.8% increase in odds of having more premature birth, and a 10% increase in exposure to PCBs increases this odds by 0.6%. We discuss the advantages and drawbacks our methods as well as potential alternatives.

1 Introduction

The data used in this analysis is a subsample of the data collected in the National Collaborative Perinatal Project (CPP), a program which enrolled women during pregnancy and followed their pregnancy outcomes. The pregnancy outcome of interest in this analysis is premature delivery, which is typically defined as having gestational age at delivery of 37 weeks or less. The project collected demographic information including the mother's race, age, education, occupation as well as a lifestyle factor that is smoking status. The subsample contains 2380 women for whom additional assaying of serum samples was perfromed and the concentration doses for DDE and PCBs as well as cholesterol and triglycerides was recorded. DDE and PCBs are of particular interest as they are breakdown products of pesticides that bioaccumulate in fatty deposits of organisms and are thought to have an impact on human reproduction (see Cohn et al. 2003, also 2011). The goal of this analysis is to assess how exposure to DDE and PCBs relates to the risk of premature delivery.

2 Materials & Methods

In order to assess how exposure to DDE and PCBs relates to the risk of premature delivery, we investigate the associations between the concentration of DDE and PCBs and categories of preterm birth, accounting for other potential confounders. Specifically, we define categories of preterm birth as severely premature for having gestational ages of less than 33 week, premature for having gestational ages of less than 38 weeks, and full- to late-term for having gestational ages of between 38 and greater based on the guidelines provided by the World Health Organization ("Preterm Birth" 2018). The natural order of preterm birth categories lead us to employ ordinal logistic regression as our modeling method (McCullagh 1980). This method provides interpretations in terms of change in risk of more premature birth as exposure to DDE and PCBs increases. This structure also addresses the non-normality, heavy-tailedness of the response which impedes the use of standard linear regression. The model has the following structural form:

$$logit(Y_i \leq j) = \alpha_j - \beta^T X_i$$

From which one derives:

$$P(Y_i \leq j) = \frac{e^{\alpha_j - \beta^T X_i}}{1 + e^{\alpha_j - \beta^T X_i}}$$

In the above, Y_i is the response category, j is some possible level of Y_i, α_j, is a constant associated with level j, β is a vector of coefficients shared across all levels, and X_i is the vector of predictors for the i^{th} observation. Note that unlike multinomial regression, in which one obtains $P(Y_i = j)$, we have an inequality instead. The phrase "proportional odds" comes from the assumption under the model that the odds ratio of $Y_i \leq j$ vs $Y_k \leq j$ is constant across the range of j.

The reasonable use ordinal logstic regression relies on several assumptions: ordered categorical response variable, proportional odds, lack of multi-collinearity, and predictors are either categorical, ordinal, or continuous. We perform a Principal Component Analysis (PCA) of the PCB predictors, which exhibit relatively high correlations (1 (right)), to produce new uncorrelated variables, guarenteeing the multi-collinearity assumption. The proportional odds assumption may be tested via the Brant test for which we obtain satisfactory results (Brant 1990). We note that "Score" variables have a high degree of missingness and use the MICE package in R to impute these missing data points. We remove one observation with missing PCB data. We also include other potential confounders such as blood cholesterol/triglyceride levels, the center where these measurements were taken and metrics capturing socioeconomic status and lifestyle in the model. Finally, we perform model selection using F-tests.

3 Results

3.1 Exploratory Data Analysis

Many of the variables in the dataset exhibit a variety of issues such as measurement errors, zero-inflation, multi-collinearity and missingness. Gestational age potentially contain measurement errors since its values range from 27 to 90 (Figure 2 left). We address this issue by keeping observations with 44 weeks or less in gestational age, a range plausible in reality.

We also see an inflation at zeros in the PCB variables due to the limitations of measurement equipment (Figure 2 right). Consequently, we add half of the smallest value of PCB measurements to the zero and log-transform these variables. As seen in Figure 1 (right), the PCB variables are highly correlated. Applying PCA to these variables addresses the issues. Among the predictors, only DDE and the first principal component of PCBs, triglycerides and cholesterol are mildly correlated (Figure 1 left).

Missingness is most significantly present in the variable albumin for which less than 10% of women had their levels recorded. Figure 3 shows that being tested for albumin seems to associate with slightly larger gestational age. Bivariate plots show associations between premature birth and predictors such as race, centers, a quadratic function of maternal age, concentration of DDE and cholesterol (Figure 3).

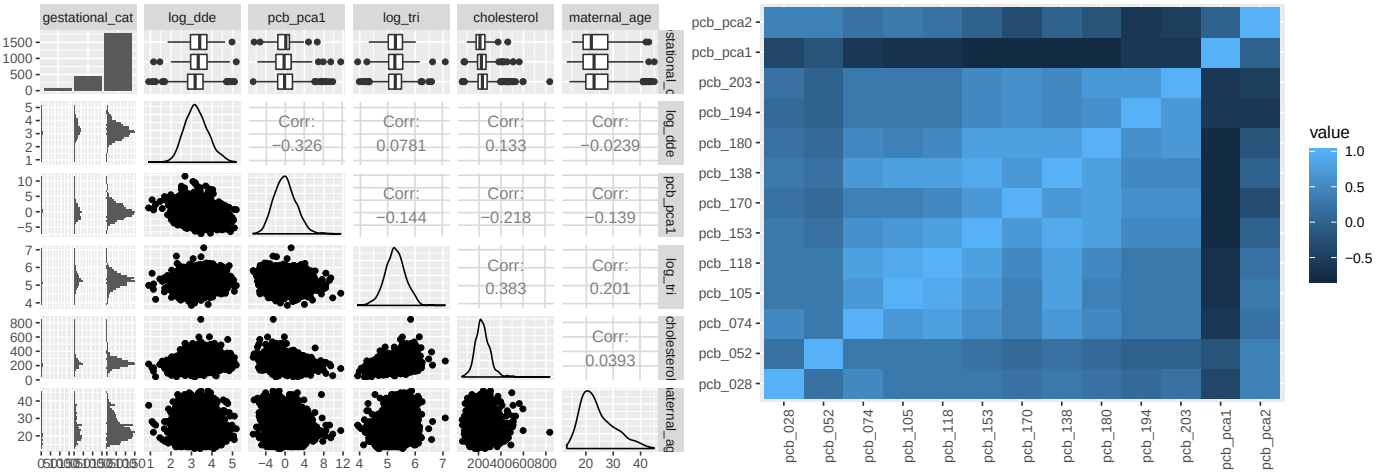

Figure 1: Correlations between predictors. First principal component of log(PCB) and log(DDE) are mildly negatively correlated (-0.3), and triglycerides and cholesterol are mildly positively correlated (0.3) (left). The PCB variables are all highly correlated (right).

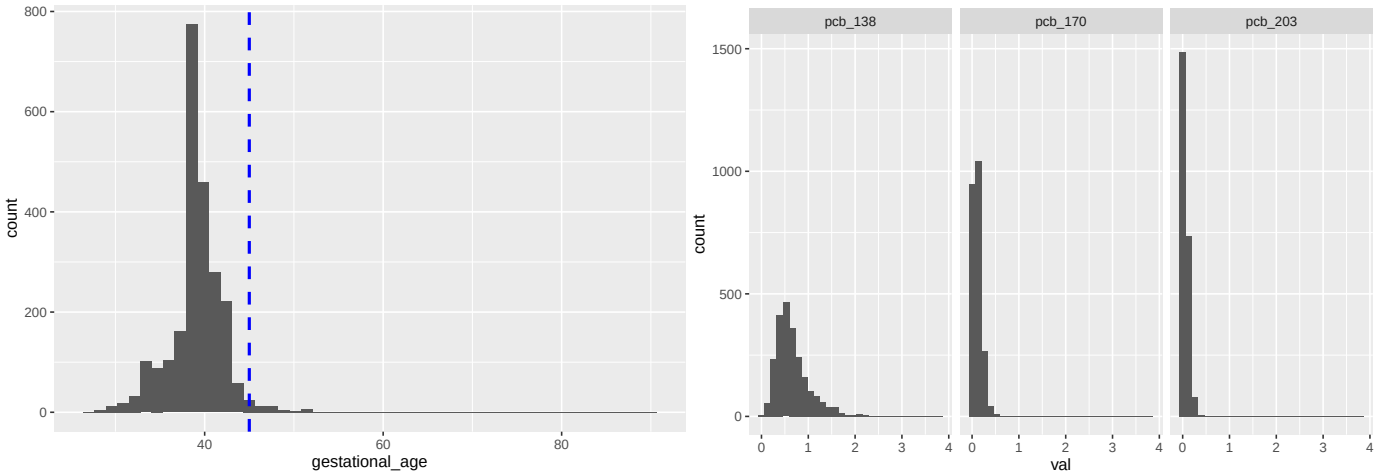

Figure 2: Gestational age ranges from 27 to 90 weeks (left). Distributions of some PCBs measurements

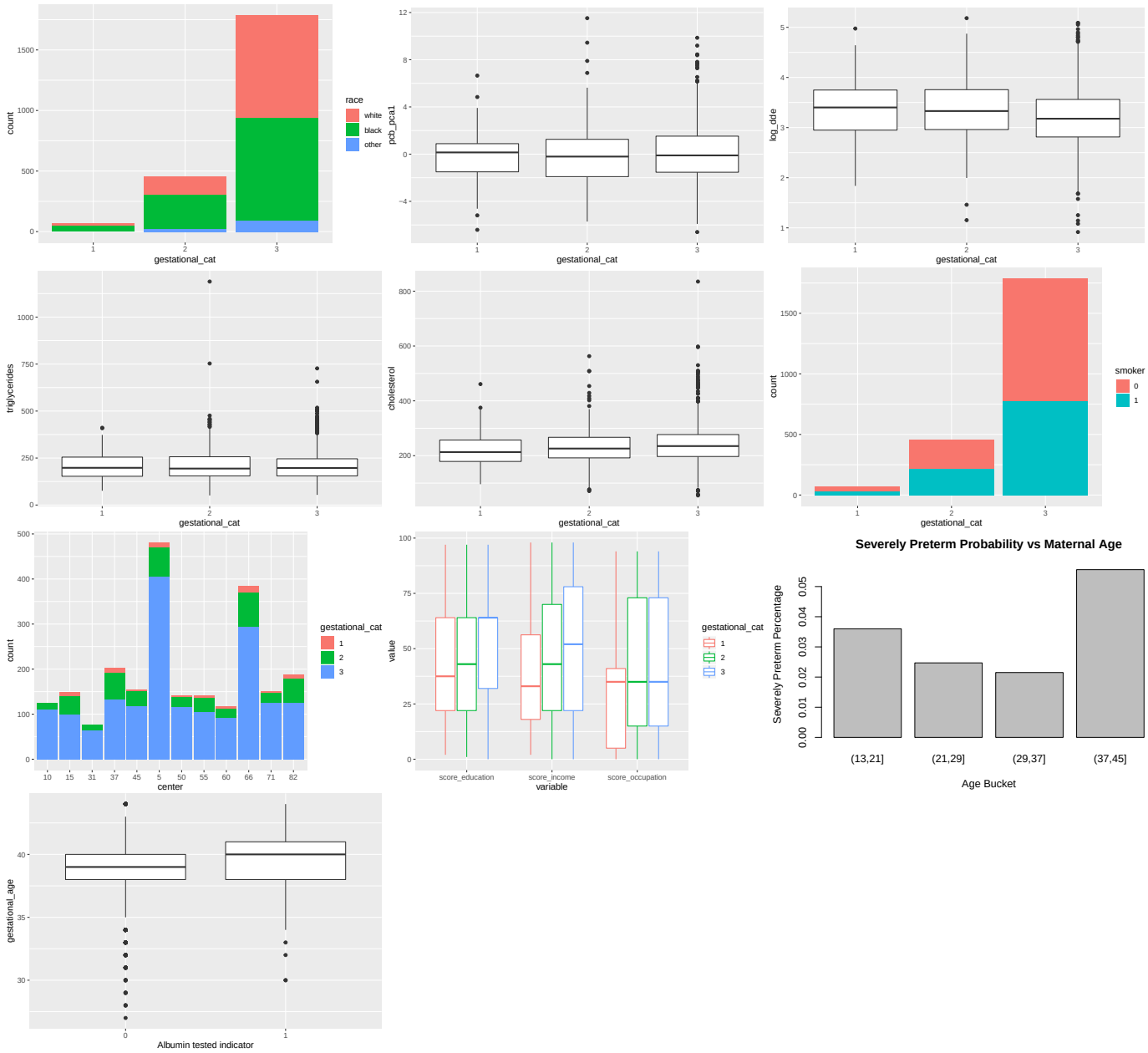

Figure 3: Bivariate plots of gestational categeries (1: serverely-premature, 2: premature, 3: normal-late) against predictors

3.2 Main Results

We perform F-tests to determine the appropriate structure and complexity of our model.

- Indication (p=0.39) that the first principle component of the pcb_* values is sufficient.
- Indication (p=0.74) against including Score Variables (post imputation)
- Indication (p=0.21) against including Center interactions with DDE/PCB
- Indication (p=0) for including Center as variable
- Indication (p=0.46) against PCB-DDE interaction effect
- Indication (p=0.15) (weakly) against Triglyceride interaction with PCE/DDE
- Indication (p=0.02) for inclusion of quadratic term in maternal age
- Strong indication that the indicator of testing for Albumin is associated with longer gestational period on the margin.

Taking into account these results, we express our final model form below in R-standard notation:

$$\begin{aligned} GestCategory \sim Center + log(dde) + logPCB1 \quad (1)\\ + log(trigl.) + Poly(MaternalAge, 2) \quad (2)\\ + Smoking + log(Cholest.) + AlbuminTested \quad (3)\\ + race \quad (4) \end{aligned}$$

Testing inclusion of both DDE and PCB vs. control model gives p=0.005. This indicates that DDE and PCB improve the explanatory power of pre-term birth risk model. Predictive check for the model is seen in Figure **??** and marginal fit showed in Figure 7. Marginally in the full model each is at the edge of significance (at the 5% level):

Table 1: Coefficient Estimates for Target Chemicals

	Coef Est	2.5 %	97.5 %
log(dde)	-0.183	-0.388	0.022
logPCB1	-0.056	-0.113	0.000

Table 2: Multiplier of odds ratio for 10 percent increase in exposure

	Coef Est	2.5 %	97.5 %
DDE	0.982	0.962	1.002
PCB1	0.994	0.989	1.000

These estimates suggests evidence for an association between exposure to DDE and PCBs and an increase in risk of premature birth. We can interprete these coefficents as follows:

- For 10% increase in DDE exposure, the odds of having more normal gestational age decreases by 1.8%.
- For 10% increase in the first principal components of PCBs, the odds of having more normal gestational age decreases by 0.6%. See Figure 4 below for a pictoral representation of fitted dose response probabilities across gestational term severities.

3.3 Sensitivity Analysis

We refit the above model for different cutoffs defining severely-premature, premature, and normal-late birth categories and plot the 95% CI's of the significant predictors in Figure 5. The plot shows that the CIs are relatively robust to reasonable change in cutoffs of the response categories.

4 Discussion

A key advantage of this approach is that its interpretation is naturally suited to the problem at hand. Our goal is to assess risks of pre-term delivery resulting from increased exposure to DDE and PCBs. An ordinal logistic regression model provides probability of different categories of preterm birth and their natural order. One disadvantage of this model is that it cannot capture different effects of the exposure on different severity of preterm birth, if such relationship exists. Another disadvantage is the necessity of cut points to define preterm birth categories. In one sense, the existence of certain natural cut points in fetal development, as provided by various reputable organizations as WHO, may make this type of categorization seem sensible. On the other hand, categorizing continuous data will always possess a degree of arbitrarity, and it may risk masking different dynamics contained within categories or across categories. Alternative modeling frameworks that address these limitations include quantile regression, generalized additive model, B-spline or density regression.

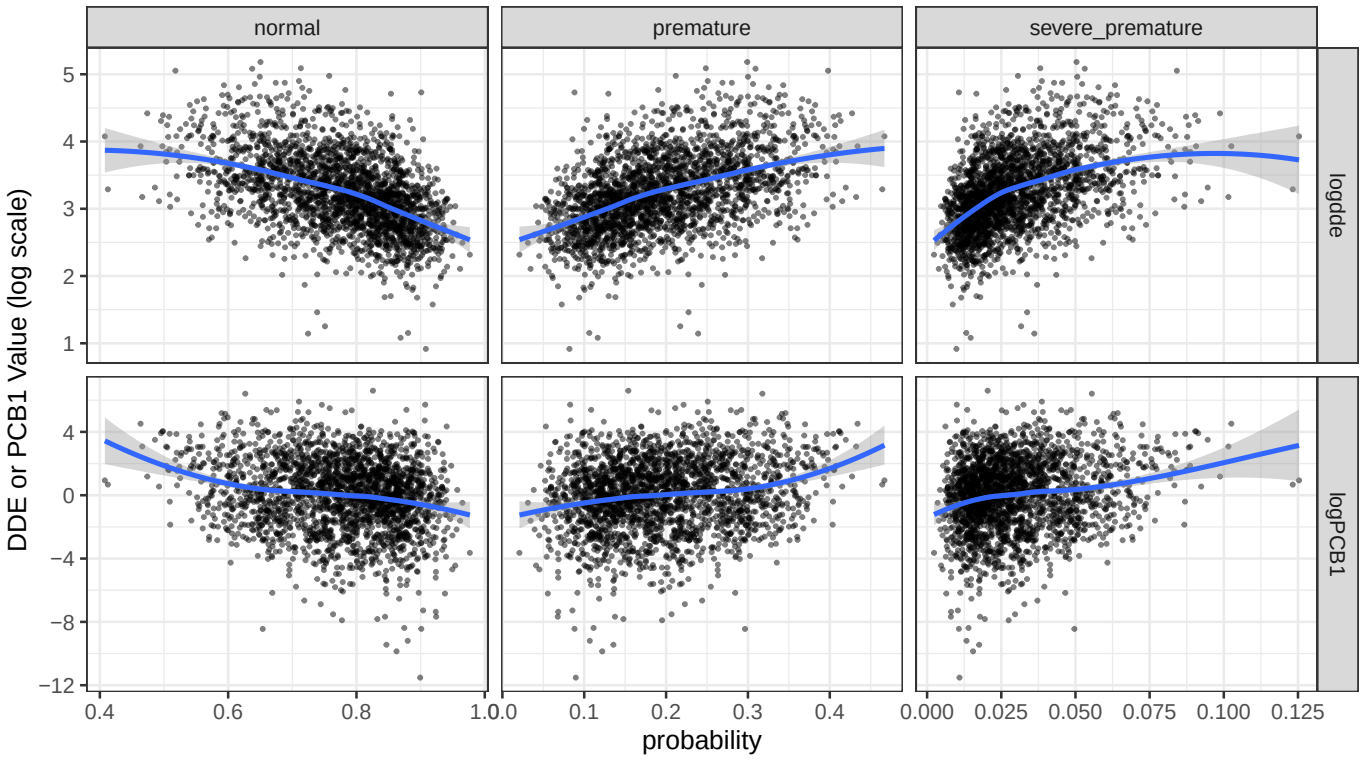

Figure 4: Predicted Dose Responses as Probabilities for Log DDE and Log PCB1

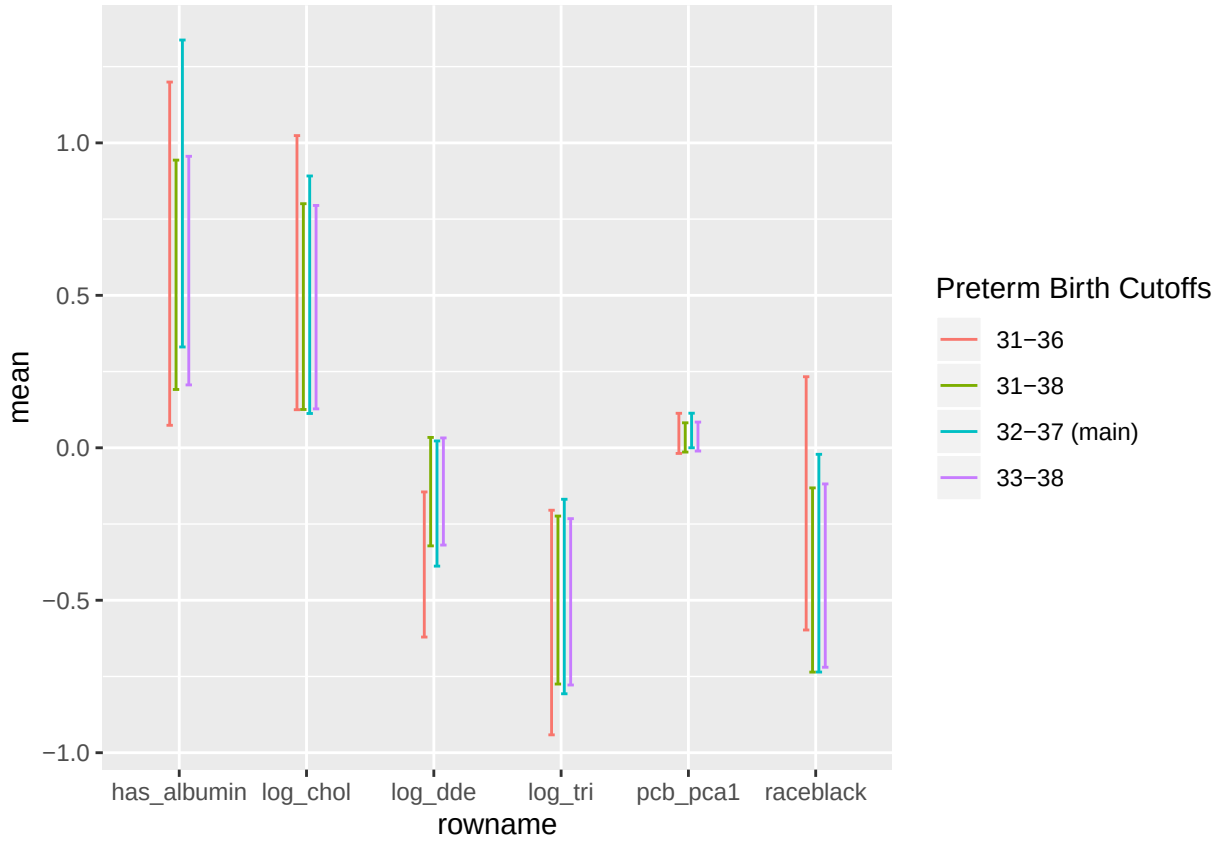

Figure 5: 95% CI of significant predictors for different cutoffs defining preterm birth categories

We find that the first principle component is sufficient. As the weights in the first principle component are all of the same sign and of broadly similar magnitude, we could have perhaps simply taken the sum of the PCBs predictors as an alternative methodology. This would improve the interpretability of the summary predictor.

5 Appendix

5.1 Final Model Summary Output

```
## Call:
## polr(formula = gestord ~ center + log(triglycerides) + log(dde) +
##     logPCB1 + poly(maternal_age, 2) + smoking_status + log(cholesterol) +
##     albuminTested + race, data = Long, Hess = T)
##
## Coefficients:
##                             Value Std. Error t value
## center15                -1.20741    0.38795 -3.1123
## center31                -0.17985    0.45518 -0.3951
## center37                -1.16104    0.35438 -3.2763
## center45                -0.53159    0.37910 -1.4022
## center5                 -0.41717    0.32132 -1.2983
## center50                -0.73297    0.37338 -1.9631
## center55                -0.85155    0.41156 -2.0691
## center60                -0.86261    0.38700 -2.2290
## center66                -0.55967    0.35415 -1.5803
## center71                -0.45049    0.36936 -1.2197
## center82                -1.10342    0.38127 -2.8941
## log(triglycerides)      -0.48705    0.16268 -2.9939
## log(dde)                -0.18279    0.10469 -1.7461
## logPCB1                 -0.05637    0.02889 -1.9512
## poly(maternal_age, 2)1  0.87514    2.47401  0.3537
## poly(maternal_age, 2)2 -5.41672    2.36687 -2.2886
## smoking_status          -0.16647    0.10515 -1.5831
## log(cholesterol)         0.50306    0.19860  2.5331
## albuminTested            0.80710    0.25551  3.1589
## raceblack               -0.37777    0.18207 -2.0749
## raceother               -0.38563    0.31451 -1.2261
##
## Intercepts:
##                   Value   Std. Error t value
## (0,32]|(32,37]   -4.9604  1.2205    -4.0643
## (32,37]|(37,45] -2.6448  1.2146    -2.1774
##
## Residual Deviance: 2752.277
## AIC: 2798.277
```

5.2 ANOVA Tests

Table 3: PCA Anova

Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
2289	2757.368		NA	NA	NA
2279	2746.764	1 vs 2	10	10.60398	0.3891918

Table 4: Center Interaction Anova

Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
2289	2757.368		NA	NA	NA
2267	2730.376	1 vs 2	22	26.99144	0.2115519

Table 5: Center Inclusion Anova

Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
2300	2791.971		NA	NA	NA
2289	2757.368	1 vs 2	11	34.6034	0.000288

Table 6: DDE-PCB1 Interaction Anova

Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
2289	2757.368		NA	NA	NA
2288	2756.817	1 vs 2	1	0.5507864	0.4579966

Table 7: Triglyceride-DDE/PCB Interaction Anova

Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
2289	2757.368		NA	NA	NA
2287	2753.636	1 vs 2	2	3.731565	0.154775

Table 8: Quadratic Maternal Age Anova

Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
2289	2757.368		NA	NA	NA
2288	2752.277	1 vs 2	1	5.090667	0.0240549

Table 9: DDE/PCB vs Control Anova

Resid. df	Resid. Dev	Test	Df	LR stat.	Pr(Chi)
2290	2762.890		NA	NA	NA
2288	2752.277	1 vs 2	2	10.61292	0.0049594

5.3 Proportional Odds Assumption

Below is the output for the Brant test mentioned above:

Table 10: Brant test for proportional odds

	X2	df	probability
Omnibus	10.0183083	20	0.9678387
raceblack	2.2812712	1	0.1309445
raceother	4.3867585	1	0.0362191
log_dde	0.6321191	1	0.4265791
pcb_pca1	0.7089018	1	0.3998086
has_albumin	0.2879753	1	0.5915209
maternal_age	0.0851589	1	0.7704242
log_tri	0.0230199	1	0.8794054
log_chol	0.3232999	1	0.5696312
smoking_status	0.2286195	1	0.6325494
center15	0.0002433	1	0.9875544
center31	0.0000005	1	0.9994108
center37	0.0002319	1	0.9878491
center45	0.0002223	1	0.9881053
center5	0.0002567	1	0.9872177
center50	0.0002384	1	0.9876820
center55	0.0002245	1	0.9880466
center60	0.0002629	1	0.9870633
center66	0.0002366	1	0.9877270
center71	0.0002437	1	0.9875445
center82	0.0002316	1	0.9878580

Note that we have a significant p-value for Race:Other; otherwise there are robust results indicating that the proportional odds assumption is not unreasonable. We note that because the Race:Other category is quite small as a percentage of the whole data set, it is likely to be acceptable to proceed with this modeling procedure. On a more general note, the proportional odds assumption is evident in the following computation:

$$\frac{odds(y_i \leq j)}{odds(y_k \leq j)} = \frac{e^{\alpha_j - \beta^T x_i}}{e^{\alpha_j - \beta^T x_k}} = e^{\beta^T (x_k - x_i)}$$

Note how in the above this ratio is a constant over the range of j. As is fairly evident, this is a rather strict requirement to place on one's data.

5.4 Ordinal Logistic Model

We perform some short calculations to demonstrate how one solve for probabilities of individual ordinal values from the modeled output of the model. We have:

$$\begin{aligned}
logit(y_i \leq j) &= \alpha_j - \beta^T x_i \Rightarrow \\
P(y_i \leq j) &= \frac{e^{\alpha_j - \beta^T x_i}}{1 + e^{\alpha_j - \beta^T x_i}} \\
P(y_i = j) &= P(y_i \leq j) - P(y_i \leq j-1) = \frac{e^{\alpha_j - \beta^T x_i}}{1 + e^{\alpha_j - \beta^T x_i}} - \frac{e^{\alpha_{j-1} - \beta^T x_i}}{1 + e^{\alpha_{j-1} - \beta^T x_i}} \\
&= \frac{e^{\alpha_j - \beta^T x_i} - e^{\alpha_{j-1} - \beta^T x_i}}{(1 + e^{\alpha_j - \beta^T x_i})(1 + e^{\alpha_{j-1} - \beta^T x_i})}
\end{aligned}$$

5.5 Model Checking

See Figure 6 and Figure 7

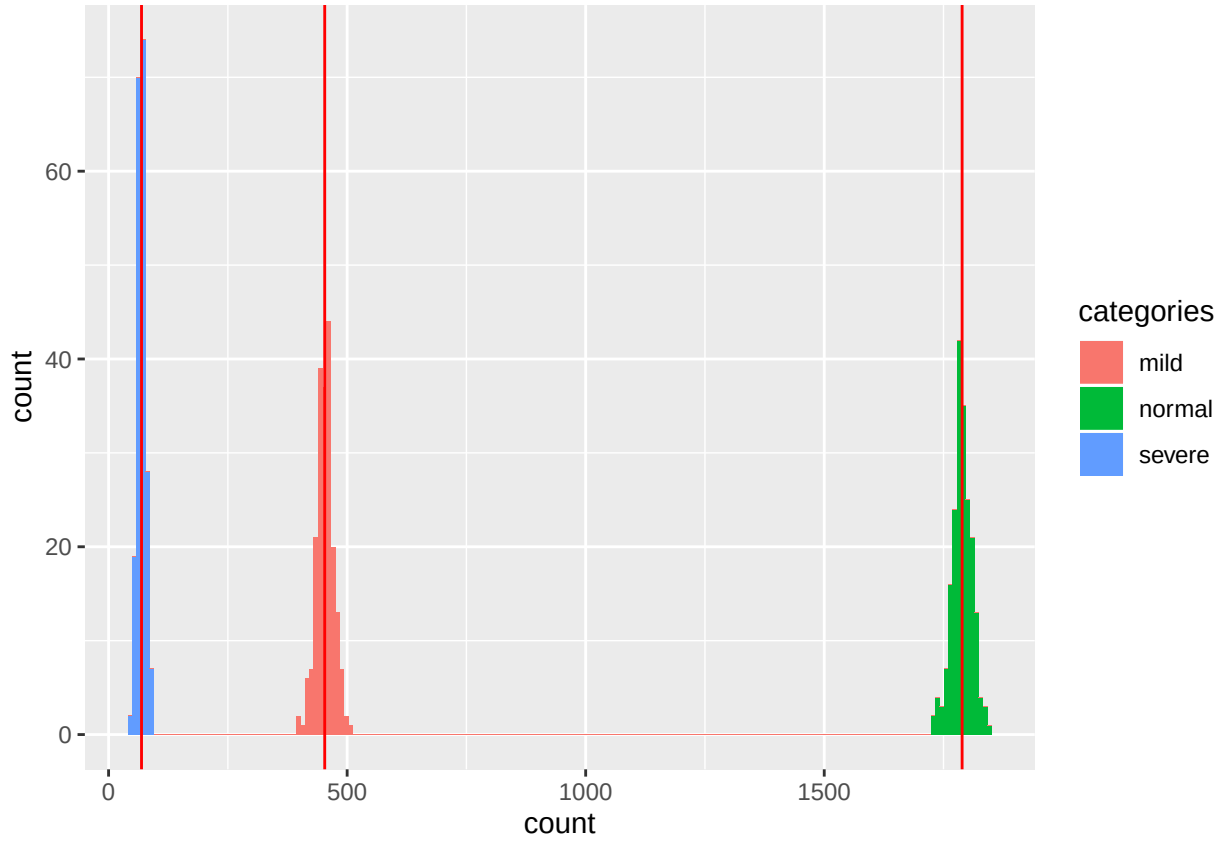

Figure 6: Predictive check for distribution of count of each preterm birth category

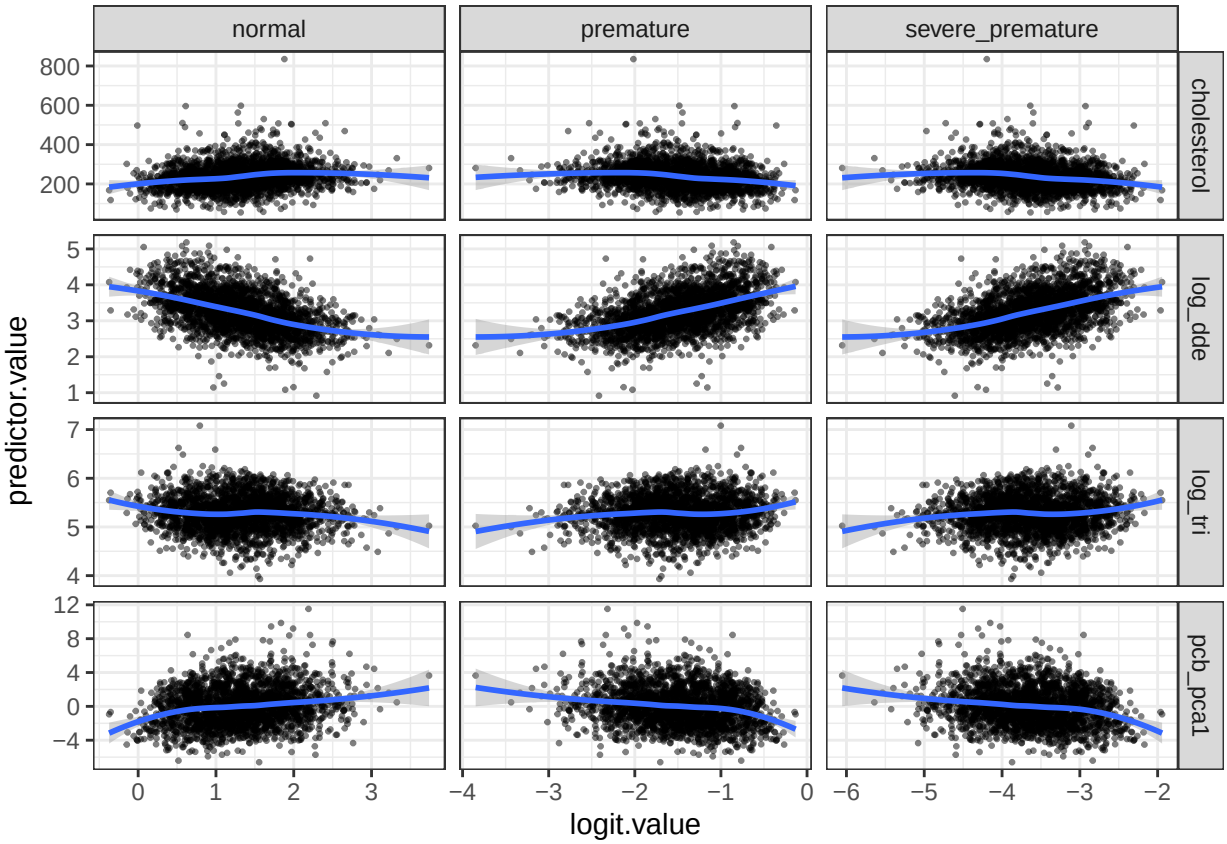

Figure 7: Visualizing predicted logit against covariates

6 Citations

Brant, Rollin. 1990. "Assessing Proportionality in the Proportional Odds Model for Ordinal Logistic Regression." *Biometrics* 46 (December). International Biometric Society: 1171–8. https://www.jstor.org/stable/pdf/2532457.pdf.

Cohn, Barbara, Piera Cirillo, Robert Sholtz, Assiamira Ferrara, June-Soo Park, and Pamela Schwingl. 2011. "Polychlorinated Biphenyl (PCB) Exposure in Mothers and Time to Pregnancy in Daughters." *Reproductive Toxicology* 31 (April). Elsevier: 290–96. https://doi.org/https://doi.org/10.1016/j.reprotox.2011.01.004.

Cohn, Barbara, Piera Cirillo, Mary wolff, Pamela Schwingl, Richard Cohen, Robert Sholtz, Assiamira Ferrara, Roberta Christianson, Barbara van den Berg, and Pentti Siiteri. 2003. "DDT and DDE Exposure in Mothers and Time to Pregnancy in Daughter." *The Lancet* 361 (June). Elsevier: 2205–6. https://doi.org/https://doi.org/10.1016/S0140-6736(03)13776-2.

McCullagh, Peter. 1980. "Models for Ordinal Data." *Journal of the Royal Statistical Society* 42 (February). Wiley: 109–42. https://www.jstor.org/stable/2984952.

"Preterm Birth." 2018. World Health Organization. February 2018. https://www.who.int/news-room/fact-sheets/detail/preterm-birth.