Data Prep

Raphal Morsomme

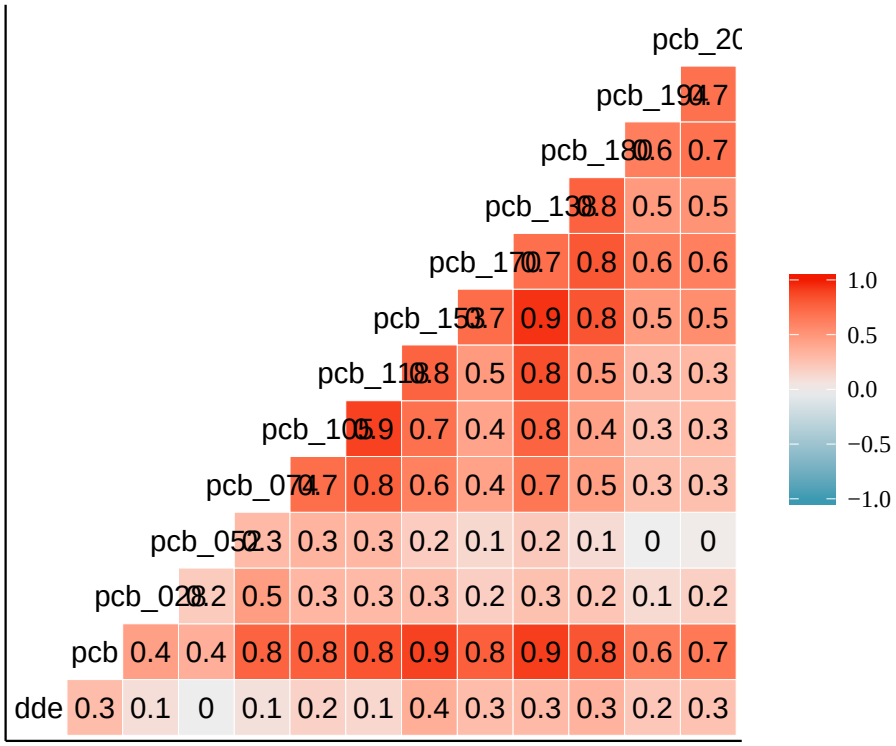
January 16, 2020

Data Preparation

- drop albumin
- remove female with length of gestation superior to 45 weeks and dichotomize the variable
- aggregate scaled PCB's (could do PCA)
- total fat
- exposure (quantity of chemical in environment)

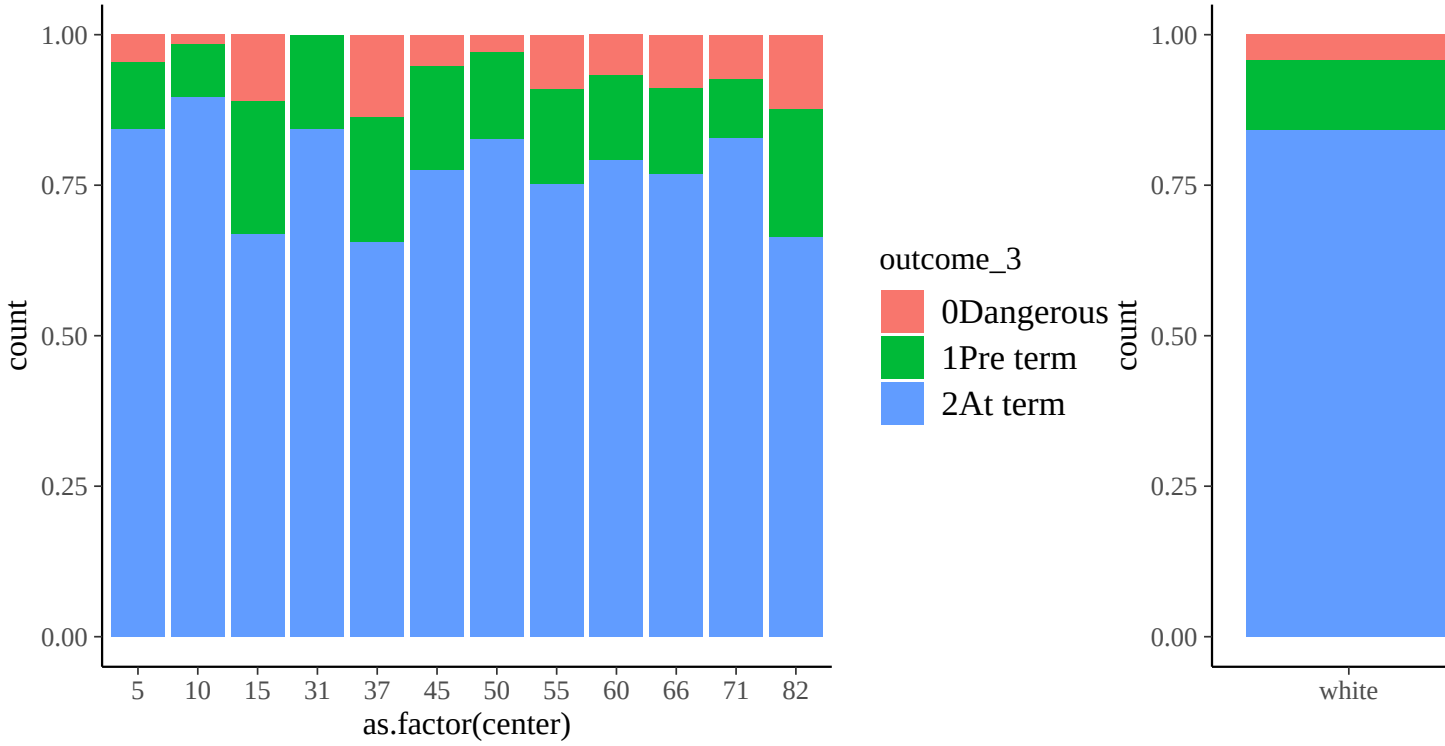
EDA

- justify aggregation of pcb: they are all correlated, tell the same story, reduce number of predictor (more stable parameter estimation).



side-by-side boxplots: - dde & pcb per outcome_3 - fat per race

side-by-side barplot - gestation per center - gestation per race



Model

Model Building

The cumulative logistics model(ordinal regression model) is fitted. The model is paramterized as:

$$logit(P(Y \leq j)) = \beta_{j0} - \eta_1 x_1 - \cdots - \eta_p x_p$$

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = outcome_3 ~ dde_env + pcb_env + race + maternal_age +
##     score_occupation + center + score_income + score_education,
##     data = d, method = c("logistic"))
##
## Coefficients:
##                       Value Std. Error    t value
## dde_env          -3.456e+00   1.717978 -2.012e+00
## pcb_env          -1.558e+02   0.013721 -1.135e+04
## raceblack        -2.767e-01   0.174361 -1.587e+00
## raceother        -3.276e-01   0.314112 -1.043e+00
## maternal_age      2.588e-03   0.008382  3.088e-01
## score_occupation  1.504e-03   0.002309  6.516e-01
## center10          3.917e-01   0.322880  1.213e+00
## center15         -6.593e-01   0.267781 -2.462e+00
## center31          2.648e-01   0.360149  7.353e-01
## center37         -7.167e-01   0.227948 -3.144e+00
## center45          1.172e-02   0.268673  4.361e-02
## center50         -2.614e-01   0.253779 -1.030e+00
## center55         -2.991e-01   0.301812 -9.909e-01
```

```
## center60         -3.433e-01   0.263377 -1.303e+00
## center66         -1.410e-01   0.225579 -6.250e-01
## center71         -7.597e-02   0.254482 -2.985e-01
## center82         -6.361e-01   0.255114 -2.493e+00
## score_income      1.367e-03   0.002291  5.966e-01
## score_education   2.543e-03   0.002554  9.958e-01
##
## Intercepts:
##                       Value       Std. Error  t value
## 0Dangerous|1Pre term     -2.8313      0.3246     -8.7236
## 1Pre term|2At term       -1.4922      0.3178     -4.6957
##
## Residual Deviance: 3042.324
## AIC: 3084.324
```

Besides, the cumulative Logit model w/o env is fitted.

```
##
## Re-fitting to get Hessian
```

```
## Call:
## polr(formula = outcome_3 ~ dde + pcb + race + maternal_age +
##     score_occupation + center + score_income + score_education,
##     data = d, method = c("logistic"))
##
## Coefficients:
##                       Value Std. Error  t value
## dde                -0.005604   0.002611 -2.14605
## pcb                -0.218864   0.080988 -2.70242
## raceblack          -0.287568   0.173128 -1.66101
## raceother          -0.339716   0.314588 -1.07987
## maternal_age        0.003227   0.008521  0.37868
## score_occupation    0.001508   0.002312  0.65238
## center10            0.399749   0.323621  1.23524
## center15           -0.664752   0.274715 -2.41979
## center31            0.286140   0.361732  0.79103
## center37           -0.732810   0.228086 -3.21287
## center45            0.015327   0.269419  0.05689
## center50           -0.281451   0.257056 -1.09490
## center55           -0.303012   0.305769 -0.99098
## center60           -0.350859   0.268494 -1.30677
## center66           -0.145896   0.226285 -0.64475
## center71           -0.087744   0.255039 -0.34404
## center82           -0.651184   0.261411 -2.49103
## score_income        0.001532   0.002294  0.66789
## score_education     0.002388   0.002558  0.93351
##
## Intercepts:
##                       Value   Std. Error t value
## 0Dangerous|1Pre term -2.8616  0.3299    -8.6743
## 1Pre term|2At term   -1.5214  0.3234    -4.7046
##
## Residual Deviance: 3039.599
## AIC: 3081.599
```

Variable Selection

-Variable Selection (AIC) Stepwise regression by AIC is used to select variables. The final model we choose is "outcome_3 ~ dde_env + pcb_env + center + score_education".

```
## Start:  AIC=3084.32
## outcome_3 ~ dde_env + pcb_env + race + maternal_age + score_occupation +
##     center + score_income + score_education
##
##                     Df    AIC
## - maternal_age       1 3082.4
## - score_income       1 3082.7
## - score_occupation   1 3082.8
## - race               2 3083.0
## - score_education    1 3083.3
## <none>                 3084.3
## - dde_env            1 3086.0
## - pcb_env            1 3090.1
## - center            11 3091.5
##
## Step:  AIC=3082.42
## outcome_3 ~ dde_env + pcb_env + race + score_occupation + center +
##     score_income + score_education
##
##                     Df    AIC
## - score_income       1 3080.8
## - score_occupation   1 3080.8
## - race               2 3081.1
## - score_education    1 3081.3
## <none>                 3082.4
## - dde_env            1 3084.2
## - pcb_env            1 3088.1
## - center            11 3089.8
##
## Step:  AIC=3080.82
## outcome_3 ~ dde_env + pcb_env + race + score_occupation + center +
##     score_education
##
##                     Df    AIC
## - score_occupation   1 3079.5
## - race               2 3079.6
## - score_education    1 3080.0
## <none>                 3080.8
## - dde_env            1 3082.6
## - pcb_env            1 3086.5
## - center            11 3088.6
##
## Step:  AIC=3079.49
## outcome_3 ~ dde_env + pcb_env + race + center + score_education
##
##                    Df    AIC
## - race              2 3078.6
## <none>                3079.5
## - score_education   1 3079.8
## - dde_env           1 3081.3
```

```
## - pcb_env          1 3084.9
## - center          11 3087.8
##
## Step:  AIC=3078.57
## outcome_3 ~ dde_env + pcb_env + center + score_education
##
##                   Df    AIC
## <none>               3078.6
## - score_education  1 3078.8
## - dde_env          1 3081.9
## - pcb_env          1 3084.5
## - center          11 3107.4
```

Model Diagnostics

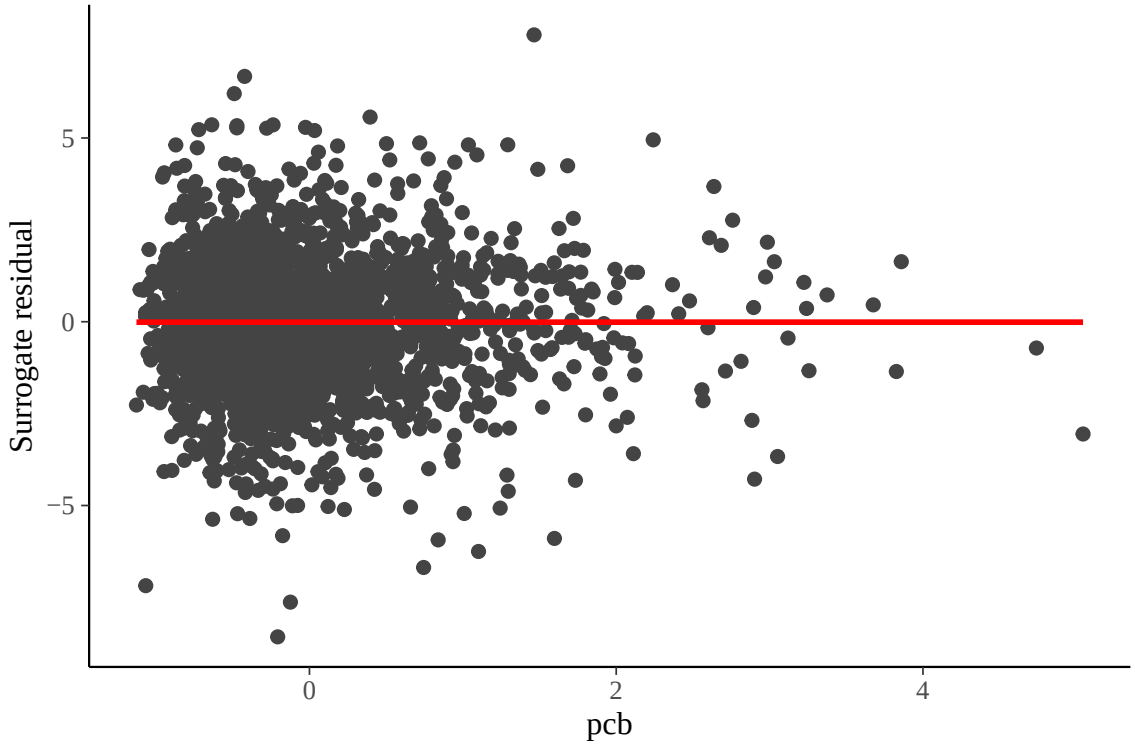
-Model Checking (i) residuals Surrogate residuals are used to check if the model assumption is correct, as suggested by (Liu and Zhang, 2017). If the model assumptions are correct, then the surrogate residuals R_S will have three properties:

- $E(R_S|X) = 0$
- $Var(R_S|X) = c$, the conditional variance of R_S is constant
- The emiprical distribution of R_S resembles an explicit distribution that is related to the link function $G^{-1}(\cdot)$. Specifically, $R_S \sim G(c + \int udG(u))$.

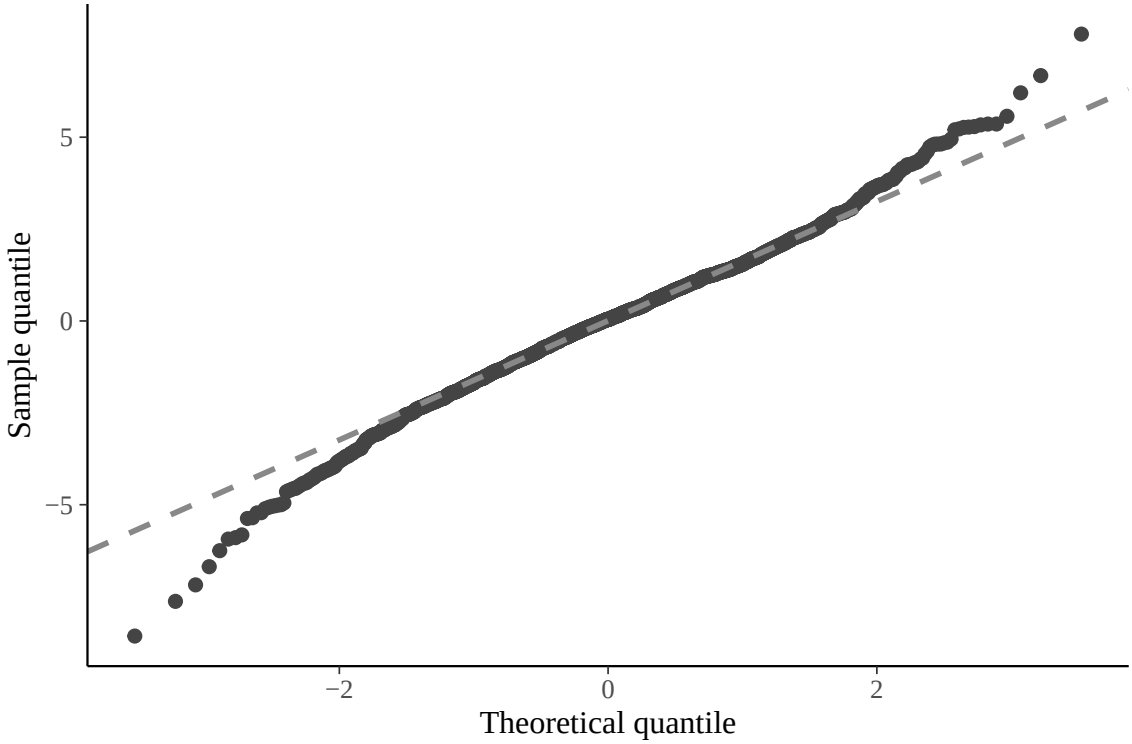
The surrogate residuals vs covariate plot is used to check the first and second properties. The QQ-plot is used to check the third property.

-Residual vs Pcb The surrogate residuals are scatter around 0 evenly. This plot indicates that $E(R_S|X) = 0$ and $Var(R_S|X) = c$ are basically satisfied.

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```



-QQ-plot Both the theoretical quantiles and sample quantiles match quite well. Our model assumptions are ap-



propriate.
(ii) (probabilistic) predictive model checking

Try to fit a Bayesian model

A Bayesian model is fitted. 5 chains with 3000 iterations each are ran. All Rhats are closed to 1 and effective sample sizes exceed 8000, so there is strong evidence that the chains converge.

```
##
## SAMPLING FOR MODEL 'polr' NOW (CHAIN 1).
## Chain 1:
## Chain 1: Gradient evaluation took 0.00052 seconds
## Chain 1: 1000 transitions using 10 leapfrog steps per transition would take 5.2 seconds.
## Chain 1: Adjust your expectations accordingly!
## Chain 1:
## Chain 1:
## Chain 1: Iteration:    1 / 3000 [  0%]  (Warmup)
## Chain 1: Iteration:  300 / 3000 [ 10%]  (Warmup)
## Chain 1: Iteration:  600 / 3000 [ 20%]  (Warmup)
## Chain 1: Iteration:  900 / 3000 [ 30%]  (Warmup)
## Chain 1: Iteration: 1200 / 3000 [ 40%]  (Warmup)
## Chain 1: Iteration: 1500 / 3000 [ 50%]  (Warmup)
## Chain 1: Iteration: 1501 / 3000 [ 50%]  (Sampling)
## Chain 1: Iteration: 1800 / 3000 [ 60%]  (Sampling)
## Chain 1: Iteration: 2100 / 3000 [ 70%]  (Sampling)
## Chain 1: Iteration: 2400 / 3000 [ 80%]  (Sampling)
## Chain 1: Iteration: 2700 / 3000 [ 90%]  (Sampling)
## Chain 1: Iteration: 3000 / 3000 [100%]  (Sampling)
## Chain 1:
## Chain 1:  Elapsed Time: 6.07811 seconds (Warm-up)
## Chain 1:                8.1301 seconds (Sampling)
```

```
## Chain 1:                14.2082 seconds (Total)
## Chain 1:
##
## SAMPLING FOR MODEL 'polr' NOW (CHAIN 2).
## Chain 2:
## Chain 2: Gradient evaluation took 0.000247 seconds
## Chain 2: 1000 transitions using 10 leapfrog steps per transition would take 2.47 seconds.
## Chain 2: Adjust your expectations accordingly!
## Chain 2:
## Chain 2:
## Chain 2: Iteration:    1 / 3000 [  0%]  (Warmup)
## Chain 2: Iteration:  300 / 3000 [ 10%]  (Warmup)
## Chain 2: Iteration:  600 / 3000 [ 20%]  (Warmup)
## Chain 2: Iteration:  900 / 3000 [ 30%]  (Warmup)
## Chain 2: Iteration: 1200 / 3000 [ 40%]  (Warmup)
## Chain 2: Iteration: 1500 / 3000 [ 50%]  (Warmup)
## Chain 2: Iteration: 1501 / 3000 [ 50%]  (Sampling)
## Chain 2: Iteration: 1800 / 3000 [ 60%]  (Sampling)
## Chain 2: Iteration: 2100 / 3000 [ 70%]  (Sampling)
## Chain 2: Iteration: 2400 / 3000 [ 80%]  (Sampling)
## Chain 2: Iteration: 2700 / 3000 [ 90%]  (Sampling)
## Chain 2: Iteration: 3000 / 3000 [100%]  (Sampling)
## Chain 2:
## Chain 2:  Elapsed Time: 6.02559 seconds (Warm-up)
## Chain 2:                8.31802 seconds (Sampling)
## Chain 2:                14.3436 seconds (Total)
## Chain 2:
##
## SAMPLING FOR MODEL 'polr' NOW (CHAIN 3).
## Chain 3:
## Chain 3: Gradient evaluation took 0.000247 seconds
## Chain 3: 1000 transitions using 10 leapfrog steps per transition would take 2.47 seconds.
## Chain 3: Adjust your expectations accordingly!
## Chain 3:
## Chain 3:
## Chain 3: Iteration:    1 / 3000 [  0%]  (Warmup)
## Chain 3: Iteration:  300 / 3000 [ 10%]  (Warmup)
## Chain 3: Iteration:  600 / 3000 [ 20%]  (Warmup)
## Chain 3: Iteration:  900 / 3000 [ 30%]  (Warmup)
## Chain 3: Iteration: 1200 / 3000 [ 40%]  (Warmup)
## Chain 3: Iteration: 1500 / 3000 [ 50%]  (Warmup)
## Chain 3: Iteration: 1501 / 3000 [ 50%]  (Sampling)
## Chain 3: Iteration: 1800 / 3000 [ 60%]  (Sampling)
## Chain 3: Iteration: 2100 / 3000 [ 70%]  (Sampling)
## Chain 3: Iteration: 2400 / 3000 [ 80%]  (Sampling)
## Chain 3: Iteration: 2700 / 3000 [ 90%]  (Sampling)
## Chain 3: Iteration: 3000 / 3000 [100%]  (Sampling)
## Chain 3:
## Chain 3:  Elapsed Time: 6.1168 seconds (Warm-up)
## Chain 3:                7.30277 seconds (Sampling)
## Chain 3:                13.4196 seconds (Total)
## Chain 3:
##
## SAMPLING FOR MODEL 'polr' NOW (CHAIN 4).
```

```
## Chain 4:
## Chain 4: Gradient evaluation took 0.000243 seconds
## Chain 4: 1000 transitions using 10 leapfrog steps per transition would take 2.43 seconds.
## Chain 4: Adjust your expectations accordingly!
## Chain 4:
## Chain 4:
## Chain 4: Iteration:    1 / 3000 [  0%]  (Warmup)
## Chain 4: Iteration:  300 / 3000 [ 10%]  (Warmup)
## Chain 4: Iteration:  600 / 3000 [ 20%]  (Warmup)
## Chain 4: Iteration:  900 / 3000 [ 30%]  (Warmup)
## Chain 4: Iteration: 1200 / 3000 [ 40%]  (Warmup)
## Chain 4: Iteration: 1500 / 3000 [ 50%]  (Warmup)
## Chain 4: Iteration: 1501 / 3000 [ 50%]  (Sampling)
## Chain 4: Iteration: 1800 / 3000 [ 60%]  (Sampling)
## Chain 4: Iteration: 2100 / 3000 [ 70%]  (Sampling)
## Chain 4: Iteration: 2400 / 3000 [ 80%]  (Sampling)
## Chain 4: Iteration: 2700 / 3000 [ 90%]  (Sampling)
## Chain 4: Iteration: 3000 / 3000 [100%]  (Sampling)
## Chain 4:
## Chain 4:  Elapsed Time: 5.97915 seconds (Warm-up)
## Chain 4:                7.2899 seconds (Sampling)
## Chain 4:                13.269 seconds (Total)
## Chain 4:
##
## SAMPLING FOR MODEL 'polr' NOW (CHAIN 5).
## Chain 5:
## Chain 5: Gradient evaluation took 0.000266 seconds
## Chain 5: 1000 transitions using 10 leapfrog steps per transition would take 2.66 seconds.
## Chain 5: Adjust your expectations accordingly!
## Chain 5:
## Chain 5:
## Chain 5: Iteration:    1 / 3000 [  0%]  (Warmup)
## Chain 5: Iteration:  300 / 3000 [ 10%]  (Warmup)
## Chain 5: Iteration:  600 / 3000 [ 20%]  (Warmup)
## Chain 5: Iteration:  900 / 3000 [ 30%]  (Warmup)
## Chain 5: Iteration: 1200 / 3000 [ 40%]  (Warmup)
## Chain 5: Iteration: 1500 / 3000 [ 50%]  (Warmup)
## Chain 5: Iteration: 1501 / 3000 [ 50%]  (Sampling)
## Chain 5: Iteration: 1800 / 3000 [ 60%]  (Sampling)
## Chain 5: Iteration: 2100 / 3000 [ 70%]  (Sampling)
## Chain 5: Iteration: 2400 / 3000 [ 80%]  (Sampling)
## Chain 5: Iteration: 2700 / 3000 [ 90%]  (Sampling)
## Chain 5: Iteration: 3000 / 3000 [100%]  (Sampling)
## Chain 5:
## Chain 5:  Elapsed Time: 6.05559 seconds (Warm-up)
## Chain 5:                6.899 seconds (Sampling)
## Chain 5:                12.9546 seconds (Total)
## Chain 5:

##
## Model Info:
##
##  function:     stan_polr
##  family:       ordered [logistic]
```

```
##  formula:      outcome_3 ~ dde_env + pcb_env + score_education + center
##  algorithm:    sampling
##  priors:       see help('prior_summary')
##  sample:       7500 (posterior sample size)
##  observations: 2337
##
## Estimates:
##                          mean    sd      2.5%    25%     50%     75%
## dde_env                   -3.6     1.6    -6.8    -4.7    -3.6    -2.5
## pcb_env                 -133.3    50.0  -229.4  -166.6  -133.2   -99.8
## score_education            0.0     0.0     0.0     0.0     0.0     0.0
## center10                   0.3     0.3    -0.2     0.1     0.3     0.5
## center15                  -0.8     0.2    -1.3    -1.0    -0.8    -0.7
## center31                   0.1     0.3    -0.5    -0.1     0.1     0.3
## center37                  -0.8     0.2    -1.2    -0.9    -0.8    -0.7
## center45                  -0.2     0.2    -0.6    -0.3    -0.2     0.0
## center50                  -0.2     0.2    -0.6    -0.3    -0.2     0.0
## center55                  -0.5     0.2    -0.9    -0.6    -0.5    -0.3
## center60                  -0.4     0.2    -0.8    -0.5    -0.4    -0.2
## center66                  -0.3     0.2    -0.7    -0.4    -0.3    -0.2
## center71                  -0.1     0.2    -0.6    -0.3    -0.1     0.0
## center82                  -0.8     0.2    -1.2    -0.9    -0.8    -0.7
## 0Dangerous|1Pre term      -2.9     0.2    -3.3    -3.1    -2.9    -2.8
## 1Pre term|2At term        -1.6     0.2    -2.0    -1.7    -1.6    -1.5
## mean_PPD:0Dangerous        0.1     0.0     0.1     0.1     0.1     0.1
## mean_PPD:1Pre term         0.2     0.0     0.1     0.1     0.2     0.2
## mean_PPD:2At term          0.8     0.0     0.8     0.8     0.8     0.8
## log-posterior          -1544.7     3.9 -1553.3 -1547.2 -1544.4 -1541.9
##                         97.5%
## dde_env                  -0.3
## pcb_env                 -35.1
## score_education            0.0
## center10                   0.9
## center15                  -0.4
## center31                   0.7
## center37                  -0.5
## center45                   0.2
## center50                   0.3
## center55                   0.0
## center60                   0.1
## center66                   0.0
## center71                   0.3
## center82                  -0.4
## 0Dangerous|1Pre term      -2.6
## 1Pre term|2At term        -1.3
## mean_PPD:0Dangerous        0.1
## mean_PPD:1Pre term         0.2
## mean_PPD:2At term          0.8
## log-posterior          -1538.1
##
## Diagnostics:
##                       mcse Rhat n_eff
## dde_env               0.0  1.0  13262
## pcb_env               0.5  1.0   9780
```

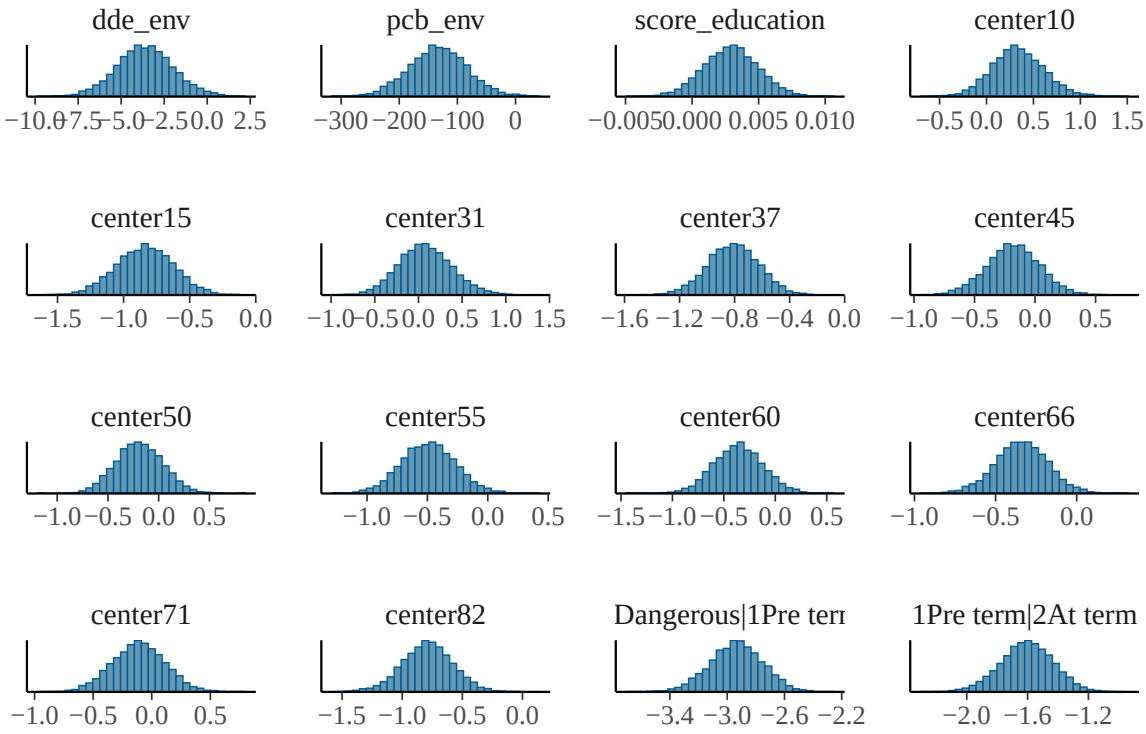
```
## score_education      0.0  1.0  12433
## center10             0.0  1.0   9247
## center15             0.0  1.0   9072
## center31             0.0  1.0  10116
## center37             0.0  1.0   9255
## center45             0.0  1.0  10222
## center50             0.0  1.0   9131
## center55             0.0  1.0   9827
## center60             0.0  1.0   9621
## center66             0.0  1.0  10111
## center71             0.0  1.0  10568
## center82             0.0  1.0  10598
## 0Dangerous|1Pre term 0.0  1.0  12049
## 1Pre term|2At term   0.0  1.0  11725
## mean_PPD:0Dangerous  0.0  1.0   7949
## mean_PPD:1Pre term   0.0  1.0   8379
## mean_PPD:2At term    0.0  1.0   8607
## log-posterior        0.1  1.0   1873
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude measure of effective sample
```

-Printing the 95% credible intervals

```
##                              2.5%         97.5%
## dde_env              -6.819856e+00  -0.291627457
## pcb_env              -2.293794e+02 -35.128117595
## score_education      -1.082122e-03   0.006988659
## center10             -1.889436e-01   0.894573351
## center15             -1.262709e+00  -0.425141827
## center31             -4.894636e-01   0.709124766
## center37             -1.178991e+00  -0.454840911
## center45             -6.157815e-01   0.247720371
## center50             -6.331921e-01   0.268160633
## center55             -9.280028e-01  -0.048990780
## center60             -8.280147e-01   0.114296851
## center66             -6.680694e-01  -0.017303912
## center71             -5.565522e-01   0.328154016
## center82             -1.210218e+00  -0.385507224
## 0Dangerous|1Pre term -3.302961e+00  -2.587067714
## 1Pre term|2At term   -1.950914e+00  -1.276271468
```

-Histrograms of posterior draws for each coefficient

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



Interpretation

-Try to summarize the result

Bayesian modelling