Guidelines to Solutions

The dataset birthwt.dta, available on the course website, is a subset of the data analyzed in Adams et al. (1997). The Adams et al. (1997) dataset consisted of the live births that occurred in the state of Georgia from 1980 to 1992. Our subset of the data contains information only on mothers who had five live births from 1980-1992 in Georgia. The variables in the dataset are as follows:

- mother: identifier for mother
- child: unique identifier for child
- birthwt: birthweight of the child (in grams)
- age: mothers age at time of birth to the child
- birthorder: child-level identifier of birth order for a given mother

Note: for multilevel modeling purposes we are interested in children 'nested' within mothers. The easiest way to operationalize this nesting structure of the data is to use 'mother' to index mothers and 'birthorder' to index children within a mother.

We are interested in using this dataset to assess whether mother's age and/or the duration of time between pregnancies are related to a child's birthweight.

Part I: EDA/Data preparation

1. **How many mothers are in this dataset? What is the distribution of the number of children per mother?**

```
. xtset mother birthorder
       panel variable:  mother (strongly balanced)
        time variable:  birthorder, 1 to 5
                delta:  1 unit

. xtdes

  mother:  80, 84, ..., 370377                             n =        878
birthorder:  1, 2, ..., 5                                  T =          5
           Delta(birthorder) = 1 unit
           Span(birthorder)  = 5 periods
           (mother*birthorder uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         5       5       5         5         5       5       5

     Freq.  Percent    Cum. |  Pattern
 ---------------------------+---------
      878    100.00  100.00 |  11111
 ---------------------------+---------
      878    100.00         |  XXXXX
```

There are 878 mothers in this dataset. Each mother has 5 children.

2. **What are the maximum and minimum ages at which a mother gives birth to a child in this dataset? What is the median age of mothers at the first recorded live birth?**

```
--------------------------------------------------------------------------------
age                                                                  Mother's Age
--------------------------------------------------------------------------------
                  type:  numeric (float)

                 range:  [12,42]                      units:  1
         unique values:  31                       missing .:  0/4390

                  mean:   21.6503
              std. dev:   4.63078

           percentiles:        10%       25%       50%       75%       90%
                                16        18        21        24        28
```

The maximum and minimum ages are 12 and 42.

```
. codebook age if birthorder==1

--------------------------------------------------------------------------------
age                                                                   Mother's Age
--------------------------------------------------------------------------------

                  type:  numeric (float)

                 range:  [12,35]                      units:  1
         unique values:  22                       missing .:  0/878

                  mean:   17.8713
              std. dev:   3.44172

           percentiles:        10%       25%       50%       75%       90%
                                15        16        17        19        23
```

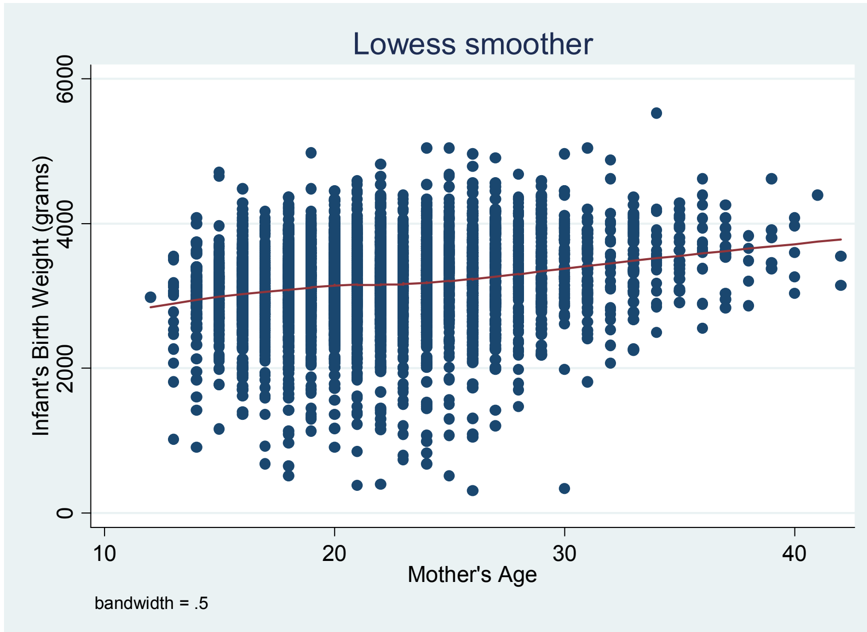
The median age of mothers at the first recorded live birth is 17 years.

3. **The original paper on this dataset, Adams et al. (1997) assessed the relationship between interpregnancy interval and infant birthweight. Create a variable called 'interval' that roughly represents the number of years between the current delivery and the previous delivery by subtracting the mother's age at a given delivery from the age at the prior delivery.**

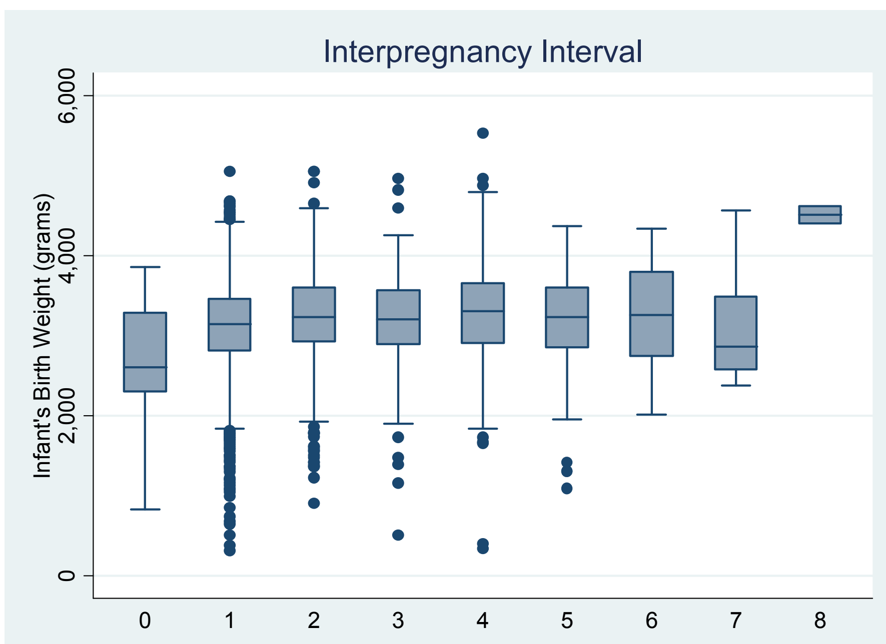
```
* generate a rough interpregnancy interval variable
. sort mother age

. by mother: gen agelag1 =age[_n-1]
(878 missing values generated)

. gen interval = age - agelag1
(878 missing values generated)
. label var interval "Interpregnancy Interval"
. drop agelag1
```

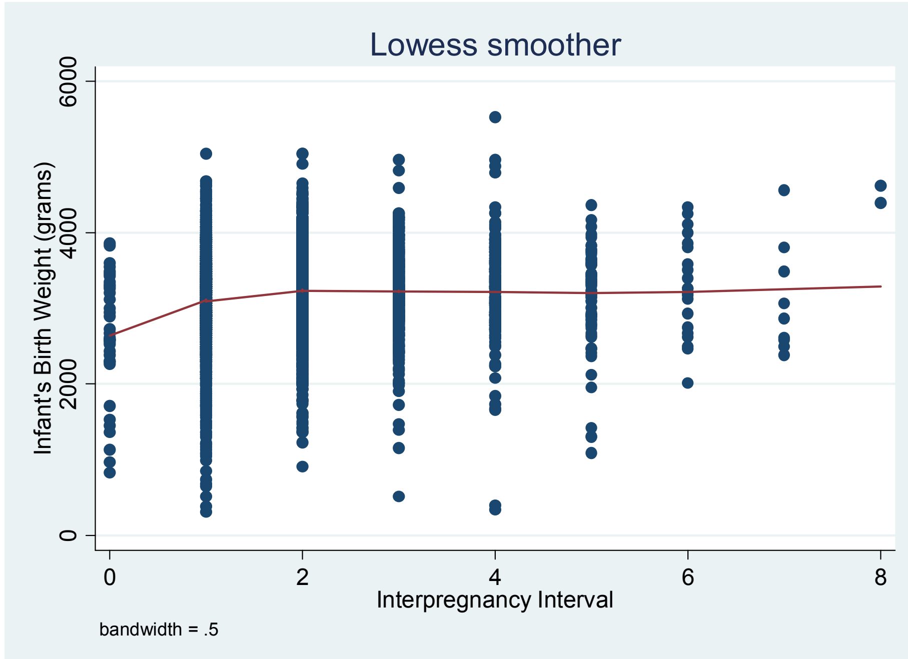
4. **Use graphs (scatterplots with lowess curves and/or boxplots) to explore the marginal relationship between birth weight and the following:**
    a. **Mother's age**
    b. **Interpregnancy interval**

    In 2-3 sentences describe the relations you see in the data (ignoring clustering).



Ignoring clustering due to mother, we observe that as mother's age increases, infant birthweight increases as well. The trend appears fairly linear.





Interpregnancy intervals of less than one year appear to be related to decreased infant birthweight. *For this exam, I'll assume the effect of interprenancy interval on birthweight is linear, but for a future analysis, we might use an indicator of short interval length.*

Part II: Variance components model with no covariates

5. **Write down the model that represents a linear regression of birthweight with an intercept but without any covariates. Account for the clustering of children(j) within mothers(i) by including a random intercept for mother. Interpret, in non-mathematical language, all of the parameters of your model: the intercept, the random intercept, the error and the variances of the random intercept and the error.**

$$y_{ij} \mid U_i = \beta_0 + U_i + \varepsilon_i$$

$$U_i \sim N(0, \tau^2)$$

$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

β_0 : average infant birthweight of an infant from a typical (U_i=0) mother.
U_i : difference between the average infant birthweight for mother(i) and the average infant birthweight of a typical (U_i=0) mother
$\beta_0 + U_i$: average infant birthweight for mother i
ε_{ij}: difference between the birthweight for child(ij) and the average infant birthweight of mother(i)
σ^2 : variance of the errors ε_{ij}, a measure of the dispersion of infant birthweights around the mother-specific mean infant birthweight

τ^2 : variance of the random deviations U_i, a measure of the dispersion of mother-specific mean infant birthweights around the 'typical' (U_i=0) mother-specific mean infant birthweight

6. **Fit this model. Calculate the ICC (show the calculation by hand even if your output gives you the ICC automatically). Interpret the ICC.**

```
. xtreg birthwt, i(mother) mle

Random-effects ML regression                    Number of obs      =      4390
Group variable: mother                          Number of groups   =       878

Random effects u_i ~ Gaussian                   Obs per group: min =         5
                                                               avg =       5.0
                                                               max =         5

                                                Wald chi2(0)       =      0.00
Log likelihood  = -33572.321                    Prob > chi2        =         .

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       _cons |   3156.304   14.06306   224.44   0.000     3128.741    3183.867
-------------+----------------------------------------------------------------
    /sigma_u |   368.4007   11.31476                      346.8785    391.2583
    /sigma_e |   435.4458   5.195672                      425.3806    445.7492
         rho |   .4171708   .0165993                      .3849558    .4499521
------------------------------------------------------------------------------
Likelihood-ratio test of sigma_u=0: chibar2(01)= 1034.16 Prob>=chibar2 = 0.000
```

$$ICC = \frac{\sigma_u^2}{\sigma_u^2 + \sigma_\varepsilon^2} = \frac{368.4^2}{368.4^2 + 435.4^2} = 0.42$$

Two equivalent interpretations of the ICC:

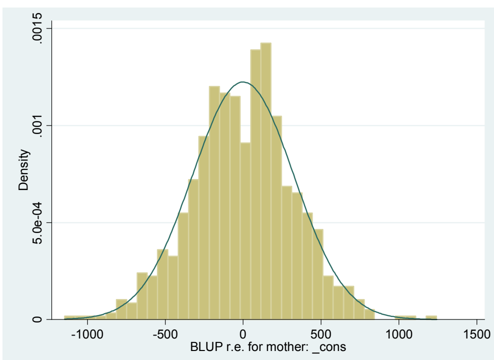
1. The proportion of the total variance in infant birthweight that is due to variation between mothers
2. The correlation of the birthweights of infants from the same mother

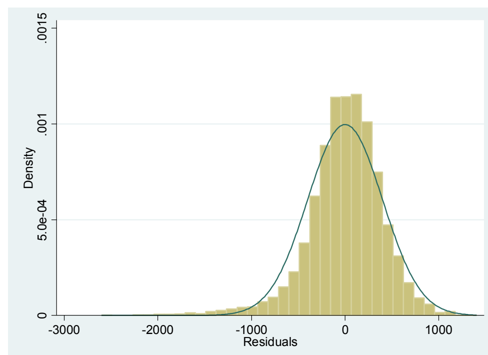
7. **Calculate and report**
   a. **The overall mean birthweight of infants in this study**
      `3156.3 grams`
   b. **The mean of the birthweights of the infants of the mother with id = 84**
      `3339.8 grams`
   c. **The fitted mean birthweight of the infants for the mother with id = 84 from the model fit in (6) where you assign values to the random intercepts using empirical Bayes.**
      `3299.7 grams`
8. **Given that the mother with id number 84 has currently had 5 children with the following birthweights: 2892, 3204, 4253, 2948 and 3402, which of the three values in (7) would you use as a prediction for the average birthweight mother 84's children, if she were to have 3 more children? Why?**

I would use the estimate from 7(c) because the empirical Bayes estimate will be a more accurate prediction of mean birthweight for mother 84 than either the overall mean or the current mother-specific mean birthweight. The empirical Bayes estimate is a weighted average of the overall mean and the current mother-specific mean. The empirical Bayes estimate introduces bias in our estimate since it shrinks the mother specific mean towards the overall mean, but this borrowing of information from the other mothers helps minimize the error in our prediction by trading bias for a reduction in variance.

9. **Assess the two normality assumptions about the errors and the random intercepts in your model specification in (5) using appropriate graphical methods. Briefly (in a sentence or two) describe what you observe.**

```
. predict ebri, reffects
. predict resid, residuals
. hist ebri, normal
. hist resid, normal
```





The normality assumption appears to be fairly well-satisfied for the empirical Bayes estimates of the random intercepts while the normality assumption is less well satisfied for the residuals. There appears to be slight left skew in the residuals. We could also plot histograms of the standardized forms of each of these variables to identify outliers.

Part III: Random Intercept model with covariates and Random Effects model

10. **Fit a random intercept model with birthweight as the outcome. Include linear terms for mother's age and interpregnancy interval. In non-mathematical terminology, interpret the estimated coefficients on mother's age and interpregnancy interval. Include 95% confidence intervals. Write a sentence or two comparing the magnitudes of the estimated coefficients to the magnitude of the estimated standard deviation of the random intercept for mother.**

```
. xtmixed birthwt age interval|| mother:, mle nolog

Mixed-effects ML regression                     Number of obs      =      3512
Group variable: mother                          Number of groups   =       878

                                                Obs per group: min =         4
```

```
                                                          avg =        4.0
                                                          max =          4

                                             Wald chi2(2)       =     52.85
Log likelihood = -26879.329                  Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   15.46804    2.53821     6.09   0.000     10.49324    20.44284
    interval |    20.1028    7.66621     2.62   0.009     5.077302    35.12829
       _cons |   2783.614   58.05281    47.95   0.000     2669.833    2897.396
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
mother: Identity             |
                   sd(_cons) |    377.431    12.0862      354.4705    401.8787
-----------------------------+------------------------------------------------
                sd(Residual) |   427.3717   5.899074      415.9647    439.0915
------------------------------------------------------------------------------
LR test vs. linear regression: chibar2(01) =   770.46 Prob >= chibar2 = 0.0000
```

We estimate that, for a mother of a given age, the birth weight of her infant will be 15.46 grams higher (95% CI: 10.49 to 20.44) than the birth weight would have been had she given birth one year prior, controlling for the interpregnancy interval preceding the delivery. We also estimate that, for a given mother of a fixed age, a one year increase in interpregnancy interval is associated with an increase in birthweight of her infant of 20.10 grams (95% CI: 5.08 to 35.13).

The magnitude of the estimated standard deviation of the random intercept for mother is much larger than the magnitudes of the effects of a one unit change in either age or interpregnancy interval (holding the other constant). This implies that there is a large amount of variability in infant birthweight that is due to unexplained differences between mothers that are represented in this model by random intercepts for mothers. In a future analysis, we might include more mother-level covariates to try to explain some of the differences between mothers.

11. Write down the model that extends your model in (10) by including a random intercept on mother's age. Interpret all the model parameters:

- **intercept**
- **fixed effect coefficients (2)**
- **random effects (2)**
- **variance of each random effect (2) and the covariance of the random effects**
- **error**
- **variance of the error**

$$y_{ij} | U_{1i}, U_{2i}, \text{age}, \text{interval} = (\beta_0 + U_{1i}) + \beta_1 \text{age}_{ij} + (\beta_2 + U_{2i}) \text{interval}_{ij} + \varepsilon_{ij}$$

$$\begin{pmatrix} U_{1i} \\ U_{2i} \end{pmatrix} \sim MVN\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \tau_{11} & \tau_{12} \\ \tau_{21} & \tau_{22} \end{pmatrix} \right)$$

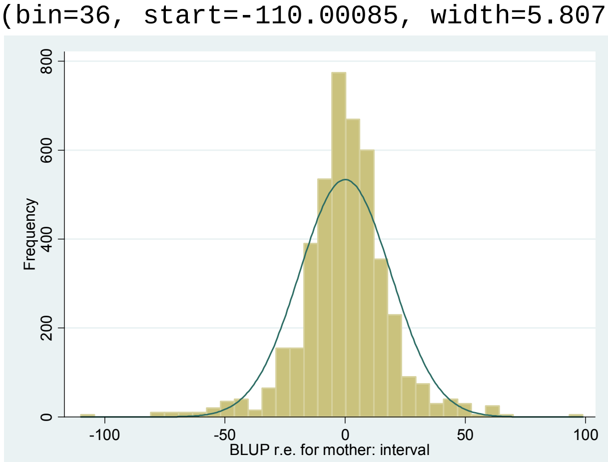
$$\varepsilon_{ij} \sim N(0, \sigma^2)$$

β_0 : average infant birthweight from a typical (U_{1i}=0) mother of age 0 and with a interpregnancy interval of 0 years. (You could make this more interpretable by centering the covariates age and interval)
U_{1i} : difference between the average infant birthweight for mother(i) and the average infant birthweight of a typical (U_{1i}=0) mother given that both the mothers are the same age and have an interpregnancy interval of zero.
β_1 : mother-specific increase in average infant birthweight associated with a one year increase in mother's age, controlling for interpregnancy interval
$\beta_2 + U_{2i}$: mother-specific increase in average infant birthweight associated with a one year increase in interpregnancy, controlling for mother's age
β_2: increase in average infant birthweight associated with a one year increase in interpregnancy interval for a mother with a typical interval effect (U_{2i}=0) of a fixed age
U_{2i}: additional increase in average infant birthweight associated with a one year increase in interpregnancy interval for mother (i) compared to the increase in birthweight for a mother with a typical interpregnancy interval effect (U_{2i}=0), with both mothers being of the same age
ε_{ij}: difference between the birthweight for child(ij) and the estimated average infant birthweight for mother(i) at that age and interpregnancy interval.
σ^2 : variance of the errors ε_{ij}, a measure of the dispersion of infant birthweights around the mother-specific mean infant birthweight for the given covariate and random effect values.
τ_{11} : variance of the random deviations U_{1i}, a measure of the dispersion of mother-specific mean infant birthweights around the 'typical'(U_i=0) mother-specific mean infant birthweight
τ_{22} : variance of the random deviations U_{2i}, a measure of the dispersion of mother-specific interpregnancy interval effects on infant birthweights around the 'typical'(U_i=0) mother-specific interpregnancy interval effect on infant birthweight
$\tau_{12} = \tau_{21}$: covariance of the random deviations U_{1i} and U_{2i}; a measure of how mother-specific deviations in baseline average birthweight relate to mother-specific deviations in the interpregnancy interval effect. If the covariance is negative, mothers who tend to have **heavier** babies than the typical mother, will have a **smaller** increases in birthweight due to increases in interpregnancy interval than the typical mother and vice versa. If the covariance is positive, mothers who tend to have **heavier** babies than the typical mother, will have a

larger increases in birthweight due to increases in interpregnancy interval than the typical mother and vice versa.

12. Fit the model in (11). Obtain the empirical Bayes estimates of the random intercept and random coefficients. Make a histogram of the empirical Bayes estimates of the random coefficients on interval.

```
. xtmixed birthwt age interval|| mother: interval, cov(unstruct)
mle nolog

Mixed-effects ML regression                     Number of obs      =      3512
Group variable: mother                          Number of groups   =       878

                                                Obs per group: min =         4
                                                               avg =       4.0
                                                               max =         4


                                                Wald chi2(2)       =     52.36
Log likelihood = -26876.762                     Prob > chi2        =    0.0000

------------------------------------------------------------------------------
     birthwt |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   15.59405   2.539878     6.14   0.000     10.61598    20.57211
    interval |   21.47751   8.269485     2.60   0.009      5.26962     37.6854
       _cons |   2778.794   58.36572    47.61   0.000     2664.399    2893.189
------------------------------------------------------------------------------

------------------------------------------------------------------------------
  Random-effects Parameters  |   Estimate   Std. Err.     [95% Conf. Interval]
-----------------------------+------------------------------------------------
mother: Unstructured         |
                sd(interval) |   67.56457   17.70042       40.4318    112.9055
                   sd(_cons) |   399.9725   24.06178      355.4864    450.0256
         corr(interval,_cons) |  -.3403248   .1472332     -.5920603   -.0280798
-----------------------------+------------------------------------------------
                sd(Residual) |   421.6683    6.39221      409.3241    434.3847
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =    775.59   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

. predict ebcoef ebri, reffects

. hist ebcoef, norm freq
(bin=36, start=-110.00085, width=5.807226)
```



13. **You have very limited funds to conduct a targeted intervention that teaches mothers the positive effects on infant birthweight of waiting longer between pregnancies. Use the empirical Bayes estimates of the random coefficients on interpregnancy interval to identify the 5 mothers whose future children might potentially benefit the most (in terms of birthweight) from the intervention. List the mother ids and explain your reasoning.**

```
. sort ebcoef
. list mother ebcoef in -25/L if birthorder==1

      +-------------------+
      | mother     ebcoef |
      |-------------------|
4369. | 122609   62.07715 |
4372. |  58339   63.76237 |
4377. |  48028   64.05288 |
4382. |   6847   67.09091 |
4386. | 187444   99.05929 |
      +-------------------+
```

The mothers with the largest values of the random coefficient on interval have the largest subject-specific effect of increasing the length of interpregnancy interval on increasing infant birthweight. Hence if I had limited funds to promote the benefits of longer interpregnancy intervals to a subset of my study sample, I would chose those mothers in the study sample for whom longer interpregnancy intervals have the potential to have the greatest impact on increasing infant birthweight.

14. **Write a one paragraph summary of your analysis and findings regarding infant birthweight and interpregnancy interval as if for a scientific journal.**

Our goal was to assess the relationship between interpregnancy interval and infant birthweight, controlling for mother's age, using data on live births that occurred in the state of Georgia from 1980 to 1992. We fit a multilevel model of birthweight on age and interpregnancy interval that accounted for clustering of infants within mother by including a random intercept for mother and allowed the effect of interpregnancy interval to vary between mothers by including a random coefficient at the mother level on interpregnancy interval. We estimate that, for a typical mother of a given age, the birth weight of her infant will be 15.59 grams higher (95% CI: 10.62 to 20.57) than the birth weight would have been had she given birth one year prior, controlling for the interpregnancy interval preceding the delivery. For a typical mother of a fixed age, a one year increase in interpregnancy interval is associated with an increase in birthweight of her infant of 21.48 grams (95% CI: 5.27 to 37.69). However, controlling for mother's age, the effect of interpregnancy interval on infant birthweight varies considerably between mothers. We estimate that, for a fixed age, a one year increase in interpregnancy interval is associated in a change in infant birthweight that ranges from -110.95 to +153.89 grams in 95% of the mothers in our sample.
Note: I obtained the numbers in the last sentence using 21.47 +/- 1.96*67.56

Reference:

Adams MM, Delaney KM, Stupp PW, McCarthy BJ, Rawlings JS. The relationship of interpregnancy interval to infant birthweight and length of gestation among low-risk women, Georgia. Paediatr Perinat Epidemiol. 1997 Jan;11 Suppl 1:48-62