Memorandum

To: Dr. Ventura

From: Diego Krenz, Billy Block, Alexander Yuan, Matthew Huang

Date: May 22nd, 2025

Re: IBaby Month One: Key Findings and Analysis

The purpose of this memorandum is to describe and summarize our statistical analysis and key findings related to your research project which examines the relationships between maternal depression, parenting stress, and technology use during the first month postpartum. We hope that our findings are helpful in informing your ongoing research in this area.

This memo is organized into five sections:

I. Abstract of Key Findings

Using data collected from 64 mothers, we computed CES-D-based depression scores, stress scores, and perceived technology-related distraction, along with demographic covariates. We used logistic regression to model the likelihood of maternal depression, defined as a CES-D score ≥13. Backward stepwise selection identified five key predictors: stress score, perceived distraction, maternal age, household income, and marital status. Results indicate that higher stress (p=.002) and distraction levels (p=.004) are significantly associated with higher odds of depression. Conversely, older maternal age reduced depression risk (p=.040). Notably, low income (p=.011) and being married to the baby's biological father (p=.037) were associated with substantially higher odds of depression.

II. Background and Data

Our understanding is that assistance is needed in answering if high maternal technology use during infant feeding/care and high levels of stress are able to predict maternal depression at one month postpartum en route to the research goal of examining the associations between maternal technology use, parenting stress, and maternal depression during the early postpartum period. Below we have listed the variables that we included in the analysis.

Depression score/indicator (depress_sum):

This depression indicator variable is calculated from the CES-D score which is calculated from 20 likert scale (0,1,2,3) questions on depression.

For questions 1-3, 5-7, 9-11, 13-15, 17-20 the scoring is:

- Rarely or none of the time (less than one day) = 0
- Some or a little of the time (1-2 days) = 1
- Occasionally or a moderate amount of time (3-4 days) = 2
- Most or all of the time (5-7 days) = 3

For questions 4, 8, 12, 16 are reverse scored as follows:

- Most or all of the time (5-7 days) = 0
- Occasionally or a moderate amount of time (3-4 days) = 1
- Some or a little of the time (1-2 days) = 2
- Rarely or none of the time (less than 1 day) = 3

These 20 scores are then summed, resulting in a range of possible scores between 0 and 60, with higher scores indicating the presence of more symptomatology. Scores are then dichotomized into a binary categorical variable with a score of 13 or more being classified as depressed.

Parental Stress Score (stress sum):

The parental stress score variable is calculated from 18 likert scale (1,2,3,4,5) questions on stress.

For questions 1, 2, 5, 6, 7, 8, 17, and 18: the scoring is reverse scored as follows:

(1=5) (2=4) (3=2) (4=1) (5=1).

All 18 scores are then summed into a single parental stress score.

Maternal Age (m0momage):

Mother’s age in years.

Mother of Hispanic or Latina ethnicity? (m0momethnic):

Coded as 1 for yes, 0 for no.

Mother Race (m0morace):

Re-coded as 1 for White, 0 for other.

Mother Education (m0momeduc):

Re-coded as 1 for at most high school diploma or GED, 0 for higher attainment.

Mother Marital Status (m1maritalstatus):

Re-coded as 1 if married to the baby's biological father, 0 otherwise.

Monthly Household Income (m1income):

Re-coded as 1 if monthly income under $4,000, 0 if monthly income is at least $4,000.

iBaby Month One: Key Findings and Analysis

In the cleaning process, we removed the columns m1mdqfeedother and m1mdqcareother due to high missingness (9 and 14 missing values, respectively). After this, we remove any observations that have missing values or preferred not to share their income. Our cleaned data set contained 65 observations, of which 16 (25%) are considered depressed by the depression score criteria (13>=). We also decided to recode the 4 variables above, as all of them had a

We decided to recode the four categorical variables mentioned above into binary (0/1) format because each had a dominant category. In other words, most observations fell into one or two categories, making it practical to simplify them into binary indicators.

Table 1 gives the mean and standard deviation of the depression and stress scores. Figure 1 gives the distribution of the depression scores with the gold dashed line representing the 13 passing cutoff. We note that this distribution looks slightly skewed to the right. Figure 2 gives a histogram displaying the monthly incomes for the mothers. We note that the $6000-$6999 monthly income category had the most mothers. Figure 3 shows a box and whisker plot displaying parenting stress score by depression status. The visual confirms our prior belief that there should be a positive correlation between depression and stress. Figure 4 displays a correlation matrix between key numeric variables. We note the moderate positive correlation of 0.45 between the depression sum and stress sum which also confirms our observation in the previous sentence.

iBaby Month One: Key Findings and Analysis

Table 1:

Variable	Mean	Standard Deviation
Depression Score	8.921	8.752
Stress Score	37.125	8.337

Figure 1

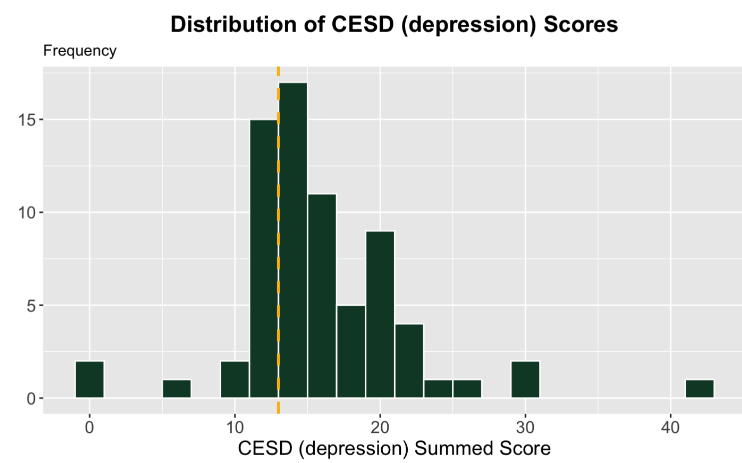


Figure 2:

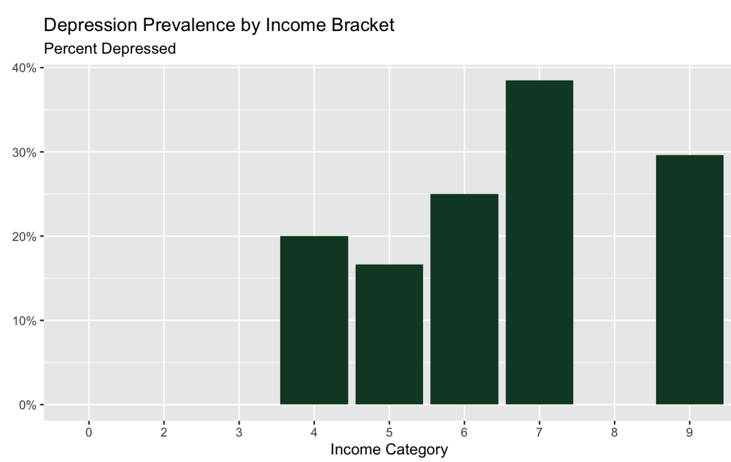


Figure 3:

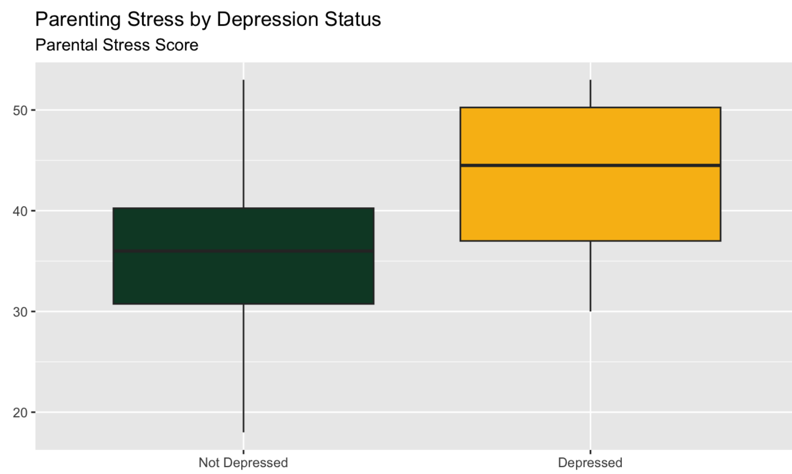
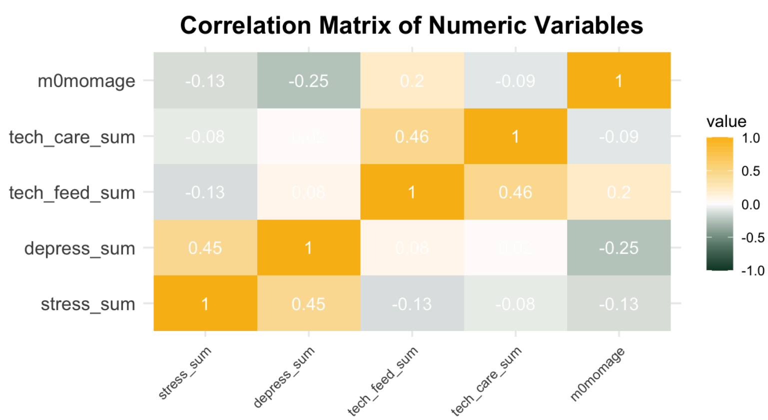


Figure 4:



III. Methods

Due to the dichotomous nature of our response variable of interest, depression, we chose **logistic regression** as an appropriate model of the relationship between the depression, technology use, and covariate variables of interest.

Logistic Regression Model

Logistic Regression is a statistical model which can be used to model how input variables relate to the odds of a binary response variable. Specifically, the model expresses the log odds of the response as a linear function of the inputs:

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n$$, where

- p is the probability of the response, with $\frac{p}{1-p}$ being the odds of the response,
- β_0 is the intercept of the linear model,
- x_i is the i'th predictor variable, and
- β_i is the i'th coefficient of the linear model, representing the change in log odds of the response given a one-unit increase in x_i.

Logistic regression models the relationship between predictors and a binary outcome (in this case, being classified as depressed) by estimating changes in the log-odds of the outcome. The odds represent the ratio of the probability of the event occurring to it not occurring, and taking the logarithm of the odds allows for a linear relationship with the predictors. While the log-odds themselves are not directly intuitive, they can be converted to odds ratios, which express how a one-unit increase in a predictor is associated with a change in the odds of the outcome, holding other variables constant.

Variable Selection

With the large number of variables relative to the sample size, it would be inappropriate to include all of the available variables since doing so can lead to overfitting, where the model learns the noise in the data rather than true underlying patterns. Additionally, including many irrelevant variables can reduce the precision and interpretability of the model. Therefore, we decided to use the summarized stress, depression, and technology variables (as detailed in the previous section), rather than individual survey questions, to reduce the number of predictors in the model while retaining most of the important information from the responses.

Additionally, we used backwards stepwise regression, which starts with a set of initial input variables and iteratively removes the least predictive variable (by p-value) until every variable was significant at the 0.1 level. This more lenient significance threshold helps retain predictors that may have moderate associations with the outcome variable but may not meet smaller significance thresholds due to factors like low sample size.

IV. Results and Discussion

Final Model

Final model equation: $log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5$, where

- p is the probability of a mother having depression,
- β_0 is the intercept of the linear model,
- β_1 represents the change in log odds of depression in a mother given a one-unit increase in the Stress Sum holding the other variables fixed,
- β_2 represents the change in log odds of depression in a mother given a one-unit increase in the Perceived Distraction holding the other variables fixed,
- β_3 represents the change in log odds of depression in a mother given a one year increase in a mother's age holding the other variables fixed,
- β_4 represents the change in log odds of depression given an income of below $40,000, holding the other variables fixed,
- β_5 represents the change in log odds of depression given a mother is married or not holding the other variables fixed.

Table 2: Logistic Regression Model Coefficients

Predictor	**β Estimate**	**Effect** (Odds ratio)	Std error	**P-value**
Intercept	1.996		4.116	0.006
Stress Sum (Standardized)	**1.944**	**6.985**	0.6378	**0.002**
Perceived Distraction (Standardized)	**1.747**	**5.738**	0.610	**0.004**
Mother Age	**- 0.255**	**0.775**	0.124	**0.040**
Low Income	**4.223**	**68.23**	1.649	**0.010**
Married	**3.038**	**20.87**	1.455	**0.036**

Interpretation:

All of the selected input variables are significant at the 5% level, with

1. *(Stress Sum)* A standard deviation increase in total stress scores corresponds with **~7 times higher** odds of depression, holding all other variables constant.
2. *(Perceived Distraction)* A standard deviation increase in perceived distraction scores corresponds with **~5.7 times higher** odds of depression, holding all other variables constant.
3. *(Mother Age)* Each additional year of a mother's age **reduces** the odds of depression by **~77%**, holding all other variables constant.
4. *(Low Income)* An annual income of less than $40,000 corresponds with **~68 times higher** odds of depression, holding all other variables constant.
5. *(Married)* A mother being married corresponds with **~21 times higher** odds of depression, holding all other variables constant.

Example:

Consider a 22-year-old mother whose standardized stress score is half a standard deviation below the mean (stress_sum = –0.5), whose perceived distress on the MDQ scale is exactly average (m1MDQpercievedist = 0), who is married (married = 1) and not low-income (low_income = 0). Plugging her characteristics into the updated logistic model and exponentiating gives odds of depression of 0.21 to 1. In other words, this profile corresponds to roughly a one-in-five chance of being classified as depressed.

Performance:

Our final model achieves an ROC AUC of 0.915 and accuracy of 74% on the dataset, indicating that the model is able to capture meaningful patterns in the data and distinguish well between individuals classified as depressed and not depressed. Additionally, we found no evidence of violations to the linearity and independent predictor assumptions for our logistic regression model.

V. Technical Output

Depressed	Frequency
1	16
0	48

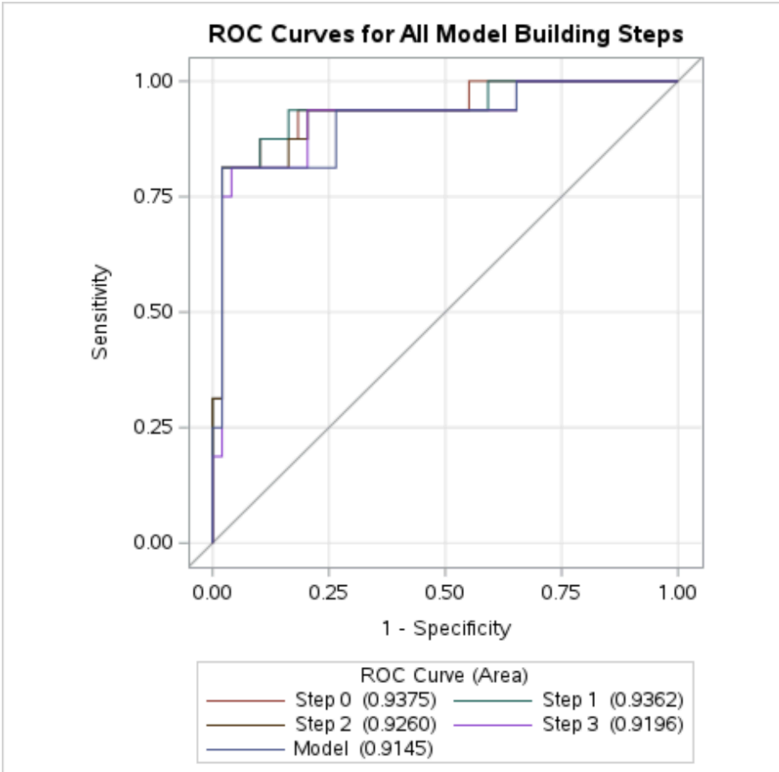
25% of mothers in our dataset are considered depressed.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	1.9961	4.1160	0.2352	0.6277
stress_sum		1	1.9437	0.6378	9.2872	0.0023
m1MDQpercievedist		1	1.7470	0.6097	8.2103	0.0042
m0momage		1	-0.2552	0.1243	4.2131	0.0401
low_income	0	1	4.2229	1.6493	6.5554	0.0105
married	0	1	3.0383	1.4551	4.3596	0.0368

Parameter estimates from the final fitted logistic regression model.

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
stress_sum	6.985	2.001	24.381
m1MDQpercievedist	5.738	1.737	18.954
m0momage	0.775	0.607	0.989
low_income 0 vs 1	68.230	2.692	>999.999
married 0 vs 1	20.870	1.205	361.526

Odds ratio estimates from the final fitted logistic regression model.

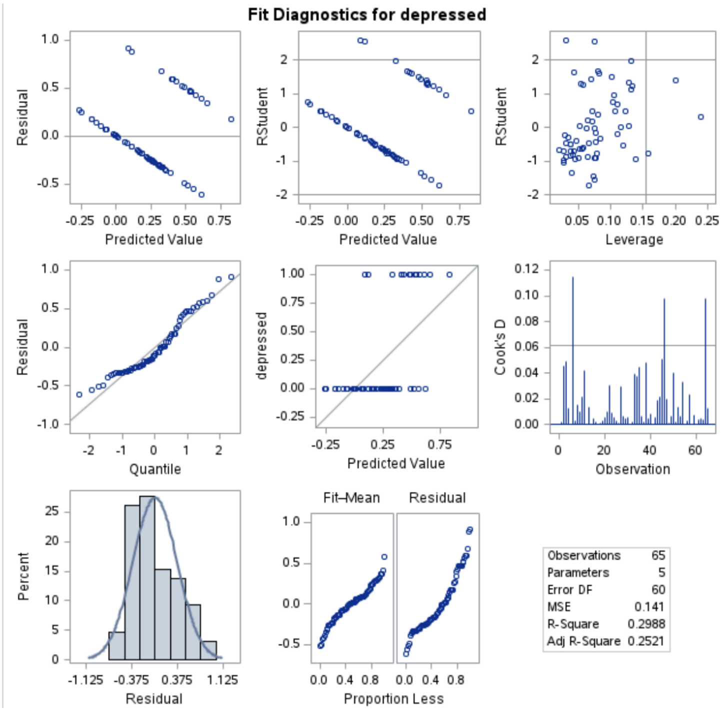


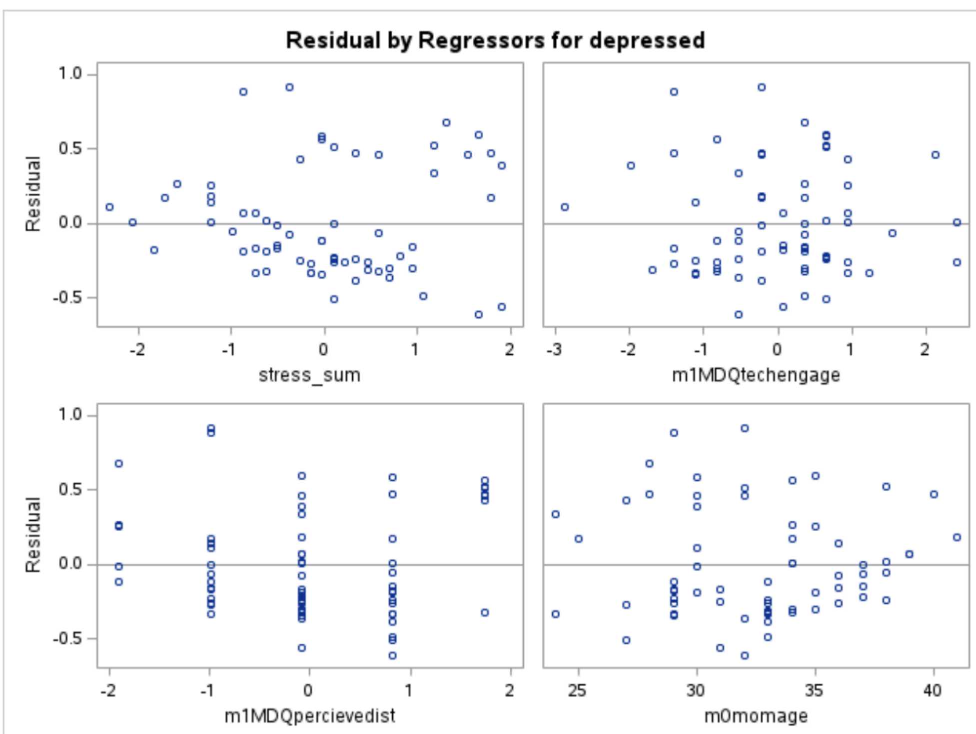
Our final model achieves a strong ROC AUC, indicating strong predictive performance.

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	1	1.16965	0.40490	2.89	0.0054	0
stress_sum	1	0.14612	0.04842	3.02	0.0037	1.06451
m1MDQtechengage	1	-0.01182	0.04762	-0.25	0.8048	1.02954
m1MDQpercievedist	1	0.12543	0.04889	2.57	0.0128	1.08508
m0momage	1	-0.02831	0.01233	-2.30	0.0252	1.05880

No issues arise with multicollinearity in our model.





No notable patterns in the residuals between quantitative predictors.

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1145.5	3635.6	0.0993	0.7527
stress_sum	1	-65.4085	51.4345	1.6172	0.2035
ln_stress	1	64.8309	49.6503	1.7050	0.1916
stress_int	1	27.9268	22.2911	1.5696	0.2103
m1MDQtechengage	1	-15.5432	19.7650	0.6184	0.4316
ln_tech	1	14.2480	18.3563	0.6025	0.4376
tech_int	1	5.7337	9.0038	0.4055	0.5242
m1MDQpercievedist	1	-29.2617	16.8383	3.0200	0.0822
ln_dist	1	24.5054	13.5033	3.2934	0.0696
dist_int	1	18.6963	10.1841	3.3703	0.0664
m0momage	1	-126.8	396.7	0.1022	0.7492
ln_age	1	743.5	2336.8	0.1012	0.7504
age_int	1	23.2741	72.9151	0.1019	0.7496

Linearity assumptions are not violated, as all interaction terms between continuous predictors and their log transformations have large p-values, indicating no evidence of non-linearity in the logit.