Facilities Accepting Oil and Gas Waste and Birthweight: An Exploratory Bayesian Analysis

by

Nicholas Tedesco

BS, BA, University of Pittsburgh, 2021

Submitted to the Graduate Faculty of the

Department of Biostatistics

School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2022

UNIVERSITY OF PITTSBURGH

SCHOOL OF PUBLIC HEALTH

This thesis was presented by

by

Nicholas Tedesco

It was defended on

December 11, 2022

and approved by

Jeanine M. Buchanich, PhD, Research Associate Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

Ada O. Youk, PhD, Associate Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

Jenna C. Carlson, PhD, Assistant Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

James P. Fabisiak, PhD, Associate Professor, Department of Environmental and Occupational Health, School of Public Health, University of Pittsburgh

Thesis Advisor: Jeanine M. Buchanich, PhD, Research Associate Professor, Department of Biostatistics, School of Public Health, University of Pittsburgh

Facilities Accepting Oil and Gas Waste and Birthweight:
An Exploratory Bayesian Analysis

Nicholas Tedesco, MS

University of Pittsburgh, 2022

Background: Previous studies have identified links between fracking and adverse health outcomes such as reduced birthweight. However, none have examined the potential health effects of exposure to facilities accepting oil and gas waste. The purpose of this work was to investigate the relationship between residential proximity to facilities accepting oil and gas waste, as defined via a binary exposure variable, and birthweight.

Methods: The relationship between exposure to facilities accepting oil and gas waste within two kilometers of the mother's residence and birthweight was examined via both linear ordinary least squares and Bayesian regression. Various Bayesian priors were specified on the exposure coefficient to understand the effects of different distributions and parameters. Models were compared by looking at differences in the exposure coefficient, its 95% confidence/credible interval, and test set root mean squared error (RMSE).

Results: Both the unadjusted and adjusted linear models found a negative relationship between proximity (exposure) and birthweight in grams (unadjusted model β = -56.87 g, 95% CI [-71.82 g, -41.93 g]; adjusted model β = -13.34 g, 95% CI [-25.02 g, -1.65 g], respectively). As we might expect, the Bayesian exposure coefficient was increasingly pulled towards its respective prior mean as the prior standard deviation decreased. In terms of test set RMSE, none of the univariate Bayesian models outperformed their corresponding linear model, but many of the adjusted Bayesian models outperformed their linear model counterpart.

Conclusion: This thesis found an association between proximity to facilities accepting oil and gas waste and lower average birthweight (grams). However, this association was relatively small and requires further support.

Public Health Significance: This work may serve to better inform our collective understanding of the impacts of fracking on birth outcomes. Furthermore, no previous study has investigated the effects of facilities accepting oil and gas waste on birthweight.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Over the last 25 years, the American energy landscape has undergone drastic evolution, perhaps most notably with the expansion of hydraulic fracturing operations. In fact, from 2000 to 2015, the number of hydraulically fractured wells in the United States increased from 23,000 to approximately 300,000 [1]. This rapid growth has corresponded to a range of economic benefits, including decreased energy costs [2] and greatly increased production of both oil and natural gas [1]. However, mounting evidence suggests that hydraulic fracturing may have adverse impacts on public health and well-being.

1.1 Fracking and its Effects

Hydraulic fracturing – also known as fracking – is the process of injecting large amounts of fluid at high pressure into dense rock in order to free trapped oil and natural gas [3]. The fluid used for injection typically consists of a mixture of water, sand, and various chemical additives, including some with potential toxicity. As one might expect, this technique raises concern over the quality and contamination of local air and water. According to a review from the Concerned Health Professionals of New York, fracking has resulted in widespread air pollution and water contamination that will only worsen with time [4]. Some research has found increased methane levels in ground and well water near fracking sites in the Marcellus Shale region of Northeastern America [5]. From a broader perspective, some articles have also expressed concern regarding the effects of fracking on the atmosphere [6].

Other studies have gone one step further to investigate the potential effects that fracking may have on health. For example, air pollution related to fracking may cause various forms of irritation and respiratory illness [7]. Some fracking pollutants have also been shown to be highly carcinogenic [8] – although it is difficult to prove that these pollutants have an immediate and direct effect on human health, this may manifest through higher rates of cancer as time continues. Finally, fracking has been linked to various effects on birth outcomes. Multiple studies have found an association between proximity to fracking drilling locations and decreased birthweight [9 – 12]. For example, Currie et al. (2017) found that the average birthweight was 39 grams lower for babies whose mothers live within one kilometer of fracking sites [12].

Most of the aforementioned studies have examined the local effects of fracking relative to fracking well sites. However, few have examined the effects of the waste produced from fracking alone. Fracking wastes – which consist of sludges/sediments, contaminated equipment or components, and produced waters – are classified as Technologically Enhanced Naturally Occurring Radioactive Material (TENORM) [13]. While there are a range of methods for disposing of TENORM, including reinjection deep underground, one common approach is to transfer these radioactive wastes to offsite waste disposal facilities. Currently, little to no research in the field has focused on exposure to facilities that accept radioactive waste from fracking operations. Therefore, the primary goal of this study was to examine the potential impact of residential proximity to facilities accepting oil and gas waste on birthweight.

1.2 Introduction to Bayesian Statistics

Ordinary least squares (OLS) regression estimates coefficients, and thereby relationships, using the current dataset alone. However, this approach ignores any prior beliefs we may have about certain patterns in the data. In comparison, Bayesian statistics allow us to incorporate prior beliefs into the model fitting process, which ultimately influences our results. In this situation, multiple studies have found a lower average birthweight for mothers who live close to fracking operations. A prior specification that captures this relationship enables us to make use of previous data, which may improve the generalizability and application of our results.

1.3 Objectives

The purpose of this project was to examine the relationship between residential proximity to facilities accepting oil and gas waste from fracking operations and birthweight. To accomplish this, a series of unadjusted and adjusted models were fit using ordinary least squares (OLS) and Bayesian regression. In order to better understand the impact of the prior, and to determine whether the prior can influence model quality, a series of Bayesian models were fit using different prior specifications. Models were compared using the primary coefficient of interest and its respective 95% confidence/credible interval and were evaluated by calculating root mean squared error on a 10% test subset of the data not used for model training. With this information in mind, the objectives of this work are outlined as follows:

1. Use linear regression (ordinary least squares) to evaluate the relationship between maternal residential proximity to facilities accepting oil and gas waste from fracking operations and birthweight.
2. Fit various Bayesian regression models using different prior specifications and compare the results. Priors were only set on the relationship between maternal residence proximity and birthweight.

1.4 Public Health Significance

This work has clear importance to the field of public health. Considering the relatively recent and widespread emergence of fracking, current research on its potential environmental and health impacts is lacking. Any additional work is a valuable contribution to our collective understanding of fracking. Furthermore, some studies have found an association between residential proximity to fracking well sites and birthweight. However, no study has extended this work to investigate the potential impact of residential proximity to waste sites accepting oil and gas waste from fracking. Therefore, this work is a logical extension of the current body of research and may help to better inform us on the health effects of fracking.

2.0 Methods

2.1 Data Source

2.1.1 Birth Data

All health-related data was obtained from the Bureau of Health Statistics and Research, Department of Health, Pennsylvania following Institutional Review Board (IRB) and Protected Access approvals. The inclusion criteria were live births occurring between January 1, 2010 and December 31, 2020 to mothers living within the following eight counties: Allegheny, Armstrong, Beaver, Butler, Fayette, Greene, Washington, and Westmoreland. The exclusion criteria were as follows:

1. Death occurred within seven days of birth
2. Infant suffered from serious birth defects
3. Multiple (non-singleton) birth
4. Unknown gestational age
5. Gestational age < 22 weeks (pre-viability) or > 41 weeks (post-term)
6. Birthweight missing or < 500 g
7. Maternal residence located outside of eight-county study area or within City of Pittsburgh (no fracking allowed within the city)

2.1.2 Facilities Accepting Oil and Gas Waste

Data on facilities accepting oil and gas waste were obtained from the Pennsylvania Department of Environmental Protection (PA DEP). Here, fracking waste loosely refers to substances such as sludge/sediment, flowback from hydraulic fracturing, and produced water coming from fracking well sites. Facilities were included if their disposition method, as defined by the PA DEP, was either landfill, public sewage treatment plant, or residual waste processing facility.

2.2 Data Definitions and Preparation

2.2.1 Predictor Variable

The main predictor of interest in this work was residential proximity to facilities accepting oil and gas waste from fracking operations. This categorical exposure metric was defined as "yes" if the mother's residence was within two kilometers of a candidate waste facility, and "no" if it was not. To determine exposure status, the distance between each maternal residence and each waste facility was calculated and summarized in a distance matrix. If a given health record had at least one waste facility within two kilometers of the mother's residence, the record was assigned a positive exposure status.

2.2.2 Outcome Variable

The outcome variable was birthweight (g). All records with missing birthweight values were dropped prior to analysis (n = 2160). A total of 183,442 records with non-missing birthweight values were used in the final analysis.

2.2.3 Covariates

Various clinical, demographic, and environmental features were included as covariates in the adjusted models to control for potential confounding. These covariates are listed as follows: gestational age, neonate sex, adequacy of prenatal care, maternal age at delivery, race, education, smoking status during pregnancy, pre-pregnancy body mass, parity, diagnosis of gestational diabetes, and receipt of WIC services. Table 1 provides more in-depth definitions for each of these variables.

Table 1: Covariate Definitions

Covariate	Definition
Neonate sex	Neonate sex (male, female, unknown)
Gestational age (weeks)	Obstetric estimate of gestation
Maternal age (years)	Mother's age at delivery
Race	Maternal single race self-designation collapsed into the following categories: • White: White • Black or African American: Black or African American • All other races: American Indian or Alaska Native, Asian Indian, Chinese, Filipino, Japanese, Korean, Vietnamese, Other Asian,

	Native Hawaiian, Guamanian or Chamorro, Samoan, Other Pacific Islander, Other • Unknown or refused: Don't know/Not sure, Refused
Education	Maternal education level collapsed into the following categories: • Less than High School: 8th grade or less, 9th-12th grade but no diploma • High School or GED: High School graduate or GED completed • Some college: Some college credit but not a degree, Associate's degree • Bachelor's or Graduate degree: Bachelor's degree, Master's degree, doctorate or professional degree • Unknown: Unknown
Smoking status during pregnancy	Categorical variable (yes, no, unknown) for smoking during the three months before pregnancy or during any trimester
Pre-pregnancy body mass index (BMI; kg/m^2)	BMI was calculated based on the mother's pre-pregnancy weight in pounds and height in feet and inches: BMI = 703 * [weight] / (12*[height ft] + [height in])^2 Then, BMI was categorized using one of two sets of criteria, depending on maternal age at birth. For births to mothers aged 20 years or younger, we used the following criteria based on the CDC's recommended youth BMI-for-age cutoffs [14]. • **Underweight:** <5th percentile • **Normal:** 5th to <85th percentile • **Overweight:** 85th to <95th percentile • **Obese:** ≥ 95th percentile • Unknown: missing height and/or weight Percentile data were available for males and females at one-month age increments with a half-month offset. Because maternal age was available only in years, we used data for females corresponding to the lower bound of

	age (e.g., for an 18-year-old mother, we used data corresponding to 216.5 months). For births to mothers aged 21 years or older, or for births in which maternal age was missing, we used the following criteria based on the CDC's recommended cutoffs for adults [15]: • **Underweight:** BMI <18.5 • **Normal:** BMI ∈ [18.5, 25) • **Overweight:** BMI ∈ [25, 30) • **Obese:** BMI ≥ 30 • Unknown: missing height and/or weight
Parity	Categorized as follows: • Nulliparous: no previous live births • Multiparous: ≥ 1 previous live birth
Gestational diabetes	Diagnosis of gestational diabetes during the pregnancy (yes, no, unknown).
Adequacy of prenatal care utilization index (APNCU)	The Adequacy of Prenatal Care Utilization (APNCU) Index [16] determines the adequacy of prenatal care utilization based on two parts: (1) the month in which prenatal care is initiated and (2) the number of prenatal care visits from initiation of care until delivery. The observed number of prenatal care visits is compared to the expected number of visits based on the schedule of prenatal care visits recommended by the American College of Obstetricians and Gynecologists (ACOG) [17]. The typical ACOG-recommended schedule is: • One visit every four weeks for the first 28 weeks of gestation • One visit every two weeks until 36 weeks of gestation • One visit every one week until birth The four categories of the APNCU Index are defined as follows:

	1. **Inadequate:** beginning care after the fourth month of pregnancy (16 weeks gestation) OR receiving less than 50% of expected prenatal care visits 2. **Intermediate:** beginning care by the fourth month of pregnancy AND receiving 50-79% of expected visits 3. **Adequate:** beginning care by the fourth month of pregnancy AND receiving 80-109% of expected visits 4. **Adequate plus:** beginning care by the fourth month of pregnancy AND receiving 110% or more of expected visits The APNCU index cannot be calculated if the number of prenatal visits and/or the date of the first prenatal visit is unknown. The only exception is that an "Inadequate" rating can be assigned if the date of the first prenatal is known and occurred after the fourth month of pregnancy, but the number of prenatal visits is unknown. If the month and year of the first prenatal visit was known but the day was missing, we assigned the visit date to the first of the month. Unknown and Inadequate were collapsed into a single category.
Receipt of WIC services	Indicates whether the mother receives WIC food (yes, no, unknown).
Community socioeconomic deprivation	For each community, we calculated an index of socioeconomic deprivation incorporating six indicators from the 2015-2019 American Community Survey 5-year estimates [18] from the US Census. These indicators include: • Percent less than high school education • Percent in poverty • Percent not in the labor force • Percent on public assistance • Percent does not own a vehicle

	• Percent civilian unemployment The six indicators were standardized for direction, natural log-transformed if necessary, z-scored using the standard deviations for Pennsylvania, and summed to create the final, unitless index for each county, township, or census tract. The total number of communities was divided into quartiles of socioeconomic deprivation index. Higher values of the index reflect greater community socioeconomic deprivation.

All continuous covariates (gestational age and maternal age) were centered prior to analysis for the sake of interpretability. All categorical covariates (neonate sex, adequacy of prenatal care, race, education, smoking status during pregnancy, pre-pregnancy body mass, parity, diagnosis of gestational diabetes, and receipt of WIC services) were factorized to ensure the correct reference category was used in modeling.

2.3 Statistical Analysis

2.3.1 Fundamentals of Bayesian Regression

Regression allows us to model the relationship between a series of predictors and an outcome variable. More specifically, regression calculates the most optimal coefficient for each predictor, where each coefficient's value determines the predictor's contribution to the outcome. Mathematically, this is written as:

Equation 1: $y = X\beta + \varepsilon$

where y is the outcome, X is the predictor matrix, beta is the coefficient vector, and epsilon is the error term, which represents the random error between the model and the actual data. The beta terms are most commonly estimated using the ordinary least squares approach, which optimizes the beta vector to produce the minimum residual sum of squares.

This above process can be more generally referred to as the frequentist approach to regression. Frequentist regression only considers the data at hand. However, if the dataset is small, this may limit the accuracy of our coefficients relevant to their true population values. Furthermore, if we have a prior understanding of the relationship between certain predictors and the outcome variable, this prior knowledge cannot be incorporated into the final model.

Bayesian regression accounts for these shortcomings by making use of prior specifications during coefficient calculation. To accomplish this, the Bayesian method takes the following distribution-based approach:

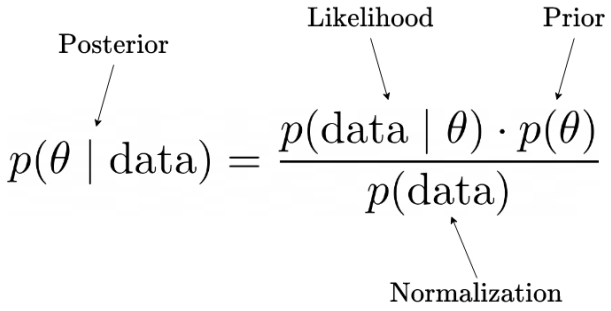


Figure 1: Fundamentals of Bayesian Inference [19]

In other words, the posterior distribution on our parameters is a balance between the likelihood of observing our data given the parameters and our prior belief on the parameters (the

distribution of the data, otherwise known as the “evidence” or “normalization” term, is included for the primary purpose of normalizing the data). In terms of regression, these concepts can be applied as follows: Bayesian regression estimates the posterior coefficient distributions by combining the likelihood of the current data at hand (likelihood) with our prior belief of the relationships in the data (prior). For more on the process of Bayesian regression, please see *An Introduction to Bayesian Thinking* [20].

It is important to emphasize that Bayesian inference yields a posterior predictive distribution, as opposed to point estimates for our coefficients, that can be used to predict the outcome. From a conceptual standpoint, this allows us to quantify the uncertainty on our predictions in a much more transparent fashion. From a logistical perspective, this means we must sample from the distribution to obtain our predictions. There are various methods for sampling from the posterior predictive distribution to approximate the outcome (see Section 2.3.2). Overall, Bayesian regression differs from the frequentist approach in many crucial ways.

2.3.2 Model Fitting with brms Package in R

This project used the brms package in R [21]. brms fits a Bayesian regression model using the probabilistic programming language Stan. More specifically, brms generates C++ code from R input to fit models in Stan, then passes the results back to R for post-processing. The primary function in brms – brm() – fits Bayesian models using the following format:

```
fit1 <- brm(formula = time | cens(censored) ~ age * sex + disease
            + (1 + age|patient),
            data = kidney, family = lognormal(),
            prior = c(set_prior("normal(0,5)", class = "b"),
                      set_prior("cauchy(0,2)", class = "sd"),
                      set_prior("lkj(2)", class = "cor")),
            warmup = 1000, iter = 2000, chains = 4,
            control = list(adapt_delta = 0.95))
```

Figure 2: brms::brm Function [21]

The formula and data arguments are standard to typical model-fitting functions in R. Family specifies the distribution of the response variable – for example, using a "gaussian" family indicates that we are performing linear regression. The prior argument is where we list our prior distributions on the model parameters, which may include coefficients, standard deviation, and so on. In this example, we specify the same normal prior on all model coefficients by indicating "class = 'b'" (where b represents beta) – if we wished to place a prior on only one coefficient, we would also include "coef = 'age'" in the set_prior function.

Finally, BRMS uses Hamiltonian Monte-Carlo (HMC) and No-U-Turn Sampling (NUTS) to sample from the posterior and thus approximate the outcome. As demonstrated in Figure 1, BRMS by default uses four Markov Chains with 2000 iterations per chain. In Monte-Carlo Markov Chain simulation, each iteration is a sample of the posterior distribution, where the samples within each chain are dependent on one another. The warmup iterations are used to allow the sampling density to converge to a stationary state, indicating that our samples have approximated the real distribution [22].

2.3.3 Modeling Process

In this thesis, a series of models were fit, compared, and evaluated. First, univariate and covariate-adjusted linear models were fit using the lm function in R. These were considered the "base" models, serving as reference points for the corresponding Bayesian models. Then, multiple Bayesian models were fit using the following prior specifications:

Table 2: Bayesian Priors

• b0: no prior • b1: uniform(-40, -38) • b2: normal(-39, 25) • b3: normal(-39, 5) • b4: normal(-39, 1) • b5: student_t(1, -39, 1) • b6: student_t(50, -39, 1)	(strong uniform) (uninformative normal) (weak normal) (strong normal) (weak t) (strong t)	*Univariate*
• b7: no prior • b8: uniform(-40, -38) • b9: normal(-39, 25) • b10: normal(-39, 5) • b11: normal(-39, 1) • b12: student_t(1, -39, 1) • b13: student_t(50, -39, 1)	(strong uniform) (uninformative normal) (weak normal) (strong normal) (weak t) (strong t)	*Adjusted*

where b0 – b13 simply refers to the name of the model.

Each prior was set on the individual coefficient between exposure and birthweight. The purpose of including a prior was to limit the possible values of the model coefficient to a certain distribution – depending on the specific prior that we use, we may influence the final coefficient differently. In this work, we used three distributions: the uniform, normal, and Student's t distributions. The parameters of each distribution were tweaked in order to specify our confidence in the prior belief. For example, as the standard deviation parameter for the normal distribution

decreases from models b2 to b4, we are suggesting an increasing level of confidence in the belief that our prior mean on the exposure coefficient is -39 g. This same notion follows for the uniform distributions, where the parameters are lower and upper bounds, respectively, and the Student's t-distributions, where the degrees of freedom represent the strength of the prior.

The primary goal of varying the prior was to see how different prior specifications impact the final coefficient, whether in terms of estimation, confidence, or quality. More specifically, the primary coefficient (exposure within 2 km) and its 95% credible interval were examined for each of the Bayesian models. Model quality was assessed using root mean squared error. The data were split at a 90:10 ratio into the training and testing groups, which were respectively used for model training and predictions. The purpose of this split was to evaluate how each model performs on "new" data, since the quality of fit relative to training data may be more reflective of overfitting as opposed to generalizability.

3.0 Results

3.1 Descriptive Statistics

3.1.1 Demographics

Table 3 outlines the demographic breakdown of the cohort, both overall and by exposure categorization. All variables listed in the table are included as covariates in the adjusted regression models. Many of the more disadvantaged factor levels, including those for maternal education, WIC status, and community SES index, were present at higher proportions in the exposed group. Higher values of the community SES index indicate greater socioeconomic deprivation.

Table 3: Participant Demographics

	No (N=177696)	Yes (N=5746)	Overall (N=183442)
Gestational age (weeks)			
Mean (SD)	38.7 (1.70)	38.7 (1.74)	38.7 (1.70)
Median [Min, Max]	39.0 [22.0, 41.0]	39.0 [23.0, 41.0]	39.0 [22.0, 41.0]
Neonate sex			
Female	86860 (48.9%)	2802 (48.8%)	89662 (48.9%)
Male	90836 (51.1%)	2944 (51.2%)	93780 (51.1%)
APNCU index (collapsed)			
Adequate	101686 (57.2%)	3207 (55.8%)	104893 (57.2%)
Adequate plus	30830 (17.3%)	1012 (17.6%)	31842 (17.4%)
Inadequate or unknown	25718 (14.5%)	867 (15.1%)	26585 (14.5%)
Intermediate	19462 (11.0%)	660 (11.5%)	20122 (11.0%)
Maternal age (years)			
Mean (SD)	29.1 (5.53)	27.8 (5.49)	29.1 (5.53)
Median [Min, Max]	29.0 [13.0, 59.0]	28.0 [13.0, 52.0]	29.0 [13.0, 59.0]
Missing	58 (0.0%)	3 (0.1%)	61 (0.0%)

	No (N=177696)	Yes (N=5746)	Overall (N=183442)
Maternal race			
All other races	7080 (4.0%)	200 (3.5%)	7280 (4.0%)
Black or African American	13782 (7.8%)	1035 (18.0%)	14817 (8.1%)
Unknown or refused	1377 (0.8%)	67 (1.2%)	1444 (0.8%)
White	155457 (87.5%)	4444 (77.3%)	159901 (87.2%)
Maternal education (collapsed)			
Less than high school	12338 (6.9%)	651 (11.3%)	12989 (7.1%)
High school or GED	38025 (21.4%)	1740 (30.3%)	39765 (21.7%)
Some college	48433 (27.3%)	1715 (29.8%)	50148 (27.3%)
Bachelor's or graduate degree	77973 (43.9%)	1604 (27.9%)	79577 (43.4%)
Unknown	927 (0.5%)	36 (0.6%)	963 (0.5%)
Received WIC			
No	126682 (71.3%)	3303 (57.5%)	129985 (70.9%)
Unknown or not classifiable	2700 (1.5%)	79 (1.4%)	2779 (1.5%)
Yes	48314 (27.2%)	2364 (41.1%)	50678 (27.6%)
Maternal BMI			
Normal	68231 (38.4%)	1956 (34.0%)	70187 (38.3%)
Obese	33203 (18.7%)	1215 (21.1%)	34418 (18.8%)
Overweight	33029 (18.6%)	1086 (18.9%)	34115 (18.6%)
Underweight	4641 (2.6%)	161 (2.8%)	4802 (2.6%)
Unknown	38592 (21.7%)	1328 (23.1%)	39920 (21.8%)
Gestational diabetes			
No	168431 (94.8%)	5451 (94.9%)	173882 (94.8%)
Yes	9265 (5.2%)	295 (5.1%)	9560 (5.2%)
Nulliparous			
No	103227 (58.1%)	3453 (60.1%)	106680 (58.2%)
Unknown	260 (0.1%)	11 (0.2%)	271 (0.1%)
Yes	74209 (41.8%)	2282 (39.7%)	76491 (41.7%)
Smoking status			
No	141046 (79.4%)	4160 (72.4%)	145206 (79.2%)
Unknown	1704 (1.0%)	61 (1.1%)	1765 (1.0%)
Yes	34946 (19.7%)	1525 (26.5%)	36471 (19.9%)
Community SES index (quartile)			
Q1	77132 (43.4%)	1201 (20.9%)	78333 (42.7%)

	No (N=177696)	Yes (N=5746)	Overall (N=183442)
Q2	38761 (21.8%)	1481 (25.8%)	40242 (21.9%)
Q3	30754 (17.3%)	663 (11.5%)	31417 (17.1%)
Q4	31049 (17.5%)	2401 (41.8%)	33450 (18.2%)

3.1.2 Distributions

Figure 3 illustrates the distribution of maternal residences (blue circles) and waste facilities (red pins) used in this analysis. As shown in both Figure 3 and Table 3, a considerable amount of the maternal residences fall within two kilometers of a facility accepting oil and gas waste (N = 5746; 3.13%).

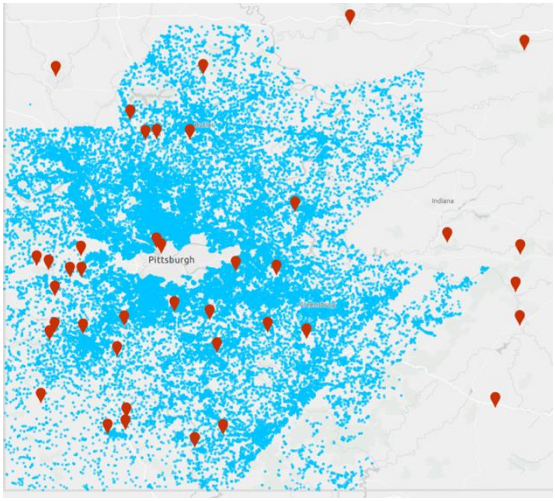


Figure 3: Geospatial Distribution of Maternal Residences (Blue) and Waste Facilities (Red)

As shown in Figure 4, the response variable (birthweight) followed a relatively normal distribution for both exposure categorizations.

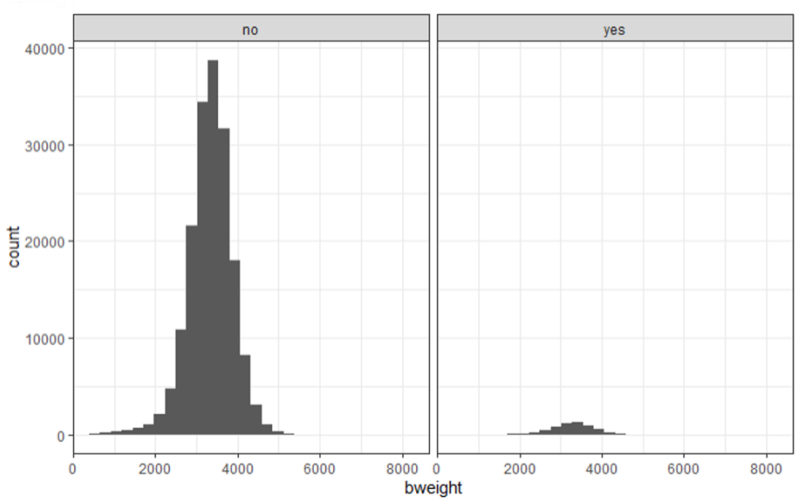


Figure 4: Distribution of Birthweight by Exposure Status

3.2 Linear Models

The results of univariate analysis indicate that mothers with homes within two kilometers of facilities accepting oil and gas waste had babies with a significantly lower average birthweight (-56.87 g) relative to those outside of the buffer zone. The coefficient estimates, confidence intervals, and p-values for the univariate model are summarized in Table 4.

Table 4: Univariate Model Results for Exposure-Birthweight Relationship

Coefficient	Estimate	95% CI[1]	P
Exposed: 2 km (Reference = No)			
Yes	-56.87	-71.82, -41.93	<0.001

Coefficient	Estimate	95% CI[1]	P
Intercept	3336.28	3333.63, 3338.92	<0.001

[1] CI = Confidence Interval

As shown in Table 5, exposure within two kilometers remained significant after covariate adjustment. However, the exposed group had a less drastic lower average birthweight (-13.34 g) relative to the unexposed group following adjustment.

Table 5: Adjusted Model Results for Exposure-Birthweight Relationship

Coefficient	Estimate	95% CI[1]	P
Exposed: 2 km (Reference = No)			
Yes	-13.34	-25.02, -1.65	0.0253
Neonate Sex (Reference = Female)			
Male	135.62	131.58, 139.66	<0.001
APNCU Index, Collapsed (Reference = Adequate)			
Inadequate or Unknown	-34.53	-40.71, -28.35	<0.001
Intermediate	-23.88	-30.71, -17.05	<0.001
Adequate Plus	2.43	-3.43, 8.3	0.4161
Maternal Race (Reference = White)			
Black or African American	-140.67	-148.63, -132.71	<0.001
All other races	-112.27	-122.74, -101.79	<0.001

Coefficient	Estimate	95% CI[1]	P
Unknown or Refused	-25.63	-48.97, -2.29	0.0314
Maternal Education, Collapsed (Reference = Bachelor's or Graduate Degree)			
Less than High School	-87.19	-97.1, -77.27	<0.001
High School or GED	-57.82	-64.49, -51.15	<0.001
Some College	-22.12	-27.65, -16.58	<0.001
Unknown	-7.05	-35.97, 21.87	0.6329
Received WIC (Reference = No)			
Yes	-11.38	-16.81, -5.96	<0.001
Unknown or Not Classifiable	-27.59	-44.52, -10.66	0.0014
Maternal BMI (Reference = Normal)			
Underweight	-120.19	-133.13, -107.25	<0.001
Overweight	78.98	73.24, 84.72	<0.001
Obese	124.58	118.75, 130.4	<0.001
Unknown	26.32	20.56, 32.08	<0.001
Gestational Diabetes (Reference = No)			
Yes	76.59	67.36, 85.81	<0.001
Nulliparous (Reference = No)			
Yes	-127.11	-131.51, -122.7	<0.001

Coefficient	Estimate	95% CI[1]	P
Unknown	-94.3	-147.44, -41.17	<0.001
Smoking, Gestation and Three Months Prior (Reference = No)			
Yes	-151.32	-156.99, -145.65	<0.001
Unknown	-42.52	-63.33, -21.71	<0.001
Community SES Index, Quartile (Reference = Q1)			
Q2	-10.46	-15.98, -4.95	<0.001
Q3	-10.69	-16.83, -4.56	<0.001
Q4	-31.28	-37.89, -24.67	<0.001
Gestational Age (Weeks)	181.13	179.9, 182.36	<0.001
Maternal Age (Years)	-1.44	-1.9, -0.98	<0.001
Intercept	3366.2	3360.36, 3372.05	<0.001

[1] CI = Confidence Interval

3.3 Bayesian Models

Three types of prior distributions were used in the Bayesian analysis: uniform, normal, and Student's t. Although each prior distribution had the same mean (-39 g), the strength of the distribution was varied by altering other distribution-specific parameters. For the uniform distribution, confidence in the prior belief was represented by the width of the upper and lower bounds (where "no prior" can be thought of as a uniform prior with infinite bounds). For the normal

distribution, decreasing standard deviation suggests increasing confidence in the mean. Finally, for the Student's t distribution, increasing degrees of freedom results in increased confidence.

The results for the univariate Bayesian models (named b0 – b6) are summarized in Table 6 and illustrated in Figure 5. In general, as the strength of the prior increased, the primary model coefficient (exposure within two kilometers) was increasingly pulled toward the prior mean, and the range of the corresponding credible interval decreased. The red line in Figure 5 represents the MLE estimate of the primary coefficient, as calculated through univariate linear regression. It is worth mentioning that b1 (prior = uniform[-40, -38]) could not be fit, since brm() does not allow the user to specify hard-set bounds on a continuous parameter.

Table 6: Univariate Bayesian Model Results

	Prior	Coefficient	95% CI[1]
b0	none	-56.84	-71.55, -42.11
b2	normal(-39, 25)	-55.2	-69.75, -41.29
b3	normal(-39, 5)	-44.38	-52.58, -35.9
b4	normal(-39, 1)	-39.31	-41.25, -37.43
b5	student_t(1, -39, 1)	-43.5	-61.37, -37.11
b6	student_t(50, -39, 1)	-39.35	-41.29, -37.42

[1] CI = Credible Interval

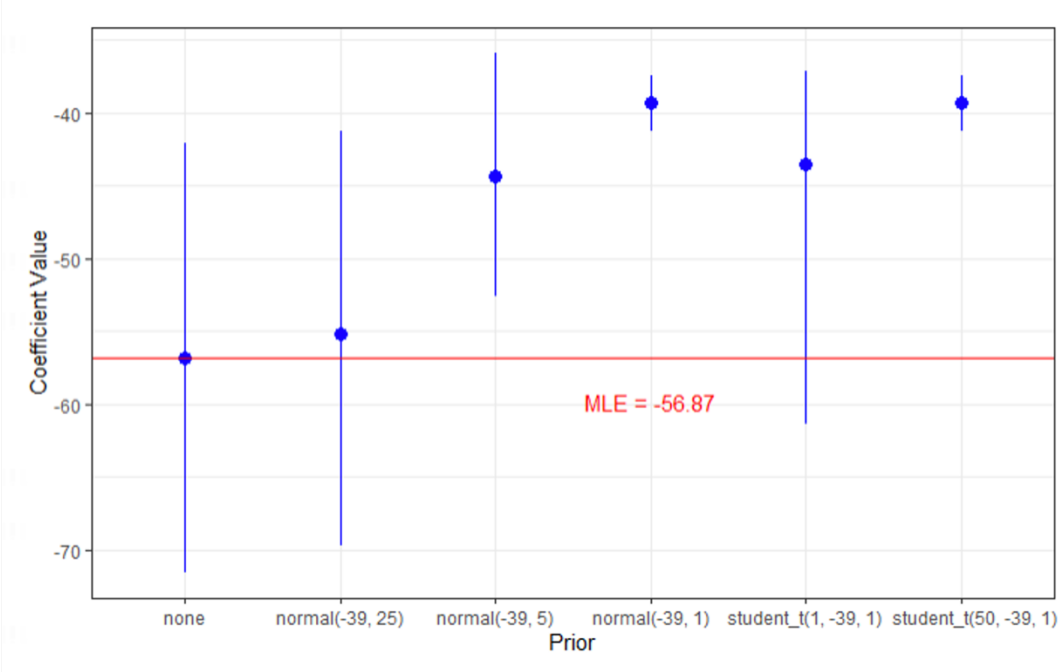


Figure 5: Univariate Bayesian Coefficient Comparison

As summarized in Table 7 and visualized in Figure 6, the results for the adjusted Bayesian models were consistent with those of the univariate models. The red line in Figure 6 represents the MLE estimate of the primary coefficient, as calculated through covariate-adjusted linear regression. Model b7 (prior = uniform[-40, -38]) could not be fit for the same reason as above.

Table 7: Adjusted Bayesian Model Results

	Prior	Coefficient	95% CI[1]
b7	none	-13.23	-24.49, -1.84
b9	normal(-39, 25)	-14.7	-25.92, -3.14
b10	normal(-39, 5)	-28.42	-35.89, -20.82

	Prior	Coefficient	95% CI[1]
b11	normal(-39, 1)	-38.29	-40.26, -36.38
b12	student_t(1, -39, 1)	-17.09	-30.84, -4.01
b13	student_t(50, -39, 1)	-38.27	-40.28, -36.23

[1] CI = Confidence Interval

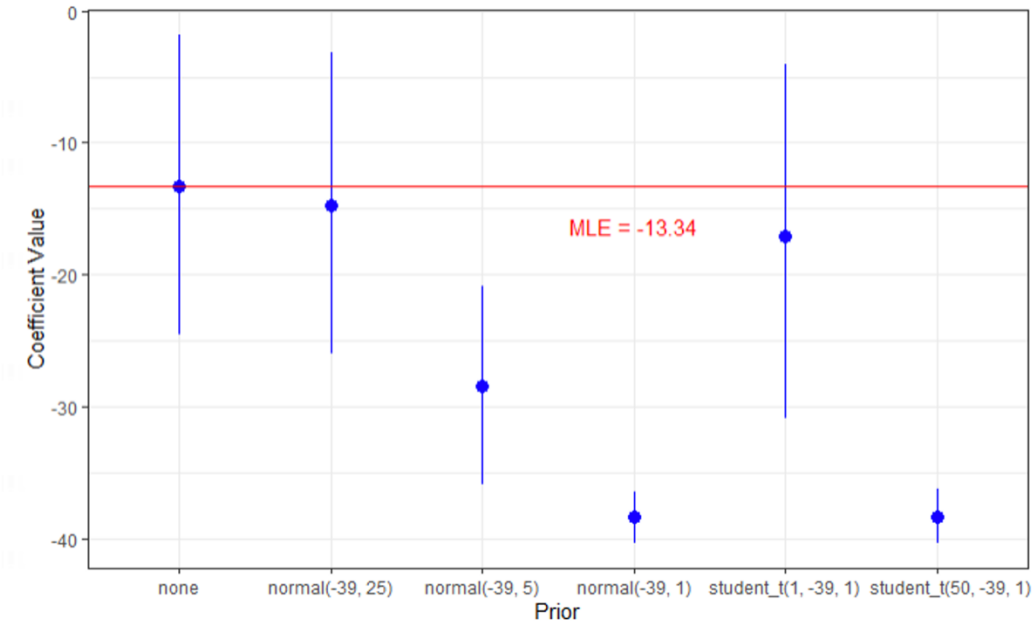


Figure 6: Adjusted Bayesian Coefficient Comparison

3.4 Model Evaluation

Finally, all models were evaluated by calculating root mean squared error (RMSE) on the test set. Table 8 summarizes the results of this analysis. None of the univariate Bayesian models

performed better than the unadjusted model fit using standard linear regression. For the adjusted models, b9 – b13 all performed slightly better than m2.

Table 8: Test Set RMSE

Model	Prior	RMSE
m1	NA	538.6731
m2	NA	420.1538
b0	none	538.6734
b2	normal(-39, 25)	538.6762
b3	normal(-39, 5)	538.7005
b4	normal(-39, 1)	538.7140
b5	student_t(1, -39, 5)	538.6876
b6	student_t(1, -39, 1)	538.7029
b7	none	420.1541
b9	normal(-39, 25)	420.1525
b10	normal(-39, 5)	420.1446
b11	normal(-39, 1)	420.1467
b12	student_t(1, -39, 5)	420.1516

Model	Prior	RMSE
b13	student_t(1, -39, 1)	420.1499

4.0 Discussion and Conclusion

4.1 Linear Model Results

The predictor of interest – residential exposure to facilities accepting oil and gas waste within two kilometers – was found to be significant in both the unadjusted and adjusted models ($P < 0.001$ and $P = 0.0253$, respectively). The relationship between exposure and birthweight was stronger for the unadjusted model compared to the adjusted version (-56.87 g vs. -13.34 g, respectively). This is most likely because the adjusted case accounted for confounding by the covariates.

Although exposure was determined to be significant, this may be a case where statistical and clinical significance are discrepant. We may consider the adjusted coefficient to be more representative of the relationship between exposure and birthweight – not only is this coefficient's magnitude extremely slight relative to the scale of birthweight (i.e., a difference of 13 for a variable with a mean of approximately 3500), but the 95% confidence interval for the adjusted coefficient was also very close to zero (95% CI [-25.02, -1.65]). These contextual clues indicate that we should be careful about making definitive conclusions based on significance alone.

On another note, every covariate was found to be highly significant ($P < 0.001$) in some regard (Table 5). The only non-significant predictors in the adjusted model were certain levels of factorized covariates (ex: the unknown level of collapsed maternal education, $P = 0.6329$). These results indicate that each of the chosen covariates are important to adjust for in the relationship between exposure to facilities accepting oil and gas waste and birthweight. It is also worth mentioning that the exposed group was more likely to have a lower level of maternal education,

receive WIC, and have a more deprived community SES index relative to the unexposed group. This may indicate that facilities accepting oil and gas waste are more likely to be located near relatively disadvantaged neighborhoods.

4.2 Bayesian Model Results

As evident from Tables 4 and 6, as well as Figure 5, the univariate exposure coefficient was increasingly pulled towards the prior mean (-39 g) as the strength of the prior distribution increased. For the case of no prior (b0), the univariate exposure coefficient was approximately equivalent to its standard linear regression counterpart. This makes sense, considering we are not specifying any prior belief on the given relationship! For the three normal priors (b2 – b4), the strength of the prior distribution, as represented through the prior's standard deviation, had an obvious impact on the exposure coefficient. The exposure coefficient was increasingly pulled away from its standard regression counterpart and towards the prior mean as the prior standard deviation decreased. Furthermore, the posterior 95% credible intervals became narrower as the prior standard deviation decreased. These findings are extremely logical, considering that the prior standard deviation conceptually resembles the confidence we have in our prior belief. Finally, the two student t priors (b5 and b6) showed a similar trend to the normal priors. Instead of standard deviation, prior strength is represented by the degrees of freedom (i.e., first argument of student_t).

Tables 5 and 7, as well as Figure 6, indicate that the results for adjusted models were very similar to those of the univariate models. Once again, no prior specification (b7) resulted in an exposure coefficient that approximated the MLE. For the normal priors (b9 – b11), as the standard deviation of the prior distribution decreased, the posterior exposure coefficient pulled away from

its MLE counterpart and towards the prior mean. The posterior 95% credible intervals for these coefficients also became increasingly tighter as the prior standard deviation decreased. Finally, the Student's t priors (b12 and b13) showed the same general trend as in the univariate case.

4.3 Model Evaluation

Test set RMSE was used to evaluate model performance. As shown in Table 8, all of the models performed relatively well on out of sample data – the highest RMSE value was slightly over a half of a kilogram, which suggests that the models were fairly good at predicting birthweight. For the univariate batch, none of the Bayesian models were able to outperform the standard regression model. In fact, as the exposure coefficient was increasingly pulled away from the MLE and towards the prior mean, the models tended to perform slightly more poorly. This is intuitive, considering that we are externally influencing the model away from the data as we increase our strength on the prior. Therefore, we would expect models more consistent with the prior to be less representative of the current dataset (for more on this thought, see the limitations section).

As for the adjusted set of models, RMSE surprisingly *decreased* as the strength of the prior increased. This is opposite to the trend observed in the univariate case. There could be a range of reasons that explain this finding – for example, perhaps this is only because the RMSE differences were extremely slight, making the trends very sensitive to the randomness in the data. From a more optimistic point of view, perhaps setting the prior allowed the results of the model to be more generalizable to "new" (out of sample) data. In either case, it is somewhat mysterious why the trends of the univariate and adjusted models were found to be opposing.

Overall, it is important to emphasize that because the differences in RMSE between the models were very small, it is difficult to draw definitive conclusions from these findings.

4.4 Limitations

There are many limitations to this report worth discussing. First and foremost, the definition of the exposure variable should be critiqued. Originally, I would have preferred to use a one-kilometer buffer zone for exposure, since this would be consistent with the prior as defined by the work of Currie et al. (2017). However, only 878 participants (of the total 183442) fell within one kilometer of a facility accepting oil and gas waste. In order to improve the ratio between exposed and unexposed, a two-kilometer buffer zone was selected. Furthermore, this definition of exposure only classifies in a binary fashion, thus failing to account for any differences in distance to the nearest facility accepting fracking waste. Perhaps future work should investigate how the relationships and models change when using nearest distance as the primary predictor.

Next, the practicality of applying Bayesian regression in this context is questionable. As evident from this work, Bayesian regression can strongly manipulate the relationships that we observe in our data. This may be useful when we have limited sample size, or strong evidence that a given belief is valid. However, in this situation, we have a relatively large dataset that could stand on its own in statistical analysis, and relatively sparse evidence to support our prior belief. The primary purpose of employing Bayesian regression in this project was to explore the statistical impacts of using varied priors, as opposed to drawing concrete conclusions about the relationships in the Bayesian models. Nevertheless, it was interesting to see the improved RMSE values for the adjusted Bayesian models.

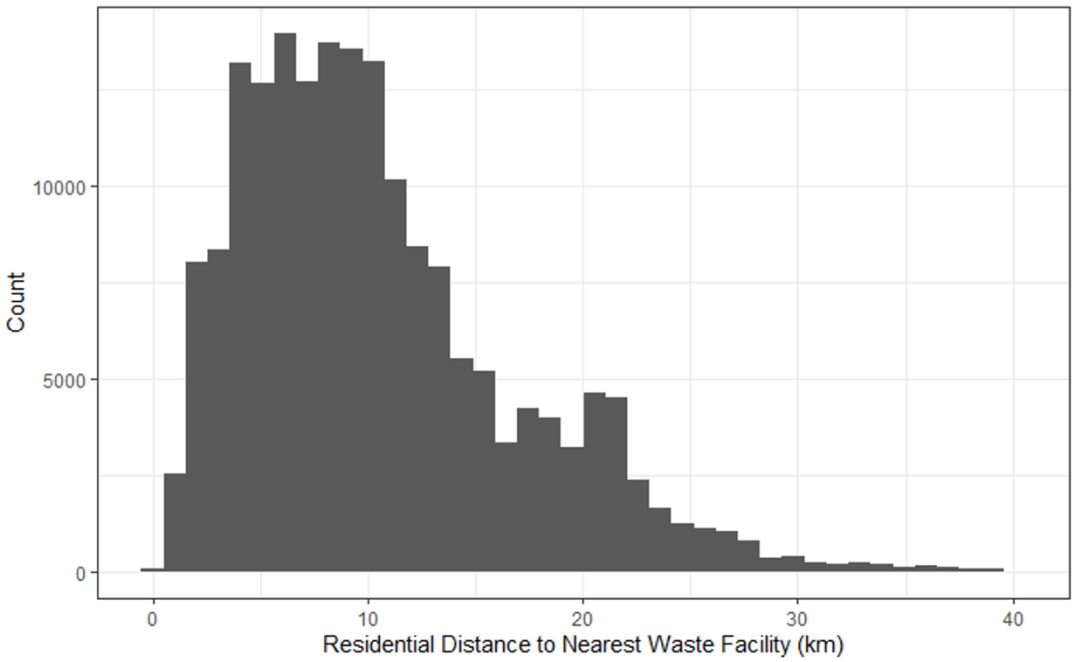
Finally, RMSE proved to be a relatively poor metric for evaluating the data. The differences in RMSE across the models were extremely slight, making it difficult to say if our trends will hold given a replicated analysis. It is likely that the RMSE values differed so slightly since the primary predictor of interest that changed across the models was binary, with a small coefficient magnitude relative to the scale of the outcome. In other words, if the difference in coefficients was ~10 between two models, and the true outcome is in the range of 3500, our prediction error will not differ very much between the two models.

4.5 Conclusion

Overall, this thesis presents an association between residential exposure to facilities accepting oil and gas waste and reduced birthweight. Considering the small nature of this association, perhaps this is a case of statistical, versus clinical or practical, significance. Future work should consider different definitions of exposure in order to better understand this relationship. Furthermore, we observed that the prior has a noticeable impact on Bayesian regression coefficients. Interestingly, the adjusted Bayesian models had better test set RMSE values relative to the adjusted model fit using linear regression – this trend was opposite for the set of univariate models.

Appendix A Distance Metric and Buffer Zones

As hinted at in the limitations section, a considerable amount of analysis and thought went into defining the exposure variable. In order to define exposure within two kilometers, or within any buffer zone distance, the distance matrix between maternal residences and waste facilities had to first be calculated. From this distance matrix, each mother was assigned a "minimum distance," which corresponded to the distance (in kilometers) between the mother's residence and the closest waste facility. Appendix Figure 1 illustrates the distribution of minimum distances across all observations.



Appendix Figure 1: Distribution of Minimum Distance

Once each observation was assigned a minimum distance, the exposure variables were easily defined. A total of four exposure variables were created, corresponding to the following four buffer zones: one kilometer, two kilometers, three kilometers, and five kilometers. Appendix Table 1 summarizes the exposed counts for each of these variables. Ultimately, the exposure variable with the two-kilometer buffer zone was chosen as a tradeoff between exposure ratio and the previous literature, which suggests that exposure outside of three kilometers has little impact on birthweight.

Appendix Table 1: Exposure Count and Percent by Buffer Zone

Buffer Zone (km)	Exposure Count	Percent Exposed[1]
1	878	0.48
2	5746	3.13
3	14225	7.75
5	36996	20.17

[1] N = 183,442

Appendix B Code

This section includes all R code used in this thesis project. This includes three main files: one used to generate the data and variables, one to create the descriptive statistics and figures, and one to run the final analyses. Each of the following subsections correspond to one of these files.

Appendix B.1 Processing Data and Creating Variables

````
---
title: "Capstone Data"
output: html_document
date: '2022-09-25'
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

# Package and Data Loading

```{r packages, echo = FALSE}
library(geosphere)
library(dplyr)
```

```{r load birthData, echo=FALSE}
birthData <- read.csv("D:\\Nick\\Capstone\\clean_birth_data.csv")
```

```{r load waste facility data, echo = FALSE}
wasteData <- read.csv("D:\\Data\\Raw\\DEP O&G Waste Facilities\\OGRE_Waste_Facilities.csv")
```

# Data Cleaning
````

Filtering Columns

```{r birthData columns, echo = FALSE}
birthData <- birthData %>%

  ## subset dataset to relevant variables
  select(
    ## birth info + outcome
    birth_id, bweight,
    ## mother's residence (location)
    arcgis_lat, arcgis_long,
    ## covariates
    gestational_age_weeks, neonate_sex, apncu_index_collapsed, nulliparous, maternal_age_years, maternal_race,
    maternal_edu_cat_collapsed, received_wic, maternal_bmi_cat, gestational_diabetes,
    overall_smoking_gestation_and_three_months_prior, community_ses_index_quartile)
```

```{r wasteData columns, echo = FALSE}
wasteData <- wasteData %>%
  rename(name = ï..WASTE_FACILITY) %>%
  select(name,  DISPOSITION_METHOD,  FACILITY_ADDRESS1,  FACILITY_ADDRESS2, FACILITY_STATE,
         FACILITY_ZIP, FACILITY_LATITUDE, FACILITY_LONGITUDE)
```

Filtering Rows

```{r birthData Missing Values, echo = FALSE}
n_missing_bw <- length(birthData$bweight[(birthData$bweight == 9999)])

birthData <- birthData[!(birthData$bweight == 9999), ]

print(paste0("Number of missing birthweight records: ", n_missing_bw))
```

```{r birthData Under 500g, echo = FALSE}
n_small_bw <- length(birthData$bweight[(birthData$bweight < 500)])

birthData <- birthData[!(birthData$bweight == 500), ]

print(paste0("Number of birthweight records under 500 g: ", n_small_bw))
```

```{r wasteData Disposition Method, echo = FALSE}
## filter waste facilities to only include those with disposition methods of interest
```

```
wasteData <- wasteData[which(wasteData$DISPOSITION_METHOD %in% c('LANDFILL', 'PUBLIC SEWAGE TREATMENT PLANT',
                                                 'RESIDUAL WASTE PROCESSING FACILITY')), ]

wasteData <- wasteData[which(wasteData$FACILITY_STATE == 'PA'), ]
```

Calculating Metrics

Gestational Age Bins

```{r Gestational Age Bins, echo = FALSE}
birthData <- birthData %>% mutate(gest_bin = cut(gestational_age_weeks, breaks = 3))
```

Residential Proximity (Distance Matrix) + Minimum Distance

```{r Distance Metric, echo = FALSE}
d <- distm(cbind(birthData$arcgis_long, birthData$arcgis_lat),
         cbind(wasteData$FACILITY_LONGITUDE, wasteData$FACILITY_LATITUDE),
         fun = distGeo) / 1000
  ## distance is originally in meters, /1000 converts to km

colnames(d) <- paste0('waste', 1:nrow(wasteData))
d <- data.frame(d)

minDist <- do.call(pmin, d)
```

Buffer Zone Counts

```{r Buffer Zone: 1 kilometer, echo = FALSE}
d1 <- d

d1[d1 <= 1] <- 1
d1[d1 > 1] <- 0

d1$count <- apply(d1, 1, sum)
d1$exposed <- case_when((d1$count == 0) ~ 'no', TRUE ~ 'yes')

table(d1$count)
table(d1$exposed)
```

```{r Buffer Zone: 2 kilometers, echo = FALSE}
d2 <- d


```

```
d2[d2 <= 2] <- 1
d2[d2 > 2] <- 0

d2$count <- apply(d2, 1, sum)
d2$exposed <- case_when((d2$count == 0) ~ 'no', TRUE ~ 'yes')

table(d2$count)
table(d2$exposed)
```

```{r Buffer Zone: 3 kilometers, echo = FALSE}
d3 <- d

d3[d3 <= 3] <- 1
d3[d3 > 3] <- 0

d3$count <- apply(d3, 1, sum)
d3$exposed <- case_when((d3$count == 0) ~ 'no', TRUE ~ 'yes')

table(d3$count)
table(d3$exposed)
```

```{r Buffer Zone: 5 kilometers, echo = FALSE}
d5 <- d

d5[d5 <= 5] <- 1
d5[d5 > 5] <- 0

d5$count <- apply(d5, 1, sum)
d5$exposed <- case_when((d5$count == 0) ~ 'no', TRUE ~ 'yes')

table(d5$count)
table(d5$exposed)
```

Creating Final Table

```{r Final Table Merging and Naming, echo = FALSE}
finalData <- cbind(birthData, minDist,
                   d1$count, d2$count, d3$count, d5$count,
                   d1$exposed, d2$exposed, d3$exposed, d5$exposed)
finalData <- rename(finalData, c('count_1_km' = 'd1$count',
                                 'count_2_km' = 'd2$count',
                                 'count_3_km' = 'd3$count',
```

```
                        'count_5_km' = 'd5$count',
                        'exposed_1_km' = 'd1$exposed',
                        'exposed_2_km' = 'd2$exposed',
                        'exposed_3_km' = 'd3$exposed',
                        'exposed_5_km' = 'd5$exposed'))
```

Data Export

```{r Export, echo = FALSE}
write.csv(finalData, "D:\\Nick\\Capstone\\capstoneData.csv")
```

Appendix B.2 Descriptive Statistics and Figures

```
---
title: "Capstone Descriptives"
output:
  html_document: default
  word_document: default
date: "2022-09-25"
---
```

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Package and Data Loading

```{r package loading}
library(ggplot2)
library(kableExtra)
library(table1)
library(dplyr)
library(flextable)
library(gt)
```

```{r data loading}
capData <- read.csv("D:\\Nick\\Capstone\\capstoneData.csv")
wasteData <- read.csv("D:\\Data\\Raw\\DEP O&G Waste Facilities\\OGRE_Waste_Facilities.csv")
```

```

# Data Cleaning

```{r wasteData Cleaning}
wasteData <- wasteData %>%
  rename(name = ï..WASTE_FACILITY) %>%
  select(name, DISPOSITION_METHOD, FACILITY_ADDRESS1, FACILITY_ADDRESS2,
FACILITY_STATE,
         FACILITY_ZIP, FACILITY_LATITUDE, FACILITY_LONGITUDE)

wasteData <- wasteData[which(wasteData$DISPOSITION_METHOD %in% c('LANDFILL',
'PUBLIC SEWAGE TREATMENT PLANT',
                                                 'RESIDUAL WASTE PROCESSING FACILITY')), ]

wasteData <- wasteData[which(wasteData$FACILITY_STATE == 'PA'), ]
```

```{r map wasteData Cleaning}
mapData <- wasteData %>% filter(FACILITY_LATITUDE < 41.2, FACILITY_LONGITUDE <
-78.7)
```

```{r mapData export}
write.csv(mapData, "D:\\Nick\\Capstone\\mapWasteData.csv")
```

Descriptive Stats

```{r birthweight stats}
summary(capData$bweight)
```

```{r distance stats}
summary(capData$minDist)
  ## units: km
```

Plots

```{r histogram of birthweight}
capData %>% ggplot(aes(x = bweight)) +
  geom_histogram() +
  theme_bw()
```
```

```{r histogram of birthweight facetted by exposure status}
capData %>% ggplot(aes(x = bweight)) +
  geom_histogram() +
  facet_wrap(~exposed_2_km, scales = 'free') +
  theme_bw()
```

```{r histogram of minDist}
capData %>% ggplot(aes(x = minDist)) +
  geom_histogram(bins = 40) +
  xlab('Residential Distance to Nearest Waste Facility (km)') +
  ylab('Count') +
  theme_bw()
```

Tables

```{r Exposure Table}
capData$exposed_2_km_LAB <- case_when(capData$exposed_2_km == "no" ~ "No",
                              capData$exposed_2_km == "yes" ~ "Yes",
                              TRUE ~ capData$exposed_2_km)

label(capData$gestational_age_weeks)                    <- "Gestational age"
label(capData$neonate_sex)                              <- "Neonate sex"
label(capData$apncu_index_collapsed)                    <- "APNCU index (collapsed)"
label(capData$maternal_age_years)                       <- "Maternal age"
label(capData$maternal_race)                            <- "Maternal race"
label(capData$maternal_edu_cat_collapsed)               <- "Maternal education (collapsed)"
label(capData$received_wic)                             <- "Received WIC"
label(capData$maternal_bmi_cat)                         <- "Maternal BMI"
label(capData$gestational_diabetes)                     <- "Gestational diabetes"
label(capData$nulliparous)                              <- "Nulliparous"
label(capData$overall_smoking_gestation_and_three_months_prior) <- "Smoking status"
label(capData$community_ses_index_quartile)             <- "Community SES index (quartile)"

units(capData$gestational_age_weeks)      <- "weeks"
units(capData$maternal_age_years)         <- "years"

tb1 <- table1(~ gestational_age_weeks + neonate_sex + apncu_index_collapsed +
            maternal_age_years + maternal_race + maternal_edu_cat_collapsed +
            received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
            overall_smoking_gestation_and_three_months_prior + community_ses_index_quartile
| exposed_2_km_LAB,
          data = capData)

tb1
```

```

```{r save t1}
t1flex(tb1) %>%
  save_as_docx(path="capstone_table1.docx")
```

```{r buffer zone table}
exposed_1km <- as.numeric(table(capData$exposed_1_km)[2])
exposed_2km <- as.numeric(table(capData$exposed_2_km)[2])
exposed_3km <- as.numeric(table(capData$exposed_3_km)[2])
exposed_5km <- as.numeric(table(capData$exposed_5_km)[2])

exposed_df <- data.frame(buffer_zone_km = c(1, 2, 3, 5),
                         exposure_count = c(exposed_1km, exposed_2km, exposed_3km, exposed_5km))
%>%
  mutate(exposure_percent = round((exposure_count / nrow(capData)) * 100, 2))

exposed_df %>%
  gt() %>%
  cols_label(buffer_zone_km = "Buffer Zone (km)",
             exposure_count = "Exposure Count",
             exposure_percent = "Percent Exposed") %>%
  cols_align(
    align = "center",
    columns = everything()) %>%
  tab_options(column_labels.font.weight = 'bold') %>%
  tab_footnote(
       footnote = 'N = 183,442',
       locations = cells_column_labels(columns = exposure_percent)) %>%
    fmt_markdown(columns = everything())
```

Appendix B.3 Final Analysis

---

title: "Capstone Analysis"

output: html_document

date: '2022-10-17'
```

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = TRUE)
```

Package and Data Loading

```{r package loading, echo = FALSE, message = FALSE, warning = FALSE}
library(ggplot2)
library(bayestestR)
library(rstan)
library(brms)
library(caTools)
library(dplyr)
library(kableExtra)
library(gt)
library(tidyverse)
```

```{r data loading, echo = FALSE}
capData <- read.csv("D:\\Nick\\Capstone\\capstoneData.csv")
```

Data Preparation

Format Data

```{r Formatting Data}
capData$neonate_sex = factor(capData$neonate_sex, levels = c('Female', 'Male'))
capData$apncu_index_collapsed = factor(capData$apncu_index_collapsed, levels = c('Adequate', 'Inadequate or unknown',
                                                                                  'Intermediate', 'Adequate plus'))
capData$nulliparous = factor(capData$nulliparous, levels = c('No', 'Yes', 'Unknown'))
capData$maternal_race = factor(capData$maternal_race, levels = c('White', 'Black or African American',
                                                                 'All other races', 'Unknown or refused'))
capData$maternal_edu_cat_collapsed = factor(capData$maternal_edu_cat_collapsed,
                                            levels = c('Bachelor\'s or graduate degree', 'Less than high school',
                                                       'High school or GED', 'Some college', 'Unknown'))
capData$received_wic = factor(capData$received_wic, levels = c('No', 'Yes', 'Unknown or not classifiable'))
capData$maternal_bmi_cat = factor(capData$maternal_bmi_cat, levels = c('Normal', 'Underweight', 'Overweight', 'Obese', 'Unknown'))
capData$gestational_diabetes = factor(capData$gestational_diabetes, levels = c('No', 'Yes'))
```

```
capData$overall_smoking_gestation_and_three_months_prior =
factor(capData$overall_smoking_gestation_and_three_months_prior,
                                        levels = c('No', 'Yes', 'Unknown'))
capData$community_ses_index_quartile = factor(capData$community_ses_index_quartile,
levels = c('Q1', 'Q2', 'Q3', 'Q4'))


## NOTE: have to manually specify factor levels to ensure correct reference (first item in list is reference level)
```

Center Continuous Predictors

```{r Center Predictors}
# center maternal age at the mean
capData$maternal_age_years_c <- capData$maternal_age_years -
mean(capData$maternal_age_years, na.rm = TRUE)

# center gestational age in weeks at the mean
capData$gestational_age_weeks_c <- capData$gestational_age_weeks -
mean(capData$gestational_age_weeks, na.rm = TRUE)
```

Train-Test Split

```{r Train-Test Split}
set.seed(25)

sample <- sample.split(capData$birth_id, SplitRatio = 0.9)

train <- data.frame(subset(capData, sample == TRUE))
test <- data.frame(subset(capData, sample == FALSE))
```

Simple Linear Regression

Checking Assumptions

```{r Checking Regression Assumptions - Linearity}
# The relationship between X and the mean of Y is linear

group_means <- aggregate(train$bweight, by = list(train$exposed_2_km), FUN = mean)

ggplot(data = group_means, aes(x = Group.1, y = x)) + geom_point()
```

Observations are assumed to be independent.

```{r Checking Regression Assumptions - Normality}
# For any fixed value of X, Y is normally distributed

ggplot(data = train, aes(x = bweight)) +
  geom_histogram() +
  facet_wrap(~exposed_2_km, scales = 'free') +
  theme_bw()
```

```{r Checking Regression Assumptions - Equal Variance}
var.test(bweight ~ exposed_2_km, data = train)
```

Fitting Univariate Model

```{r Fitting Simple Univariate Model}
m1 <- lm(data = train,
         formula = bweight ~ exposed_2_km)

summary(m1)
```

Fitting Adjusted Model

```{r Fitting Simple Adjusted Model}
m2 <- lm(data = train,
         formula = bweight ~ exposed_2_km +
           # infant characteristics
           gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
           # maternal characteristics
           maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
           received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
           overall_smoking_gestation_and_three_months_prior +
           # environmental characteristics
           community_ses_index_quartile)

summary(m2)
```

Bayesian Regression Analysis

Fitting Univariate (Unadjusted) Models

```{r Fitting Univariate Bayesian Model with No Prior (Infinitely Diffuse Uniform), message = FALSE, output = FALSE}
b0 <- brm(formula = bweight ~ exposed_2_km,
          data = train,
          file = "b0",
          seed = 123)
```

```{r b0 Results}
summary(b0)
```

```{r Fitting Univariate Bayesian Model with Strong Informative Uniform Prior, message = FALSE, output = FALSE}
strong_informative_uniform_prior <- c(set_prior("uniform(-38, -40)", class = "b", coef = "exposed_2_kmyes"))

b1 <- brm(formula = bweight ~ exposed_2_km,
          data = train,
          file = "b1_updated2",
          prior = strong_informative_uniform_prior,
          seed = 123)
```

```{r b1 Results}
summary(b1)
```

```{r Fitting Univariate Bayesian Model with Uninformative Normal Prior, message = FALSE, output = FALSE}
uninformative_normal_prior <- c(set_prior("normal(-39, 25)", class = "b", coef = "exposed_2_kmyes"))

b2 <- brm(formula = bweight ~ exposed_2_km,
          data = train,
          prior = uninformative_normal_prior,
          file = "b2",
          seed = 123)
```

```{r b2 Results}
summary(b2)
```

```{r Fitting Univariate Bayesian Model with Weak Informative Normal Prior, message = FALSE, output = FALSE}
```

```
weak_informative_normal_prior <- c(set_prior("normal(-39, 5)", class = "b", coef = "exposed_2_kmyes"))

b3 <- brm(formula = bweight ~ exposed_2_km,
          data = train,
          prior = weak_informative_normal_prior,
          file = "b3",
          seed = 123)
```

```{r b3 Results}
summary(b3)
```

```{r Fitting Univariate Bayesian Model with Strong Informative Normal Prior, message = FALSE, output = FALSE}
strong_informative_normal_prior <- c(set_prior("normal(-39, 1)", class = "b", coef = "exposed_2_kmyes"))

b4 <- brm(formula = bweight ~ exposed_2_km,
          data = train,
          prior = strong_informative_normal_prior,
          file = "b4",
```

```
        seed = 123)
```

```{r b4 Results}
summary(b4)
```

```{r Fitting Univariate Bayesian Model with Weak t Prior, message = FALSE, output = FALSE}
weak_t_prior <- c(set_prior("student_t(1, -39, 1)", class = "b", coef = "exposed_2_kmyes"))

b5 <- brm(formula = bweight ~ exposed_2_km,
        data = train,
        prior = weak_t_prior,
        file = "b5.1",
        seed = 123)
```

```{r b5 Results}
summary(b5)
```

```{r Fitting Univariate Bayesian Model with Strong t Prior, message = FALSE, output = FALSE}
strong_t_prior <- c(set_prior("student_t(50, -39, 1)", class = "b", coef = "exposed_2_kmyes"))
```

```
b6 <- brm(formula = bweight ~ exposed_2_km,
          data = train,
          prior = strong_t_prior,
          file = "b6.1",
          seed = 123)
```

```{r b6 Results}
summary(b6)
```

Fitting Covariate-Inclusive (Adjusted) Models

```{r Fitting Adjusted Bayesian Model with No Prior (Infinitely Diffuse Uniform), message = FALSE, output = FALSE}
b7 <- brm(formula = bweight ~ exposed_2_km +
                # infant characteristics
                gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
                # maternal characteristics
                maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
                received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
                overall_smoking_gestation_and_three_months_prior +
```

```
        # environmental characteristics
        community_ses_index_quartile,
    data = train,
    file = "b7",
    seed = 123)
```

```{r b7 Results}
summary(b7)
```

```{r Fitting Adjusted Bayesian Model with Strong Informative Uniform Prior, message = FALSE, output = FALSE, eval = FALSE}
strong_informative_uniform_prior <- c(set_prior("uniform(-38, -40)", class = "b", coef = "exposed_2_kmyes"))

b8 <- brm(formula = bweight ~ exposed_2_km +
        # infant characteristics
        gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
        # maternal characteristics
        maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
        received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
        overall_smoking_gestation_and_three_months_prior +
```

```
        # environmental characteristics
        community_ses_index_quartile,
    data = train,
    prior = strong_informative_uniform_prior,
    file = "b8_updated",
    seed = 123)
```

```{r b8 Results, eval = FALSE}
summary(b8)
```

```{r Fitting Adjusted Bayesian Model with Uninformative Normal Prior, message = FALSE, output = FALSE}
uninformative_normal_prior <- c(set_prior("normal(-39, 25)", class = "b", coef = "exposed_2_kmyes"))

b9 <- brm(formula = bweight ~ exposed_2_km +
        # infant characteristics
        gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
        # maternal characteristics
        maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
        received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
```

```
        overall_smoking_gestation_and_three_months_prior +
        # environmental characteristics
        community_ses_index_quartile,
    data = train,
    prior = uninformative_normal_prior,
    file = "b9",
    seed = 123)
```

```{r b9 Results}
summary(b9)
```

```{r Fitting Adjusted Bayesian Model with Weak Informative Normal Prior, message = FALSE, output = FALSE}
weak_informative_normal_prior <- c(set_prior("normal(-39, 5)", class = "b", coef = "exposed_2_kmyes"))

b10 <- brm(formula = bweight ~ exposed_2_km +
        # infant characteristics
        gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
        # maternal characteristics
        maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
```

```
        received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
        overall_smoking_gestation_and_three_months_prior +
        # environmental characteristics
        community_ses_index_quartile,
    data = train,
    prior = weak_informative_normal_prior,
    file = "b10",
    seed = 123)
```

```{r b10 Results}
summary(b10)
```

```{r Fitting Adjusted Bayesian Model with Strong Informative Normal Prior, message = FALSE, output = FALSE}
strong_informative_normal_prior <- c(set_prior("normal(-39, 1)", class = "b", coef = "exposed_2_kmyes"))

b11 <- brm(formula = bweight ~ exposed_2_km +
        # infant characteristics
        gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
        # maternal characteristics
```

```
        maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
        received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
        overall_smoking_gestation_and_three_months_prior +
        # environmental characteristics
        community_ses_index_quartile,
    data = train,
    prior = strong_informative_normal_prior,
    file = "b11",
    seed = 123)
```

```{r b11 Results}
summary(b11)
```

```{r Fitting Adjusted Bayesian Model with Weak Informative t Prior, message = FALSE, output = FALSE}
weak_t_prior <- c(set_prior("student_t(1, -39, 1)", class = "b", coef = "exposed_2_kmyes"))

b12 <- brm(formula = bweight ~ exposed_2_km +
        # infant characteristics
        gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
        # maternal characteristics
```

```
        maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
        received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
        overall_smoking_gestation_and_three_months_prior +
        # environmental characteristics
        community_ses_index_quartile,
    data = train,
    prior = weak_t_prior,
    file = "b12.1",
    seed = 123)
```

```{r b12 Results}
summary(b12)
```

```{r Fitting Adjusted Bayesian Model with Right Skew Normal Prior, message = FALSE, output = FALSE}
strong_t_prior <- c(set_prior("student_t(50, -39, 1)", class = "b", coef = "exposed_2_kmyes"))

b13 <- brm(formula = bweight ~ exposed_2_km +
        # infant characteristics
        gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
        # maternal characteristics
```

```
        maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
        received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
        overall_smoking_gestation_and_three_months_prior +
        # environmental characteristics
        community_ses_index_quartile,
    data = train,
    prior = strong_t_prior,
    file = "b13.1",
    seed = 123)
```

```{r b13 Results}
summary(b13)
```

Linear Model Summary

```{r linear_model_summary function}
linear_model_summary <- function(model_object){

  ## retrieve number of coefficients for table
  num_coefficients <- length(coef(model_object))
```

```
## retrieve coefficient names
names <- names(model_object$coefficients)

## initialize variables
coef <- rep(NA, num_coefficients)
upper_bound <- rep(NA, num_coefficients)
lower_bound <- rep(NA, num_coefficients)
p_value <- rep(NA, num_coefficients)

## retrieve coefficient estimate, confidence interval, and p values
for(i in 1:num_coefficients){

  coef[i] <- round(coef(model_object)[i], 2)

  lower_bound[i] <- round(confint(object = model_object, parm = names[i], level = 0.95)[1], 2)
  upper_bound[i] <- round(confint(object = model_object, parm = names[i], level = 0.95)[2], 2)

  p_value[i] <- round(summary(model_object)$coefficients[i, 4], 4)

}

## format p values
p <- case_when(p_value == 0 | p_value < 0.001 ~ "<0.001",
```

```
            TRUE ~ as.character(p_value))



 ## create data frame for tabular output
 df <- data.frame(Names = names,
                  Estimate = coef,
                  CI = paste(lower_bound, upper_bound, sep = ", "),
                  "P" = p) %>%
   mutate(Coefficient = case_when(Names == "exposed_2_kmyes" ~ "Yes",
                                  Names == "(Intercept)" ~ "Intercept",
                                  Names == "gestational_age_weeks_c" ~ "Gestational Age (Weeks)",
                                  Names == "neonate_sexMale" ~ "Male",
                                  Names == "apncu_index_collapsedInadequate or unknown" ~ "Inadequate
or Unknown",
                                  Names == "apncu_index_collapsedIntermediate" ~ "Intermediate",
                                  Names == "apncu_index_collapsedAdequate plus" ~ "Adequate Plus",
                                  Names == "maternal_age_years_c" ~ "Maternal Age (Years)",
                                  Names == "maternal_raceBlack or African American" ~ "Black or African
American",
                                  Names == "maternal_raceAll other races" ~ "All other races",
                                  Names == "maternal_raceUnknown or refused" ~ "Unknown or Refused",
                                  Names == "maternal_edu_cat_collapsedLess than high school" ~ "Less than
High School",
```

```
                Names == "maternal_edu_cat_collapsedHigh school or GED" ~ "High
School or GED",
                Names == "maternal_edu_cat_collapsedSome college" ~ "Some College",
                Names == "maternal_edu_cat_collapsedUnknown" ~ "Unknown",
                Names == "received_wicYes" ~ "Yes",
                Names == "received_wicUnknown or not classifiable" ~ "Unknown or Not
Classifiable",
                Names == "maternal_bmi_catUnderweight" ~ "Underweight",
                Names == "maternal_bmi_catOverweight" ~ "Overweight",
                Names == "maternal_bmi_catObese" ~ "Obese",
                Names == "maternal_bmi_catUnknown" ~ "Unknown",
                Names == "gestational_diabetesYes" ~ "Yes",
                Names == "nulliparousYes" ~ "Yes",
                Names == "nulliparousUnknown" ~ "Unknown",
                Names == "overall_smoking_gestation_and_three_months_priorYes" ~
"Yes",
                Names == "overall_smoking_gestation_and_three_months_priorUnknown"
~ "Unknown",
                Names == "community_ses_index_quartileQ2" ~ "Q2",
                Names == "community_ses_index_quartileQ3" ~ "Q3",
                Names == "community_ses_index_quartileQ4" ~ "Q4",
                TRUE ~ Names))
```

```
## sort order of data frame for table
df <- df[, c(5, 1, 2, 3, 4)]
df <- df[c(2:nrow(df), 1), ]

## final table
gt <- df %>%
  gt() %>%
  cols_hide(columns = Names) %>%
  cols_label(CI = '95% CI') %>%
  tab_options(column_labels.font.weight = 'bold') %>%
  tab_row_group(
    label = "Community SES Index, Quartile (Reference = Q1)",
    rows = str_detect(Names, "community_ses_index_quartile")) %>%
  tab_row_group(
    label = "Smoking, Gestation and Three Months Prior (Reference = No)",
    rows = str_detect(Names, "overall_smoking_gestation_and_three_months_prior")) %>%
  tab_row_group(
    label = "Nulliparous (Reference = No)",
    rows = str_detect(Names, "nulliparous")) %>%
  tab_row_group(
    label = "Gestational Diabetes (Reference = No)",
    rows = str_detect(Names, "gestational_diabetes")) %>%
  tab_row_group(
```

```
  label = "Maternal BMI (Reference = Normal)",
  rows = str_detect(Names, "maternal_bmi_cat")) %>%
tab_row_group(
  label = "Received WIC (Reference = No)",
  rows = str_detect(Names, "received_wic")) %>%
tab_row_group(
  label = "Maternal Education, Collapsed (Reference = Bachelor's or Graduate Degree)",
  rows = str_detect(Names, "maternal_edu_cat_collapsed")) %>%
tab_row_group(
  label = "Maternal Race (Reference = White)",
  rows = str_detect(Names, "maternal_race")) %>%
tab_row_group(
  label = "APNCU Index, Collapsed (Reference = Adequate)",
  rows = str_detect(Names, "apncu_index_collapsed")) %>%
tab_row_group(
  label = "Neonate Sex (Reference = Female)",
  rows = str_detect(Names, "neonate_sex")) %>%
tab_row_group(
  label = "Exposed: 2 km (Reference = No)",
  rows = str_detect(Names, "exposed")) %>%
tab_footnote(
  footnote = 'CI = Confidence Interval',
  locations = cells_column_labels(columns = CI)) %>%
```

```
    cols_align(
      align = "center",
      columns = c(Estimate, CI, P)) %>%
    tab_style(
      style = cell_text(align = "center"),
      locations = cells_column_labels()) %>%
    tab_style(
      style = cell_text(align = "left", indent = px(20)),
      locations = cells_body(columns = Coefficient,
                             rows = !(Names %in% c("(Intercept)", "gestational_age_weeks_c",
"maternal_age_years_c")))) %>%
    tab_style(
      style = "padding-right:175px;",
      locations = cells_body(columns = Coefficient, rows = everything())) %>%
    fmt_markdown(columns = everything())

  ## return table
  gt

}
```

```{r m1 summary}
```

```
linear_model_summary(m1)
```

```{r m2 summary}
linear_model_summary(m2)
```

Coefficient Summary

```{r coef_summary}
coef_summary <- function(brms_mod_list){

  ## retrieve number of models for visualization
  num_mods <- length(brms_mod_list)

  ## initialize variables
  coef <- rep(NA, num_mods)
  upper_bound <- rep(NA, num_mods)
  lower_bound <- rep(NA, num_mods)

  ## retrieve coefficient estimate and upper/lower credible interval bounds for each model
  for(i in 1:num_mods){
```

```
  temp_mod <- eval(as.name(brms_mod_list[i]))

  coef[i] <- round(fixef(temp_mod)[2, 1], 2)
  upper_bound[i] <- round(fixef(temp_mod)[2, 3], 2)
  lower_bound[i] <- round(fixef(temp_mod)[2, 4], 2)

}

## define prior (brms model list must be input in same order and with same length as priors)
prior <- c("none", "normal(-39, 25)", "normal(-39, 5)",
           "normal(-39, 1)", "student_t(1, -39, 1)", "student_t(50, -39, 1)")

## create df for table
df <- data.frame(model = brms_mod_list, Prior = prior,
                 Coefficient = coef, CI = paste(upper_bound, lower_bound, sep = ", "))

## final table
df %>%
  gt(rowname_col = 'model') %>%
  cols_label(CI = '95% CI') %>%
  cols_align(
    align = "center",
    columns = everything()) %>%
```

```
    tab_options(column_labels.font.weight = 'bold') %>%
      tab_footnote(
        footnote = 'CI = Credibility Interval',
        locations = cells_column_labels(columns = CI)) %>%
    fmt_markdown(columns = everything())

}
```

```{r univariate coefficient summary}
coef_summary(c('b0', 'b2', 'b3', 'b4', 'b5', 'b6'))
```

```{r adjusted coefficient summary}
coef_summary(c('b7', 'b9', 'b10', 'b11', 'b12', 'b13'))
```

Coefficient Visualization

```{r coef_viz function}
coef_viz <- function(reg_mod, brms_mod_list){

  ## retrieve number of models for visualization


```

```
  num_mods <- length(brms_mod_list)

  ## initialize variables
  coef <- rep(NA, num_mods)
  upper_bound <- rep(NA, num_mods)
  lower_bound <- rep(NA, num_mods)

  ## retrieve coefficient estimate and upper/lower credible interval bounds for each model
  for(i in 1:num_mods){

    temp_mod <- eval(as.name(brms_mod_list[i]))

    coef[i] <- fixef(temp_mod)[2, 1]
    upper_bound[i] <- fixef(temp_mod)[2, 3]
    lower_bound[i] <- fixef(temp_mod)[2, 4]

  }

  ## create df to use in ggplot
  df <- data.frame(model = brms_mod_list, coef = coef, upper_bound = upper_bound,
lower_bound = lower_bound)

  ## retrieve mle (linear model coefficient estimate)
```

```
  mle <- as.numeric(coef(reg_mod)[2])


  ## create final graph
  df %>% ggplot() +
    geom_point(aes(x = as.factor(reorder(model,
                                         order(as.numeric(gsub(x = model, pattern = "b", replacement = ""))))),
                   y = coef), size = 3, color = 'blue') +
    geom_linerange(aes(x = as.factor(reorder(model,
                                             order(as.numeric(gsub(x = model, pattern = "b", replacement = ""))))),
                       ymin = lower_bound, ymax = upper_bound), color = 'blue') +
    geom_hline(yintercept = mle, color = 'red') +
    scale_x_discrete(labels=c("none", "normal(-39, 25)", "normal(-39, 5)",
                              "normal(-39, 1)", "student_t(1, -39, 1)", "      student_t(50, -39, 1)")) +
    xlab('Prior') +
    ylab('Coefficient Value') +
    theme_bw() +
    annotate("text", x = as.factor(brms_mod_list[4]), y = mle - 3,
             label = paste0('MLE = ', as.character(round(mle, 2))), color = 'red')


}
```

```'{r Coefficient Visualization: b0 - b6}

coef_viz(m1, c('b0', 'b2', 'b3', 'b4', 'b5', 'b6'))

```

```{r Coefficient Visualization: b7 - b13}

coef_viz(m2, c('b7', 'b9', 'b10', 'b11', 'b12', 'b13'))

```

Model Evaluation

```{r test_noNA}

test_noNA <- test %>% select(c(exposed_2_km, gestational_age_weeks_c, neonate_sex, apncu_index_collapsed, maternal_age_years_c, maternal_race, maternal_edu_cat_collapsed, received_wic, maternal_bmi_cat, gestational_diabetes, nulliparous, overall_smoking_gestation_and_three_months_prior, community_ses_index_quartile, bweight)) %>% na.omit()

```

```{r m1 rmse}

m1_preds <- predict(m1, newdata = test)

m1_rmse <- sqrt(mean((m1_preds - test$bweight)^2))

```

```{r m2 rmse}
m2_preds <- predict(m2, newdata = test_noNA)

m2_rmse <- sqrt(mean((m2_preds - test_noNA$bweight)^2))
```

```{r b0 rmse}
betas0 <- as.matrix(fixef(b0)[, 1])

X0 <- model.matrix(~ exposed_2_km, data = test)

preds0 <- X0 %*% betas0

rmse0 <- sqrt(mean((preds0 - test$bweight)^2))
```

```{r b1 rmse, eval = FALSE}
betas1 <- as.matrix(fixef(b1)[, 1])

X1 <- model.matrix(~ exposed_2_km, data = test)

preds1 <- X1 %*% betas1
```

```
rmse1 <- sqrt(mean((preds1 - test$bweight)^2))
```

```{r b2 rmse}
betas2 <- as.matrix(fixef(b2)[, 1])

X2 <- model.matrix(~ exposed_2_km, data = test)

preds2 <- X2 %*% betas2

rmse2 <- sqrt(mean((preds2 - test$bweight)^2))
```

```{r b3 rmse}
betas3 <- as.matrix(fixef(b3)[, 1])

X3 <- model.matrix(~ exposed_2_km, data = test)

preds3 <- X3 %*% betas3

rmse3 <- sqrt(mean((preds3 - test$bweight)^2))
```

```{r b4 rmse}

betas4 <- as.matrix(fixef(b4)[, 1])

X4 <- model.matrix(~ exposed_2_km, data = test)

preds4 <- X4 %*% betas4

rmse4 <- sqrt(mean((preds4 - test$bweight)^2))
```

```{r b5 rmse}

betas5 <- as.matrix(fixef(b5)[, 1])

X5 <- model.matrix(~ exposed_2_km, data = test)

preds5 <- X5 %*% betas5

rmse5 <- sqrt(mean((preds5 - test$bweight)^2))
```

```{r b6 rmse}

betas6 <- as.matrix(fixef(b6)[, 1])
```

```
X6 <- model.matrix(~ exposed_2_km, data = test)

preds6 <- X6 %*% betas6

rmse6 <- sqrt(mean((preds6 - test$bweight)^2))
```

```{r b7 rmse}
betas7 <- as.matrix(fixef(b7)[, 1])

X7 <- model.matrix(~ exposed_2_km +
                   # infant characteristics
                   gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
                   # maternal characteristics
                   maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
                   received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
                   overall_smoking_gestation_and_three_months_prior +
                   # environmental characteristics
                   community_ses_index_quartile, data = test)

preds7 <- X7 %*% betas7
```

```
rmse7 <- sqrt(mean((preds7 - test_noNA$bweight)^2))
```

```{r b8 rmse, eval = FALSE}
betas8 <- as.matrix(fixef(b8)[, 1])

X8 <- model.matrix(~ exposed_2_km +
                   # infant characteristics
                   gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
                   # maternal characteristics
                   maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
                   received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
                   overall_smoking_gestation_and_three_months_prior +
                   # environmental characteristics
                   community_ses_index_quartile, data = test)

preds8 <- X8 %*% betas8

rmse8 <- sqrt(mean((preds8 - test_noNA$bweight)^2))
```

```{r b9 rmse}
betas9 <- as.matrix(fixef(b9)[, 1])
```

```
X9 <- model.matrix(~ exposed_2_km +
                   # infant characteristics
                   gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
                   # maternal characteristics
                   maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
                   received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
                   overall_smoking_gestation_and_three_months_prior +
                   # environmental characteristics
                   community_ses_index_quartile, data = test)

preds9 <- X9 %*% betas9

rmse9 <- sqrt(mean((preds9 - test_noNA$bweight)^2))
```

```{r b10 rmse}

betas10 <- as.matrix(fixef(b10)[, 1])

X10 <- model.matrix(~ exposed_2_km +
                   # infant characteristics
                   gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
                   # maternal characteristics
```

```
            maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
            received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
            overall_smoking_gestation_and_three_months_prior +
            # environmental characteristics
            community_ses_index_quartile, data = test)

preds10 <- X10 %*% betas10

rmse10 <- sqrt(mean((preds10 - test_noNA$bweight)^2))
```

```{r b11 rmse}
betas11 <- as.matrix(fixef(b11)[, 1])

X11 <- model.matrix(~ exposed_2_km +
            # infant characteristics
            gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
            # maternal characteristics
            maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
            received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
            overall_smoking_gestation_and_three_months_prior +
            # environmental characteristics
            community_ses_index_quartile, data = test)
```

```
preds11 <- X11 %*% betas11

rmse11 <- sqrt(mean((preds11 - test_noNA$bweight)^2))
```

```{r b12 rmse}
betas12 <- as.matrix(fixef(b12)[, 1])

X12 <- model.matrix(~ exposed_2_km +
                    # infant characteristics
                    gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
                    # maternal characteristics
                    maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
                    received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
                    overall_smoking_gestation_and_three_months_prior +
                    # environmental characteristics
                    community_ses_index_quartile, data = test)

preds12 <- X12 %*% betas12

rmse12 <- sqrt(mean((preds12 - test_noNA$bweight)^2))
```

```{r b13 rmse}
betas13 <- as.matrix(fixef(b13)[, 1])

X13 <- model.matrix(~ exposed_2_km +
                    # infant characteristics
                    gestational_age_weeks_c + neonate_sex + apncu_index_collapsed +
                    # maternal characteristics
                    maternal_age_years_c + maternal_race + maternal_edu_cat_collapsed +
                    received_wic + maternal_bmi_cat + gestational_diabetes + nulliparous +
                    overall_smoking_gestation_and_three_months_prior +
                    # environmental characteristics
                    community_ses_index_quartile, data = test)

preds13 <- X13 %*% betas13

rmse13 <- sqrt(mean((preds13 - test_noNA$bweight)^2))
```

```{r rmse df}
## create rmse df to pipe into gt function
rmse_df <- data.frame(Model = c('m1', 'm2', 'b0', 'b2', 'b3', 'b4', 'b5',
                                'b6', 'b7', 'b9', 'b10', 'b11', 'b12', 'b13'),
```

```
                Prior = c("NA", "NA", "none", "normal(-39, 25)", "normal(-39, 5)",
                          "normal(-39, 1)", "student_t(1, -39, 1)", "student_t(50, -39, 1)",
                          "none", "normal(-39, 25)", "normal(-39, 5)",
                          "normal(-39, 1)", "student_t(1, -39, 1)", "student_t(50, -39, 1)"),
                RMSE = c(m1_rmse, m2_rmse, rmse0, rmse2, rmse3, rmse4, rmse5,
                         rmse6, rmse7, rmse9, rmse10, rmse11, rmse12, rmse13))


## final table
rmse_df %>%
  gt() %>%
  cols_align(
    align = "center",
    columns = everything()) %>%
  tab_options(column_labels.font.weight = 'bold')
```

Bibliography

[1] Cook, T., & Perrin, J. (2016, March 15). *Hydraulic fracturing accounts for about half of current U.S. crude oil production*. U.S. Energy Information Administration (EIA). Retrieved November 2022, from https://www.eia.gov/todayinenergy/detail.php?id=25372

[2] Gecan, R., Tawil, N., & Lasky, M. (2014). (rep.). *The Economic and Budgetary Effects of Producing Oil and Natural Gas From Shale*. Congress of the United States Congressional Budget Office. Retrieved November 2022, from https://www.cbo.gov/sites/default/files/113th-congress-2013-2014/reports/49815-effectsofshaleproduction.pdf.

[3] Denchak, M. (2022, April 13). *Fracking 101*. Natural Resources Defense Council. Retrieved December 2022, from https://www.nrdc.org/stories/fracking-101#whatis

[4] Concerned Health Professionals of New York, & Physicians for Social Responsibility. (2019, June). *Compendium of scientific, medical, and media findings demonstrating risks and harms of fracking (unconventional gas and oil extraction) (6th ed.).* http://concernedhealthny.org/compendium/

[5] Howarth, R., Ingraffea, A. & Engelder, T. *Should fracking stop?* Nature 477, 271–275 (2011). https://doi.org/10.1038/477271a

[6] Qingmin Meng. The impacts of fracking on the environment: A total environmental study paradigm. *Science of The Total Environment*, Volume 580, 2017, Pages 953-957, ISSN 0048-9697. https://doi.org/10.1016/j.scitotenv.2016.12.045.

[7] John L. Adgate, Bernard D. Goldstein, and Lisa M. McKenzie. *Potential Public Health Hazards, Exposures and Health Effects from Unconventional Natural Gas Development.* Environmental Science & Technology (2014). 48 (15), 8307-8320. DOI: 10.1021/es404621d

[8] Elliott, E. G., Trinh, P., Ma, X., Leaderer, B. P., Ward, M. H., & Deziel, N. C. (2016). Unconventional oil and gas development and risk of childhood leukemia: Assessing the evidence. *Science of The Total Environment, 576*, 138–147. https://doi.org/10.1016/j.scitotenv.2016.10.072

[9] Hill, E. L. (2018). Shale gas development and Infant Health: Evidence from Pennsylvania. *Journal of Health Economics, 61*, 134–150. https://doi.org/10.1016/j.jhealeco.2018.07.004

[10] Cushing, L. J., Vavra-Musser, K., Chau, K., Franklin, M., & Johnston, J. E. (2020). Flaring from unconventional oil and gas development and birth outcomes in the Eagle Ford Shale

in South Texas. *Environmental Health Perspectives*, *128*(7). https://doi.org/10.1289/ehp6394

[11] Tran, K. V., Casey, J. A., Cushing, L. J., & Morello-Frosch, R. (2020). Residential proximity to oil and gas development and birth outcomes in California: A retrospective cohort study of 2006–2015 births. *Environmental Health Perspectives, 128*(6). https://doi.org/10.1289/ehp5842

[12] Currie, J., Greenstone, M., & Meckel, K. (2017). Hydraulic fracturing and Infant Health: New Evidence from Pennsylvania. *Science Advances*, *3*(12). https://doi.org/10.1126/sciadv.1603021

[13] *TENORM: Oil and Gas Production Wastes*. United States Environmental Protection Agency. (n.d.). Retrieved December 2022, from https://www.epa.gov/radiation/tenorm-oil-and-gas-production-wastes

[14] *Recommended BMI-for-age Cutoffs*. Centers for Disease Control and Prevention. (2022, February 3). Retrieved October 2022, from https://www.cdc.gov/nccdphp/dnpao/growthcharts/training/bmiage/page4.html

[15] *Defining adult overweight & obesity*. Centers for Disease Control and Prevention. (2022, June 3). Retrieved October 2022, from https://www.cdc.gov/obesity/basics/adult-defining.html?CDC_AA_refVal=https%3A%2F%2Fwww.cdc.gov%2Fobesity%2Fadult%2Fdefining.html

[16] Kotelchuck, M. (1994, September). *Overview of adequacy of prenatal care utilization index*. Maternal and Child Health Library. Retrieved October 2022, from https://www.ncemch.org/databases/HSNRCPDFs/Overview_APCUIndex.pdf

[17] American Academy of Pediatrics. (2017). *Guidelines for perinatal care*.

[18] *2015-2019 ACS 5-Year estimates*. United States Census Bureau. (2021, December 8). Retrieved October 2022, from https://www.census.gov/programs-surveys/acs/technical-documentation/table-and-geography-changes/2019/5-year.html

[19] Kubler, R. (2022, September 18). *Beginner-friendly Bayesian inference*. Towards Data Science. Retrieved December 2022, from https://towardsdatascience.com/beginner-friendly-bayesian-inference-2e2839a9ae18

[20] Clyde, M., Çetinkaya-Rundel, M., Rundel, C., Banks, D., Chai, C., & Huang, L. (2022, June 15). *Chapter 6: Introduction to Bayesian Regression*. An Introduction to Bayesian Thinking. Retrieved November 2022, from https://statswithr.github.io/book/introduction-to-bayesian-regression.html

[21] Bürkner PC (2017). "brms: An R Package for Bayesian Multilevel Models using Stan." Journal of Statistical Software, 80(1), 1–28. doi:10.18637/jss.v080.i01.

[22] Lai, M. (2019, December 13). *Markov Chain Monte Carlo*. Bayesian Data Analysis. Retrieved November 2022, from https://bookdown.org/marklhc/notes_bookdown/