52542 - Generalized Linear Models Final Paper

Analyzing Racial Disparities in Low Birth Weight Using GLM

April 20, 2025

Prof. Samuel Oman

Itay Ben Avraham (206628943) & Eden Malka (318849940)

Table of Contents

Introduction

The Problem

Low birth weight is more than just a medical statistic — it's a critical indicator of infant health and survival. Babies born under 2.5 kg face significantly higher risks of developmental complications, long-term health issues, and even infant mortality. But low birth weight isn't solely determined by genetics or chance — it might be deeply tied to maternal health, prenatal care, and broader social determinants of health. Thus, at the heart of this study lies a research question we formulated after reviewing the data and variables—one we believe the original researchers might have sought to answer:

Do racial differences exist in the likelihood of low birth weight, even after accounting for maternal health and behavior?

This question is not only statistical — it reflects a deeper concern with understanding whether disparities in birth outcomes reflect underlying social inequalities, biological mechanisms, or differences in access to healthcare.

Research Objective

Our objective is to investigate whether race is an independent predictor of low birth weight after controlling for other maternal risk factors. In other words, we aim to determine whether the racial disparities observed in birth weight outcomes persist even when accounting for variables such as maternal age, smoking behavior, hypertension, and prenatal care.

Research Hypotheses

- **Null Hypothesis (H_0):**
  There are no racial differences in the likelihood of low birth weight after accounting for maternal health and behavioral factors.

- **Alternative Hypothesis (H_1):**
  Racial differences in the likelihood of low birth weight persist even after controlling for maternal health and behavioral factors.

Data and Variables

To answer this question, we utilize data from a **sample of 137 births** collected at *Baystate Medical Center in Springfield, Massachusetts*, during 1986. The dataset includes a binary response variable for low birth weight and a range of maternal and pregnancy-related predictors.

The variables included in the analysis are:

Dependent Variable

- **low** — Binary response variable *(1 = birth weight < 2.5 kg, 0 = birth weight ≥ 2.5 kg)*

Independent Variable of Interest

- **race** — Mother's race *(Categorical: White, Black, Other)*

- **age** — Mother's age in years *(Continuous)*
- **lwt** — Mother's weight in pounds at last menstrual period *(Continuous)*
- **smoke** — Smoking status during pregnancy *(Binary: 0 = No, 1 = Yes)*
- **ht** — History of hypertension *(Binary: 0 = No, 1 = Yes)*
- **ui** — Presence of uterine irritability *(Binary: 0 = No, 1 = Yes)*
- **ftv** — Number of physician visits during the first trimester *(Count)*
- **ptl** — Number of previous premature labors *(Count)*
- **bwt** — Actual birth weight in grams *(Continuous; used for descriptive analysis only)*

The dataset allows us to explore both medical and behavioral factors that may influence birth outcomes — and critically, how these intersect with racial identity.

Preliminary Exploratory Analysis

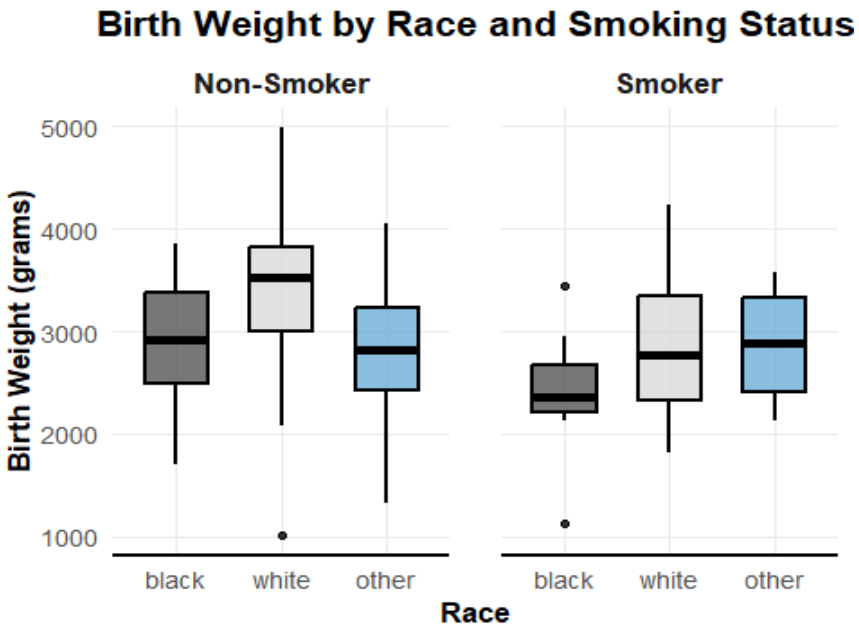
To explore the relationship between race and birth weight, we conducted a series of graphical analyses to uncover patterns and potential disparities. The goal was to assess whether racial differences in birth weight persist after accounting for other maternal factors, such as smoking.

Figure 1: Descriptive Statistics by Race

Race	Num	Age (mean ± SD)	LWT (mean ± SD)	BWT (mean ± SD) (kg)	Low BW *Independent Variable*	Smoker	Hypertension	Uterine Irritability
black	20	21.8 ± 5.4	149.8 ± 41.8	2.7 ± 0.7	45%	35%	15%	15%
other	48	21.8 ± 4.3	118.8 ± 19.0	2.8 ± 0.6	35%	17%	8%	10%
white	69	24.4 ± 5.8	129.8 ± 25.5	3.1 ± 0.7	25%	51%	4%	14%

Figure 2: Boxplot - Birth Weight by Race and Smoking Status

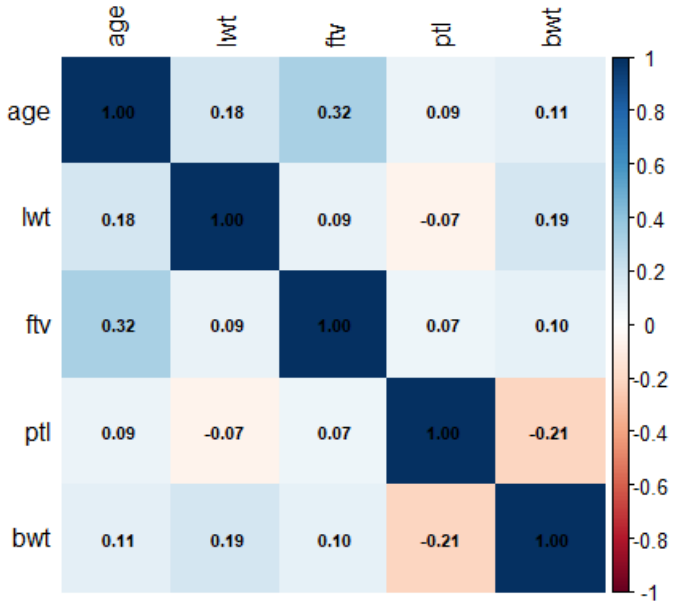
We begin with a boxplot {*Figure 2*} that compares birth weight across different racial groups, separated by smoking status. This plot was chosen because it allows us to simultaneously evaluate the impact of both race and smoking on birth weight. *Why smoking?* Because it is a well-established risk factor for low birth weight — it reduces fetal growth by limiting oxygen supply. By introducing smoking as a second dimension, we aim to explore whether racial disparities in birth weight are potentially mediated by behavioral differences, or whether race itself contributes independently to these outcomes. If racial disparities persist within both smoking and non-smoking groups, it would suggest that race may have an independent association with birth weight beyond the behavioral influence of smoking.



(!) The results reveal clear differences: babies born to Black mothers tend to have lower birth weights compared to those born to White and Other race mothers. Smoking further reduces birth weight across White and Black racial groups, highlighting a potential compounding effect.

Figure 3: Correlation Heatmap

To assess the overall relationship between maternal characteristics and birth weight, we generated a correlation heatmap {*Figure 3*}. This plot helps us identify which predictors are most strongly related to birth weight and whether any multicollinearity issues exist.



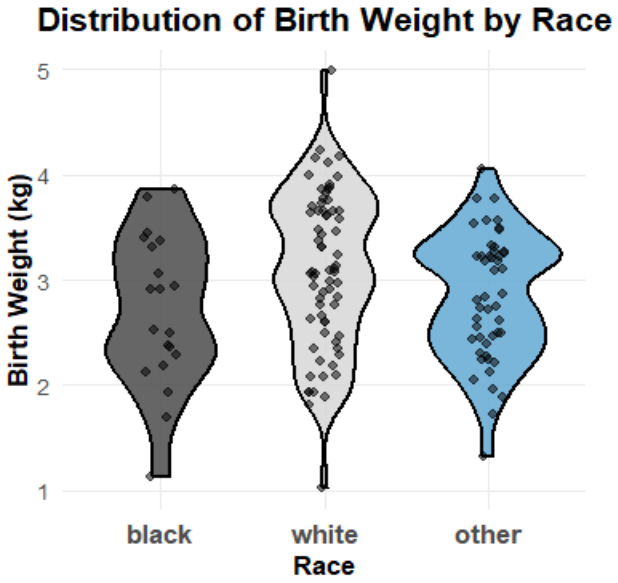
Interestingly, maternal weight (lwt) shows the highest positive correlation with birth weight - (r = 0.19), while the number of previous premature labors (ptl) has a modest negative correlation -

(r = -0.21). The absence of strong correlations between predictors reassures us that multicollinearity is unlikely to undermine the stability of the model.

💡 Fun Fact: We noticed that the highest correlation in the dataset is between maternal age and the number of physician visits during the first trimester. This makes intuitive sense — as age increases, so does pregnancy-related risk, leading to more frequent medical monitoring. A nice alignment between data and common sense!

Figure 4: Violin Plot

To further explore racial differences in birth weight distribution, we created a violin plot {*Figure 4*}. While the boxplot shows central tendency and variation, the violin plot allows us to see the full **shape** of the distribution within each racial group, which might be informative.

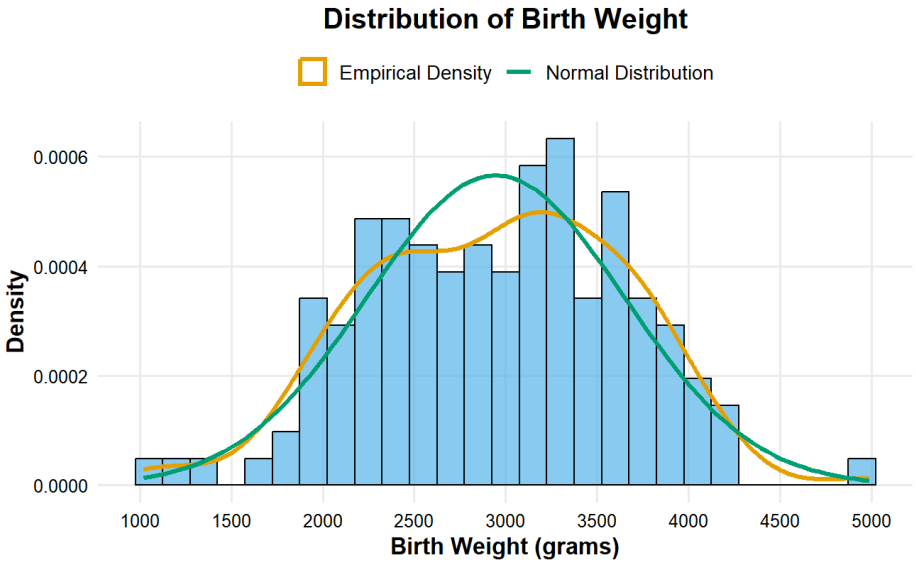


The violin plot {*Figure 4*} reveals clear visual differences in birth weight distributions by race. First, the number of Black mothers in the dataset is smaller (20 compared to 69 White babies), which is visually apparent in the plot. Despite the smaller sample size, the distribution for Black babies is noticeably shifted lower, with fewer births reaching higher weights. Interestingly, about half of the births in the Black group fall below the low-birth-weight threshold.

(!) The individual points within the violin plot are slightly spread out along the x-axis due to jittering — a technique that introduces small random variation to prevent overlap and make the individual data points more visible. The x-axis position of the points carries no actual meaning — the meaningful information comes from the y-axis (birth weight) and the overall shape of the distribution.

Figure 5: Histogram of birth weight

Finally, we included a histogram {*Figure 5*} to provide a general sense of the overall birth weight distribution.



Birth weight follows a roughly normal distribution, with a slight skew to the right. This suggests that birth weight itself is a reasonably well-behaved predictor for modeling purposes. The binary outcome (low birth weight) will be modeled using a generalized linear model with a logit link, which is appropriate given the binary nature of the response variable.

For summary, these exploratory findings provide a clear foundation for the next stage of analysis. The boxplot and violin plot confirm that racial differences in birth weight persist even after accounting for smoking, while the correlation heatmap highlights the key maternal predictors of birth weight. The overall shape of the birth weight distribution, as shown in the histogram, supports the appropriateness of using a generalized linear model with a logit link. The next step is to fit the model and test whether racial disparities remain significant after adjusting for maternal health and behavioral factors.

Formulation of a Generalized Linear Model (GLIM)

To evaluate whether racial differences exist in the likelihood of low birth weight, we formulate a **Generalized Linear Model (GLIM)** based on a **logistic regression** framework. Since the outcome variable (*low*) is binary (*0* = normal birth weight, *1* = low birth weight), the logistic regression model is appropriate because it models the **probability** of low birth weight using a **logit link function**.

1. Structure of the Model

The general form of a logistic regression model is:

$$P(Y = 1 \mid X) = \frac{\exp(\eta)}{1 + \exp(\eta)}$$

Where:

- Y is a binary response vector of dimension $n \times 1$, where $Y = 1$ if low birth weight and 0 otherwise
- X is the design matrix of predictors
- β is a coefficient vector
- $\eta = X\beta$ is the linear predictor

- $P(Y = 1 \mid X)$ is a vector of probabilities for each observation, representing the likelihood that the baby is of low birth weight given the predictors in X

The linear predictor η takes the form:

$$\eta = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{lwt} + \beta_3 \cdot \text{race}_{\text{black}} + \beta_4 \cdot \text{race}_{\text{other}} + \beta_5 \cdot \text{smoke} + \beta_6 \cdot \text{ht} + \beta_7 \cdot \text{ui} + \beta_8 \cdot \text{ftv} + \beta_9 \cdot \text{ptl}$$

Where:
- $\boldsymbol{\beta_0}$ — *Intercept*
- $\boldsymbol{\beta_i}$ — *Regression coefficient* for predictor i
- race_white is treated as the **reference category** (*baseline*)
- smoke, ht, ui are **binary predictors**
- age, lwt are **continuous predictors**
- ftv, ptl are **discrete numeric predictors** (*count variables*)

2. Link Function

We use the **logit link function**, which transforms the probability into a linear form:

$$\text{logit}\big(P(Y = 1 \mid X)\big) = \log \frac{P(Y = 1 \mid X)}{1 - P(Y = 1 \mid X)} = \eta$$

The response is modeled as the **log-odds** of low birth weight — the logarithm of the ratio between the probability of the event and its complement. This transformation maps probabilities (ranging from 0 to 1) onto a continuous linear scale, allowing us to use linear modeling while keeping predicted values within valid probability bounds.

3. Model Assumptions

The logistic regression model relies on the following key assumptions:

1. **Linearity in the log-odds** – The relationship between the predictors and the log-odds of low birth weight is linear.
2. **Independent observations** – The birth outcomes are assumed to be independent.
3. **No severe multicollinearity** – Predictors should not be highly correlated with each other (confirmed by the correlation heatmap).
4. **Outcome is binary** – The response variable (*low*) is binary and correctly coded as *0* or *1*.

4. Interpretation of Coefficients

The estimated coefficients β_i can be interpreted as follows:
- β_i represents the **log-odds change** associated with a one-unit increase in the predictor.
- After fitting the model, we will exponentiate the coefficients to obtain the **odds ratios**:

$$e^{\beta_i}$$

which represents the **multiplicative effect** on the odds of low birth weight for a one-unit increase in the predictor, holding other variables constant.

For example:
- If $e^{\beta_3} > 1$, it would mean that babies born to Black mothers have **higher odds** of low birth weight compared to White mothers, holding other factors constant.
- If e^{β_5} (smoke) is significant and greater than 1, it would suggest that smoking increases the likelihood of low birth weight.

5. Complete Model Specification

The final model is specified as:

$$\log\frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{lwt} + \beta_3 \cdot \text{race}_{\text{black}} + \beta_4 \cdot \text{race}_{\text{other}} + \beta_5 \cdot \text{smoke} + \beta_6 \cdot \text{ht} + \beta_7 \cdot \text{ui} + \beta_8 \cdot \text{ftv} + \beta_9 \cdot \text{ptl}$$

where:
- $\text{race}_{\text{white}}$ is the **reference category**
- The coefficients β_i are estimated using **maximum likelihood estimation (MLE)**, which finds the values that maximize the **log-likelihood function** — the probability of observing the data given the model.

Why a Logit Link?

We chose a logit link because it maps the probability of low birth weight to a scale where the relationship with the predictors is linear. It ensures that the predicted probability remains between **0** and **1**, which aligns with the binary nature of the outcome variable.

The model

Initial Model Fitting

Again, to evaluate the relationship between maternal characteristics, behavioral factors, and the likelihood of low birth weight, we fitted an initial logistic regression model. The dependent variable was low birth weight (binary outcome: 0 = normal birth weight, 1 = low birth weight). The independent variables included maternal age, weight at last menstrual period, race, smoking status, history of hypertension, uterine irritability, number of physician visits during the first trimester, and the number of previous premature labors. **Note:** $\text{race}_{\text{white}}$ is the reference category. Coefficients for $\text{race}_{\text{black}}$ and $\text{race}_{\text{other}}$ are interpreted relative to this group.

$$\log\frac{P(Y=1)}{1-P(Y=1)} = \beta_0 + \beta_1 \cdot \text{age} + \beta_2 \cdot \text{lwt} + \beta_3 \cdot \text{race}_{\text{black}} + \beta_4 \cdot \text{race}_{\text{other}} + \beta_5 \cdot \text{smoke} + \beta_6 \cdot \text{ht} + \beta_7 \cdot \text{ui} + \beta_8 \cdot \text{ftv} + \beta_9 \cdot \text{ptl}$$

Here is the table with the coefficients, odds ratios, highlighted p-values and CI bounds:

Term	Coefficient (β)	Estimate	Std. Error	z value	p-value	CI Lower	CI Upper
(Intercept)	**1.43**	1.43	1.53	0.23	0.81	0.07	30.31
raceblack	**1.39**	4.02	0.64	2.17	0.03	1.17	14.78
raceother	**0.69**	2.00	0.55	1.25	0.21	0.68	6.09
age	**-0.01**	0.99	0.05	-0.13	0.9	0.91	1.09
lwt	**-0.02**	0.98	0.01	-1.97	0.05	0.96	1.00
smokeSmoker	**0.82**	2.26	0.51	1.60	0.11	0.83	6.23
ht1	**1.82**	6.15	0.78	2.33	0.02	1.38	31.46
ui1	**0.86**	2.37	0.58	1.48	0.14	0.74	7.53
ftv	**-0.21**	0.81	0.46	-0.45	0.65	0.32	2.01
ptl	**1.31**	3.69	0.52	2.49	0.01	1.34	10.63

Interpretation of Model Coefficients

- The effect of **race (Black vs White)** was found to be statistically significant (p = 0.03), meaning we **reject the null hypothesis** that race has no association with low birth weight. Babies born to Black mothers are approximately **4.02 times more likely** to be of low birth weight compared to those born to White mothers ($e^{1.39}$ = 4.02). This supports our central research question — suggesting that racial disparities in birth outcomes **persist even after controlling maternal health and behavior**.

- **Hypertension:** Mothers with a history of hypertension had significantly increased odds of delivering a low-birth-weight baby. The odds were approximately **6.15 times higher** ($e^{1.82}$ = 6.15, p = 0.02) compared to those without hypertension.

- **Previous Premature Labors (ptl):** A history of premature labor increased the odds of low birth weight by a factor of **3.69** ($e^{1.31}$ = 3.69, p = 0.01), suggesting a strong association between past obstetric complications and adverse birth outcomes.

- **Maternal Weight (lwt):** For each additional pound of maternal weight at the last menstrual period, the odds of low birth weight decrease by about 2%, holding all else constant ($e^{-0.018}$ = 0.98, p = 0.05), indicating a small but statistically significant protective effect.

Other predictors — such as maternal age, smoking status, uterine irritability, and first-trimester physician visits — did not reach statistical significance (p>0.05). In summary, the model highlights several maternal health indicators — particularly **race, hypertension**, and **past premature labors** — as significant predictors of birth outcomes.

Evaluating a Refined Sub model

To improve model parsimony and fit, we applied AIC-based stepwise selection to the initial model. Stepwise selection evaluates the trade-off between model complexity and fit by iteratively adding and removing predictors based on their contribution to the model's log-likelihood and AIC (Akaike

Information Criterion). This stepwise model selection procedure functions as a form of deviance analysis, which serves as the GLM equivalent of ANOVA in classical linear modeling.

Here is the new table:

Term	Coefficient (β)	Std. Error	z value	p-value	Odds Ratio	CI Lower	CI Upper
(Intercept)	**0.103**	1.242	0.083	0.9338	1.11	0.10	12.66
ptl	**1.249**	0.508	2.457	0.0140	3.49	1.29	9.44
ht	**1.845**	0.776	2.378	0.0174	6.33	1.38	28.97
lwt	**-0.018**	0.009	-2.022	0.0431	0.98	0.96	1.00
ui	**0.901**	0.571	1.578	0.1145	2.46	0.80	7.53
raceblack	**1.418**	0.630	2.251	0.0244	4.13	1.20	14.20
raceother	**0.776**	0.528	1.469	0.1417	2.17	0.77	6.12
smokeSmoker	**0.874**	0.489	1.786	0.0740	2.40	0.92	6.25

1. Model Fit:

- Stepwise selection reduced the AIC from **159.94** to **156.21**, indicating a more parsimonious and better-fitting model.
- The slight increase in residual deviance (**139.94** to **140.21**) suggests a small loss in explanatory power, but the lower AIC indicates that the model is more efficient and better balanced in terms of fit and complexity.

2. Key Predictors:

- **Hypertension** (p = 0.0174) and **previous premature labors** (p = 0.0140) remain statistically significant.
- **Maternal weight** at the last menstrual period remains significant (p = 0.0431), suggesting that higher maternal weight continues to reduce the likelihood of low birth weight.
- **Race (black)** remains statistically significant (p = 0.0244), indicating that babies born to Black mothers have higher odds of low birth weight compared to White mothers.
- The consistency of hypertension, previous premature labors, and race effects after removing non-significant terms suggests that these predictors are **robust and stable**.

3. Dropped Predictors:

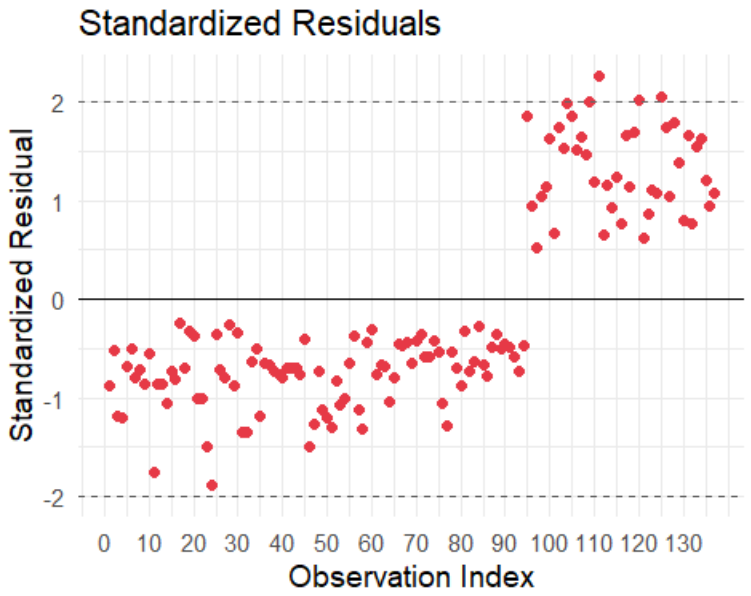
- Age and the number of physician visits (ftv) were dropped during stepwise selection – indicating that it does not meaningfully contribute to predicting low birth weight after accounting for other factors.

Residuals and Influence Diagnostics

Residual analysis was conducted based on the reduced model obtained via stepwise selection, rather than the full model, to evaluate the fit and assumptions of the final selected model. Thus, the number of predictors was reduced from 9 to 7 through this process.

Standardized Residuals

To evaluate the **individual prediction accuracy** of our model, we examined the **standardized residuals**. These residuals show how far off each observation is from the model's prediction, expressed in units of standard deviation. This allows us to identify **outliers** — data points where the model made unusually large prediction errors. In logistic regression, residuals should generally fall within the range of -2 to +2. The plot below displays each observation's standardized residual:



Interpretation:

Most residuals are well within the acceptable range of -2 to +2, and **no observations exceed ±3**, indicating that the model fits the vast majority of cases reasonably well.

However, starting around **observation index 95**, we notice a **clear split in the residuals**: One group of points lies **above the zero line**, with residuals between 1 and 2, and another group lies **below**, around -1.

This pattern likely reflects the **binary nature of the response variable** (low) and how the model fits the two outcome groups:

- For babies predicted to have **low birth weight** (1), the model **underestimates the probability** in some cases, leading to positive residuals.

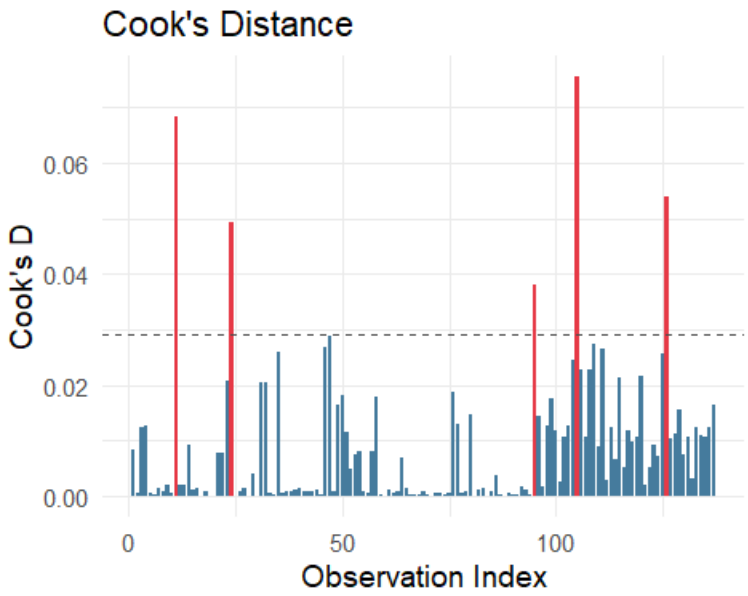
- For babies predicted to have **normal birth weight** (0), the model **slightly overestimates the risk**, resulting in negative residuals.

This type of "layered" residual pattern is **not necessarily a red flag** in logistic regression — it's a natural result of modeling a binary outcome. More importantly, **no extreme residuals** are present, and the structure appears symmetric and stable.

We conclude that the residuals are well-behaved and that this observed split reflects **expected variation across response categories**, not model misfit.

Cook's Distance

To investigate the potential influence of these high residuals, we calculated Cook's Distance for each observation. Cook's Distance measures how much an individual data point affects the model's overall fit. The plot below shows Cook's Distance for all observations, with the red line representing the threshold for identifying influential points



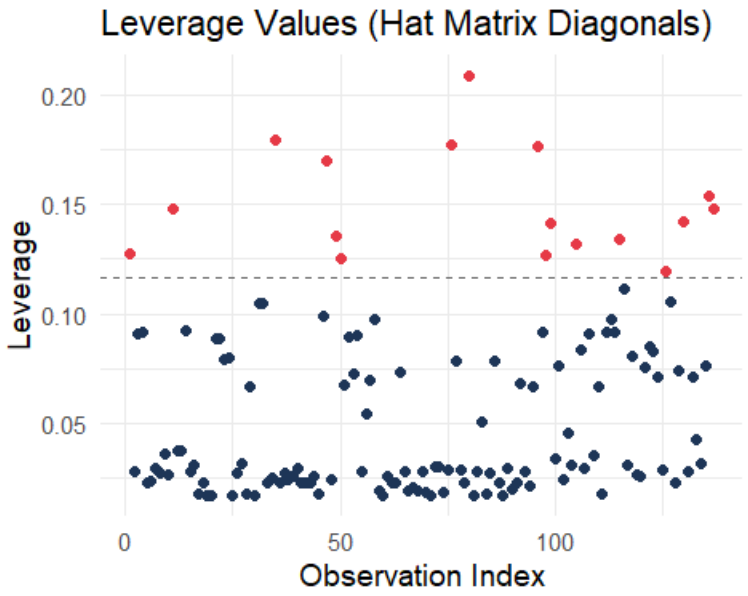
Interpretation:

Most points have low Cook's Distance values, indicating they have little influence on the model's fit. However, observations 11, 24, 95, 105, and 126 exceed the typical threshold — meaning they have disproportionate influence on the model's coefficients. Their Cook's Distance values are not extreme, suggesting that these points are not distorting the model, and may reflect meaningful variation rather than data errors.

To be thorough, we also fit the model excluding these five observations. The resulting estimates were nearly identical to those of the original model, and thus we decided not to include the alternative version in the report to avoid unnecessary complexity (The resulting model yielded nearly identical coefficients (to two decimal places), stable odds ratios, and preserved statistical significance — with race (Black) still significant at $p \approx 0.025$). Still, we note that this sensitivity check reinforces the robustness of our model's results.

Leverage

To assess which observations have **unusual combinations of predictor values**, we examined the **leverage values** (diagonal entries of the hat matrix). High leverage points are **not necessarily outliers**, but they have the potential to **pull the regression surface toward themselves**, especially when paired with large residuals. A common rule of thumb is that observations with leverage greater than $2(p + 1)/n$ are worth investigating. For our model (with ~7 predictors and 137 observations), that threshold is roughly **0.127**, marked as a dashed line below.



Interpretation:

Most observations fall well below the leverage threshold, indicating they have typical predictor profiles. However, we observe a number of **high-leverage points** (highlighted in red) — these are observations with **unusual combinations of maternal characteristics**, such as rare configurations of race, hypertension, prior labors, or smoking status.

Importantly, these high-leverage points are **not necessarily problematic on their own**. In fact, they do **not coincide with large residuals** or outliers in Cook's Distance, which means they aren't distorting the model. Their influence is primarily in the **space of the predictors**, not in error magnitude. As such, these points are **informative**, not alarming — and they reflect legitimate variability in patient profiles. Their inclusion improves generalizability of the model rather than threatening it, and no corrective action is warranted.

Model Fit and Goodness-of-Fit Tests

To assess the overall fit and calibration of the model, we evaluated two key diagnostic measures:

Overdispersion Test

- The **overdispersion ratio** was **1.087**, which is very close to 1.
- This indicates that the variance of the data aligns well with the assumed binomial distribution.

- An overdispersion ratio near 1 suggests that the model is **well-calibrated** and that there is **no meaningful overdispersion** in the data.

- Therefore, the model's estimated standard errors and significance levels are likely reliable.

Hosmer-Lemeshow Test

- The **Hosmer-Lemeshow test** produced a chi-square value of **7.80** with a **p-value of 0.454**.

- Since the p-value is large (greater than 0.05), we **fail to reject the null hypothesis** that the model fits the data well.

- This indicates that the model's **predicted probabilities are consistent with the observed outcomes** — confirming that the model is well-calibrated and not systematically misrepresenting the data.

Summary: Residuals and Influence Diagnostics

Overall, diagnostic checks suggest that the logistic regression model is **statistically sound and well-calibrated**. Residual plots show no severe outliers or patterns indicating misspecification. A few high-leverage points and moderately influential observations were identified, but they do not distort the model and likely reflect valid variation in the data. The **overdispersion ratio** is close to 1, and the **Hosmer-Lemeshow test** supports a good fit, indicating that the model's predicted probabilities align well with observed outcomes. Together, these diagnostics confirm that the model is reliable for inference and interpretation.

Conclusions

This study investigated whether racial disparities in low birth weight persist after accounting for maternal health and behavioral factors. Using logistic regression models, we examined a broad set of predictors and found that **race—particularly being Black—remained a statistically significant factor**, even after adjusting for maternal weight, hypertension, and previous premature labor. These findings lead us to reject the null hypothesis and directly support our central research question: **racial disparities in birth weight outcomes persist and cannot be fully explained by individual-level clinical or behavioral characteristics**.

Our final model, refined through AIC-based stepwise selection, struck an effective balance between parsimony and predictive accuracy. It identified **race, hypertension, prior premature labors, maternal weight, and smoking** as the most influential predictors of low birth weight. Model diagnostics—including residual analysis, leverage, and Cook's distance—indicated a robust and well-behaved fit. Furthermore, the **Hosmer-Lemeshow test** and the **overdispersion check** confirmed that the model was appropriately specified and well-calibrated for binary outcomes.

In conclusion, our analysis offers statistically grounded evidence that **race remains an independent and meaningful predictor of low birth weight**, even in the presence of medical and behavioral controls. This persistent disparity suggests that broader, unmeasured structural forces—such as systemic inequities, differential access to healthcare, and psychosocial stressors—may contribute to adverse birth outcomes among marginalized groups.