PHP 2550: Worksheet 7
Due: October 18th at 11:59pm

Many Analysts Recap

Summarize the data and research question posed to the teams in the paper "Many Analysts, One Data Set". What do you notice about the difference in methodological approaches teams had? How different are the resulting estimated odds ratios? Overall, what do you think about this experiment? (~2 paragraphs)

In the paper "Many Analysts, One Data Set", researchers applied a crowdsourcing data analysis approach to investigate the same research question: whether soccer referees are more likely to give red cards to dark-skin-toned players than to light-skin-toned players. The dataset covers 2,053 players from the top male leagues in England, Germany, France, and Spain during the 2012-2013 season, along with 3,147 referees they encountered, creating 146,028 player-referee dyads. Key variables included the players' position, weight, height, and skin tone, and referee data such as country of origin. Additionally, each player-referee dyad provided information on how many games they interacted in, as well as the frequency of yellow and red cards issued to each player. To be noticed, the skin tone ratings, originally coded on a scale from 1 (very light) to 5 (very dark), were later standardized to a 0 to 1 scale. Additional control variables like age, club, and league which would change over players' career are received at the time of data collection but not at the times the red cards are awarded. The same dataset was assigned to 29 independent research teams working from 13 different counties and came from various disciplinary backgrounds to investigate the same research questions. Teams are free to decide which variables they would like to include, their statistical methods to use, and their methods to handle nonindependence of players and referees. Among the 29 teams, 21 unique combinations of covariates were used, with some groups including a variety of control variables while others only used a few key variables. Additionally, statistical methods vary as well. Six teams chose to use poisson models, fifteen chose logistic models, six used linear models, and the remaining two teams used specialized methods which are classified as "miscellaneous." Also, groups chose various methods to address the nonindependence between players and referees, including introducing fixed effects, variance components, and adding clustered standard errors.

The resulting estimated odds ratio varies among teams significantly, ranging from 0.89 to 2.93, and the median estimate is 1.31. Among the 29 teams, 20 found significant positive relationships between players' skin tone and red card issuance, other groups resulted in insignificance with no group resulting in significant negative relationship. Logistic and Poisson models generally produced higher odds ratios, with most teams finding significant effects (median ORs around 1.34–1.36), while linear models showed lower median ORs (1.21) and fewer significant results. Methods to handle nonindependence, such as fixed effects or clustered standard errors, also influenced outcomes, with teams using these methods reporting median ORs from 1.28 to 1.39. These significance variations imply how differences in variable selection and statistical model or approach would lead to various conclusions. This study introduces a really innovative approach for us to show how research is influenced by researchers' decisions and subjective choices. Even with identical dataset, different researchers could report significant various conclusions due to different choices of variables and statistical methods or analytic approaches. The paper mentions that "in some cases, authors use a particular analytic strategy because it is the one they know how to use, rather than because they have a specific rationale for using it." This sentence does inspire us about a human side of research that we have never realized before. It implies the hidden subjective choices and personal limitations that can shape the results. By employing a crowdsourced data analysis approach, this study incorporates diverse analytic strategies and perspectives from researchers worldwide, each with unique disciplinary backgrounds. This diversity encourages transparency, highlighting the flexibility in data interpretation and strengthening the robustness and reliability of their results. The collective approach in this study implies benefits of open collaboration, as it enriches our understanding of data by drawing from a wide range of analytic viewpoints.

Answering Scientific Questions with Regression

Answer the following questions about the difference-in-differences paper you were assigned. (~1 paragraph per question)

1. What was the motivating research question? How was this translated to a scientific question and analytic approach?

   Previous studies have found that beverage taxes are associated with increased prices and reductions in the volume of beverages sold. However, this tax effect at small, independent stores has not been thoroughly studied. This study was conducted in the purpose to examine whether an increase in excise tax on sweetened beverages was associated with the sustained changes in beverage prices and purchases as well as calories purchased from beverages and high-sugar foods, over 2 years at independent stores in Philadelphia and Baltimore, Maryland. To approach this question, a cross-sectional study was conducted, with Baltimore as the control and Philadelphia as the treated city. A difference-in-difference approach was used to examine the difference in the pre-tax and post tax beverage prices, fluid ounces purchased, and total calories purchased from beverages and high-sugar foods between the two cities.

2. What is the underlying model(s) used? Be as specific as possible and explain how you determined the model.

This paper uses a Difference-in-Difference approach combines with generalized linear mixed-effects model. This study has two time periods: pre-tax period (October to December 2016), and post-tax period (6 months, 12 months, and 24 months after the tax implementation). Data are collected at Philadelphia (intervention city where implements tax) and Baltimore (control city where doesn't implement tax) before and after the tax implementation. Our interested outcomes are change in beverage prices (cents per fluid ounce), change in fluid ounces of taxed and non-taxed beverages purchased per customer, and change in total calories purchased from sweetened beverages and high-sugar foods. In addition to the main variables of interest (time period, city location, and beverage tax status), the study also includes other covariates, like income level, customer demographics (gender, race, education level, age, and frequency of store visit), and purchasing behaviors (purchased goods and total spending), to control for potential confounding. The outcomes are analyzed using a DiD approach, which compares the difference in changes over time between the two cities, aiming to isolate the effect of the tax from other factors that might have influenced beverage prices and consumption in both cities over time. Additionally, the model used generalized linear mixed-effects model which introduces random intercepts for stores to adjust for unobserved heterogeneity among stores at baseline. The underlying model(s) used in this study is:

$$Y_{it} = \alpha + \beta_1 \text{PostTax}_t + \beta_2 \text{Treatment}_i + \beta_3 (\text{PostTax}_t \times \text{Treatment}_i) + X_{it}\gamma + \epsilon_{it}$$

Here Y_{it} can be our three interested outcomes including beverage price, beverage quantity purchased, and the calories purchased for store or individual i at time t. $PostTax_t$ is a binary indicator where it is 1 if the observation is post-tax and 0 if the observation is pre-tax. $Treatment_i$ is also a binary indicator where it is 1 if the observation is from the intervention city, Philadelphia, and 0 if the observation is from the control city, Baltimore. The interaction term between $PostTax_t$ and $Treatment_i$ is for the difference-in-difference approach which is used to capture the effect of the tax implementation across time and location. X_{it} represents a series of demographic characteristics of the customers including income, education, race, age, and gender. We generate this model using information given in the results table output and the statistical analysis section descriptions.

3. How were the results used to answer the question and what was the conclusion?

- The results in the "Change in Beverage Price" can be used to answer the question of whether an increase in excise tax on sweetened beverages was associated with the sustained changes in beverage prices of beverages and high-sugar foods. In the paper, the research mentions that "there was a 2.06 cents per fl oz (95% CI, 1.75 to

2.38 cents per fl oz; p <.001) increase for taxed beverages in Philadelphia compared with Baltimore, an increase of 33%, indicating a 137.3% pass-through of tax." This implies that stores not only passed the 1.5-cent-per-ounce tax onto customers, but also charged even more than the required tax. Since Baltimore, the control city, does not show price increases and non-taxed beverages in Philadelphia also does not exhibit this price increase, the price increases in Philadelphia would be the result of the tax implementation. Also, the price increases were consistent across income levels, with no significant difference between low- and high-income neighborhoods. By using DiD approach combined with generalized linear mixed-effects models, researchers are able to compare the price change on taxed and non-taxed beverages in Philadelphia and Baltimore before and after the tax was implemented. Also, the adding random intercepts across stores is able to adjust for differences in store-specific factors. From the results, we can conclude that the tax implement significantly increases the price of taxed beverage overtime, and the tax are overly passed on to the customers which indicates the direct and substantial impact of tax on beverage prices.

- The results in the "Change in Volume of Beverage Purchased" can be used to answer the question of whether an increase in excise tax on sweetened beverages was associated with the sustained changes in purchases from beverages and high-sugar foods. The paper mentions that "there was a 6.12–fl oz decline (95% CI, −9.88 to −2.37 fl oz; P < .001), or a 41.9% decrease, in the ounces of taxed beverages purchased per person in Philadelphia compared with Baltimore." This refers to a significant drop in the purchase of taxed beverages. This significant drop was driven mainly by a reduction in the purchase of sugar-sweetened beverages (SSBs), with a 6.17 fluid ounce per person reduction, corresponding to a 47.3% decrease. Also, the paper said there was no significant change on the volume of non-taxed beverages in Philadelphia, indicates the implementation of tax would only increase the sweet beverages. In addition, the analysis showed that the tax's effects were more pronounced in low-income neighborhoods and lower education levels, with purchases dropped by around 40%. From their results, we can conclude that the tax implementation in Philadelphia significantly reduced the consumption amount of taxed beverages where the effect is more pronounced on vulnerable population. Form this, we can see that tax implementation is able to help reducing sugar consumptions and addressing the health disparity problem caused by excessive sugar consumption.

- The results in the "Changes in Calories and Spending on Beverages and High-Sugar Foods" can be used to answer the question of whether an increase in excise tax on sweetened beverages was associated with the sustained changes in calories purchased from beverages and high-sugar foods. The paper mentions that "there was a 69-calorie decrease (95% CI, −132 to −5 calories; P = .04) in the total calories purchased from SSBs and high-sugar foods combined, a 22.6% decline." In addition, "the grams of sugar from these items declined by 19.9 g (95% CI, −31.7 to −8.2 g; P

4

= .002), or 34.1% per person." These results indicate that the tax implementation significantly reduces calorie and sugar intake. There wasn't a change on people's spending with their shopping post tax, but the frequency of neighboring counties purchasing increases slightly as people chose to make purchases in other countries to avoid tax. This reduction was more noticeable among low-income neighborhoods and people with lower levels of education as well. From these, we can conclude that tax significantly decreases people's calorie and sugar intake from sweetened goods, especially strong for people with lower-income and lower levels of education. However, this decrease does not influence their total spendings.

4. Overall, how do you evaluate this paper? Think about strengths and weaknesses of the approach and any remaining questions you have.

This paper evaluates the impact of Philadelphia's sweetened beverage tax on beverage prices and purchases at independent stores over a two-year period. By implementing a difference-in-difference approach incorporated with mixed-effect models, the study compares Philadelphia to Baltimore, a nearby city without the tax, and finds that the tax led to significant changes. The inclusion of random intercept by the mixed effect model helps manage repeated measures and capture variability across stores and customers. Additionally, the analysis also incorporates data on prices, items purchased (price and items purchased by each customer), and cross-border shopping, providing a more comprehensive understanding of how tax implementation influences consumer behavior. However, there are also some limitations. It excludes evening and weekend purchases, which could introduce bias, and focuses only on independent stores, leaving questions about potential shifts to supermarkets. Moreover, the assumption of parallel trends between large chains and independent stores was not directly tested. Additionally, the lower response rate in Philadelphia and limited data on artificially sweetened beverages could affect the generalizability of the findings.

Model Evaluation Example

These questions are on the paper 'Predicting lung cancer prior to surgical resection in patients with lung nodules' by Deppen et al. This paper introduces a model called TREAT that is currently used in practice to predict lung cancer.

1. Compare the Mayo model to the TREAT model in terms of the initial goals of building the model, the population the training data represented, the variables included, and the resulting model. (~2 paragraphs)

The Mayo model focused on improving biopsy screening or referral in the general medical population. The model was designed to evaluate patients with nodules selected from the general population whose lesions were found on imaging. The prevalence of disease in

the population considered in the training data is 23%. The model contained six variables: age, smoking history, previous cancer, lesion size, spiculated edge and location.

The TREAT model was constructed based on the need to reduce unnecessary surgery for benign disease and was calibrated to be used in the preoperative evaluation of suspicious lesions in the lung. The model was designed to help surgeons obtain an accurate and well-calibrated predictive model to facilitate the diagnosis of suspected lung cancer without missing early stage disease. The model was trained on people undergoing thoracic surgical evaluation of lung nodules or masses for known or suspected non-small cell lung cancer with a prevalence of 72%. Compared with the Mayo Clinic model, the TREAT model added six variables: gender, body mass index, chronic obstructive pulmonary disease (COPD), lesion growth, FDG-PET positivity, and hemoptysis in preoperative symptoms. For variables related to smoking, The TREAT model included pack-years of smoking, which took into account both duration and intensity of smoking.

The TREAT lung cancer model demonstrated superior performance (AUC = 0.87) compared to the Mayo Clinic model (AUC = 0.80) and was validated in a separate, higher-risk cohort (AUC = 0.89). While the Mayo Clinic model performed well in a Vanderbilt University Medical Center (VUMC) population, its accuracy decreased (AUC = 0.73) as disease prevalence increased in a Veterans Affairs cohort with 95% lung cancer prevalence. The Mayo Clinic model showed poorer calibration, underestimating cancer risk in lower-risk patients, limiting its use in surgical populations with higher cancer prevalence.

2. What measures or visuals were used to evaluate the models? How do we interpret these? Why do you think these measures were chosen for comparison? (1 paragraph)

The models were evaluated using three key measures: the area under the receiver-operating-characteristic curve (AUC), Brier score, and bootstrapping. AUC was used to assess the model's discrimination ability, or its capacity to differentiate between cancer and benign cases, with higher values indicating better performance. The AUC values were visualized as AUC curves. The Brier score measured model calibration by comparing predicted probabilities with actual outcomes, where lower scores reflect better alignment between predictions and reality. In the box plots that were used to visualize the Brier scores, it can be clearly seen that the TREAT model has lower Brier scores, indicating better calibration compared to the Mayo Clinic model in both cohorts. Bootstrapping was employed to estimate the standard errors of model parameters and predictions, as well as to assess the degree of optimism of the model's accuracy when predicting cancer. The estimated model coefficients were shown in a summary table, along with the odds ratios and p-values. The odds ratios for the TREAT model provided a more intuitive interpretation of the effect size. For example, an OR of 1.05 for age suggests that each additional year increases the odds of lung cancer by 5%. The p-values indicate the statistical significance of each predictor in the TREAT model. A p-value less than 0.05 suggests that the variable is a statistically significant predictor of malignancy.

3. What were some limitations that the paper addressed? (1 paragraph)

6

The paper addressed several limitations, including the retrospective nature of the cohort used for model development, which was drawn from a single tertiary academic medical center. The external validation cohort, which had a high prevalence of disease, may have contributed to improved discrimination but was primarily composed of male smokers from a Veteran population. While these factors did not affect the model's AUC or Brier score, the authors note that missing data for key predictors was a limitation, and the data may not have been missing at random. To improve generalizability, future validations of the TREAT model should use cohorts with different disease prevalences and from various regions. Additionally, prospective evaluation is needed to determine risk cut points that balance the risks of missing a lung cancer with unnecessary surgeries.

Model Building Practice

Read the NEJM editorial to understand the background of developing the equations that are used to calculate the estimated glomerular filtration rate (eGFR). Then, load in the data `baseseg.csv` and run the pre-processing below. The data contains the following variables.

1. Base serum Creatinine (bascre)

2. Systolic blood pressure (sbase)

3. Diastolic blood pressure (dbase)

4. Urine protein (baseu)

5. Age (age)

6. Sex (Sex = 1 if male; = 0 if female)

7. Indicator if African-American (black)

8. Measured glomerular filtration rate (gfr)

In this worksheet, we will build a model to calculate the eGFR and practice our model evaluation skills.

1. Conduct a brief exploratory data analysis (EDA). Check the distributions of the variables and consider whether transformations are necessary. Hints: Log transformations or polynomial transformations may be helpful.

2. Build a linear regression model with appropriate variable selection. Check the model assumptions using model diagnostics. You may also consider including interaction terms.

3. Evaluate the performance of the model using evaluation measures. Using your evaluation and your estimated model, comment on how you useful you expect the model to be in practice.

4. Evaluate the performance of the model now between race populations. In particular, compare the measured and estimated GFR using mean squared error (MSE), bias, and the percentage of estimates within 10% and 30% of the measured GFR (P_{10} (%) and P_{30} (%), respectively). Visualize the comparison of the measured and estimated GFR.

5. Repeat steps 1-4 but remove the race variable (black) from consideration. Interpret your results and relate them back to the discussion in the editorial.

6. Last, write a non-technical summary of one of your models and its evaluation (1 paragraph) for a clinical audience.

1.

In `Figure 1`, we first examine the distribution of the response variable, `gfr`, across sex and race groups. Observing the boxplot, we found there is an outlier in the non-black male group which is later removed from our analysis. The distribution does not exhibit significant differences between sex groups, except males have a slightly wider range of values. In addition, males have slightly higher median values compared to females. For race, non-Black participants exhibit more variability in their `gfr` values with wider range showing on the plot. The distribution across all groups seems to be skewed to the right, especially With longer right whiskers in the male and the non-black groups.

Figure 1: Boxplot of GFR by Sex and Race
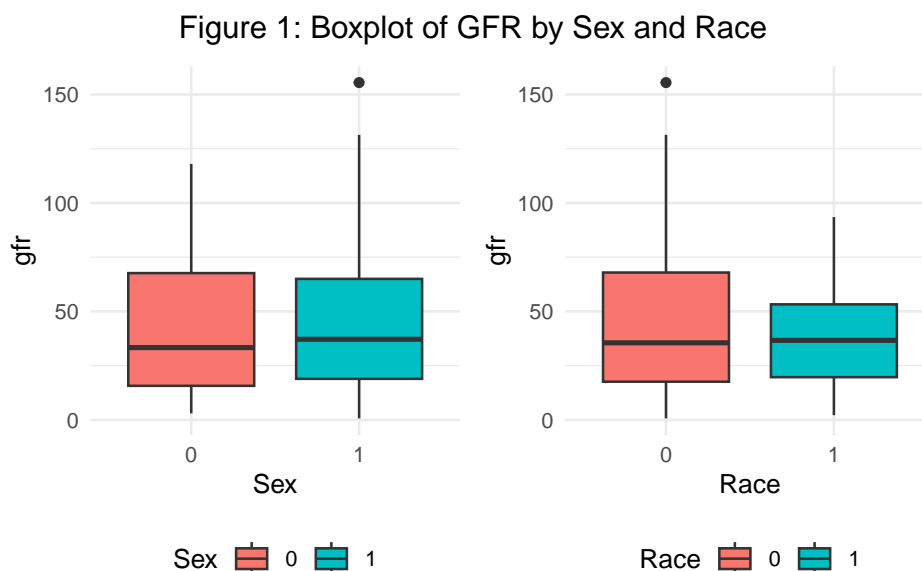
Next, we plot the distribution of the response and all key continuous variables in `Figure 2`. Among those variables, `sbase` and `dbase` appear close to normal and symmetric. `gfr` and `age` exhibit slight skewness, with `gfr` skewed to right and `age` skewed to left. `bascre` and `baseu`

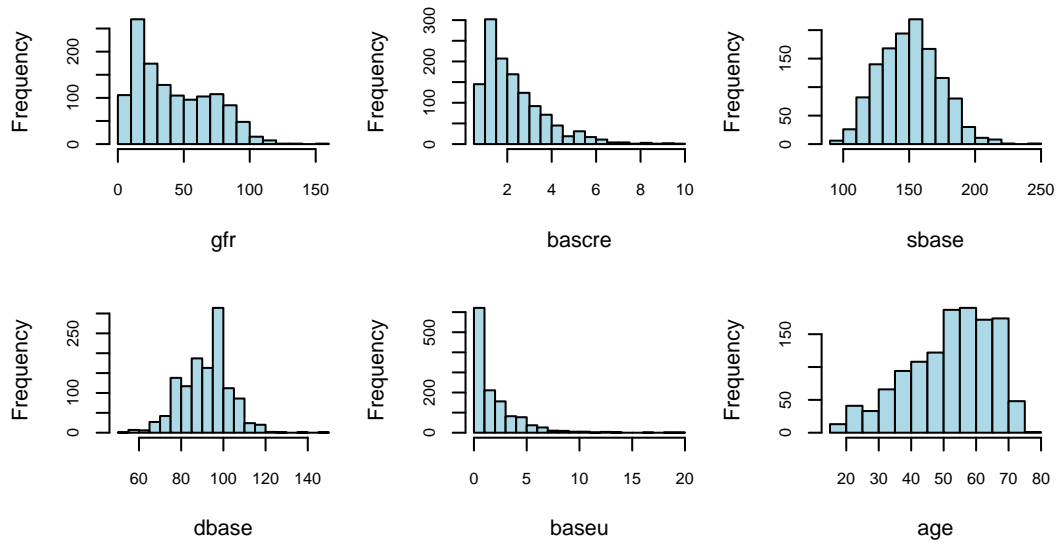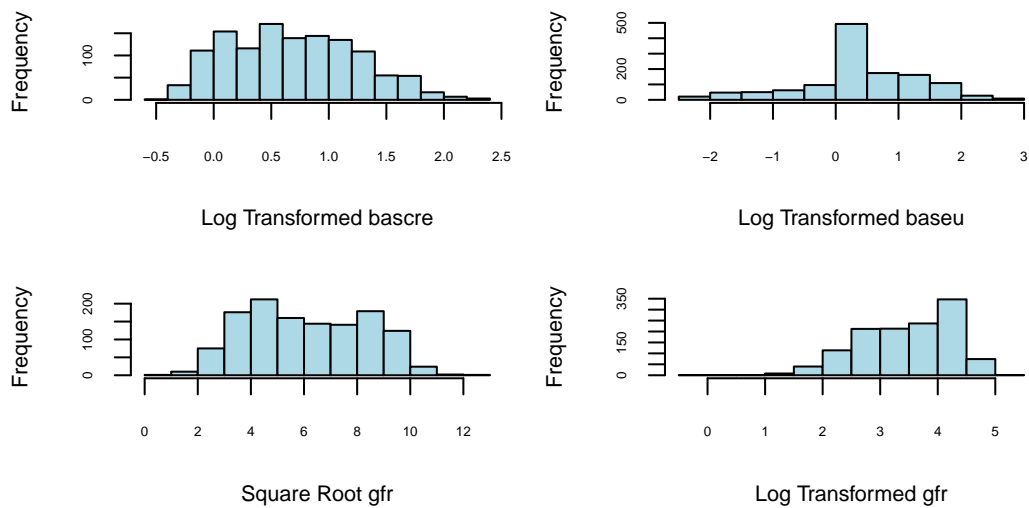Figure 2: Distribution of Key Variables (Before Transformation)

Figure 3: Distribution of Transformed Variables

9

appear to be highly right-skewed with most observations close to 0. Based on the distribution plot, we decide to perform transformation on `gfr`, `bascre`, and `baseu`.

For `bascre`, we directly perform log transformation on the value to normalize its distribution. However, exploring the summary statistics of `baseu`, we found over 25% of observations have 0.1 which is near to zero. Thus, we set a threshold of 0.1 to have all observations with `baseu` less than or equal to 0.1 to remain their original value and perform log transformation on others. Additionally, we tried both log transformation and square root transformation on our response. Observing the result distribution plot, we decided to use the square root transformation since it appeared more normal compared to the log transformation. All transformed variables' distributions are presented in `Figure 3`.

Table 1: Statistical Summary of Key Variables by Race and Transformation

Characteristic	Original Scale		Transformed Scale	
	Non-Black, N = 1,135	Black, N = 114	Non-Black, N = 1,135	Black, N = 114
age	52 (13)	52 (11)	52 (13)	52 (11)
sex				
Female	425 (37%)	41 (36%)	425 (37%)	41 (36%)
Male	710 (63%)	73 (64%)	710 (63%)	73 (64%)
Measured glomerular filtration rate(gfr)	43 (29)	38 (21)	6.16 (2.27)	5.90 (1.77)
Base serum Creatinine(bascre)	2.31 (1.41)	2.55 (1.32)	0.67 (0.56)	0.83 (0.43)
Systolic blood pressure(sbase)	154 (23)	139 (24)	154 (23)	139 (24)
Diastolic blood pressure(dbase)	93 (11)	89 (12)	93 (11)	89 (12)
Urine protein(baseu)	1.91 (2.32)	0.94 (1.99)	0.41 (0.89)	-0.42 (1.07)

[1] Mean (SD) for continuous; n (%) for categorical

We also generate a summary statistics table stratified by race in `Table 1`. From the table we see there are 1135 non-black observations and 114 black observations. Age shows similar distribution between race groups with mean age of 52 years old and standard deviation of 13 for non-black observations and 11 for black observations. There are generally more male observations compared to female observations in both race groups; around 60% of participants are males and 40% are females in both groups. For measured glomerular filtration rate (`gfr`), the mean value for non-black observation is higher than that for black observations (43 vs. 38) with a greater standard deviation as well (29 vs. 21). This pattern remains consistent after applying the square root transformation. Black people exhibit a higher base serum creatinine (`bascre`) value compared to non-black groups, but with lower variability. For the two blood pressure variables, systolic blood pressure (`sbase`) and diastolic blood pressure (`dbase`), non-black people appear to have higher value on average with lower variability. Finally, non-black people exhibit higher urine protein (`baseu`) with greater variability compared to black observations.

Finally, we plot the correlation value among the response and continuous key variables. The `sbase` and `dbase` have a strong positive correlation of 0.7 which makes sense since the systolic and diastolic blood pressure often vary together. In addition, our response variables `sqrt_gfr`

and `log_bascre` show strong negative correlation of -0.86 which indicates that higher creatinine levels might be associated with lower kidney filtration rates.

Figure 4: Correlation Matrix
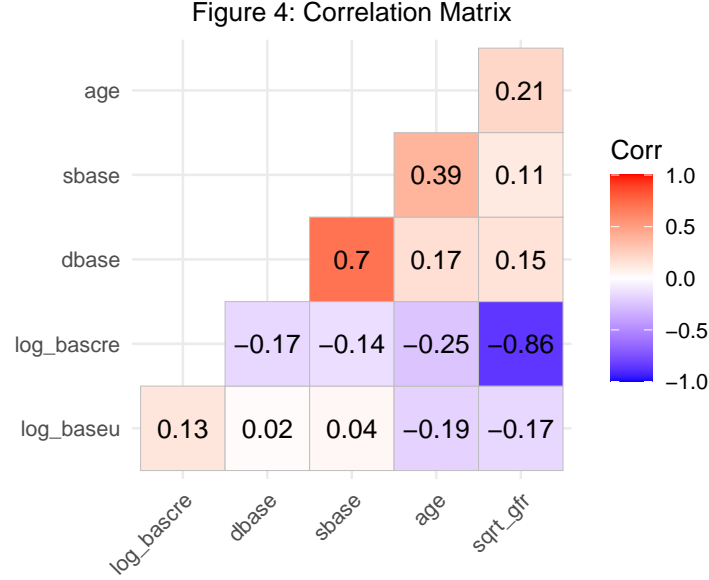
2.

To build a linear regression model and perform a variable selection process, we start from a `full_model` with all main effects and all of their possible interaction terms. Here, We first create a cross validation train and test data with 80% and 20% of our original data, respectively. To be noticed, since we found `dbase` and `sbase` are highly correlated from the EDA part, we decided to include only `dbase` in our model to avoid multicollinearity. Then, using the full model, we perform a backward model selection using `step()` and get the optimal model with the following format, and the estimated coefficient of the optimal model is presented in `Table 2`.

$$
\begin{aligned}
\text{sqrt\_gfr} = {} & \beta_0 + \beta_1 \cdot \text{dbase} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex}_1 + \beta_4 \cdot \text{black}_1 \\
& + \beta_5 \cdot \log(\text{bascre}) + \beta_6 \cdot \log(\text{baseu}) + \beta_7 \cdot \text{dbase} \times \log(\text{bascre}) \\
& + \beta_8 \cdot \text{dbase} \times \log(\text{baseu}) + \beta_9 \cdot \text{age} \times \text{sex}_1 \\
& + \beta_{10} \cdot \text{age} \times \log(\text{baseu}) + \beta_{11} \cdot \text{sex}_1 \times \text{black}_1 \\
& + \beta_{12} \cdot \text{sex}_1 \times \log(\text{bascre}) + \beta_{13} \cdot \text{black}_1 \times \log(\text{bascre}) \\
& + \beta_{14} \cdot \log(\text{bascre}) \times \log(\text{baseu}) + \epsilon
\end{aligned}
$$

Observing the optimal model output in `Table 2`, we see main effects, `dbase`, `sex`, `black`, `log_bascre`, and `log_baseu`, are significant predictors of the response variable, sqrt_gfr.

Among these main effect variables, `black`, `log_bascre`, and `log_baseu` have negative coefficient estimates, which implies a negative association between these variables and the response variables.

Table 2: Coefficient Estimate of the Best Model

	Estimate	Standard Error	T Statistics	P Value
(Intercept)	6.0379881	0.5681945	10.626622	0.0000000
dbase	0.0196513	0.0055729	3.526249	0.0004409
age	0.0051667	0.0045765	1.128948	0.2591941
sex1	2.0026398	0.3445638	5.812102	0.0000000
black1	-0.9237535	0.2910695	-3.173653	0.0015519
log_bascre	-1.1812960	0.5629000	-2.098589	0.0361067
log_baseu	-0.6911598	0.1742049	-3.967511	0.0000779
dbase:log_bascre	-0.0236057	0.0059399	-3.974103	0.0000758
age:sex1	-0.0208493	0.0057512	-3.625218	0.0003035
age:log_baseu	0.0060477	0.0030267	1.998112	0.0459786
sex1:black1	0.4823647	0.2365373	2.039275	0.0416885
sex1:log_bascre	-0.5409619	0.1341886	-4.031355	0.0000597
black1:log_bascre	0.8780471	0.2633363	3.334319	0.0008869
log_bascre:log_baseu	0.3146070	0.0796565	3.949545	0.0000839

For example, black with a coefficient estimate of -0.774 indicates black people tend to have a lower outcome value by 0.77 units compared to the non-black participants. The coefficients for `log_bascre` (-1.216) and `log_baseu` (-1.025) are significant with p-values of 0.0137 and 0.0007, respectively. This indicates that increase in log-transformed base serum creatinine and urine protein values are associated with decrease in `sqrt_gfr`. The coefficient of `dbase`, 0.019, is positive and highly significant. The `sqrt_gfr` would increase by 0.019 units as `dbase` increases by one unit. Additionally, sex shows that males' `sqrt_gfr` values are approximately 1.75 units higher than females and this difference is statistically significant.

Interaction terms play a significant role in our model as well which indicates the complex relationships among our key covariates. The interaction term between `dbase` and `log_bascre` has a significant negative estimate, indicating that higher blood pressure with higher creatinine would lead to decrease in the response variable, `sqrt_gfr`. The significant interaction between age and sex implies that the effect of age on our outcome would vary by sex. Since this interaction term shows a significant negative estimate, male participants experience a greater decrease (or a smaller increase) in the response variable with age compared to female participants. Moreover, the interactions between sex and `log_bascre`, and race and `log_bascre`, suggest that both gender and race influence how creatinine affects the response variable. Finally, the interaction between `log_bascre` and `log_baseu` exhibits a significant positive coefficient estimate showing that the combined effects of these two biomarkers is greater than their individual effects.
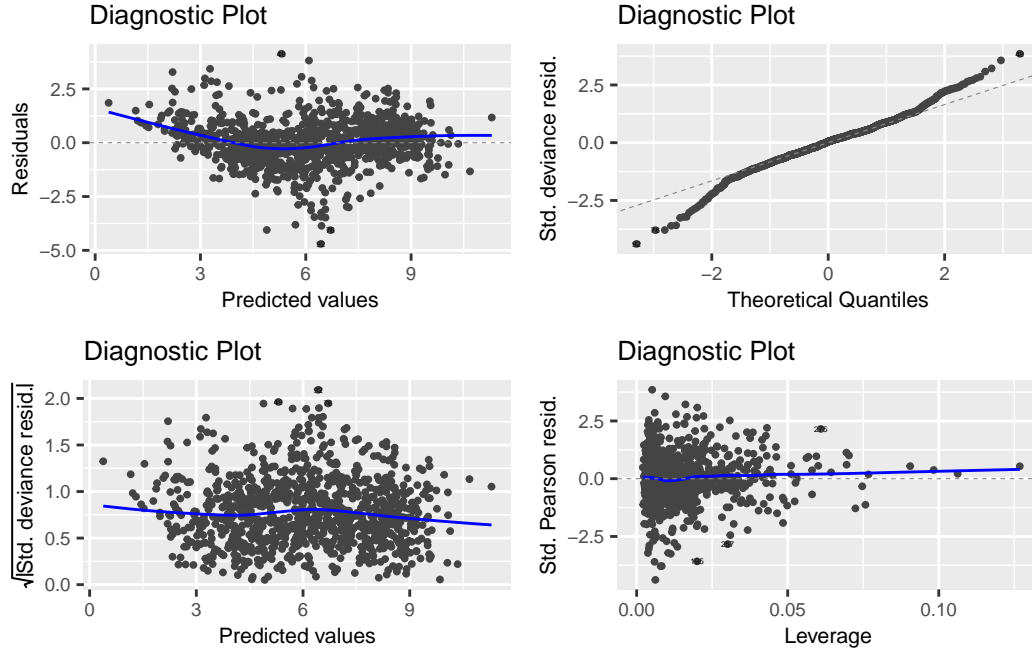
To check the essential assumptions of our model, we first observe the residual plot to check for linearity and homoscedasticity. Our points here on the left top plot generally have no obvious pattern and they randomly scattered around the center line of zero. Additionally, we look at the Q-Q plot on the right top to assess the normality assumption. Our points mostly fall along the diagonal line in the middle of the graph, but curve off in the extremities. This suggests that the residuals follow normal distribution but the presence of more extreme values than expected lead to the deviation on the tails. The scale-location plot on the left bottom checks for homoscedasticity. In this plot, we see the points appear mostly randomly scattered and it shows a reasonably flat blue line with no pronounce variation. This indicates that the variance of residuals remains roughly constant across predicted values. The residual vs. leverage plot on the right bottom aims to identify influential points. Here, most points clustered around zero and there are a few points which exhibit relatively high leverage but none of these points appear to have extremely large residuals.

3.

Here, to evaluate the model performance using evaluation measures, we choose to use the Mean Square Error (MSE) and Mean Absolute Error (MAE) presented in `Table 3`. Using the test set, we calculated the predicted value from our optimal model and squared each predicted value to the original phase of `gfr`. Then, we calculate the model MSE and MAE using the predicted and observed `gfr` values.

Table 3: Evaluation Measures

Measures	Values
MSE	166.755389
MAE	9.406527

In our model, the MSE is 166.76, meaning that on average, the squared difference between our predictions and the actual GFR values is around 166. In addition, the MAE is 9.41, meaning that on average, the model's predictions are about 9 units off from the true GFR values.

Our observed GFR value ranges from 0.7 to 155.50 with a standard deviation of 28.38. Our MAE value, 9.41, is relatively small compared to the standard deviation of GFR in the data. The Root Mean Squared Error (RMSE), which is the square root of the MSE, is around 12.91. The RMSE is close to the MAE, further suggesting that our model performs reasonably well in capturing the variability of GFR. However, although the relatively smaller prediction error value suggests that our model performs reasonably well to capture the variability of GFR, it still can be clinically important. Especially for patients with lower GFR value, an 9 units error would significantly influence the medical decision. Thus, we still need further investigation on how to improve our model, especially for patients with extreme GFR values like what we mentioned in the model assumption part.

Table 4: Percentage of Estimates within Various Ranges

Range	Value
Within 10%	29.43548
Within 30%	63.30645
Within 50%	79.43548

In `Table 4`, we present the percentage of estimates within 10%, 30%, and 50% of the measured GFR (P_{10} (%), P_{30} (%), and P_{50} (%), respectively). About 29.44% of the model predictions fall within 10% of the actual observed GFR values that one-third of our predictions are very close to the true values. About 63.31% of predicted values are within 30% of the observed values, and the majority of our predictions are within a reasonable range of our observed values. Finally, around 79.44% of our predictions are within 50% of the true values. This indicates that most predictions are moderately close to the actual values, but there are still 20% of them which deviate by more than 50% from the observed value.

Overall, while the model generally captures the variability in GFR, further investigation and improvements are needed to refine its accuracy, particularly for observations with extreme values.

4.

Table 5: Evaluation Measures by Race

black	MSE	Bias	P10	P30
0	157.9485	0.1182867	29.48718	63.67521
1	313.9562	-8.5498308	28.57143	57.14286

`Table 5` shows the mean-square error between the measured and estimated values of GFR, the bias, and the percentage of the estimated values within 10% and 30% of the measured GFR. The MSE was much greater in black patients than in non-black patients, suggesting that the model is notably less accurate in predicting GFR in black patients and that the model may not generalize well across racial groups. This might be due to an imbalance in the sample size of the data between the non-black and black groups, with black patients comprising only 10% of the data. The GFR values for non-blacks were slightly positively biased (0.12), while the model underestimated the GFR values for blacks with a negative bias of -8.55. For the percentages of the estimates, we can see that the predicted values are similar for both groups, with the predicted values for the non-black group being slightly better than the predicted values for the black group. This suggests that the model is able to predict GFR more precisely for non-blacks, but still captures the overall trend in GFR for both groups reasonably well.
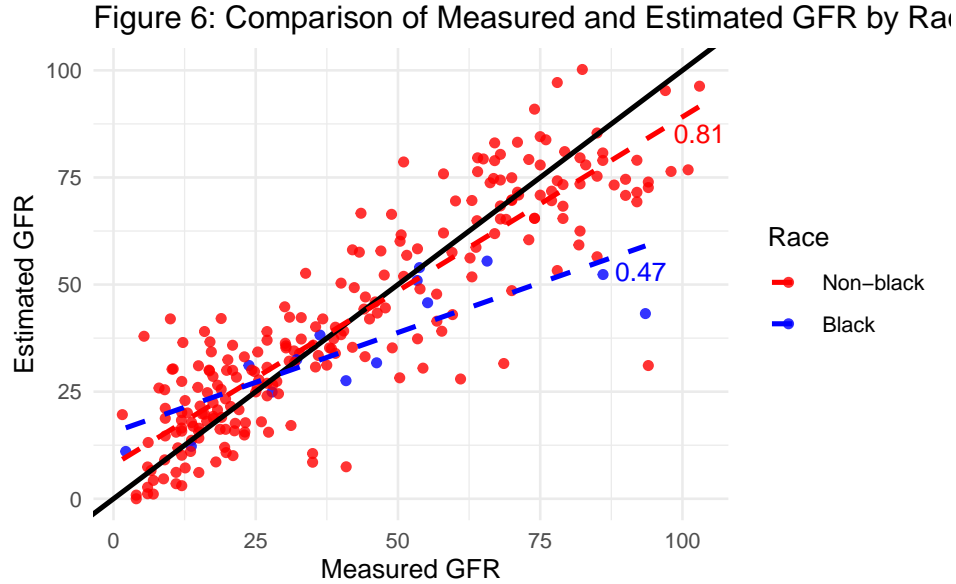

Figure 6: Comparison of Measured and Estimated GFR by Race

`Figure 6` shows the relationship between measured and estimated GFR by race, with the x-axis being the measured GFR and the y-axis being the estimated GFR. The black diagonal solid line represents a perfect fit, i.e., the estimated GFR is equal to the measured GFR. As we can see from the figure, the red line for non-black patients is closer to the line of perfect

fit than the blue line for black patients. This difference suggests that the accuracy of the model's predictions may differ between these racial groups, and that the model appears to better predict non-black patients.

5.

`Figure` 7 illustrates the distribution of the square root of GFR (SR_GFR) across three population groups: all individuals, black individuals, and non-black individuals. We can see that the SR_GFR distribution of the non-black group highly overlaps with the SR_GFR distribution of the overall population, whereas the SR_GFR distribution of the black group is significantly different compared to these two groups. The SR_GFR for the non-black and overall population showed a bimodal distribution, while the SR_GFR for the black group was more concentrated, with more values around 6. This coincided with the statement in the editorial that there was a difference in GFR between blacks and nonblacks when GFR was measured directly from endogenous clearance of creatinine.

Figure 7: Density Plot of Square Root GFR by Population Groups
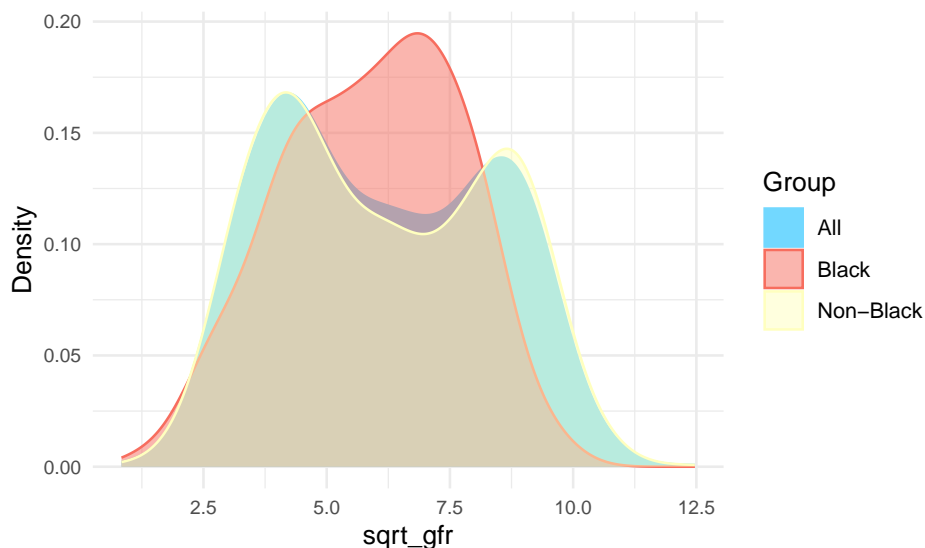
We also generate a summary statistics table with all observations without stratification by race in `Table` 6. In the race removed column, there are totally 1249 observations with 466 females (37%) and 783 males (63%). The mean value of age is 52 with a standard deviation of 13 which matches the Non-Black group in the "Transformed Scale" column. This suggests that the overall age distribution is influenced more by the Non-Black group, likely due to its larger size. The square root `gfr` value is 6.13 on average with a standard deviation of 2.23 which falls between the non-black and black groups under the transformed scale. The remaining biomarkers, such as base serum creatinine, systolic blood pressure, diastolic blood pressure, and urine protein, exhibit values that fall between the non-Black and black groups values.

However, these values are closer to the non-black group's values due to the larger proportion of non-black observations.

Table 6: Summary Table by Population Groups

Characteristic	Transformed Scale		Transformed Scale (Race Removed)
	Non-Black, N = 1,135	Black, N = 114	N = 1,249
age	52 (13)	52 (11)	52 (13)
sex			
Female	425 (37%)	41 (36%)	466 (37%)
Male	710 (63%)	73 (64%)	783 (63%)
Measured glomerular filtration rate(gfr)	6.16 (2.27)	5.90 (1.77)	6.13 (2.23)
Base serum Creatinine(bascre)	0.67 (0.56)	0.83 (0.43)	0.69 (0.55)
Systolic blood pressure(sbase)	154 (23)	139 (24)	152 (23)
Diastolic blood pressure(dbase)	93 (11)	89 (12)	93 (12)
Urine protein(baseu)	0.41 (0.89)	-0.42 (1.07)	0.33 (0.94)

[1] Mean (SD) for continuous; n (%) for categorical

To build a linear regression model and perform a variable selection process, we start from a full model with all main effects except for black and all of their possible interaction terms. As what we done previously, We again create a cross validation train and test data with 80% and 20% of our original data, respectively. Then, using the full model, we perform a backward model selection using the step() and get the optimal model with the following format, and the estimated coefficient of the optimal model is presented in `Table 7`.

$$
\begin{aligned}
\text{sqrt\_gfr} = {} & \beta_0 + \beta_1 \cdot \text{dbase} + \beta_2 \cdot \text{age} + \beta_3 \cdot \text{sex}_1 + \beta_4 \cdot \log(\text{bascre}) \\
& + \beta_5 \cdot \log(\text{baseu}) + \beta_6 \cdot \text{dbase} \times \text{sex}_1 + \beta_7 \cdot \text{dbase} \times \log(\text{bascre}) \\
& + \beta_8 \cdot \text{dbase} \times \log(\text{baseu}) + \beta_9 \cdot \text{age} \times \text{sex}_1 \\
& + \beta_{10} \cdot \text{age} \times \log(\text{baseu}) + \beta_{11} \cdot \text{sex}_1 \times \log(\text{bascre}) \\
& + \beta_{12} \cdot \log(\text{bascre}) \times \log(\text{baseu}) + \epsilon
\end{aligned}
$$

Among the five main effects produced by the optimal model, diastolic blood pressure (`dbase`), gender, and urine protein appeared to have a significant effect on the outcome GFR, with a p-value of less than 0.05. GFR was significantly higher in males than in females, with lower urine protein associated with higher GFR, and diastolic blood pressure associated with higher GFR. The main effect of basal serum creatinine became insignificant compared with the model that included "black". The interaction terms in the coefficient estimation tables reveal how different variables work together to influence the results. For example, the negative coefficient (-0.0261) for the interaction between diastolic blood pressure and basal serum creatinine suggests that the combined effect of `dbase` and `log_bascre` on the outcome decreases as they increase.

Table 7: Coefficient Estimate of the Best Model without Race

	Estimate	Standard Error	T Statistics	P Value
(Intercept)	5.2876638	0.6831349	7.740292	0.0000000
dbase	0.0281360	0.0070534	3.988984	0.0000713
age	0.0033208	0.0046435	0.715167	0.4746748
sex1	2.9465307	0.6812559	4.325145	0.0000168
log_bascre	-0.8970941	0.5656845	-1.585856	0.1130920
log_baseu	-1.0704228	0.3410005	-3.139066	0.0017452
dbase:sex1	-0.0110526	0.0066745	-1.655948	0.0980499
dbase:log_bascre	-0.0260957	0.0059738	-4.368364	0.0000138
dbase:log_baseu	0.0046629	0.0032307	1.443321	0.1492471
age:sex1	-0.0187454	0.0058393	-3.210209	0.0013692
age:log_baseu	0.0055721	0.0030406	1.832563	0.0671685
sex1:log_bascre	-0.5225945	0.1354892	-3.857093	0.0001222
log_bascre:log_baseu	0.2901869	0.0790381	3.671483	0.0002541

To check the essential assumptions of our model without the race variable, we first observe the residual plot to check for linearity and homoscedasticity. In general, these diagnostic plots appear similar to what we have for the optimal model with the race variable. In the residual plot, points generally have no obvious pattern and they randomly scattered around the center line of zero. The Q-Q plot has points mostly falling along the diagonal line in the middle of the graph, but curve off in the extremities. This suggests that the residuals follow normal distribution but the presence of more extreme values than expected lead to the deviation on the tails. In the scale-location plot, we see the points appear mostly randomly scattered and it shows a reasonably flat blue line with no pronounced variation. This indicates that the variance of residuals remains roughly constant across predicted values. Finally, in the residual vs. leverage plot, most points clustered around zero and there are a few points which exhibit relatively high leverage but none of these points appear to have extremely large residuals.
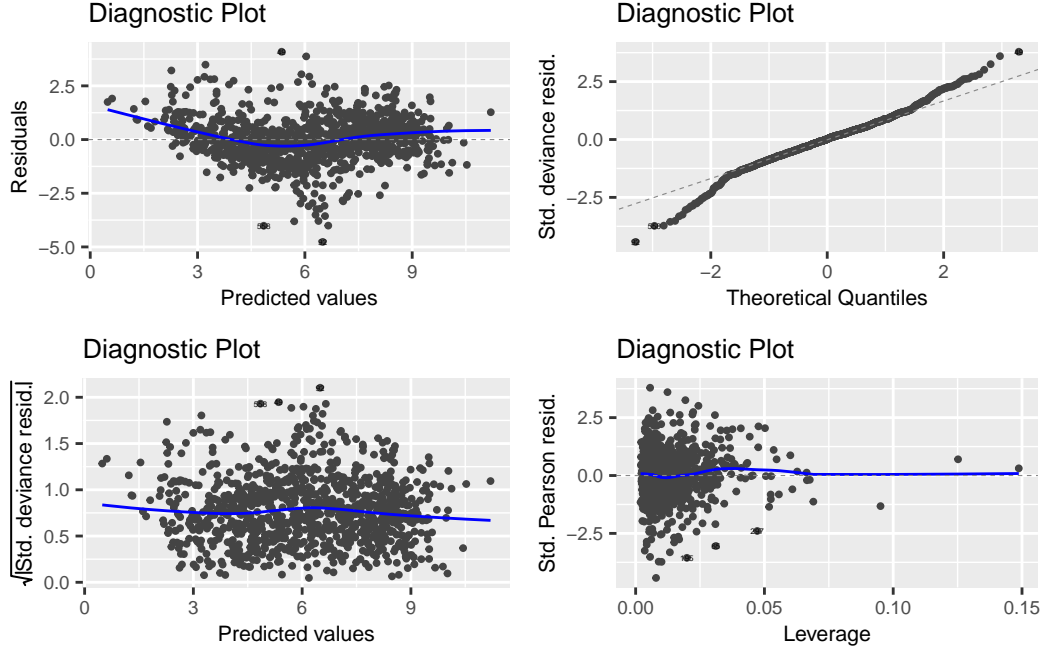
Table 8 shows the evaluation measures for models including or excluding the race variable `black`, a binary variable that indicates whether a person is Black. The mean squared error (MSE) of the model excluding `black` was 168.63, indicating a poor fit because exclusion leads to an increase in prediction error. The MSEs of both models appeared to be more closely aligned with the MSEs of non-Black individuals. The MAEs for the two models are almost identical, both indicating an average deviation of 9 units between the estimated and measured values of GFR. In addition, both models appeared to underestimate the GFR by 0.4 units. After removing Black from the model, the percentage of individuals with estimates within 10% and 30% of the measured value did not change much. In both models, around 63% of individuals had estimates within 30% of the measured value.

Table 8: Evaluation Measures for Models With and Without Black

Model	MSE	MAE	Bias	P10	P30
With Black	**166.7554**	**9.406527**	**-0.3710425**	**29.43548**	**63.30645**
Non-black	157.9485	9.299234	0.1182867	29.48718	63.67521
Black	313.9562	11.199852	-8.5498308	28.57143	57.14286
Without Black	**168.6285**	**9.481655**	**-0.4374376**	**29.83871**	**62.50000**

Figure 8 shows the relationship between the measured and estimated GFR , with x-axis being the observed GFR values and the y-axis being the estimated GFR values. The black solid diagonal line represents a perfect fit which indicates perfect agreement between measured and estimated GFR. The blue dashed line represents the regression line fitted to all observations. This line captures the relationship between the observed and predicted GFR values generated

by the model. The slope of the fitted line appears flatter compared to the diagonal line which indicates that our model underestimated the outcome GFR values.

Figure 8: Comparison of Measured and Estimated GFR (Excluding Race in Model)

6.

In one of our optimal models for evaluating patient GFR measurements, we examined the effect of key demographic and clinical factors (e.g., age, gender, race, and baseline kidney function) on the prediction of the outcome glomerular filtration rate (GFR). The model showed that GFR was significantly higher in male patients than in females, whereas black patients had lower GFR values than non-black patients. Kidney function, as measured by creatinine and urine protien levels, was also important; higher values were associated with lower GFR values, which meant that the kidneys were not working well or the patients were more likely to have kidney disease. Importantly, the model allowed us to capture complex interactions, such as the combined effects of baseline kidney function and patient demographic characteristics. These findings suggest that patient characteristics, particularly gender and ethnicity, as well as baseline clinical metrics, can have a meaningful impact on predicted health outcomes, thereby guiding a more personalized approach to treatment.

`Figure 6` shows the relationship between measured GFR and predicted GFR by race using the optimal model, with the black diagonal solid line representing a perfect fit, i.e., the estimated GFR is equal to the measured GFR. The red line for non-black patients is closer to the line of perfect fit than the blue line for black patients. This discrepancy suggests that the accuracy of the predictions of the model may differ between these racial groups and that the model appears to better predict the non-black patients. One of the limitations was the imbalance between the non-black and black subgroups of the population included in the model, which suggested that the model might be biased, preventing the predictions from generalizing to the entire population. In addition, the model appeared to correctly estimate GFR in about

two-thirds of the patients, with estimates within 30% of the measured values, but with a slight tendency to underestimate, on average, by about 8 unit (`Table 5`).

Appendix

```r
suppressPackageStartupMessages(library(tidyverse))
suppressPackageStartupMessages(library(moments))
suppressPackageStartupMessages(library(stats))
suppressPackageStartupMessages(library(kableExtra))
suppressPackageStartupMessages(library(dplyr))
suppressPackageStartupMessages(library(gtsummary))
suppressPackageStartupMessages(library(gt))
suppressPackageStartupMessages(library(corrplot))
suppressPackageStartupMessages(library(ggcorrplot))
suppressPackageStartupMessages(library(MASS))
suppressPackageStartupMessages(library(ggplot2))
suppressPackageStartupMessages(library(ggfortify))
suppressPackageStartupMessages(library(caret))
suppressPackageStartupMessages(library(kableExtra))
suppressPackageStartupMessages(library(ggpubr))
suppressPackageStartupMessages(library(patchwork))
suppressPackageStartupMessages(library(gridExtra))
suppressPackageStartupMessages(library(grid))


kidney_df <- read.csv("baseseg.csv")
kidney_df <- kidney_df %>%
  dplyr::select(gfr, bascre, sbase, dbase, baseu, AGE, SEX, black) %>%
  rename(sex = SEX, age = AGE) %>%
  na.omit()

kidney_df$black <- as.factor(kidney_df$black)
kidney_df$sex <- as.factor(kidney_df$sex)
# Create the sex plot
sex_plot <- ggplot(kidney_df) +
  geom_boxplot(aes(x = as.factor(sex), y = gfr, fill = as.factor(sex))) +
  theme_minimal() +
  labs(x = "Sex", y = "gfr", fill = "Sex")

# Create the race plot
race_plot <- ggplot(kidney_df) +
  geom_boxplot(aes(x = as.factor(black), y = gfr, fill = as.factor(black))) +
  theme_minimal() +
  labs(x = "Race", y = "gfr", fill = "Race")
```

```
# Set the aspect ratio
sex_plot <- sex_plot + theme(aspect.ratio = 1)
race_plot <- race_plot + theme(aspect.ratio = 1)

# Arrange the plots and add a title
figure <- ggarrange(sex_plot, race_plot,
                    ncol = 2, nrow = 1,
                    align = "hv",
                    legend = "bottom")

# Add title using annotate_figure()
annotate_figure(figure,
                top = text_grob("Figure 1: Boxplot of GFR by Sex and Race", size = 13))

par(mfrow = c(2, 3), mar = c(4, 4, 2, 1), oma = c(0, 0, 4, 0))
hist(kidney_df$gfr, main = "", xlab = "gfr", col = "lightblue", breaks = 15,
     cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
hist(kidney_df$bascre, main = "", xlab = "bascre", col = "lightblue", breaks = 15,
     cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
hist(kidney_df$sbase, main = "", xlab = "sbase", col = "lightblue", breaks = 15,
     cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
hist(kidney_df$dbase, main = "", xlab = "dbase", col = "lightblue", breaks = 15,
     cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
hist(kidney_df$baseu, main = "", xlab = "baseu", col = "lightblue", breaks = 15,
     cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)
hist(kidney_df$age, main = "", xlab = "age", col = "lightblue", breaks = 15,
     cex.main = 1, cex.lab = 1, cex.axis = 0.8, font.main = 1)

mtext("Figure 2: Distribution of Key Variables (Before Transformation)", outer = TRUE, cex

kidney_df$log_bascre <- log(kidney_df$bascre)
kidney_df$log_baseu <- ifelse(kidney_df$baseu <= 0.1, kidney_df$baseu, log(kidney_df$baseu
kidney_df$sqrt_gfr <- sqrt(kidney_df$gfr)
kidney_df$log_gfr <- log(kidney_df$gfr)

par(mfrow = c(2, 2), mar = c(4, 4, 2, 1), oma = c(0, 0, 4, 0))
hist(kidney_df$log_bascre, main = "", xlab = "Log Transformed bascre", col = "lightblue",
     cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.5, font.main = 1)
hist(kidney_df$log_baseu, main = "", xlab = "Log Transformed baseu", col = "lightblue", br
     cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.5, font.main = 1)
hist(kidney_df$sqrt_gfr, main = "", xlab = "Square Root gfr", col = "lightblue", breaks =
```

```r
      cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.5, font.main = 1)
hist(kidney_df$log_gfr, main = "", xlab = "Log Transformed gfr", col = "lightblue", breaks
      cex.main = 0.8, cex.lab = 0.8, cex.axis = 0.5, font.main = 1)

mtext("Figure 3: Distribution of Transformed Variables", outer = TRUE, cex = 0.8, font = 1
# Creating first summary table (original)
kidney_tbl <- kidney_df %>%
  dplyr::select(black, age, sex, gfr, bascre, sbase, dbase, baseu) %>%
  mutate(black = recode(black, `0` = "Non-Black", `1` ="Black"),
         sex = recode(sex, `0` = "Female", `1` ="Male")) %>%
  tbl_summary(by = black,
              label = list(gfr ~ "Measured glomerular filtration rate(gfr)",
                           bascre ~ "Base serum Creatinine(bascre)",
                           sbase ~ "Systolic blood pressure(sbase)",
                           dbase ~ "Diastolic blood pressure(dbase)",
                           baseu ~ "Urine protein(baseu)"
                           ),
              statistic = all_continuous() ~ "{mean} ({sd})") %>%
  modify_spanning_header(update =  all_stat_cols() ~  "**Black**") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categori
  bold_labels()

# Creating second summary table (transformed)
kidney_tbl_transformed <- kidney_df %>%
  dplyr::select(black, age, sex, sqrt_gfr, log_bascre, sbase, dbase, log_baseu) %>%
  rename(gfr=sqrt_gfr, bascre=log_bascre, baseu=log_baseu) %>%
  mutate(black = recode(black, `0` = "Non-Black", `1` ="Black"),
         sex = recode(sex, `0` = "Female", `1` ="Male")) %>%
  tbl_summary(by = black,
              label = list(gfr ~ "Measured glomerular filtration rate(gfr)",
                           bascre ~ "Base serum Creatinine(bascre)",
                           sbase ~ "Systolic blood pressure(sbase)",
                           dbase ~ "Diastolic blood pressure(dbase)",
                           baseu ~ "Urine protein(baseu)"
                           ),
              statistic = all_continuous() ~ "{mean} ({sd})") %>%
  modify_spanning_header(update =  all_stat_cols() ~  "**Black**") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categori
  bold_labels()

tbl_combined1 <- tbl_merge(
```

```
  tbls = list(kidney_tbl, kidney_tbl_transformed),
  tab_spanner = c("**Original Scale**", "**Transformed Scale**")
)

tbl_combined1 <- tbl_combined1 %>%
  as_kable_extra(booktabs = TRUE, caption = "Statistical Summary of Key Variables by Race
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 7, position = "center",
                            latex_options = c("repeat_header", "HOLD_position", "scale_dow
  column_spec(1, width = "5cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  column_spec(4, width = "2cm") %>%
  column_spec(5, width = "2cm") %>%
  row_spec(0, bold = TRUE)

tbl_combined1 <- tbl_combined1 %>%
  row_spec(5, background = "#e0ecf4") %>%
  row_spec(6, background = "#e0ecf4") %>%
  row_spec(9, background = "#e0ecf4")

tbl_combined1
cor_m <- cor(kidney_df[, -c(1, 2, 5, 7, 8, 12)])
variable_order <- c("log_baseu", "log_bascre", "dbase", "sbase", "age", "sqrt_gfr")
r_reordered <- cor_m[variable_order, variable_order]

ggcorrplot(r_reordered,
           hc.order = TRUE,
           type = "lower",
           lab = TRUE) +
  ggtitle("Figure 4: Correlation Matrix") +
  theme(plot.title = element_text(hjust = 0.5, size = 11),
        axis.text.x = element_text(size = 9),
        axis.text.y = element_text(size = 9))
set.seed(2550)
index <- createDataPartition(kidney_df$gfr, p = 0.8, list = FALSE)
train <- kidney_df[index, ]
test <- kidney_df[-index, ]

full_model <- glm(sqrt_gfr ~ dbase + age + sex + black + log_bascre + log_baseu +
                  dbase:age + dbase:sex + dbase:black + dbase:log_bascre + dbase:log_baseu +
```

```r
                  age:sex + age:black + age:log_bascre + age:log_baseu +
                  sex:black + sex:log_bascre + sex:log_baseu +
                  black:log_bascre + black:log_baseu + log_bascre:log_baseu,
                  data = train)

step_model <- step(full_model, direction = "backward", trace = 0)

summary_model <- as.data.frame(summary(step_model)$coefficient)
colnames(summary_model) <- c("Estimate", "Standard Error", "T Statistics", "P Value")

summary_model %>%
  kbl(booktabs = TRUE, caption = "Coefficient Estimate of the Best Model",
      longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 10,
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))
# Create the diagnostic plot
autoplot(step_model, label.size = 1, size = 0.7) +
  ggtitle("Diagnostic Plot") +
  theme(
    plot.title = element_text(size = 10),
    axis.title = element_text(size = 8),
    axis.text = element_text(size = 7),
    main = "Diagnostic Plot"
  )
predictions <- predict(step_model, newdata = test)
test$predictions_transformed <- predictions^2
mse <- mean((test$predictions_transformed - test$gfr)^2)
mae <- mean(abs(test$predictions_transformed - test$gfr))

evaluation_measure_df <- data.frame(Measures = c("MSE", "MAE"),
                                    Values = c(mse, mae))

evaluation_measure_df %>%
  kbl(booktabs = TRUE, caption = "Evaluation Measures",
      longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 10,
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))
range_df <- data.frame(Range = c("Within 10%", "Within 30%", "Within 50%"),
                       Value = c(mean(abs(test$predictions_transformed - test$gfr) / test$
                                 mean(abs(test$predictions_transformed - test$gfr) / test$
                                 mean(abs(test$predictions_transformed - test$gfr) / test$
```

```r
range_df %>%
   kbl(booktabs = TRUE, caption = "Percentage of Estimates within Various Ranges",
      longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 10,
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))
# Split data by race and calculate performance metrics
test$estimated_gfr <- predict(step_model, newdata = test)
performance_by_race <- test %>%
  group_by(black) %>%
  summarise(MSE = mean((predictions_transformed - gfr)^2),
            Bias = mean(predictions_transformed - gfr),
            P10 = mean(abs(predictions_transformed - gfr) / gfr <= 0.10) * 100,
            P30 = mean(abs(predictions_transformed - gfr) / gfr <= 0.30) * 100)
performance_by_race %>%
  kbl(caption = "Evaluation Measures by Race") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))
# Calculate the slope for each group
linear_model_0 <- lm(predictions_transformed ~ gfr, data = test, subset = black == 0)
linear_model_1 <- lm(predictions_transformed ~ gfr, data = test, subset = black == 1)

slope_0 <- round(coef(linear_model_0)[2], 2)
slope_1 <- round(coef(linear_model_1)[2], 2)

# Create the plot
ggplot(test, aes(x = gfr, y = predictions_transformed, color = factor(black))) +
  geom_point(alpha = 0.8) +
  geom_abline(intercept = 0, slope = 1, linetype = "solid", color = "black", size = 1) +
  geom_smooth(method = "lm", linetype = "dashed", se = FALSE, alpha = 0.8) +
  scale_color_manual(values = c("0" = "red", "1" = "blue"),
                     labels = c("0" = "Non-black", "1" = "Black"),
                     name = "Race") +
  labs(title = "Figure 6: Comparison of Measured and Estimated GFR by Race",
       x = "Measured GFR",
       y = "Estimated GFR") +
  theme_minimal() +
  annotate("text", x = max(test$gfr) * 1, y = max(test$predictions_transformed) * 0.85,
           label = slope_0, color = "red", size = 4) +
  annotate("text", x = max(test$gfr) * 0.9, y = max(test$predictions_transformed) * 0.53,
           label = slope_1, color = "blue", size = 4)
cols <- c("#F76D5E", "#FFFFBF")
```

```
#, "#72D8FF"
# Basic density plot in ggplot2
cols <- c("Black" = "#F76D5E", "Non-Black" = "#FFFFBF", "All" = "#72D8FF")
ggplot() +
  geom_density(data = kidney_df, aes(x = sqrt_gfr, fill = "All"),
               alpha = 1, color = "#72D8FF") +  # Entire dataset
  geom_density(data = kidney_df %>% filter(black == 1),
               aes(x = sqrt_gfr, fill = "Black"),
               alpha = 0.5, color = "#F76D5E") +  # Black group
  geom_density(data = kidney_df %>% filter(black == 0),
               aes(x = sqrt_gfr, fill = "Non-Black"),
               alpha = 0.5, color = "#FFFFBF") +  # Non-Black group
  scale_fill_manual(values = cols) +
  labs(title = "Figure 7: Density Plot of Square Root GFR by Population Groups",
       x = "sqrt_gfr",
       y = "Density",
       fill = "Group") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 12),
        axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8))
# Creating third summary table (transformed)
kidney_tbl_nb <- kidney_df %>%
  dplyr::select(sex, age, sqrt_gfr, log_bascre, sbase, dbase, log_baseu) %>%
  rename(gfr=sqrt_gfr, bascre=log_bascre, baseu=log_baseu) %>%
  mutate(sex = recode(sex, `0` = "Female", `1` ="Male")) %>%
  tbl_summary(label = list(gfr ~ "Measured glomerular filtration rate(gfr)",
                           bascre ~ "Base serum Creatinine(bascre)",
                           sbase ~ "Systolic blood pressure(sbase)",
                           dbase ~ "Diastolic blood pressure(dbase)",
                           baseu ~ "Urine protein(baseu)"
                           ),
              statistic = all_continuous() ~ "{mean} ({sd})") %>%
  modify_footnote(update = all_stat_cols() ~ "Mean (SD) for continuous; n (%) for categori
  bold_labels()

tbl_combined2 <- tbl_merge(
  tbls = list(kidney_tbl_transformed, kidney_tbl_nb),
  tab_spanner = c("**Transformed Scale**", "**Transformed Scale (Race Removed)**")
)
```

```r
tbl_combined2 <- tbl_combined2 %>%
  as_kable_extra(booktabs = TRUE, caption = "Summary Table by Population Groups",
                 longtable = TRUE, linesep = "") %>%
  kableExtra::kable_styling(font_size = 7.5, position = "center",
                            latex_options = c("repeat_header", "HOLD_position", "scale_dow
  column_spec(1, width = "5cm") %>%
  column_spec(2, width = "2cm") %>%
  column_spec(3, width = "2cm") %>%
  row_spec(0, bold = TRUE)

tbl_combined2 <- tbl_combined2 %>%
  row_spec(5, background = "#e0ecf4") %>%
  row_spec(6, background = "#e0ecf4") %>%
  row_spec(9, background = "#e0ecf4")

tbl_combined2
set.seed(2550)
index_norace <- createDataPartition(kidney_df$gfr, p = 0.8, list = FALSE)
train_norace <- kidney_df[index_norace, ]
test_norace <- kidney_df[-index_norace, ]

model_norace <- glm(sqrt_gfr ~ dbase + age + sex + log_bascre + log_baseu + dbase:age +
                      dbase:sex + dbase:log_bascre + dbase:log_baseu + age:sex +
                      age:log_bascre + age:log_baseu + sex:log_bascre + sex:log_baseu +
                      log_bascre:log_baseu,
             data = train_norace)

step_model_norace <- step(model_norace, direction = "backward", trace = 0)
summary_model_norace <- as.data.frame(summary(step_model_norace)$coefficient)
colnames(summary_model_norace) <- c("Estimate", "Standard Error", "T Statistics", "P Value

summary_model_norace %>%
  kbl(booktabs = TRUE, caption = "Coefficient Estimate of the Best Model without Race",
      longtable = TRUE, linesep = "") %>%
  kable_styling(font_size = 10,
                latex_options = c("repeat_header", "HOLD_position", "scale_down"))
autoplot(step_model_norace, label.size = 1, size = 0.7) +
  ggtitle("Diagnostic Plot") +
  theme(
    plot.title = element_text(size = 10),
    axis.title = element_text(size = 8),
```

```r
    axis.text = element_text(size = 7),
    main = "Diagnostic Plot"
  )
# Predictions and performance metrics by race (grouped by black)
prediction <- predict(step_model, newdata = test)
test$estimated_gfr <- prediction
test$predictions_transformed <- prediction^2
performance_by_race <- test %>%
  group_by(black) %>%
  summarise(MSE = mean((predictions_transformed - gfr)^2),
            MAE = mean(abs(predictions_transformed - gfr)),
            Bias = mean(predictions_transformed - gfr),
            P10 = mean(abs(predictions_transformed - gfr) / gfr <= 0.10) * 100,
            P30 = mean(abs(predictions_transformed - gfr) / gfr <= 0.30) * 100) %>%
  mutate(Model = ifelse(black == 1, "Black", "Non-black")) %>%  # Rename 'black' as 'Model
  dplyr::select(-black) %>%
  dplyr::select(Model, everything()) # Move 'Model' to the first column

# Performance metrics for the model that includes the race variable
performance_by_race2 <- test %>%
  summarise(MSE = mean((predictions_transformed - gfr)^2),
            MAE = mean(abs(predictions_transformed - gfr)),
            Bias = mean(predictions_transformed - gfr),
            P10 = mean(abs(predictions_transformed - gfr) / gfr <= 0.10) * 100,
            P30 = mean(abs(predictions_transformed - gfr) / gfr <= 0.30) * 100) %>%
  mutate(Model = "With Black") %>%
  dplyr::select(Model, everything())

prediction_norace <- predict(step_model_norace, newdata = test_norace)
test_norace$estimated_gfr <- prediction_norace
test_norace$predictions_transformed <- prediction_norace^2

performance_by_race_norace <- test_norace %>%
  summarise(MSE = mean((predictions_transformed - gfr)^2),
            MAE = mean(abs(predictions_transformed - gfr)),
            Bias = mean(predictions_transformed - gfr),
            P10 = mean(abs(predictions_transformed - gfr) / gfr <= 0.10) * 100,
            P30 = mean(abs(predictions_transformed - gfr) / gfr <= 0.30) * 100) %>%
  mutate(Model = "Without Black") %>%
  dplyr::select(Model, everything())
```

```r
# Combine the performance tables
combined_performance <- rbind(performance_by_race2, performance_by_race, performance_by_ra

# Display the combined table
combined_performance %>%
  kbl(caption = "Evaluation Measures for Models With and Without Black") %>%
  kable_styling(bootstrap_options = c("striped", "hover", "condensed", "responsive"),
                latex_options = c("repeat_header", "HOLD_position", "scale_down"),
                font_size = 9) %>%
  row_spec(which(combined_performance$Model == "With Black"), bold = TRUE) %>%
  row_spec(which(combined_performance$Model == "Without Black"), bold = TRUE)
linear_model <- lm(predictions_transformed ~ gfr, data = test_norace)
slope <- round(coef(linear_model)[2], 2)

# Create a scatter plot to compare measured and estimated GFR
ggplot(test_norace, aes(x = gfr, y = predictions_transformed)) +
  geom_point(alpha=0.7, color="skyblue") +
  geom_abline(intercept = 0, slope = 1, linetype = "solid", size=1) +  # Perfect predictio
  geom_smooth(method = "lm", linetype = "dashed", se = FALSE, alpha=0.8, color="blue") +
  labs(title = "Figure 8: Comparison of Measured and Estimated GFR (Excluding Race in Mode
       x = "Measured GFR",
       y = "Estimated GFR") +
  theme_minimal() +
  theme(plot.title = element_text(hjust = 0.5, size = 10),
        axis.text.x = element_text(size = 8),
        axis.text.y = element_text(size = 8)) +
  annotate("text", x = max(test_norace$gfr) , y = max(test_norace$predictions_transformed)
           label = slope, color = "blue", size = 4)
```