Assessing the Limitations of the Public Medicare Part B Physician Payment Data Set in Terms of Linear Regression Analysis

Vivek Kantamani (vk2389)

Introduction

Medicare is the federal health insurance program for Social Security recipients ages 65 and older and for individuals under the age of 65 with a long-term disability. The Medicare program covers a variety of healthcare costs, including inpatient and outpatient hospital care, physician services, and prescription drug coverage, divided into various coverage components. In 2017, Medicare accounted for 15 percent of total federal spending and 20 percent of total national health spending.

The two primary coverage components of Medicare are Part A and Part B. Medicare Part A covers inpatient hospital stays, skilled nursing facility care, hospice care, and some home health visits. Medicare Part B covers medical care and services provided by physicians and other medical providers, medical equipment costs, outpatient care, and preventative care. While Part A of the Medicare program is financed through federal payroll tax contributions towards Social Security, Part B is financed by federal government general revenues and by monthly premiums paid by Medicare beneficiaries.

Cost-control measures are particularly important because only a portion of medical costs are covered by Medicare, with deductibles and co-insurance often required for most individuals. As a result, many Medicare beneficiaries maintain supplemental secondary insurance at their own expense. The primary contributor to rising monthly premiums as well as rising out-of-pocket costs for beneficiaries is the high unit prices of medical products and services - namely prescription drugs and physician services. Medicare establishes its payment rates for products and services according to contract negotiations with pharmaceutical companies and medical practitioners that take place on a regular basis.

Because Medicare as well as most private insurance companies employ a fee-for-service model, a concerning trend in healthcare over the last two decades has been the declining compensation of primary care physicians (i.e. internal medicine, family medicine, pediatric, and geriatric physicians) who provide fewer discrete services and more comprehensive, preventative care compared to specialists.

In this report, we seek to examine public Medicare Part B physician payment data to evaluate whether primary care physicians receive less compensation than physicians of other medical specialties. In particular, we will assess the limited utility of the public Medicare Part B physician payment data set (in terms of linear regression analysis) in answering this research question.

Data Collection and Description

The Centers for Medicare and Medicaid Services (CMS) release numerous public data sets of Medicare data annually to facilitate research. The "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" data set provides information on services and procedures provided to Medicare beneficiaries by physicians and other medical providers. Thus, this data set constitutes the primary source of data on payment and utilization associated with Medicare Part B services.

It is important to note that this data set is based on aggregated CMS administrative claims data for services offered to Medicare beneficiaries by healthcare providers. As a result, the data set lacks the specificity in claims data necessary to more accurately determine how claims for services contributed to Medicare payments for each physician.

CMS provides more detailed Medicare claims data in the "Basic Stand Alone (BSA) Medicare Claims Public Use Files" data sets. However, these data sets are composed of procedure-level claims data of random samples of beneficiaries (whose data have been de-identified to preserve confidentiality) and lack sufficient information regarding overall physician compensation beyond compensation for individual procedures. Although attempts were made, it was not possible to merge the data sets.

We now consider the 2017 "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" data set in more detail.

The data set was cleaned by

(1) Removing 15 variables associated with healthcare provider demographic information (e.g. name of the provider, national provider identifier number, address of the provider, etc.).

(2) Reducing the data set to observations corresponding only to various physician specialties using the Provider Type (x1) variable.

(3) Recoding the Provider Type (x1) variable to the levels "PCP" (primary care provider) and "Specialist" to facilitate our research question.

The resulting data set has a sample size of $n = 4,374,694$ observations and variables as follows.

Table 1: Variables

	Variable	Data Type	Description
x1	Provider Type	Factor	Provider (i.e. physician) type associated with the claim. Levels: PCP (0), Specialist (1)
x2	HCPCS Code	Factor	HCPCS code specifying the medical service rendered by the healthcare provider. Levels: HCPCS Codes
x3	Prescription Indicator	Factor	Identifies whether the provider dispensed prescriptions for the service provided. Levels: Y (0), N (1)
x4	Number of Services	Numeric	Number of physician services associated with the claim.
x5	Number of Medicare Beneficiaries	Numeric	Number of Medicare beneficiaries receiving the service from the provider.
x6	Number of Distinct Medicare Beneficiary Per-Day Services	Numeric	Number of distinct Medicare per-day services provided.
x7	Average Medicare Allowed Amount	Numeric	Average of the Medicare allowed amount payable from the service provided.
x8	Average Submitted Charge Amount	Numeric	Average of the charges submitted by the provider to Medicare for the service provided.
y1	Average Medicare Payment	Numeric	Average amount that Medicare paid for the service provided.
y2	Average Medicare Standardized Payment	Numeric	Average amount that Medicare paid for the service provided, standardized for number of services to facilitate comparisons.

Recall that our research question seeks to evaluate whether primary care physicians receive less compensation than physicians of other medical specialties based on Medicare Part B physician payment data. For the purposes of our analysis, two options exist for the response variable: Average Medicare Payment (y1) and Average Medicare Standardized Payment (y2). In this report, Average Medicare Standardized Payment (y2) will be utilized as the default response because the variable has been standardized to facilitate accurate comparisons between different years (should the need arise).

Exploratory Data Analysis and Variable Selection

x1, x2, x3 Considerations

The explanatory factor variables Provider Type (x1), HCPCS Code (x2), and Prescription Indicator (x3) have significant limitations that should be considered prior to inclusion in a linear regression model.

The variable Provider Type (x1) was originally a factor variable including 43 levels corresponding to a variety of physician specialties, which was then recoded to the levels "PCP" and "Specialist" to facilitate our research question.

(1) "PCP" (0) - including the levels "Internal Medicine", "Family Practice", "General Practice", "Pediatric Medicine", and "Geriatric Medicine".

(2) "Specialist" (1) - including the levels "Anesthesiology", "Cardiology", "Dermatology", "Gastroenterology", "General Surgery", etc.

The consolidation of 43 levels into 2 levels could artificially produce a statistically significant difference between the groups "PCP" and "Specialist".

The variable Prescription Indicator (x3) is a binary indicator variable indicating whether a physician dispensed prescriptions for the service provided. The binary nature of the variable may contribute to the limited utility of the variable in explaining variation in the response in a regression model.

The variable HCPCS Code (x2) is composed of 4511 levels corresponding to the HCPCS codes utilized by physicians to indicate the type of service rendered (e.g. routine physical examination, ultrasound, etc.). Although the type of service offered would likely explain a significant proportion of the variation in the response Average Medicare Standardized Payment (y2), we believe that the large number of levels relating to HCPCS codes will result in overparametrization of our linear regression model, which can lead to poor generalizability of regression coefficients.

Additionally, the interrelated nature of the HCPCS codes among themselves and with other explanatory variables results in collinearity. Calculating the generalized VIF values for Provider Type (x1), HCPCS Code (x2), and Average Submitted Charge Amount (x8) demonstrates collinearity among the variables in a regression model fit on a randomly sampled subset of our data set. In particular, $\max\{GVIF_k\} = 781.478 \gg 10$, and $\overline{GVIF_k} = 13.476 \gg 1$, as demonstrated by the summary output below. Note that generalized VIF allows for the calculation of VIF-like values for categorical variables through the addition of random noise and random reclassification of observations.

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.640   2.003   3.007  13.476   8.995 781.478
```

As a result of these considerations, we believe that HCPCS Code (x2) is not well suited for inclusion in a linear regression model. While forcing shrinkage on the parameters through the inclusion of an L2 penalty (as in ridge regression) could improve the utility of the variable, we believe that the overwhelming number of levels of the variable relative to the number of other available explanatory variables in the data set would still make this a poor choice. The inclusion of HCPCS Code (x2) is better suited for inclusion in a PCA model (which is outside the scope of this report), as linear combinations of the HCPCS codes could be easily interpreted, and because features in PCA are uncorrelated.

Due to the expected limitations of the explanatory factor variables, only the recoded variable Provider Type (x1) will be utilized in our regression model, as it is integral in addressing our research questions despite its faults.

x4, x5, x6 Considerations

The explanatory variables Number of Services (x4), Number of Medicare Beneficiaries (x5), and Number of Distinct Medicare Beneficiary Per-Day Services (x6) are interrelated and therefore demonstrate collinearity.

The variable Number of Distinct Medicare Beneficiary Per-Day Services (x6) incorporates the variables Number of Services (x4) and Number of Medicare Beneficiaries (x5) while controlling for how many services were rendered to beneficiaries over time. Because Distinct Medicare Beneficiary Per-Day Services (x6) condenses the information in the variables Number of Services (x4) and Number of Medicare Beneficiaries (x5), as indicated by the large correlation coefficients of $r = 0.980$ and $r = 0.872$ respectively, Distinct Medicare Beneficiary Per-Day Services (x6) is more viable as an explanatory variable in order to control for collinearity.

However, intuition and prior experience with claims data motivates our exclusion of the explanatory variable Number of Distinct Medicare Beneficiary Per-Day Services (x6) from our model - it is unlikely to be a useful explanatory variable for Average Medicare Standardized Payment (y2). This is because Average Medicare Standardized Payment (y2) indicates average Medicare payments per beneficiary, making the number of beneficiaries treated and the rate at which they were treated largely irrelevant.

```
##           x4        x5        x6
## x4 1.0000000 0.8547853 0.9797566
## x5 0.8547853 1.0000000 0.8718179
## x6 0.9797566 0.8718179 1.0000000
```

x7, x8 Considerations

The explanatory variables Average Medicare Allowed Amount (x7) and Average Submitted Charge Amount (x8) are interrelated and therefore demonstrate collinearity.

Average Submitted Charge Amount (x8) represents the charges billed to Medicare by physicians, which is an amount calculated by physician practices based on the services rendered to patients. Average Medicare Allowed Amount (x7) is an amount that Medicare specifies, representing the maximum amount that Medicare is willing to reimburse physicians for the services rendered to patients. The Average Medicare Allowed Amount (x7) is calculated based on the Average Submitted Charge Amount (x8) submitted by a physician, and is almost always smaller in value (because Medicare will almost never reimburse services at the rate requested by a physician).

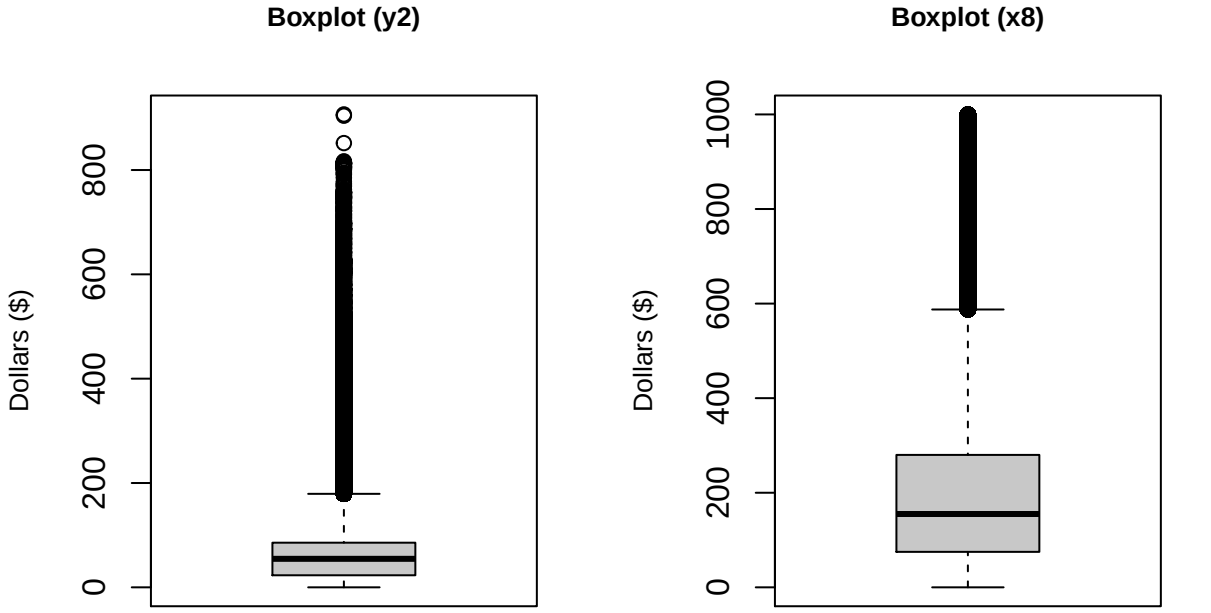
Average Submitted Charge Amount (x8) is the precursor to Average Medicare Allowed Amount (x7) which makes it more viable as an explanatory variable to be included in our regression model in order to control for collinearity.

```
##           x7        x8
## x7 1.0000000 0.7186103
## x8 0.7186103 1.0000000
```

y2 Considerations

The response Average Medicare Standardized Payment (y2) demonstrates right-skew.

In linear regression analysis, normality of the errors is assumed in order to facilitate inferences on the parameters. It is important to recognize that controlling for appropriate sources of variation in the response will typically result in normality of the errors (residuals). The variable Average Submitted Charge Amount (x8) demonstrates similar right-skew to the response Average Medicare Standardized Payment (y2), and should result in residuals that are approximately normal.



Average Medicare Standardized Payment (y2) Average Submitted Charge Amount (x8)

Methodology

For explanatory observational studies, establishing causation entails controlling for confounding variables. However, model selection among several candidate models could potentially impact the validity and interpretability of the p-values associated with our inferential testing procedure. In particular, model building involving the introduction and removal of covariates from a model results in p-values that are not entirely representative of the inferential procedure being performed (and affects the power of the test being performed). As a result, we instead establish a "Baseline Model" motivated primarily by the exploratory data analysis and intuition discussed in the previous section, in order to control for collinear variables which could cause instability in slope and standard error estimates.

$$Baseline\ Model: \quad Y_{i2} = \beta_0 + \beta_1 x_{i1} + \beta_8 x_{i8} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$x_1 = \begin{cases} 0 & if\ PCP \\ 1 & if\ Specialist \end{cases}$$

Recall that in this report we seek to

(1) Address our research question: Do primary care physicians receive less compensation than physicians of other medical specialties based on Medicare Part B physician payment data?

(2) Assess the limited utility of public Medicare Part B physician payment data (in terms of linear regression analysis) in addressing this research question.

We recognize the following general limitations of the 2017 "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" data set.

(1) Whereas a common problem in modern data analysis is high-dimensional data ($p \gg n$), our data set is low-dimensional ($n \gg p$) with an extremely large sample size of $n = 4,374,694$ observations relative to the number of available variables. Larger sample sizes typically result in smaller p-values if the null hypothesis of the inferential test is false, suggesting that p-values are a problematic measure of statistical significance, especially for large samples.

(2) The limited number of viable explanatory variables in our data set are unlikely to explain a sufficient proportion of the variation in the response to facilitate inferential procedures.

In order to examine these limitations in greater detail, we proceed with the following approach.

(1) In the "Appendix - Candidate Model Selection and Diagnostics" section of our report, we will utilize model diagnostics and remedial measures on our Baseline Model to determine possible candidate models for our inferential testing procedure. Recognizing the difficulty in interpreting statistical significance of testing procedures using p-values with large sample sizes, and due to R's memory allocation limitations, candidate models will be determined based on a randomly sampled subset of our data set.

$$Baseline\ Model: \quad Y_{i2} = \beta_0 + \beta_1 x_{i1} + \beta_8 x_{i8} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n$$

$$x_1 = \begin{cases} 0 & if\ PCP \\ 1 & if\ Specialist \end{cases}$$

(2) In the "Data Set Limitations - Candidate Model Evaluation and Inference" section of our report, we will examine the limited utility of the 2017 "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" data set in discerning an appropriate final model (from among the candidate models) that could be utilized in an inferential testing procedure to address our research question. Motivated by methods presented in statistical literature, we will utilize an approach akin to k-fold cross-validation, drawing multiple randomly sampled subsets from our data set in order to validate our analysis.

Data Set Limitations - Candidate Model Evaluation and Inference

In the "Appendix - Candidate Model Selection and Diagnostics" section of our report, we established two candidate models that could be utilized in an inferential testing procedure to address our research question: Do primary care physicians receive less compensation than physicians of other medical specialties based on Medicare Part B physician payment data? We will assess the utility of these models in testing this research question through an inferential testing procedure performed on the same randomly sampled subset of size $n^* = 20,000$ used in model diagnostics. In particular, we will evaluate the following equivalent null and alternative hypotheses:

$$H_0 : \beta_1 = 0 \quad vs \quad H_A : \beta_1 \neq 0$$

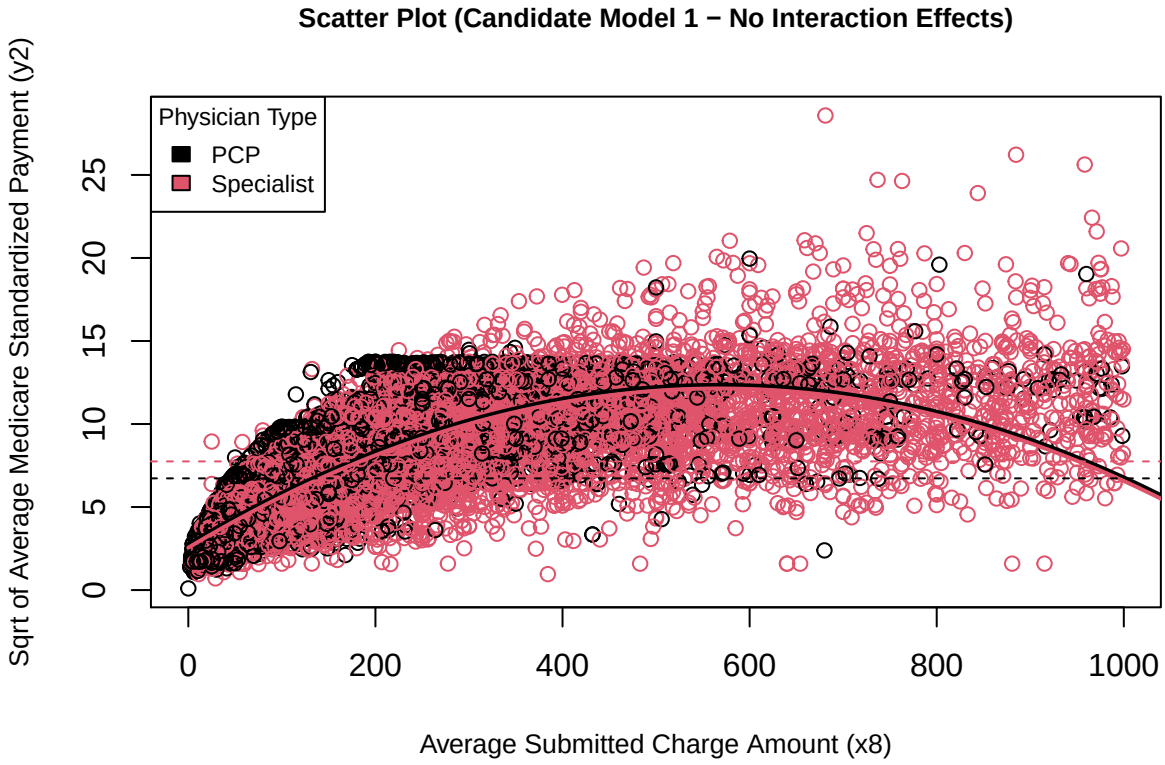
$$H_0 : \mu_{PCP} = \mu_{Spec} \quad vs \quad H_A : \mu_{PCP} \neq \mu_{Spec}$$

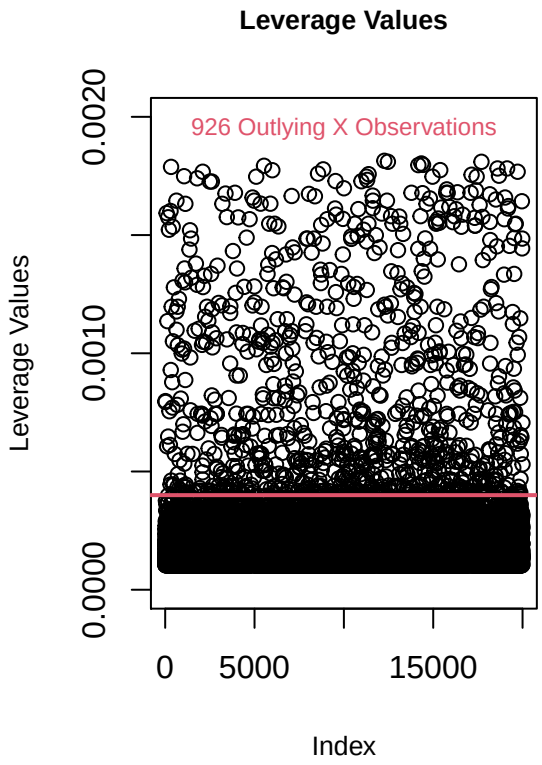
Analysis of Candidate Model 1

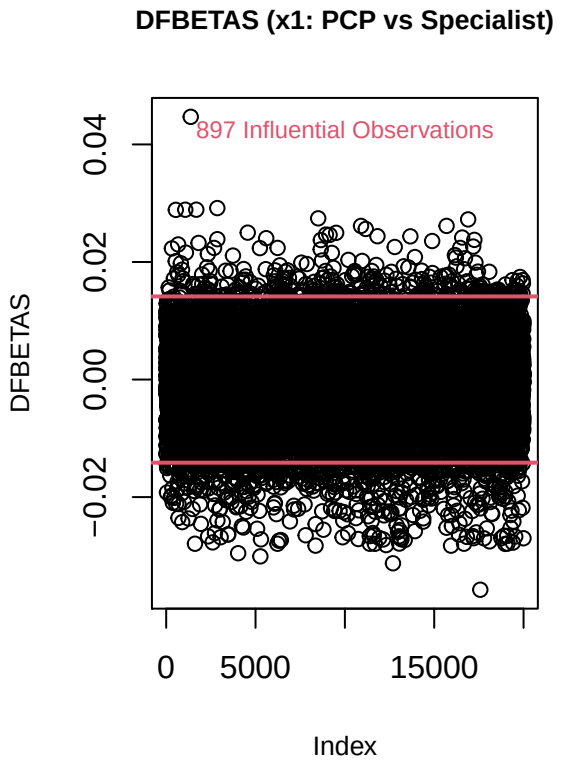
$$Candidate\ Model\ 1\ (No\ Interaction\ Effects): \quad Y'_{i2} = \sqrt{Y_{i2}} = \beta_0 + \beta_1 x_{i1} + \beta_8 x_{i8} + \beta_9 x_{i8}^2 + \varepsilon_i, \quad \varepsilon_i \overset{ind}{\sim} N(0, \sigma_i^2), \quad i = 1, \ldots, n$$

$$x_1 = \begin{cases} 0 & if\ PCP \\ 1 & if\ Specialist \end{cases}$$

```
##
## Call:
## lm(formula = sqrt(y2) ~ x1 + x8 + I(x8^2), data = df_analysis,
##     weights = wt)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9561  -1.6233  -0.0547   1.4942  13.3044
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.766e+00  2.429e-02 113.873   <2e-16 ***
## x1Specialist -2.144e-01  2.432e-02  -8.817   <2e-16 ***
## x8            3.379e-02  2.066e-04 163.515   <2e-16 ***
## I(x8^2)      -2.977e-05  3.205e-07 -92.883   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 19996 degrees of freedom
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.692
## F-statistic: 1.498e+04 on 3 and 19996 DF,  p-value: < 2.2e-16
```





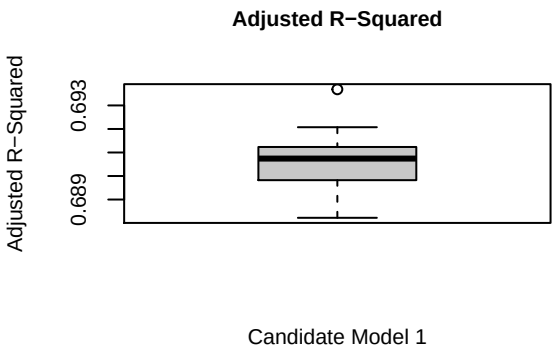


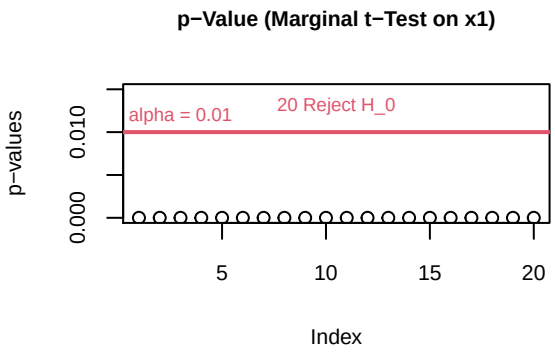
For Candidate Model 1, the marginal t-test on the β_1 slope parameter indicates a p-value of $< 2e-16$, suggesting that we reject the null hypothesis $H_0 : \mu_{PCP} = \mu_{Spec}$ at 5% significance. This result implies that there is a statistically significant difference in the Average Medicare Standardized Payment (y2) made to physicians depending on their designation as primary care physicians or specialists. However, upon visual inspection, the difference in the response functions for the two groups seems less pronounced than the difference in the unconditional averages in the response for the two groups - this suggests that there is not a significant difference in the compensation of primary care physicians and specialists. It is plausible that this result stems from the smaller p-values typically associated with larger samples when the null hypothesis is false - although a subset of the original data set was utilized, the large sample size and simplicity of the model resulted in all terms demonstrating statistical significance. Moreover, we note that the regression model suggests that specialists receive less compensation than primary care physicians although this appears to contradict trends apparent in the data.

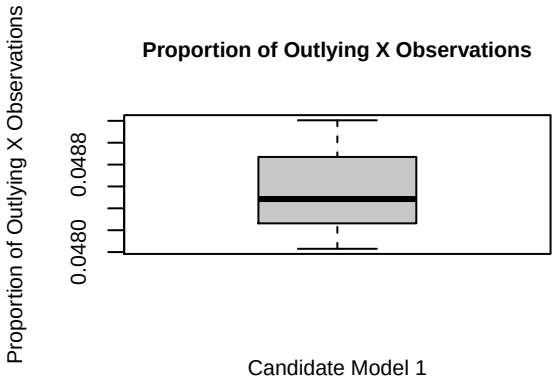
To analyze this further, we note that an analysis using leverage values suggests that there are 926 outlying X observations, representing approximately 4.63% of our sample. An analysis using the DFBETAS measure suggests that there are 897 observations influencing the $\hat{\beta}_1$ regression coefficient used in our inference procedure, representing approximately 4.48% of our sample. These outlying X observations and influential observations likely had a noteworthy impact on our inferential testing procedure that could skew results. In particular, from a model validation perspective, perhaps the influential observations affected our regression coefficient estimates.

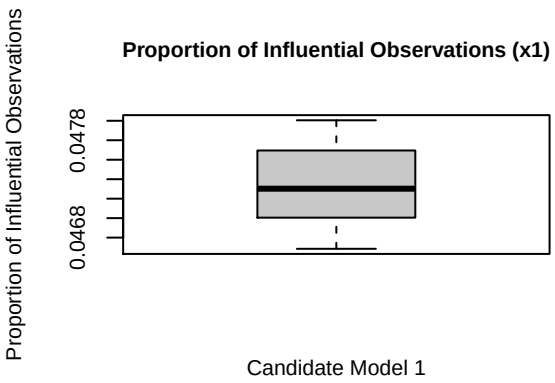
We note that the weights in weighted least squares are inversely proportional to the error variance of a given observation. It is plausible that the weighted least squares procedure was not appropriate for this data set - while this measure resulted in a moderate adjusted R-squared value of 0.6920, the "outliers" indicated by our model fit using this procedure are likely valid data points whose structure could not be accounted for by our model. As noted in the "Exploratory Data Analysis and Variable Selection" section of our report, the limited number of viable explanatory variables in the 2017 "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" data set limits our ability to construct a more appropriate linear regression model.

Motivated by k-fold cross-validation procedures, we split our data set of size $n = 4,374,694$ into $k = 20$ folds, and validated our analysis of Candidate Model 1 using these folds. This replication (summarized below) demonstrates similar adjusted R-squared values, similar proportions of outlying X observations, and similar proportions of influential observations to our analysis. We note that all 20 hypothesis tests resulted in rejection of the null hypothesis.







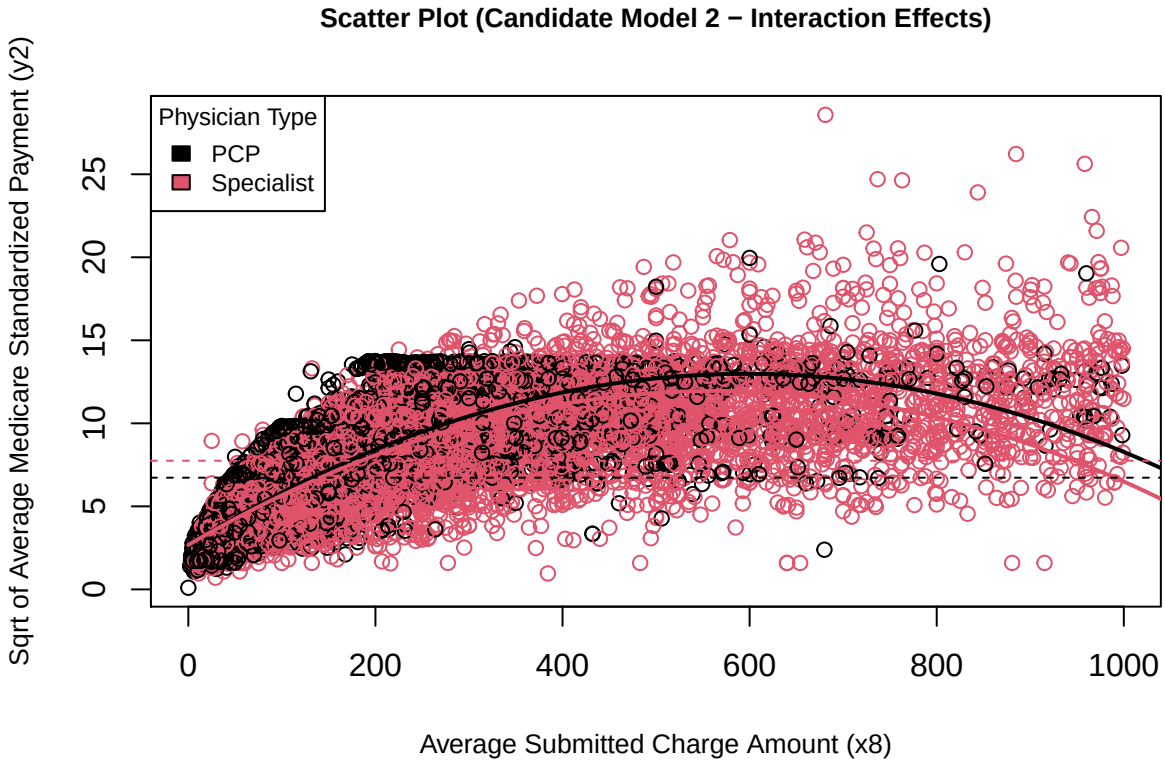


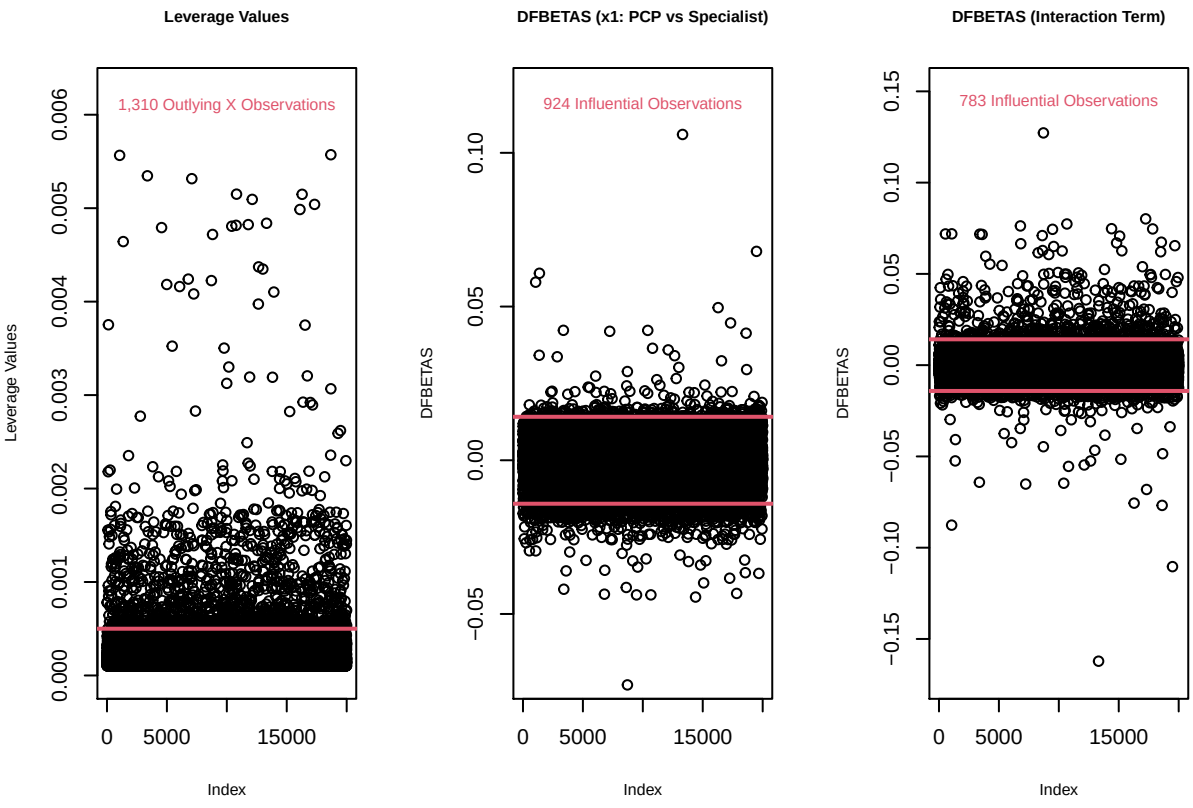
Analysis of Candidate Model 2

$$Candidate\ Model\ 2\ (Interaction\ Effects):\ \ Y'_{i2} = \sqrt{Y_{i2}} = \beta_0 + \beta_1 x_{i1} + \beta_8 x_{i8} + \beta_9 x_{i8}^2 + \beta_{10} x_{i1} x_{i8} + \varepsilon_i,\ \ \varepsilon_i \overset{ind}{\sim} N(0, \sigma_i^2),\ \ i = 1, \ldots, n$$

$$x_1 = \begin{cases} 0 & if\ PCP \\ 1 & if\ Specialist \end{cases}$$

```
##
## Call:
## lm(formula = sqrt(y2) ~ x1 + x8 + I(x8^2) + x1 * x8, data = df_analysis,
##     weights = wt)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3991  -1.6243  -0.0672   1.4985  12.9740
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.662e+00  2.798e-02  95.142   <2e-16 ***
## x1Specialist     1.706e-02  3.496e-02   0.488    0.625
## x8               3.458e-02  2.222e-04 155.589   <2e-16 ***
## I(x8^2)         -2.896e-05  3.243e-07 -89.324   <2e-16 ***
## x1Specialist:x8 -1.800e-03  1.915e-04  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 19995 degrees of freedom
## Multiple R-squared:  0.6948, Adjusted R-squared:  0.6948
## F-statistic: 1.138e+04 on 4 and 19995 DF,  p-value: < 2.2e-16
```





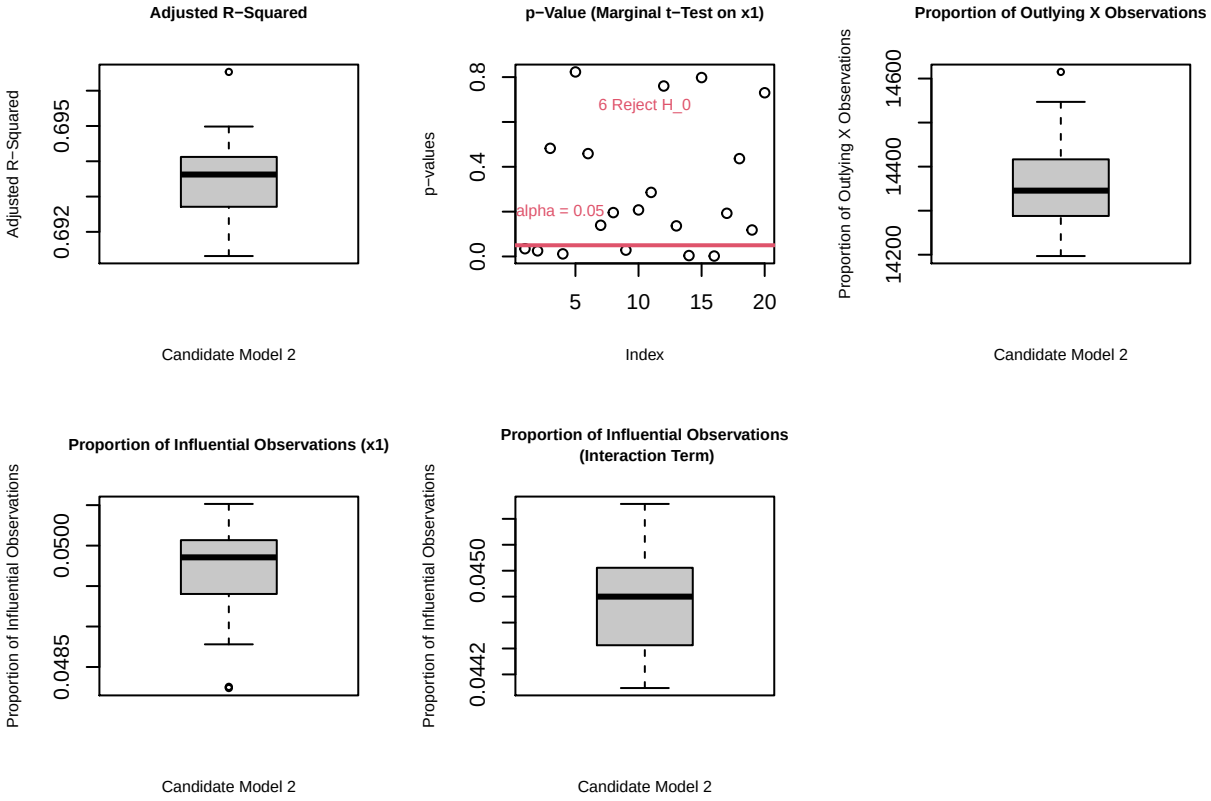
For Candidate Model 2, the marginal t-test on the β_1 slope parameter indicates a p-value of 0.625, suggesting that we fail to reject the null hypothesis $H_0 : \mu_{PCP} = \mu_{Spec}$ at 5% significance. This result implies that there is not a statistically significant difference in the Average Medicare Standardized Payment (y2) made to physicians depending on their designation as primary care physicians or specialists. However, we note that this regression model also suggests that specialists receive less compensation than primary care physicians although this is appears to contradict trends apparent in the data. In fact,

the interaction term in Candidate Model 2 exaggerates this effect even more so than Candidate Model 1, due to a more pronounced quadratic trend.

Analyses using leverage values and DFBETAS measures suggests that there are 1,310 outlying X observations (6.55% of our sample), 924 observations (4.62% of our sample) influencing the $\hat{\beta}_1$ regression coefficient, and 783 observations (3.92% of our sample) influencing the interaction term regression coefficient, $\hat{\beta}_{10}$. Similarly to Candidate Model 1, these outlying and influential observations likely impacted our inferential testing procedure, once again calling into question the use of weighted least squares.

Furthermore, it is plausible that the 783 observations influencing the interaction term regression coefficient, $\hat{\beta}_{10}$, led to its statistical significance in the model - after all, Candidate Model 2's adjusted R-squared value of 0.6948 is virtually identical to that of Candidate Model 1, despite the inclusion of the supposedly significant interaction term. If the interaction term was truly necessary to describe trends in the data, we would expect its impact on adjusted R-squared to be more pronounced. Furthermore, the inclusion of this questionable interaction term led to failing to reject the null hypothesis of our inferential testing procedure, contradicting the conclusion of Candidate Model 1. It is difficult to determine which of the candidate models is more appropriate and which conclusion is more robust due to the large number of outlying and influential observations in both models affecting the regression coefficients of both our variable of interest Provider Type (x1) and the interaction term.

Our 20-fold replication analysis resulted in similar adjusted R-squared values, similar proportions of outlying X observations and similar proportions of influential observations to our original analysis. We note that 6 of the 20 hypothesis tests resulted in rejections of the null hypothesis, which contradicts our results above. The inconsistency in the results of our hypothesis tests despite the large sample size of the data set and each fold further calls into question both our model and the inclusion of the interaction term.



Discussion

In the process of examining the diagnostics of our candidate models, we noted that the candidate models demonstrated a sufficiently normal and independent error (residual) structure, and had reasonable residual plots following the application of weighted least squares, a transformation in the response, and the introduction of a non-linear response surface. Although some outliers were apparent, it was assumed that these outliers were representative of a systematic error variance function and were suitably controlled for by the use of weighted least squares. Finally, both models explained a sufficient proportion of variation in the response on the basis of the adjusted R-squared measure, so proceeding with an inferential testing procedure seemed reasonable.

However, both candidate models suggested that specialists received less compensation (in terms of average Medicare Part

B payments) than primary care physicians, which contradicts both intuition and exploratory data analysis of trends in the data. It would appear that although weighted least squares improved the adjusted R-squared measure in terms of model diagnostics, it was an inappropriate remedial measure for this analysis as it "punished" observations with large error (residual) variances with low weights when these points should have been incorporated into the model using a different remedial structure. Furthermore, the large number of outlying X observations and influential observations on regression coefficients of interest calls into question the stability of our regression coefficient estimates and conclusions of our inferential testing procedure. This instability is evident in the fact that the inclusion of a statistically significant interaction term (whose regression coefficient had an inordinate number of influential observations) resulted in statistical conclusions that contradicted the conclusions of our influential testing procedure without the interaction term. It is unclear which conclusion is legitimate due to the excessive number of outliers and influential observations in both models.

These issues could be resolved with the selection of a more appropriate model. However, the 2017 "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" data set is low-dimensional ($n \gg p$) with a limited number of explanatory variables that are highly collinear. We note that collinear variables cause instability in slope and standard error estimates, and that large sample sizes typically result in smaller p-values if the null hypothesis of the inferential test is false. This suggests that it is difficult to confirm the robustness of our statistical conclusions given the small number of viable parameters in our models and the large sample sizes of even randomly sampled subsets of our data set. These limitations of the data set are indicative of its limited utility in an inferential procedure predicated on linear regression analysis. The data set simply requires more explanatory variables addressing more sources of variation in Average Medicare Standardized Payments (y2) for linear regression analysis to be useful as part of an inferential testing procedure.

We believe that introducing shrinkage or sparsity (as in ridge regression or lasso), or using a more robust regression method are not particularly viable in this instance due to the limited number of explanatory variables available. As noted previously, PCA is a potentially useful unsupervised learning technique for this data set (although it is typically used with high-dimensional data) because the linear combinations of features used in PCA are uncorrelated (possibly resolving the high collinearity of the available explanatory variables). Simulation analysis using the formulas that Medicare uses to reimburse physicians could also be informative, although this information is not publicly available.

In conclusion, we recognize that developing an intuition and understanding of the data being analyzed (i.e. through exploratory data analysis and research) is more important than blindly relying on model diagnostics and remedial measures in the process of model building and evaluation. While model diagnostics are informative measures that can aid in resolving noteworthy violations of particular model assumptions, relying solely on these techniques without considering the limitations of the data set could result in models that do not accurately reflect the underlying data leading to flawed inferential conclusions and inaccurate predictions.

Appendix - Candidate Model Selection and Diagnostics

The Baseline Model established through exploratory data analysis and intuition is stated as follows:

$$Baseline\ Model: \quad Y_{i2} = \beta_0 + \beta_1 x_{i1} + \beta_8 x_{i8} + \varepsilon_i, \quad \varepsilon_i \overset{iid}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n$$

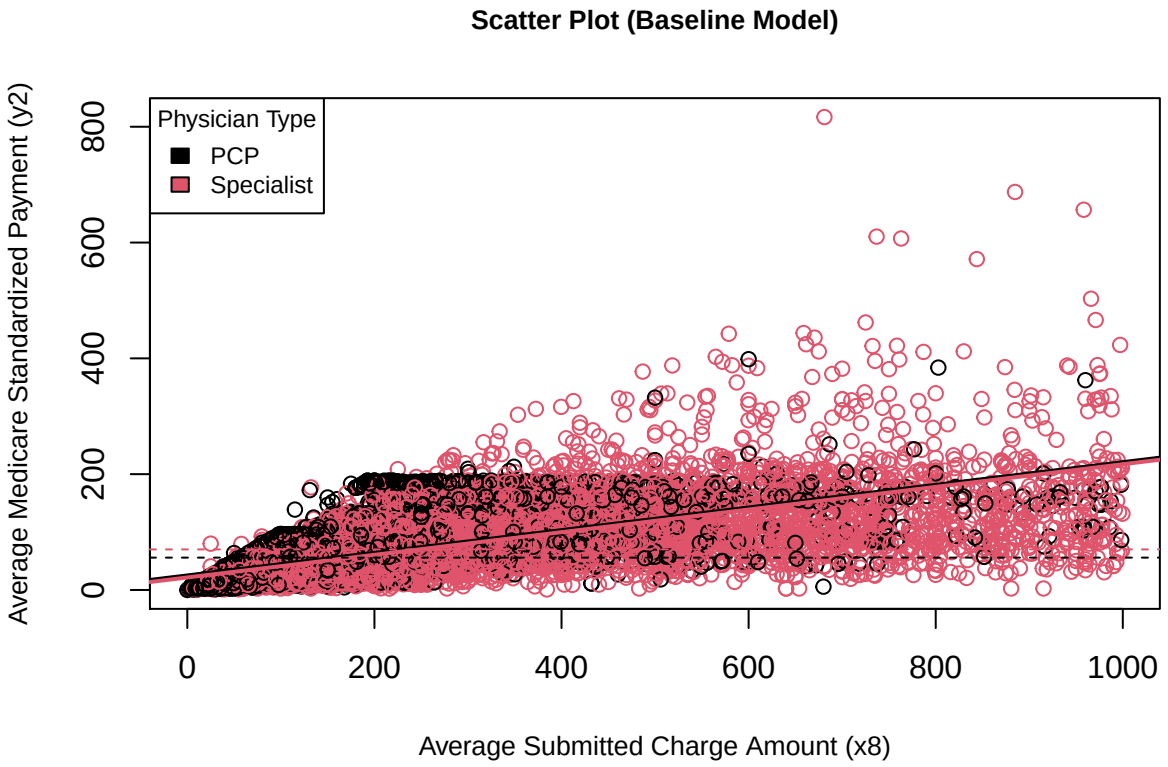
$$x_1 = \begin{cases} 0 & if\ PCP \\ 1 & if\ Specialist \end{cases}$$

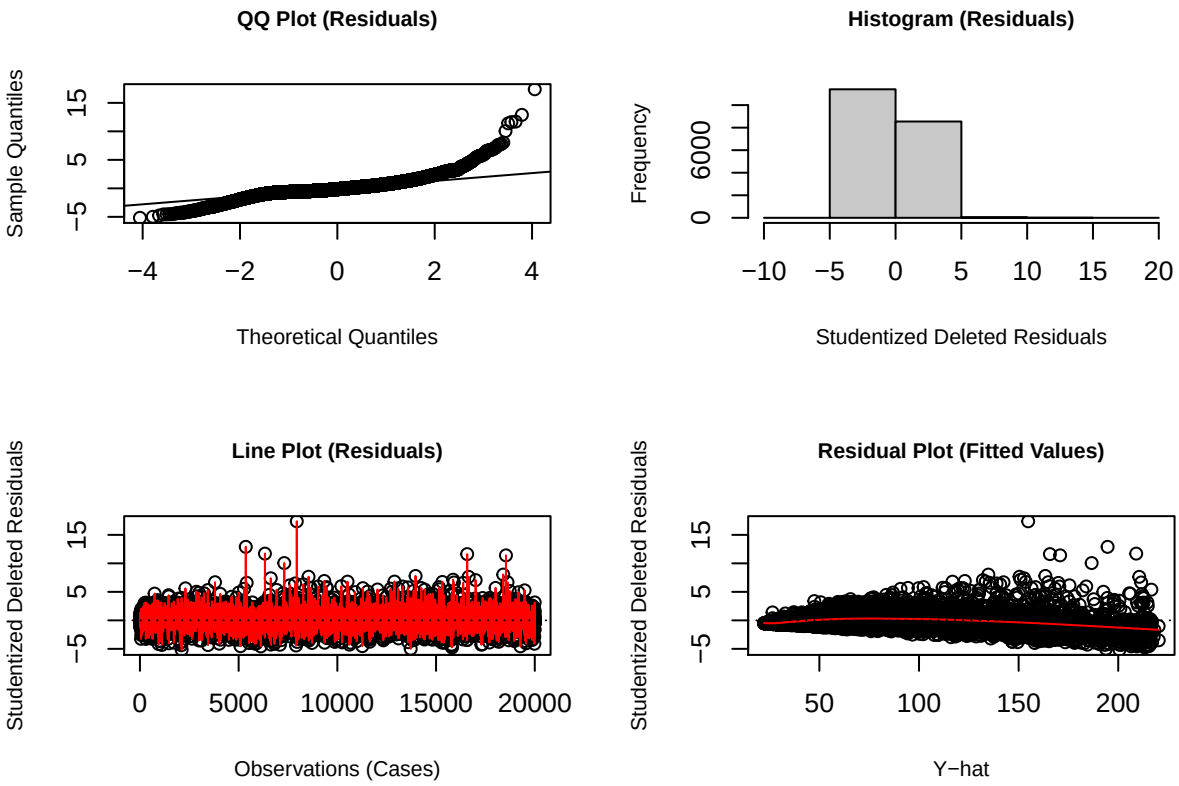
We will utilize diagnostic and remedial measures (interactions, functional forms, transformations, etc.) as necessary to determine possible candidate models for our inferential testing procedure. Recognizing the difficulty in interpreting statistical significance of testing procedures using p-values with large sample sizes, and due to R's memory allocation limitations, we randomly sample a subset of size $n^* = 20,000$ from our data set of size $n = 4,374,694$.

We proceed with diagnostic measures on our Baseline Model in order to determine the regression assumptions that have been violated, and the remedial measures (interactions, functional forms, transformations, etc.) needed to resolve these violations.

```
##
## Call:
## lm(formula = y2 ~ x1 + x8, data = df_analysis)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -198.07  -20.09   -5.55   15.36  661.93
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  25.068073   0.492260   50.92  < 2e-16 ***
```

```
## x1Specialist -3.475348   0.572543   -6.07  1.3e-09 ***
## x8            0.195600   0.001433  136.50  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.38 on 19997 degrees of freedom
## Multiple R-squared:  0.4908, Adjusted R-squared:  0.4908
## F-statistic:  9638 on 2 and 19997 DF,  p-value: < 2.2e-16
```






```
## [1] "Box-Cox Procedure Lambda Estimate = 0.4242"
```

The Baseline Model has an adjusted R-squared value of 0.4908, which indicates that the model does not explain a large proportion of variation in the response.

The scatter plot and residual plot versus fitted values suggest a quadratic trend to the data as well as the presence of outliers. The plots also indicate heteroscedasticity of the residuals (errors). In particular, the error variance changes in a systematic fashion (in the traditional 'megaphone' shape) suggesting that weighted least squares is an appropriate remedial measure.

The QQ plot and the histogram of deleted residuals clearly suggest that the distribution of the residuals has heavy tails, and violates normality. It is very plausible that this structure is due to the heteroscedasticity of the residuals. Additionally, the Box-Cox procedure produced a lambda estimate of $\lambda = 0.4242 \approx 0.5$, suggesting that the transformation $Y'_{i2} = \sqrt{Y_{i2}}$ is appropriate. We note that transforming the response could remedy the non-normality of the errors (residuals).

The line plot of residuals is suitably noisy, although several outliers are clearly evident. Therefore, the absolute residuals (as opposed to the squared residuals) will be utilized when performing iteratively reweighted least squares to determine an appropriate error variance function to address the heteroscedasticity of the errors because absolute residuals are more robust to outliers than squared residuals.

This discussion motivates the following remedial measures.

(1) A square root transformation, $Y'_{i2} = \sqrt{Y_{i2}}$, of the response Average Medicare Standardized Payment (y2). Although transformations of the response can affect the interpretability of the model, this transformation should not be particularly impactful in this regard.

(2) The introduction of a squared Average Submitted Charge Amount (x8) term into the model to account for the quadratic trend in our data. Note that the Provider Type (x1) variable is a factor, and it does not make sense to introduce a squared factor variable into our model to introduce a quadratic structure.

(3) The use of iteratively reweighted least squares to determine weights of the form $w_i = \frac{1}{\hat{s}_i^2}$ using a standard deviation function $\hat{s}_i$ of σ_i^2 computed from the regression $|e_i| \sim X_k$.

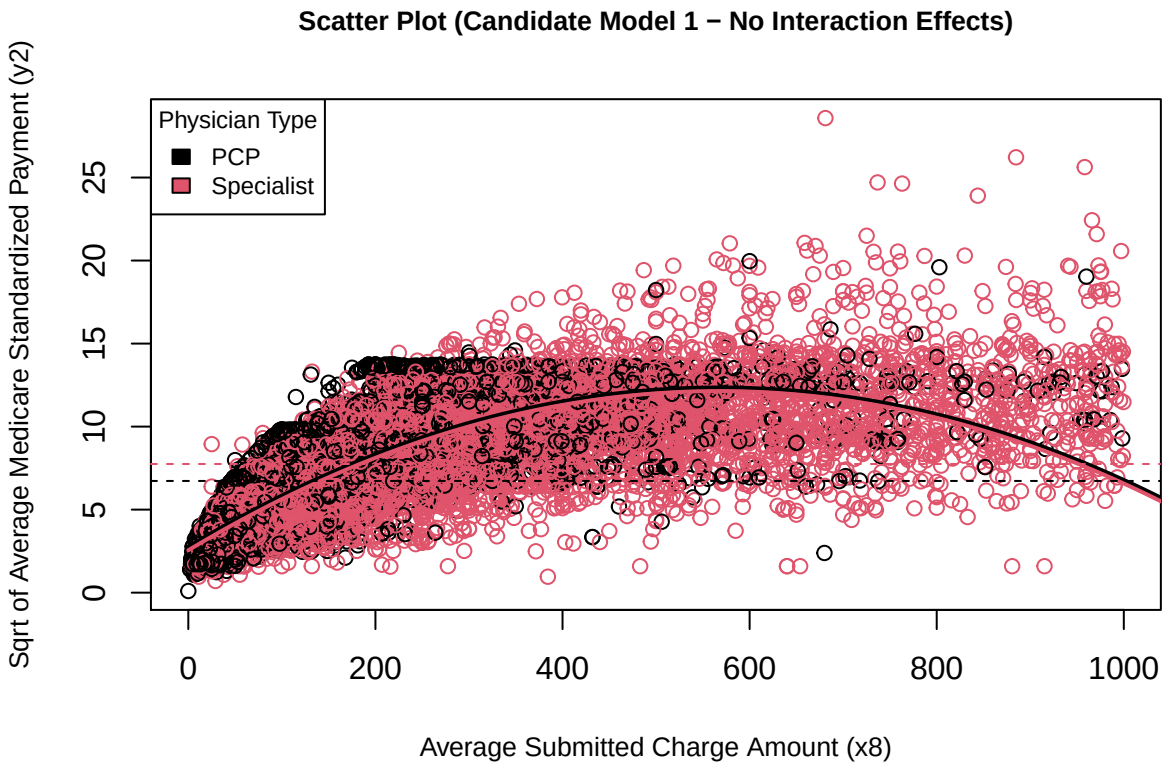
This results in two possible candidate models, differing only in the inclusion of interaction effects.

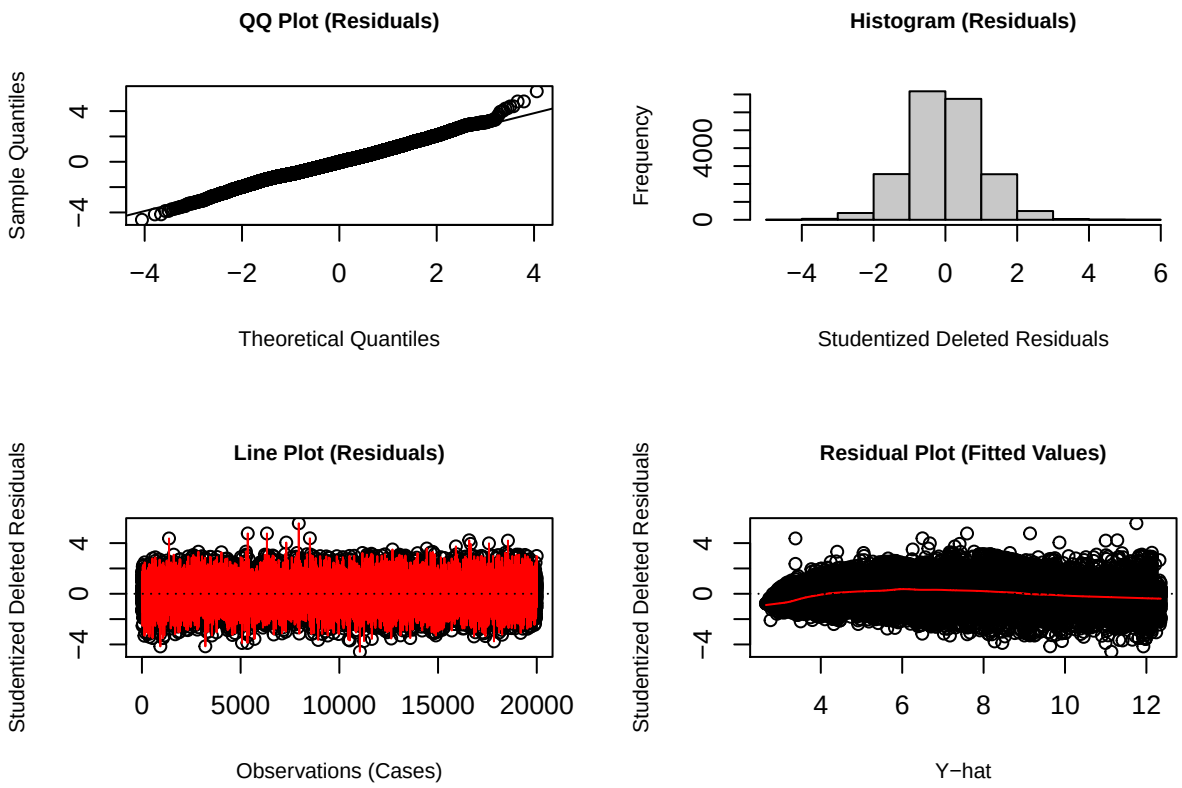
(1) Candidate Model 1 - No Interaction Effects

Candidate Model 1 (No Interaction Effects) : $Y'_{i2} = \sqrt{Y_{i2}} = \beta_0 + \beta_1 x_{i1} + \beta_8 x_{i8} + \beta_9 x_{i8}^2 + \varepsilon_i, \;\; \varepsilon_i \overset{ind}{\sim} N(0, \sigma_i^2), \;\; i = 1, \ldots, n$

$$x_1 = \begin{cases} 0 & if\ PCP \\ 1 & if\ Specialist \end{cases}$$

```
##
## Call:
## lm(formula = sqrt(y2) ~ x1 + x8 + I(x8^2), data = df_analysis,
##     weights = wt)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -10.9561  -1.6233  -0.0547   1.4942  13.3044
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.766e+00  2.429e-02 113.873   <2e-16 ***
## x1Specialist -2.144e-01  2.432e-02  -8.817   <2e-16 ***
## x8            3.379e-02  2.066e-04 163.515   <2e-16 ***
## I(x8^2)      -2.977e-05  3.205e-07 -92.883   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 19996 degrees of freedom
## Multiple R-squared:  0.692,  Adjusted R-squared:  0.692
## F-statistic: 1.498e+04 on 3 and 19996 DF,  p-value: < 2.2e-16
```



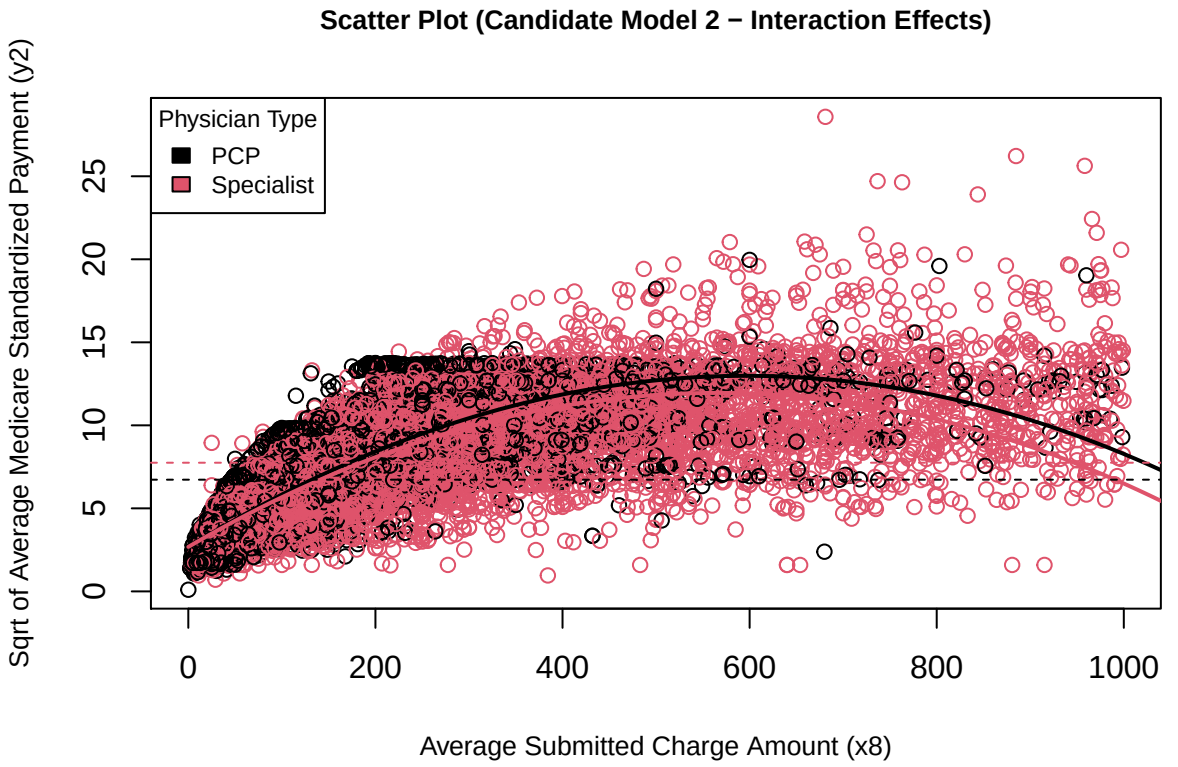


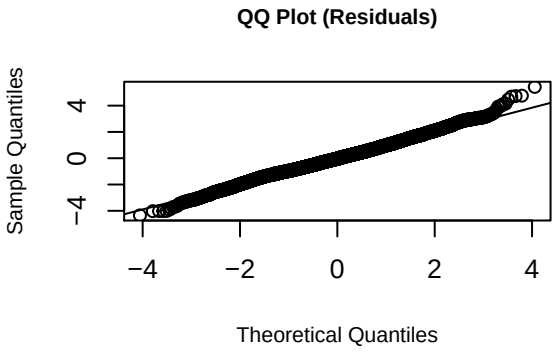
(2) Candidate Model 2 - Interaction Effects

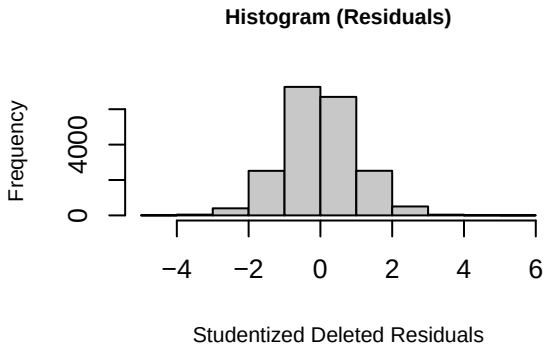
$$Candidate\ Model\ 2\ (Interaction\ Effects):\ \ Y'_{i2} = \sqrt{Y_{i2}} = \beta_0 + \beta_1 x_{i1} + \beta_8 x_{i8} + \beta_9 x_{i8}^2 + \beta_{10} x_{i1} x_{i8} + \varepsilon_i,\ \ \varepsilon_i \overset{ind}{\sim} N(0, \sigma_i^2),\ \ i = 1, \ldots, n$$

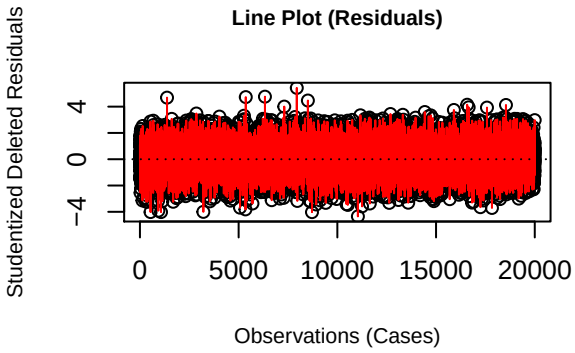
$$x_1 = \begin{cases} 0 & if\ PCP \\ 1 & if\ Specialist \end{cases}$$

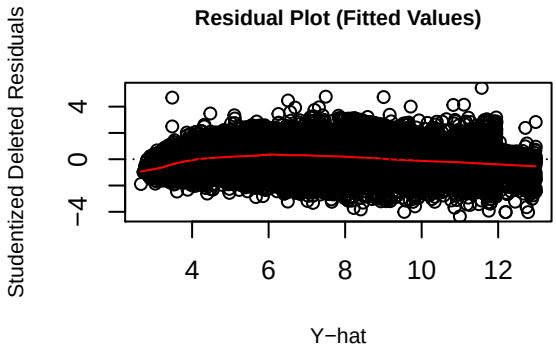
```
##
## Call:
## lm(formula = sqrt(y2) ~ x1 + x8 + I(x8^2) + x1 * x8, data = df_analysis,
##     weights = wt)
##
## Weighted Residuals:
##      Min       1Q   Median       3Q      Max
## -10.3991  -1.6243  -0.0672   1.4985  12.9740
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      2.662e+00  2.798e-02  95.142   <2e-16 ***
## x1Specialist     1.706e-02  3.496e-02   0.488    0.625
## x8               3.458e-02  2.222e-04 155.589   <2e-16 ***
## I(x8^2)         -2.896e-05  3.243e-07 -89.324   <2e-16 ***
## x1Specialist:x8 -1.800e-03  1.915e-04  -9.396   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.393 on 19995 degrees of freedom
## Multiple R-squared:  0.6948, Adjusted R-squared:  0.6948
## F-statistic: 1.138e+04 on 4 and 19995 DF,  p-value: < 2.2e-16
```











Candidate Model 1 (including no interaction effects) and Candidate Model 2 (including interaction effects) have similar adjusted R-squared values of 0.6920 and 0.6948 respectively, which indicate that the models explain only a moderate proportion of variation in the response. The scatter plots of both models appear reasonable, although both plots indicate a sizable number of outliers.

The shape of the QQ plots and histograms of both models suggest that the distribution of the residuals is approximately normal. The line plots of residuals for both models are suitably noisy, although several outliers are evident in both models. Finally, the residual plots versus the fitted values of both models also indicate the presence of several outliers, despite the use of weighted least squares.

Both Candidate Model 1 and Candidate Model 2 represent our attempt at constructing a model that could be utilized in addressing our research question using model diagnostics and remedial measures. However, in the "Data Set Limitations - Candidate Model Evaluation and Inference" section of our report, we will examine the limitations of the 2017 "Medicare Provider Utilization and Payment Data: Physician and Other Supplier" data set (in terms of both its low-dimensional nature and its limited number of useful explanatory variables) in discerning a suitable final model from among the candidate models to facilitate inferential testing procedures.