Learning from Insurance Data: Robust Regression Through the Origin

Victor Wright

February 2020

Abstract

In linear regression modeling, it is the analyst's responsiblility to make various assumptions about the the relationship between a dependent variable Y, a set of independent variables X, and the unobservable statistical errors ϵ_i. In particular, one assumes that the true relationship Y between and X is linear, the residuals have constant variance, the errors are independent, and they should follow a distribution that is symmetric around a mean of $\mu = 0$ and have variance σ^2. I.e., follow a normal distribution. As discussed in *Machine Learning: Learning from Data Using Linear Regression*, variable transformations can be used to satisfy the linearity condition as well as the constant variance and normality assumptions in some cases. This is necessary because we assume that the true regression model errors have such properties as well. However, interpretations of such models can become increasingly difficult when the regression model has a large number of variables as well as interaction terms. This is because the interpretation of the impact of the transformed independent variables have on the dependent variable have to be interpreted in terms of the transformations. For example, when logarithms are applied to the data in the variables, independent or dependent, the interpretation of how the independent variable(s) impact the dependent variable are given as a percent change.

In this paper, we analyze a complex data set that contains information that describes medical expenses for over thirteen hundred beneficiaries on a insurance plan and build linear a regression model that attempts to explain the variability in the project target variable *Expenses*. Additionally, we do not fit a regression model with an intercept. Instead, we regress *Expenses* on a set of independent variables through the origin. Most of which we engineer based on facts we previously discovered while conducting research for our previous project titled *Machine Learning: Using Logistic Regression to Predict Coronary Heart Disease*. Secondly, the enineered features are all binary valued categorical variables that represent various groups of beneficiary body mass indices for smoker and nonsmoker benficiaries. Therefore, these indicator variables allow us to model the impact of various *BMI* levels on *Expenses* for smoker and nonsmoker benficiaries. Thirdly, we also discover that there are *multiple linear* relationships

between *Expenses* and a continuous variable *Age* when the pairs (*Age*, *Expenses*) are plotted and grouped by the categories we engineered. Finally, *Regression through the origin* (RTO) is thought to be controversial and we discuss why it is thought to be so by touching on statements in contained Eisenhauer's (2003) discussion. Then, give a brief argument as to why we believe RTO is appropriate for our problem.

The estimated RTO model appears to perform very well on all of the training data at a first glance. For example, we are able to obtain $\boldsymbol{R^2} = 93.75\%$. However, the residual plot of the estimated residuals vs the model's fitted values shows that the likelihood of nonconstant variance in the residuals is extremely high and we test for the condition using a technique constructed by Weisberg and Cook (2014). Since the majority of the independent variables are binary valued, we could not apply natural logarithm transforms to them and did not bother applying root transforms since $\sqrt{0} = 0$ and $\sqrt{1} = 1$ in hopes to achieve a constant variance. Additionally, a combination of natural logarithm transforms to *Expenses* and *Age* returned residual plots that exhihibted very bad behavior. Additionally, we identified a large amount of outliers. Near the end of the paper, we briefly describe weighted least squares and iteratively reweighted least squares (IRLS). We make use of the IRLS fit and make a prediction with the model.

An Explanation of Variables, Feature Engineering, Data Visualization with R, and Regression Through the Origin

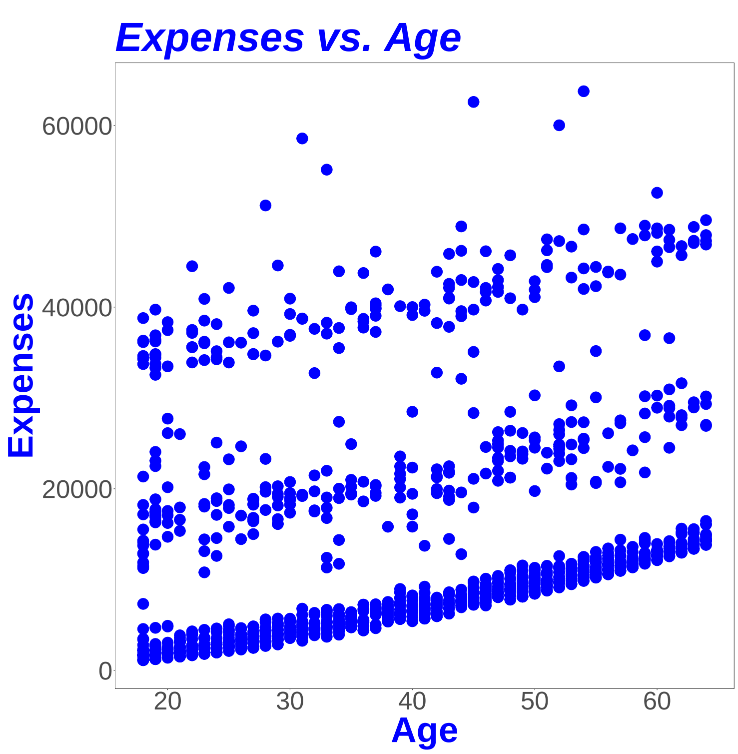
An Explanation of Variables

In this section, we provide a brief discussion of the information we have obtained. We also provide scatter plots as well as grouped scatter plots of the data and identify trends in the information and estimate a regression through the origin via OLS. The variables in the data set are, *Age*, *Sex*, *BMI*, *Children*, *Smoker*, *Region*, and *Expenses*. The data set we have obtained, *Expenses.csv*, can be found at https://www.kaggle.com/mirichoi0218/insurance or https://github.com/stedy/Machine-Learning-with-R-datasets and are part of *Machine Learning with R: Learn how to use R to apply powerful machine learning methods and gain insight into real-world applications* by Brett Lantz. The data were collected from the US Census Bureau and based on demographic statistics computed by the organization [1]. Since the data were collected by the US Census Bureau, the data contain some level of realism regarding patient medical costs in the United States [1]. A description of the variables can be found below.

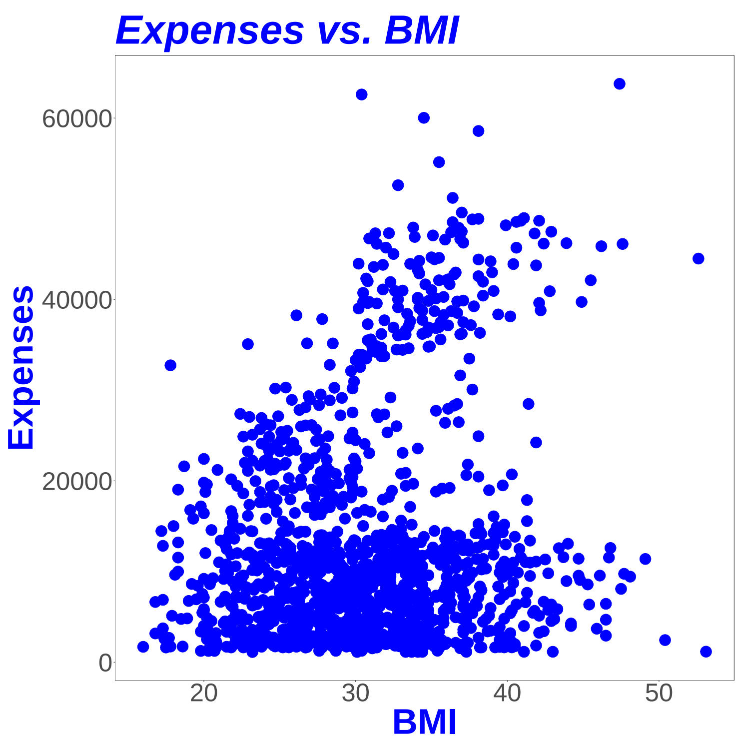
We name our data set *Expenses* which contains insurance plan beneficiary information, some patient qualities, total medical expenses charged to the plan over the course of a yearand, and 1338 training examples [1]. In the *Expenses* data set, primary beneficiary ages are contained in *Age* which are recorded as integers ranging from 18 to 64, *Sex* which is a qualitative variable indicating *male* or *female*, *BMI* which is the body mass index of the person measured in $\frac{kg}{m^2}$, the number of children that the beneficiary has which are also covered by the plan (children/dependents) is *Children*, *Smoker* indicating whether or not the beneficiary smokes, *Region* which divides the country into the regions northeast, northwest, southwest, and southeast which classifies where the beneficiary's place of residence lies in the United States, and *Expenses* which are charged to the plan per year [1]. Even though age is reported as an integer, the concept of age is clearly continuous. *E.g.*, age can be finely divided into months, weeks, days, hours, etc. The other continuous variables in the data set are clearly *BMI* and *Expenses*. A scatter plot of *Expenses vs. Age* can be seen below along with the code used to contruct the image.

Data Visualization

```
> ggplot(Expenses, aes(x = Expenses$age, y= Expenses$expenses)) +
+   geom_point(col = "blue", xlab = "age", size = 10) + theme_bw() +
+   theme(plot.title = element_text(colour = "blue", face = "bold.italic",
+                                   size = 80),
+         axis.title.x = element_text(colour = "blue", face = "bold",
+                                     size = 70),
+         axis.title.y = element_text(colour = "blue", face = "bold",
+                                     size = 70),
+         axis.text.x = element_text(size = 50),
+         axis.text.y = element_text(size = 50),
+         panel.grid = element_blank()) +
+         ggtitle(label = "Expenses vs. Age") +
+         xlab("Age") + ylab("Expenses")
```



Above is the scatterplot of *Expenses vs. Age*. Both variables are continuous variables. It is immediate that the relationship between *Expenses* and *Age* is linear. Moreover, there appears to be three linear trends in the scatter plot indicating that claims may be higher for some groups vs. others that we have yet to identify. Below a scatter plot of *Expenses vs. BMI* can be seen. Code to produce the image is omitted because it is very similar to the code that produced the scatter plot of *Expenses vs. Age*.



It appears that the relationship betweem *Expenses* and *BMI* is linear over the entire domain of *BMI* for only *some* of the observations.

Feature Engineering, Continued Data Visualization and Regression through the Origin

We have seen that there are three linear trends in the scatterplot of *Expenses vs. Age* and some linearity in the scatterplot of *Expenses vs. BMI*. Since the linearity between *Expenses* and *Age* appears to break into *groups*, this is indicative that we may be able to find *categorical variables* through feature engineering that uniquely defines a linear trend for each group. Visitors who have read *Machine Learning: Using the Logistic Regression Model to Predict Coronary Heart Disease*, may recall that the act of smoking and being overweight were some patient characteristics that were identified, and agreed upon by many health care professionals, to be major risk factors of absolute short-term risk of CHD development [2]. Since this is the case, and based on intuition, it is logical to assume that claims are likely to be higher for benficiaries that exhibit such characteristics. Thus, the beneficiaries can be separated into such groups.

In this section, we construct a handful of indicator variables that places subsets of the ordered pairs of *BMI* and the levels of *Smoker* into different groups. Recall, observations that belong to some category can be represented by an indicator variable that has been coded to indicate which group the observation belongs to [3]. Finally, from the scatterplot of *Expenses vs. Age*, it appears that *Expenses* increases at the same *constant rate* as *Age* increases in each linear trend. On the other hand, the variable *Expenses* tends to be automatically higher for some *groups* in the population of beneficiaries compared to others. If we were to model the trends with a linear model, the forms of appropriate models to capture the linear trends are

$$E[Y \mid X] = \beta_0 + \beta_{1,X_1} I_{X_1} + \beta_{2,X_2} I_{X_2} + ... + \beta_{K,X_K} I_{X_K} + \beta_{S_1} * X_1 + \beta_{S_2} X_2 \quad (1)$$

where (1) is a *multiple linear regression model* [3, 4, 5]. The β_{k,X_k} are intercepts in (1) for $k = 1, 2, ..., K$ possible groupings in the relationship of *Expenses vs. Age*, I_{X_K} for $k = 1, 2, ..., K$ are binary-valued indicator variables, and β_{S_1} is the slope for *Age* and β_{S_2} is the slope for the *Children* variable.

From our previous work in *Machine Learning: Using the Logistic Regression Model to Predict Coronary Heart Disease*, a patient is said to have a *normal body weight* when their $BMI \in [18.5000, 25.0000)$, *overweight* when their $BMI \in [25.0000, 29.9999)$, and *obese* when $bmi \geq 29.9999$), and *BMI* is measured in $\frac{kg}{m^2}$ [2]. In addition to the mentioned intervals, certain levels of *BMI* can put one at greater risk of developing diabetes. For example, it was found through the analysis of the *Study to Help Improve Early evaluation and management of risk factors Leading to Diabetes* (SHILED) study that persons with $BMI \geq 28$ in $\frac{kg}{m^2}$ were at a high risk of developing diabetes [6, 7]. In this section, we create indicator variables that represent such weight conditions and classify each person as either a smoker or non-smoker. We provide code that accomplishes this below and comment when appropriate.

```
> Expenses <- data.frame(age, bmi, children, smokeryes, expenses)
> attach(Expenses)
> #
> # Note, smokeryes = 1 when the person is a smoker and 0 if
> # they are not a smoker.
> #
> # The variables that indicate persons who can be classified as
> # underweight, having a normal body weight, being overweight, being
> # overweight and having an increased risk of developing diabetes,
> # or obese AND non-smokers are created below.
```

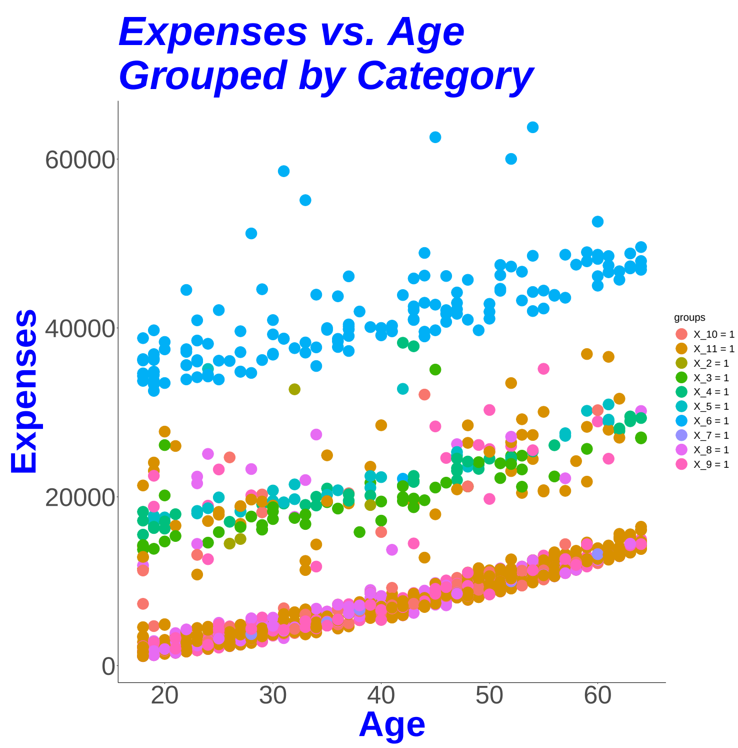
```
> #
>
> non_under <- ifelse(bmi <18.5000 & smokeryes == 0, 1, 0)
> non_normal <- ifelse(bmi < 25.0000 & bmi >= 18.5000 & smokeryes == 0,
+                       1, 0)
> non_over <- ifelse(bmi < 28.0000 & bmi >= 25.0000 & smokeryes == 0,
+                       1, 0)
> dm_risk_non <- ifelse(bmi < 29.9999 & bmi >= 28.0000 & smokeryes == 0,
+                       1, 0)
> non_obese <- ifelse(bmi >= 29.9999 & smokeryes == 0, 1, 0)
> #
> # The variables that indicate persons who are underweight
> # have a normal body weight, are overweight, have a bmi where
> # being diagnosed with diabetes is common but not obese, and
> # being obese AND each person is a smoker are created below.
> #
>
> smoke_under <- ifelse(bmi <18.5000 & smokeryes == 1,1,0)
> smoke_normal <- ifelse(bmi < 25.0000 & bmi >= 18.5000 & smokeryes == 1,
+                       1, 0)
> smoke_overweight <- ifelse(bmi >=25.0000 & bmi < 28.0000 & smokeryes==1,
+                             1,0)
> dm_risk_smoke <- ifelse( bmi < 29.9999 & bmi >= 28.0000 & smokeryes == 1,
+                         1, 0)
> smoke_obese <- ifelse(bmi >= 29.9999 & smokeryes == 1,1,0)
> Expenses <- data.frame(age, smoke_under, smoke_normal,
+                       smoke_overweight, dm_risk_smoke,
+                       smoke_obese, non_under, non_normal, non_over,
+                       dm_risk_non, non_obese, children, expenses)
> colnames(Expenses) <- c("X_1","X_2","X_3","X_4","X_5","X_6","X_7",
+                         "X_8","X_9","X_10",
+                         "X_11", "X_12", "expenses")
> attach(Expenses)
```

Before fitting a linear model, we describe the aliases of the independent variables we use to model *Expenses*. The list of aliases is found below.

- X_1 is the age of the primary beneficiary.
- $X_2 = 1$ if the person is a smoker and underweight. $X_2 = 0$ otherwise.
- $X_3 = 1$ if the person is a smoker and has a normal body weight. $X_3 = 0$ otherwise.
- $X_4 = 1$ if the person is a smoker, overweight, and is not diabetic or unlikely or undiagnosed diabetes. $X_4 = 0$ otherwise.

- $X_5 = 1$ if the person is a smoker, overweight, and is likely to develop diabetes. $X_5 = 0$ otherwise.
- $X_6 = 1$ if the person is a smoker and obese $X_6 = 0$ otherwise.
- $X_7 = 1$ if the person is a non-smoker and underweight $X_7 = 0$ otherwise.
- $X_8 = 1$ if the person is a non-smoker and has a normal body weight. $X_8 = 0$ otherwise.
- $X_9 = 1$ if the person is a non-smoker, overweight, and is not diabetic or unlikely to have undiagnosed diabetes. $X_9 = 0$ otherwise.
- $X_{10} = 1$ if the person is a non-smoker, overweight, and is likely to develop diabetes. $X_{10} = 0$ otherwise.
- $X_{11} = 1$ if the person is a non-smoker and obese. $X_{11} = 0$ otherwise.
- X_{12} is the number of dependents covered.

The grouped scatter plot of the *Expenses* vs. *Age* is found below.



The fitted regression model is

```
> OLS <- lm(expenses ~. + 0, data = Expenses)
> summary(OLS)

Call:
lm(formula = expenses ~ . + 0, data = Expenses)

Residuals:
     Min       1Q   Median       3Q      Max
-19610.3  -1844.4  -1297.7   -502.5  24426.4

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
X_1    265.686      8.858  29.995  < 2e-16 ***
X_2   9918.211   2048.579   4.842 1.44e-06 ***
X_3   9223.495    728.435  12.662  < 2e-16 ***
X_4  11318.625    785.355  14.412  < 2e-16 ***
X_5  12424.002    864.220  14.376  < 2e-16 ***
X_6  30595.517    519.187  58.930  < 2e-16 ***
X_7  -3687.904   1202.670  -3.066  0.00221 **
X_8  -2632.521    481.401  -5.468 5.42e-08 ***
X_9  -2421.257    491.358  -4.928 9.37e-07 ***
X_10 -2914.988    528.763  -5.513 4.24e-08 ***
X_11 -2533.612    418.543  -6.053 1.84e-09 ***
X_12   517.741    102.934   5.030 5.58e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4510 on 1326 degrees of freedom
Multiple R-squared:  0.9375,	Adjusted R-squared:  0.937
F-statistic:  1658 on 12 and 1326 DF,  p-value: < 2.2e-16

>
```

. Note, this is linear regression is a *multiple linear regression through the origin* because we have specified in the ***R*** *lm* function that we do not wish to include an intercept by including *+ 0*. We have computed a regression through the origin because we have strong reason to believe that $Expenses = 0$ for any $Age = 0$ making it appropriate to exclude β_0 from the model [8, 9].

In *Regression through the Origin (RTO)* by Eishenhauer includes a statement from Hocking (1996, pg. 177) concerning RTO when the data are far from the origin [8]. The argument is as follows: there is no certainty or proof that linearity exists in the dependency between Y and any X close to the origin if it so happens that the observed data are distant from the origin [8]. In other words, the the behavior of the process under study near the origin may exhibit some other *unobserved* functional behavior [8]. Again, in this application, it makes sense to conclude that $Expenses = 0$ when $Age = 0$ because the person is unborn.

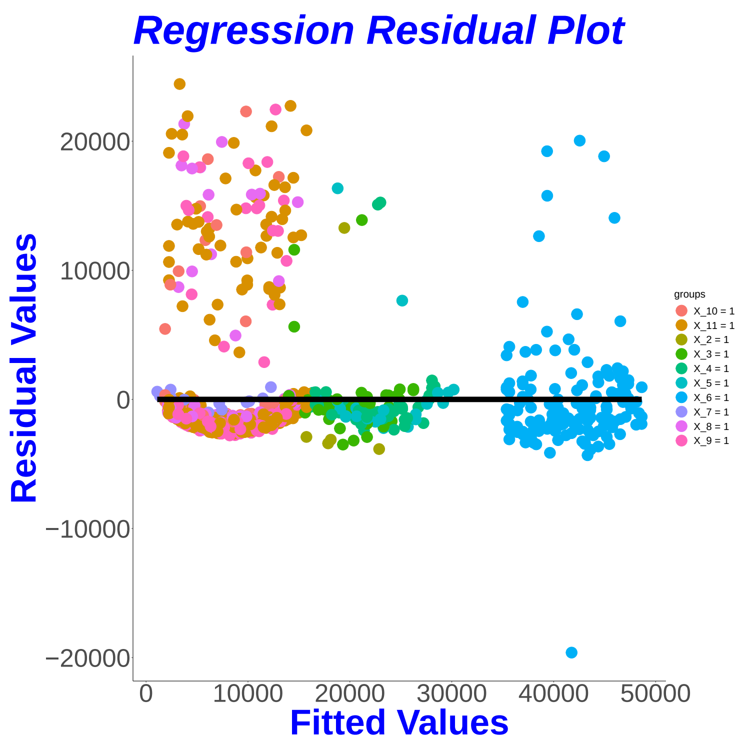
Secondly, we are confident that for $Age \in (0, 18)$ we must have $Expenses = 0$ because you have to be eighteen or older in the US to have an insurance policy. Therefore, the relationship between *Expenses* and *Age* must be constant at zero when $Age \in [0, 18)$ and *linear* when $Age \in [18, 64]$. Since this is the case, RTO is appropriate and the model will estimate a value of \$0.00 for persons with that have $0 \leq X_1 < 18$. Finally, we are not saying that minors do not have *any* medical expenses/costs. The multiple RTO model, with coefficients rounded to two decimal places, is

$$\begin{aligned} E[Y \mid \vec{X}] = 265.69 * X_1 + 9918.21 * X_2 + 9223.49 * X_3 + 11318.62 * X_4 \\ +12424.00 * X_5 + 30595.52 * X_6 - 3687.90 * X_7 - 2632.52 * X_8 \\ -2421.26 * X_9 - 2914.99 * X_{10} - 2533.612 * X_{11} + 517.74 * X_{12} \end{aligned} \tag{2}$$

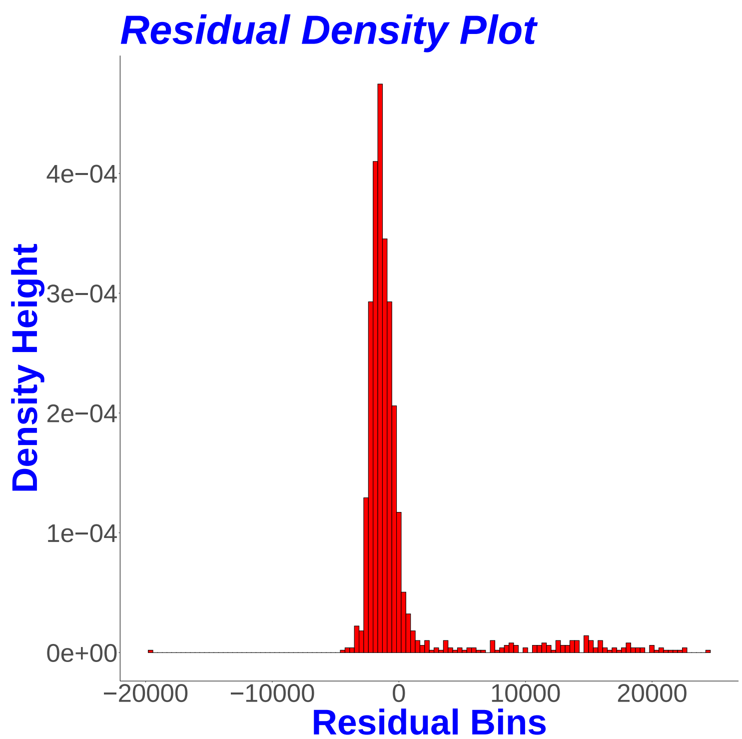
.

Regression Diagnostics

We have obtained a linear model that explains a significant percentage of variability in the target variable *Expenses*. Further, it appears that all the slopes for the independent variables are significantly different from zero. This can be seen from the summary print out. All *p-values*, or $\Pr(>\mid t \mid)$, are approximately zero. Secondly, $\boldsymbol{R^2} = 0.9375$ or 93.75% of the variability in *Expenses* is explained by this regression. In this section, we first produce a residual plot and histogram of the errors for this model. The residuals are color-coded according to group. This plot is found below.



The grouped residual plot shows that the variance may not be constant, there appears to be a large amount of *outliers*, and it appears that $\sum_{i=1}^{1338} \hat{e}_i \neq 0$. However, in regression through the origin $\sum_{i=1}^{1338} \hat{e}_i \neq 0$ [5]. Finally, it looks to be that there many large residuals for some of the nonsmoker cases, a few large residuals for some of the smoker cases as well as obsese smokers. The histogram of the residuals is found below.



It appears that the errors follow a normal distribution centered at $\mu < 0$ which is heavy tailed. To conclude this section, we test for nonconstant variance using a test based on work by *Cook* and *Weisberg* (1983) using their model that is comparable to what was discovered by *Breusch* and *Pagan* (1979) [5]. If $Var[e_i] \neq 0$ for all data cases $i = 1, 2, ..., n$ then the variance function of the errors is given by

$$Var[e_i] = \sigma^2 * e^{(\lambda^T \vec{z_i})} \quad (3)$$

and $Var[e_i] = \sigma^2$ for all i if $\lambda = 0$ [5]. So the hypotheses tested are $H_0 : \lambda = 0$ vs. $H_A : \lambda \neq 0$ [5]. The computations needed to obtain the test statistic(s) for the test are described below.

- Store the e_is obtained from regressing a response Y on the independent variables.
- Compute $\delta^2 = \frac{\sum_{i=1}^{n} \hat{e_i}^2}{n}$ and the vector α whose elements are $\frac{\hat{e_i}^2}{\delta^2}$ for $i = 1, 2, ..., n$ cases.
- Regress α on the vector z whose elements are either $\hat{y}_i$ if the variance is thought to depend on fitted values of the model. Or, z contains the regressors.
- Obtain the *regression sum of squares* from α regressed on z. Compute the test statistic C which follows a χ^2_1 distribution if z contains fitted

values. On the other hand, if z consists of p regressors, then C follows a χ^2_p distribution. In either case, C is the regression sum of squares divided by 2.

[5]. We carry out the test for our model below with $z = X$.

```
> rm1 <- residuals(OLS)
> deltasq <- (sum(rm1^2))/(nrow(Expenses))
> testalpha <- (rm1^2)/(deltasq)
> testreg <- lm(testalpha ~ X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7
+                + X_8 + X_9 + X_10 + X_11 + X_12 + 0)
> summary(testreg)

Call:
lm(formula = testalpha ~ X_1 + X_2 + X_3 + X_4 + X_5 + X_6 +
    X_7 + X_8 + X_9 + X_10 + X_11 + X_12 + 0)

Residuals:
    Min      1Q  Median      3Q     Max
-1.7666 -1.0067 -0.8953 -0.7163 28.4733

Coefficients:
      Estimate Std. Error t value Pr(>|t|)
X_1  -0.002342   0.006714  -0.349 0.727335
X_2   2.126723   1.552833   1.370 0.171051
X_3   0.457438   0.552158   0.828 0.407562
X_4   0.652547   0.595303   1.096 0.273208
X_5   0.592754   0.655083   0.905 0.365708
X_6   1.020200   0.393546   2.592 0.009638 **
X_7   0.065828   0.911630   0.072 0.942447
X_8   1.048076   0.364904   2.872 0.004141 **
X_9   1.260206   0.372451   3.384 0.000736 ***
X_10  0.880413   0.400805   2.197 0.028221 *
X_11  1.123161   0.317258   3.540 0.000414 ***
X_12  0.050580   0.078025   0.648 0.516935
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.419 on 1326 degrees of freedom
Multiple R-squared:  0.08343,       Adjusted R-squared:  0.07514
F-statistic: 10.06 on 12 and 1326 DF,  p-value: < 2.2e-16

> SSE <- 3.419^2*1326
> SYY <- (SSE)/(1 - 0.08343)
> C <- 0.5*(SYY - SSE)
> pvalue <- pchisq(q = C, df = 12, lower.tail = FALSE)
> pvalue
```

```
[1] 2.998693e-143

>
>
```

so we conclude that nonconstant variance is an issue because we reject $H_0 : \lambda = 0$ and is likely to be caused by our choice of independent variables. We next conduct the test to see if nonconstant variance is also caused by the fitted values. To do this, we make a minor adjustment to the code above where z contains fitted values from the model instead of the independent variables. We omit the code and report only the p-value of the test below.

```
[1] 0.1715319

Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 1.869569, Df = 1, p = 0.17152
```

which indicates that it is highly unlikey that nonconstant variance of the residuals is created by the fitted values returned by our multiple linear regression model. Note, the p-value from the procedure given by Weisberg (2014) is similar to the p-value returned from the ncvTest (nonconstant variance test) function.

A Discussion of Weighted Least Squares and Robust Regression

Weighted Least Squares

We saw in the previous section that nonconstant variance in the estimated residuals is most certainly an issue after we tested for the condition. Moreover, it seems that our choice of independent variables to model *Expenses* is most likely the cause of this violated OLS assumption. This is because the test for nonconstant variance we described in the previous section returned an extremely small p-value forcing us to reject $H_0 : \lambda = 0$. Additionally, has lead us to believe that $\lambda \neq 0$ due to our choice of indepedent variables. In this section, we discuss *weighted least squares* and *iteratively reweighted least squares* (IRLS) and estimate the linear model with the IRLS method.

Recall that in linear regression $Var[Y \mid X] = \sigma^2$ or $Var[e_i] = \sigma^2$ is assumed when we believe that a dependent variable can be modeled with a linear mean function [5]. In some cases, our assumption about the variance function may be wrong and the true variance function is

$$Var[Y \mid X] = \frac{\sigma^2}{w_i} \tag{4}$$

so σ^2 still plays a role in describing the variance function[5]. However, for $i = 1, 2, ..., n$ the value of the function is not constant and is dependent on the *realized* numbers $w_1, w_2, ..., w_n$ which are greater than zero [5]. If we assume that the variance function has this form, we must use the method of weighted least squares to compute approximations to the constants in the linear function instead of the ordinary least squares approach [5]. We know that $\hat{\boldsymbol{\beta}} = (X^T X)^{-1} X^T Y$ for linear regression but, for weighted least squares

$$\hat{\boldsymbol{\beta}} = (X^T W X)^{-1} X^T W Y \tag{5}$$

where each w_i appear as diagnonal entries of the $n \times n$ matrix W and $(W)_{i,j} = 0$ when $i \neq j$ so W is a digagonal matrix [5]. That is the appearance of W is

$$W = \begin{bmatrix} w_{1,1} & 0 & \cdots & 0 & 0 & 0 \\ 0 & w_{2,2} & \cdots & 0 & 0 & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 \\ \vdots & \vdots & \cdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & 0 & w_{n-1,n-1} & 0 \\ 0 & 0 & \cdots & 0 & 0 & w_{n,n} \end{bmatrix}$$

[5, 9]. Moreover, if W is invertible, the elements of W^{-1} must be $\frac{1}{w_i}$ for $i = 1, 2, ..., n$ since W is a diagonal matrix and $w_i > 0$ [9, 10].

As previously stated, we also suspect that we are dealing with many outliers in this problem as well. This is another issue we must account for. To identify outliers, we compute

$$\hat{\sigma^2_{stan}} = \frac{\hat{e_i}}{\hat{\sigma^2}\sqrt{1 - h_i}} \tag{6}$$

which is called a *standardized residual* for $i = 1, 2, ..., 1338$ cases [11]. Additionally, if case i has $\hat{\sigma^2_{stan}} > 2$ or $\hat{\sigma^2_{stan}} < -2$, then case i is sometimes called a *moderate outlier*. Finally, if $\hat{\sigma^2_{stan}} > 3$ or $\hat{\sigma^2_{stan}} < -3$ indicates case i is a serious outlier [11]. We compute the standardized residuals below and count how many moderate and serious outliers are present when using the OLS estimator below.

```
> library(MASS)
> library(sqldf)
> standardized_residuals <- data.frame(stdres(OLS))
> colnames(standardized_residuals) <- c("standardized_residuals")
> standardres <- data.frame(stdres(OLS))
> colnames(standardres) <- c("standardized_residuals")
> moderate_outliers <- sqldf(x = "SELECT standardized_residuals FROM
+                       standardres WHERE standardized_residuals < -2
+                       OR standardized_residuals > 2")
> serious_outliers <- sqldf(x = "SELECT standardized_residuals FROM
+                       standardres WHERE standardized_residuals
+                       < -3 OR standardized_residuals > 3")
> #
```

```
> # The number of moderate outliers is
> nrow(moderate_outliers)

[1] 91

> # The number of serious outliers is
> nrow(serious_outliers)

[1] 63
```

Robust Regression: Iteravtively Reweighted Least Squares

If OLS is used to estimate $\boldsymbol{\beta}$, outliers are present, and the residuals deviate from normality, $\hat{\boldsymbol{\beta}}$ can be severely distorted [12, 13, 14, 15]. Also, parameter estimated standard errors suffer because the variances of their errors increase in the presence of outliers can and contain *bias* [16]. In other words, even if it is found that it is reasonable to believe that the ϵ_is are normal, the standard error of $\hat{\boldsymbol{\beta}}$ may be quite large in the presence of outliers [17]. This appears to be the case in our problem as well.

In cases where there are many outliers, *robust regression* can be used to obtain healthy parameter estimates [18]. Secondly, since we have found evidence of nonconstant variance, *Strickland* claims that robust regression can be considered and employed in such scenarios [19]. In summary, one alternative to OLS regression when the constant variance and normality of the residuals are violated is robust regression [12, 19]. One robust method minimizes the function

$$\sum_{i=1}^{n} \rho(\frac{y_i - x_i^T \boldsymbol{\beta}}{\sigma}) \tag{7}$$

[12, 15]. Secondly, σ in (7) must also be estimated. This estimate is given by $\hat{\sigma} = \frac{median|\hat{e_i} - median(\hat{e_i})|}{0.6745}$ [12]. Finally, partial derivatives of (7) are calculated with respect to each β and then set to zero [12, 15]. The derivative of ρ is known as the *influence function* [12]. The partial derivative of (7) with respect to independent variable j is

$$\sum_{i=1}^{n} x_{i,j} \mu(\frac{y_i - x_i^T \boldsymbol{\beta}}{\sigma}) \tag{8}$$

where μ is the influence function or *weight function* which depends the residuals [15]. Note, If there is no intercept, p partial derivatves are calculated and set to zero for $j = 1, 2, ..., p$ independent variables. $\boldsymbol{\beta}$ is found by the method of *iteratively reweighted least squares* where

$$\boldsymbol{\beta_{t+1}} = (X^T W_t X)^{-1} X^T W_t Y \tag{9}$$

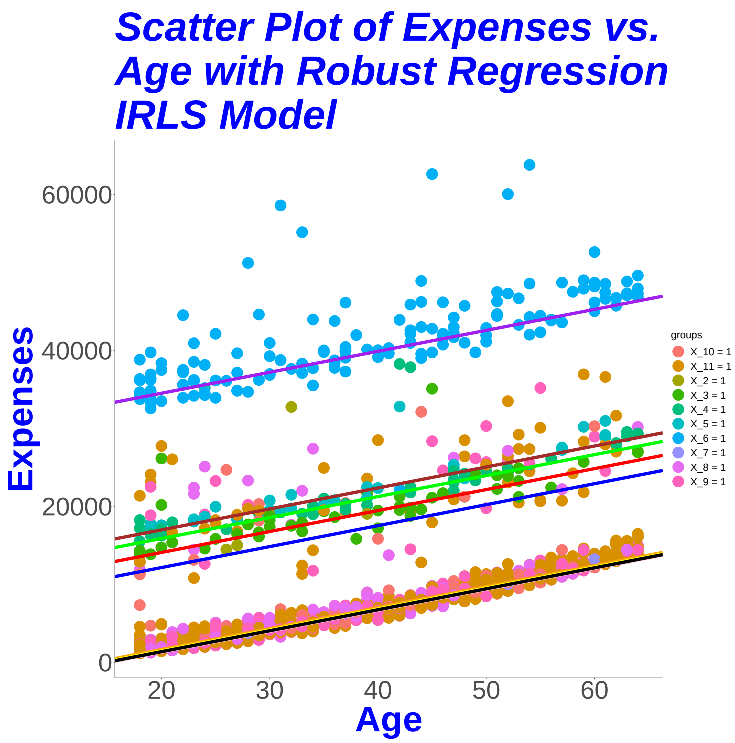
is calculated at each iteration and W_t is a diagonal weight matrix at iteration t [12, 15]. The elements of W_t are

$$w_{i,t} = \begin{cases} \mu(\frac{y_i - x^T \boldsymbol{\beta_t}}{\sigma_t}) \text{ when } y_i \neq x^T \boldsymbol{\beta} \\ 1 \text{ when } y_i = x^T \boldsymbol{\beta} \end{cases}$$

[12]. The algorithm is the following: at the first iteration, estimate the ordinary least squares coefficients to start the iterative process. Then, calculate the elements of W_{t-1} as well as the the residuals $\hat{e_{t-1}}$ for $i = 1, 2, ..., n$ cases from the previous iteration. Define a stopping criterion. The criterion is to stop when the change between $\boldsymbol{\beta_t}$ and $\boldsymbol{\beta_{t+1}}$ is small indicating that the algorithm has found a solution. Otherwise, continue the process until the change is is below some threshold [12, 15]. Using the Tukey's Biweight we have

```
> library(MASS)
> weightrob <- rlm(expenses ~ X_1 + X_2 + X_3 + X_4 + X_5 + X_6 +
+                    X_7 + X_8 + X_9 + X_10 + X_11 + X_12 + 0,
+                  data = Expenses,  method = "M",
+                  psi = psi.bisquare
+                  , wt.method = "case",
+                  maxit = 30, acc = 0.001)
> summary(weightrob)

Call: rlm(formula = expenses ~ X_1 + X_2 + X_3 + X_4 + X_5 + X_6 +
    X_7 + X_8 + X_9 + X_10 + X_11 + X_12 + 0, data = Expenses,
    psi = psi.bisquare, maxit = 30, acc = 0.001, method = "M",
    wt.method = "case")
Residuals:
      Min        1Q    Median        3Q       Max
-18260.16   -455.21     86.24    728.27  25946.61

Coefficients:
     Value      Std. Error t value
X_1    268.8105     1.5553  172.8326
X_2   6761.0615   359.7119   18.7958
X_3   8696.1243   127.9067   67.9880
X_4  10487.2831   137.9013   76.0492
X_5  11594.4741   151.7493   76.4055
X_6  29114.1457    91.1645  319.3584
X_7  -3744.6293   211.1780  -17.7321
X_8  -3923.2030    84.5296  -46.4122
X_9  -4059.0921    86.2780  -47.0467
X_10 -4093.0734    92.8461  -44.0845
X_11 -4025.8740    73.4924  -54.7795
X_12   427.3424    18.0744   23.6436

Residual standard error: 827.8 on 1326 degrees of freedom

>
```

Compared to the OLS fit, the iteratively least squares solution appears to be much better because the parameter standard errors are much smaller. Secondly, the residual standard error is significantly smaller compared to the OLD residual standard error which is 4510. Below one can find the model fit to the data below where the value of *Children* is held constant at zero.



Unfortunately the plot above is very busy and complicated. The takeaway is that it the linearity assumption appears to hold and the model passes through the bulk of the linear trends in the grouped data. Note, this is a *two dimensional plot*. In this two dimensional plot, we hold X_{13} constant at zero. Similar plots can be created for different values of *Children*. In other words, this is the multiple linear regression model fit to the data where none of the primary beneficiaries have children. Below we return 95% confidence intervals for the model coefficients of the IRLS model.

```
          2.5 %     97.5 %
X_1    265.7621   271.8589
X_2   6056.0390  7466.0839
X_3   8445.4319  8946.8168
X_4  10217.0014 10757.5647
X_5  11297.0510 11891.8972
X_6  28935.4666 29292.8248
```

```
X_7  -4158.5306 -3330.7280
X_8  -4088.8781 -3757.5280
X_9  -4228.1938 -3889.9904
X_10 -4275.0484 -3911.0984
X_11 -4169.9164 -3881.8316
X_12   391.9173   462.7676
```

The model's R-squared value is computed below.

```
> SST <- sum((expenses - mean(expenses))^2)
> RRegSSE <- sum((expenses - fitted.values(weightrob))^2)
> RRegRSQ <- 1 - (RRegSSE)/(SST)
> RRegRSQ
```

```
[1] 0.8493014
```

or $\boldsymbol{R^2} = 84.93\%$ which is lower than the OLS R-squared. Finally, we estimate the test mean square error with a 50-50 split.

```
> set.seed(3)
> train <- sample(x = nrow(Expenses),
+                 size = round(x = 0.50* nrow(Expenses), digits = 0),
+                 replace = FALSE)
> test <- Expenses[-train, ]
> test.rlm <- rlm(expenses ~ X_1 + X_2 + X_3 + X_4 + X_5 + X_6 +
+                     X_7 + X_8 + X_9 + X_10 + X_11 + X_12 + 0,
+                   data = Expenses,  method = "M",
+                   psi = psi.bisquare
+                   , wt.method = "case",
+                   maxit = 30, acc = 0.001,
+                   subset = train)
> test.predictions <- predict(object = test.rlm, newdata = test,
+                              interval = "none",
+                              type = "response")
> Test.MSE <- mean((test$expenses - fitted.values(test.rlm))^2)
> Test.MSE
```

```
[1] 263268233
```

So the average squared error is $263,268,233$ dollars squared and the root mean squared error is $\$16,225.54$. The training MSE is

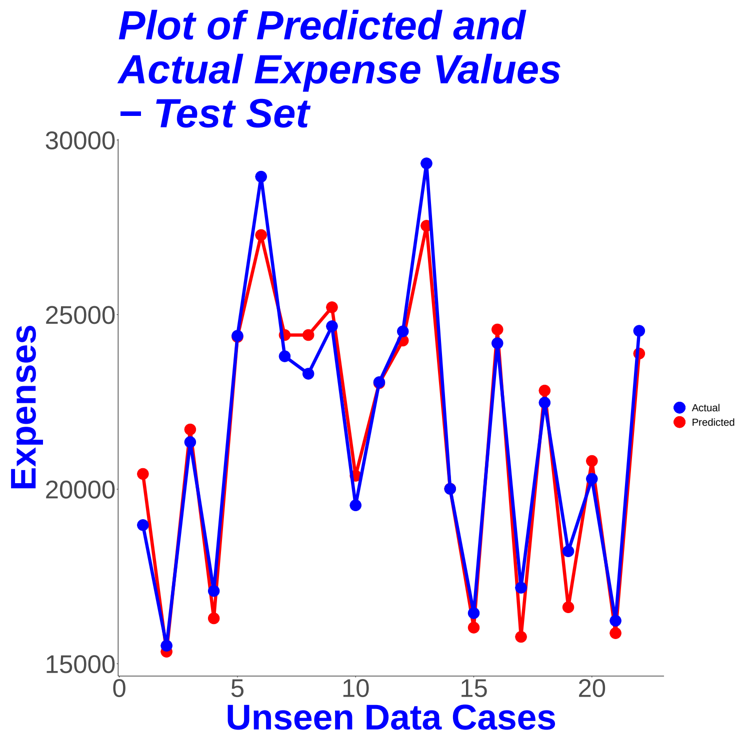
```
> r <- predict(object = weightrob, newdata = Expenses,
+               interval = "none", type = "response")
> Train.MSE <- mean((Expenses$expenses - fitted.values(weightrob))^2)
> Train.MSE
```

```
[1] 22083789
```

The training MSE is much smaller than that of the test MSE. With the validation set approach, the error may be quite large due to some of the erratic behavior of the data. In general, we do not expect the test error to be much larger than the training MSE and root MSE. The summary of the training IRLS model is seen below.

```
Call: rlm(formula = expenses ~ X_1 + X_2 + X_3 + X_4 + X_5 + X_6 +
    X_7 + X_8 + X_9 + X_10 + X_11 + X_12 + 0, data = Expenses,
    psi = psi.bisquare, maxit = 30, acc = 0.001, subset = train,
    method = "M", wt.method = "case")
Residuals:
     Min       1Q   Median       3Q      Max
-3077.60  -411.83    78.98   662.22 24134.24

Coefficients:
     Value      Std. Error t value
X_1    265.1525     2.1068   125.8560
X_2   6924.2144   433.6041    15.9690
X_3   8822.5008   159.4394    55.3345
X_4  10574.3056   186.3066    56.7576
X_5  12016.4095   218.3365    55.0362
X_6  29110.6421   122.4992   237.6394
X_7  -3872.3776   309.4929   -12.5120
X_8  -3741.0493   110.8177   -33.7586
X_9  -3981.6716   112.9287   -35.2583
X_10 -3980.9725   125.4807   -31.7258
X_11 -3929.5068    98.1382   -40.0405
X_12   424.1781    23.7419    17.8662

Residual standard error: 768.2 on 657 degrees of freedom
```

The change in the coeffcients of the model fit on the training set seems to be small. Instead of relying on the estimation of the test MSE and test root mean squared error, we examine the closeness between the test observations and predicted values with the graph below.



Above we see that the majority of the predicted values are quite close to the *actual* expense values. Note, predictions in the plot were made on *unseen* data cases. Additionally, in this plot, we subsetted the test set such that $X_4 = 1$ in each data case and all other indicator variables were not *activated*. In other words, the test cases only consisted of insurance beneficiaries who are overweight smokers who are not likely to suffer from diabetes. Of course, the other independent variable is *Age*. The largest test error on under these circumstances is

```
[1] 1786.916
```

which is $\$1,786.92$. So the model appears to perform reasonably well on unseen cases that lie in this particular subset of X. Suppose that we want to predict *Expenses* for a policyholder who is 33 years old, has no children, and is a smoker but, is not likely to develop diabetes. Additionally, highly unlikely to have undiagnosed diabetes. The prediction is

$$E[Y \mid X^*] = 268.81 * 33 + 10487.28 + 427.34 * 0 \tag{10}$$

which rounds to $\$19,358.01$ for an entire year.

Conclusion

In this paper, we conducted a regression through the origin using ordinary least squares to predict medical expenses for a population of beneficiaries on an insurance plan. Along the way, we ran into some complications. I.e., we witnessed that some of the ordinary least squares assumptions were violated. Namely the assumption of normality and constant variance of the errors. We then briefly described weighted least squares and iteratively reweighted least squares. After estimating the regression parameters and standard errors via IRLS, we saw a significant decrease in the standard errors of the parameter estimates. We then assessed the accuracy of the model using the validation set approach. Additionally, we examined the closeness of the fitted values to actual values for the IRLS model for unseen data cases. From what was observed, it appeared that the model performed well on these particular unseen data. Similar visual analysis should be conducted when making predictions about *Expenses* for beneficiaries that fall into different groupings. Finally, it appears that the IRLS model is superior compared to the OLS fit.

References

[1] Brett Lantz. *Machine learning with R: learn how to use R to apply powerful machine learning methods and gain an insight into real-world applications.* Packt Publishing, 2013.

[2] Victor Wright. Machine learning: Using the logistic regression model to predict coronary heart disease, 2019.

[3] Ann R. Cannon, George W. Cobb, Bradley A. Hartlaub, Julie M. Legler, Robin H. Lock, Thomas L. Moore, Allan J. Rossman, and Jeffrey A. Witmer. *STAT2: building models for a world of data.* W.H. Freeman, 2013.

[4] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning with applications in R.* Springer, 2017.

[5] Sanford Weisberg. *Applied linear regression.* Wiley, 2014.

[6] N. G. Clark, K. M. Fox, and S. Grandy. Symptoms of diabetes and their association with the risk and presence of diabetes: Findings from the study to help improve early evaluation and management of risk factors leading to diabetes (shield). *Diabetes Care*, 30(11):2868–2873, 2007. doi: 10.2337/dc07-0816.

[7] Harold E Bays, Debbra D Bazata, Nathaniel G Clark, James R Gavin, Andrew J Green, Sandra J Lewis, Michael L Reed, Walter Stewart, Richard H Chapman, Kathleen M Fox, and et al. Prevalence of self-reported diagnosis of diabetes mellitus and associated risk factors in a national survey in the

us population: Shield (study to help improve early evaluation and management of risk factors leading to diabetes). *BMC Public Health*, 7(1), Mar 2007. doi: 10.1186/1471-2458-7-277.

[8] Joseph G. Eisenhauer. Regression through the origin. *Teaching Statistics*, 25(3):76–80, 2003. doi: 10.1111/1467-9639.00136.

[9] Sanford Weisberg. *Applied linear regression*. John Wiley and Sons, 1985.

[10] Howard Anton. *Elementary linear algebra*. John Wiley Sons Inc., 2014.

[11] Ann R. Cannon. *STAT2: building models for a world of data*. W.H. Freeman, 2013.

[12] *Chapter 308 Robust Regression*.

[13] Y. Susanti, H. Pratiwi, S. Sulistijowati H., and T. Liana. M estimation, s estimation, and mm estimation in robust regression. *International Journal of Pure and Apllied Mathematics*, 91(3), Aug 2014. doi: 10.12732/ijpam.v91i3.7.

[14] John Fox. Robust regression: Appendix to an r and s-plus companion to applied linear regression, January 2002.

[15] John Fox and Sanford Weisberg. Robust regression*, October 2013.

[16] Robert Yaffee. Robust regression analysis: Some popular statistical packages. Connect, 12 2002.

[17] Rand R. Wilcox and H. J. Keselman. Robust regression methods: Achieving small standard errors when there is heteroscedasticity. *Understanding Statistics*, 3(4):349–364, 2004. doi: 10.1207/s15328031us0304$_8$.

[18] Colin Chen. Robust regression and outlier detection with the robustreg procedure. 01 2002.

[19] PRESIDENT JEFFREY. STRICKLAND. *PREDICTIVE ANALYTICS USING R*. LULU COM, 2015.