Brave Wanderers Final Project: A Linear Regression Model for the Prediction of Glucose Levels

GU-GPH-2353
Spring 2022

Alex Holbrook, Mandy Lacroix, Walter Carlos

Introduction

The Pima Indians are a nation of indigenous people who live in both southern Arizona, United States and northwestern Mexico. A notable public health challenge among the Pima people is their disproportionately high rate of diabetes; they have among the highest reported prevalence levels of diabetes worldwide. Diabetes is a heterogeneous disease, meaning there are several different pathways that can result in its development (Staimez et al, 2019). Research suggests that the Pima people are more likely to develop diabetes due to a combination of genetic factors (Nair et al, 2016, Muller et al, 2010) and environmental factors (Esparza-Romero et al., 2015).

Known as the Akimel O’odham or “River people" in their native language, the Pima people are descendants of the Hohokam culture and have traditionally lived around the Gila and Salt Rivers. Historically, local chiefs would elect a tribal chief to oversee the wellbeing of the entire community. They are related to the Tohono O’odham (Papago) people, whose traditional homeland is near that of the Pima. The population of Pima descendents was estimated to be approximately 11,000 in the early 21st century (Pauls, 2019).

The Pima are traditionally sedentary farmers, although sometimes they engage in hunting and gathering, especially during droughts (Pauls, 2019). As many Pima people in the United States have adopted a more modern lifestyle (less reliance on manual labor, consumption of processed foods, etc.), they have experienced a higher prevalence of type 2 diabetes and obesity than their counterparts in Mexico—who have lived a more traditional lifestyle. However, Mexican Pima people have seen an increase in diabetes and obesity between 1995 and 2010 correlated with their own adoption of a modernized lifestyle (Esparza-Romero et al, 2015).

Research has shown that Pima people have high insulin resistance at an average BMI (Staimez et al., 2019), suggesting that they are more susceptible to type 2 diabetes. Pima people were reported to have high insulin resistance across BMI strata, including average weight and normoglycemic individuals (Staimez et al., 2019). A number of genes present in the Pima population have been considered linked to type 2 diabetes (Muller et al., 2010), including some that are related to insulin resistance in Pima women (Nair et al., 2016).

Diabetes is associated with the development of vascular disease and its subsequent abnormalities, as prolonged high levels of glucose damage the body's vasculature (Adeva-Andany et al., 2019). Pregnancy causes massive changes to a person's vasculature in order to support the growing fetus, and is thus a risk factor for the development of type 2 diabetes post-birth (Bellamy et al., 2009). Additionally, if a person was diagnosed with gestational diabetes during their pregnancy, their risk of developing type 2 diabetes post-pregnancy has been reported to be either 7.43 [95% CI, 4.79-11.51] or 9.51 [95% CI, 7.14 to 12.67] by two different meta-analysis (Bellamy et al., 2009 & Vounzoulaki et al., 2020).

Obesity has long been a known risk factor for the development of type 2 diabetes, the codiagnosis of which is colloquially known as “Diabesity” (Karla et al., 2013). The pathophysiological mechanisms around which obesity predisposes a person to type 2 diabetes is hypothesized to mainly involve insulin resistance. Moreover, high BMI and pregnancy are both risk factors for the development of type 2 diabetes.

In 1999, Sievers et al. published a study on the effects of glycemia on mortality in the Pima people with type 2 diabetes. The researchers were interested in the possibility that glycemic levels influence cardiovascular mortality among people with type 2 diabetes and chose to study this population. As part of the study, the researchers conducted laboratory examinations of 5,506 people of at least half Pima or Tohono O'odham heritage who had been diagnosed with type 2 diabetes. These included a glucose test and an insulin test. The researchers also collected demographic information about the participants (Sievers et al., 1999). The purpose of this paper is to use the information collected from this study to conduct statistical analyses and draw conclusions about the relationship between glucose and the predictor variables detailed below .

Variables

Figure 1 shows histograms for all seven variables included in this model. These histograms are of the pre-cleaned, untransformed variables.

For the variable pregnant, the average number of pregnancies was 3.8, with a median of 3 and a range of 0-17. The distribution of the data is positively skewed. While it seems unusual to have such a high number of pregnancies in some of the study participants, it should be emphasized that it was not clear whether the respondent answered for the number of pregnancies that were carried to term, or had included those that did not.

The variable diastolic has a data range of 0-122, with a minimum value that is incompatible with life. Other than the zero values, the data seems to be normally distributed with an average and median of 69.1 and 72, respectively.

The variable triceps also has a minimum of 0 and a maximum of 99. The distribution is positively skewed with a median of 23 and an average of 20.54.

The variable insulin has a range of 0-846, with a mean of 79.8, a median of 30.5, and the distribution is also positively skewed. The zero values of insulin are not incompatible with life and do not represent missing data.

Body mass index (BMI) has a range of 0-67.1, an average of 31.99, and a median of 32. Excluding the 0 values for BMI, which are also incompatible with life, it seems to be normally distributed.

The final predictor variable considered for this analysis, age, consists of a range of 21 to 81, with a median age of 29 and an average age of 33.24. The distribution is positively skewed.

The outcome, glucose, showed a normal distribution with an average glucose concentration value of 120.9, a median of 117, and a range of 0-199. While a value of 0 is not impossible, it would indicate that the patient was having a hypoglycemic episode and was unlikely to have been included in this study.

From these summary statistics it was determined that the zero values of glucose, triceps and BMI were to be recoded as "NA" in the data set. Due to their distributions, the variables triceps, pregnant, and age were considered for transformation to categorical variables. Both triceps and age were recoded along their percentiles to yield four groups with an equal number of observations. Pregnancy was recoded along its percentiles, but the observation of zero pregnancies was also included as its own category. The categories for these three recoded variables are listed in the data dictionary below.

Insulin was considered for transformation to a higher power—(insulin)^2— due to its distribution with glucose. Figure 2 shows the pairwise correlation plot of all the variables in the data set. There was a seeming positive correlation between insulin and glucose. Figure 3 shows insulin and insulin^2 plotted against glucose. In the plot of insulin vs. glucose, there seems to be a possible parabolic relationship. The plot of insulin^2 vs. glucose seems to have made this relationship more linear, but there still seems to be a curvilinear relationship between the two variables. The higher power of insulin was not more significant than the untransformed variable in the models developed in this investigation, nor did it improve the model fit. Therefore, it is not included in the discussion of model development.

Insulin was transformed later in the analysis to the value (insluin +1). This was done so that a BoxCox analysis could be carried out. Running the initial model with insulin and running it with (insulin + 1) did not change the coefficient nor the p-value of insulin.

Data Dictionary		
Variable	**Description**	**Units**
glucose	Plasma glucose concentration at 2 hours in an oral glucose tolerance test	mmol/liter
pregnant	The number of times a participant has been pregnant	pregnancy
pregnant_cat	Values: 0 → never been pregnant 1 → 1 pregnancy 2 → 2-3 pregnancies 3 → 4-6 pregnancies 4 → more than 6 pregnancies	
diastolic	The diastolic blood pressure of a participant	mm/Hg
triceps	The skinfold thickness of a participant's triceps	mm
triceps_cat	Values: 1 → value of 0 2 → value greater than 0 and less than 23.00 3 → value between 23 and 32 4 → value greater than 32	
insulin	The result of a two hour serum insulin test	mU/ml
Insulin +1	All values were increased by a value of 1	

bmi	Body Mass index as measured by a participant's height and weight	kg/m2
age	The age of a participant in years	years
age_cat	Values: 1 → less than 24 2 → 24-29 3 → 30-41 4 → greater than 41	

Multicollinearity and VIF

Figure 2. shows the analysis for multicollinearity of the variables included in the model. The pairwise correlation chart showed possible relationships between glucose and insulin, between triceps and BMI, and between age and diabetes. Additionally there was a significant difference between the initial model and the null model when compared via ANOVA, and the adjusted-R^2 value of the initial model was quite low with insulin, age, and diastolic all being statistically significant. Therefore, there was an implication of multicollinearity.

However, when checking the condition numbers and variable inflation factors (VIF), multicollinearity did not appear to be a problem. Figure 2 shows that the condition numbers were not overtly overwhelming. Additionally, the VIF's were small, all being in the range of 1.178746-1.506947.

Thus, multicollinearity did not seem to be a significant problem with this data set.

Initial Model and Interpretation

Figure 4. shows three of the models run for the initial model. To determine the best initial model, we ran models with all variations of the categorical and untransformed variables. The model that had the best fit was the model that included the categorically transformed pregnancy variable, insulin, age, diastolic, bmi and tricep predictors; given by lmod <- lm(glucose~factor(pregnancy_cat)+ insulin + age + diastolic + bmi + triceps, data=pima)

This model has an R-squared of 0.3131, an adjusted R-squared of 0.3012, a residual standard error of 25.91, 522 degrees of freedom and a significant p-value of 2.2e^{-16}. The significant predictors—with alpha values less than 0.05— were categories 1 and 3 for pregnancy, insulin, age, and diastolic blood pressure.

All of the categories for the pregnancy variable had negative coefficients. These coefficients became less negative with each category. Insulin had a positive coefficient of 0.10 at a p-value of 2e^{-16}. Age had a positive coefficient of 0.596759 with a p-value of 4.21e^{-}5. Diastolic also had a positive coefficient of 0.256783 with a p-value of 0.0133.

Neither BMI nor triceps were significant predictors in this model. BMI had a positive coefficient of 0.221454 and a p-value of 0.3324. Triceps had a positive coefficient of 0.221092 and a p-value of 0.1202.

Figure 5. shows the model diagnostic plots for this initial model. The residuals vs. fitted plot shows possible non-constant variance. The values of the residuals seem to narrow, moving toward larger fitted values. The ab line for the residuals vs. fitted plot does look fairly linear, with a few points of large fitted values pulling it down toward the right of the graph.

The Q-Q plot shows that the data is roughly linear until the larger values in the data set, as the values of the standardized residuals start to deviate around a theoretical quantile value of 1. The scale-location plot shows that there also is some nonconstant variance. Finally, the residuals vs. leverage plot shows that there are values that should be considered as possible outliers that may be affecting the linearity assumptions of this model.

Box Cox Transformation

Figure 6. Shows the values returned for the Box Cox transformation. Running Box Cox the first time, the results showed a 95%CI centered around a lambda of 0. This result suggests a logarithmic transformation of the outcome variable glucose.

When this transformation was applied to the model, a second Box Cox analysis was run to determine if any further transformations were indicated. The figure on the top left of Figure 6 shows the results for this secondary Box Cox. The value returned a 95% CI that was centered around 1, which suggests that no further transformation was necessary.

The last image is the updated Q-Q plot after the transformation. While the tails of the data still are not perfectly normal, the line fits much better than the untransformed model. Not shown is the residuals vs. fitted plot. The log transformation of glucose did help decrease the observed heteroscedasticity, but there still was a slight trend of decreasing variance with increasingly large fitted values.

Identifying Outliers and Rerunning the Model

Figure 7. shows a histogram of the hatv values calculated for the data set. Calculating the leverage cut-off value as 2*(p/n) returned a value of 0.015625. Of the original 768 observations in the data set, 96 were above this threshold value. This is far too many values to consider as outliers, so alternative methods of identifying values of high leverage and influence were used.

Looking back to Figure 5. these diagnostic plots identify some extreme values. In the Residuals vs. Fitted plot, the observations 662, 400 and 186 had extreme residual values. The Q-Q plot identified these same extreme values, as did the scale location plot. The Residuals vs. Leverage plot identified points 550 and 585 as having high standardized residual values. Additionally, it found that point 580 had a very large leverage value. Finally, this plot did not find any points that had a Cook's distance greater than 0.5.

Figure 7b. shows the half-norm plot for the initial model. It identified observations 403 and 382 as being extreme values. Thus, these 8 observations were initially dropped from the data set, and the initial model was rerun.

Rerunning this model and revisiting the model diagnostic plots, observation 14 came under scrutiny. Observation 14 has an extreme value of insulin (846 mU/ml), and so it was also dropped from the data set.

Figure 7c. shows the model summary statistics after the outliers were removed. This model had a residual standard error of 0.2047 with 513 degrees of freedom; the overall values in the model were reduced because of the log transform of glucose. Its R-squared value was 0.3313 and the adjusted R-squared value was 0.3195. The model has the same p-value as the initial model, at 2.2e^{-16}.

The second category for pregnancy became significant in this model. Additionally, BMI, triceps and the fourth category of pregnancy remained insignificant predictor values.

AIC and Backward Elimination

The AIC results are not shown in the figures below, but are accessible in the attached R markdown file.

Moving toward building a final model an AIC was run to help assess if any of the variables should be dropped from the model. In running the AIC, the model with 4 predictors had the lowest score. The 4 predictor model that had the lowest AIC score was the one which excluded pregnant and triceps as predictor variables.

Running this 4 predictor model resulted in a model that did not meet the linearity assumptions more than the 6 predictor model, and had a reduced adjusted R-squared value of 0.2243. For these reasons, this model was not favored over the previous one.

To further see if any variables should be dropped from the model, backward elimination was implemented. In this process, the predictors BMI and triceps were dropped because they were not significant predictors. This resulted in a model that again did not meet the linearity assumptions more than the 6 predictor model, and also had a reduced adjusted R-squared value of 0.22. Again, this model was not favored.

Final Model and Interpretation

The final model for this analysis is shown in Figure 8. and given by the following code: lm(log(glucose) ~factor(pregnant_cat) + insulin + age +diastolic + bmi +triceps, data=pima2). The summary statistics are described above in the identifying outliers section. The diagnostic plots show improvement in meeting the linearity assumptions from the initial model. There is still evidence of heteroscedasticity in the residuals vs. fitted plot, and the tails of the Q-Q plot still show signs of non-linear data.

This model states that glucose levels are expected to decrease with successive pregnancies, however the level of effect successive pregnancies have on glucose levels decreases. Insulin, age, diastolic blood pressure, BMI, and triceps all have positive coefficients and so, increases in any of these variables are predicted to increase glucose levels. Moreover, BMI and triceps are not significant predictors in the model with p-values of 0.28872 and 0.60536 respectively.

This final model ultimately shows that it is difficult to predict glucose values with these predictor variables.

Citations:

1. Adeva-Andany MM, Funcasta-Calderón R, Fernández-Fernández C, Ameneiros-Rodríguez E, Domínguez-Montero A. Subclinical vascular disease in patients with diabetes is associated with insulin resistance. Diabetes Metab Syndr. 2019 May-Jun;13(3):2198-2206. doi: 10.1016/j.dsx.2019.05.025. Epub 2019 May 23. PMID: 31235157.
2. Bellamy L, Casas JP, Hingorani AD, Williams D. Type 2 diabetes mellitus after gestational diabetes: a systematic review and meta-analysis. Lancet. 2009 May 23;373(9677):1773-9. doi: 10.1016/S0140-6736(09)60731-5. PMID: 19465232.
3. Esparza-Romero J, Valencia ME, Urquidez-Romero R, et al. Environmentally Driven Increases in Type 2 Diabetes and Obesity in Pima Indians and Non-Pimas in Mexico Over a 15-Year Period: The Maycoba Project. Diabetes Care. 2015;38(11):2075-2082. doi:10.2337/dc15-0089
4. Kalra S. Diabesity. J Pak Med Assoc. 2013 Apr;63(4):532-4. PMID: 23905459.
5. Knowler WC, Saad MF, Pettitt DJ, Nelson RG, and Bennett PH. Determinants of Diabetes Mellitus in the Pima Indians. Diabetes Care. 1993; 16(1): 216-227.
6. Lavery JA, Friedman AM, Keyes KM, Wright JD, and Ananth CV. Gestational diabetes in the United States: temporal changes in prevalence rates between 1979 and 2010. BJOG: Int J Obstet Gynaecol. 2016; 124(5): 804-813.
7. Moses RG. The Recurrence Rate of Gestational Diabetes in Subsequent Pregnancies. Diabetes Care. 1996; 19(12): 1348-1350.
8. Muller YL, Hanson RL, Bian L, et al. Functional variants in MBL2 are associated with type 2 diabetes and pre-diabetes traits in Pima Indians and the old order Amish. Diabetes. 2010;59(8):2080-2085. doi:10.2337/db09-1593
9. Nair AK, Piaggi P, McLean NA, et al. Assessment of established HDL-C loci for association with HDL-C levels and type 2 diabetes in Pima Indians. Diabetologia. 2016;59(3):481-491. doi:10.1007/s00125-015-3835-x
10. Pauls EP. Pima. Encyclopædia Britannica. https://www.britannica.com/topic/Pima-people. Published May 27, 2019. Accessed May 13, 2022.
11. Sievers ML, Bennett PH, Nelson RG. Effect of Glycemia on Mortality in Pima Indians with Type 2 Diabetes. Diabetes. 1999; 48: 896-902.
12. Staimez LR, Deepa M, Ali MK, Mohan V, Hanson RL, Narayan KMV. Tale of two Indians: Heterogeneity in type 2 diabetes pathophysiology. Diabetes Metab Res Rev. 2019;35(8):e3192. doi:10.1002/dmrr.3192
13. Vounzoulaki E, Khunti K, Abner SC, Tan BK, Davies MJ, Gillies CL. Progression to type 2 diabetes in women with a known history of gestational diabetes: systematic review and meta-analysis. BMJ. 2020 May 13;369:m1361. doi: 10.1136/bmj.m1361. PMID: 32404325; PMCID: PMC7218708.

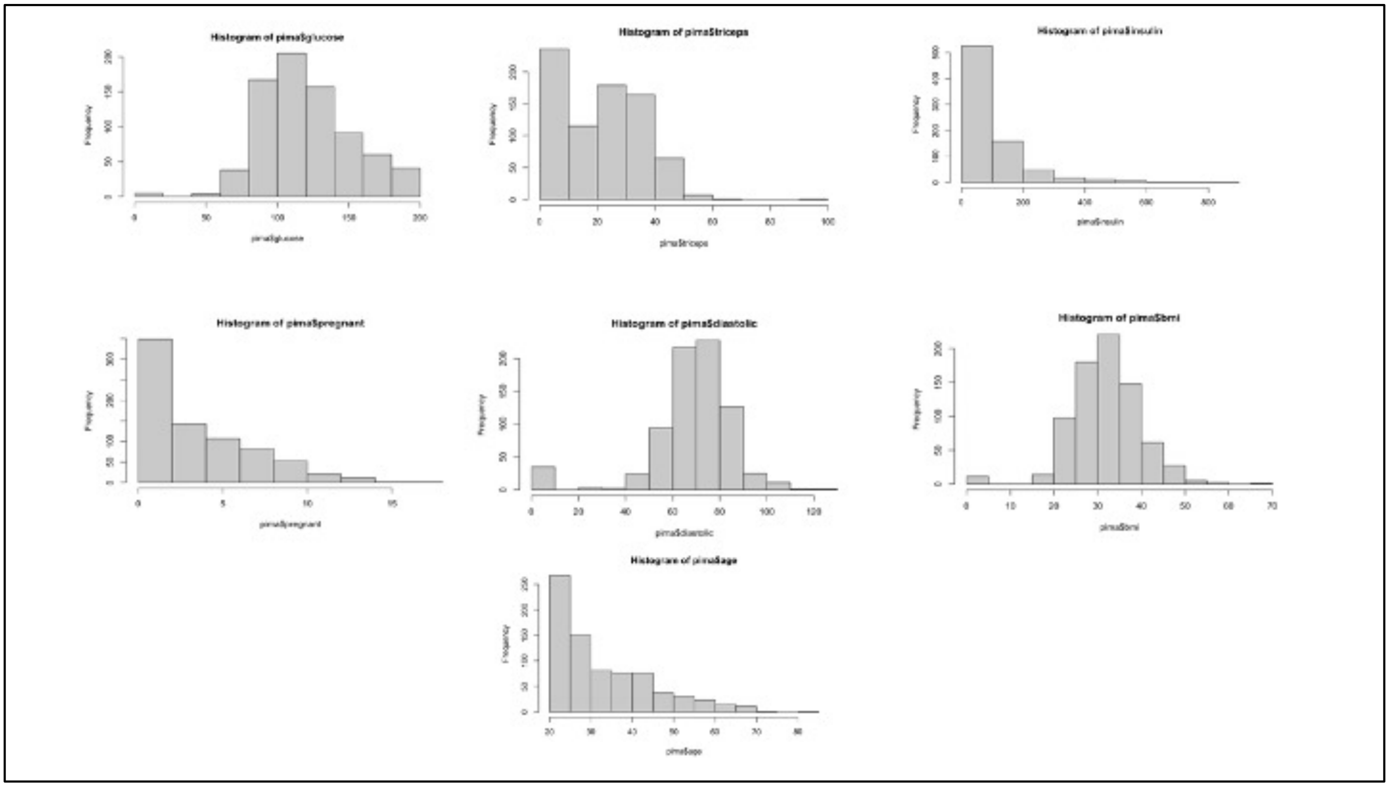
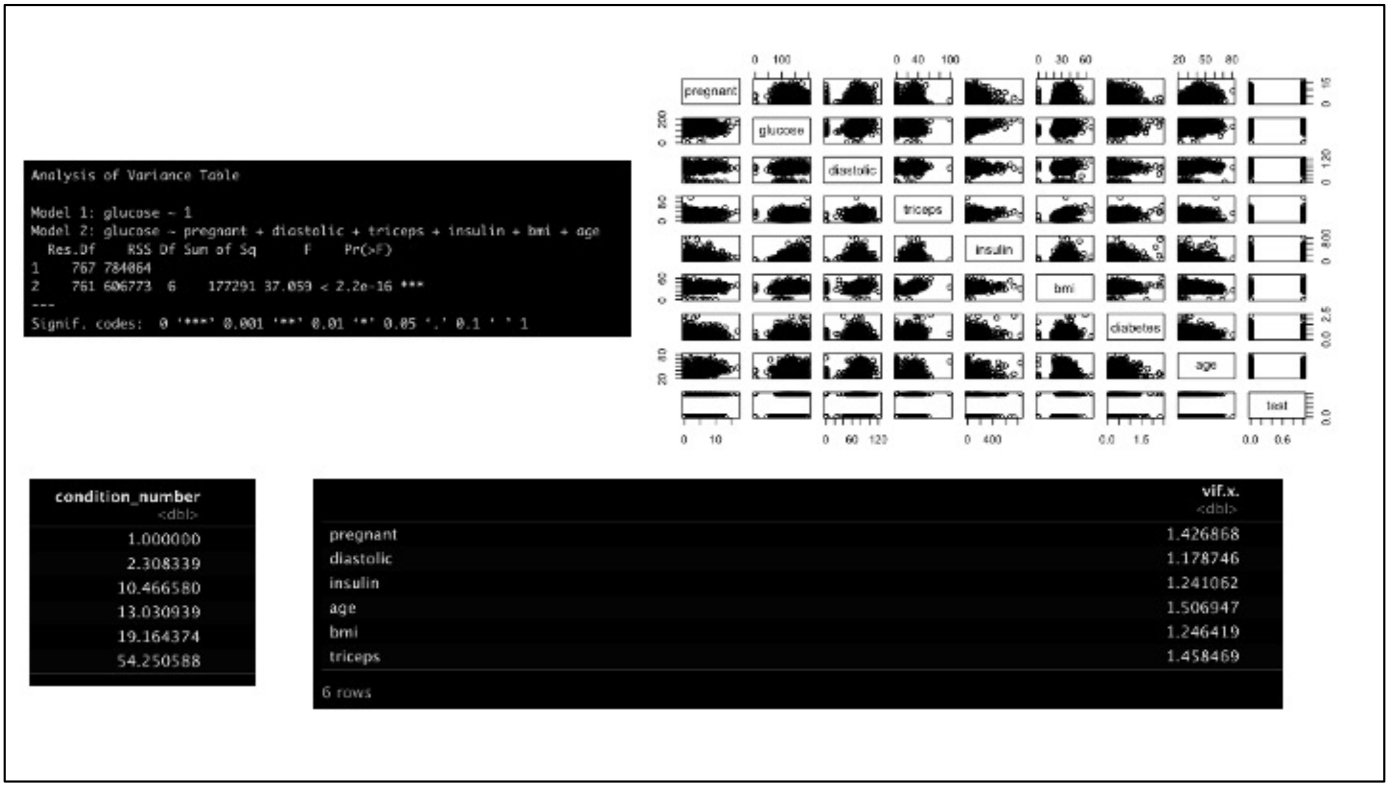


Figure 1. Histograms of the variables considered in the model before data cleaning and transformations.



```
Analysis of Variance Table

Model 1: glucose ~ 1
Model 2: glucose ~ pregnant + diastolic + triceps + insulin + bmi + age
  Res.Df    RSS Df Sum of Sq      F    Pr(>F)
1    767 784064
2    761 606773  6    177291 37.059 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

condition_number <dbl>
1.000000
2.308339
10.466580
13.030939
19.164374
54.250588

	vif.x. <dbl>
pregnant	1.426868
diastolic	1.178746
insulin	1.241062
age	1.506947
bmi	1.246419
triceps	1.458469

6 rows

Figure 2. Assessment of Multicollinearity by Anova between the initial model and a null model and pairwise correlation plot.

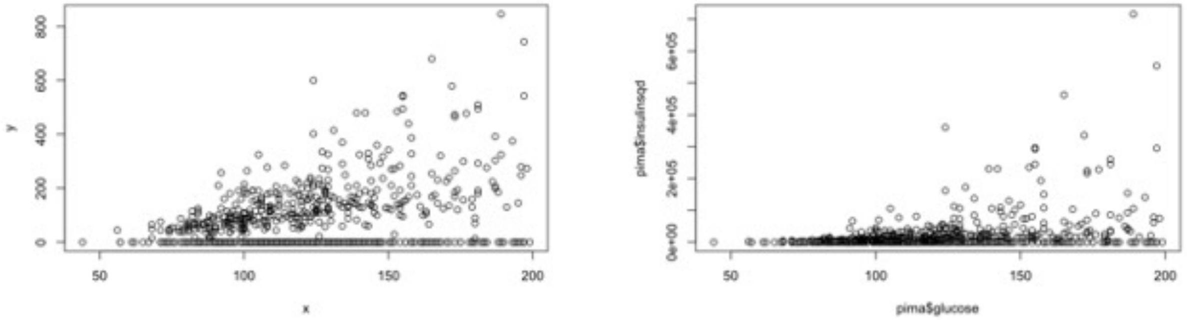
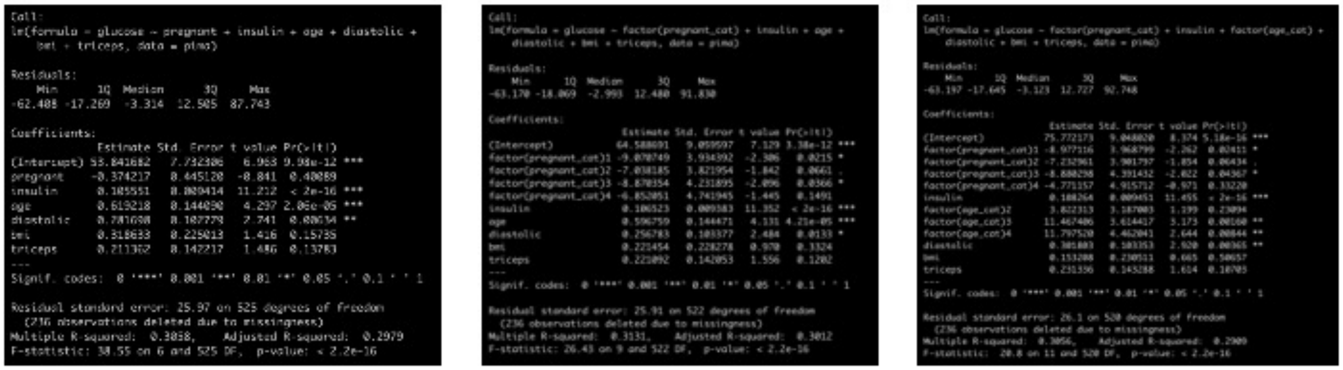


Figure 3. Left is a lot of the untransformed insulin variable (y-axis) and glucose (x-axis). Right is a graph of the transformed insulin variable to insulin^2 (y-axis) and glucose (x-axis).



```
Call:
lm(formula = glucose ~ pregnant + insulin + age + diastolic +
    bmi + triceps, data = pima)

Residuals:
    Min      1Q  Median      3Q     Max 
-62.408 -17.269  -3.314  12.505  87.743 

Coefficients:
             Estimate Std. Error t value Pr(>|t|)    
(Intercept) 53.841682   7.732306   6.963 9.98e-12 ***
pregnant    -0.374217   0.445120  -0.841  0.40089    
insulin      0.105551   0.009414  11.212  < 2e-16 ***
age          0.619218   0.144090   4.297 2.06e-05 ***
diastolic    0.281698   0.102779   2.741  0.00634 ** 
bmi          0.318633   0.225013   1.416  0.15735    
triceps      0.211362   0.142217   1.486  0.13783    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.97 on 525 degrees of freedom
  (236 observations deleted due to missingness)
Multiple R-squared:  0.3058,	Adjusted R-squared:  0.2979 
F-statistic: 38.55 on 6 and 525 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = glucose ~ factor(pregnant_cat) + insulin + age + 
    diastolic + bmi + triceps, data = pima)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.170 -18.069  -2.993  12.480  91.830 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           64.588691   9.059597   7.129 3.38e-12 ***
factor(pregnant_cat)1 -9.070749   3.934392  -2.306   0.0215 *  
factor(pregnant_cat)2 -7.038185   3.821954  -1.842   0.0661 .  
factor(pregnant_cat)3 -8.870354   4.231895  -2.096   0.0366 *  
factor(pregnant_cat)4 -6.852051   4.741945  -1.445   0.1491    
insulin                0.106523   0.009383  11.352  < 2e-16 ***
age                    0.596759   0.144471   4.131 4.21e-05 ***
diastolic              0.256783   0.103377   2.484   0.0133 *  
bmi                    0.221454   0.228278   0.970   0.3324    
triceps                0.221092   0.142053   1.556   0.1202    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 25.91 on 522 degrees of freedom
  (236 observations deleted due to missingness)
Multiple R-squared:  0.3131,	Adjusted R-squared:  0.3012 
F-statistic: 26.43 on 9 and 522 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = glucose ~ factor(pregnant_cat) + insulin + factor(age_cat) + 
    diastolic + bmi + triceps, data = pima)

Residuals:
    Min      1Q  Median      3Q     Max 
-63.197 -17.645  -3.123  12.727  92.748 

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)    
(Intercept)           75.772173   9.048020   8.374 5.18e-16 ***
factor(pregnant_cat)1 -8.977116   3.968799  -2.262  0.02411 *  
factor(pregnant_cat)2 -7.232961   3.901797  -1.854  0.06434 .  
factor(pregnant_cat)3 -8.880298   4.391432  -2.022  0.04367 *  
factor(pregnant_cat)4 -4.771157   4.915712  -0.971  0.33220    
insulin                0.108264   0.009451  11.455  < 2e-16 ***
factor(age_cat)2       3.822313   3.187003   1.199  0.23094    
factor(age_cat)3      11.467406   3.614417   3.173  0.00160 ** 
factor(age_cat)4      11.797520   4.462041   2.644  0.00844 ** 
diastolic              0.301803   0.103353   2.920  0.00365 ** 
bmi                    0.153208   0.230511   0.665  0.50657    
triceps                0.231336   0.143288   1.614  0.10705    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 26.1 on 520 degrees of freedom
  (236 observations deleted due to missingness)
Multiple R-squared:  0.3056,	Adjusted R-squared:  0.2909 
F-statistic:  20.8 on 11 and 520 DF,  p-value: < 2.2e-16
```

Figure 4. Running initial model with different transformations of variables.

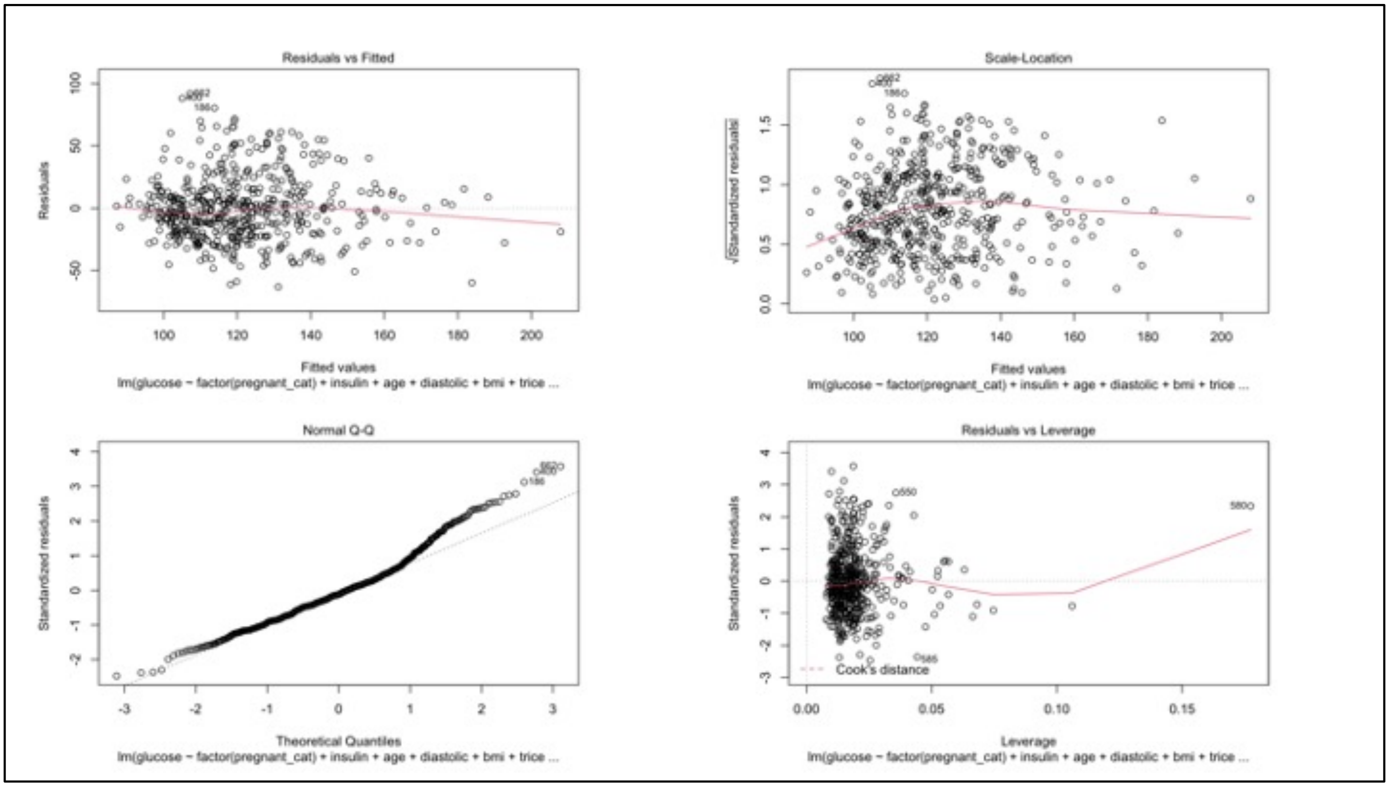


Figure 5. Model diagnostic plots of the initial model before Box Cox and before outlier identification

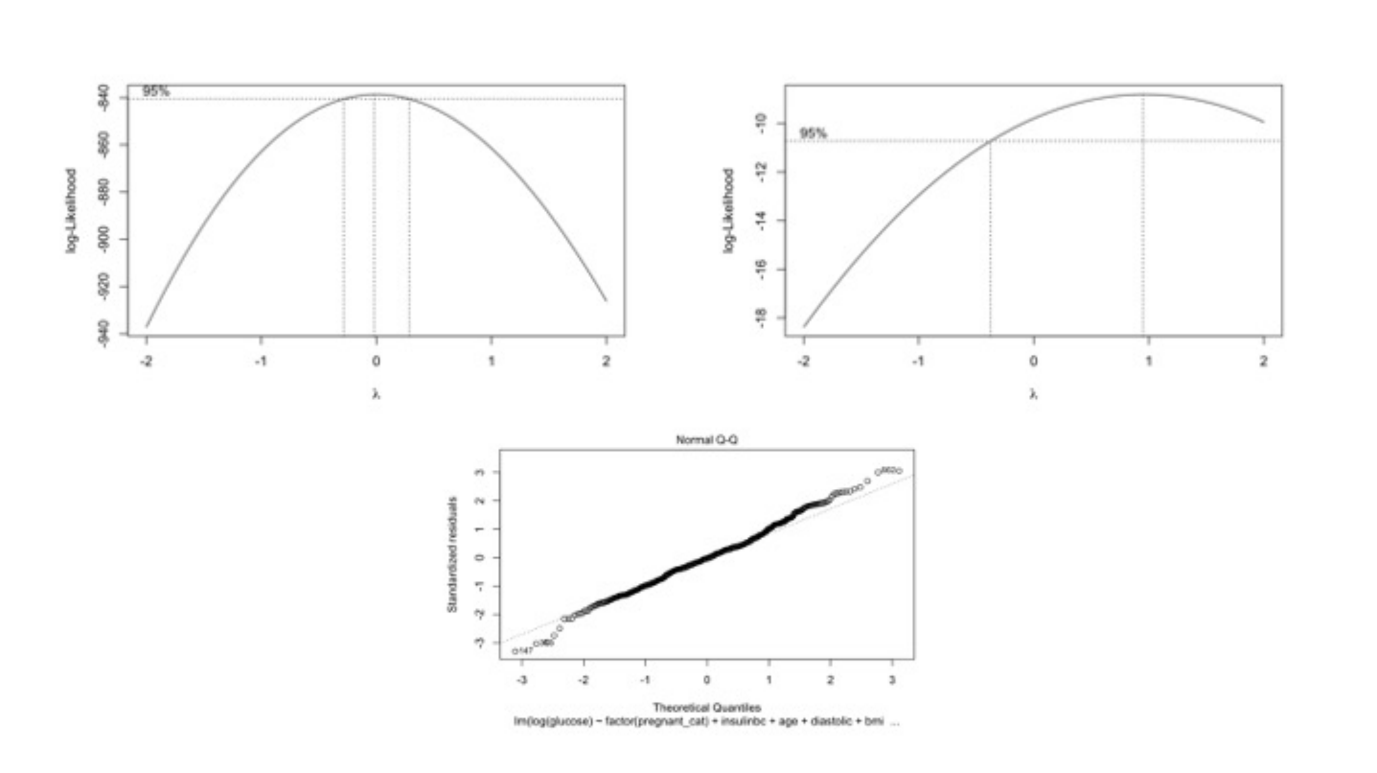
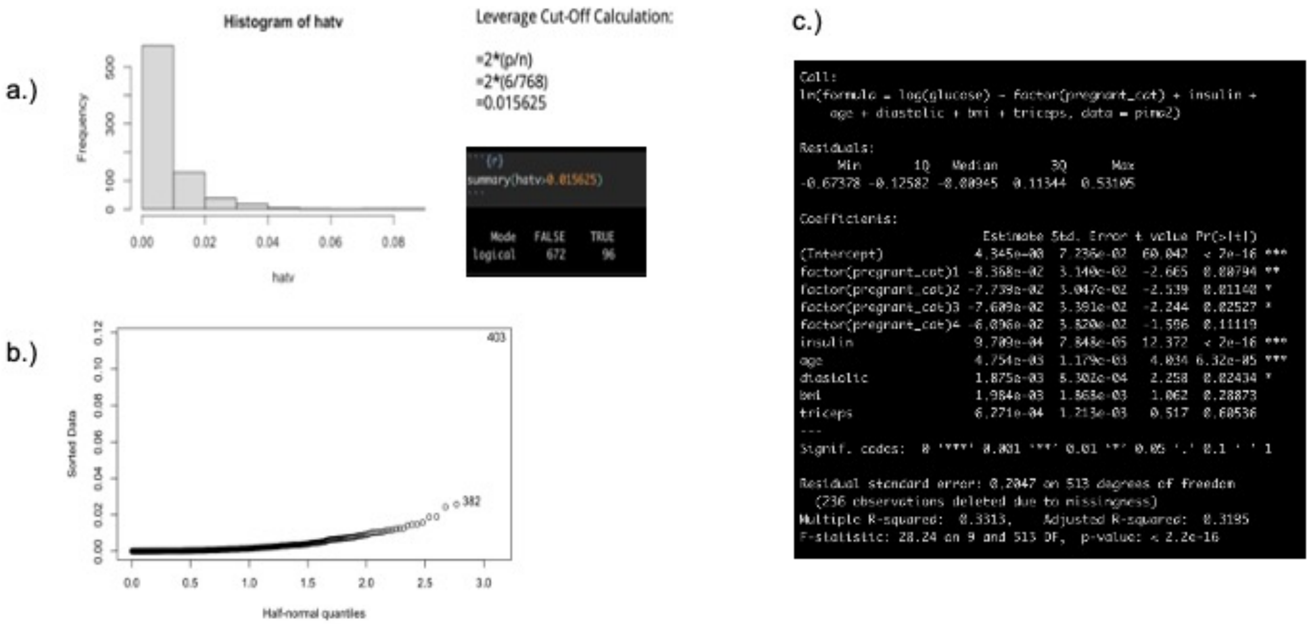


Figure 6. Box Cox Transformation of the initial model. Left is the initial Box Cox. Right is the Box Cox of the transformed model to determine if any further transformation was necessary. Bottom shows the Q-Q plot for the model after the transformation was applied.



Leverage Cut-Off Calculation:

=2*(p/n)
=2*(6/768)
=0.015625

~~~
```{r}
summary(hatv>0.015625)
```

   Mode   FALSE    TRUE
logical     672      96
~~~

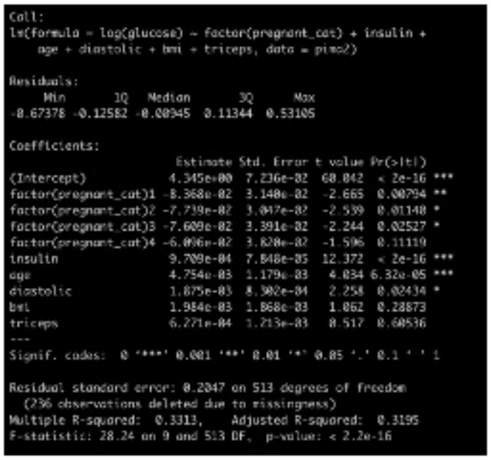
```
Call:
lm(formula = log(glucose) ~ factor(pregnant_cat) + insulin +
    age + diastolic + bmi + triceps, data = pima2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67378 -0.12582 -0.00945  0.11344  0.53105

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.345e+00  7.236e-02  60.042  < 2e-16 ***
factor(pregnant_cat)1 -8.368e-02  3.140e-02  -2.665  0.00794 **
factor(pregnant_cat)2 -7.739e-02  3.047e-02  -2.539  0.01140 *
factor(pregnant_cat)3 -7.609e-02  3.391e-02  -2.244  0.02527 *
factor(pregnant_cat)4 -6.096e-02  3.820e-02  -1.596  0.11119
insulin               9.709e-04  7.848e-05  12.372  < 2e-16 ***
age                   4.754e-03  1.179e-03   4.034 6.32e-05 ***
diastolic             1.875e-03  8.302e-04   2.258  0.02434 *
bmi                   1.984e-03  1.868e-03   1.062  0.28873
triceps               6.271e-04  1.213e-03   0.517  0.60536
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2047 on 513 degrees of freedom
  (236 observations deleted due to missingness)
Multiple R-squared:  0.3313,	Adjusted R-squared:  0.3195
F-statistic: 28.24 on 9 and 513 DF,  p-value: < 2.2e-16
```

Figure 7. A.) Calculation of leverage cut off values, and number of values that exceeded the limit in the untransformed data set. B.) The Half-norm plot of the initial model. C.) The model summary statistics after the outliers were removed.



```
Call:
lm(formula = log(glucose) ~ factor(pregnant_cat) + insulin +
    age + diastolic + bmi + triceps, data = pima2)

Residuals:
     Min       1Q   Median       3Q      Max
-0.67378 -0.12582 -0.00945  0.11344  0.53105

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           4.345e+00  7.236e-02  60.042  < 2e-16 ***
factor(pregnant_cat)1 -8.368e-02  3.140e-02  -2.665  0.00794 **
factor(pregnant_cat)2 -7.739e-02  3.047e-02  -2.539  0.01140 *
factor(pregnant_cat)3 -7.609e-02  3.391e-02  -2.244  0.02527 *
factor(pregnant_cat)4 -6.096e-02  3.820e-02  -1.596  0.11119
insulin               9.709e-04  7.848e-05  12.372  < 2e-16 ***
age                   4.754e-03  1.179e-03   4.034 6.32e-05 ***
diastolic             1.875e-03  8.302e-04   2.258  0.02434 *
bmi                   1.984e-03  1.868e-03   1.062  0.28873
triceps               6.271e-04  1.213e-03   0.517  0.60536
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2047 on 513 degrees of freedom
  (236 observations deleted due to missingness)
Multiple R-squared:  0.3313,    Adjusted R-squared:  0.3195
F-statistic: 28.24 on 9 and 513 DF,  p-value: < 2.2e-16
```

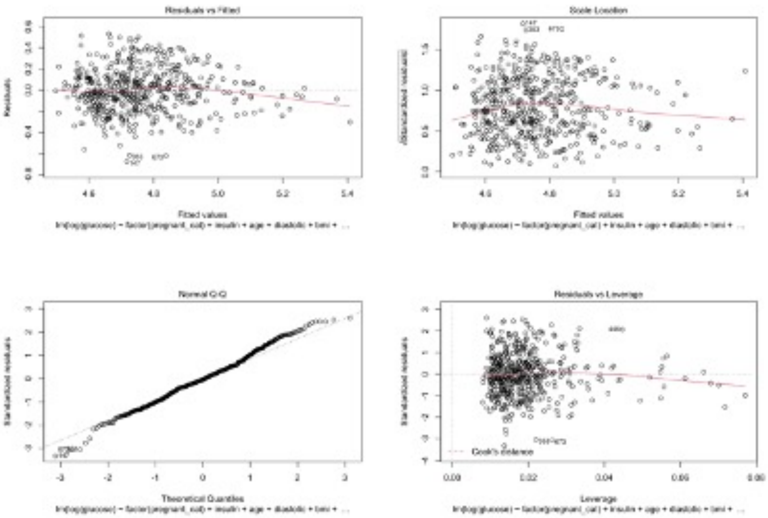


Figure 8. Summary statistics and model diagnostic plots of the final model.