PREDICTING GLUCOSE LEVEL BASED ON DIABETES RELATED VARIABLES

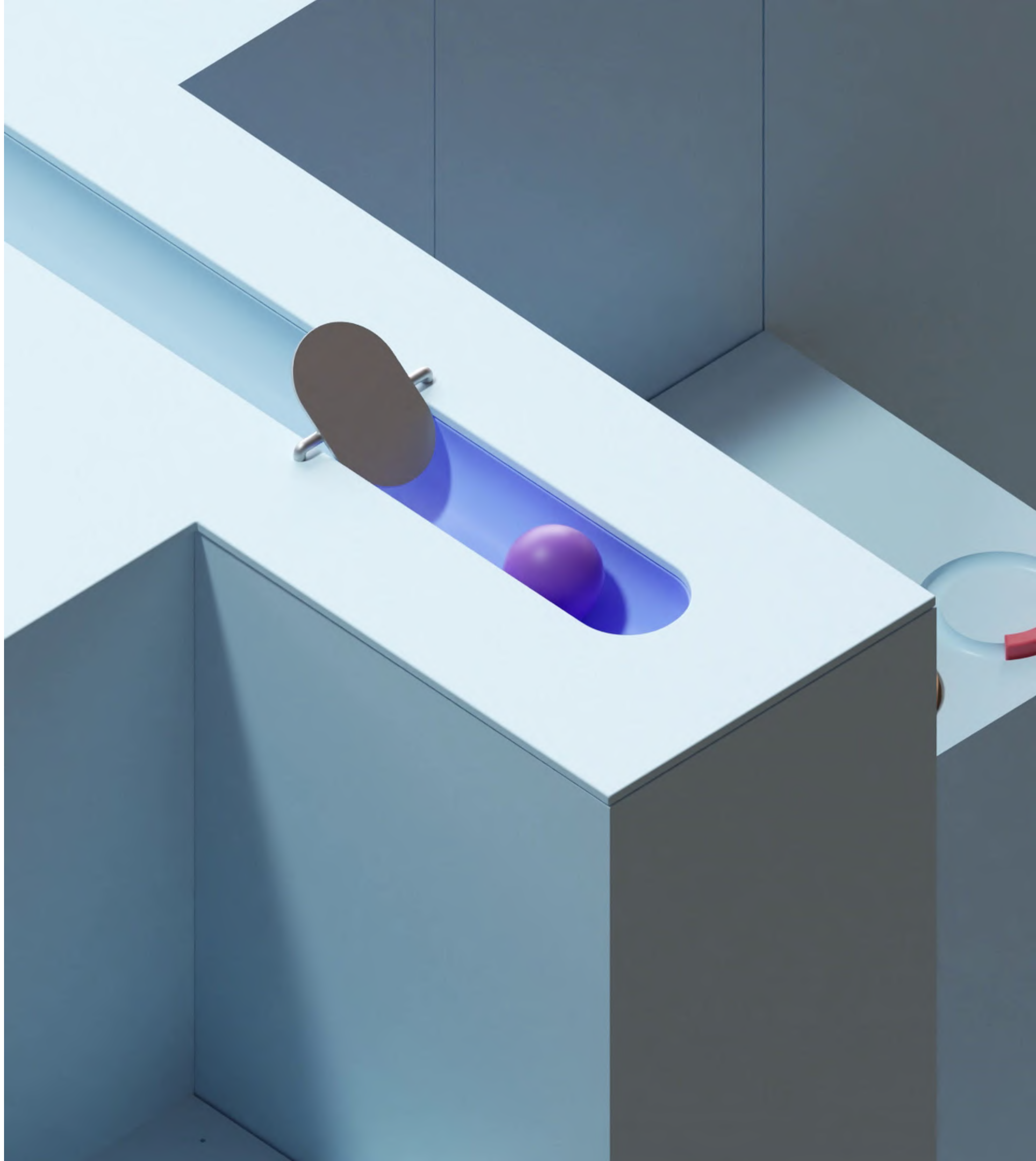
Aldo Untoro Edward Halim Ernest Salim Pierre Tanaya

WHY?

The growth of people diagnosed with diabetes tends to grow every year and the data provided reveals that glucose level is the key factor in explaining these occurrences.

Thus, we're trying to observe all possible factors and create a prediction model that has the capability to describe the relationship between glucose level and other biological traits

	Glucose	BloodPressure	Insulin	BMI	DiabetesPedigreeFunction	Age
1	89	66	94	28.1	0.167	21
2	137	40	168	43.1	2.288	33
3	78	50	88	31.0	0.248	26
4	197	70	543	30.5	0.158	53
5	189	60	846	30.1	0.398	59
6	166	72	175	25.8	0.587	51
7	118	84	230	45.8	0.551	31
8	103	30	83	43.3	0.183	33
9	115	70	96	34.6	0.529	32
10	126	88	235	39.3	0.704	27



Methods

Variables

- Glucose Level (Response Variable)
- Body Mass Index (BMI), Insulin, Blood Pressure, Age, and Diabetes Pedigree Function (Explanatory Variables)

Equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + e$$

$Y = \textit{Glucose Level}$	β_0 = Slope Intercept
$X_1 = \textit{BMI}$	β_1 = BMI Slope
$X_2 = \textit{Insulin}$	β_2 = Insulin Slope
$X_3 = \textit{Blood Pressure}$	β_3 = Blood Pressure Slope
$X_4 = \textit{Age}$	β_4 = Age Slope
$X_5 = \textit{Diabetes Pedigree Function}$	β_5 = Diabetes Pedigree Function Slope

- **Adjusted R2**: 0.3931
- Test Statistic (measure of significance based on p-value of T-test)
  - Insulin : 2 e-16
  - BMI : 0.2883
  - BloodPressure : 0.0469
  - DiabetesPedigreeFunction : 0.2377
  - Age : 3.6210 e-6

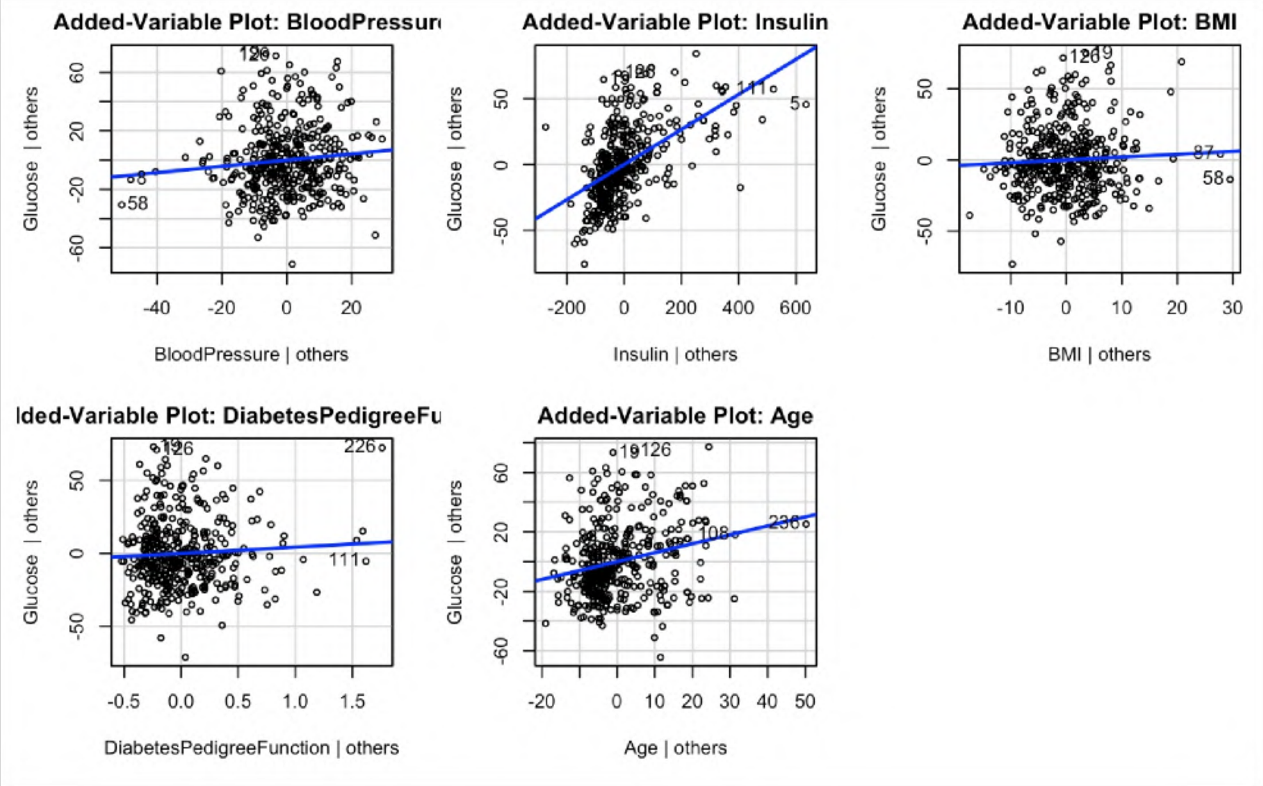
Based on T-statistic, Insulin, Blood Pressure, and Age are statistically significant, while BMI and Diabetes Pedigree Function are not.

Model Summary

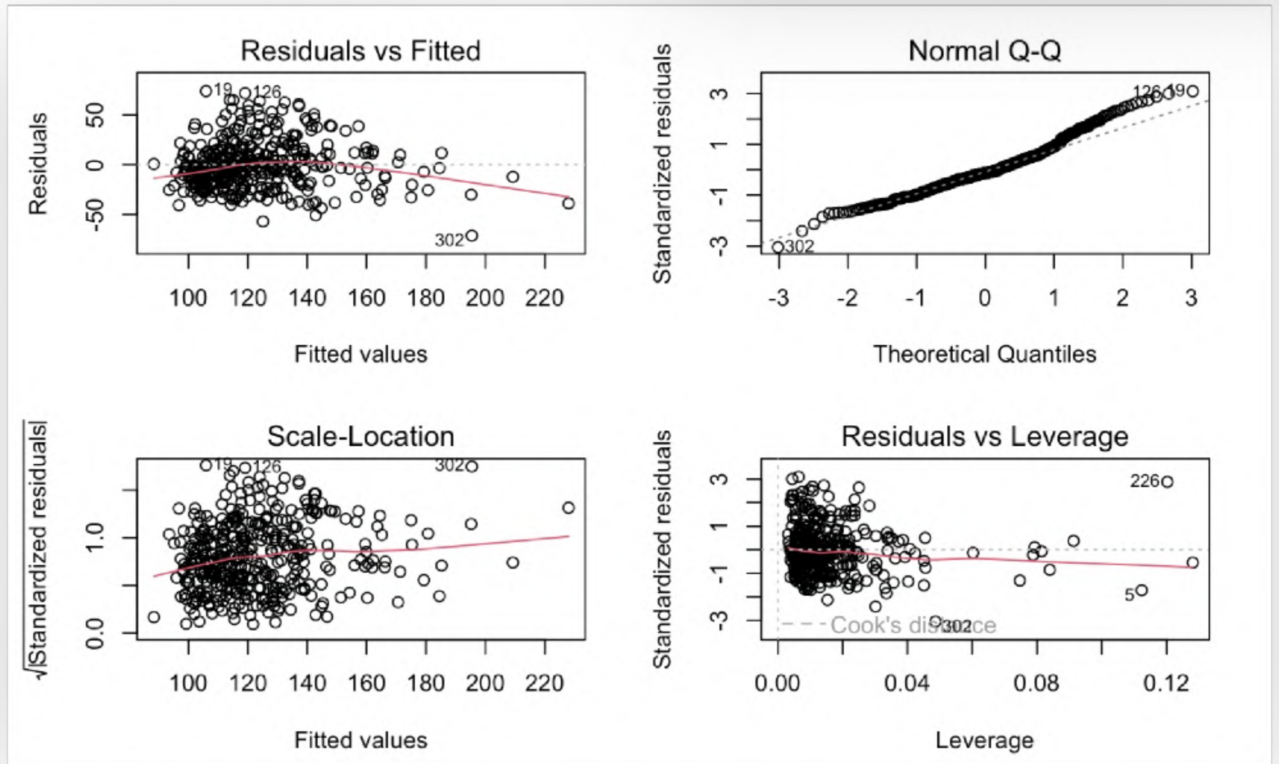
```
Call:
lm(formula = Glucose ~ Insulin + BMI + BloodPressure + DiabetesPedigreeFunction +
    Age)

Residuals:
    Min      1Q  Median      3Q     Max
-71.422 -15.736  -3.137  12.078  74.077

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)              59.28180    8.23946   7.195 3.28e-12 ***
Insulin                   0.13313    0.01078  12.345  < 2e-16 ***
BMI                       0.20022    0.18830   1.063   0.2883
BloodPressure             0.21404    0.10736   1.994   0.0469 *
DiabetesPedigreeFunction  4.26284    3.60479   1.183   0.2377
Age                       0.60237    0.12816   4.700 3.62e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 24.04 on 386 degrees of freedom
Multiple R-squared:  0.4008,    Adjusted R-squared:  0.3931
F-statistic: 51.65 on 5 and 386 DF,  p-value: < 2.2e-16
```

Anova Table

```
Analysis of Variance Table

Response: Glucose
                          Df Sum Sq Mean Sq  F value    Pr(>F)
Insulin                    1 125799  125799 217.6333 < 2.2e-16 ***
BMI                        1   2384    2384   4.1237 0.0429710 *
BloodPressure              1   6900    6900  11.9374 0.0006113 ***
DiabetesPedigreeFunction   1   1412    1412   2.4431 0.1188589
Age                        1  12769   12769  22.0906 3.624e-06 ***
Residuals                386 223120     578
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



Diagnostic Tools



- There exists a quadratic relationship between the predictors and the response variable.
- Observations are not distributed normally in the Q-Q plot.
- There are bad leverages: cases 226 and 302.
- Variance is constant.

Transformation

Model Transformation

```
bcPower Transformations to Multinormality
                         Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
Insulin                     0.0402           0      -0.0677       0.1480
BMI                         0.1524           0      -0.2081       0.5129
BloodPressure               1.3039           1       0.9377       1.6701
DiabetesPedigreeFunction    0.0572           0      -0.0794       0.1937
Age                        -1.6341          -2      -2.0239      -1.2443

Likelihood ratio test that transformation parameters are equal to 0
 (all log transformations)

Likelihood ratio test that no transformations are needed
```

lambda <dbl>	RSS <dbl>
-0.5073371	87456.67
-1.0000000	87922.92
0.0000000	87920.84
1.0000000	91188.64

4 rows

Based on the output above, we can use **log transformation** for the response variable. On the other hand, in terms of explanatory variables, we can implement **log transformation** for Insulin, BMI, and Diabetes Pedigree Function, and **power transformation of -2** for age..

Summary and ANOVA table for the transformed model

```
Call:
lm(formula = tGlucose ~ tInsulin + tBMI + tBloodPressure + tDiabetesPedigreeFunction +
    tAge)

Residuals:
     Min       1Q   Median       3Q      Max
-0.48799 -0.11932 -0.01062  0.11483  0.84733

Coefficients:
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)               3.742e+00  1.691e-01  22.127  < 2e-16 ***
tInsulin                  2.068e-01  1.449e-02  14.271  < 2e-16 ***
tBMI                      5.866e-03  4.854e-02   0.121    0.904
tBloodPressure            1.761e-03  8.273e-04   2.128    0.034 *
tDiabetesPedigreeFunction 1.173e-02  1.533e-02   0.765    0.445
tAge                     -7.108e+01  1.675e+01  -4.243 2.77e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1866 on 386 degrees of freedom
Multiple R-squared:  0.4491,    Adjusted R-squared:  0.442
F-statistic: 62.94 on 5 and 386 DF,  p-value: < 2.2e-16
```

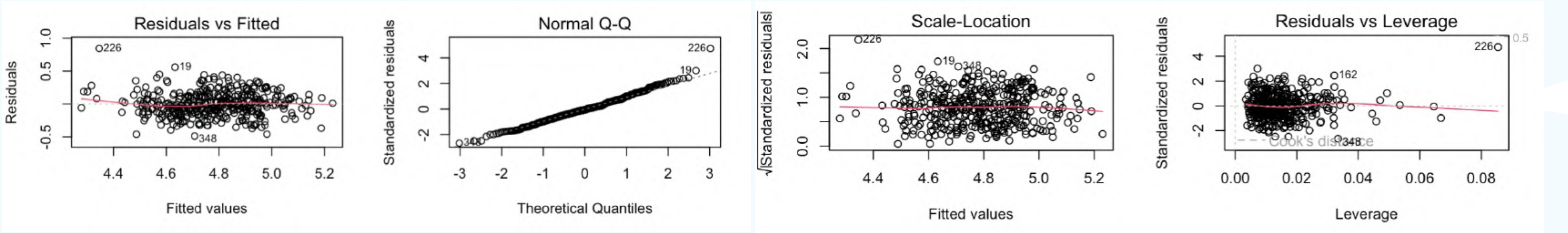
```
Analysis of Variance Table

Response: tGlucose
                           Df  Sum Sq Mean Sq  F value    Pr(>F)
tInsulin                    1  9.8231  9.8231 282.1808 < 2.2e-16 ***
tBMI                        1  0.0533  0.0533   1.5317 0.2166104
tBloodPressure              1  0.4057  0.4057  11.6535 0.0007087 ***
tDiabetesPedigreeFunction   1  0.0464  0.0464   1.3323 0.2491023
tAge                        1  0.6267  0.6267  18.0015 2.767e-05 ***
Residuals                 386 13.4371  0.0348
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Equation

$$log(Y) = \beta_0 + \beta_1 log(x_1) + \beta_2 log(x_2) + \beta_3 x_3 + \beta_4 log(x_4) + \beta_5 (x_5)^{-2} + e$$

- Adjusted R2 : 0.442. This means that roughly 44% is explained by the model.
- Test Statistic (measure of significance based on p-value of T-test)
  - log(Insulin) : 2e-16
  - log(BMI) : 0.904
  - BloodPressure : 0.034
  - log(DiabetesPedigreeFunction) : 0.445
  - (Age)-2 : 2.77 e-5

Based on T-statistic **Insulin, Blood Pressure, and Age** are significant, while **BMI and DiabetesPedigreeFunction** are not significant.



Insulin	BMI	BloodPressure	DiabetesPedigreeFunction	Age
1.110958	1.184590	1.217488	1.049184	1.156149

- There exists a linear relationship between predictors and response variable
- Normality exists since the points follow along the line in the Q-Q plot.
- Based on our observation, leverage points still exist, specifically, point 226.
- Constant variance exists since we can see in the scale location plot that there is no pattern.
- VIF value indicates that there's no multicolinearity.

Transformed Model

Variable Selection

we still have two predictors that are not significant, we can use variable selection to see whether we can remove the variable or not. In this project, we use all possible subset methods to find the best model formula. We implemented all possible subsets because it visits all models to find the highest adjusted R-squared, and lowest value of AICc, AIC, and BIC.

```
Subset selection object
5 Variables  (and intercept)
                          Forced in Forced out
tInsulin                      FALSE      FALSE
tBMI                          FALSE      FALSE
tBloodPressure                FALSE      FALSE
tDiabetesPedigreeFunction     FALSE      FALSE
tAge                          FALSE      FALSE
1 subsets of each size up to 5
Selection Algorithm: exhaustive
         tInsulin tBMI tBloodPressure tDiabetesPedigreeFunction tAge
1  ( 1 ) "*"      " "  " "            " "                       " "
2  ( 1 ) "*"      " "  " "            " "                       "*"
3  ( 1 ) "*"      " "  "*"            " "                       "*"
4  ( 1 ) "*"      " "  "*"            "*"                       "*"
5  ( 1 ) "*"      "*"  "*"            "*"                       "*"
```

	Size	Adj R2	AIC	AICc	BIC
[1,]	1	0.4011813	-1286.603	-1286.542	-1278.661
[2,]	2	0.4384887	-1310.826	-1310.620	-1298.912
[3,]	3	0.4439588	-1313.673	-1313.209	-1297.787
[4,]	4	0.4434073	-1312.296	-1311.427	-1292.439
[5,]	5	0.4419865	-1310.310	-1308.860	-1286.483

Based on the table above, the best adj R2 value, AIC, and AICc are founded in a model with p = 3, which has the formula,

$$log(Y) = \beta_0 + \beta_1 log(x_1) + \beta_3 x_3 + \beta_5 (x_5) + e$$

Based on the value of BIC, a model with p = 2 is considered the best model, which has the formula:

$$log(Y) = \beta_0 + \beta_1 log(x_1) + \beta_5 (x_5)^{-2} + e$$

In combination with results from R-summaries, the model with p = 3 is considered to be the better model than the model with p = 2. Then, we can conclude that the model **p=3** as the final model for this project,

$$log(Y) = \beta_0 + \beta_1 log(x_1) + \beta_3 x_3 + \beta_5 (x_5)^{-2} + e$$

Final Model

Based on our summary table above, we can conclude that:

- Adjusted R2 : 0.444. This means that 44% can be explained by the model.
- Test Statistic (a measure of significance based on p-value of T-test)
  - log(Insulin) : 2 e-16
  - BloodPressure : 0.0286
  - (Age)-2 : 1.81 e-5

Final Model Summary and ANOVA Table

```
Call:
lm(formula = tGlucose ~ tInsulin + tBloodPressure + tAge)

Residuals:
     Min       1Q   Median       3Q      Max
-0.49159 -0.11770 -0.00977  0.11373  0.87452

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.748e+00  9.630e-02  38.925  < 2e-16 ***
tInsulin        2.082e-01  1.400e-02  14.869  < 2e-16 ***
tBloodPressure  1.746e-03  7.946e-04   2.197   0.0286 *
tAge           -7.229e+01  1.665e+01  -4.341 1.81e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.1862 on 388 degrees of freedom
Multiple R-squared:  0.4482,    Adjusted R-squared:  0.444
F-statistic: 105.1 on 3 and 388 DF,  p-value: < 2.2e-16


Analysis of Variance Table

Response: tGlucose
                Df  Sum Sq Mean Sq F value    Pr(>F)
tInsulin         1  9.8231  9.8231 283.182 < 2.2e-16 ***
tBloodPressure   1  0.4566  0.4566  13.162 0.0003239 ***
tAge             1  0.6536  0.6536  18.842 1.814e-05 ***
Residuals      388 13.4590  0.0347
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Conclusion

What did we learn from the results?

We can confidently conclude that **log(Insulin), Blood Pressure, and (Age)-2** are the appropriate predictor variables to estimate the **Glucose** level. This model is determined by several approaches and indicators; R-summary, ANOVA table, and diagnostic plots to test the fit and validity of each model (initial, transformed, and final).

We can safely conclude that **Insulin has the highest influence on the model**. The relationship between the variables can be interpreted as:

- **With every 1% increase in the Insulin level, there will be a 0.2082% increase in the Glucose level.**
- With every 1 unit increase in the Blood Pressure level, there will be a 0.0017% increase in the Glucose level.
- With every 1 unit increase in Age, there will be a 0.7229% decrease in the Glucose level.

This experiment can be concluded in the finalization of the model which implements

$$log(Glucose) = 3.748 + 0.2082 log(Insulin) + 0.001 BloodPressure - 0.7229(Age)^{-2} + e$$

As the best model in terms of goodness-of-fit and explaining the dataset with Adjusted R2 of 0.444. Thus, intuitively speaking, from the analysis that we have performed, it is true that these biological traits which are possessed by diabetic patients have a strong correlation with the amount of glucose in their body.

To tie our understanding to the real world scenario our model succesfully captures the correlation among those variables. **If a patient has more Insulin in their body, they will have relatively higher Glucose Level.** Based on our project we found certain limitations that out model cannot explain such as; the low R2. To overcome this limitation we need to access more robust data sets in order to improve our model.

Thank you