Diabetes prevalence in Pima Indian women

Sara Dabbagh

2024-05-07

R Markdown

Load in & preview data

```
library("MASS")
data(Pima.tr)
diabetes <- Pima.tr

head(diabetes)
```

```
##   npreg glu bp skin  bmi   ped age type
## 1     5  86 68   28 30.2 0.364  24   No
## 2     7 195 70   33 25.1 0.163  55  Yes
## 3     5  77 82   41 35.8 0.156  35   No
## 4     0 165 76   43 47.9 0.259  26   No
## 5     0 107 60   25 26.4 0.133  23   No
## 6     5  97 76   27 35.6 0.378  52  Yes
```

Part 1: Model exploration

In part 1, exploring Pima.tr dataset.

1a: first model

```
diabetes_model1 = lm(glu ~ npreg + bp + skin +
                       bmi + ped + age + type,
                     data = diabetes)
summary(diabetes_model1)
```

```
##
## Call:
## lm(formula = glu ~ npreg + bp + skin + bmi + ped + age + type,
```

```
##     data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.595 -17.396  -1.641  12.952  89.977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.96159   15.62700   4.413 1.70e-05 ***
## npreg       -0.84245    0.72494  -1.162   0.2466
## bp           0.31786    0.18792   1.691   0.0924 .
## skin         0.07046    0.22559   0.312   0.7551
## bmi          0.23301    0.43405   0.537   0.5920
## ped         -2.36903    6.64389  -0.357   0.7218
## age          0.55832    0.24166   2.310   0.0219 *
## typeYes     26.29928    4.64856   5.658 5.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.26 on 192 degrees of freedom
## Multiple R-squared:  0.2853, Adjusted R-squared:  0.2592
## F-statistic: 10.95 on 7 and 192 DF,  p-value: 1.312e-11
```

Interpretations

a. The fitted coefficient value for my quantitative predictor bmi (body mass index), is 0.23301, meaning that for every 1 unit increase in bmi, I'd expect the estimated glucose level to increase by 0.23301 on average, holding all other predictors constant.

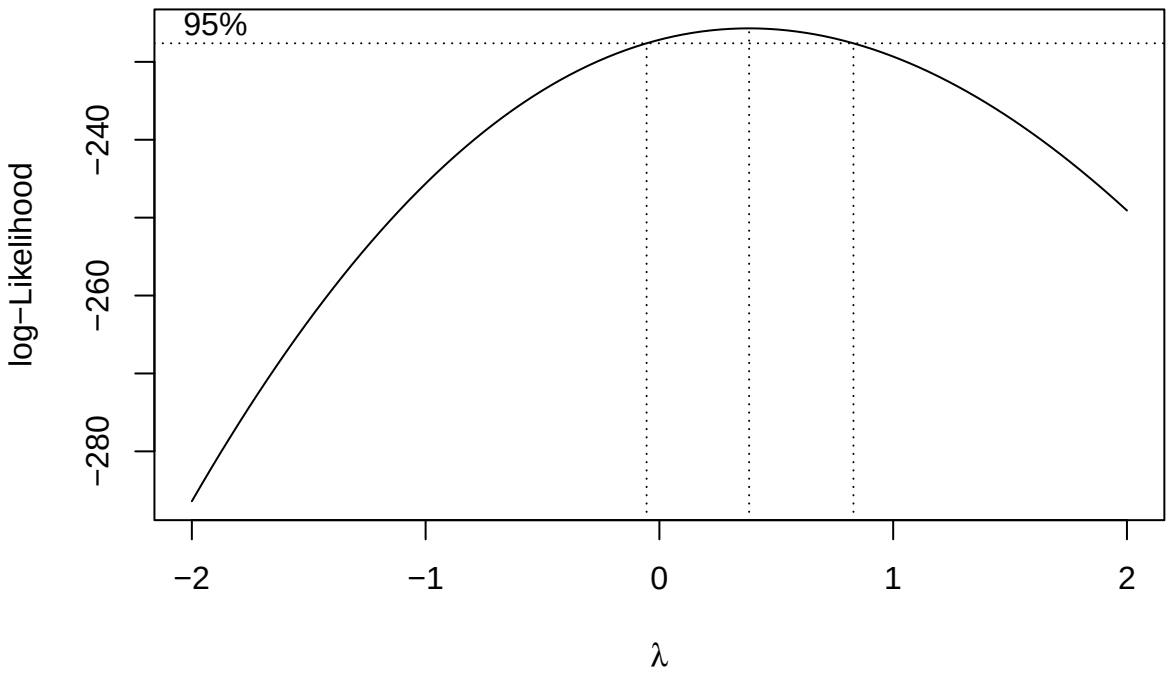
b. For my categorical predictor, type, (diabetes type), the fitted coefficients for its level "typeYes," has an estimated value of 26.29928. This means that those with diabetes (typeYes, type = 1) have an estimated glucose level of 26.29928 higher compared to those without diabetes (typeNo, type = 0), on average.

c. The contribution to the p-value from my categorical predictor "type" is given by the p-value associated with its coefficient in the summary output. Rhe p-value for "typeYes" is 5.51e-08, indicating that having diabetes significantly contributes to explaining the variation in glucose levels.

d. The baseline level for my categorical predictor "type" is "typeNo," since typeYes is included in my model but the other level (type = 0) is not present, thus it must be representative in my baseline level.

1b: Box-Cox transformation

```
boxcox(lm(glu ~ npreg + bp + bmi + skin + age + ped + type, data = diabetes),
       plotit = TRUE)
```




```
transformed_model = lm(log(glu) ~ npreg + bp + skin + bmi +
                         ped + age + type, data = diabetes)
summary(transformed_model)
```

```
##
## Call:
## lm(formula = log(glu) ~ npreg + bp + skin + bmi + ped + age +
##     type, data = diabetes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.71051 -0.12649  0.00669  0.12592  0.64370
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.338e+00  1.288e-01  33.673  < 2e-16 ***
## npreg       -5.792e-03  5.977e-03  -0.969   0.3337
## bp           2.682e-03  1.549e-03   1.731   0.0851 .
## skin         2.263e-05  1.860e-03   0.012   0.9903
## bmi          2.711e-03  3.579e-03   0.758   0.4496
## ped         -2.290e-02  5.478e-02  -0.418   0.6764
## age          4.029e-03  1.992e-03   2.022   0.0446 *
## typeYes      2.103e-01  3.833e-02   5.488 1.28e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 0.2247 on 192 degrees of freedom
## Multiple R-squared:  0.268,  Adjusted R-squared:  0.2413
## F-statistic: 10.04 on 7 and 192 DF,  p-value: 1.122e-10
```

```
summary(diabetes_model1)
```

```
##
## Call:
## lm(formula = glu ~ npreg + bp + skin + bmi + ped + age + type,
##     data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -66.595 -17.396  -1.641  12.952  89.977
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 68.96159   15.62700   4.413 1.70e-05 ***
## npreg       -0.84245    0.72494  -1.162   0.2466
## bp           0.31786    0.18792   1.691   0.0924 .
## skin         0.07046    0.22559   0.312   0.7551
## bmi          0.23301    0.43405   0.537   0.5920
## ped         -2.36903    6.64389  -0.357   0.7218
## age          0.55832    0.24166   2.310   0.0219 *
## typeYes     26.29928    4.64856   5.658 5.51e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.26 on 192 degrees of freedom
## Multiple R-squared:  0.2853, Adjusted R-squared:  0.2592
## F-statistic: 10.95 on 7 and 192 DF,  p-value: 1.312e-11
```

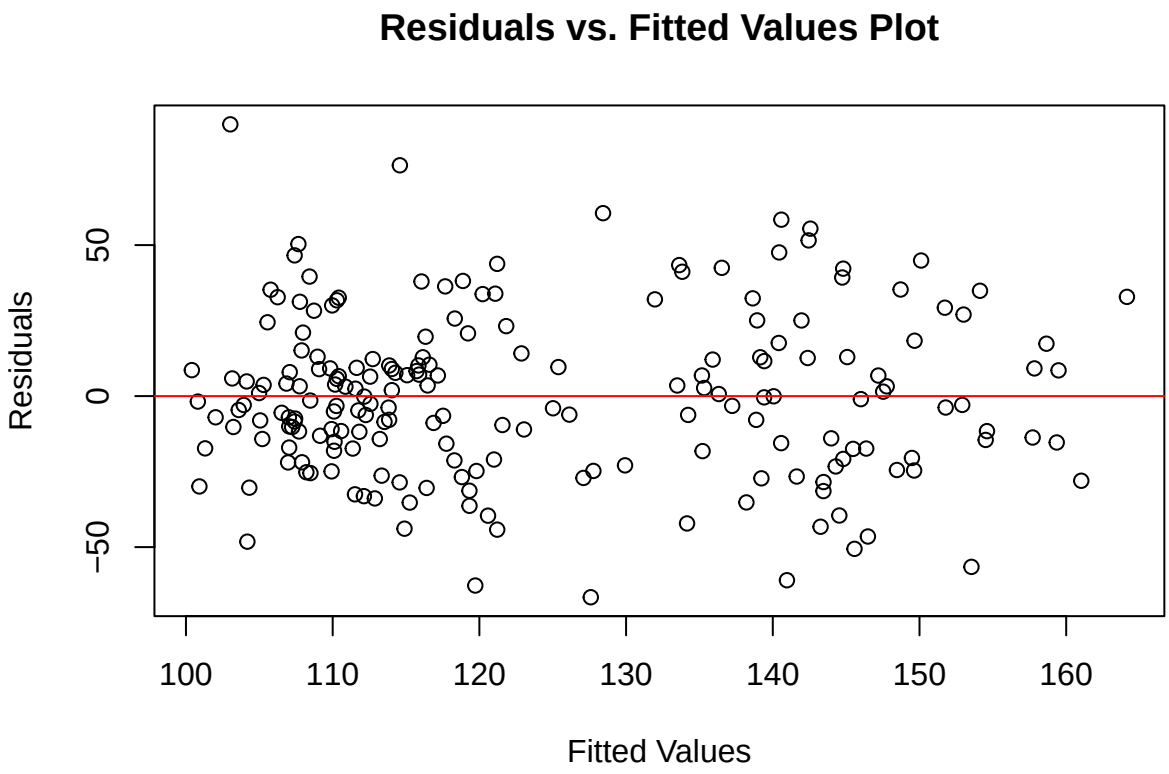
```
summary(diabetes_model1)$sigma
```

```
## [1] 27.25572
```

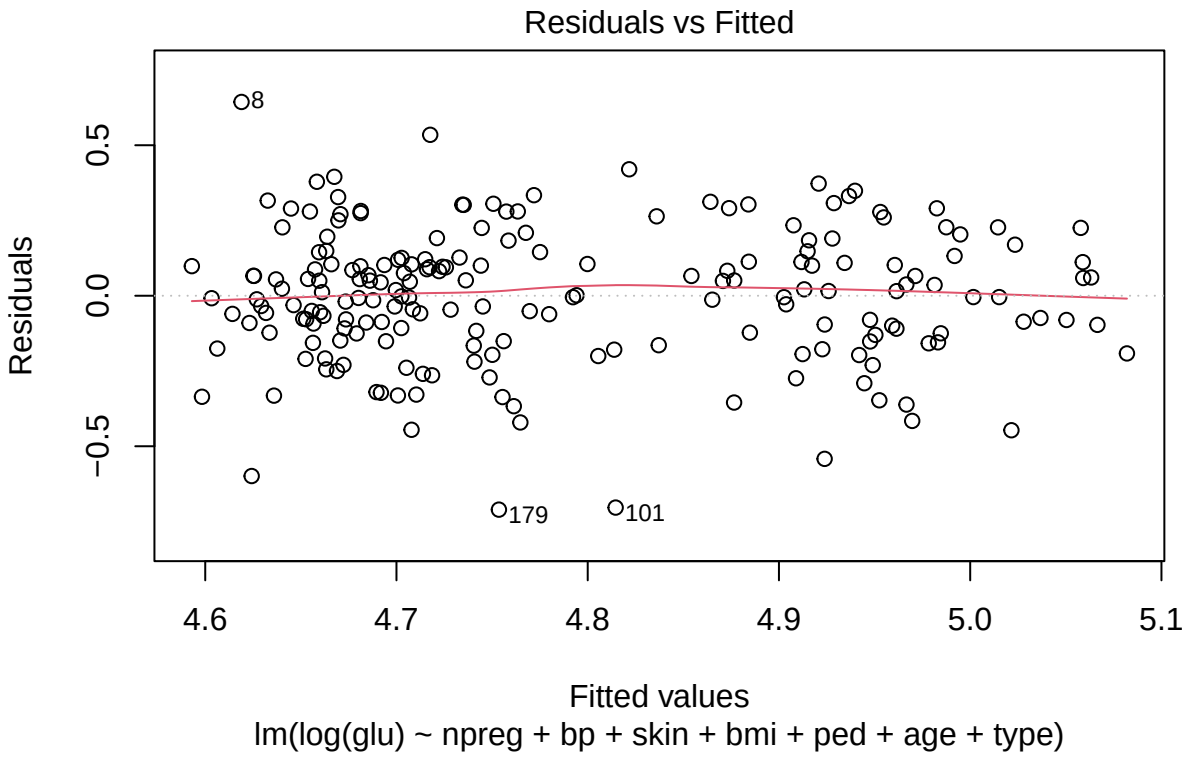
```
summary(transformed_model)$sigma
```

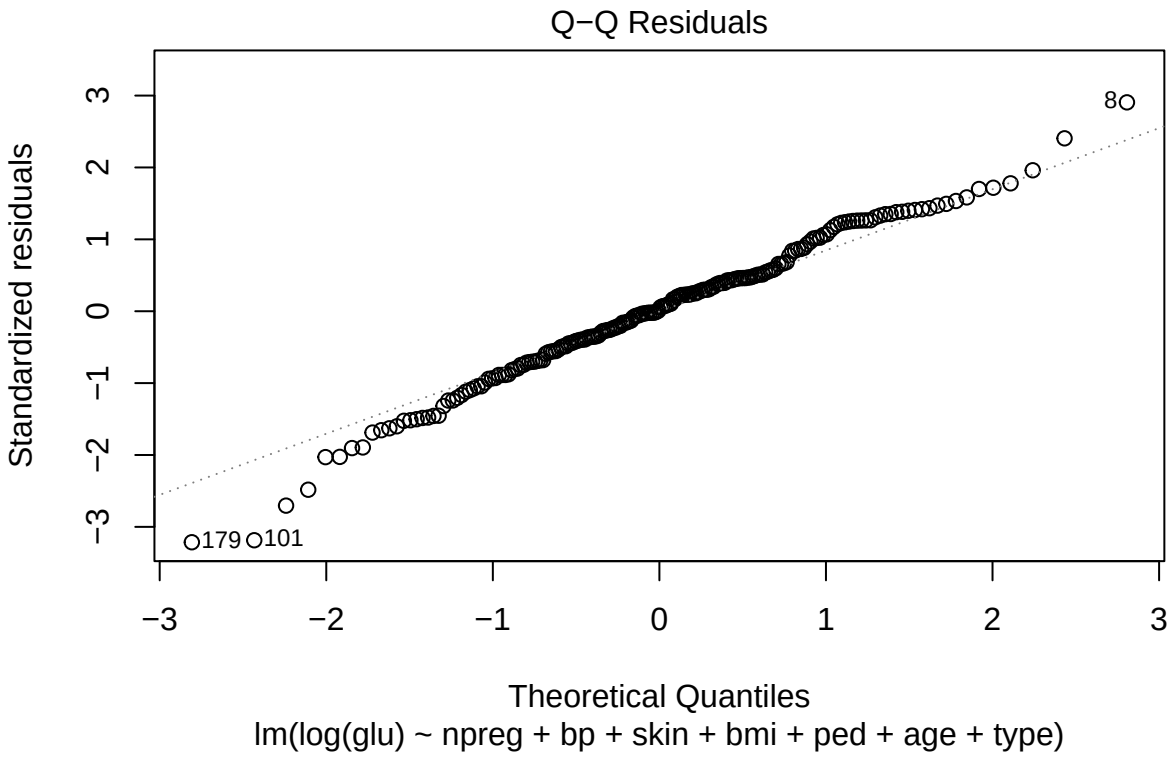
```
## [1] 0.2247124
```

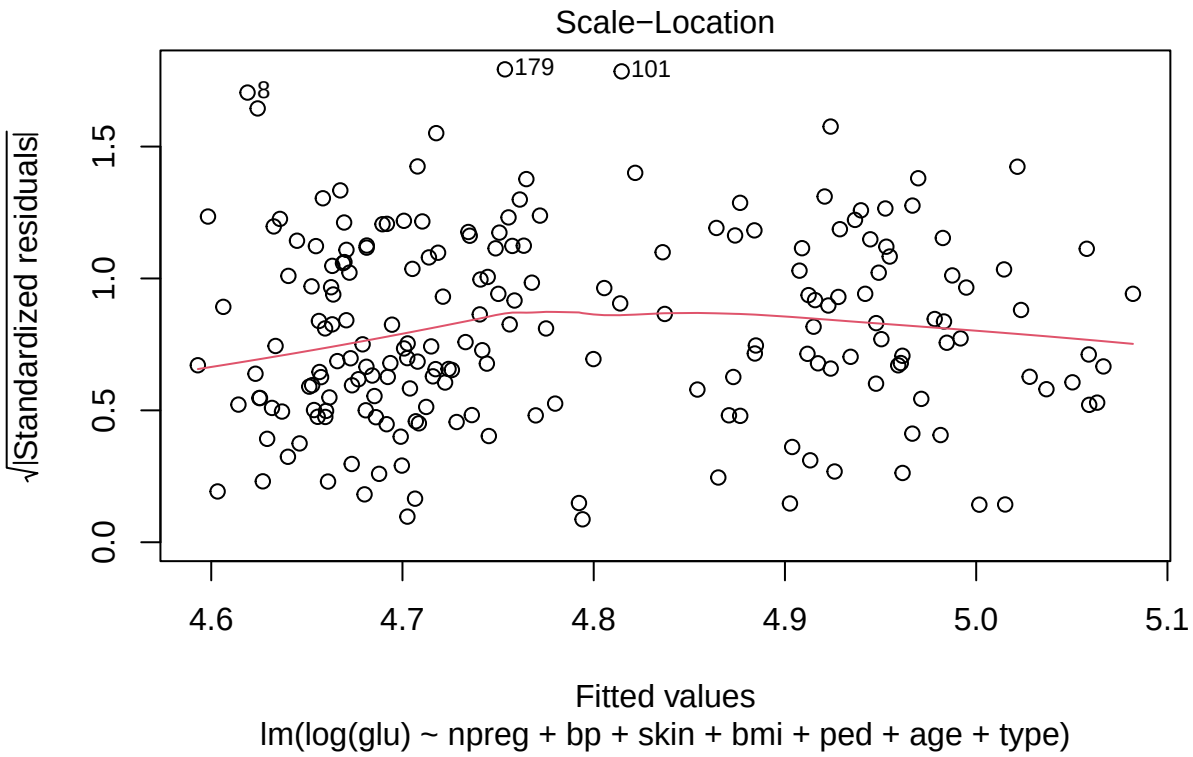
```
plot(fitted(diabetes_model1), resid(diabetes_model1),
      xlab = "Fitted Values", ylab = "Residuals",
     main = "Residuals vs. Fitted Values Plot")
abline(h = 0, col = "red")
```

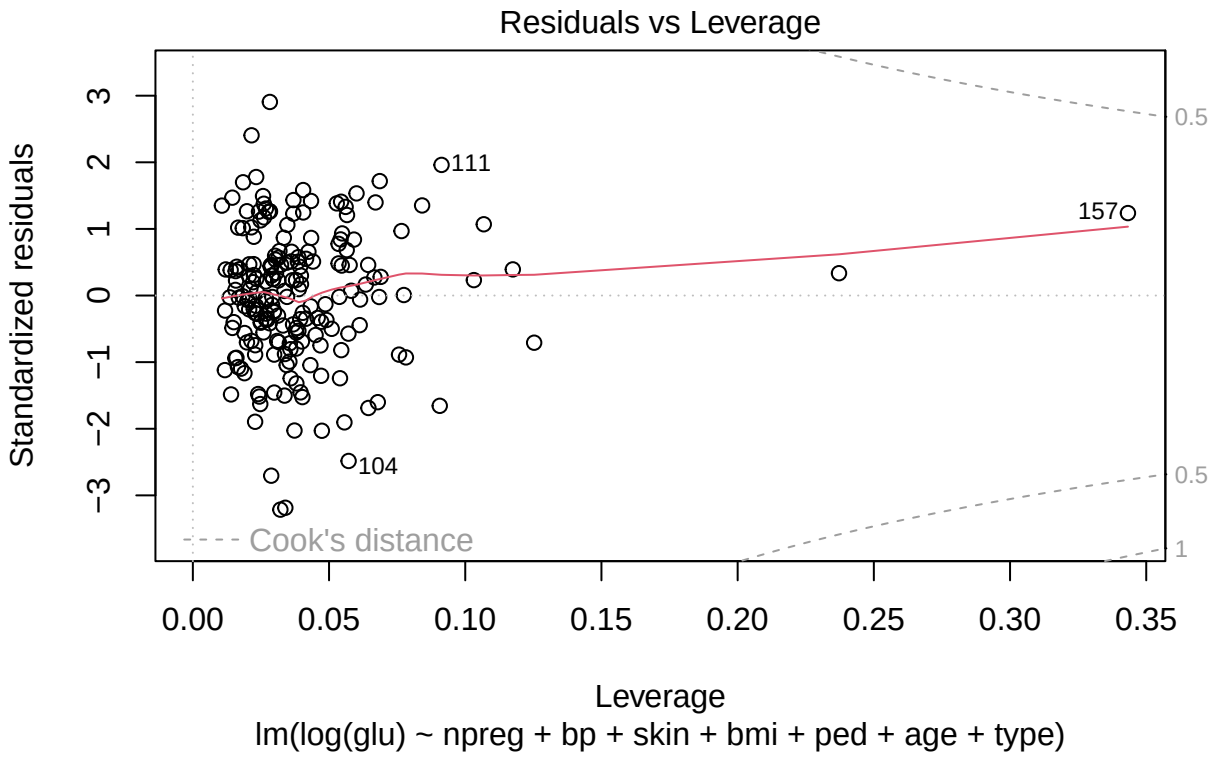



```
plot(transformed_model)
```


Residuals vs Fitted
Residuals
0.5
0.0
−0.5
8
179
101
4.6
4.7
4.8
4.9
5.0
5.1
Fitted values
lm(log(glu) ~ npreg + bp + skin + bmi + ped + age + type)


Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(log(glu) ~ npreg + bp + skin + bmi + ped + age + type)
179
101
8


Scale−Location
179
101
8
√|Standardized residuals|
0.0
0.5
1.0
1.5
4.6
4.7
4.8
4.9
5.0
5.1
Fitted values
lm(log(glu) ~ npreg + bp + skin + bmi + ped + age + type)



What does my Box-Cox plot tell me? Firstly, my Box-Cox plot tells me that my confidence interval for lambda falls between λ = -0.01 and $\lambda = 0.09$, with my optimal $\lambda = 0.04$. Based on this, a log transformation ($\lambda = 0$) is in this range and is reasonable.

Looking at my summary of the transformed model, I can see that my residual standard error for my transformed model is much lower, going from 27.25572 in my original model to 0.2247 in my newer model.This indicates that the transformed model fits my data to a much greater extent because the predictions it makes will be closer to the observed glucose values. The R^2 of both is very similar, meaning the variance explained by the linear relationship between glucose and my predictors for both is to the same extent (approximately 28%), therefore I might prefer to use my transformed model over my original from now on.

Looking at other plots for the transformed model Next, I want to look at some visuals to determine if I should continue with using my transformed model. Looking at my residuals vs. fitted plot for linearity, I see that my red line is not as straight at y=0 as the line in my residuals vs fitted plot for my original diabetes model, meaning that diabetes_model1 is a better fit in this case. Looking at the normality of residuals in my Q-Q plot, the points don't fall on the line towards the beginning and end of the graph. Thus, my errors aren't normally distributed. My scale-location plot tells me this model lacks homoscedasticity, seeing that I lack a parallel, horizontal lines around my values. My equal variance assumption is not met, and I'd prefer to stick to my original diabetes model for the rest of my model exploration, until I find a better fit.

1c: second model

What interaction term would I include? Conceptually speaking, when going about deciding on an interaction term to add to my fitted model, I want to analyze my relationships between variables. Adding

an interaction term can be helpful if one of my predictors is dependent on the value of another. I notice that my p-value for BMI is high, and when I look at the individual linear relationship between glucose and BMI (glu ~ bmi), I see that the p-value for bmi is statistically significant ($0.00205 > 0.05$) and greatly impacts my model when considered on its own. This tells me that I might need to introduce an interaction variable that accounts for a multicollinearity and the influence of any interaction effects when BMI is incorporated into my bigger model. Conceptually, out of all my predictor variables, I believe age plays the biggest role on BMI since older age affects the rate of metabolization within the body.

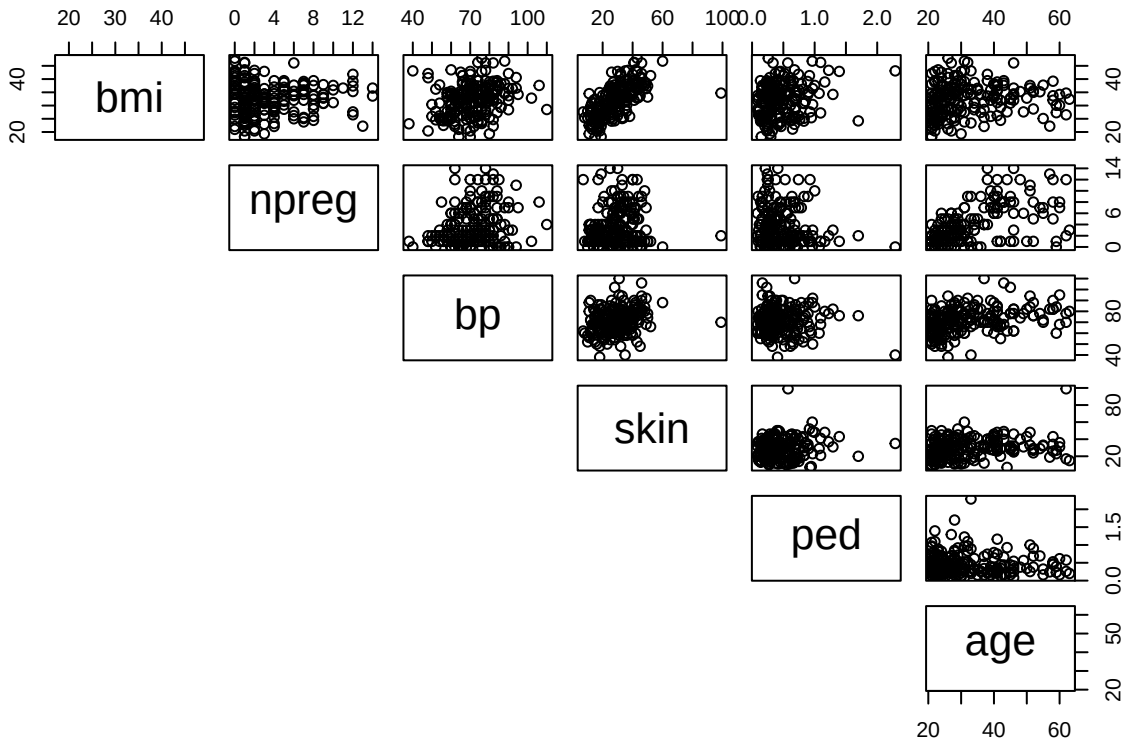
Plot diagnostics My matrix for bmi and age suggests that a linear relationship between these two variables might not explain any patterns or variance in my data; in other words, there is a weak linear relationship. Introducing an interaction term (bmi:age) could potentially help strengthen their relationship in my model; it may help capture any non linear / interactive effects that could be influencing my data..

Should I keep this interaction variable? Looking at my summary, I see that my p-value for bmi:age is 0.1367, demonstrating that the interaction between bmi and age in predicting glucose levels is not statistically significant. Because it is greater than my desired significance level of 0.05, there isn't sufficient evidence to reject the null hypothesis, indicating that the relationship between BMI and glucose levels does not significantly vary depending on age. However, when I conduct a backwards step on this model (metric being AIC), with the interaction variable, it is kept in my final model. My result for keeping this interaction variable is inconclusive, and I'd have to compare it with my previous models to further determine its relevance.

```
summary(lm(glu~bmi, data = diabetes))
```

```
##
## Call:
## lm(formula = glu ~ bmi, data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -67.519 -20.292  -5.239  19.483  79.104
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  87.7865    11.7849   7.449 2.84e-12 ***
## bmi           1.1199     0.3584   3.125  0.00205 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.99 on 198 degrees of freedom
## Multiple R-squared:  0.047,  Adjusted R-squared:  0.04218
## F-statistic: 9.764 on 1 and 198 DF,  p-value: 0.002046
```

```
pairs(bmi ~ npreg + bp +skin + bmi + ped + age, data = diabetes,
      lower.panel = NULL)
```




```
diabetes_model2 = lm(glu ~ npreg + bp + bmi + age + type + bmi:age,
                     data = diabetes)
summary(diabetes_model2)
```

```
##
## Call:
## lm(formula = glu ~ npreg + bp + bmi + age + type + bmi:age, data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.868 -18.378  -1.118  13.227  90.669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 18.36628   35.49741   0.517   0.6055
## npreg       -0.80171    0.71412  -1.123   0.2630
## bp           0.33205    0.18558   1.789   0.0751 .
## bmi          1.82065    1.07147   1.699   0.0909 .
## age          2.21471    1.12136   1.975   0.0497 *
## typeYes     26.38315    4.50890   5.851 2.06e-08 ***
## bmi:age     -0.05171    0.03460  -1.495   0.1367
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.05 on 193 degrees of freedom
```

```
## Multiple R-squared:  0.2926, Adjusted R-squared:  0.2706
## F-statistic: 13.31 on 6 and 193 DF,  p-value: 1.343e-12
```

```
step(diabetes_model2, direction = "backward")
```

```
## Start:  AIC=1325.88
## glu ~ npreg + bp + bmi + age + type + bmi:age
##
##           Df Sum of Sq    RSS    AIC
## - npreg    1     921.9 142090 1325.2
## <none>                  141168 1325.9
## - bmi:age  1    1633.9 142802 1326.2
## - bp       1    2341.7 143510 1327.2
## - type     1   25043.3 166212 1356.5
##
## Step:  AIC=1325.18
## glu ~ bp + bmi + age + type + bmi:age
##
##           Df Sum of Sq    RSS    AIC
## <none>                  142090 1325.2
## - bmi:age  1    1689.7 143780 1325.5
## - bp       1    2262.4 144353 1326.3
## - type     1   24456.8 166547 1354.9

##
## Call:
## lm(formula = glu ~ bp + bmi + age + type + bmi:age, data = diabetes)
##
## Coefficients:
## (Intercept)           bp          bmi          age      typeYes      bmi:age
##    19.14403      0.32625      1.86467      2.10214     25.99605     -0.05257
```

```
new_diabetes_model2 = lm(glu ~ bp + age +
                           type + bmi + bmi:age,
                         data = diabetes)
```

1d: model comparison

My models so far

Thus far, I have four models: my initial model predicting glu from all predictors, my transformed model predicting log(glu) from all predictors, my second model predicting glu from all predictors and the interaction term bmi:age, and my new second model attained from backwards selection (Aic as metric) that predicts glu from bp, type, and bmi*age.

Two models to look at

I want to further analyze my new second model from the previous part, comparing its R^1 to another model. However, R^2 is penalized by the increase in predictors for a model, so I want to proceed to find a model that'll offer a fair comparison in their strengths. I will fit a fifth model by using the same backwards stepping proccess on my original model. This will help me determine if the previous interaction term helps explain variance within my model.

```r
step(diabetes_model1, direction = "backward")
```

```
## Start:  AIC=1329.94
## glu ~ npreg + bp + skin + bmi + ped + age + type
##
##         Df Sum of Sq    RSS    AIC
## - skin   1      72.5 142704 1328.0
## - ped    1      94.5 142726 1328.1
## - bmi    1     214.1 142846 1328.2
## - npreg  1    1003.2 143635 1329.3
## <none>               142632 1329.9
## - bp     1    2125.4 144757 1330.9
## - age    1    3965.4 146597 1333.4
## - type   1   23777.5 166409 1358.8
##
## Step:  AIC=1328.04
## glu ~ npreg + bp + bmi + ped + age + type
##
##         Df Sum of Sq    RSS    AIC
## - ped    1      97.9 142802 1326.2
## - bmi    1     650.0 143354 1327.0
## - npreg  1    1034.9 143739 1327.5
## <none>               142704 1328.0
## - bp     1    2193.0 144897 1329.1
## - age    1    4280.3 146985 1332.0
## - type   1   23790.3 166495 1356.9
##
## Step:  AIC=1326.18
## glu ~ npreg + bp + bmi + age + type
##
##         Df Sum of Sq    RSS    AIC
## - bmi    1     588.3 143390 1325.0
## - npreg  1     977.7 143780 1325.5
## <none>               142802 1326.2
## - bp     1    2269.2 145071 1327.3
## - age    1    4367.0 147169 1330.2
## - type   1   24305.5 167108 1355.6
##
## Step:  AIC=1325
## glu ~ npreg + bp + age + type
##
##         Df Sum of Sq    RSS    AIC
## - npreg  1    1057.4 144448 1324.5
## <none>               143390 1325.0
## - bp     1    2846.0 146236 1326.9
## - age    1    4351.3 147742 1329.0
## - type   1   27887.2 171278 1358.5
##
## Step:  AIC=1324.47
## glu ~ bp + age + type
##
##        Df Sum of Sq    RSS    AIC
## <none>              144448 1324.5
```

```
## - bp    1    2781.5 147229 1326.3
## - age   1    3302.0 147750 1327.0
## - type  1   27279.2 171727 1357.1

##
## Call:
## lm(formula = glu ~ bp + age + type, data = diabetes)
##
## Coefficients:
## (Intercept)           bp          age      typeYes
##     75.9991       0.3549       0.4250      26.5690
```

```
new_diabetes_model1 = lm(glu ~ bp + age + type, data = diabetes)
summary(new_diabetes_model1)$r.squared
```

```
## [1] 0.2761677
```

```
summary(new_diabetes_model2)$r.squared
```

```
## [1] 0.2879822
```

```
anova(new_diabetes_model1, new_diabetes_model2)
```

```
## Analysis of Variance Table
##
## Model 1: glu ~ bp + age + type
## Model 2: glu ~ bp + age + type + bmi + bmi:age
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    196 144448
## 2    194 142090  2    2357.7 1.6095 0.2026
```

Results of comparison

Looking at R^2

Upon comparing my two new models, I see that their only difference is bmi and bmi * age. I want to test if my model with or without bmi and bmi * age is a better fit to my data. Looking at my R^2, it is a bit higher for my second model indicating that new_diabetes_model2 explains a larger proportion (28.79822%) of the variance in the response variable compared to new_diabetes_model1 (27.61677% explained). Therefore, I might prefer to use new_diabetes_model2 for modeling the linear relationship between the predictors and glucose levels. For further confirmation, I'll perform a statistical test.

Looking at my anova test

My hypotheses for this test:

H_0 : y = $\beta_0 + \beta_1 * bp + \beta_2 * age\ \beta_3 * type + \epsilon$ and $\beta_4 = \beta_5 = 0$

H_1 : y = $\beta_0 + \beta_1 * bp + \beta_2 * age\ \beta_3 * type + \beta_4 * bmi + \beta_5 * bmi : age + \epsilon$, *where at least* β_4, β_5 *are nonzero*

My test statistic: 1.6095;

My p-value: 0.2026;

Decision at 10% level: fail to reject null hypothesis given my high p-value of 0.2026, thus I would proceed with my smaller model 1. I do not have evidence to suggest that I should include bmi and bmi:age in my final model.

Conclusion of model exploration

I've analyzed several models thus far. Firstly, I looked at my initial model containing all predictors. This model proved to be pretty strong and explains the variance in my model to a similar extent as the rest of my models. Next, I considered log transforming my response variable. Though my residual standard errors decreased greatly, my plots told me that the underlying assumptions of the regression model were not met, such as linearity, homoscedasticity, and normality of residuals. Next, I introduced an interaction variable, did a backwards selection on the new model containing it, and ended with a new strong model that may be my best fit. I compared this model to my original model after a backwards selection, and I found that I didn't have significant evidence that this interaction model added anything relevant to my model (with a statistical test). I am left with two models that may be my best fit going forward: my original model and my backwards selected original model. I will conduct a final test to determine which is more significant:

```
anova(diabetes_model1, new_diabetes_model1)
```

```
## Analysis of Variance Table
##
## Model 1: glu ~ npreg + bp + skin + bmi + ped + age + type
## Model 2: glu ~ bp + age + type
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    192 142632
## 2    196 144448 -4     -1816 0.6111 0.6551
```

Best fitted model

My final anova test tells me that I fail to reject my null hypothesis that the additional variables in model 1 (npreg, bp, skin, ped) make my model significantly stronger. I know this looking at my p-value of 0.6551, which is significantly greater than our desired significance level. Thus, I would proceed with my smaller model.

Statistical test for this conclusion:

My hypotheses for this test:

H_0 : y = $\beta_0 + \beta_1 * bp + \beta_2 * age\ \beta_3 * type + \epsilon$

H_1 : y = $\beta_0 + \beta_1 * npreg + \beta_2 * bp\ \beta_3 * skin + \beta_4 * bmi + \beta_5 * ped + \beta_6 * age + \beta_7 * type + \epsilon$

Test-statistic, P-value, and conclusion:

My test statistic is 0.6111 and p-value is 0.6551, conclusion at 10% significance level is to reject alternate hypothesis and stick to my smaller model. **best fit model:**

y_hat = 75.9991 + 0.3549 * bp + 0.4250 * age + 26.5690 * typeYes

---

Part 2: Model analysis

2a

Explanation for best fit model

Prevously, I decided my best fit model predicted glucose levels from my predictor variables blood pressure, age, and diabetes type. My conclusion in part 1 explains this, as through a combination of statistical tests and analyzing plots, I realized that my other predictor variables (npreg, skin, bmi, and ped) did not have

significant evidence to suggest they strengthened my model. Thus, through a backwards selection with an AIC metric, I was able to see that my strongest predictors in predicting glucose levels were bp, age, and type. My numerical evidence is present in the previous part.

2b

My fitted model:

y_hat = 75.9991 + 0.3549 * bp + 0.4250 * age + 26.5690 * typeYes

```
summary(new_diabetes_model1)
```

```
##
## Call:
## lm(formula = glu ~ bp + age + type, data = diabetes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -64.817 -18.320  -2.178  14.546  89.056
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  75.9991    12.2421   6.208 3.14e-09 ***
## bp            0.3549     0.1827   1.943   0.0535 .
## age           0.4250     0.2008   2.117   0.0355 *
## typeYes      26.5690     4.3670   6.084 6.04e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 27.15 on 196 degrees of freedom
## Multiple R-squared:  0.2762, Adjusted R-squared:  0.2651
## F-statistic: 24.93 on 3 and 196 DF,  p-value: 1.048e-13
```

2c

Number observations and coefficients I have 200 observations, n, and 4 coefficients, p.

2d

```
sd(diabetes$glu)
```

```
## [1] 31.66723
```

```r
summary(new_diabetes_model1)$sigma
```

```
## [1] 27.14735
```

Standard deviation interpretations The standard deviation of the glucose levels (31.66723) in my diabetes dataset represents the variability of the observed glucose values/observations, around the mean. The standard deviation (27.14735), or se_2, of my best fit model represents the variability of the observed glucose values around the predicted values from new_diabetes_model1.

Seeing how my se_2 is smaller (27.14735) than the standard deviation for glucose levels indicates that my best fit model explains some of the variability in the glucose levels, but there is still some unexplained variability that remains. The values are pretty similar, though, meaning that my model does not do a great job in explaining this variance.

2e

```r
new_diabetes_model1vif = lm(glu ~ bp + age, data = diabetes)
r_squaredj = summary(new_diabetes_model1vif)$r.squared
print(r_squaredj)
```

```
## [1] 0.1394707
```

```r
1 / (1- r_squaredj)
```

```
## [1] 1.162075
```

Collinearity? After removing my categorical predictor, type, my R^2 is 0.1394707, and upon calculating my VIF, I get 1.162075. Our threshold for a concerning VIF is any VIF above 5, and considering that my calculated VIF is less than that, I have no concerns for collinearity in this model.

2f

*L*inearity: true relationship between x & y is linear.

*I*ndependent: true errors are independent.

*N*ormal: true errors are normally distributed.

*E*qual varance: the variance of y at each x is the same σ^2.

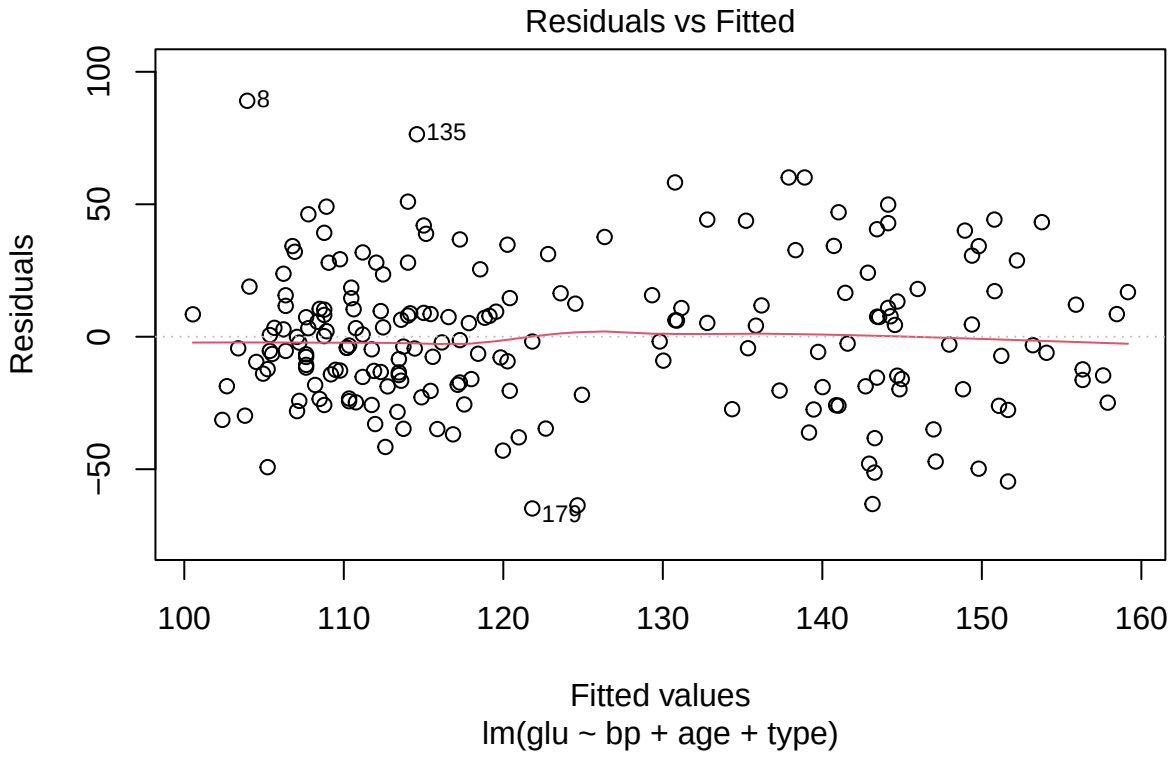
```r
library(lmtest)
```

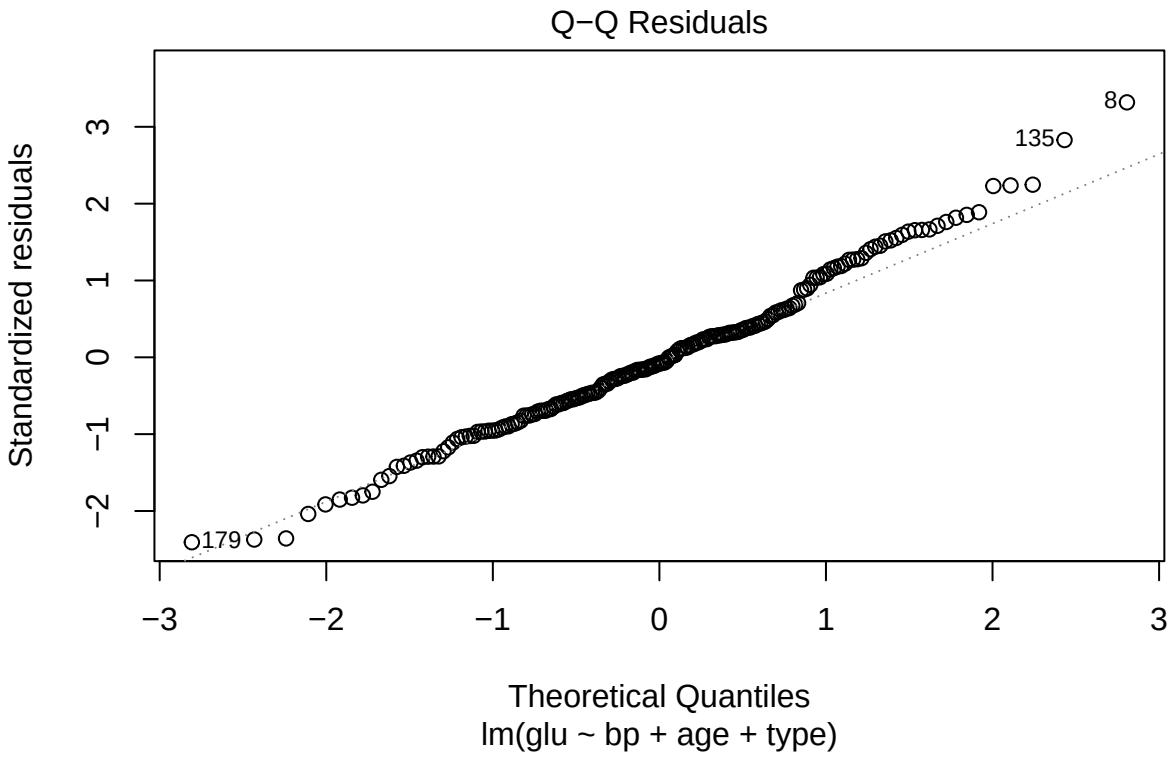
```
## Loading required package: zoo
```

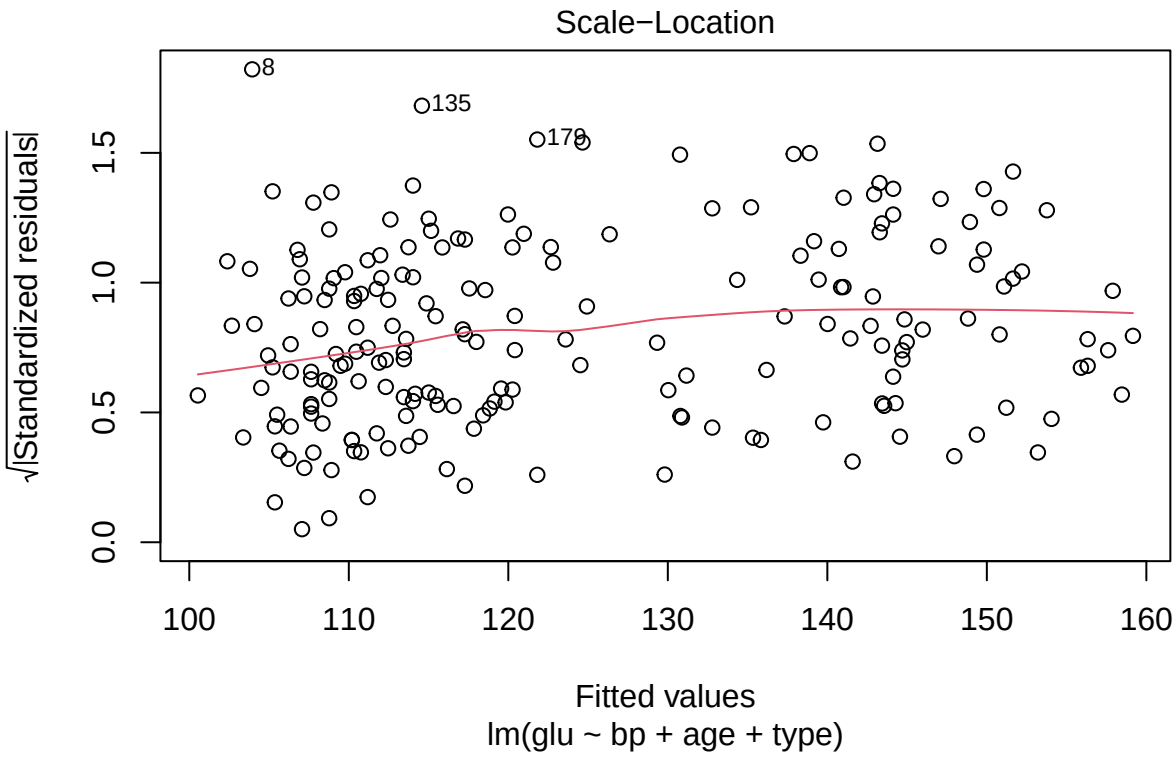
```
##
## Attaching package: 'zoo'
```

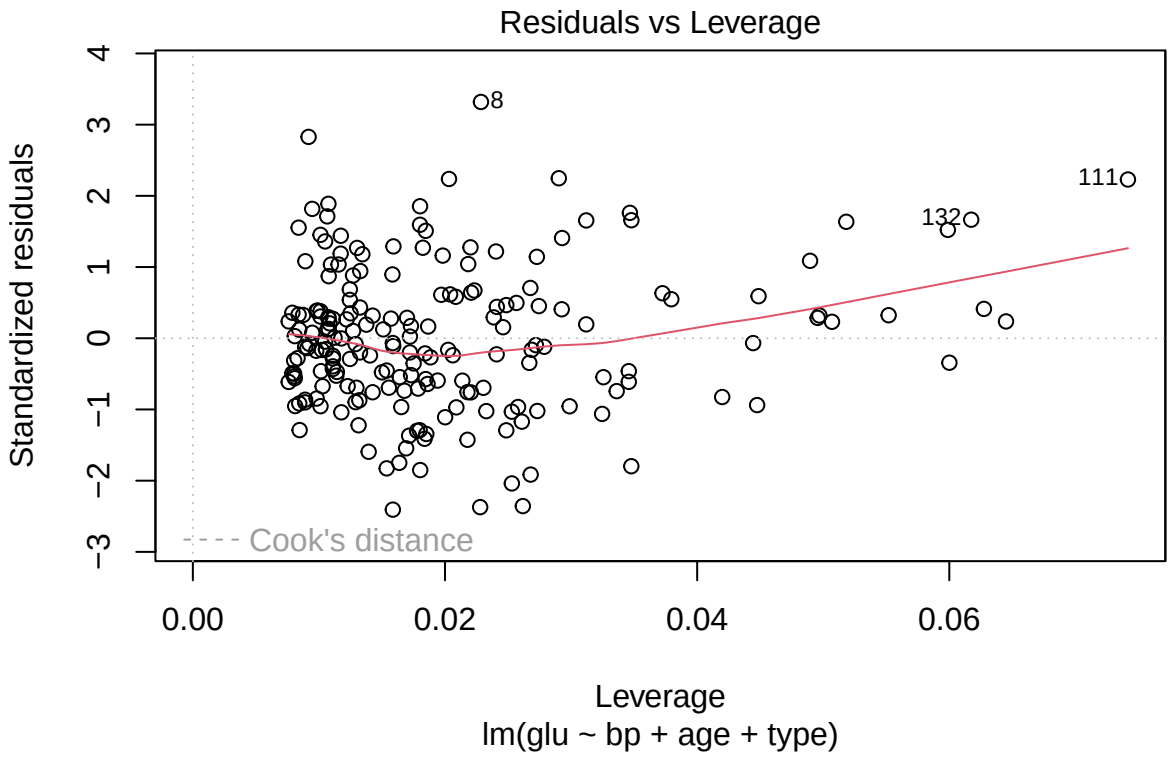
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
plot(new_diabetes_model1)
```




Q−Q Residuals
Standardized residuals
Theoretical Quantiles
lm(glu ~ bp + age + type)
179
135
8
−3
−2
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
8
135
179
100
110
120
130
140
150
160
Fitted values
lm(glu ~ bp + age + type)




```
shapiro.test(resid(new_diabetes_model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(new_diabetes_model1)
## W = 0.98986, p-value = 0.1703
```

```
bptest(new_diabetes_model1)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  new_diabetes_model1
## BP = 2.8359, df = 3, p-value = 0.4176
```

L: Looking at my residuals vs fitted plot, I can visualize an almost straight line centered at y=0, indicating that there is a true linear relationship between glucose and my predictor variables.

I: This depends on how my data was collected. If none of my observations depend on another observation, I can say this assumption can be met. However, I'm unable to tell.

N: Looking at my Q-Q plot and Shapiro Wilk test results, I can say that my true errors are not normally distributed. My true errors don't follow the line in my q-q plot around my theoretical quantatities greater than 1. As well, my p-value of 0.1703 (greater than 0.05) from my Shapiro test tells me that I can reject my null hypothesis that true errors are normally distributed.

E: Looking at my residuals vs fitted and my scale-location plot, I can see that I cannot draw a straight line across the top and bottom of my values. There are 3 outliers that affect my equal variance assumption, 8, 135, and 179. If I remove these values from my dataset, I can draw parallel lines around y=50 and y=-50 in my residuals vs fitted plot. However, as is, this model does not meet my equal variance assumption. As well, my bptest's p value of 0.4176 (greater than 0.05) tells me that I can reject my null hypothesis of homoskedasticity.

2g

```
n = length(resid(new_diabetes_model1))
p = length(coef(new_diabetes_model1))

hatvalues(diabetes_model1)[which(hatvalues(diabetes_model1) > 2 * (p / n))]
```

```
##          2          4          9         10         11         13         14
## 0.05609018 0.06008883 0.11746954 0.05097422 0.23721427 0.10680333 0.09064965
##         15         18         26         28         33         36         38
## 0.04708833 0.04584057 0.05423941 0.05813870 0.04003076 0.07660452 0.04022710
##         44         48         50         56         58         59         60
## 0.04035737 0.12530211 0.06861665 0.05756941 0.05466501 0.05649300 0.06663681
##         69         70         73         75         76         79         80
## 0.04866797 0.05350349 0.05480760 0.06895618 0.06125949 0.04686479 0.07745426
##         96        100        104        106        108        110        111
## 0.04712231 0.05376802 0.05718584 0.05448785 0.04506444 0.06445758 0.09129732
##        115        116        123        125        129        131        132
## 0.04736450 0.10316848 0.06435725 0.04903784 0.06840356 0.05712165 0.06706514
##        139        140        141        142        146        148        150
## 0.04152116 0.07822270 0.04343226 0.05351612 0.05284122 0.04338596 0.04315119
##        151        152        153        154        157        158        159
## 0.05401983 0.04411249 0.04049402 0.05442789 0.34328772 0.04051882 0.04156600
##        163        165        167        169        175        182        187
## 0.05659340 0.04243340 0.06348107 0.04325378 0.05565181 0.05914021 0.06790012
##        188        190        193
## 0.08416463 0.07569253 0.06139517
```

```
cooks_distance = cooks.distance(diabetes_model1)
which(cooks_distance > 4 / n)
```

```
##   2   4   6   8  13  14  50 101 104 111 132 135 153 154 157 175 179 187 188
##   2   4   6   8  13  14  50 101 104 111 132 135 153 154 157 175 179 187 188
```

Unusual observations

I can see 52 observations with high leverage. This may indicate that my leverage point is too low (currently 0.04), and I may want to raise it. When looking at my cooks distance, I see 19 observations whose distance from the mean is above my threshold. More importantly, in my residuals vs leverage plot from the previous part, I see 3 influential points, 8, 132, and 111. Within these 3 points, it appears that 132 and 111 have the highest leverage and seem to be affecting the line of my graph. I may want to adjust my model by treating these outliers, by excluding them; or, I can perform cross-validation on my model to assess its ability to predict new data, and potentially identify overfitting.

2h

```
calc_loocv_rmse = function(model) {
     sqrt(mean((resid(model) / (1 - hatvalues(model))) ^ 2))
 }
calc_loocv_rmse(diabetes_model1)
```

LOOCV error result

On average, the model's predictions are off by approximately 27.97 units in glucose levels when applied to new data points.

2i

Model complexity

Am I concerned if the number of observations is sufficient to fit my data? Our rule of thumb is that we'd like about 5 to 10 observations per coefficient, and seeing how in my best fit model I have 4 coefficients with 200 observations (obtained from part 2c), I can say I have enough data to fit this model.
