Pima Indian Women and their High Prevalence of Type 2 Diabetes

Presented by: Sharon Nelson

The Story

Native American Indians have participated in longitudinal studies concerning diabetes since the 1970s. Over the years, Pima Indians have shown greater prevalence for the disease, especially their women. In this study, I will be exploring a dataset collected on a population of the Pima women.

Questions:

1. What are the contributing factors for type 2 diabetes in Pima Indian women?
2. Knowing that insulin resistance is associated with type-2 diabetes, what is the variation in Glucose~Insulin in nondiabetic Pima women compared to diabetic Pima women?

Data Exploration

```
library(readxl)
diabetes=read_excel("diabetes.xlsx")
diabetes_n= subset(diabetes, Glucose!="0"& diabetes$BloodPressure!="0"& diabetes$SkinThickness!="0" & diabetes$Insulin!="0"&
diabetes$BMI!="0")

diabetes_clean= select(diabetes_n, -DiabetesPedigreeFunction)

head(diabetes_clean)
```

Pregnancies <dbl>	Glucose <dbl>	BloodPressure <dbl>	SkinThickness <dbl>	Insulin <dbl>	BMI <dbl>	Age <dbl>	Outcome <dbl>
1	89	66	23	94	28.1	21	0
0	137	40	35	168	43.1	33	1
3	78	50	32	88	31.0	26	1
2	197	70	45	543	30.5	53	1
1	189	60	23	846	30.1	59	1
5	166	72	19	175	25.8	51	1

6 rows

Pregnancies: # of pregnancies
Glucose: glucose level after 2 hours in an oral glucose tolerance test, *mg/dL*
Blood Pressure: diastolic blood pressure, *mmHg*
Skin Thickness: triceps skinfold thickness, *mm*
Insulin: 2-hour serum insulin, *Units/ml* of liquid
BMI: body mass index, *kg/m^2*
Age: years
Outcome - 1 means person has diabetes; 0 means no diabetes

Summary: a few descriptive statistics on entire data

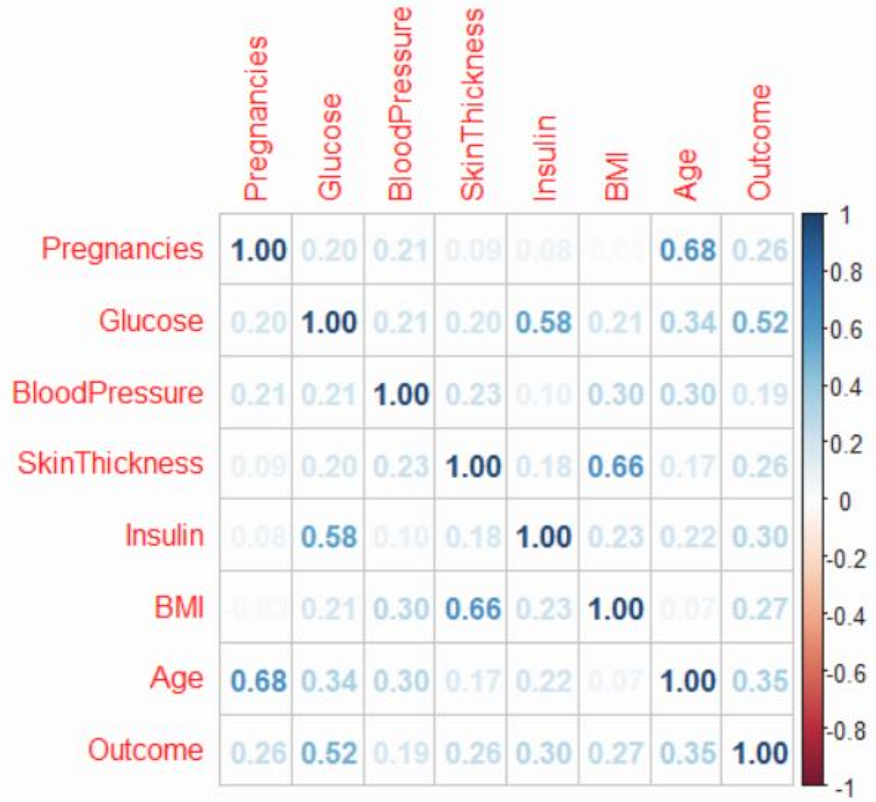
```
summary(diabetes_clean)
```

```
  Pregnancies        Glucose      BloodPressure    SkinThickness      Insulin           BMI        
 Min.   : 0.000   Min.   : 56.0   Min.   : 24.00   Min.   : 7.00   Min.   : 14.00   Min.   :18.20  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.:21.00   1st Qu.: 76.75   1st Qu.:28.40  
 Median : 2.000   Median :119.0   Median : 70.00   Median :29.00   Median :125.50   Median :33.20  
 Mean   : 3.301   Mean   :122.6   Mean   : 70.66   Mean   :29.15   Mean   :156.06   Mean   :33.09  
 3rd Qu.: 5.000   3rd Qu.:143.0   3rd Qu.: 78.00   3rd Qu.:37.00   3rd Qu.:190.00   3rd Qu.:37.10  
 Max.   :17.000   Max.   :198.0   Max.   :110.00   Max.   :63.00   Max.   :846.00   Max.   :67.10  
      Age           Outcome     
 Min.   :21.00   Min.   :0.0000  
 1st Qu.:23.00   1st Qu.:0.0000  
 Median :27.00   Median :0.0000  
 Mean   :30.86   Mean   :0.3316  
 3rd Qu.:36.00   3rd Qu.:1.0000  
 Max.   :81.00   Max.   :1.0000  
```

Finding Relationships: Question 1

- Skin thickness and BMI have a strong correlation, followed by pregnancy and age, **insulin and glucose, and glucose and outcome.**
- With an R-squared of 0.270*, glucose alone can not sufficiently explain outcome.
- What other factors help predict outcome?

```
cor.table = cor(select(diabetes_clean,1:8))
corrplot(cor.table, method="number")
```



*R-squared was calculated by squaring the r-value provided in the corrplot.

Outcome ~ Multiple Variables

```
summary(lm(data=diabetes_clean,formula = Outcome ~ Glucose*Age))
```

```
Call:
lm(formula = Outcome ~ Glucose * Age, data = diabetes_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-0.91813 -0.26516 -0.07884  0.30317  1.01907

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -8.107e-01  2.749e-01  -2.950 0.003375 **
Glucose      7.016e-03  2.112e-03   3.322 0.000979 ***
Age          9.828e-03  8.775e-03   1.120 0.263429
Glucose:Age -5.465e-06  6.367e-05  -0.086 0.931636
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3959 on 388 degrees of freedom
Multiple R-squared:  0.3001,    Adjusted R-squared:  0.2947
F-statistic: 55.46 on 3 and 388 DF,  p-value: < 2.2e-16
```

```
summary(lm(data=diabetes_clean,formula = Outcome ~ BMI*Age*Glucose))
```

```
Call:
lm(formula = Outcome ~ BMI * Age * Glucose, data = diabetes_clean)

Residuals:
     Min       1Q   Median       3Q      Max
-1.00056 -0.24283 -0.06076  0.25989  1.01914

Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)       1.789e+00  1.452e+00   1.232   0.2186
BMI              -7.323e-02  4.322e-02  -1.694   0.0910 .
Age              -9.264e-02  5.029e-02  -1.842   0.0662 .
Glucose          -1.255e-02  1.079e-02  -1.164   0.2452
BMI:Age           2.967e-03  1.485e-03   1.998   0.0464 *
BMI:Glucose       5.444e-04  3.157e-04   1.724   0.0855 .
Age:Glucose       6.522e-04  3.565e-04   1.830   0.0681 .
BMI:Age:Glucose -1.878e-05  1.043e-05  -1.801   0.0725 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3876 on 384 degrees of freedom
Multiple R-squared:  0.3362,    Adjusted R-squared:  0.3241
F-statistic: 27.78 on 7 and 384 DF,  p-value: < 2.2e-16
```

- Variation in outcome is better explained by glucose*age*BMI
  - Not too significant but better
- Note:
  - The more variables →higher adjusted R-squared
  - Prevalence of type-2 diabetes not based on just one factor

Outcome as a function is not providing enough information…..

Switching Directions

- How about glucose as a function of insulin, an indirect predictor of type 2 diabetes?
  - High levels of insulin after fasting→ larger amount of glucose in blood → higher risk of Type 2 diabetes

Glucose ~ Insulin

```
ggplot(data=diabetes_clean, aes(x=Insulin, y=Glucose, col=as.factor(Outcome)))+geom_point(size=2)+labs(title="Insulin vs. Gl
ucose", x="Insulin(units/mL)", y="Glucose(mg/dL)")
```

- A moderately linear trend with positive correlation (0.58).
- More individuals with type-2 diabetes on the higher side of glucose (>140 mg/dL) (MayoClinic).
- Cluster of individuals without diabetes on the lower side of both glucose and insulin.

Glucose ~ Insulin

```
dc_lm=lm(data=diabetes_clean,formula = Glucose ~ Insulin)
summary(dc_lm)
```

```
Call:
lm(formula = Glucose ~ Insulin, data = diabetes_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-65.633 -17.361  -5.807  12.626  78.813

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  99.0737     2.0979   47.23   <2e-16 ***
Insulin       0.1509     0.0107   14.11   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 25.14 on 390 degrees of freedom
Multiple R-squared:  0.3378,    Adjusted R-squared:  0.3361
F-statistic:   199 on 1 and 390 DF,  p-value: < 2.2e-16
```

- R-squared is low but high F-statistic suggests meaningful relationship between the two variables.
- Still worth analyzing…..

Question 2: What is the variation in glucose~insulin in healthy(outcome 0) individuals compared to diabetic(outcome 1) individuals?

Insulin Resistance

- Studies have shown that insulin levels in individuals with type-two diabetes can be high regardless of glucose level in blood– insulin resistance (*CDC.gov*).

- Based on the Pima Indian women dataset, how does this phenomena look in outcome 0 individuals compared to outcome 1 individuals?

	Normal Insulin Level
Fasting	< 25 mIU/L
30 minutes after glucose	30-230 mIU/L
1 hour after glucose	18-276 mIU/L
2 hours after glucose	16-166 mIU/L
3 hours or more after glucose	< 25 mIU/L

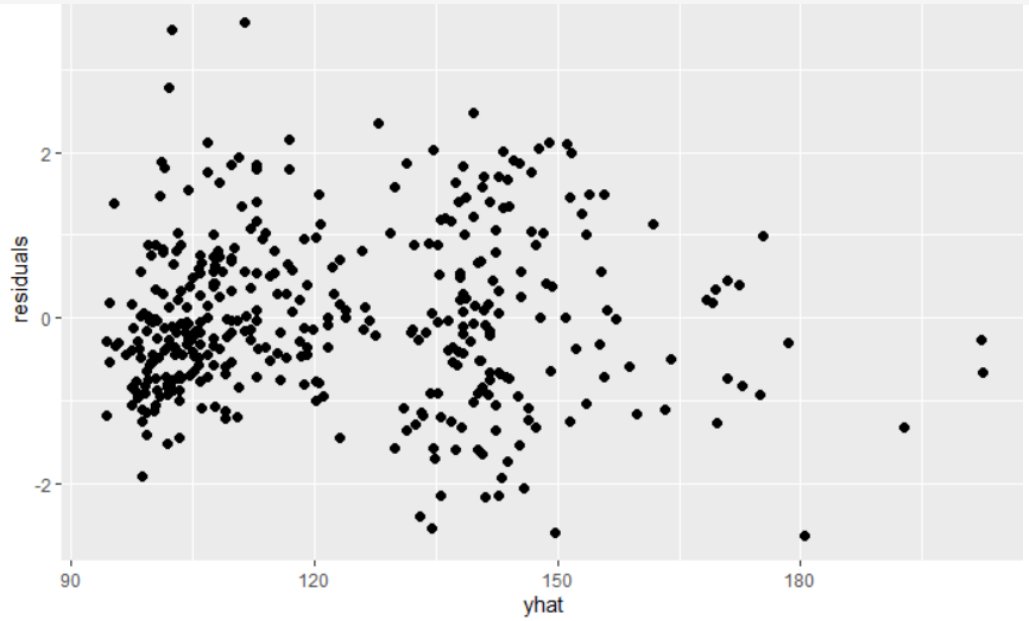
What Are Normal Levels of Insulin? | New Health Advisor, mIU/L = micro units/liter

```
Call:
lm(formula = Glucose ~ Insulin * Outcome, data = diabetes_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-57.771 -15.388  -2.971  13.427  79.695

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     92.04509    2.24590  40.984  < 2e-16 ***
Insulin          0.14815    0.01352  10.962  < 2e-16 ***
Outcome         34.63356    4.28527   8.082 8.18e-15 ***
Insulin:Outcome -0.05865    0.02009  -2.919  0.00372 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 22.41 on 388 degrees of freedom
Multiple R-squared:  0.4768,    Adjusted R-squared:  0.4728
F-statistic: 117.9 on 3 and 388 DF,  p-value: < 2.2e-16
```

- Glucose as a function of Insulin*Outcome produces greater explanation for variation in glucose levels and provides significant insight into the glucose~insulin relationship for both non-diabetics and diabetics.

Glucose ~ Insulin*Outcome:

```
r=rstandard(dc_lm2)
yhat=fitted(dc_lm2)
df=data.frame(residuals=r,yhat=yhat)

#Checking for linearity and equal standard deviations - should see spread, no patterns
ggplot(df,aes(x=df$yhat,y=df$residuals))+geom_point(size=2)+labs(x="yhat", y="residuals")
```

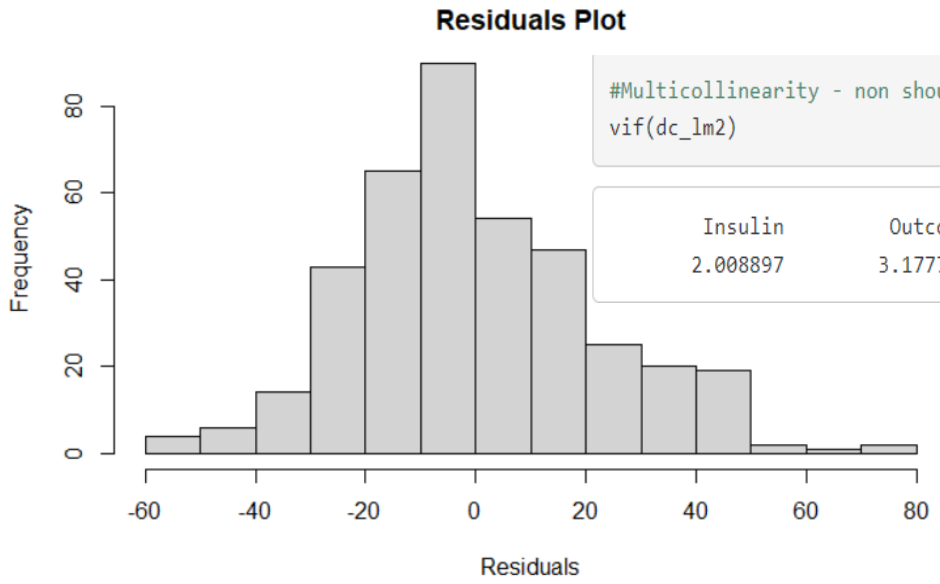


yhat = fitted values for glucose as a function of Insulin*Outcome

Multiple Linear Regression Assumptions:

1. Linearity
2. Equal standard deviations
3. Independence
4. Normality of residuals
5. Non-Multicollinearity: explanatory variables should not be more correlated than either is to the dependent variable

```
#Checking for normality
hist(dc_lm2$residuals, main="Residuals Plot", xlab="Residuals")
```

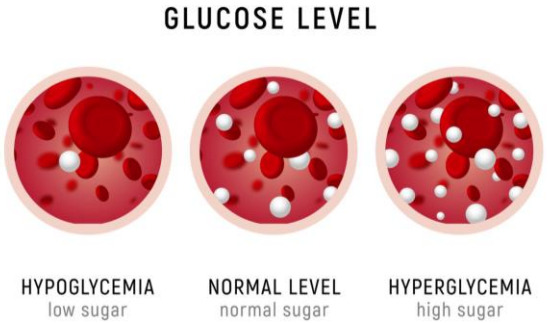


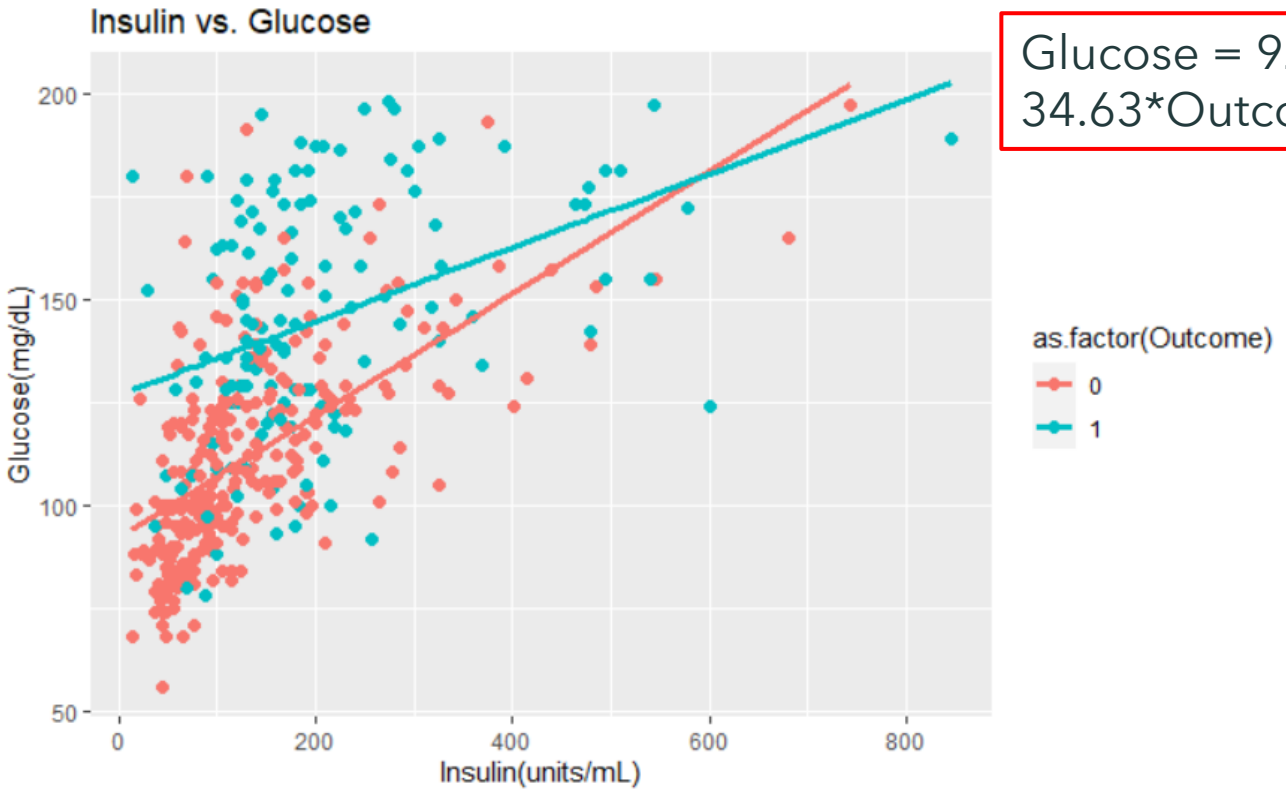

```
#Multicollinearity - non should be 10 or >10
vif(dc_lm2)
```

Insulin	Outcome	Insulin:Outcome
2.008897	3.177755	4.815267

Glucose ~ Insulin*Outcome

```
dc_predict=predict(dc_lm2)
dc_df2=mutate(dc_df,Predicted2_Values=dc_predict)

ggplot(data=dc_df2, aes(x=Insulin, y=Glucose, col=as.factor(Outcome)))+geom_point(size=2)+labs(title="Insulin vs. Glucose",
x="Insulin(units/mL)", y="Glucose(mg/dL)")+geom_line(data=dc_df2, aes(y=Predicted2_Values, col=as.factor(Outcome)), size=1.
2)
```



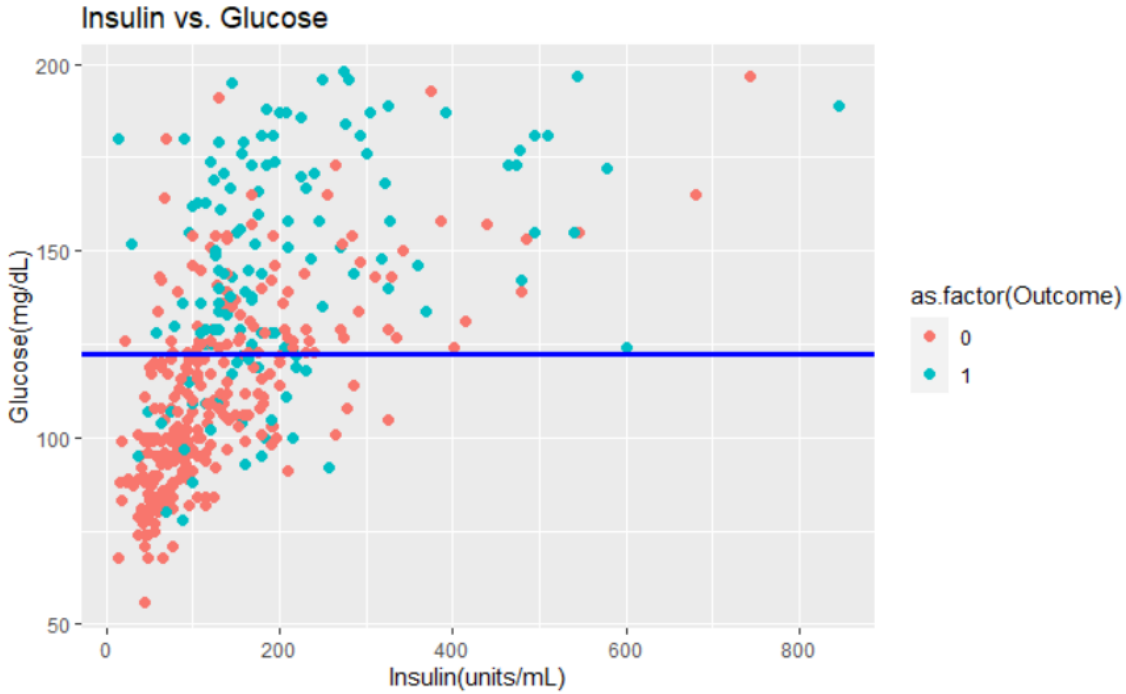


Glucose = 92.05 + 0.15*Insulin + 34.63*Outcome - 0.059*Insulin*Outcome

- Model for outcome 0: starts at lower value for glucose and considers the increase in insulin as glucose increases
- Model for outcome 1: accounts for the higher values of glucose and higher values in insulin altogether

Reference Model vs. New Model

```
ggplot(data=diabetes_clean, aes(x=Insulin, y=Glucose, col=as.factor(Outcome)))+geom_point(size=2)+labs(title="Insulin vs. Gl
ucose", x="Insulin(units/mL)", y="Glucose(mg/dL)")+geom_hline(yintercept = 122.6, size=1.1, col="blue")
```



Reference model

New model

Testing Significance of New Model by P-value

- Null: The fit of the reduced model(reference) and full model (new model) are equal.
- Alternative: The fit of the new model yields significant improvements over the reference model.
  - Alpha is 0.05.
- By F statistic and p-value, reject null and accept that the fit of the new model yields significant improvements over the reference model.
  - Large F means good amount of variation in glucose can be explained by insulin*outcome as a predictor.
- It is easier to predict where a diabetic person's glucose~insulin level may be for regulating insulin medication
- It is easier to assess non-diabetics to evaluate whether they are at risk for diabetes or not

```
Call:
lm(formula = Glucose ~ Insulin * Outcome, data = diabetes_clean)

Residuals:
    Min      1Q  Median      3Q     Max
-57.771 -15.388  -2.971  13.427  79.695

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     92.04509    2.24590  40.984  < 2e-16 ***
Insulin          0.14815    0.01352  10.962  < 2e-16 ***
Outcome         34.63356    4.28527   8.082 8.18e-15 ***
Insulin:Outcome -0.05865    0.02009  -2.919  0.00372 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 22.41 on 388 degrees of freedom
Multiple R-squared:  0.4768,    Adjusted R-squared:  0.4728
F-statistic: 117.9 on 3 and 388 DF,  p-value: < 2.2e-16
```

Conclusion

- It is obvious that diabetes is not dependent upon one factor
- As the R-squared value increased with increasing explanatory variables, it is prevalent that diabetes is dependent upon multiple factors.
  - Glucose alone was not sufficient to predict outcome
  - BMI and age were also contributing factors
- Diabetics have a higher glucose~insulin relationship while non-diabetics have a lower glucose~insulin relationship
- Overall, diabetes may be more complicated than we think

Citations

- Mayo Clinic. https://www.mayoclinic.org/tests-procedures/glucose-tolerance-test/about/pac-20394296#:~:text=%20If%20you%27re%20being%20tested%20for%20type%202,mmol%2FL%29%20or%20higher%20may%20indicate%20diabetes.%20More%20. Accessed April 24, 2022
- CDC. https://www.cdc.gov/diabetes/basics/insulin-resistance.html. Accessed April 24,2022
- NewHealthAdvisor. https://www.newhealthadvisor.org/Normal-Insulin-Levels.html. Accessed April 24,2022