Variations in Diabetes Progression

Jeet Patel

April 2, 2025

1. **Introduction** Diabetes is one of the top ten leading cause of death and an ongoing pandemic. However, there are a lot of factors that contribute to diabetes and its progression. In our experiment, we take 442 patients and measure 10 predictor variables: age, sex, body mass index, average blood pressure, serum total cholesterol level, low-density lipoproteins, high-density lipoproteins, total cholesterol/HDL ratio, serum triglyceride level, and blood sugar level. Our goal is to interpret the data and figure out which predictor variable influences the response variable, diabetes progression level. We utilize multiple statical measurements including mean, variance, correlation, multicollinearity, and coefficient of determination; our conclusions are reinforced with graphs including histograms and boxplots. In addition, we check the models for linearity, independence, normality, and equal variance of residuals. After confirming the criteria is met, we try to apply backwards elimination to take out insignificant predictor variables to produce the best model. We conclude by interpreting the coefficient of determination and better understanding the statically significant variables.

2. **Data Introduction and Description** Our analysis focuses on 10 predictor variables: age, sex, body mass index, average blood pressure, serum total cholesterol level, low-density lipoproteins, high-density lipoproteins, total cholesterol/HDL ratio, serum triglyceride level, and blood sugar level and only one response variable: diabetes progression indicator (the target). We look at the mean and standard deviation in order to estimate predictor variables that are potentially influential before we conduct higher level statistical tests. In our appendix, there are histograms and boxplots that showcase that our data is overall normally distributed in all 10 categories that removes bias and other confounding variables.

   Table 1. The dataset

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	59	2	32.1	101	157	93.2	38	4	4.9	87	151
2	48	1	21.6	87	183	103.2	70	3	3.9	69	75
3	72	2	30.5	93	156	93.6	41	4	4.7	85	141
4	24	1	25.3	84	198	131.4	40	5	4.9	89	206
5	50	1	23.0	101	192	125.4	52	4	4.3	80	135
6	23	1	22.6	89	139	64.8	61	2	4.2	68	97
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	36	1	30.0	95	201	125.2	42	5	5.1	85	220
442	36	1	19.6	71	250	133.2	97	3	4.6	92	57

Table 2. Five Number Summaries (including mean)

Statistic	AGE	SEX	BMI	BP	S1	S2	S3	S4	S5	S6	Target
Min.	19.00	1.000	18.00	62.00	97.0	41.60	22.00	2.00	3.258	58.00	25.0
1st Qu.	38.25	1.000	23.20	84.00	164.2	96.05	40.25	3.00	4.277	83.25	87.0
Median	50.00	1.000	25.70	93.00	186.0	113.00	48.00	4.00	4.620	91.00	140.5
Mean	48.52	1.468	26.38	94.65	189.1	115.44	49.79	4.07	4.641	91.26	152.1
3rd Qu.	59.00	2.000	29.27	105.00	209.8	134.50	57.75	5.00	4.997	98.00	211.5
Max.	79.00	2.000	42.20	133.00	301.0	242.40	99.00	9.09	6.107	124.00	346.0

Table 2 Analysis. The mean age of patients is 48.5 and the median age is 50 with a range interquartile range of 38.2-59. Therefore, our dataset covers a middle-aged population. The mean of sex is 1.46 (male = 1, female = 2), indicating there is slightly more males than females in the dataset.

Table 3. Standard Deviation and Variance

Variable	Standard Deviation (SD)	Variance (SD^2)

AGE	13.1090	171.85
SEX	0.4996	0.25
BMI	4.4181	19.52
BP	13.8313	191.30
S1	34.6081	1197.83
S2	30.4131	924.98
S3	12.9342	167.26
S4	1.2904	1.66
S5	0.5224	0.27
S6	11.4963	132.38
Target	77.0930	5943.34

Table 3 Analysis. We notice there is a high variance in S1 and S2 indicating they might be rational indicators of the target variable. In contrast, there is a low variance in S4 and S5 indicating there might not be rational indicators of the target variable. What is most significant is the target is the variable with the highest variance meaning there are potentially multiple predictor variables influencing disease progression indicator that contributes the wide range of diversity. Therefore, it is important to conduct a multiple linear regression model in order to figure out what are the significant predictor variables.

3. **Multiple Linear Regression Analysis** To determine a relationship between the explanatory variable and response variable, it is important to conduct a hypotheses test composed of a preliminary hypothesis and primary hypothesis.

   The first step is the preliminary hypothesis. We ought to check the conditions are met including linearity, independence, normality, and equal variance (L.I.N.E.). We utilize multiple tools including plot graphs, the Durbin-Watson test, and Shapiro-Wilk test.

   Figure 1. Linearity: plot of predictor variable vs residuals

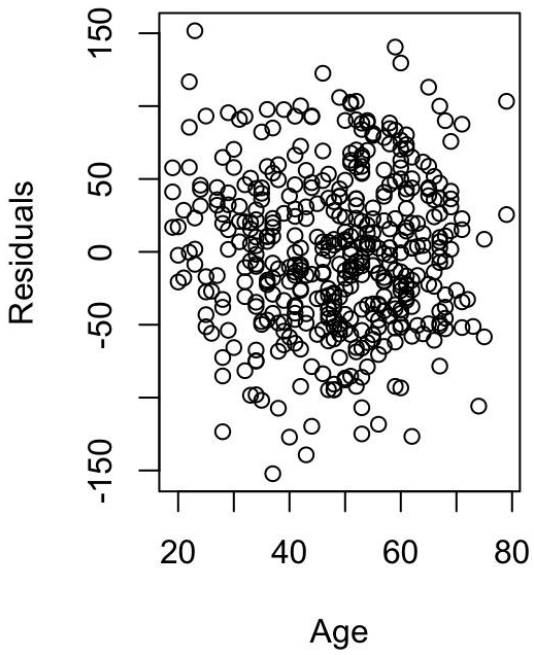


Figure 1 Analysis. The linearity condition requires making a plot for all 10 predictor variables. Figure 1 focuses on Age where the other predictor variable graphs are in the appendix. We notice that the points are scattered with no pattern implying there is linearity. The same pattern appears in all other variables. Therefore, we are able to conclude linearity.

Figure 2. Independence: Durbin -Watson test

Test	Statistic	p-value	Interpretation
Durbin-Watson Test	2.0285	0.8906	No significant autocorrelation detected (residuals appear independent).

Figure 2 Analysis. Our p-value is .8906 and our significance level is .05. Since our p-value is greater than the significance level (.8906 > .05), we fail to reject the Ho implying independence.

Figure 3. Normality: Shapiro-Wilk test

Test	Statistic	p-value	Interpretation

Shapiro-Wilk Test	.99647	.443	Residuals appear to be normally distributed (fail to reject normality).

Figure 3 Analysis. Our p-value is .443 and our significance level is .05. Since our pvalue is greater than the significance level (.443 > .05), we fail to reject the Ho implying normality.

Figure 4. Equal Variance: plot of fitted values vs residuals

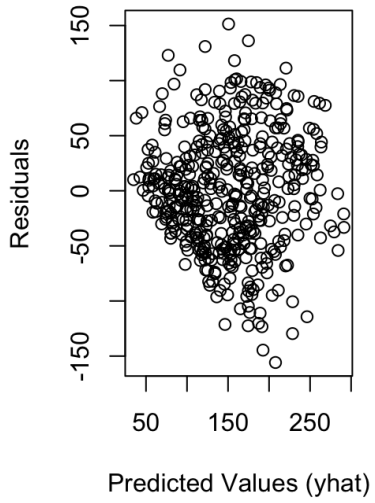


Figure 4 Analysis. The main goal is to see no pattern and equal variance. The plot points are clearly random and scattered and therefore we fulfil the equal variance condition.

Another factor we inspect is correlation and multicollinearity (VIF) to prevent confounding variables that mess up our hypothesis test. It is futile that all the predictor variables are independent, where a high correlation like a r greater than .7 or multicollinearity with a VIF greater than 10 might negativity influence the validity of an identical and independent distribution.

Figure 5. Correlation Matrix

	AGE	**SEX**	**BMI**	**BP**	**S1**	**S2**	**S3**	**S4**	**S5**	**S6**	**Target**
AGE	1.000	0.174	0.185	0.335	0.260	0.219	-0.075	0.204	0.271	0.302	0.188
SEX	0.174	1.000	0.088	0.241	0.035	0.143	-0.379	0.332	0.150	0.208	0.043

BMI	0.185	0.088	1.000	0.395	0.250	0.261	-0.367	0.414	0.446	0.389	0.586
BP	0.335	0.241	0.395	1.000	0.242	0.186	-0.179	0.258	0.393	0.390	0.441
S1	0.260	0.035	0.250	0.242	1.000	0.897	0.052	0.542	0.516	0.326	0.212
S2	0.219	0.143	0.261	0.186	0.897	1.000	-0.196	0.660	0.318	0.291	0.174
S3	-0.075	-0.379	-0.367	-0.179	0.052	-0.196	1.000	-0.738	-0.399	-0.27	-0.395
S4	0.204	0.332	0.414	0.258	0.542	0.660	-0.738	1.000	0.618	0.417	0.430
S5	0.271	0.150	0.446	0.393	0.516	0.318	-0.399	0.618	1.000	0.465	0.566
S6	0.302	0.208	0.389	0.390	0.326	0.291	-0.274	0.417	0.465	1.000	0.382
Target	0.188	0.043	0.586	0.441	0.212	0.174	-0.395	0.430	0.566	0.382	1.000

Figure 6. Original Multicollinearity

Variable	VIF Value
AGE	1.217307
SEX	1.278071
BMI	1.509437
BP	1.459428
S1	59.202510
S2	39.193370
S3	15.402156
S4	8.890986
S5	10.075967
S6	1.484623

Figure 7. Multicollinearity without S1

Variable	VIF Value
AGE	1.216892
SEX	1.275049
BMI	1.502320
BP	1.457413
S2	2.926535
S3	3.736890
S4	7.818670
S5	2.172865
S6	1.484410

Figure 5-7 Analysis. All the correlations in the correlation matrix are under .7 (the correlation with itself is always 1 and not a key factor in determining dependence between the predictor variables). The exception is S1 and S2 where r = .897 implying a high correlation and thus potential dependence between S1 and S2. On top of that, there are multiple VIF values above 10 indicating multicollinearity: S1, S2, S3, and S5. To fix that, we plan to remove S1 from our linear model. After that, our VIF values are all below 10. By testing correlation and multicollinearity, we uphold the independence criteria by removing S1.

The next step is the primary hypotheses test. In our multiple linear regression, our H0 is all the regression coefficients are equal to zero ($\beta 1 = \beta 2 = \ldots = \beta n = 0$). The null hypothesis means there is no relationship between any predictor variables and the response variable. Our H1 is at least one regression coefficient is not equal to zero (at least one $\beta_j != 0$). The alternative hypothesis means there is a minimum of one predictor variable that affects the response variable. Now, we take a look at the ANOVA table.

Figure 8. ANOVA Table

Predictor	Df	Sum Sq	Mean Sq	F Value	Pr(>F)	Significance
AGE	1	92,527	92,527	31.5504	3.49e-08	*** (Significant)
SEX	1	293	293	0.1000	0.7519	Not Significant
BMI	1	826,955	826,955	281.9792	< 2.2e-16	*** (Highly Significant)
BP	1	129,312	129,312	44.0934	9.448e-11	*** (Significant)
S1	1	1,791	1,791	0.6108	0.4349	Not Significant
S2	1	5,058	5,058	1.7246	0.1898	Not Significant
S3	1	237,329	237,329	80.9257	< 2.2e-16	*** (Highly Significant)
S4	1	1,821	1,821	0.6210	0.4311	Not Significant
S5	1	58,856	58,856	20.0690	9.582e-06	*** (Significant)
S6	1	3,080	3,080	1.0504	0.3060	Not Significant
Residual	43	1,263,981	2,933	-	-	-

Our test statistic is the F Value, and the significance level is.05. We reject the null hypothesis when the test static is smaller than the significance level. Furthermore, we reject the null hypothesis for the predictor variables Age, BMI, BP, S3, and S5. Thus, we conclude that the variables Age, BMI, BP, S3, and S5 statically contribute to the variance of the diabetes progression indicator. In other words, we reason that Age, BMI, BP, S3, and S5 influence diabetes progression indicator and therefore the variables are correlated.

4. **Variable Selection** In order to make the best linear regression model, it is important to eliminate statistically insignificant variables for the best model. We already eliminated S1 while checking for correlation and multilinearity in order to fulfil the independence criteria. Now we continue the process via backward elimination. The steps of backward elimination include:

1) Start with a model with all of variables
2) Omit insignificant variables one at a time
3) Refit the model after the variable is omitted
4) Repeat the process as necessary

Since our significance is .10 or lower, we eliminate variables with a pvalue greater than .1. However, it is important not to make the mistake of eliminating all the insignificant variables at the same time instead of one at a time.

Model 1. Original Model – Without S1

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -267.5419    37.1322  -7.205 2.59e-12
AGE           -0.1119     0.2199  -0.509   0.6113
BMI            6.0800     0.7216   8.426 5.33e-16
BP             0.9397     0.2252   4.172 3.65e-05
S2            -0.2152     0.1477  -1.457   0.1459
S3            -0.6570     0.3862  -1.701   0.0896
S4             1.1514     5.6788   0.203   0.8394
S5            45.1410     7.3679   6.127 2.02e-09
S6             0.2099     0.2776   0.756   0.4501
---
```

Caption: Our largest pvalue is .8394 for S4. Thus, S4 is statistically insignificant.

Model 2. Eliminate S4

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -265.6117    35.8515  -7.409 6.71e-13
AGE           -0.1126     0.2197  -0.512  0.60862
BMI            6.0712     0.7195   8.439 4.84e-16
BP             0.9364     0.2244   4.173 3.63e-05
S2            -0.1921     0.0936  -2.052  0.04077
S3            -0.7201     0.2287  -3.149  0.00175
S5            45.8751     6.4100   7.157 3.55e-12
S6             0.2142     0.2765   0.775  0.43903
```

Caption: Our largest pvalue is .60862 for Age. Thus, Age is statistically insignificant.

Model 3. Eliminate Age

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -264.68365   35.77534  -7.398 7.16e-13 
BMI            6.07775    0.71873   8.456 4.22e-16 
BP             0.91171    0.21897   4.164 3.78e-05 
S2            -0.19769    0.09288  -2.128  0.03386 
S3            -0.72767    0.22801  -3.191  0.00152 
S5            45.57435    6.37765   7.146 3.79e-12 
S6             0.19441    0.27359   0.711  0.47771   
---
```

Caption: Our largest pvalue is .711 for S6. Thus, S6 is statistically insignificant.

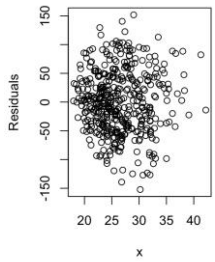
Model 4. Eliminate S6 – Final Model

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -257.58289   34.33196  -7.503 3.54e-13
BMI            6.14848    0.71140   8.643  < 2e-16
BP             0.94387    0.21412   4.408 1.31e-05
S2            -0.18870    0.09196  -2.052  0.04077
S3            -0.73778    0.22744  -3.244  0.00127
S5            46.69411    6.17640   7.560 2.40e-13
```

Caption: There are no pvalues greater than .10. Therefore, BMI, BP, S2, S3, and S5 are all statically significant.

Now, it is time to check the criteria of linearity, independence, normality, and equal variance for the final model after backwards elimination. In addition, we plan to compare the original linear model with the final original model with an ANOVA table to detect a significant difference.

Model 5. Linearity: plot of predictor variable vs residuals



Model 5 Analysis. The linearity condition requires making a plot for all 5 predictor variables. Figure 1 focuses on BMI where the other predictor variable graphs are in the appendix. We notice that the points are scattered with no pattern implying there is linearity. The same pattern appears in all other variables. Therefore, we are able to conclude linearity.

Model 6. Independence: Durbin -Watson test

```
        Durbin-Watson test

data:  model_4
DW = 2.0118, p-value = 0.8992
alternative hypothesis: true autocorrelation is not 0
```

Model 6 Analysis. Our p-value is .8992 and our significance level is .05. Since our p-value is greater than the significance level (.8992 > .05), we fail to reject the Ho implying independence.

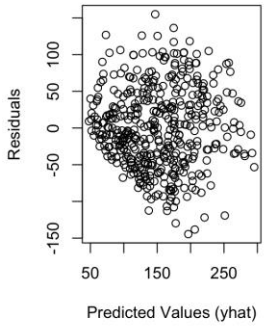
Model 7. Normality: Shapiro-Wilk test

```
            Shapiro-Wilk normality test

data:  residuals(model_4)
W = 0.99443, p-value = 0.1084
```

Model 7 Analysis. Our p-value is .1084 and our significance level is .05. Since our pvalue is greater than the significance level (.1084 > .05), we fail to reject the Ho implying normality.

Model 8. Equal Variance: plot of fitted values vs residuals



Model 8 Analysis. The main goal is to see no pattern and equal variance. The plot points are clearly random and scattered and therefore we fulfil the equal variance condition.

Model 9. ANOVA Tables – Original vs Final

```
Analysis of Variance Table

Model 1: Target ~ AGE + BMI + BP + S2 + S3 + S4 + S5 + S6
Model 2: Target ~ BMI + BP + S2 + S3 + S5
  Res.Df     RSS Df Sum of Sq      F Pr(>F)
1    433 1317587
2    436 1320040 -3   -2452.9 0.2687  0.848
```

Model 9 Analysis. Since the pvalue is .848, there is no significant difference between the original and final model. Consequently, removing the variables, Age and S4 does not reduce the model's explanatory power.

5. Conclusion

Models 10-13. R^2 -backward elimination models

```
Residual standard error: 55.16 on 433 degrees of freedom
Multiple R-squared:  0.4973,    Adjusted R-squared:  0.488
F-statistic: 53.54 on 8 and 433 DF,  p-value: < 2.2e-16
```

```
Residual standard error: 55.1 on 434 degrees of freedom
Multiple R-squared:  0.4973,    Adjusted R-squared:  0.4891
F-statistic: 61.32 on 7 and 434 DF,  p-value: < 2.2e-16
```

```
Residual standard error: 55.06 on 435 degrees of freedom
Multiple R-squared:  0.4969,    Adjusted R-squared:   0.49
F-statistic: 71.62 on 6 and 435 DF,  p-value: < 2.2e-16
```

```
Residual standard error: 55.02 on 436 degrees of freedom
Multiple R-squared:  0.4964,    Adjusted R-squared:  0.4906
F-statistic: 85.94 on 5 and 436 DF,  p-value: < 2.2e-16
```

One key observation between all the models is the coefficient of determination is relatively the same in that it is .49. That means that taking out the variables S1, Age, S4, and S6 does not majorly impact that model and consequently confirms that the selected predictor variables are statistically insignificant. Utilizing the coefficient of determination, we interpret 49.64% of the variation in the Target variable can be explained by the independent variables BMI, BP, S2, S3, and S5 in my model. On top of that, the high F-statistic (85.94) and low pvalue (2.2e-16) indicate the model is statically significant as a whole.

Appendix

```
#Name: Jeet Patel
#Project: Regression Analysis on Diabetes
#Summary: Determine what variables influence diabetes progression indicator
#We plan to conduct hypothesis test utilizing an ANOVA table
#In addition, we look at correlation and multicollinearity
#And we validate the criteria of linearity, independence, normality, and equal variance



#Part 1: Data Introduction and Description
diabetes <- read.delim("/Users/jeetpatel/Desktop/Diabetes.csv", header = TRUE)
diabetes[] <- lapply(diabetes, function(x) if(is.character(x)) as.numeric(x) else x)

#Five Number Summary
summary(diabetes)

#Mean
sapply(diabetes, mean, na.rm=TRUE)

#Variance
sapply(diabetes, sd, na.rm=TRUE)

#Histogram or Boxplot
#Age
hist(diabetes$AGE, main="Histogram of Age", xlab="Age", col="blue", border="black")
boxplot(diabetes$AGE, main="Boxplot of Age", ylab="Age", col="red")

#Sex
hist(diabetes$SEX, main="Histogram of Sex", xlab="Sex", col="blue", border="black")
boxplot(diabetes$SEX, main="Boxplot of Sex", ylab="Sex", col="red")

#BMI
hist(diabetes$BMI, main="Histogram of BMI", xlab="BMI", col="blue", border="black")
boxplot(diabetes$BMI, main="Boxplot of BMI", ylab="BMI", col="red")

#BP
hist(diabetes$BP, main="Histogram of BP", xlab="BP", col="blue", border="black")
boxplot(diabetes$BP, main="Boxplot of BP", ylab="BP", col="red")

#S1
hist(diabetes$S1, main="Histogram ofS1", xlab="S1", col="blue", border="black")
boxplot(diabetes$S1, main="Boxplot of S1", ylab="S1", col="red")

#S2
hist(diabetes$S2, main="Histogram of S2", xlab="S2", col="blue", border="black")
boxplot(diabetes$S2, main="Boxplot of S2", ylab="S2", col="red")

#S3
hist(diabetes$S3, main="Histogram of S3", xlab="S3", col="blue", border="black")
boxplot(diabetes$S3, main="Boxplot of S3", ylab="S3", col="red")

#S4
hist(diabetes$S4, main="Histogram of S4", xlab="S4", col="blue", border="black")
boxplot(diabetes$S4, main="Boxplot of S4", ylab="S4", col="red")

#S5
hist(diabetes$S5, main="Histogram of S5", xlab="S5", col="blue", border="black")
boxplot(diabetes$S5, main="Boxplot of S5", ylab="S5", col="red")

#S6
hist(diabetes$S6, main="Histogram of S6", xlab="S6", col="blue", border="black")
```

```
boxplot(diabetes$S6, main="Boxplot of S6", ylab="S6", col="red")



#Part 2: Multiple Linear Regression Analysis
full_model <- lm(Target ~ AGE + SEX + BMI + BP + S1 + S2 + S3 + S4 + S5 + S6,
data=diabetes)
summary(full_model)

#ANOVA
anova(full_model)

# Correlation Matrix
cor_matrix <- cor(diabetes[, sapply(diabetes, is.numeric)])
print(cor_matrix)

# Variance Inflation Factor (VIF)
install.packages("car")
library(car)

vif_values <- vif(full_model)
print(vif_values)

#Take out S1
full_model <- lm(Target ~ AGE + SEX + BMI + BP + + S2 + S3 + S4 + S5 + S6, data=diabetes)

vif_values <- vif(full_model)
print(vif_values)

#L.I.N.E.
#Linearity
plot(x = diabetes$AGE, y =full_model$residuals, xlab = "Age", ylab = "Residuals")
plot(x = diabetes$SEX, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$BMI, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$BP, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$S2, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$S3, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$S4, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$S5, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$S6, y =full_model$residuals, xlab = "x", ylab = "Residuals")

#Independence
install.packages("lmtest")
library(lmtest)
dwtest(full_model, alternative = "two.sided")

#Normality
qqnorm(full_model$residuals)
shapiro_test <- shapiro.test(residuals(full_model))
print(shapiro_test)

#Equal Variance
plot(x = full_model$fitted.values, y = full_model$residuals, xlab = "Predicted Values
(yhat)", ylab = "Residuals")



#Part 3: Variable Selection
#Original: Removed S1
model_1 <- lm(Target ~ AGE + BMI + BP + S2 + S3 + S4 + S5 + S6, data = diabetes)
```

```
summary(model_1)

#Remove S4
model_2 <- lm(Target ~ AGE + BMI + BP + S2 + S3 + S5 + S6, data = diabetes)
summary(model_2)

#Remove Age
model_3 <- lm(Target ~ BMI + BP + S2 + S3 + S5 + S6, data = diabetes)
summary(model_3)

#Remove S6
model_4 <- lm(Target ~ BMI + BP + S2 + S3 + S5, data = diabetes)
summary(model_4)

#ANOVA Comparison
anova(model_1, model_4)

#L.I.N.E.
#Linearity
plot(x = diabetes$BMI, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$BP, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$S2, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$S3, y =full_model$residuals, xlab = "x", ylab = "Residuals")
plot(x = diabetes$S5, y =full_model$residuals, xlab = "x", ylab = "Residuals")

#Independence
dwtest(model_4, alternative = "two.sided")

#Normality
qqnorm(model_4$residuals)
shapiro_test <- shapiro.test(residuals(model_4))
print(shapiro_test)

#Equal Variance
plot(x = model_4$fitted.values, y = model_4$residuals, xlab = "Predicted Values (yhat)",
ylab = "Residuals")
```