NCD_Analysis

Omar Alsubaihi, Abdullah Bin Afif, Yousef Alharbi, Abdulqudus Idris Abiri

2025-11-09

CRISP-DM Cycle 1

1- Business Understanding

The aim of this project to establish a predictive relationship by modeling Diabetes Prevalence in adults as the response variable (Y) . The explanatory variables (X) chosen reflect the major upstream risk factors: Obesity Prevalence, Mean Total Cholesterol, and Prevalence of Raised Blood Pressure, controlled for a fixed characteristic Sex. This linear regression model, utilizing data spanning the period 1990 to 2015, will test the hypothesis that these established risk factors are significant predictors of diabetes rates. Our success criterion in the Modeling phase is to fit a robust model capable of explaining a substantial portion of the variance in Diabetes Prevalence, quantified by achieving a high goodness-of-fit with a R squared value greater than 0.50.

2- Data Understanding

The analysis is based on four seperate datasets sourced from the NCD Risk Factor Collaboration (NCD-RisC), representing national age-standardized estimates for major adult non-communicable disease risk factors. The files utilized are: Diabetes Prevalence (from Lancet 2024), Adult BMI (from Lancet 2024), Total Cholesterol (from Nature 2020), and Prevalence of Raised Blood Pressure (from Lancet 2017).

A crucial aspect of data understanding involves verifying the data's structure. All four files share the core identifiers of Country, Sex, and Year, which serves as the foundation for merging. The data is available across 200 countries and covers multiple years (ranging from the late 1970s/1980s up to 2022, depending on the file). Critically, all variables represent age-standardized mean values or prevalence rates for the adult population (18+ years), ensuring that direct comparisons can be made across different countries and time points without confounding due to varying age structures.

1.3- Data Preperation

We created a reliable, consistent dataset for the linear regression model. This process involved three major steps: selection and renaming, merging, and final filtering/transformation.

First, to resolve R's automatic renaming of column headers (which converts special characters like / and spaces to periods), the data was loaded, and the four key variables were explicitly selected and renamed for clarity: Diabetes_Prev, Obesity_Prev, Mean_Cholestrol, and Raised_BP_Prev.

```
# load all data files
diabetes_data <- read.csv('data/NCD_RisC_Lancet_2024_Diabetes_age_standardised_countries.csv')
BMI_data <- read.csv('data/NCD_RisC_Lancet_2024_BMI_age_standardised_country.csv')
Ch_data <- read.csv('data/NCD_RisC_Nature_2020_Cholesterol_age_standardised_countries.csv')
```

```r
BP_data <- read.csv('data/NCD_RisC_Lancet_2017_BP_age_standardised_countries.csv')

#filtering for needed columns
diabetes_filterd <- diabetes_data %>%
  select( Country = Country.Region.World,
          Year,
          Sex,
          Diabetes_Prev = Prevalence.of.diabetes..18..years.)

BMI_filtered <- BMI_data %>%
  select(Country = Country.Region.World,
         Year,
         Sex,
         Obesity_Prev = Prevalence.of.BMI..30.kg.m...obesity.)


Ch_filtered <- Ch_data %>%
  select(Country = Country.Region.World,
         Year,
         Sex,
         Mean_Cholestrol = Mean.total.cholesterol..mmol.L.)

BP_filtered <- BP_data %>%
  select(Country = Country.Region.World,
         Year,
         Sex,
         Raised_BP_Prev = Prevalence.of.raised.blood.pressure)
```

Second, the four separate data frames were joined into a single, wide-format dataset (df_merged_adult_ncd) using a series of left_join operations, ensuring the merge keys were consistent across the Year, Sex, and Country columns.

```r
df_merged_adult_ncd <- BMI_filtered %>%
  left_join(Ch_filtered, by = c("Year", "Sex", "Country")) %>%
  left_join(BP_filtered, by = c("Year", "Sex", "Country")) %>%
  left_join(diabetes_filterd, by = c("Year", "Sex", "Country"))
```

Finally, the resulting merged data was filtered to meet the modeling scope: we restricted the time range to 1990–2015 to ensure consistency across all primary data sources and removed the non-country aggregate row ("World"). The final step involved converting the categorical predictors, Sex and Country, into factor variables, which is necessary for their proper interpretation as fixed effects in the linear regression model. The final prepared dataset was saved as clean_data.

```r
#filter for year 1990 to 2015 and remove world row
clean_data <- df_merged_adult_ncd %>%
  filter(
    Year >= 1990 & Year <= 2015,
    #Country != "World"
  )

clean_data <- clean_data %>%
  mutate(
    Sex = as.factor(Sex),
```

```
    Country = as.factor(Country)
  )

cache('clean_data')
```

1.4- Modeling

The core of the analysis involves fitting a multiple linear regression model using the prepared data. This model is designed to test the strength of the relationship between established NCD risk factors and the prevalence of diabetes.

The chosen model uses Diabetes Prevalence as the response variable (Y), and all other NCD risk factors, along with categorical control variables, as the explanatory variables (X). The model is formulated as follows:

$$\mathbf{Y}_i = \beta_0 + \beta_1 \mathbf{X}_{i1} + \beta_2 \mathbf{X}_{i2} + \beta_3 \mathbf{X}_{i3} + \beta_4 \mathbf{X}_{i4} + \epsilon_i$$

Where: * i indexes the observation (Country and Year combination). * $\beta_1, \beta_2, \beta_3 \beta_4, \beta_5$ are the coefficients for the continuous risk factors. * $\mathbf{X}_1$ is Obesity $\mathbf{X}_2$ is the mean Cholestrol $\mathbf{X}_3$ is high Blood Pressure $\mathbf{X}_4$ is Sex * ϵ_i is the error term.

The model is implemented in R using the lm() function:

```
lm_ncd_fit <- lm(
  Diabetes_Prev ~ Obesity_Prev + Mean_Cholestrol + Raised_BP_Prev +
                  Sex,
  data = clean_data
)

summary(lm_ncd_fit)
```

```
##
## Call:
## lm(formula = Diabetes_Prev ~ Obesity_Prev + Mean_Cholestrol +
##     Raised_BP_Prev + Sex, data = clean_data)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.118280 -0.026076 -0.007667  0.021012  0.166599
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.2193055  0.0042834  51.199   <2e-16 ***
## Obesity_Prev     0.3499738  0.0030794 113.651   <2e-16 ***
## Mean_Cholestrol -0.0359888  0.0008406 -42.814   <2e-16 ***
## Raised_BP_Prev  -0.0091487  0.0081143  -1.127     0.26
## SexWomen        -0.0140776  0.0008377 -16.806   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.0382 on 10031 degrees of freedom
##   (364 observations deleted due to missingness)
## Multiple R-squared:  0.5749, Adjusted R-squared:  0.5747
## F-statistic:  3391 on 4 and 10031 DF,  p-value: < 2.2e-16
```

1.5- Evaluation

We assess the quality of the model against the objective defined in the Business Understanding phase (adjusted R square >0.50).

```
summary(lm_ncd_fit)$adj.r.squared
```

```
## [1] 0.5747141
```

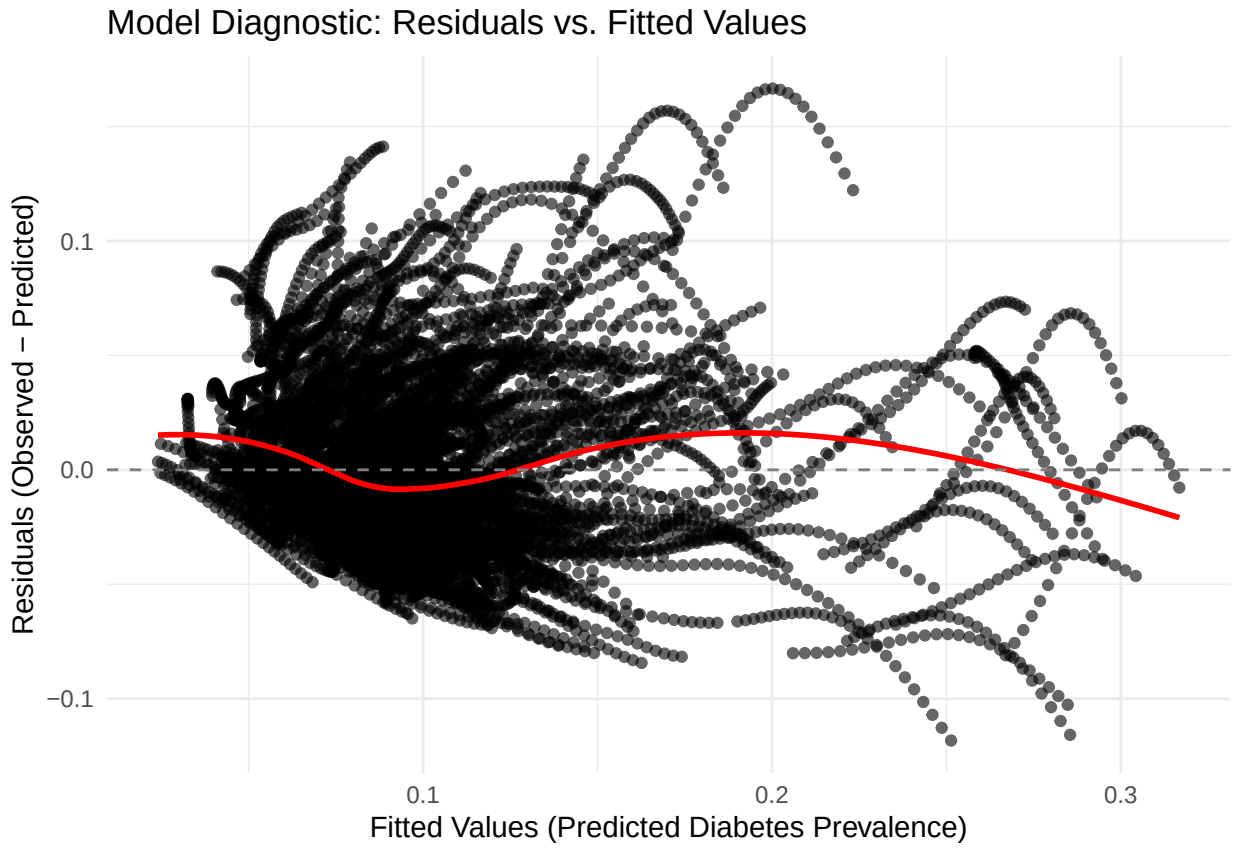
Upon fitting the linear regression model (excluding the Year and Country fixed effects to focus on the contemporaneous relationship between risk factors), the results demonstrate a good fit.

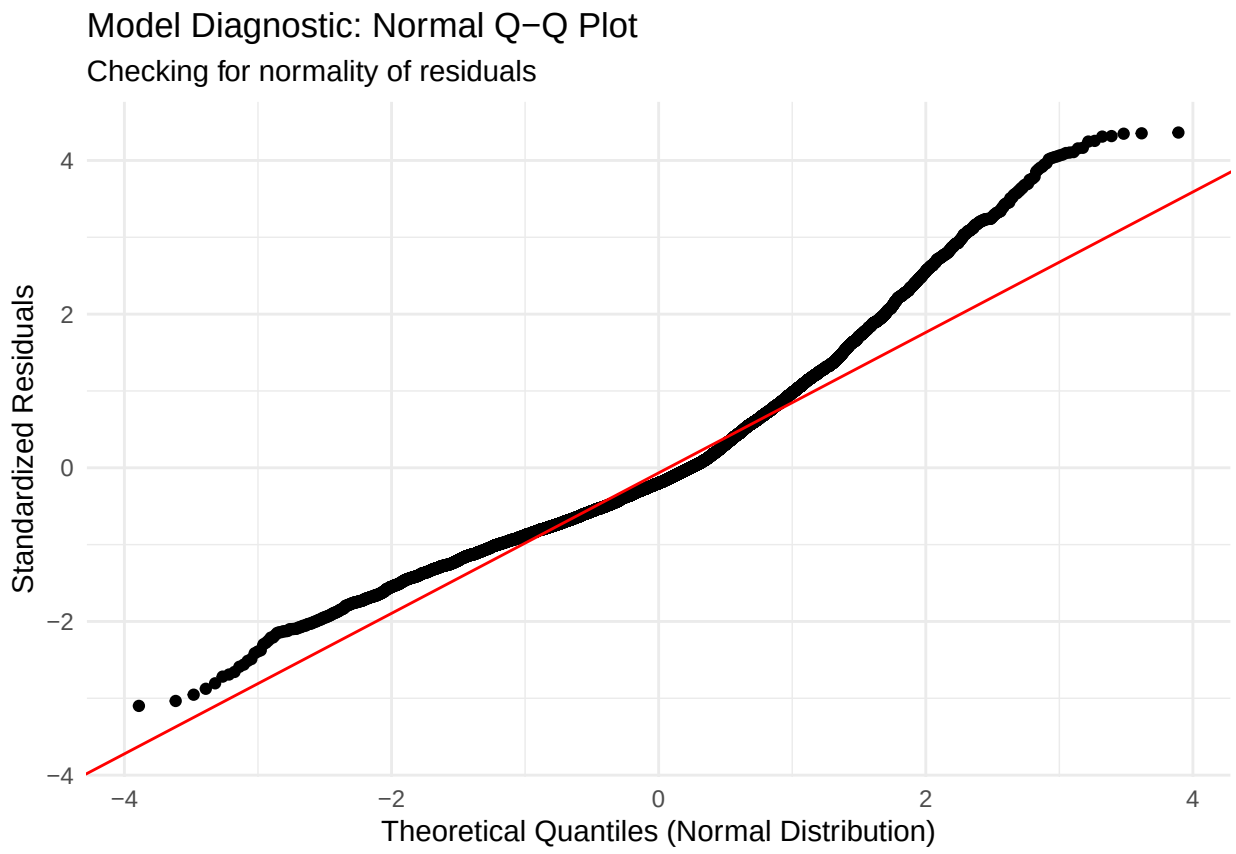
The model yielded an Adjusted R square of 0.5747 . This value meets the target of 0.50, indicating that 57.47% of the variance in global Diabetes Prevalence between 1990 and 2015 can be statistically explained by the combined effect of Obesity, Mean Total Cholesterol,Raised Blood Pressure, and Sex.

This outcome validates the core assumption that these established NCD risk factors are strong, quantifiable predictors of Diabetes Prevalence and provides a robust foundation for the final Deployment phase

1.6- Deployment

```
## `geom_smooth()` using formula = 'y ~ x'
```





CRISP-DM Cycle 2

2.1- Business Understanding

Following the successful completion of Cycle 1, which established that a combined model of risk factors could explain a substantial portion of the variance in diabetes (R-squared > 0.50), we identified a key limitation: the model assumes the relationships between these risk factors and diabetes are identical for both genders. The Sex variable was treated only as a fixed-effect predictor (an intercept shift), which may obscure clinically significant differences in disease pathways.

2.2- Data Understanding

The data understanding for Cycle 2 builds directly upon the foundation established in the previous cycle. The four core NCD-RisC datasets (Diabetes, BMI, Cholesterol, and BP) remain our primary sources.

While the data sources are unchanged, our understanding of them is refocused for the new objective. The critical variable Sex is present in all four datasets, containing two distinct levels: 'Men' and 'Women'. In Cycle 1, we understood this variable primarily as a necessary key for merging the data. In this cycle, we now re-evaluate it as our primary moderating variable.

2.3- Data Preperation

Since we already prepared the data we will use in the second cycle from the first cycle with the name "clean_data". Now we just have to classify them into two new sets; one for men with the name "clean_data_men" and the second set will be for women data with the name "clean_data_women".

```
#filter for year 1990 to 2015 and remove world row
clean_data_men <- clean_data %>%
  filter(
    Sex = "Men",

  )

clean_data_women <- clean_data %>%
  filter(
    Sex = "Women",

  )

cache('clean_data_men')
cache('clean_data_women')
```

Then we saved the new data sets in the cache to enhance the smoothness of our code.

2.4- Modeling

To test the hypothesis from our new Business Understanding—that the relationships themselves differ by gender. This method involves splitting the clean_data into two separate datasets (clean_data_men and clean_data_women) and fitting two independent linear models.

This allows each coefficient to be estimated separately for each gender, giving us two distinct models to compare.

```
lm_ncd_fit_men <- lm(
  Diabetes_Prev ~ Obesity_Prev + Mean_Cholestrol + Raised_BP_Prev,
  data = clean_data_men
)

summary(lm_ncd_fit_men)
```

```
##
## Call:
## lm(formula = Diabetes_Prev ~ Obesity_Prev + Mean_Cholestrol +
##     Raised_BP_Prev, data = clean_data_men)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.120026 -0.025793 -0.006984  0.020613  0.152514
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.204978   0.005183  39.549   <2e-16 ***
## Obesity_Prev     0.353329   0.004976  71.010   <2e-16 ***
```

```
## Mean_Cholestrol -0.025927   0.001213 -21.375   <2e-16 ***
## Raised_BP_Prev  -0.120864   0.012092  -9.995   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03788 on 5014 degrees of freedom
##   (182 observations deleted due to missingness)
## Multiple R-squared:  0.5371, Adjusted R-squared:  0.5369
## F-statistic:  1940 on 3 and 5014 DF,  p-value: < 2.2e-16
```

```
lm_ncd_fit_women <- lm(
  Diabetes_Prev ~ Obesity_Prev + Mean_Cholestrol + Raised_BP_Prev,
  data = clean_data_women
)

summary(lm_ncd_fit_women)
```

```
##
## Call:
## lm(formula = Diabetes_Prev ~ Obesity_Prev + Mean_Cholestrol +
##     Raised_BP_Prev, data = clean_data_women)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.113034 -0.025203 -0.007552  0.019877  0.171160
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      0.229269   0.007365  31.130  < 2e-16 ***
## Obesity_Prev     0.344458   0.003857  89.308  < 2e-16 ***
## Mean_Cholestrol -0.043538   0.001290 -33.746  < 2e-16 ***
## Raised_BP_Prev   0.041260   0.011886   3.471 0.000522 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.03764 on 5014 degrees of freedom
##   (182 observations deleted due to missingness)
## Multiple R-squared:  0.6227, Adjusted R-squared:  0.6224
## F-statistic:  2758 on 3 and 5014 DF,  p-value: < 2.2e-16
```

```
# 1. Calculate the mean of each NCD for Men
# We also use pivot_longer() to turn the data from wide to long
# This is the format ggplot needs for plotting
men_means <- clean_data_men %>%
  summarise(
    #Diabetes = mean(Diabetes_Prev, na.rm = TRUE),
    Obesity_mean1 = mean(Obesity_Prev, na.rm = TRUE),
    Cholesterol_mean1 = mean(Mean_Cholestrol, na.rm = TRUE),
    Blood_Pressure_mean1 = mean(Raised_BP_Prev, na.rm = TRUE)
  ) %>%
  pivot_longer(
    cols = everything(),
    names_to = "NCD_Metric",
```

```
    values_to = "Mean_Value"
  ) %>%
  mutate(Sex = "Men")

# 2. Do the exact same thing for Women
women_means <- clean_data_women %>%
  summarise(
    #Diabetes = mean(Diabetes_Prev, na.rm = TRUE),
    Obesity_mean1 = mean(Obesity_Prev, na.rm = TRUE),
    Cholesterol_mean1 = mean(Mean_Cholestrol, na.rm = TRUE),
    Blood_Pressure_mean1 = mean(Raised_BP_Prev, na.rm = TRUE)
  ) %>%
  pivot_longer(
    cols = everything(),
    names_to = "NCD_Metric",
    values_to = "Mean_Value"
  ) %>%
  mutate(Sex = "Women")

# 3. Combine the two summary tables into one
comparison_data <- bind_rows(men_means, women_means)

# 4. Now, create the plot
ggplot(comparison_data, aes(x = NCD_Metric, y = Mean_Value, fill = Sex)) +

  # geom_bar() creates the bar chart
  # stat = "identity" means "use the value in the 'y' column as the height"
  # position = "dodge" is the command that places the bars side-by-side
  geom_bar(stat = "identity", position = "dodge") +

  # Add clear labels
  labs(
    title = "Mean NCD Risk Factor Prevalence by Gender (1990-2015)",
    x = "NCD Risk Factor",
    y = "Mean Prevalence / Value",
    fill = "Gender"
  ) +

  # Use a clean theme
  theme_minimal() +

  # (Optional) Clean up the x-axis labels
  scale_x_discrete(labels = c(
    "Blood_Pressure" = "Raised BP",
    "Cholesterol" = "Mean Cholesterol",
    "Diabetes" = "Diabetes",
    "Obesity" = "Obesity"
  ))
```

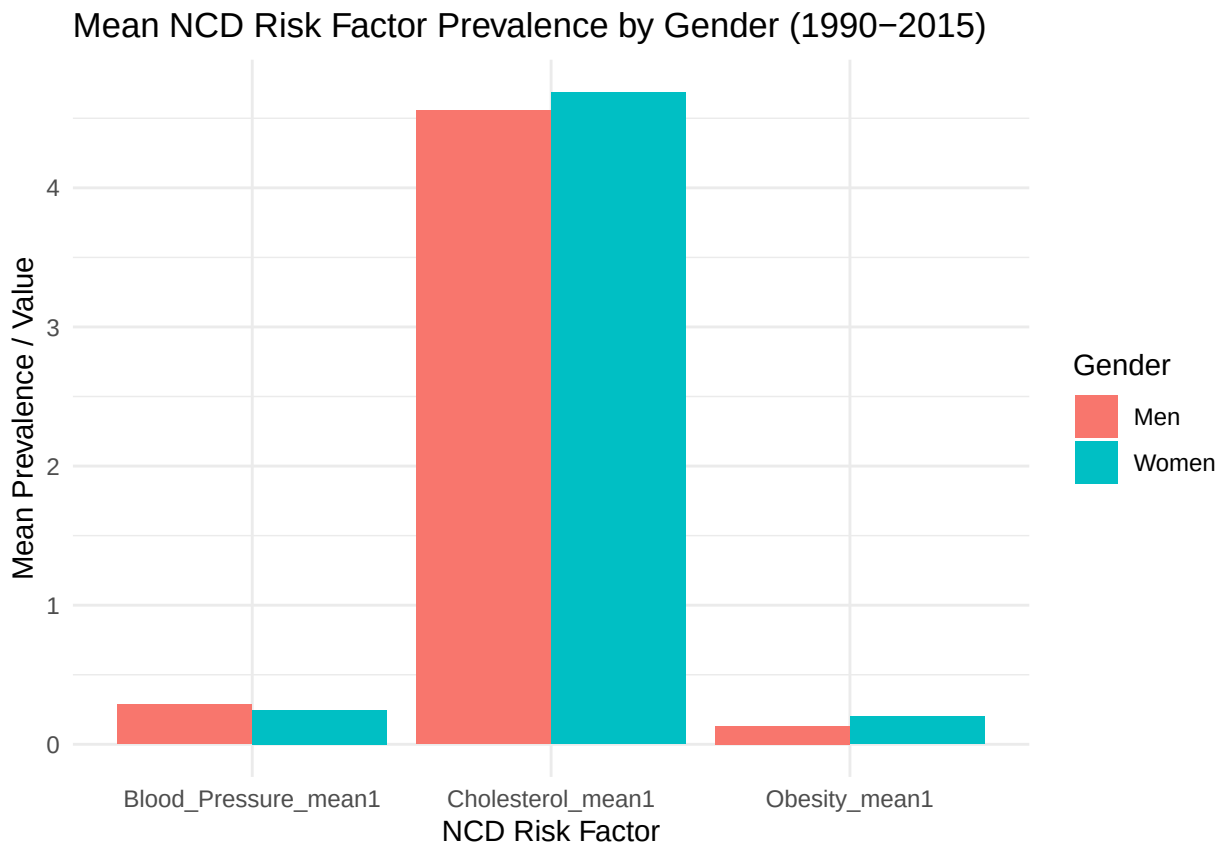
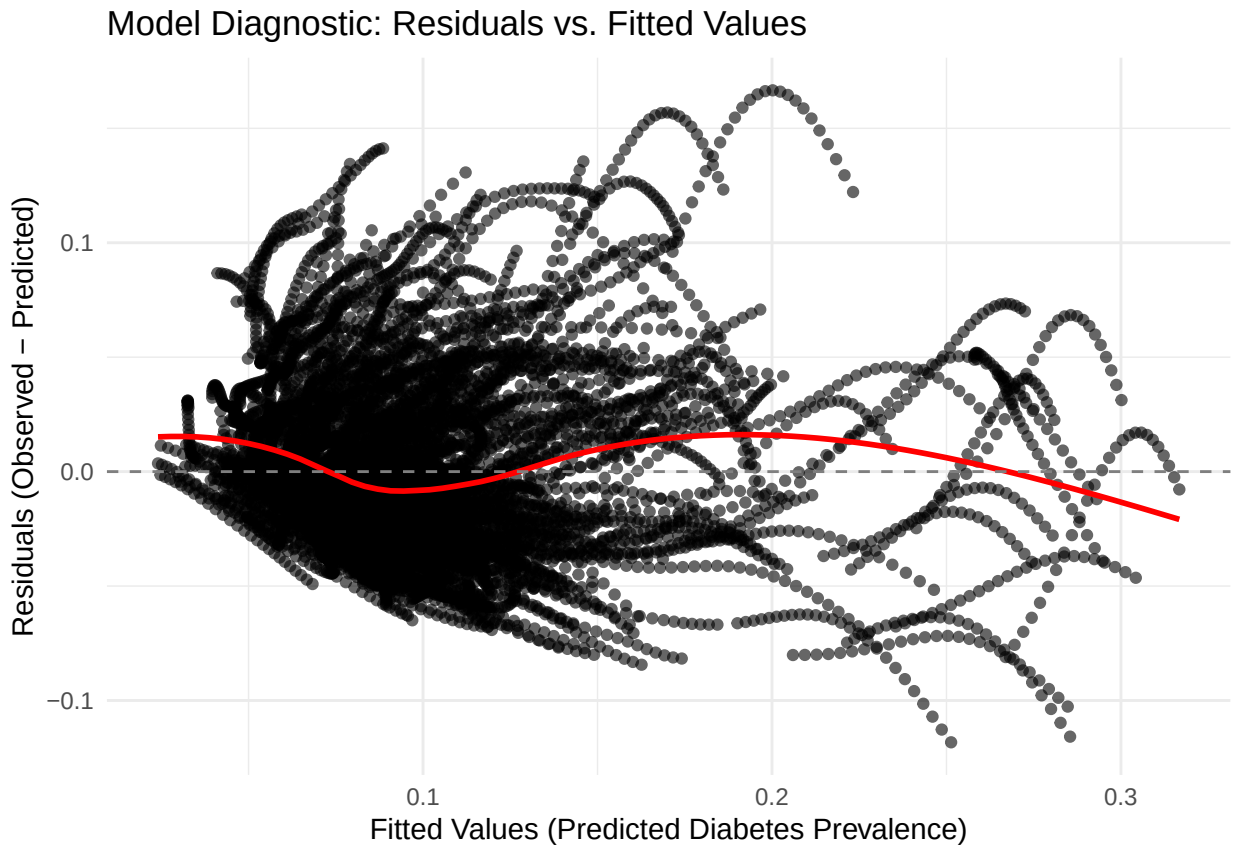


Figure: this figure shows a comparison in the three main factors for Diabetes Prevalence.

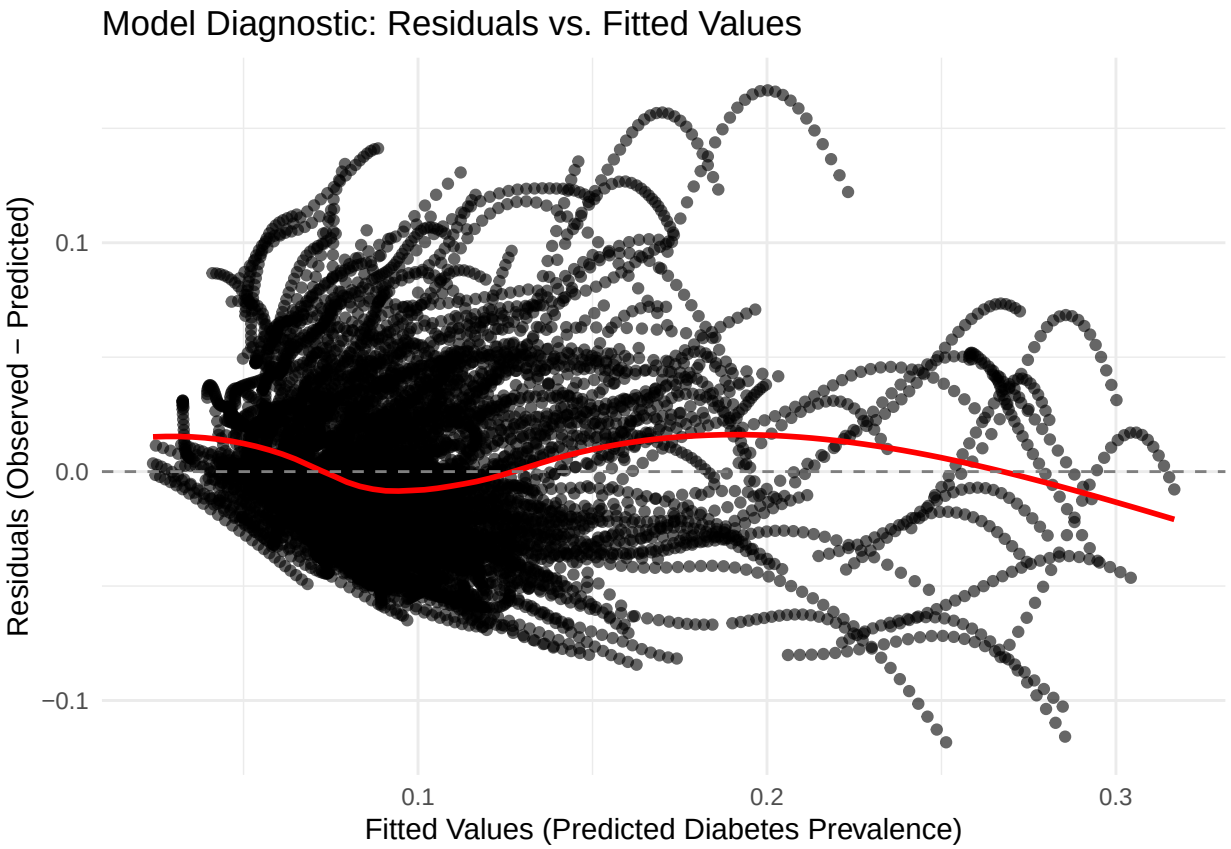
2.5- Evaluation

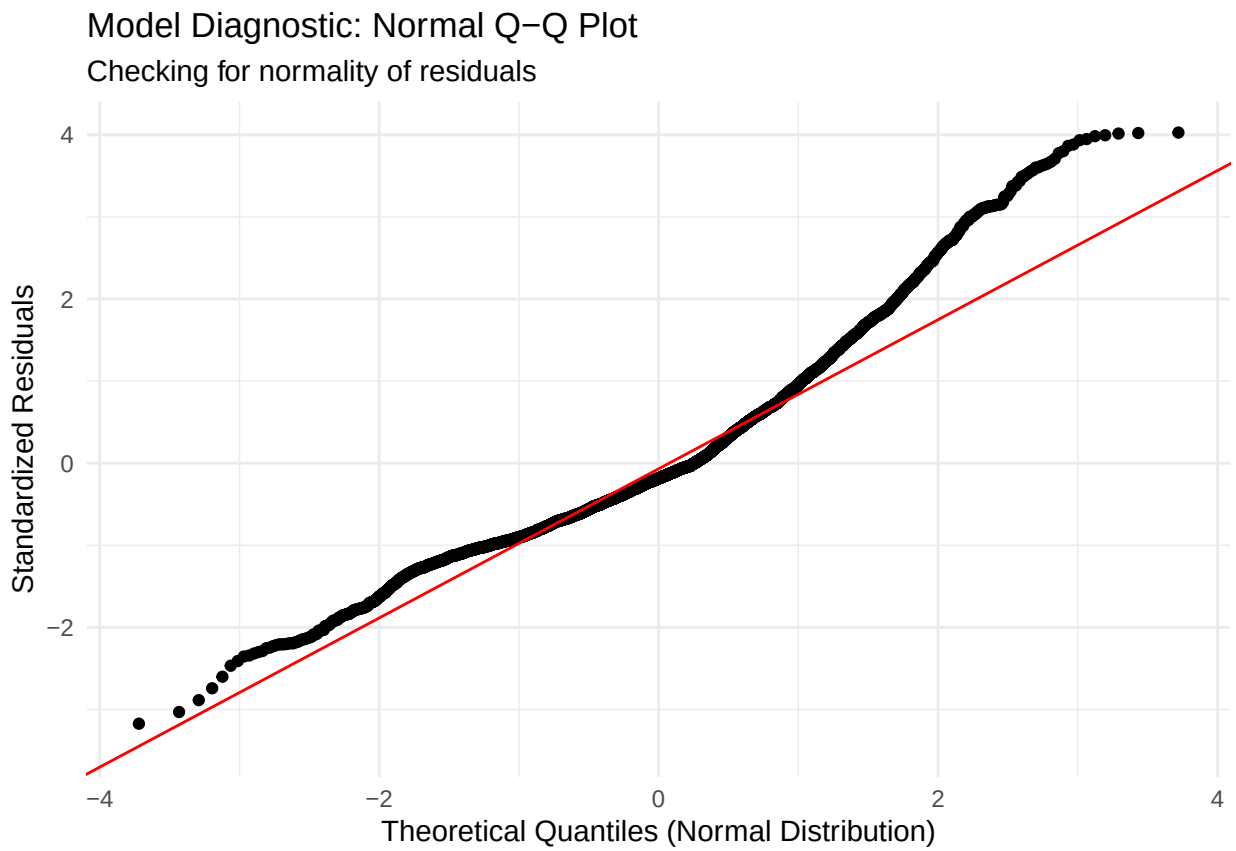
2.6- Deployment

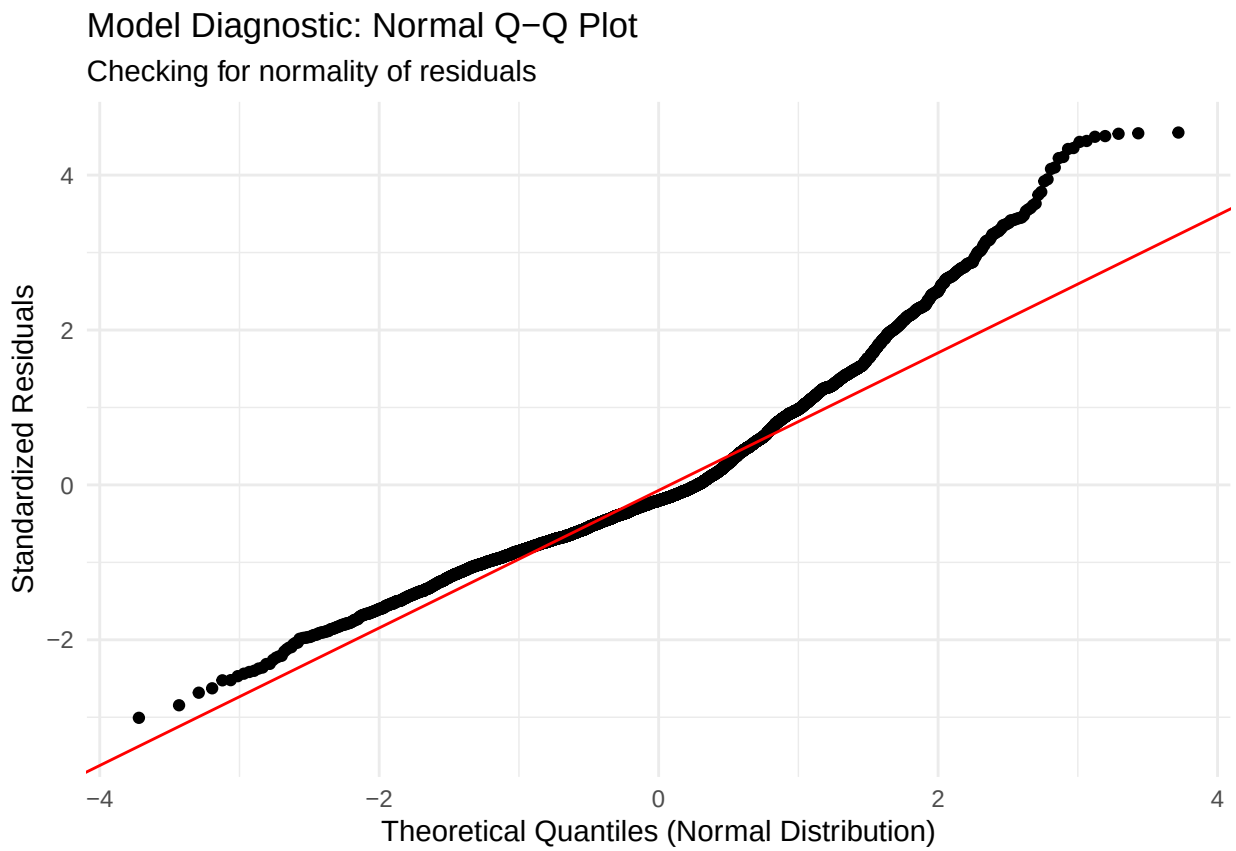
```
## `geom_smooth()` using formula = 'y ~ x'
```




```
## `geom_smooth()` using formula = 'y ~ x'
```


Model Diagnostic: Residuals vs. Fitted Values
Residuals (Observed − Predicted)
0.1
0.0
−0.1
0.1
0.2
0.3
Fitted Values (Predicted Diabetes Prevalence)


Model Diagnostic: Normal Q−Q Plot
Checking for normality of residuals
Standardized Residuals
4
2
0
−2
−4
−2
0
2
4
Theoretical Quantiles (Normal Distribution)


Model Diagnostic: Normal Q−Q Plot
Checking for normality of residuals
Standardized Residuals
4
2
0
−2
−4
−2
0
2
4
Theoretical Quantiles (Normal Distribution)