HW Solution 8to10 KB

Kimberly Brooks

2025-11-10

1.

Ans: My interests lie in mostly clinical research, and with the prevalence of diabetes I thought it would be interesting to see the strongest predictors for diabetes given certain lifestyle factors.

My General Questions: What are the strongest modifiable lifestyle factors that increase someone's risk for diabetes type 2 .

Do people who pay less attention to lifestyle factors like sleep quality, diet score and screen time have an increased risk for developing diabetes (i.e increases diabetes risk score)?

2.

The data set obtained from kaggle titled "Health & Lifestyle Data for Diabetes Prediction" is intended for diabetes risk prediction. Each record in the data set reflects an individual's health profile with attributes combining demographics,lifestyle,behavioral,family history and physiological measures of health.

3. Dependent Variable : Diabetes risk score

Higher diabetes risk score is associated with a higher risk of developing diabetes.I am interested to see the effects of modifiable lifestyle risk factors like screen time and sleep score on diabetes risk.

4. Independent variables: Modifiable Risk factors education level, income level, employment, diet score, sleep hours per day, screen time hours per day,alcohol consumption per week, smoking status I will include one Non Modifiable risk factor i.e. Age, for completeness

Hypotheses: a. Persons with higher income levels are at less risk for diabetes compared to those with lower incomes b. Persons with higher education levels are at less risk for diabetes compared to those with less than a high school education. c. Persons who get less sleep are at an increased risk for diabetes than those who get a normal amount of sleep. d. There is a relationship between screen time hours per day and diabetes risk e. I think income level and education level may interact with each other f. Overall I think diet score and smoking status may have the highest impact on diabetes risk score in this model

5. Have to create dummy variables for most of these as they are categorical Units: Screen time in hours Alcohol consumption per week in volume Sleep per day in hours

```
diabetes_data <-read.csv("/Users/kimberlybrooks/Desktop/Diabetes_and_LifeStyle_Dataset\ .csv",
                         stringsAsFactors = FALSE ) #load diabetes data set in R

head(diabetes_data) #viewing my data set
```

```
##   Age gender ethnicity education_level income_level employment_status
## 1  58   Male     Asian      Highschool Lower-Middle          Employed
## 2  52 Female     White      Highschool       Middle          Employed
## 3  60   Male  Hispanic      Highschool       Middle        Unemployed
## 4  74 Female     Black      Highschool          Low           Retired
## 5  46   Male     White        Graduate       Middle           Retired
## 6  46 Female     White      Highschool Upper-Middle          Employed
##   smoking_status alcohol_consumption_per_week
## 1          Never                            0
## 2         Former                            1
## 3          Never                            1
## 4          Never                            0
## 5          Never                            1
## 6          Never                            2
##   physical_activity_minutes_per_week diet_score sleep_hours_per_day
## 1                                215        5.7                 7.9
## 2                                143        6.7                 6.5
## 3                                 57        6.4                10.0
## 4                                 49        3.4                 6.6
## 5                                109        7.2                 7.4
## 6                                124        9.0                 6.2
##   screen_time_hours_per_day family_history_diabetes hypertension_history
## 1                       7.9                       0                    0
## 2                       8.7                       0                    0
## 3                       8.1                       1                    0
## 4                       5.2                       0                    0
## 5                       5.0                       0                    0
## 6                       5.4                       0                    0
##   cardiovascular_history  bmi waist_to_hip_ratio systolic_bp diastolic_bp
## 1                      0 30.5              0.89         134           78
## 2                      0 23.1              0.80         129           76
## 3                      0 22.2              0.81         115           73
## 4                      0 26.8              0.88         120           93
## 5                      0 21.2              0.78          92           67
## 6                      0 26.1              0.85          95           81
##   heart_rate cholesterol_total hdl_cholesterol ldl_cholesterol triglycerides
## 1         68               239              41             160           145
## 2         67               116              55              50            30
## 3         74               213              66              99            36
## 4         68               171              50              79           140
## 5         67               210              52             125           160
## 6         57               218              61             119           179
##   glucose_fasting glucose_postprandial insulin_level hba1c diabetes_risk_score
## 1             136                  236         6.36  8.18                29.6
## 2              93                  150         2.00  5.63                23.0
## 3             118                  195         5.07  7.51                44.7
## 4             139                  253         5.28  9.03                38.2
## 5             137                  184        12.74  7.20                23.5
## 6             100                  133         8.77  6.03                23.5
##   diabetes_stage diagnosed_diabetes
## 1         Type 2                 1
## 2    No Diabetes                 0
## 3         Type 2                 1
## 4         Type 2                 1
```

```
## 5         Type 2                  1
## 6   Pre-Diabetes                  0
```

```r
str(diabetes_data) #checking structure of the data
```

```
## 'data.frame':    97297 obs. of  31 variables:
##  $ Age                              : int  58 52 60 74 46 46 75 62 37 59 ...
##  $ gender                           : chr  "Male" "Female" "Male" "Female" ...
##  $ ethnicity                        : chr  "Asian" "White" "Hispanic" "Black" ...
##  $ education_level                  : chr  "Highschool" "Highschool" "Highschool" "Highschool" ...
##  $ income_level                     : chr  "Lower-Middle" "Middle" "Middle" "Low" ...
##  $ employment_status                : chr  "Employed" "Employed" "Unemployed" "Retired" ...
##  $ smoking_status                   : chr  "Never" "Former" "Never" "Never" ...
##  $ alcohol_consumption_per_week     : int  0 1 1 0 1 2 0 1 1 3 ...
##  $ physical_activity_minutes_per_week: int  215 143 57 49 109 124 53 75 114 86 ...
##  $ diet_score                       : num  5.7 6.7 6.4 3.4 7.2 9 9.2 4.1 6.7 8.2 ...
##  $ sleep_hours_per_day              : num  7.9 6.5 10 6.6 7.4 6.2 7.8 9 8.5 5.3 ...
##  $ screen_time_hours_per_day        : num  7.9 8.7 8.1 5.2 5 5.4 8 12.9 8.5 7.4 ...
##  $ family_history_diabetes          : int  0 0 1 0 0 0 0 0 0 0 ...
##  $ hypertension_history             : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ cardiovascular_history           : int  0 0 0 0 0 0 0 1 1 0 ...
##  $ bmi                              : num  30.5 23.1 22.2 26.8 21.2 26.1 25.1 23.9 24.7 26.7 ...
##  $ waist_to_hip_ratio               : num  0.89 0.8 0.81 0.88 0.78 0.85 0.88 0.86 0.84 0.81 ...
##  $ systolic_bp                      : int  134 129 115 120 92 95 129 128 103 124 ...
##  $ diastolic_bp                     : int  78 76 73 93 67 81 77 83 71 81 ...
##  $ heart_rate                       : int  68 67 74 68 67 57 81 76 72 70 ...
##  $ cholesterol_total                : int  239 116 213 171 210 218 238 241 187 188 ...
##  $ hdl_cholesterol                  : int  41 55 66 50 52 61 46 49 33 52 ...
##  $ ldl_cholesterol                  : int  160 50 99 79 125 119 161 159 132 103 ...
##  $ triglycerides                    : int  145 30 36 140 160 179 155 120 98 104 ...
##  $ glucose_fasting                  : int  136 93 118 139 137 100 101 110 116 76 ...
##  $ glucose_postprandial             : int  236 150 195 253 184 133 100 189 172 109 ...
##  $ insulin_level                    : num  6.36 2 5.07 5.28 12.74 ...
##  $ hba1c                            : num  8.18 5.63 7.51 9.03 7.2 6.03 5.24 7.04 6.9 4.99 ...
##  $ diabetes_risk_score              : num  29.6 23 44.7 38.2 23.5 23.5 36.1 34.2 26.7 30 ...
##  $ diabetes_stage                   : chr  "Type 2" "No Diabetes" "Type 2" "Type 2" ...
##  $ diagnosed_diabetes               : int  1 0 1 1 1 0 0 1 1 0 ...
```

```r
# create dummy variables: education level, income level, employment, gender, smoking

#gender
diabetes_data$Male <- ifelse(diabetes_data$gender == "Male", 1, 0)
diabetes_data$Other <- ifelse(diabetes_data$gender == "Other", 1, 0)

# Income level dummies
diabetes_data$Income_LowerMiddle <- ifelse(diabetes_data$income_level == "Lower-Middle", 1, 0)
diabetes_data$Income_Middle <- ifelse(diabetes_data$income_level == "Middle", 1, 0)
diabetes_data$Income_UpperMiddle <- ifelse(diabetes_data$income_level == "Upper-Middle", 1, 0)
diabetes_data$Income_High <- ifelse(diabetes_data$income_level == "High", 1, 0)
diabetes_data$Income_Low <- ifelse(diabetes_data$income_level == "Low", 1, 0)


#Education dummies
```

```r
diabetes_data$Education_Graduate <- ifelse(diabetes_data$education_level == "Graduate", 1, 0)
diabetes_data$Education_Postgraduate <- ifelse(diabetes_data$education_level == "Postgraduate", 1, 0)
diabetes_data$Education_Highschool <- ifelse(diabetes_data$education_level == "Highschool", 1, 0)
diabetes_data$Education_Noformal <- ifelse(diabetes_data$education_level == "No formal", 1, 0)


#employment level dummies
diabetes_data$Employed <- ifelse(diabetes_data$employment_status == "Employed", 1, 0)
diabetes_data$Unemployed <- ifelse(diabetes_data$employment_status == "Unemployed", 1, 0)
diabetes_data$Retired <- ifelse(diabetes_data$employment_status == "Retired", 1, 0)
diabetes_data$employment_status_Student <- ifelse(diabetes_data$employment_status == "Student", 1, 0)

#Smoking dummies
diabetes_data$smoking_status_Former <- ifelse(diabetes_data$smoking_status == "Former", 1, 0)
diabetes_data$smoking_status_Current <- ifelse(diabetes_data$smoking_status == "Current", 1, 0)
diabetes_data$smoking_status_Never <- ifelse(diabetes_data$smoking_status == "Never", 1, 0)
```

6. In this simple regression, there is no evidence of an association between low income status or sleep hours per day and diabetes risk score, both are not statistically significant.But we can assume that income may have a greater effect in the multiple regression when we control other variables

```r
bivarate1 <- lm(diabetes_risk_score ~ sleep_hours_per_day, data = diabetes_data)
summary(bivarate1)
```

```
##
## Call:
## lm(formula = diabetes_risk_score ~ sleep_hours_per_day, data = diabetes_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -27.501  -6.427  -1.261   5.383  37.015
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         30.03865    0.18820 159.613   <2e-16 ***
## sleep_hours_per_day  0.02621    0.02657   0.986    0.324
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.065 on 97295 degrees of freedom
## Multiple R-squared:  9.997e-06,  Adjusted R-squared:  -2.811e-07
## F-statistic: 0.9726 on 1 and 97295 DF,  p-value: 0.324
```

```r
bivarate2 <- lm(diabetes_risk_score ~ Income_Low, data = diabetes_data)
summary(bivarate2)
```

```
##
## Call:
## lm(formula = diabetes_risk_score ~ Income_Low, data = diabetes_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -27.522  -6.422  -1.225   5.378  36.975
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.221560   0.031489 959.744   <2e-16 ***
## Income_Low   0.003152   0.081790   0.039    0.969
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.065 on 97295 degrees of freedom
## Multiple R-squared:  1.527e-08,  Adjusted R-squared:  -1.026e-05
## F-statistic: 0.001485 on 1 and 97295 DF,  p-value: 0.9693
```

7.

Age has the largest impact (beta = 0.290, meaning each decade of life increases risk by ~3 points), followed by diet score (beta = -0.729, so improving diet by 5 points could reduce risk by ~3.6 points) and screen time (beta = 0.270, so reducing screen time by 10 hours/day could lower risk by 2.7 points).

The most modifiable variables are diet score, screen time, and smoking status however these behavioral interventions would likely show cumulative benefits over time rather than immediate changes just based on real world scenarios.

```
mr1 <- lm(diabetes_risk_score ~ Age + education_level + income_level + employment_status +
            diet_score + screen_time_hours_per_day + sleep_hours_per_day +smoking_status +
            alcohol_consumption_per_week,  data=diabetes_data)
```

```
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2022). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.3. https://CRAN.R-project.org/package=stargazer
```

```
stargazer(mr1, type = "latex", no.space=TRUE,
          dep.var.labels=c("Diabetes Risk Score"),
          covariate.labels=c("Age","Education Level: Highschool",
                             "Education Level: No formal",
                             "Education Level: Postgraduate",
                             "Income: Low",
                             "Income: Lower-Middle",
                             "Income: Middle",
                             "Income: Upper-Middle",
                             "Employment: Retired",
                             "Employment: Student",
                             "Employment: Unemployed",
                             "Diet Score",
                             "Screen Time Per Day (Hours)",
                             "Sleep Hours Per Day",
                             "Smoking Status (Former)",
```

```
                        "Smoking Status (Never)",
                        "Alcohol Consumption Per Week"),
          omit.stat=c("LL","ser","f"),header = FALSE)
```

Table 1:

	Dependent variable:
	Diabetes Risk Score
Age	0.290^{***}
	(0.002)
Education Level: Highschool	-0.037
	(0.056)
Education Level: No formal	-0.023
	(0.117)
Education Level: Postgraduate	0.044
	(0.077)
Income: Low	0.243^{*}
	(0.128)
Income: Lower-Middle	0.327^{***}
	(0.122)
Income: Middle	0.252^{**}
	(0.119)
Income: Upper-Middle	0.288^{**}
	(0.124)
Employment: Retired	-0.012
	(0.062)
Employment: Student	-0.115
	(0.105)
Employment: Unemployed	-0.072
	(0.079)
Diet Score	-0.729^{***}
	(0.014)
Screen Time Per Day (Hours)	0.270^{***}
	(0.010)
Sleep Hours Per Day	0.040^{*}
	(0.023)
Smoking Status (Former)	-0.099
	(0.078)
Smoking Status (Never)	-0.127^{**}
	(0.064)
Alcohol Consumption Per Week	0.065^{***}
	(0.018)
Constant	17.892^{***}
	(0.244)
Observations	97,297
R^2	0.271
Adjusted R^2	0.271
Note:	*p<0.1; **p<0.05; ***p<0.01

8. In comparing the simple and multiple regressions, I noticed that income levels became statistically significant in the multiple regression when controlling for other factors, whereas they showed little

to no association in simple regressions.I assume that this could be an due to supression rather than interaction.

9.

Hypothesis (Income): This hypothesis was not supported as lower and middle-income groups actually had higher diabetes risk scores compared to the high-income reference group (e.g., Lower-Middle beta = 0.327, p = 0.007), which is the opposite of what I predicted.

Hypothesis (Education): This hypothesis was not supported as education level showed no significant association with diabetes risk at all (all $p > 0.05$), suggesting that education does not independently predict diabetes risk when controlling for other lifestyle factors.

Hypothesis (Sleep): This hypothesis was partially supported as sleep hours showed a positive association with diabetes risk (beta = 0.040, p = 0.080), meaning more sleep was marginally associated with higher risk, which is the opposite direction of what I expected, though the effect was only marginally significant.

Hypothesis (Screen time): This hypothesis was strongly supported, screen time had a significant positive association with diabetes risk (beta = 0.270, $p < 0.001$), confirming that more screen time is associated with higher diabetes risk.

Hypothesis (Diet and smoking impact): This was partially supported - diet score had the largest modifiable effect (beta = -0.729, $p < 0.001$), but smoking status showed only modest effects with "never smoking" being protective (beta = -0.127, p = 0.047), making diet clearly more impactful than smoking in this model; however, Age actually had the largest overall coefficient (beta = 0.290), though it's non-modifiable.

10.

The R squared (0.2709) and adjusted R squared (0.2708) tells us that the model explains approximately 27% of the variation in diabetes risk scores, meaning that 73% of the variation remains unexplained by the variables I included.

11.

```r
#removing those with p value over 0.05 (education,employment, sleep hours)
mr3 <- lm(diabetes_risk_score ~ Age + Income_High + Income_UpperMiddle + Income_Middle +
            Income_LowerMiddle+ diet_score + screen_time_hours_per_day
          +smoking_status_Current +smoking_status_Former +
            alcohol_consumption_per_week,  data=diabetes_data)
summary(mr3)
```

```
##
## Call:
## lm(formula = diabetes_risk_score ~ Age + Income_High + Income_UpperMiddle +
##     Income_Middle + Income_LowerMiddle + diet_score + screen_time_hours_per_day +
##     smoking_status_Current + smoking_status_Former + alcohol_consumption_per_week,
##     data = diabetes_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.901  -5.053  -1.844   2.644  24.599
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  18.257224   0.151453 120.547  < 2e-16 ***
```

```
## Age                            0.289532   0.001602 180.743  < 2e-16 ***
## Income_High                   -0.241131   0.128394  -1.878 0.060377 .
## Income_UpperMiddle             0.045148   0.085174   0.530 0.596070
## Income_Middle                  0.009350   0.076853   0.122 0.903167
## Income_LowerMiddle             0.084945   0.081266   1.045 0.295899
## diet_score                    -0.729087   0.013942 -52.295  < 2e-16 ***
## screen_time_hours_per_day      0.270119   0.010049  26.881  < 2e-16 ***
## smoking_status_Current         0.126900   0.063938   1.985 0.047176 *
## smoking_status_Former          0.027645   0.064066   0.432 0.666103
## alcohol_consumption_per_week  0.065272   0.017501   3.730 0.000192 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.741 on 97286 degrees of freedom
## Multiple R-squared:  0.2709, Adjusted R-squared:  0.2708
## F-statistic:  3615 on 10 and 97286 DF,  p-value: < 2.2e-16
```

```
#ran the regression again removing income which became not significant after removing the previous vari
mr4 <- lm(diabetes_risk_score ~ Age + diet_score + screen_time_hours_per_day
          +smoking_status_Current +smoking_status_Former
          + alcohol_consumption_per_week,data=diabetes_data)
summary(mr4)
```

```
##
## Call:
## lm(formula = diabetes_risk_score ~ Age + diet_score + screen_time_hours_per_day +
##     smoking_status_Current + smoking_status_Former + alcohol_consumption_per_week,
##     data = diabetes_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.838  -5.057  -1.845   2.639  24.662
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  18.276643   0.139160 131.335  < 2e-16 ***
## Age                           0.289528   0.001602 180.739  < 2e-16 ***
## diet_score                   -0.728942   0.013942 -52.285  < 2e-16 ***
## screen_time_hours_per_day     0.270316   0.010048  26.901  < 2e-16 ***
## smoking_status_Current        0.127291   0.063939   1.991  0.04650 *
## smoking_status_Former         0.028419   0.064066   0.444  0.65734
## alcohol_consumption_per_week  0.065311   0.017501   3.732  0.00019 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.741 on 97290 degrees of freedom
## Multiple R-squared:  0.2708, Adjusted R-squared:  0.2708
## F-statistic:  6023 on 6 and 97290 DF,  p-value: < 2.2e-16
```

Answer :

I used backward elimination with alpha = 0.05 as the retention criterion. This method would lead me to eliminate all income variables, as they became non-significant ($p > 0.05$) after removing education, employment, and sleep hours. However, I disagree with completely removing income as it is a well-established

social determinant of health and diabetes risk in the literature, regardless of statistical significance in this particular model.Also removing the income variables did not change R squared (remained 0.2708), suggesting they neither help nor hurt prediction in this case

12.

Overall Interpretation :

Diet score seems to be the most impactful non modifiable risk factor in terms of its contribution to overall diabetic risk. While income level and education level were hypothesized to be a contributor to higher diabetes risk only income level more so Lower Middle income was the most significant for that group. Screen time was also pretty significant in this model and had a negative effect in terms of increased screen time leading to higher risk of diabetes. I however did also control for one non modifiable risk factor i.e. age for completeness which (not shown here) increased the R squared signficantly and was the most signficant positive contributor to diabetes risk. Therefore a potential weakness in this model is that known modifiable risk factors like age and gender may play a more significant overall picture when predicting diabetes risk, which is also shown in the literature. For this particular model possibly adding physiological factors as well like waist-hip-ratio, bmi etc could also possibly strengthen the models predictive capability.

13. Calculations (using R):

a. Derive the coefficients from your regression using the formula. (If you run into problems using solve(), try using ginv() instead, which does the same thing but is a bit more robust.)

```
mrmatrix <- as.matrix(cbind(
  diabetes_data$Age,
  diabetes_data$Education_Highschool,
  diabetes_data$Education_Noformal,
  diabetes_data$Education_Postgraduate,
  diabetes_data$Income_Low,
  diabetes_data$Income_LowerMiddle,
  diabetes_data$Income_Middle,
  diabetes_data$Income_UpperMiddle,
  diabetes_data$Retired,
  diabetes_data$employment_status_Student,
  diabetes_data$Unemployed,
  diabetes_data$diet_score,
  diabetes_data$screen_time_hours_per_day,
  diabetes_data$sleep_hours_per_day,
  diabetes_data$smoking_status_Former,
  diabetes_data$smoking_status_Never,
  diabetes_data$alcohol_consumption_per_week
))

mrmatrix <- cbind(1, mrmatrix)

beta_hat <- solve( t(mrmatrix) %*% mrmatrix ) %*% t(mrmatrix) %*% diabetes_data$diabetes_risk_score
beta_hat
```

```
##               [,1]
##  [1,] 17.89244707
##  [2,]  0.28954737
##  [3,] -0.03705278
```

```
##  [4,] -0.02314017
##  [5,]  0.04361811
##  [6,]  0.24301203
##  [7,]  0.32738216
##  [8,]  0.25191134
##  [9,]  0.28766600
## [10,] -0.01159256
## [11,] -0.11453599
## [12,] -0.07220681
## [13,] -0.72899365
## [14,]  0.27001106
## [15,]  0.03967217
## [16,] -0.09874647
## [17,] -0.12672899
## [18,]  0.06534021
```

b. For one of the coefficients, confirm its p value as shown in the regression output using the coefficient, its standard error, and pt() in R.

```
# Using sleep_hours_per_day which has p = 0.080418
coef_sleep <- 0.039672
se_sleep <- 0.022692
t_stat_sleep <- coef_sleep / se_sleep
p_value_sleep <- 2 * pt(abs(t_stat_sleep), df = 97279, lower.tail = FALSE)
p_value_sleep
```

```
## [1] 0.08041848
```

c. Calculate the R2 and adjusted R2 using R, and confirm that your results match the regression output.

```
y <- diabetes_data$diabetes_risk_score
ypred <- mrmatrix %*% beta_hat

tss <- sum((y - mean(y))^2)
sse <- sum((y - ypred)^2)
r2 <- 1 - (sse/tss)
r2
```

```
## [1] 0.2709399
```

d. Calculate the F statistic using R and confirm it against the regression output.

```
#first calculating the adjusted R2
n <- length(y)
k <- ncol(mrmatrix)-1
dft <- n - 1
dfe <- n - k - 1
(tss/dft - sse/dfe)/ (tss/dft)
```

```
## [1] 0.2708125
```

```r
#now calculating the f statistic
f <- (r2/k) / ((1-r2)/(n-k-1))
f
```

```
## [1] 2126.571
```

14. Add at least one quadratic term into your model and interpret the results. Is it significant? What is the effect of a 1-unit increase in that variable at its mean value?

Sleep hours does not have a meaningful/ statistically significant relationship with diabetes risk in this model.

```r
#adding sleep hours per day as quadratic term
quadratic <- lm(diabetes_risk_score ~ Age + education_level + income_level + employment_status
                + diet_score + screen_time_hours_per_day + I(sleep_hours_per_day^2) + +smoking_status
                + sleep_hours_per_day + alcohol_consumption_per_week,  data=diabetes_data)
summary(quadratic)
```

```
##
## Call:
## lm(formula = diabetes_risk_score ~ Age + education_level + income_level +
##     employment_status + diet_score + screen_time_hours_per_day +
##     I(sleep_hours_per_day^2) + +smoking_status + sleep_hours_per_day +
##     alcohol_consumption_per_week, data = diabetes_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.895  -5.059  -1.842   2.646  24.580
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  17.131192   0.759586  22.553  < 2e-16 ***
## Age                           0.289542   0.001602 180.744  < 2e-16 ***
## education_levelHighschool    -0.036979   0.055949  -0.661 0.508654
## education_levelNo formal     -0.022840   0.117258  -0.195 0.845563
## education_levelPostgraduate   0.043836   0.076653   0.572 0.567413
## income_levelLow               0.242336   0.128403   1.887 0.059123 .
## income_levelLower-Middle      0.326762   0.121577   2.688 0.007196 **
## income_levelMiddle            0.251237   0.118672   2.117 0.034257 *
## income_levelUpper-Middle      0.287177   0.124223   2.312 0.020792 *
## employment_statusRetired     -0.011311   0.062115  -0.182 0.855502
## employment_statusStudent     -0.115258   0.105121  -1.096 0.272895
## employment_statusUnemployed  -0.072181   0.078672  -0.917 0.358887
## diet_score                   -0.728984   0.013942 -52.285  < 2e-16 ***
## screen_time_hours_per_day     0.269945   0.010049  26.862  < 2e-16 ***
## I(sleep_hours_per_day^2)     -0.015982   0.015098  -1.059 0.289821
## smoking_statusFormer         -0.098253   0.078322  -1.254 0.209670
## smoking_statusNever          -0.126796   0.063939  -1.983 0.047363 *
## sleep_hours_per_day           0.263156   0.212345   1.239 0.215242
## alcohol_consumption_per_week  0.065410   0.017501   3.737 0.000186 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 7.741 on 97278 degrees of freedom
## Multiple R-squared:  0.2709, Adjusted R-squared:  0.2708
## F-statistic:  2008 on 18 and 97278 DF,  p-value: < 2.2e-16
```

15. Add at least one interaction term to you model and interpret the results. Is it significant? What is the effect of a 1-unit increase in one of those interacted variables holding the other at its mean value?

The effect of diet score and sleep hours per day on diabetes risk in this model with the negative interaction coefficient could suggest that better diet (higher diet score) has a stronger protective effect for people who sleep more. However this is not statistically significant in this model.

```r
#testing the interaction between diet score and sleep hours per day.
interact <- lm(diabetes_risk_score ~ Age + education_level + income_level + employment_status
               + (diet_score*sleep_hours_per_day) + diet_score + sleep_hours_per_day + smoking_status
               + screen_time_hours_per_day+ smoking_status + alcohol_consumption_per_week, data=diabetes
summary(interact)
```

```
##
## Call:
## lm(formula = diabetes_risk_score ~ Age + education_level + income_level +
##     employment_status + (diet_score * sleep_hours_per_day) +
##     diet_score + sleep_hours_per_day + smoking_status + screen_time_hours_per_day +
##     smoking_status + alcohol_consumption_per_week, data = diabetes_data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.878  -5.055  -1.846   2.648  24.602
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                    17.389325   0.589032  29.522  < 2e-16 ***
## Age                             0.289545   0.001602 180.747  < 2e-16 ***
## education_levelHighschool      -0.037257   0.055950  -0.666 0.505474
## education_levelNo formal       -0.023660   0.117259  -0.202 0.840090
## education_levelPostgraduate     0.043390   0.076654   0.566 0.571361
## income_levelLow                 0.243509   0.128403   1.896 0.057905 .
## income_levelLower-Middle        0.327630   0.121576   2.695 0.007043 **
## income_levelMiddle              0.252212   0.118671   2.125 0.033564 *
## income_levelUpper-Middle        0.288121   0.124223   2.319 0.020376 *
## employment_statusRetired       -0.011617   0.062114  -0.187 0.851637
## employment_statusStudent       -0.114095   0.105120  -1.085 0.277760
## employment_statusUnemployed    -0.072382   0.078673  -0.920 0.357554
## diet_score                     -0.645117   0.090423  -7.134 9.78e-13 ***
## sleep_hours_per_day             0.111573   0.079877   1.397 0.162471
## smoking_statusFormer           -0.098607   0.078321  -1.259 0.208025
## smoking_statusNever            -0.126807   0.063940  -1.983 0.047344 *
## screen_time_hours_per_day       0.270017   0.010049  26.870  < 2e-16 ***
## alcohol_consumption_per_week    0.065397   0.017501   3.737 0.000187 ***
## diet_score:sleep_hours_per_day -0.011990   0.012771  -0.939 0.347819
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.741 on 97278 degrees of freedom
```

```
## Multiple R-squared:  0.2709, Adjusted R-squared:  0.2708
## F-statistic:  2008 on 18 and 97278 DF,  p-value: < 2.2e-16
```

16. Test either the model in 14 or the model in 15 using the F test for nested models. That is, estimate the full model with the variable and quadratic term, or the variable and interaction, and then estimate the reduced model without either, and run the F test to establish whether those variables significantly improve your model.

```
complete <- lm(diabetes_risk_score ~ Age + education_level + income_level +
                 employment_status + (diet_score*sleep_hours_per_day)
               + diet_score + sleep_hours_per_day + smoking_status
               + screen_time_hours_per_day+ smoking_status
               + alcohol_consumption_per_week, data=diabetes_data)
reduced <- lm(diabetes_risk_score ~ Age + education_level + income_level + employment_status
              + smoking_status + screen_time_hours_per_day+ smoking_status
              + alcohol_consumption_per_week, data=diabetes_data)
anova(reduced, complete)
```

```
## Analysis of Variance Table
##
## Model 1: diabetes_risk_score ~ Age + education_level + income_level +
##     employment_status + smoking_status + screen_time_hours_per_day +
##     smoking_status + alcohol_consumption_per_week
## Model 2: diabetes_risk_score ~ Age + education_level + income_level +
##     employment_status + (diet_score * sleep_hours_per_day) +
##     diet_score + sleep_hours_per_day + smoking_status + screen_time_hours_per_day +
##     smoking_status + alcohol_consumption_per_week
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1  97281 5993057
## 2  97278 5829026  3    164031 912.48 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```