Results

Maggie McCracken, Harrison Clayman, Sarah Wright

While body mass index (BMI) is determined by height and weight, the goal of this data analysis was to create a model predicting BMI through other factors, allowing this calculation to be made when body measurements are unavailable. Exploratory data analysis showed that the BMI of the 1807 subjects (907 female) included in the dataset was 28.71 on average (with a *SD =* 6.582) (Table 1). The mean of this dataset falls into the overweight categorization of BMI measurements (Ardern et al., 2004). 1.6% (*SD =* 0.016) of the subjects had a BMI of under 18.5, which is categorized as underweight. 28.89% (*SD =* 0.45) of the subjects had a BMI between 18.5 and 24.9, which is categorized as normal or healthy weight range. 30.88% (SD = 0.46) of the subjects had a BMI between 25.0 and 29.9, which is categorized as overweight. 35.25% (*SD =* 0.48) of the subjects had a BMI over 30.0, which is categorized as obese. The age ranged from 20 to 80 years old (*M =* 47.40, *SD =* 17.156) (Table 2).

**Table 1**

*Descriptions of numerical variables in multiple regression model*

| **Variables** | ***Mean*** | ***SD*** |
| --- | --- | --- |
| BMI | 28.7103 | 6.5820 |
| Age | 47.4018 | 17.1563 |
| Cholesterol Levels | 1.3637 | 0.3920 |
| Testosterone Levels | 217.827 | 232.511 |
| Systolic Blood Pressure | 121.283 | 16.9098 |
| Age of Diabetes Diagnosis | 48.7438 | 14.0311 |
| Hours of Sleep Per Night | 6.8747 | 1.3433 |

**Table 2**

*Descriptions of categorical variables in multiple regression model*

| **Variables** | ***n*** | **%** |
| --- | --- | --- |
| Gender |  |  |
| Female | 907 | 50.19 |
| Male | 900 | 49.81 |
| General Health |  |  |
| Excellent | 192 | 10.63 |
| Very Good | 538 | 29.77 |
| Good | 654 | 36.19 |
| Fair | 220 | 12.17 |
| Poor | 39 | 2.16 |
| Smoke Habit |  |  |
| Current Smoker | 347 | 19.20 |
| Smoked in Past | 1009 | 55.84 |
| Never Smoked | 451 | 24.96 |

Our multiple regression model has a total of seven variables: cholesterol, general health, testosterone levels, systolic blood pressure, age of diabetes diagnosis, hours of sleep per night, and smoke habits. The adjusted R-squared of the final, log-transformed model was 0.257, which means that this model predicts 25.7% of the variability in this dataset. The p-value of the final model was really small, *F*(11, 1508) = 48.76, *p* < .001, showing that there is very strong evidence that there is one or more predictors in the final model that are significant for predicting BMI. The only p-value for a coefficient slope that provided moderate evidence for being a meaningful addition to the model was the slope for poor general health, *p* = 0.065418 (Table 3). This means that there is only moderate evidence for poor general health being statistically different from the reference category, which for general health is a rating of excellent. The p-value for the other coefficient slopes were very small, *p* < .001, showing that there is very strong evidence that the remaining coefficients are meaningful to the model (Table 3). Because of this, changes in the explanatory variables are strongly related to changes in BMI.

The following interpretations are considering all other variables are held constant. With the cholesterol level coefficient transformed because of the log-transform on BMI, when mmol/L in cholesterol levels decrease by 18.97%, BMI increases by 1 unit. When ng/dL in testosterone levels decrease by 0.0002%, BMI increases by 1 unit. When mmHg in systolic blood pressure increases by 0.1185702%, BMI increases by 1 unit. When the age of diabetes diagnosis increases by 0.1420007%, BMI increases by 1 unit. When the hours of sleep per night decreases by 1.247158%, BMI increases by 1 unit.

The following interpretations with the categorical variables are the difference between the reference group and consider all other variables are held constant. With general health, the reference group is excellent health. The difference with very good health is 5.75%, with good health is 11.43%, with fair health is 13.22%, and with poor health is 6.82%. The overall trend with general health is that the poorer the health is, the greater the difference between excellent health. The reference group for smoking habits is currently smoking. The difference between smoking in the past and never smoking is about 8.10%.

**Table 3**

*Significance of coefficients of final prediction model*

| **Coefficients** | **Estimate** | **Std. Error** | **p-value** | **CI(Lower, Upper)** |
| --- | --- | --- | --- | --- |
| Intercept | 3.453e+00 | 4.987e-02 | < 2e-16 \*\*\* | ( 3.3550, 3.5506) |
| Cholesterol Levels | -2.104e-01 | 1.265e-02 | < 2e-16 \*\*\* | (-0.2352, -0.1856) |
| Hours of Sleep Per Night | -1.26E-02 | 3.56E-03 | 0.000440 \*\*\* | (-0.0195, -0.0056) |
| General Health - Very Good | 5.59E-02 | 1.61E-02 | 0.000534 \*\*\* | (0.0243, 0.0875) |
| General Health - Poor | 6.60E-02 | 3.58E-02 | 0.065418 \* | (-0.0042, 0.1361) |
| Smoke Habit - Smoked in Past | 7.79E-02 | 1.47E-02 | 1.30e-07 \*\*\* | (0.0491, 0.1067) |
| Age of Diabetes Diagnosis | 1.42E-03 | 3.34E-04 | 2.23e-05 \*\*\* | (0.0008, 0.0021) |
| Testosterone Levels | -1.56E-04 | 2.10E-05 | 2.29e-13 \*\*\* | (-0.0002, -0.0001) |
| General Health - Good | 1.08E-01 | 1.62E-02 | 3.25e-11 \*\*\* | (0.0764, 0.1399) |
| General Health - Fair | 1.24E-01 | 1.97E-02 | 3.79e-10 \*\*\* | (0.0856, 0.16286) |
| Systolic Blood Pressure | 1.19E-03 | 2.94E-04 | 5.66e-05\*\*\* | (0.0006, 0.0018) |
| Smoke Habit - Never Smoked | 7.80E-02 | 1.34E-02 | 7.29e-09 \*\*\* | (0.0517, 0.1042) |

Significance codes: \*p < .07, \*\* p < 0.03, \*\*\* p < .001

*Note:* The reference group for general health is excellent and for smoke habit is current smoker.

The confidence intervals of the coefficients’ slope provided further evidence for their significance in this multiple regression model. The 95% confidence interval for the slope coefficient of poor general health contained zero (Table 3), meaning that we are 95% confident that there is no predicting ability of poor general health on BMI, which is consistent with the findings with the slope being moderate evidence of meaningful addition based on the p-value.

However, the other confidence intervals in the categorical variable of general health did not contain zero. For example, we are 95% confident that as BMI increases by 1 unit, sleep decreases between 0.56 and 1.93% For a categorical example, we are 95% confident that the difference between excellent and fair health is between 8.94 and 17.69%.

**References**

Ardern, C. I., Janssen, I., Ross, R., & Katzmarzyk, P. T. (2004). Development of health‐related waist circumference thresholds within BMI categories. *Obesity research, 12*(7), 1094-1103. https://doi.org/10.1038/oby.2004.137