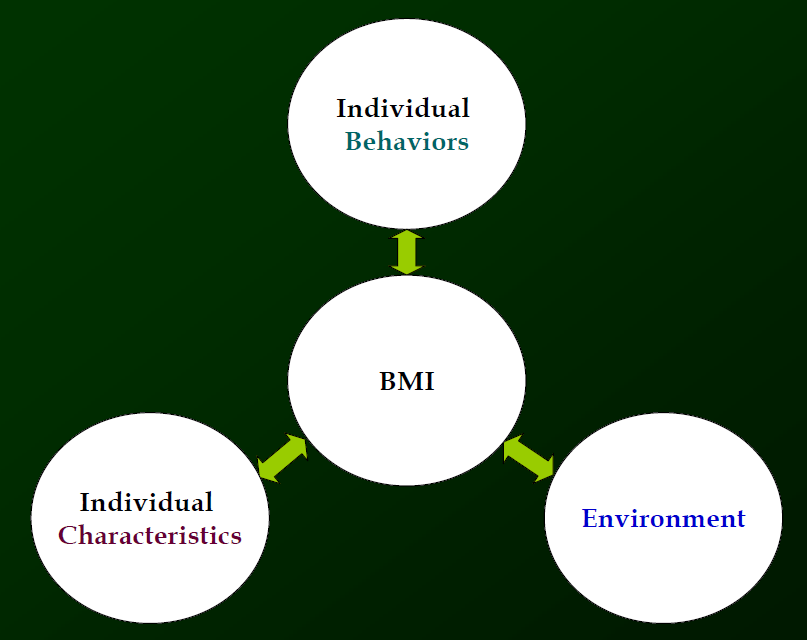
**SPSS Practical 7:**

**Part A:** Consider the following conceptualizing problem:



* Individual behaviors are associated with BMI
  + Eating behavior: daily calorie intake is positively associated with BMI
  + Exercising behavior: level of exercise is negatively associated with BMI
* Individual characteristics are associated with BMI
  + Sex
  + Income
  + Education level
  + Occupation
* Environment is associated with BMI
  + Physical environment
  + Social environment

In a study, 1000 adults aged 18+ (males and females) were recruited to study factors associated with BMI. Data collected are based on the following variables:

–BMI

–Sex (female=1) –individual characteristics

–Calorie (calorie intake daily) –individual behaviors

–Exercise (minutes of exercise per week) –individual behaviors

–Income (monthly salary in dollars $) –individual characteristics

–Expenditure on food (monthly food expense in dollars $) –individual behaviors

–Education (education level in years) –individual characteristics

–Residential density (high, median, low) –physical environment

1. Write a multiple regression model for four selected factors: (1) calorie, (2) exercise, (3) income and (4) education.

// 꼭 둘 다 scale data 가 아니여도됨

// B 나 베타

// factor 이야기 안하면 다 때려박음

BMI =

1. Check multicollinearity of independent variables based on

// multicollinearity = 너무 영향이 큰 것을 걸러냄

* 1. Pearson Correlation
* If the absolute value of Pearson correlation is greater than 0.8, collinearity is very likely to exist.
* If the absolute value of Pearson correlation is close to 0.8 (such as 0.7±0.1), collinearity is likely to exist.

// 0.8 넘으면 말도안됨

// pearson’s test에 여러 개 넣음

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Correlations** | | | | | | |
|  | | BMI | calorie | exercise | income | education |
| Pearson Correlation | BMI | 1.000 | .784 | -.310 | .033 | .011 |
| calorie | .784 | 1.000 | -.193 | -.009 | .004 |
| exercise | -.310 | -.193 | 1.000 | -.030 | -.046 |
| income | .033 | -.009 | -.030 | 1.000 | .069 |
| education | .011 | .004 | -.046 | .069 | 1.000 |
| Sig. (1-tailed) | BMI | . | .000 | .000 | .148 | .361 |
| calorie | .000 | . | .000 | .391 | .451 |
| exercise | .000 | .000 | . | .175 | .072 |
| income | .148 | .391 | .175 | . | .014 |
| education | .361 | .451 | .072 | .014 | . |
| N | BMI | 1000 | 1000 | 1000 | 1000 | 1000 |
| calorie | 1000 | 1000 | 1000 | 1000 | 1000 |
| exercise | 1000 | 1000 | 1000 | 1000 | 1000 |
| income | 1000 | 1000 | 1000 | 1000 | 1000 |
| education | 1000 | 1000 | 1000 | 1000 | 1000 |

The Pearson Correlations confirm that there are no serious problems with multicollinearity. All Pearson Correlation *r*<0.8, indicating that the predictors are not intercorrelated and that small changes in the data values will not lead to large changes in the estimates of the coefficients.

* 1. Condition Index
* A condition index greater than 15 indicates a possible problem.
* An index greater than 30 suggests a serious problem with collinearity.

// regression -> lean~~ -> statistic 오른쪽 1,3,4,5 번째꺼 체크

// 계층 구조 알려면 한 개씩 넣고 넥스트 두 개 넣고 넥스트

//

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Collinearity Diagnosticsa** | | | | | | | | |
| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | | | |
| (Constant) | calorie | exercise | income | education |
| 1 | 1 | 4.778 | 1.000 | .00 | .00 | .00 | .00 | .00 |
| 2 | .110 | 6.584 | .00 | .10 | .72 | .02 | .01 |
| 3 | .060 | 8.924 | .00 | .41 | .03 | .56 | .00 |
| 4 | .041 | 10.842 | .01 | .21 | .05 | .26 | .55 |
| 5 | .011 | 21.197 | .99 | .28 | .19 | .16 | .44 |
| a. Dependent Variable: BMI | | | | | | | | |

The Collinearity Diagnostics confirm that there are no serious problems with multicollinearity. Most of the condition indexes < 15, indicating that the predictors are not intercorrelated and that small changes in the data values will not lead to large changes in the estimates of the coefficients.

// 전체가 중요해서 한 개라도 15넘으면 영향

* 1. Tolerance and VIF
* Tolerance should be more than 0.2
* VIF should be less than

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 20.693 | .208 |  | 99.404 | .000 |  |  |
| calorie | .002 | .000 | .753 | 38.969 | .000 | .962 | 1.039 |
| exercise | -.027 | .003 | -.163 | -8.434 | .000 | .960 | 1.042 |
| income | 8.819E-5 | .000 | .035 | 1.837 | .067 | .994 | 1.006 |
| education | -.001 | .006 | -.002 | -.086 | .932 | .993 | 1.007 |
| a. Dependent Variable: BMI | | | | | | | | |

The Collinearity statistics confirm that there are no serious problems with multicollinearity.All tolerance indexes > 0.2 and VIF < 10, indicating that the predictors are not intercorrelated and that small changes in the data values will not lead to large changes in the estimates of the coefficients.

1. Check the goodness of fit of model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .801a | .642 | .641 | .77095 |
| a. Predictors: (Constant), education, calorie, income, exercise | | | | |

/\* note

The adjusted adjusted for the number of explanatory terms (independent variables) in a model and increases only if the new independent variable(s), improve(s) the model more than would be expected by chance.

\*/

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 1062.377 | 4 | 265.594 | 446.853 | .000b |
| Residual | 591.394 | 995 | .594 |  |  |
| Total | 1653.771 | 999 |  |  |  |
| a. Dependent Variable: BMI | | | | | | |
| b. Predictors: (Constant), education, calorie, income, exercise  // significant means predict model better than mean. | | | | | | |

The ANOVA table reports a significant F statistic, indicating that using the model is better than guessing the mean.

1. Analyze the coefficient of each independent variable.

* Unstandardized coefficients used in the prediction and interpretation.
* Standardized coefficients used for comparing the effects of independent variables.
* Compared Sig. with alpha 0.05. If Sig. <0.05, the coefficient is statistically significant from zero.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Collinearity Statistics | |
| B | Std. Error | Beta | Tolerance | VIF |
| 1 | (Constant) | 20.693 | .208 |  | 99.404 | .000 |  |  |
| calorie | .002 | .000 | .753 | 38.969 | .000 | .962 | 1.039 |
| exercise | -.027 | .003 | -.163 | -8.434 | .000 | .960 | 1.042 |
| income | 8.819E-5 | .000 | .035 | 1.837 | .067 | .994 | 1.006 |
| education | -.001 | .006 | -.002 | -.086 | .932 | .993 | 1.007 |
| * 1. Dependent Variable: BMI | | | | | | | | |

// significant means important

For income and education, sig > 0.05

These predictors are not significant.

1. What is your estimated multiple regression model?

BMI =

1. Interpret all estimated coefficients for predictors in Part (5).

Controlling for other variables constant, if a person increases 1 calorie intake per day, the BMI of the person will increase by 0.002

Controlling for other variables constant, if a person increases 1 min exercise per week, the BMI of the person will decrease by 0.027

7) Complete your analysis report for multiple regression

A multiple regression was run to predict BMI from calorie, exercise, income and education. These variables statistically predicted BMI, F(4,995) = 446.853, sig = 0.00 (<0.05), = 0.642, only two variables (calorie and exercise) added statistically significantly to the prediction, .

// p == 알파 p로써야함

**Part B:** Repeat questions in Part (A) for Smoking\_Survey dataset.

1. Write a multiple regression model for three selected factors: (1) age, (2) cigarette smoked per day, (3) smoked per year

Health = + (age) + (cigarette smoked per day) + (smoked per year)

2. Check multicollinearity of independent variables based on

a. Pearson Correlation

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Correlations** | | | | | |
|  | | Health | The age of the subject | The smoked per day of the subject | The number of years smoked |
| Health | Pearson Correlation | 1 | .324\* | .442\*\* | .527\*\* |
| Sig. (2-tailed) |  | .012 | .000 | .000 |
| N | 59 | 59 | 59 | 57 |
| The age of the subject | Pearson Correlation | .324\* | 1 | .086 | .355\*\* |
| Sig. (2-tailed) | .012 |  | .519 | .007 |
| N | 59 | 59 | 59 | 57 |
| The smoked per day of the subject | Pearson Correlation | .442\*\* | .086 | 1 | .646\*\* |
| Sig. (2-tailed) | .000 | .519 |  | .000 |
| N | 59 | 59 | 59 | 57 |
| The number of years smoked | Pearson Correlation | .527\*\* | .355\*\* | .646\*\* | 1 |
| Sig. (2-tailed) | .000 | .007 | .000 |  |
| N | 57 | 57 | 57 | 57 |
| \*. Correlation is significant at the 0.05 level (2-tailed). | | | | | |
| \*\*. Correlation is significant at the 0.01 level (2-tailed). | | | | | |

The Pearson Correlations confirm that there are no serious problems with multicollinearity. All Pearson Correlation *r*<0.8, indicating that the predictors are not intercorrelated and that small changes in the data values will not lead to large changes in the estimates of the coefficients.

b. Condition Index

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Collinearity Diagnosticsa** | | | | | | | |
| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | | |
| (Constant) | The age of the subject | The smoked per day of the subject | The number of years smoked |
| 1 | 1 | 3.058 | 1.000 | .01 | .01 | .03 | .03 |
| 2 | .638 | 2.188 | .05 | .04 | .15 | .13 |
| 3 | .242 | 3.554 | .02 | .02 | .66 | .63 |
| 4 | .062 | 7.027 | .92 | .93 | .16 | .22 |
| a. Dependent Variable: Health | | | | | | | |

The Collinearity Diagnostics confirm that there are no serious problems with multicollinearity. Most of the condition indexes < 15, indicating that the predictors are not intercorrelated and that small changes in the data values will not lead to large changes in the estimates of the coefficients.

c. Tolerance and VIF

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Correlations | | | Collinearity Statistics | |
| B | Std. Error | Beta | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | (Constant) | .523 | .324 |  | 1.614 | .112 |  |  |  |  |  |
| The age of the subject | .013 | .009 | .180 | 1.443 | .155 | .311 | .194 | .164 | .830 | 1.204 |
| The smoked per day of the subject | .012 | .009 | .205 | 1.347 | .184 | .432 | .182 | .153 | .553 | 1.809 |
| The number of years smoked | .034 | .017 | .331 | 2.035 | .047 | .527 | .269 | .231 | .486 | 2.059 |
| a. Dependent Variable: Health | | | | | | | | | | | |

The Collinearity statistics confirm that there are no serious problems with multicollinearity.All tolerance indexes > 0.2 and VIF < 10, indicating that the predictors are not intercorrelated and that small changes in the data values will not lead to large changes in the estimates of the coefficients.

1. Check the goodness of fit of model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .565a | .319 | .281 | .873 |
| a. Predictors: (Constant), The number of years smoked, The age of the subject, The smoked per day of the subject | | | | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 18.909 | 3 | 6.303 | 8.278 | .000b |
| Residual | 40.355 | 53 | .761 |  |  |
| Total | 59.263 | 56 |  |  |  |
| a. Dependent Variable: Health | | | | | | |
| b. Predictors: (Constant), The number of years smoked, The age of the subject, The smoked per day of the subject | | | | | | |

The ANOVA table reports a significant F statistic, indicating that using the model is better than guessing the mean.

1. Analyze the coefficient of each independent variable

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Coefficientsa** | | | | | | | | | | | |
| Model | | Unstandardized Coefficients | | Standardized Coefficients | t | Sig. | Correlations | | | Collinearity Statistics | |
| B | Std. Error | Beta | Zero-order | Partial | Part | Tolerance | VIF |
| 1 | (Constant) | .523 | .324 |  | 1.614 | .112 |  |  |  |  |  |
| The age of the subject | .013 | .009 | .180 | 1.443 | .155 | .311 | .194 | .164 | .830 | 1.204 |
| The smoked per day of the subject | .012 | .009 | .205 | 1.347 | .184 | .432 | .182 | .153 | .553 | 1.809 |
| The number of years smoked | .034 | .017 | .331 | 2.035 | .047 | .527 | .269 | .231 | .486 | 2.059 |
| a. Dependent Variable: Health | | | | | | | | | | | |

For constant age and smoked per day, sig > 0.05

These predictors are not significant

5. What is your estimated multiple regression model?

Health = 0.523 + 0.34\*(smoked year)

6. Interpret all estimated coefficients for predictors in Part (5)

Controlling for other variables constant, if a Health will increase by 0.34

7. Complete your analysis report for multiple regressions.

A multiple regression was run to predict Health from age, smoked\_per\_day, smoked\_year. These variables statistically predicted Health, F(4,995) = 8.278, sig = 0.00 (<0.05), = 0.319, only one variables (smoked\_year) added statistically significantly to the prediction, .