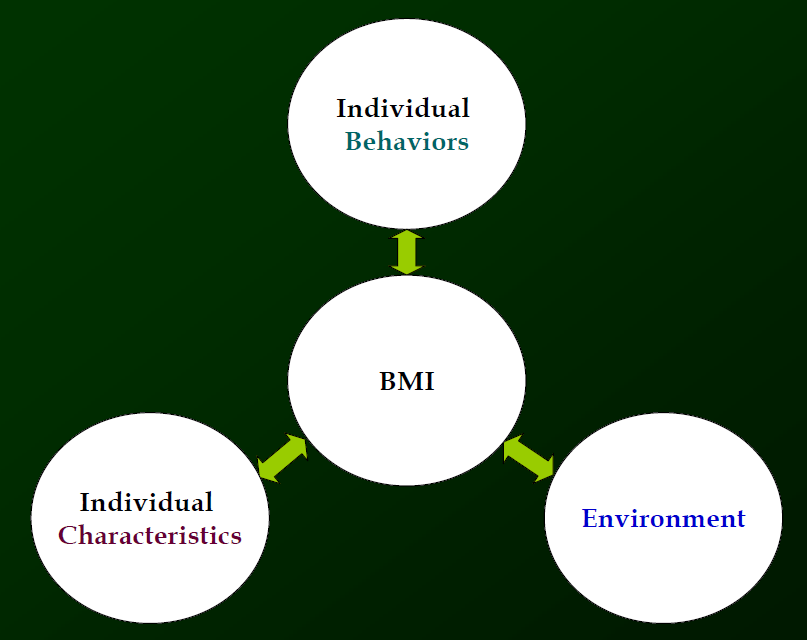
**SPSS Practical 7:**

**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.

BMI =

1. Check multicollinearity of independent variables based on

// 3가지 방법 중에서 2개 이상이 일치하는 방법사용해야함.

* 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **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.

// person 진행하면 됨.

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

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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.

//linear regression 에서 dependent 에 outcome 넣고 나머지는 independent 에 넣고 statistics 에서 1,3,4,5, 체크

* 1. Tolerance and VIF
* Tolerance should be more than 0.2
* VIF should be less than 10

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **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 | | | | |

//R제곱값이 적은 것이 안좋은 것 ,값이 커야 좋은 결과임.

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 | | | | | | |

//여기서는 model 사용이 좋을지, mean 사용이 좋을지 결정.

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 |
| a. Dependent Variable: BMI | | | | | | | | |

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, .

//p는 CI 사용시 남는 구간, 여기서는 95%CI이기 때문에 0.05로 사용.

**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

Wegiht =

2. Check multicollinearity of independent variables based on

a. Pearson Correlation

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Correlations** | | | | |
|  | | The age of the subject | The frequency of smoking per day of the subject | The frequency of smoking per year of the subject |
| The age of the subject | Pearson Correlation | 1 | .086 | .355\*\* |
| Sig. (2-tailed) |  | .519 | .007 |
| N | 59 | 59 | 57 |
| The frequency of smoking per day of the subject | Pearson Correlation | .086 | 1 | .646\*\* |
| Sig. (2-tailed) | .519 |  | .000 |
| N | 59 | 59 | 57 |
| The frequency of smoking per year of the subject | Pearson Correlation | .355\*\* | .646\*\* | 1 |
| Sig. (2-tailed) | .007 | .000 |  |
| N | 57 | 57 | 57 |
| \*\*. 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.

So, Because there is no r >0.8, it means that there is no multicollinearity.

b. Condition Index

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Collinearity Diagnosticsa** | | | | | | | |
| Model | Dimension | Eigenvalue | Condition Index | Variance Proportions | | | |
| (Constant) | The age of the subject | The frequency of smoking per day of the subject | The frequency of smoking per year of the subject |
| 1 | 1 | 3.073 | 1.000 | .01 | .01 | .03 | .02 |
| 2 | .625 | 2.217 | .05 | .04 | .16 | .13 |
| 3 | .242 | 3.562 | .02 | .02 | .65 | .63 |
| 4 | .060 | 7.142 | .92 | .93 | .17 | .21 |
| a. Dependent Variable: The weight of the subject | | | | | | | |

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.

Because All condition index < 15, so there is no serious problem with multicollinearity.

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) | 73.586 | 21.117 |  | 3.485 | .001 |  |  |  |  |  |
| The age of the subject | 1.645 | .563 | .394 | 2.925 | .005 | .420 | .379 | .360 | .835 | 1.198 |
| The frequency of smoking per day of the subject | .670 | .567 | .194 | 1.181 | .243 | .245 | .163 | .145 | .557 | 1.794 |
| The frequency of smoking per year of the subject | .313 | 1.062 | .052 | .295 | .769 | .309 | .041 | .036 | .494 | 2.022 |
| a. Dependent Variable: The weight of the subject | | | | | | | | | | | |

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.

Because all tolerance indexs >0.2, and VIF <10 there is no serious problem with multicollinearity.

3. Check the goodness of fit of model

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Model Summary** | | | | |
| Model | R | R Square | Adjusted R Square | Std. Error of the Estimate |
| 1 | .478a | .229 | .183 | 54.861 |
| a. Predictors: (Constant), The frequency of smoking per year of the subject, The age of the subject, The frequency of smoking per day of the subject | | | | |

So, is not a good model.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **ANOVAa** | | | | | | |
| Model | | Sum of Squares | df | Mean Square | F | Sig. |
| 1 | Regression | 45516.572 | 3 | 15172.191 | 5.041 | .004b |
| Residual | 153498.955 | 51 | 3009.783 |  |  |
| Total | 199015.527 | 54 |  |  |  |
| a. Dependent Variable: The weight of the subject | | | | | | |
| b. Predictors: (Constant), The frequency of smoking per year of the subject, The age of the subject, The frequency of smoking per day of the subject | | | | | | |

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

Because SSR is bigger than SSM, F value is so small.

So, This model is very bad model. We don’t use this model. We will use Average(Mean).

4. 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) | 73.586 | 21.117 |  | 3.485 | .001 |  |  |  |  |  |
| The age of the subject | 1.645 | .563 | .394 | 2.925 | .005 | .420 | .379 | .360 | .835 | 1.198 |
| The frequency of smoking per day of the subject | .670 | .567 | .194 | 1.181 | .243 | .245 | .163 | .145 | .557 | 1.794 |
| The frequency of smoking per year of the subject | .313 | 1.062 | .052 | .295 | .769 | .309 | .041 | .036 | .494 | 2.022 |
| a. Dependent Variable: The weight of the subject | | | | | | | | | | | |

For smoked per day,smoked per year , sig > 0.05

These predictors are not significant.

5. What is your estimated multiple regression model?

Wegiht =

Wegiht = 73.586 + 1.645(age)

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

Controlling for other variables constant, if a person getting older 1year, the Weight of the person will increase by 1.645

7. Complete your analysis report for multiple regressions.

A multiple regression was run to predict Weight from age, smoked per day, and smoked per year. These variables statistically predicted Weight, F(3,51) = 5.041, sig = 0.004(<0.05), = , only one variable (age) added statistically significantly to the prediction, .