Memorandum

TO: Oregon Health Authority, Public Health Division

FROM: Catherine Susman

DATE: November 23, 2021

RE: Regression Analysis of Obesity and Life Expectancy

BACKGROUND/PURPOSE

Obesity is known to contribute to a variety of illness and other health-related problems. Due to these effects, it is theorized that obesity adversely affects life expectancy. To help determine if this causal theory is correct, the following three analyses were performed and analyzed: (1) a scatter plot; (2) a simple (bivariate) regression; and (3) a multiple regression. These analyses were performed using JASP software with data from the 2019 US Cities Sustainable Development Report ("Report.") For all three analyses, the variable obesity[1] was utilized as the independent variable (cause) and life expectancy[2] was utilized as the dependent variable (effect). Additionally, for the multiple regression, the variable food insecurity[3] was added to the regression analysis to address and control for a potential common cause. See the footnotes below for how these variables were defined and measured in the Report.

SUMMARY

All three analyses show that the rate of obesity in the population statistically significantly predicts life expectancy in the population. Obesity in the population was shown to have a strong negative correlation to life expectancy in the population, with an R^2 value of 0.751 and a p-value of <.001. As for practical significance, the analysis shows for each percentage increase in obesity in the population, there is a corresponding decrease by just under ½ year in life expectancy in the population.

The analyses below show a strong negative correlation and support a causal relationship between obesity in the population and life expectancy in the population; however, caution is advised. For the analysis to be fully accurate it needs to include all applicable common causes, and only the applicable common causes (Remler & Van Ryzin, 2015, p. 409). Without the exact variables, the regression analysis may be biased due to missing control variables, including unmeasured variables (p. 409). Such bias will affect the accuracy of the results of the analysis. Due to these limitations, while the analyses below provide compelling evidence supporting a causal relationship between a population's prevalence of obesity and its life expectancy, statistical analysis alone cannot prove causation.

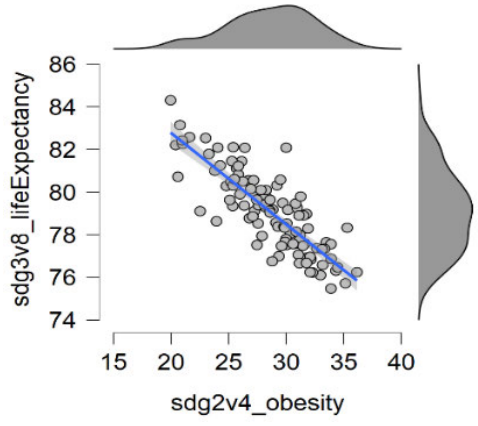
[1] The variable obesity, sdg2v4, is the prevalence of obesity in the population measured as the percentage of adults that report BMI ≥ 30 (Lynch, LoPresti, & Fox, 2019, p. 40).
[2] The variable life expectancy, sdg3v8, is the age-adjusted life expectancy from birth measured in years (Lynch, LoPresti, & Fox, 2019, p. 41).
[3] The variable food insecurity, sdg2v1, is the prevalence of food insecurity experienced by the population measured as the percentage estimates of individuals experiencing food insecurity (Lynch, LoPresti, & Fox, 2019, p. 40).

DISCUSSION

Scatter Plot

To start the analysis of the potential cause-effect relationship between obesity and life expectancy, the following scatter plot was created to allow for visualization of the data:

obesity – life expectancy



The data visualization in the above scatter plot shows the relationship between obesity and life expectancy is linear. Since the later regression and correlation analyses will be most accurate and descriptive if the relationship between obesity and life expectancy is linear, it is helpful to have that aspect determined at the outset (Remler & Van Ryzin, 2015, p. 260). Additionally, the scatter plot shows that most of the data points are closely clustered around the linear regression line, which shows there are few residuals or outliers. Again, the fewer the residual, the better the fit of the regression line, the more informative the correlation and regression analyses (p.268). Next, the scatter plot also shows that the direction of the linear relationship between obesity and life expectancy is a negative one. As shown from the scatter plot, as the percentage of obese adults increases, life expectancy decreases. Lastly, the shaded portions on the top and right-hand side of the scatter plots are a histogram of each of the variables. As shown by both histograms, they take the shape of a normal distribution with little to no skewness at either end (p 283). This is important to know because it tells us the mean of the data is an accurate mean of the population (p. 283). This will also make the analysis for statistical significance easier and more accurate (p. 283).

Regression Analyses

The use of regression analysis is a core component of significance testing between two or more variables (Remler & Van Ryzin, 2015, pp. 321-322). In this case, significance testing was undertaken to determine whether there was a causal relationship between the independent variable, obesity, and the dependent variable, life expectancy (p. 320). Significance testing is undertaken to prove a relationship between two variables through the "logic of falsification," by showing what does not exist (p. 292). To do this, one must first develop the null hypothesis, the hypothesis that is the opposite of the relationship believed to exist (p. 295). In this case, the null hypothesis was that there is no relationship between obesity and life expectancy. Next, regression analyses, both a simple regression and multiple regression, were run to determine whether the null hypothesis, there is no relationship, is true (p. 293). The discussion and results of those regression analyses are outlined below.

Simple Regression Analysis

As noted above, as part of significance testing, a simple (bivariate) regression analysis of the relationship between obesity and life expectancy was performed. The table showing the results of that regression analysis is as follows:

Model Summary: Obesity (cause) – Life Expectancy (effect)

Model	R	R^2	Adjusted R^2	RMSE
H_1	0.836	0.698	0.696	1.039

Simple Regression: Obesity (cause) – Life Expectancy (effect)

Model		Unstandardized Coefficient (Slope)	Standard Error	Standardized Coefficient	t	p
H_1	(Intercept)	91.337	0.800		114.238	< .001
	Pct. of Obese Adults	-0.428	0.028	-0.836	-15.446	< .001

As noted above, the null hypothesis for the regression analyses is that there is no relationship between obesity and life expectancy. Part of the significance testing, the testing of the null hypothesis, is to examine the R^2 value. R^2 value tells "the proportion of variation in the dependent variable explained (or predicted) by the variation in all independent variables" (Remler & Van Ryzin, 2015, p. 268).

As shown in the Model Summary table above, the R^2 value is 0.698, meaning that 70% of the variation in average life expectancy in the population is explained by or predicted by obesity. Further, the correlation coefficient, or r value, shows the direction and magnitude of the relationship between two variables (Remler & Van Ryzin, 2015, p. 262). In interpreting the magnitude or strength of the relationship, the following general rules apply: r value of .10 is considered small/weak; r value of .30 is considered moderate; and r value of .50 is considered large/strong (p. 262). With simple regression, the R^2 value is determined by squaring the r value (p. 268). An R^2 value over 0.25 or 25% (0.50 x 0.50 = 0.25,) therefore, shows a strong correlation (p. 268). Since the above R^2 value of 0.698 greater than 0.25, it shows there is a strong correlation between obesity and life expectancy.

The next part of the analysis of the simple regression is examining the unstandardized coefficient or slope. The unstandardized coefficient/slope is "the change in the dependent variable for an associated with one-unit increase in the independent variable" (Remler & Van Ryzin, 2015, p. 266). In this case, the slope is the change in life expectancy given a one-percentage point increase in obesity. If null hypothesis, that there is no relationship between obesity and life expectancy, is true, then the slope (coefficient) will be zero (p. 295).

The first line of Simple Regression table above shows the results for the intercept of obesity if there is 0% obesity in the population. With 0% obesity, the average life expectancy in the population is just over 91 years old. Next, the second line of the Simple Regression table shows that for each percent increase of obesity in the population, there is a negative 0.428-year decrease (just under a ½ year decrease) in life expectancy. The slope is not zero, therefore, not consistent

with the null hypothesis. The next step is to determine if the slope value shown is significant or just random chance. To determine significance, review the standard error of the coefficient and the corresponding t-statistic and p-values.

The standard error tells the variability in the slope (coefficient) that is due to random sampling error (Remler & Van Ryzin, 2015, p. 320). The lower the standard error rate, the more precise the estimated coefficient (p. 320). In this case, the standard error is 0.028, which is a fairly low, thus the estimated coefficient of 0.428 is fairly precise.

Using the slope and dividing by its standard error results in the t-statistic (Remler & Van Ryzin, 2015, p. 293). This t-statistic is used to judge if the regression results are statistically significant or the result of random chance (p. 293). In interpreting the t-statistic, a general rule is if the t-statistic is greater than two, it is statistically significant (Rutgers University, 2021). In this case, the t-statistic is negative 15.466. This result tells us that the slope shown in the regression analysis is over 15 standard errors away from the null of zero (p. 296). The t-statistic, therefore, shows the relationship between obesity and life expectancy is statistically significant. Additionally, the t-statistic result is negative, which tells that the direction of the regression line is downward sloping, as shown on the scatter plot (Rutgers University, 2021).

More commonly examined for determining significance is the p-value that is associated with the t-statistic (Remler & Van Ryzin, 2015, p. 293). The p-value, or probability value "represents the probability of observing the sample estimate. . . when the null hypothesis is true" (p. 293). The lower the p-value, the less probable it is that the sample estimate is due to random chance (p. 294). The most commonly used p-value for determining significance is 0.05, or 5% (p. 294). An even more stringent standard is 0.01, or 1% (p. 294). If the p-value is less than .05, the relationship between the variables is interpreted to be statistically significant (p. 294). In this case, the p-value is < .001, meaning there is less than ⅒ of 1% chance that the regression results shown would have occurred due to random chance. Since it is so unlikely results would have occurred by chance, the null hypothesis that there is no relationship between obesity and life expectancy can be rejected (p. 294). The p-value of < .001 shows the relationship between obesity and life expectancy is statistically significant.

Lastly, the standardized coefficient is examined. The standardized coefficient tells "the predicted standard deviations y changes, given a 1-standard deviation increase in x" (Remler & Van Ryzin, 2015, p. 319). In simple regression, the standardized coefficient is equivalent to r, the correlation (p. 319). As shown in the Simple Regression table above, the r and standardized coefficient value is 0.836. The standardized coefficient is a negative value indicating the direction of the regression line is downward sloping. The above standardized coefficient results are interpreted as meaning when the percentage of obesity in the population increases by one standard deviation, the life expectancy decreases by over ¾ of a standard deviation.

Multiple Regression Analysis

The last part of the analysis was to perform a multiple regression analysis of the relationship between obesity and life expectancy after controlling for the possible influence of food insecurity. A multiple regression analysis is used for two reasons: one to predict the dependent variable given a combination of independent variables; and two, to estimate causal effects of the independent variable on the dependent variable (Remler & Van Ryzin, 2015, p. 317). For the

purposes of this analysis, the multiple regression is used for the latter purpose. The table showing the results of that multiple regression analysis is as follows:

Model Summary - Obesity (cause) – Life Expectancy (effect) controlling for Food Insecurity (common cause)

Model	R	R^2	Adjusted R^2	RMSE
H_1	0.867	0.751	0.746	0.948

Multiple Regression – Obesity (cause) – Life Expectancy (effect) controlling for Food Insecurity (common cause)

Model		Unstandardized Coefficient (Slope)	Standard Error	Standardized Coefficient	t	p
H_1	(Intercept)	91.994	0.743		123.757	< .001
	Pct. of Obese Adults	-0.350	0.030	-0.684	-11.538	< .001
	Pct. of Food Insecure Individuals	-0.222	0.048	-0.275	-4.649	< .001

Unlike the simple regression analysis first undertaken, a multiple regression allows for adding an additional independent variable, a common cause, to control for that cause and prevent confounding (Remler & Van Ryzin, 2015, p. 404). With the added control variable, first, examine the R^2 value under the multiple regression. With the common cause added to the analysis, the R^2 value has increased. Note, however, that the R^2 value will always increase when variables are added (p. 316). The adjusted R^2 is the R^2 value adjusted to help account for the automatic increase to due to adding variables (p. 316). However, for our analysis, since only one additional independent variable was added, we can continue to use the R^2 for the purposes of our analysis. As seen in the Model Summary table above, R^2 increased from 0.698 to 0.751, or from 70% to 75%. With the control variable, food insecurity added, the model shows 75% of the variation in average life expectancy in the population is explained or predicted by obesity and food insecurity. As noted in the simple regression analysis, an R^2 value of over.25 or 25% is a strong correlation. (p. 268). The multiple regression analysis shows the addition of the control variable further strengthened an already strong correlation.

While R^2 value shows a strong correlation and is important for seeking to identify and understand the different possible causes of a dependent variable, when the regression analysis is for causation, the R^2 value is not that helpful (Remler & Van Ryzin, 2015, p. 417). Rather the focus of the analysis should be on the unstandardized coefficient or slope and the corresponding t-statistic and p-values (p. 417). Since the purpose of this regression analysis is to estimate whether obesity effects life expectancy causally and the magnitude of the effect, the focus of this analysis is on the unstandardized coefficient or slope and its corresponding t-statistic and p-value of the independent variable of interest, obesity.

In examining the slope from this multiple regression compare it to the slope from the simple regression analysis. As shown in the tables above, the slope changed from negative 0.428 under the simple regression to negative 0.350 under the multiple regression. This shows that the effect of obesity, when food insecurity is controlled for, is smaller than initially found in the simple

regression. Under the multiple regression, for each percent increase of obesity in the population, there is a negative .350-year decrease (just over ¼ year decrease) in life expectancy. While the overall effect of obesity is weaker than initially shown, it is still a significant and substantively large effect.

Next, examine the standard error and the resulting t-statistic and p-values and compare the results of the simple regression to this multiple regression. The comparison is done to determine if, and to what extent, the addition of the control variable of food insecurity effected the relationship between obesity and life expectancy. The t-statistic changed from negative 15.446 under the simple regression to negative 11.538 under the multiple regression. However, the p-values under both the simple regression and the multiple regression are < .001. While the t-statistic is lower under the multiple regression, both the t-statistic and p-value show the relationship between obesity and life expectancy remains statistically significant.

While not necessary to the examination of the causal effect of obesity on life expectancy, the same type of examination of the slope, standard error, t-statistic, and p-value can be made to analyze the relationship between the control variable, food insecurity, and life expectancy (Remler & Van Ryzin, 2015, p. 404). In analyzing the slope for food insecurity, it shows that for every percent increase in food insecurity in the population, the average life expectancy decreases by just under ¼ a year. Its corresponding t-statistic is over negative 4.649 and its p-value is < .001, so the variable of food insecurity, like obesity, is statistically significant.

Lastly, examine the standardized coefficient of obesity and compare the value under the simple regression to the value under the multiple regression. The standardized coefficient value changed from negative 0.836 under the simple regression to negative 0.684 under the multiple regression. Similar to the change in the slope, the effect of obesity, when food insecurity is controlled for, is smaller than initially found under the simple regression. With food insecurity controlled for, when the obesity percentage in the population increases by one standard deviation, the life expectancy in the population decreases by just under ¾ of a standard deviation. Again, while lower, the results remain significant.

Again, while not necessary to the examination of the causal effect of obesity on life expectancy, the standardized coefficients of obesity and food insecurity may be compared to each other to determine which variable has the greater effect on life expectancy (Remler & Van Ryzin, 2015, p. 319). The standardized coefficient of obesity is negative 0.684 while standardized coefficient of food insecurity is negative 0.275. The standardized coefficient of obesity is larger, showing that obesity is a stronger predictor of life expectancy than food insecurity.

References

Lynch, A., LoPresti, A., & Fox, C. (2019). *The 2019 US Cities Sustainable Development Report.* New York: Sustainable Development Solutions Network (SDSN). Retrieved November 2021, from https://s3.amazonaws.com/sustainabledevelopment.report/2019/2019USCitiesReport.pdf

Remler, D. K., & Van Ryzin, G. G. (2015). *Research Methods in Practice: Strategies for Description and Causation* (Second ed.). Thousand Oaks, CA: Sage Publications, Inc.

Rutgers University. (2021). *Regression Analysis* [Video]. Newark, New Jersey. Retrieved November 2021, from https://rutgers.instructure.com/courses/143272/pages/10-regression-analysis-lectures?module_item_id=4919372