Impact of Age on VPCS42: A Simple Linear Regression Analysis

Jacqueline V. Gomez

Email: gomez.jacqueline14@gmail.com

Linkedin: linkedin.com/in/jacquelinevgomez

28 January 2024

Impact of Age on VPCS42: A Simple Linear Regression Analysis

Contents

Introduction

This document explores the relationship between age and physical health scores (VPCS42) using simple linear regression analysis. The analysis is based on a dataset comprising **14,423 observations** and **23 variables**, drawn from the **2020 Medical Expenditure Panel Survey (MEPS)**.

The key variables of interest in this analysis include:

- **VPCS42 (Physical Component Score):** This numeric variable reflects the physical health component of the VR-12 health survey, providing insights into an individual's physical well-being.
- **Age:** A continuous numeric variable representing the respondent's age, specifically including adults aged 18 and older in this dataset.

Real-World Applications

Conducting a simple linear regression with age as the independent variable (X) and VPCS42 as the dependent variable (Y) can help answer several real-world questions, such as:

(1) How does age influence physical health scores (VPCS42)?
(2) Are there significant differences in physical health outcomes across different age groups?
(3) Can age be used to predict physical health outcomes?
(4) Should resources be allocated differently based on age-related physical health scores?

**Responses to these questions have been addressed formally in the PowerPoint presentation titled: "Impact of Age on Physical Health Scores (VPCS42)"*

Output Analysis

Analysis Of Variance (ANOVA)

Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	309146	309146	2955.92	<.0001
Error	14421	1508226	104.58536		
Corrected Total	14422	1817371			

- **Source**: This indicates the different sources of variation in the model (Model, Error, and Corrected Total).
  - **Model**
    - **DF (Degrees of Freedom):** 1 (corresponds to the one predictor, age).
    - **Sum of Squares:** 309,146 (variation explained by the model).
    - **Mean Square:** 309,146.
  - **Error**
    - **DF:** 14,421 (# of observations minus the number of parameters estimated).
    - **Sum of Squares:** 1,508,226 (variation that is not explained by the model).
    - **Mean Square:** 104.58536.
  - **Corrected Total**
    - **Sum of Squares:** 1,817,371 (total variation in the dependent variable).
- **F Value:** 2955.92 (ratio of the Model Mean Square to the Error Mean Square). This F-value is considered high and indicates that the model explains a significant amount of variation relative to the error.
- **Pr > F:** <.0001 (p-value associated with the F-test). Since it's very low (less than 0.05), it indicates that the model is statistically significant; therefore, age is a significant predictor of VPCS42.

Model Summary

Root MSE	10.22670	R-Square	0.1701
Dependent Mean	48.37372	Adj R-Sq	0.1700
Coeff Var	21.14102		

- **Root MSE**: 10.2267 (standard deviation of the residuals, approximately the average distance that the observed values differ from the regression line).
- **Dependent Mean**: 48.37372 (represents the average VPCS42 score).
- **R-Square**: 0.1701 (approximately 17% of the variance in VPCS42 can be explained by age).
- **Adj R-Sq**: 0.17 (accounts for the number of predictors in the model, is similar here since there's only one predictor).
- **Coeff Var**: 21.14102 (indicates the extent of variability in relation to the mean of the dependent variable).

Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	62.05995	0.26574	233.53	<.0001
age	Continuous age	1	-0.25586	0.00471	-54.37	<.0001

- **Intercept**:
  - **Parameter Estimate**: 62.05995 (predicted value of VPCS42 when age is 0).
  - **Standard Error**: 0.26574.
  - **t Value**: 233.53 (this very high t-value suggests the intercept is significantly different from zero).
  - **Pr > |t|**: <.0001 (p-value indicates the intercept is statistically significant).
- **age**:
  - **Parameter Estimate**: -0.25586 (for each year, VPCS42 decreases by about 0.256 units).
  - **Standard Error**: 0.00471.
  - **t Value**: -54.37 (indicates a very strong negative relationship).
  - **Pr > |t|**: <.0001 (confirms that age is a significant predictor).

Conclusion

Simple linear regression was used to test whether age significantly predicted VPCS42. The fitted regression model was:

VPCS42 = 62.05995 - 0.25586 * age

It was found that age significantly predicts VPCS42, with a coefficient $of\ \beta = -0.25586$ and a p-value of < 0.0001. The overall goodness of fit statistic, represented by R-squared, is 0.1701. This indicates that age accounts for approximately 17% of the variance in VPCS42.

While the low p-value demonstrates that age is indeed a significant predictor of VPCS42, the relatively low R-squared value suggests that age alone is not a strong predictor, as it explains only 17% of the variance. This indicates that there remains a considerable amount of variation (about 83%) in VPCS42 that is not accounted for by age alone.

Further steps may include conducting multivariate regression to include other relevant variables from the dataset (e.g., income, gender, chronic health conditions). This approach can help identify whether additional factors improve the model's predictive ability.