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.

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?

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. The two main sources here are the Model and Error.
  - **Model**
    - **DF (Degrees of Freedom):** 1 (there's one predictor, age).
    - **Sum of Squares:** 309,146 (this reflects the variation explained by the model).
    - **Mean Square:** 309,146.
  - **Error**
    - **DF (Degrees of Freedom):** 14,421 (this reflects the number of observations minus the number of parameters estimated).
    - **Sum of Squares:** 1,508,226 (this reflects the variation that is not explained by the model).
    - **Mean Square:** 104.58536.
  - **Corrected Total**
    - **Sum of Squares:** 1,817,371 (the total variation in the dependent variable).
- **F Value:** 2955.92 (this is the ratio of the Model Mean Square to the Error Mean Square). A high F-value indicates that the model explains a significant amount of variation relative to the error.
- **Pr > F:** <.0001 (this is the p-value associated with the F-test). Since it's very low, it indicates that the model is statistically significant; 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 (this is the standard deviation of the residuals, giving an idea of the average distance that the observed values fall from the regression line).
- **Dependent Mean**: 48.37372 (This value represents the average VPCS42 score across all observations in your dataset. It serves as a baseline for understanding how individual scores compare to this average.)
- **R-Square**: 0.1701 (approximately 17% of the variance in VPCS42 can be explained by age. This suggests that age is a weak predictor).
- **Adj R-Sq**: 0.17 (the adjusted R-squared, which accounts for the number of predictors in the model, is similar here since there's only one predictor).
- **Coeff Var**: 21.14102 (this is the coefficient of variation, which indicates the extent of variability in relation to the mean of the dependent variable).

Parameter Estimates

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 (this is the predicted value of VPCS42 when age is 0).
  - ○ **Standard Error**: 0.26574.
  - ○ **t Value**: 233.53 (a very high t-value suggests the intercept is significantly different from zero).
  - ○ **Pr > |t|**: <.0001 (indicating the intercept is statistically significant).
- **age**:
  - ○ **Parameter Estimate**: -0.25586 (for each additional year of age, VPCS42 decreases by about 0.256 units).
  - ○ **Standard Error**: 0.00471.
  - ○ **t Value**: -54.37 (indicating a very strong negative relationship).
  - ○ **Pr > |t|**: <.0001 (confirming 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 predicted VPCS42 (β = -0.25586, p-value < .0001). The overall goodness of fit statistic is R-squared = 0.1701. The residual plot shows some spread around zero, indicating that age accounts for approximately 17% of the variance in VPCS42. The relatively low R-squared suggests that there is still a considerable amount of variation (about 83%) that is not explained by age alone.