Exploring Health Discrepancies: Forecasting Diabetes Rates through an Examination of Obesity and Physical Inactivity

The issues:

In 2018, the CDC gathered extensive data on Diabetes, Inactivity, and Obesity in the United States. This investigation aims to uncover connections between these health indicators. It seeks to pinpoint the root causes of rising diabetes rates, explore links between obesity, inactivity, and diabetes, and develop predictive models. Data will be visualized creatively, and regional disparities will be analyzed. Additionally, the study delves into the impact of inactivity on public health and endeavors to construct precise predictive models for early intervention. The goal is to contribute insights for a healthier future.

The investigation also delves into the relationship between physical inactivity and the rates of diabetes and obesity, focusing on understanding how higher levels of inactivity impact public health. The challenge is to construct highly accurate predictive models, forecasting diabetes rates through a synthesis of obesity and inactivity data. These models can empower healthcare professionals with valuable tools for early intervention.

This investigation seeks to not only address the pressing issue of rising diabetes rates in the United States but also to deepen understanding of the intricate interplay between diabetes, inactivity, and obesity. Through rigorous data analysis and research methods, the aim is to uncover the multifaceted nature of these health challenges and pave the way for proactive strategies to create a healthier and more resilient future for all.

Findings:

My discovery indicates that the degree of physical inactivity exerts a more pronounced influence on diabetes when contrasted with obesity. The data implies that a lack of physical activity is more strongly associated with the risk of diabetes. This assertion is statistically substantiated by correlation coefficients and R-squared scores, affirming the heightened impact of inactivity on diabetes rates within our dataset.

In addition to that, this analysis reveals significant connections between diabetes and factors like physical inactivity and obesity. Addressing these factors can greatly impact diabetes management. Regional variations exist, guiding targeted interventions. Predictive models help identify potential hotspots for timely interventions, akin to discovering a rare gem. These insights, when applied, can lead to a healthier future.
Discussions:

Positive Coefficients and Diabetes: The presence of positive coefficients for inactivity and obesity in our analysis implies a clear link with diabetes rates. This observation aligns with established research, which has consistently shown a connection between sedentary behavior, obesity, and the prevalence of diabetes.

Model's Explanatory Power: The R-squared values indicate that our model explains a moderate portion of the variation in diabetes rates. This suggests that while inactivity and obesity are influential, other factors contribute significantly to diabetes rates. To enhance the model's effectiveness, it may be valuable to explore and incorporate additional variables.

Addressing Variability: The initial model highlighted the presence of heteroskedasticity, implying that error variability is not consistent across all levels of independent variables. I mitigated this issue to some extent by applying a log transformation to diabetes and a square root transformation to obesity.

Potential for Intervention: Given the positive associations observed between inactivity, obesity, and diabetes rates, there is potential for targeted interventions aimed at reducing diabetes rates. Public health policies promoting physical activity and healthy weight management could yield positive outcomes in this regard.

Room for Further Investigation: Our findings suggest avenues for further exploration. It is advisable to investigate interactions between variables, incorporate additional features, and assess the impact of potential outliers or influential data points. Collaborating with local stakeholders and health authorities can provide a more comprehensive understanding of the specific factors influencing diabetes rates in outlier counties.

Assessing Model Accuracy: To gauge the accuracy of our predictive model, I used the metric known as R-squared. Before applying transformations, the Test p-value was approximately 3.5 for the training set and 0.27 for the test set. This indicates that our model explains around 35% of the variation in diabetes rates within the training set and 27% within the test set.

Appendix A: Method

Data Collection:

This data collection process involved surveying adults aged 18 or older in 2018 to gather information on diagnosed diabetes, new cases, obesity (defined as BMI ≥30), and physical inactivity at the county level. Participants provided self-reports regarding their diabetes status, new cases, obesity, and physical inactivity.
Variable Collection: The dataset is centered around three primary variables: "Inactivity Percentage," "Obesity Percentage," and "Diabetes Percentage." These variables were recorded at the county level, and each record was uniquely identified by a corresponding FIPS code. To construct a comprehensive dataset with shared data points, i executed a merging operation on three distinct tables using FIPS codes as common identifiers.

Data Analysis Methods:

Data Merging: To consolidate the data effectively, i merged three separate tables, each representing inactivity percentage, obesity percentage, and diabetes percentage. This merging process enabled us to create a unified dataset that incorporated relevant information from all three tables.

Descriptive Statistics: Descriptive statistics were computed, including a 5-point summary (minimum, 25th percentile, median, 75th percentile, and maximum) for the variables: inactivity, obesity, and diabetes percentages. Additionally, i generated individual histogram plots to visualize the distribution of each variable. Our analysis revealed that inactivity exhibited a slight left-skew, obesity displayed a more pronounced left-skew, while diabetes skewed to the right.

Data Splitting: To evaluate the performance of our model on unseen data, i divided the dataset into training and testing sets. The training set comprised 80% of the data, while the testing set constituted the remaining 20%. i ensured reproducibility by specifying the 'random state' parameter.

Linear Regression Modeling: Multiple linear regression was applied, using inactivity and obesity (X) as predictors and diabetes percentage (y) as the target variable. i derived coefficients for the intercept, inactivity, and obesity. The model was fitted to the training set, and its performance was assessed using R-squared values on both the training and testing sets.

Residual Analysis: Residuals were calculated for the training set. Additionally, we conducted the Breusch-Pagan test to identify heteroskedasticity. The test revealed evidence of heteroskedasticity, indicated by a significant p-value.

Cross-Validation: To evaluate the generalizability of our model, i employed 5-fold cross-validation. This process yielded an array of R-squared values for each fold. i then computed the range, standard deviation, and interquartile range of these cross-validated R-squared values.

Transformation to Address Heteroskedasticity  In response to the identified heteroskedasticity, i applied transformations to the data. Specifically, a log transformation was applied to the diabetes data, and a square root transformation was applied to the obesity data. New coefficients were obtained for the transformed model, which was subsequently fitted to the training set and evaluated on the testing set.
Residual Analysis Post-Transformation: Following the transformation, recalculated residuals for the training set. The Breusch-Pagan test was conducted once more, revealing no evidence of heteroskedasticity, as indicated by a Breusch-Pagan Test p-value: 3.55

Cross-Validation Post-Transformation: The transformed model underwent cross-validation, and we assessed the range, standard deviation, and interquartile range of the cross-validated R-squared values.

Appendix B: Results

Simple Linear Regression Model:
I generated basic linear regression plots to gain a clearer view of the data, providing some insights into its arrangement, although it did not offer a comprehensive understanding of its underlying characteristics.

Mean Squared Error (MSE): 0.34883328631124316

R-squared (R2) Value: 0.34073967115731396

Coefficients: [0.11106297 0.23246992]
Intercept: 1.6535991518559383

Figures I use a histograms to visually and convey these distributions
Multiple Linear Regression Model:
Basic Statistics for Obesity and Inactivity vs. Diabetes:

<table>
<thead>
<tr>
<th>% OBESE</th>
<th>% INACTIVE</th>
<th>% DIABETIC</th>
</tr>
</thead>
</table>

Mean Squared Error (MSE): 0.34883328631124316

R-squared (R²) Value: 0.34073967115731396

Coefficients: [0.11106297, 0.23246992]
  Intercept: 1.6535991518559383

count 354.000000 354.000000 354.000000
mean 18.252542 14.776271 7.115819 std
1.029484 1.544542 0.728442 min 10.500000
8.800000 3.800000
25% 17.900000 14.000000 6.800000
50% 18.300000 14.400000
  7.000000 75% 18.975000
  15.475000 max 7.400000

The multiple linear regression graph utilizes %Inactivity and %Diabetes as independent variables to predict %Diabetes, offering us a means to contrast the actual and anticipated %Diabetes. However, this model produces an R-squared score of 0.34, which falls short of meeting our expected level of contentment.

**Quadratic model on log-transformed data:**
R-squared (quadratic model on log-transformed data): 0.43. The model we constructed by implementing a logarithmic transformation to the %Diabetes data, in combination with Inactivity and Obesity, generated an R-squared score of 0.43, which is the highest we have observed thus far. Unless an alternative approach yields a superior result, we will regard this model as the most appropriate for forecasting %Diabetes using our dataset.

**Breusch-Pagan Test Output:**

<table>
<thead>
<tr>
<th>Breusch-Pagan Test p-value</th>
<th>3.555846910402186e-05</th>
</tr>
</thead>
<tbody>
<tr>
<td>Heteroscedasticity is present (reject the null hypothesis).</td>
<td></td>
</tr>
</tbody>
</table>

The Breusch-Pagan Test, which assesses the p-value and the pattern of residuals, reveals a significantly low p-value, indicating the presence of heteroscedasticity in the

**K-fold validation test Output:**

<table>
<thead>
<tr>
<th>Fold</th>
<th>R-squared</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.2670</td>
</tr>
<tr>
<td>2</td>
<td>0.1851</td>
</tr>
<tr>
<td>3</td>
<td>0.2947</td>
</tr>
<tr>
<td>4</td>
<td>0.4568</td>
</tr>
<tr>
<td>5</td>
<td>0.1535</td>
</tr>
</tbody>
</table>
Mean R-squared: 0.2732
Standard Deviation of R-squared: 0.1352

We applied K-fold cross-validation using 5 iterations to assess the R-squared value for improved %Diabetes predictions. However, the obtained R-squared value (0.27) fell short of our best result (0.45 with the quadratic model). As a result, we decided not to proceed with this method.

**T-Test output:**

Reject the null hypothesis. There is a significant difference between the datasets.

T-statistic: -8.586734600367794
P-value: 1.960253729590773e-17

When conducting the t-test, we obtained a significantly low p-value, indicating that the data exhibits heteroscedasticity. Nevertheless, it's crucial to acknowledge that the t-test results may not be highly dependable due to the non-normal distribution of the data points.

**Appendix C: code**

```python
import numpy as np
import pandas as pd
from sklearn.linear_model import LinearRegression
from sklearn.preprocessing import PolynomialFeatures
from sklearn.metrics import sm
```

```python
import numpy as np
import pandas as pd
import statsmodels.api as sm
```
import matplotlib.pyplot as plt

# Load your dataset containing 'INACTIVITY', '%OBESITY',
'%DIABETES'

data = pd.read_csv('your_dataset.csv')  # Replace 'your_dataset.csv'
with your dataset file

# Perform the log transformation

data['% OBESITY_log'] = np.log(data['%OBESITY'])

data['% INACTIVITY_log'] = np.log(data['INACTIVITY'])

data['% DIABETES_log'] = np.log(data['%DIABETES'])

# Create squared terms for log-transformed variables

data['% OBESITY_log_squared'] = data['% OBESITY_log'] ** 2
data['% INACTIVITY_log_squared'] = data['% INACTIVITY_log'] ** 2

data['% DIABETES_log_squared'] = data['% DIABETES_log'] ** 2

# Define your independent variables (including the squared terms)

X = data[['% OBESITY_log', '% INACTIVITY_log', '% DIABETES_log',
          '% OBESITY_log_squared', '% INACTIVITY_log_squared', '% DIABETES_log_squared']]

# Add a constant to the independent variables for the intercept

X = sm.add_constant(X)

# Define your dependent variable

y = data['% DIABETES_log']
# Fit the quadratic model

model = sm.OLS(y, X).fit()

# Print the model summary

print(model.summary())

import statsmodels.api as sm

from statsmodels.stats.diagnostic import het_breuschpagan

# Assuming you have already defined your independent variables X
and dependent variable y

# Fit your linear regression model

model = sm.OLS(y, X).fit()
# Collect the residuals

residuals = model.resid

# Run the Breusch-Pagan test

lm, lm_p_value, f_value, f_p_value = het_breuschpagan(residuals, X)

# Print the test results

print(f"LM Statistic: {lm}")
print(f"LM-Test p-value: {lm_p_value}")
print(f"F-Statistic: {f_value}")
print(f"F-Test p-value: {f_p_value}")

# Interpret the results

if lm_p_value < 0.05 or f_p_value < 0.05:
print("Heteroscedasticity detected: The residuals have varying variances across levels of the independent variables.")
else:
    print("No significant evidence of heteroscedasticity: The residuals have roughly constant variances.")

Output: R-squared=0.45

Contribution:

Eyob Tsegaye Yosefe