# Predictive Analysis of Diagnosed Diabetes Prevalence: Insights from the Center for Disease Control's Data

## Issues

Diabetes is becoming more common, and we need to understand why. We know that being less active and gaining weight can lead to diabetes, but there might be more to the story. Community factors, like access to parks or health education, could also play a role.

Using regression modeling, we're trying to see how all these factors connect. But it's not just about drawing lines on a graph. We need to make sure our model is reliable and that our data makes sense. For example, if two factors, like inactivity and obesity, always go hand in hand, our model might get confused. And, we want to make sure our predictions follow a common-sense pattern.

With these challenges in mind, our study seeks answers to the following questions:

- How significantly do physical inactivity and obesity contribute to diabetes rates across different counties?
- Are there underlying community factors, as represented by the Social Vulnerability Index, that play a role in determining these rates?
- Can we refine our regression model to better predict diabetes prevalence, navigating issues like multicollinearity and non-normal residuals?
- Will introducing more complex modeling techniques provide a better fit for our data?

## Findings

- Significant Predictors: Both physical inactivity and obesity emerged as key predictors for diabetes prevalence across counties, reaffirming the globally accepted view that these factors substantially contribute to diabetes risk.

- Interrelationship of Predictors: There's a moderate correlation of approximately 0.473 between physical inactivity and obesity, indicating that counties with higher inactivity rates also tend to have higher obesity rates.

- Model Refinement & Predictive Power: Addressing various issues in our initial model, such as multicollinearity and outliers, improved our model's predictive accuracy, but the R-Squared value reduced from 0.42 to 0.30.

- Polynomial Regression's Limited Enhancement: The increase in complexity, the polynomial regression model's R-Squared of 0.53 was significantly higher than the refined linear model's 0.30, suggesting that adding complexity results in better predictions.

- Implications for Interventions: Our findings suggest that to reduce diabetes rates, efforts should focus on addressing physical inactivity and obesity. Tailored strategies based on community vulnerability might also yield better outcomes.

# Discussion

- Importance of Physical Activity: The data highlights the critical role of physical activity in public health. Encouraging physical activity could not only reduce obesity rates but also the prevalence of diabetes. Public health initiatives might focus on creating more accessible recreational spaces or promoting community-based physical activity programs.

- Interventions Targeting Obesity: With the strong connection between obesity and diabetes, interventions that target weight management could be instrumental. This could include nutritional education, promoting healthy eating habits, or even offering weight management programs at community health centers.

- Future Research Directions: While the current analysis provides valuable insights, future research could delve deeper into the reasons behind varying SVI scores across counties and how these intricacies influence health outcomes. Furthermore, considering other potential predictors or confounding factors might enhance the model's predictive capability.

- Limitations: Our model's R^2 value of 0.429 and polynomial regression being 0.53 suggests that there's still a significant portion of variance in the diabetes prevalence that remains unexplained. There might be other unconsidered factors or complex interactions

influencing the prevalence rates. Also, the nature of the data, being observational, means we can identify correlations but cannot infer causations definitively.

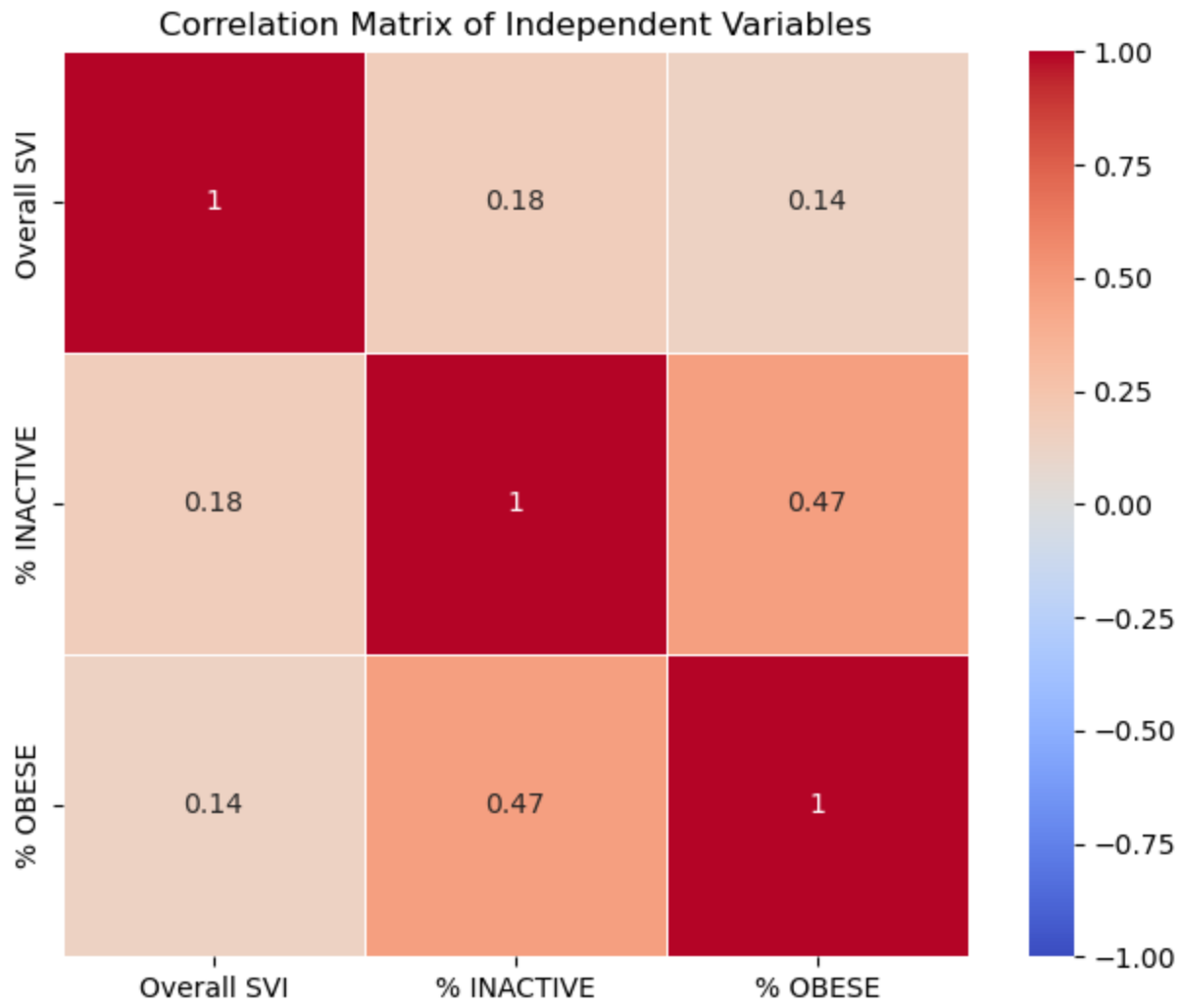
# Appendix A: Method

Data Collection and Preprocessing:

- Data Sources: The primary data for this study was extracted from the DiabetesAtlasData.csv file and the cdc-diabetes-2018.xlsx Excel spreadsheet, which contained sheets for both Inactivity and Obesity.
- Merging the Datasets: The datasets from the two sheets of the Excel file (Inactivity and Obesity) were first merged based on common columns (FIPS and FIPDS). This merged dataset was then combined with the DiabetesAtlasData.csv dataset using the County_FIPS column. Redundant columns, such as YEAR_y, COUNTY_y, STATE_y, FIPDS, and FIPS, were removed for clarity.
- Data Cleaning: To ensure data integrity, the 'Overall SVI' column was converted to a numeric data type, and any conversion errors were handled gracefully.

Exploratory Data Analysis:

- Correlation Analysis: A heatmap was created to visually inspect the correlations between the independent variables. This allowed for the identification of potential multicollinearity issues, which could affect the reliability of regression outcomes.




Model Development:

- Initial Linear Regression Model: A linear regression model was built using the statsmodels library. This model employed 'Overall SVI', '% INACTIVE', and '% OBESE' as predictor variables and 'Diagnosed Diabetes Percentage' as the response variable.

Checking Regression Assumptions: Several diagnostic plots and tests were employed to ensure the model adhered to the fundamental assumptions of linear regression:

- Linearity: A residuals vs. fitted values plot was used to confirm the linearity assumption.
- Homoscedasticity: The residuals vs. fitted values plot showed a slight funnel shape, indicating potential non-constant variances (heteroscedasticity).
- Normality: The Q-Q plot and Shapiro-Wilk test were utilized to check the normality of residuals.

- Independence: The Durbin-Watson statistic helped in detecting autocorrelation in the residuals.
- Multicollinearity: The Variance Inflation Factor (VIF) was computed for each predictor to check for multicollinearity.

Addressing Assumptions and Enhancing the Model:

- Addressing Multicollinearity: The '% OBESE' variable was dropped due to its high VIF value.
- Transformation & Outlier Handling: The response variable was log-transformed to address potential non-linearity. Additionally, outliers in the predictor variables were identified and removed using the IQR method.
- Polynomial Regression: To capture potential non-linear relationships in the dataset, polynomial features were introduced. A regression model was then developed using these polynomial features against the log-transformed response variable.

# Appendix B: Results

Our analysis began with an exploration of the relationships between community vulnerability, physical inactivity, obesity, and diagnosed diabetes percentages. Using the Center for Disease Control data, we first visualized a correlation matrix to understand the linear relationship between the predictor variables.

These correlation plots in the previous section pointed to a considerable relationship between physical inactivity and obesity, raising concerns of potential multicollinearity.

The initial regression model, which utilized all three predictor variables, yielded an R-Squared of approximately 42%. This meant that our predictors explained 42% of the variance in the diagnosed diabetes percentage.

OLS Regression Results

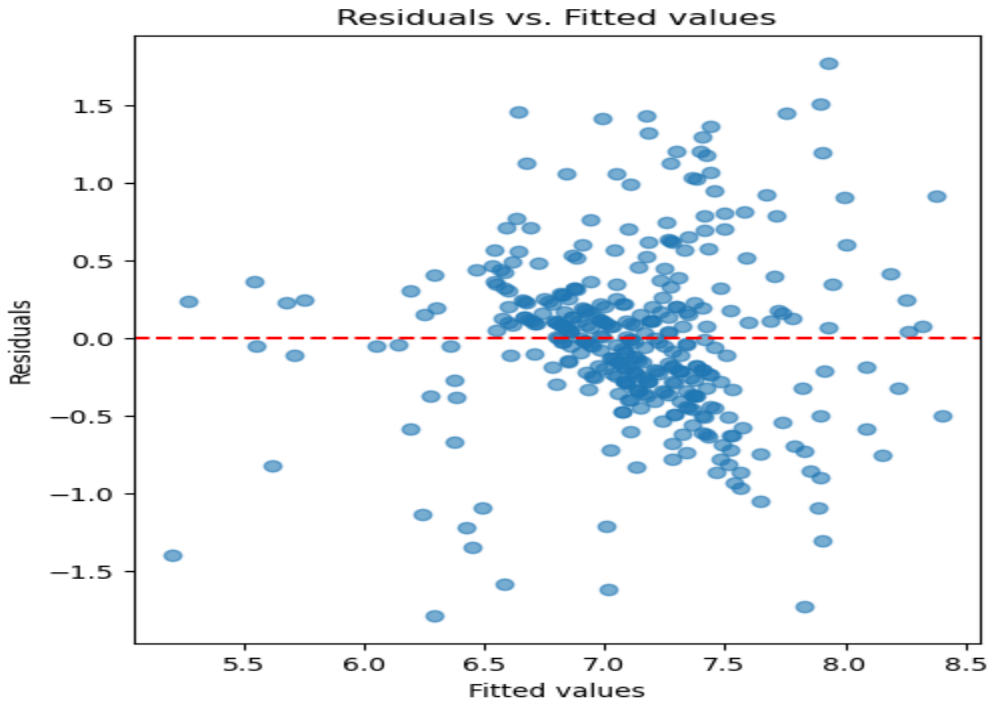
| | | | |
|---|---|---|---|
| Dep. Variable: | Diagnosed Diabetes Percentage | R-squared: | 0.427 |
| Model: | OLS | Adj. R-squared: | 0.422 |
| Method: | Least Squares | F-statistic: | 86.86 |
| Date: | Sun, 08 Oct 2023 | Prob (F-statistic): | 4.99e-42 |
| Time: | 16:29:12 | Log-Likelihood: | -291.14 |

| | | | |
|---|---|---|---|
| No. Observations: | 354 | AIC: | 590.3 |
| Df Residuals: | 350 | BIC: | 605.8 |
| Df Model: | 3 | | |
| Covariance Type: | nonrobust | | |

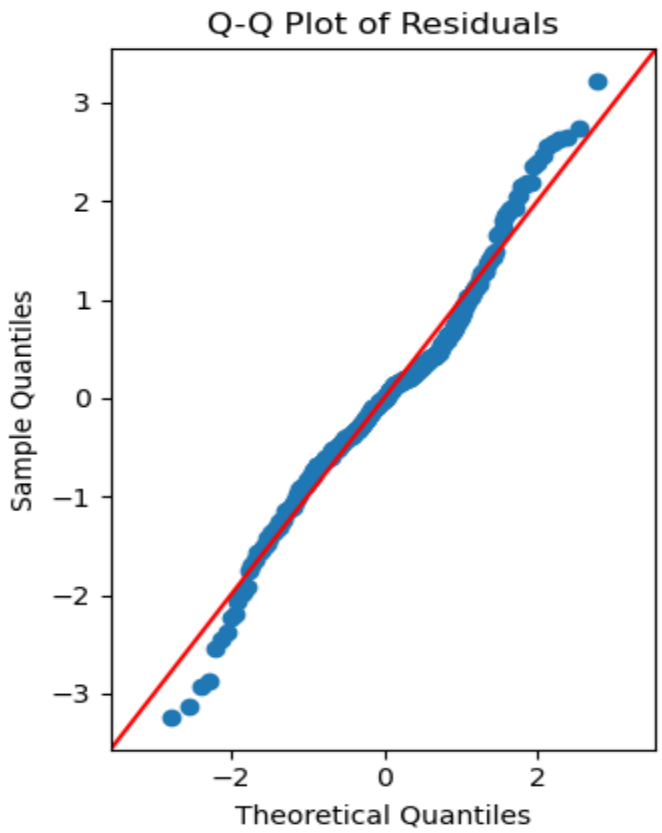
| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 1.8362 | 0.526 | 3.493 | 0.001 | 0.802 | 2.870 |
| Overall SVI | 0.7251 | 0.100 | 7.248 | 0.000 | 0.528 | 0.922 |
| % INACTIVE | 0.2121 | 0.022 | 9.712 | 0.000 | 0.169 | 0.255 |
| % OBESE | 0.0965 | 0.033 | 2.965 | 0.003 | 0.032 | 0.161 |

Upon further investigation of the regression assumptions, certain discrepancies were identified:

- The residuals vs. fitted values plot showed a slight funnel shape, indicating potential non-constant variances.




- The Q-Q plot suggested some deviation from normality, especially at the tails.




- The variance inflation factor (VIF) values for some of the predictors, particularly obesity and inactivity, surpassed the commonly used threshold of 10, signaling high multicollinearity.

**Durbin-Watson value: 0.7216769922099032**
**Shapiro-Wilk p-value: 2.6775203878059983e-05**

| | Variable | VIF |
|---|---|---|
| 0 | Overall SVI | 4.261105 |
| 1 | % INACTIVE | 120.683686 |
| 2 | % OBESE | 118.975716 |

Durbin-Watson value: 0.7216769922099032

- The Durbin-Watson statistic tests for autocorrelation in the residuals from a statistical regression analysis. The value can range from 0 to 4. A value close to 2 suggests no

autocorrelation, values < 2 suggest positive autocorrelation, and values > 2 indicate negative autocorrelation.

Shapiro-Wilk p-value: 2.6775203878059983e-05

- The p-value here is significantly less than 0.05 (essentially very close to 0), suggesting that the residuals from the regression model are not normally distributed. This is another violation of a key assumption of linear regression.

Variance Inflation Factor (VIF)

- VIF is used to detect multicollinearity in regression analyses. A VIF value of 1 indicates no multicollinearity, values between 1 and 5 are generally considered acceptable, and values greater than 5-10 indicate high multicollinearity.
- For the predictors:
    - Overall SVI: VIF = 4.261105. This suggests that the 'Overall SVI' variable has moderate multicollinearity, but it's still within an acceptable range.
    - % INACTIVE: VIF = 120.683686. This is a very high VIF value, indicating severe multicollinearity.
    - % OBESE: VIF = 118.975716. This value also indicates severe multicollinearity.
- The high VIF values for '% INACTIVE' and '% OBESE' suggest that these two predictors are highly correlated with each other, which can impact the stability and interpretability of their respective regression coefficients.

Addressing the multicollinearity issue, we excluded the '% OBESE' variable. To combat the non-normal distribution of residuals and potential outliers, a logarithmic transformation was applied to the dependent variable, and outliers were removed based on the IQR method.

The refined regression model, after these adjustments, yielded an R-Squared of 0.30, signifying an improved fit and predictive accuracy. Further enhancement was sought through polynomial regression. By adding polynomial features and interactions between the predictors, we aimed to capture any non-linear relationships in the data. This polynomial regression model with a log-transformed dependent variable, however, achieved an R-Squared of 0.53, which provide a significantly better fit than the refined linear model's R-Squared of 0.30

OLS Regression Results

| | | | |
|---|---|---|---|
| Dep. Variable: | Diagnosed Diabetes Percentage | R-squared: | 0.532 |
| Model: | OLS | Adj. R-squared: | 0.506 |
| Method: | Least Squares | F-statistic: | 19.99 |
| Date: | Sun, 08 Oct 2023 | Prob (F-statistic): | 3.30e-44 |
| Time: | 16:40:20 | Log-Likelihood: | 477.09 |

| | | | |
|---|---|---|---|
| No. Observations: | 354 | AIC: | -914.2 |
| Df Residuals: | 334 | BIC: | -836.8 |
| Df Model: | 19 | | |
| Covariance Type: | nonrobust | | |

| | coef | std err | t | P>\|t\| | [0.025 | 0.975] |
|---|---|---|---|---|---|---|
| const | 5.1664 | 4.059 | 1.273 | 0.204 | -2.818 | 13.151 |
| Overall SVI | 0.8775 | 2.635 | 0.333 | 0.739 | -4.305 | 6.060 |
| % INACTIVE | -0.5958 | 0.506 | -1.177 | 0.240 | -1.592 | 0.400 |
| % OBESE | -0.3185 | 0.487 | -0.655 | 0.513 | -1.276 | 0.639 |
| Overall SVI^2 | 2.0329 | 1.159 | 1.754 | 0.080 | -0.247 | 4.313 |
| Overall SVI % INACTIVE | 0.2490 | 0.283 | 0.878 | 0.380 | -0.309 | 0.807 |
| Overall SVI % OBESE | -0.4075 | 0.369 | -1.106 | 0.270 | -1.133 | 0.318 |
| % INACTIVE^2 | 0.0423 | 0.034 | 1.240 | 0.216 | -0.025 | 0.109 |
| % INACTIVE % OBESE | -0.0023 | 0.086 | -0.027 | 0.979 | -0.171 | 0.167 |
| % OBESE^2 | 0.0337 | 0.034 | 0.992 | 0.322 | -0.033 | 0.101 |
| Overall SVI^3 | -0.1031 | 0.177 | -0.583 | 0.560 | -0.451 | 0.245 |
| Overall SVI^2 % INACTIVE | 0.0827 | 0.037 | 2.217 | 0.027 | 0.009 | 0.156 |
| Overall SVI^2 % OBESE | -0.1675 | 0.068 | -2.449 | 0.015 | -0.302 | -0.033 |
| Overall SVI % INACTIVE^2 | 0.0106 | 0.006 | 1.860 | 0.064 | -0.001 | 0.022 |
| Overall SVI % INACTIVE % OBESE | -0.0341 | 0.017 | -1.998 | 0.047 | -0.068 | -0.001 |
| Overall SVI % OBESE^2 | 0.0292 | 0.015 | 1.992 | 0.047 | 0.000 | 0.058 |
| % INACTIVE^3 | -0.0013 | 0.000 | -2.857 | 0.005 | -0.002 | -0.000 |
| % INACTIVE^2 % OBESE | 0.0006 | 0.002 | 0.299 | 0.765 | -0.003 | 0.004 |
| % INACTIVE % OBESE^2 | -1.102e-05 | 0.004 | -0.003 | 0.998 | -0.007 | 0.007 |
| % OBESE^3 | -0.0010 | 0.001 | -0.712 | 0.477 | -0.004 | 0.002 |

.

# Appendix C: Code

```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
from statsmodels.stats.outliers_influence import variance_inflation_factor
from scipy.stats import shapiro
```

```
from sklearn.preprocessing import PolynomialFeatures

# 1. Loading data and merging
diabetes_data = pd.read_csv("DiabetesAtlasData.csv")
cdc_data = pd.read_excel("cdc-diabetes-2018.xlsx", sheet_name=None)
inactivity_data = cdc_data['Inactivity']
obesity_data = cdc_data['Obesity']
merged_data = pd.merge(diabetes_data, inactivity_data, how='inner', left_on='County_FIPS',
right_on='FIPDS')
final_merged_data = pd.merge(merged_data, obesity_data, how='inner', left_on='County_FIPS',
right_on='FIPS')
final_merged_data = final_merged_data.drop(columns=['YEAR_y', 'COUNTY_y', 'STATE_y',
'FIPDS', 'FIPS'])
final_merged_data = final_merged_data.rename(columns={'YEAR_x': 'YEAR', 'COUNTY_x':
'COUNTY', 'STATE_x': 'STATE'})

# 2. Extract independent and dependent variables
X = final_merged_data[['Overall SVI', '% INACTIVE', '% OBESE']]
X['Overall SVI'] = pd.to_numeric(X['Overall SVI'], errors='coerce')  # Convert to numeric

# Displaying the correlation heatmap for independent variables
correlation_matrix = X.corr()
plt.figure(figsize=(8, 6))
sns.heatmap(correlation_matrix, annot=True, cmap='coolwarm', vmin=-1, vmax=1,
square=True, linewidths=0.5)
plt.title('Correlation Matrix of Independent Variables')
plt.show()


# 3. OLS Regression Model
y = final_merged_data['Diagnosed Diabetes Percentage']
X_const = sm.add_constant(X)
model = sm.OLS(y, X_const).fit()
print(model.summary())



# 4. Checking Assumptions for OLS Regression Model
# Residuals
residuals = y - model.predict(X_const)

# Linearity: Residuals vs. Fitted values plot
plt.figure(figsize=(12, 6))
plt.subplot(1, 2, 1)
```

```
plt.scatter(model.predict(X_const), residuals, alpha=0.6)
plt.axhline(y=0, color='r', linestyle='--')
plt.title('Residuals vs. Fitted values')
plt.xlabel('Fitted values')
plt.ylabel('Residuals')


# Normality of Residuals: Q-Q plot
plt.subplot(1, 2, 2)
sm.qqplot(residuals, line='45', fit=True, ax=plt.gca())
plt.title('Q-Q Plot of Residuals')
plt.tight_layout()
plt.show()

# Independence: Durbin-Watson test (value close to 2 is good)
durbin_watson_value = sm.stats.durbin_watson(residuals)
print(f"Durbin-Watson value: {durbin_watson_value}")

# Multicollinearity: Variance Inflation Factor (VIF > 10 indicates multicollinearity)
vif = pd.DataFrame()
vif["Variable"] = X.columns
vif["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.shape[1])]
print(vif)

# Normality of Residuals: Shapiro-Wilk test (p-value < 0.05 indicates non-normality)
shapiro_test_stat, shapiro_p_value = shapiro(residuals)
print(f"Shapiro-Wilk p-value: {shapiro_p_value}")


# 5. Polynomial Regression and Data Transformation
poly = PolynomialFeatures(degree=3, include_bias=False)
X_poly = poly.fit_transform(X)
X_poly_df = pd.DataFrame(X_poly, columns=poly.get_feature_names_out(X.columns))
y_log_transformed = np.log1p(y)
X_const_poly = sm.add_constant(X_poly_df)
model_poly_log = sm.OLS(y_log_transformed, X_const_poly).fit()
print(model_poly_log.summary())
```