An Examination of CDC Data on Diabetes, Obesity, and Physical Inactivity

Co-Authored By:

Aditya Domala - 02096198

Mokesh Balakrishnan - 02126912

Ruksar Lukade - 02137513

Issues:

This report explores the intricate link between lifestyle choices, focusing on obesity and limited physical activity, and their profound impact on diabetes prevalence. Our goal is to unravel the nuanced ways in which these factors influence the progression of diabetes. We've closely scrutinized the comprehensive and up-to-date database maintained by the Centers for Disease Control and Prevention (CDC) in the United States. Our analysis of this database covers data on obesity, sedentary lifestyles, and county-level diabetes rates for the year 2018. Through this investigation, we are committed to uncovering key insights and answers to crucial questions.

How does the percentage of physical inactivity vary across different states and counties? How does this inactivity contribute to the prevalence of obesity, high blood pressure, high cholesterol, and diabetes?

What is the distribution of obesity percentages across different states and counties?

How does obesity correlate with the incidence of diabetes and cardiovascular diseases?

How does the percentage of diabetes vary across different states and counties?

How do obesity and physical inactivity contribute to the prevalence of diabetes?

Through the comprehensive analysis, we intend to illuminate the critical health issues related to diabetes, obesity, and physical inactivity in the United States, ultimately offering insights that can inform evidence-based policies and interventions for better public health outcomes.

The ultimate goal of this research is to shed light on the pressing health issues of diabetes, obesity, and lack of physical activity in the United States.

Findings:

Our analysis provides illuminating insights into the relationship between lifestyle choices and diabetes. A key discovery is the statistically significant impact of physical inactivity on diabetes rates, which is approximately 1.5 times more influential than obesity. Despite the strong association between inactivity and obesity, our findings demonstrate that inactivity is an independent and more powerful predictor of diabetes prevalence. This suggests that while

obesity is undeniably a factor in diabetes development, interventions focused on increasing physical activity could be more impactful.

Inactivity vs. Obesity: Our findings indicate that physical inactivity contributes to 34.1% of the variation in obesity rates across various counties, signifying a moderate level of association. The statistical significance of this association is highlighted by a low p-value, emphasizing a substantial connection between inactivity and obesity. The analysis of residual distributions revealed a slight negative skew and a leptokurtic trend, indicating the presence of several outliers.

Inactivity vs. Diabetes: The investigation showed that inactivity is responsible for 24.5% of the variability in diabetes rates. While significant, this correlation is less pronounced compared to that with obesity. The statistical significance of this finding reinforces the impact of inactivity on diabetes prevalence. The distribution of residuals was found to be left-skewed with a high kurtosis, suggesting the presence of outliers in the data.

Obesity vs. Diabetes: obesity is associated with 39.6% of the variability in diabetes rates, marking a stronger link compared to inactivity. Nonetheless, the absence of statistical significance in this correlation suggests that obesity might not be a consistent predictor of diabetes prevalence across all counties. The residual analysis showed a right-skewed distribution and a leptokurtic pattern, indicating the presence of outliers in this aspect of the data as well.

Discussions:

Our study's analysis reveals critical insights into the interplay between obesity, inactivity, and diabetes prevalence within various counties.We explored the dynamics of three key relationships:

Inactivity and Obesity's Role in Diabetes: Our primary focus was to evaluate how the rates of physical inactivity and obesity impact diabetes prevalence. By considering the percentage of inactive and obese individuals as independent variables, and the diabetes rate as the dependent variable, we aimed to quantify their combined effect on diabetes occurrence. This assessment helps in understanding how lifestyle factors contribute to diabetes risk.

Inactivity and Diabetes as Predictors of Obesity: The study also investigated how physical inactivity and diabetes prevalence might forecast obesity levels. Here, the percentages of inactive individuals and those with diabetes were the independent variables, while the obesity rate was the dependent variable. This analysis is crucial for understanding whether diabetes and lack of exercise can be significant predictors of obesity.

Obesity and Diabetes as Indicators of Physical Inactivity: Lastly, we examined how obesity and diabetes rates might predict physical inactivity levels. This aspect of the study, using obesity and diabetes percentages as independent variables and the inactivity rate as the dependent variable, aimed to determine if health conditions like obesity and diabetes are reliable indicators of a sedentary lifestyle.

Appendix A: Method

Data Collection

The dataset for this study was acquired to explore health metrics across 3,143 counties in the United States, focusing on the year 2018. This data, detailing diabetic rates, physical inactivity, and obesity statistics, was sourced to understand the health dynamics within the counties. The CDC dataset was segregated into spreadsheets, containing 1,361 counties with inactivity rates and 364 counties with obesity rates. This process aimed to utilize comprehensive and up-to-date health data to inform our understanding of health behaviours and outcomes at the county level.

Variable Creation

In this analysis, we defined several key variables to investigate the complex relationships between health-related factors:

% DIABETIC: This variable represents the percentage of the population within a county diagnosed with diabetes, a chronic health condition that affects the body's ability to regulate blood sugar levels.

% OBESE: This variable quantifies the percentage of the county's population that falls into the category of obesity, which is typically defined by a Body Mass Index (BMI) of 30 or higher. Obesity is a critical health concern associated with various diseases, including diabetes and cardiovascular issues.

% INACTIVE: This variable indicates the percentage of individuals in the county who are not engaging in regular physical activity. Physical inactivity is a risk factor for many health conditions, including obesity and diabetes.

To explore potential nonlinear relationships and interactions among these variables, we engineered new variables such as '% OBESE squared' and '% INACTIVE squared'. These were calculated by squaring the original percentage values, allowing us to assess how incremental changes in obesity and inactivity levels might have disproportionate effects on health outcomes.

Analytic Methods

We employed Ordinary Least Squares (OLS) Regression to investigate the associations between the independent variables—percentages of obesity (% OBESE) and inactivity (% INACTIVE)—and the dependent variable, the percentage of the diabetic population (% DIABETIC). OLS regression is particularly useful for determining the magnitude and direction of relationships between variables. To ensure the model's reliability and predictive accuracy, cross-validation was conducted. This process involved partitioning the data into separate subsets to train and test the model, providing an evaluation of its generalizability to new, unseen data. Additionally, to ascertain the independence of the predictors and to detect multicollinearity, we calculated the Variance Inflation Factor (VIF) for each variable. High VIF values indicate potential redundancy among independent variables, which can compromise the model's estimates. Lastly, visualisation techniques, specifically scatter plots, were utilised to compare the actual data points against the model's predictions, offering a visual assessment of the model's accuracy and the fit of the regression line.

Model Evaluation

Metrics including Rsquared, standard error, and significance tests on coefficients were used to assess the model's performance. To evaluate the model's underlying assumptions and the features of the data distribution, residual analysis was used. We used k fold cross validation, and the findings offer a thorough evaluation of our model's functionality and appropriateness for the task at hand. We are able to make well-informed conclusions regarding the model's

suitability for deployment or further development by employing this thorough evaluation method.

Appendix B: Result

OLS Regression Results

Dep. Variable:	% DIABETIC	R-squared:	0.341
Model:	OLS	Adj. R-squared:	0.337
Method:	Least Squares	F-statistic:	90.71
Date:	Sun, 03 Dec 2023	Prob (F-statistic):	1.76e-32
Time:	20:59:26	Log-Likelihood:	-315.89
No. Observations:	354	AIC:	637.8
Df Residuals:	351	BIC:	649.4
Df Model:	2		
Covariance Type:	nonrobust		

OLS Regression Results of Diabetic

Dep. Variable indicates the dependent variable, In this case, the percentage of diabetics is individuals.The model that states the OLS regression model has been used. OLS is a common method for estimating the linear regression equation.The Least Squares approach is utilised, which means that the coefficients are determined by minimising the sum of squared residuals.The adjusted R-squared which takes into account the number of predictors in the model and adjusts for the sample size. It's a more accurate measure when comparing models with different numbers of predictors.F-statistic analyses A greater number generally suggests that the model has a better fit for the data if at least one predictor variable has a non-zero coefficient. We obtained an F-statistic of 90.71, which is statistically significant according to the Prob (F-statistic).

	coef	std err	t	P>|t|	[0.025	0.975]
const	1.6536	0.562	2.941	0.003	0.548	2.759
% INACTIVE	0.2325	0.023	10.023	0.000	0.187	0.278
% OBESE	0.1111	0.035	3.192	0.002	0.043	0.180

Omnibus:	17.281	Durbin-Watson:	1.673
Prob(Omnibus):	0.000	Jarque-Bera (JB):	45.622
Skew:	-0.042	Prob(JB):	1.24e-10
Kurtosis:	4.757	Cond. No.	421.

Shows Inactive (%) vs Obese(%)

Coefficients are the estimated values for the regression coefficients. Each coefficient represents the change in the dependent variable for a one-unit change in the corresponding independent variable, holding other variables constant. const intercept, the estimated value of the dependent variable when all the independent variables are zero. our value is 1.6536. The coefficient for the independent variable "% INACTIVE" is 0.2325. This suggests that a one percentage point increase in inactivity is associated with a 0.2325 unit increase in the percentage of diabetic individuals. The coefficient for "% OBESE" is 0.1111, indicating that a one percentage point increase in obesity is associated with a 0.1111 unit increase in the percentage of diabetic individuals.For "% INACTIVE" and "% OBESE", the p-values are 0.000 and 0.002, these coefficients are statistically significant .

OLS Regression Results

Dep. Variable:	% OBESE	R-squared:	0.245
Model:	OLS	Adj. R-squared:	0.241
Method:	Least Squares	F-statistic:	57.04
Date:	Sun, 03 Dec 2023	Prob (F-statistic):	3.54e-22
Time:	20:59:26	Log-Likelihood:	-462.27
No. Observations:	354	AIC:	930.5
Df Residuals:	351	BIC:	942.2
Df Model:	2		
Covariance Type:	nonrobust		

OLS Regression Results of Obese

The coefficient of determination is a measure of the proportion of variance in the dependent variable that is predictable from the independent variables. We got 0.245, and 24.5% of the variance in "% OBESE" can be explained by the model. We have Obtained Adj.R-squared of 0.241 and the F-statistic value of 57.04, So we can conclude the model is statistically significant.

	coef	std err	t	P>|t|	[0.025	0.975]
const	12.7940	0.524	24.433	0.000	11.764	13.824
% INACTIVE	0.2471	0.038	6.586	0.000	0.173	0.321
% DIABETIC	0.2539	0.080	3.192	0.002	0.097	0.410

Omnibus:	144.747	Durbin-Watson:	1.956
Prob(Omnibus):	0.000	Jarque-Bera (JB):	837.488
Skew:	-1.618	Prob(JB):	1.39e-182
Kurtosis:	9.805	Cond. No.	182.

The results of an OLS regression analysis focusing on the relationship between obesity and two predictors:inactivity and diabetes

Coefficients indicate the estimated effect of each independent variable on the dependent variable.The constant coefficient is 12.7940, the values for all independent variables are zero, and the predicted value of the dependent variable is 12.7940. The coefficient for the "% INACTIVE" variable is 0.2471, which means a one-unit increase in inactivity, there is an estimated increase of 0.2471 units in the percentage of obesity. The coefficient for "% DIABETIC" is 0.2539, indicating an increase in the percentage of diabetics, the percentage of obesity is predicted to increase by 0.2539 units.

OLS Regression Results

Dep. Variable:	% INACTIVE	R-squared:	0.396
Model:	OLS	Adj. R-squared:	0.393
Method:	Least Squares	F-statistic:	115.2
Date:	Sun, 03 Dec 2023	Prob (F-statistic):	3.51e-39
Time:	20:59:26	Log-Likelihood:	-566.39
No. Observations:	354	AIC:	1139.
Df Residuals:	351	BIC:	1150.
Df Model:	2		
Covariance Type:	nonrobust		

OLS Regression Results of Inactive

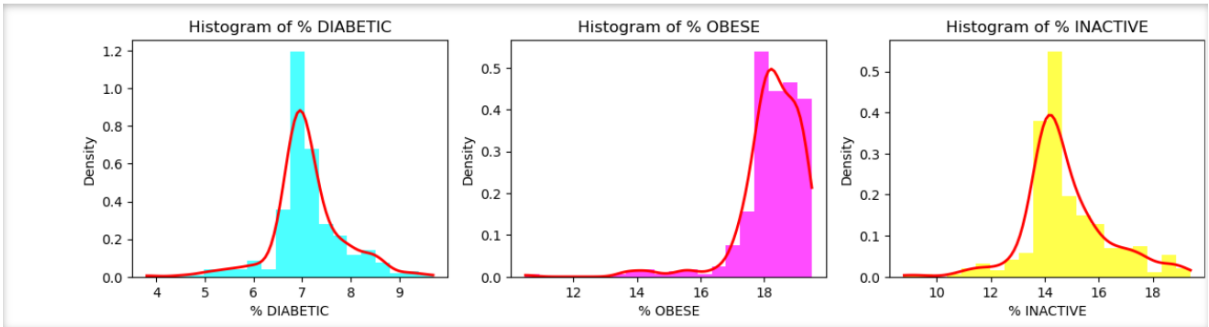
R-squared is the proportion of variance in the dependent variable that can be explained by the independent variables in the model. An R-squared value of 0.396 and 39.6% of the variability in the percentage of inactivity can be explained by the model.Adj. R-squared provides a more accurate measure when comparing models with a different number of independent variables, With a value of 0.393, it's close to the R-squared.F-statistic tests the null hypothesis that all regression coefficients are equal to zero, which means no relationship exists between the predictors and the dependent variable. The value of 115.2 is considerably high, and the model is statistically significant.

	coef	std err	t	P>|t|	[0.025	0.975]
const	-0.1578	1.155	-0.137	0.891	-2.429	2.113
% OBESE	0.4450	0.068	6.586	0.000	0.312	0.578
% DIABETIC	0.9572	0.096	10.023	0.000	0.769	1.145

Omnibus:	63.469	Durbin-Watson:	1.930
Prob(Omnibus):	0.000	Jarque-Bera (JB):	122.388
Skew:	0.969	Prob(JB):	2.65e-27
Kurtosis:	5.130	Cond. No.	355.

Output of an OLS Regression Model

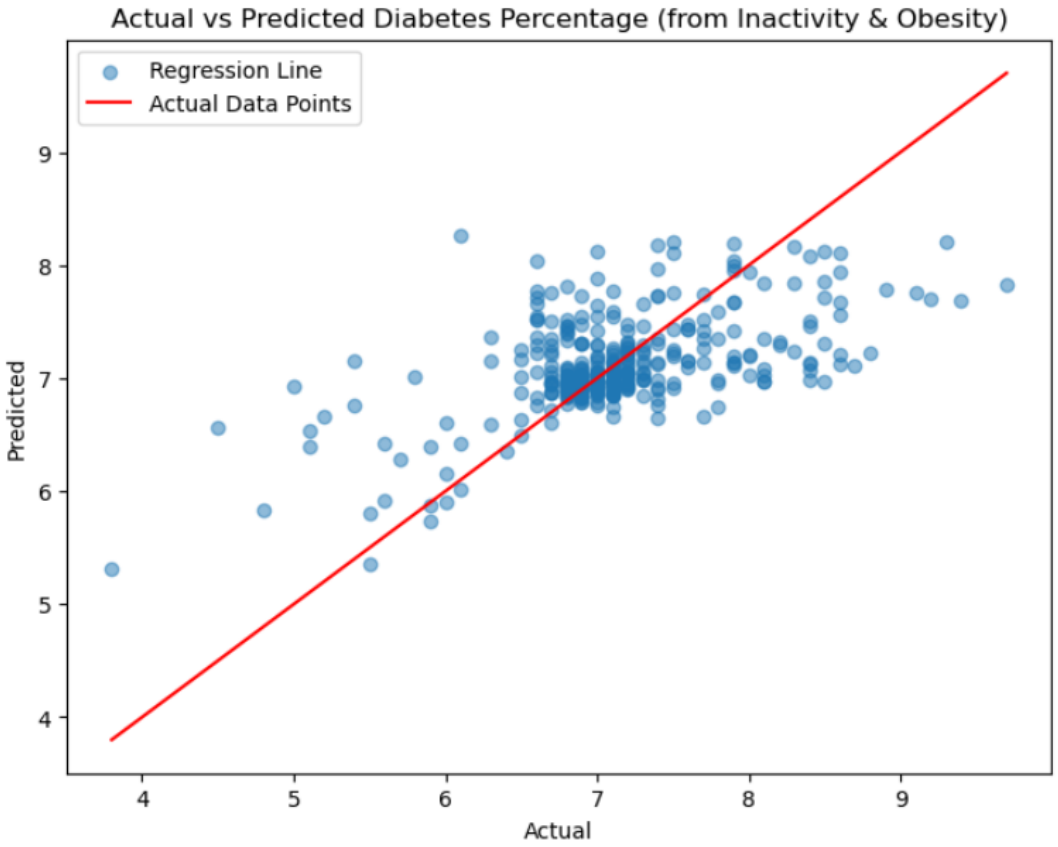
The constant is -0.1578, the value of the dependent variable when all independent variables are zero. The coefficient for the variable "% OBESE" is 0.4450, implying that each one-unit increase in the percentage of obese individuals is associated with a 0.4450 unit increase in the dependent variable. The coefficient for "% DIABETIC" is 0.9572, the percentage of diabetics is associated with a 0.9572 unit increase in the dependent variable.The 95% confidence intervals for the coefficients represent the range of values within which we may be 95% confident that the true value of the coefficient is located.



Shows three histograms, each representing the distribution of percentages for three different health-related variables: diabetic, obese, and inactive populations.

Distribution of percentages for three different health-related variables: diabetic, obese, and inactive populations. Each histogram is overlaid with a kernel density estimate, a

smooth curve that estimates the probability density function of the variable. Histogram of % DIABETIC: The histogram depicts the distribution of the percentage of diabetic individuals within a population or dataset. The distribution appears to be roughly normally distributed, centered around 6-7%. The kernel density plot, represented by the curve, confirms the approximate normality of the data distribution. Histogram of % OBESE: This histogram represents the distribution of the percentage of obese individuals. The distribution seems to be skewed to the right, indicating that there are more data points on the right side of the mode of the distribution. The peak is around 14-15%, but there is a long tail extending towards higher percentages. Histogram of % INACTIVE: The third histogram shows the distribution of the percentage of inactive individuals. This distribution also appears to be positively skewed, with the bulk of the data clustered around 12-14% but with a tail that extends toward higher values.



Shows a scatter plot that compares actual percentages of diabetes with those predicted by a regression model based on inactivity and obesity.

Each Blue dot represents an observed data point in your dataset, where the x-coordinate is the actual percentage of diabetes and the y-coordinate is the percentage of diabetes

predicted by the model.Red line represents the relationship that the regression model has found.The actual(X-axis) observed values of the diabetes percentage in your dataset. The values predicted(Y-axis) by our regression model are based on the independent variables, which in this case are inactivity and obesity rates.

Displaying the relationship between the actual and predicted percentages of inactivity in a population, with obesity and diabetes as predictor variables.

The Blue dots(Actual Data Points) represent individual observations in your dataset. The position on the x-axis indicates the actual observed percentage of inactivity, while the position on the y-axis represents the percentage of inactivity predicted by your regression model based on obesity and diabetes rates. The Red line(Regression Line) shows the predicted value of inactivity given the actual values. The actual(X-axis) observed percentage of inactivity from your data. The Predicted(Y-axis) shows the predicted percentage of inactivity based on our regression model.

Appendix B: Data And Code

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
import numpy as np
from scipy.stats import gaussian_kde
from sklearn.linear_model import LinearRegression
import statsmodels.api as sm
from sklearn.model_selection import cross_val_score
from statsmodels.stats.outliers_influence import variance_inflation_factor
from scipy.stats import skew, kurtosis

# Load the dataset
data = pd.read_excel("cdc.xlsx")
data.head()
```

Out[1]:

	YEAR	FIPS	COUNTY	STATEW	% DIABETIC
0	2018	1001	Autauga County	Alabama	9.5
1	2018	1003	Baldwin County	Alabama	8.4
2	2018	1005	Barbour County	Alabama	13.5
3	2018	1007	Bibb County	Alabama	10.2
4	2018	1009	Blount County	Alabama	10.5

In [13]:
```
input_text = """
1011
2068
2105
2195
2230
5097
6003
6027
6041
6043
6051
6057
6061
6063
6075
6087
6091
8013
8014
8015
8019
8023
8027
8033
8035
51720
51735
51820
51830
51840
53055
56039
"""

numbers = input_text.strip().split('\n')
total_numbers = len(numbers)

# Print the total number of numbers
print("Total number of numbers:", total_numbers)
fips_numbers = list(map(int, numbers))
```

```
Total number of numbers: 354
```

In [3]:
```
input_data = pd.ExcelFile("cdc.xlsx")
```

In [4]:
```
input_data.sheet_names
```

Out[4]:
```
['Diabetes', 'Obesity', 'Inactivity']
```

The Excel file contains three sheets: "Diabetes", "Obesity", and "Inactivity".

In [5]:

```
# Load data from each sheet
df_diabetic = input_data.parse('Diabetes')
df_obese = input_data.parse('Obesity')
df_inactive = input_data.parse('Inactivity')

print('Diabetes\n',df_diabetic.head())
print('Obesity\n',df_obese.head())
print('Inactivity\n',df_inactive.head())
```

```
Diabetes
    YEAR  FIPS          COUNTY   STATEW  % DIABETIC
0  2018  1001  Autauga County  Alabama         9.5
1  2018  1003  Baldwin County  Alabama         8.4
2  2018  1005  Barbour County  Alabama        13.5
3  2018  1007     Bibb County  Alabama        10.2
4  2018  1009   Blount County  Alabama        10.5
Obesity
    YEAR  FIPS                    COUNTY    STATE  % OBESE
0  2018  1011             Bullock County  Alabama     18.7
1  2018  2068             Denali Borough   Alaska     18.9
2  2018  2105  Hoonah-Angoon Census Area   Alaska     19.4
3  2018  2195     Petersburg Census Area   Alaska     17.2
4  2018  2230       Skagway Municipality   Alaska     18.3
Inactivity
    YEAR  FIPS                 COUNTY    STATE  % INACTIVE
0  2018  1011          Bullock County  Alabama        17.0
1  2018  1029         Cleburne County  Alabama        19.3
2  2018  1037            Coosa County  Alabama        16.8
3  2018  1063           Greene County  Alabama        16.8
4  2018  2013  Aleutians East Borough   Alaska        19.2
```

In [6]:

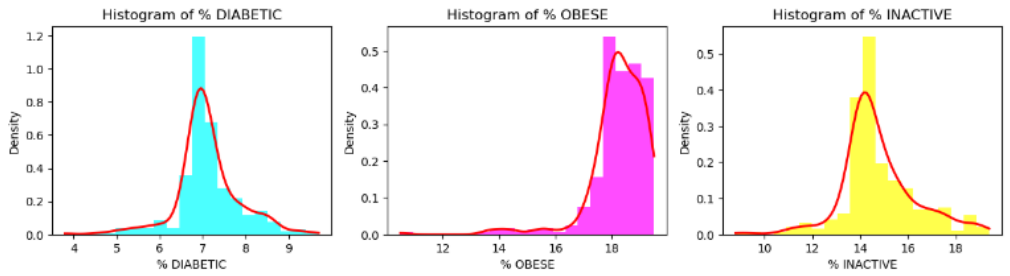
```
# Creating a function to plot smooth histograms
def plot_histogram(data, title, xlabel, color):
    plt.hist(data, bins=20, density=True, alpha=0.7, color=color)
    kde = gaussian_kde(data)
    x_vals = np.linspace(data.min(), data.max(), 100)
    smooth_curve = kde(x_vals)
    plt.plot(x_vals, smooth_curve, color='red', linewidth=2)
    plt.title(title)
    plt.xlabel(xlabel)
    plt.ylabel('Density')

# Loading the datasets and merging them
df_diabetic_filtered = df_diabetic[df_diabetic['FIPS'].isin(fips_numbers)]
df_obese_filtered = df_obese[df_obese['FIPS'].isin(fips_numbers)]
df_inactive_filtered = df_inactive[df_inactive['FIPS'].isin(fips_numbers)]
df = df_diabetic_filtered.merge(df_obese_filtered[['FIPS', '% OBESE']], on='FIPS').merge(df_inactive_filtered[['FIPS','% INAC

plt.figure(figsize=(12, 6))

plt.subplot(2, 3, 1)
plot_histogram(df['% DIABETIC'], 'Histogram of % DIABETIC', '% DIABETIC', 'cyan')
plt.subplot(2, 3, 2)
plot_histogram(df['% OBESE'], 'Histogram of % OBESE', '% OBESE', 'magenta')
plt.subplot(2, 3, 3)
plot_histogram(df['% INACTIVE'], 'Histogram of % INACTIVE', '% INACTIVE', 'yellow')


plt.tight_layout()
plt.show()
```



In [7]:
```
df
```

Out[7]:

	YEAR	FIPS	COUNTY	STATEW	% DIABETIC	% OBESE	% INACTIVE
0	2018	1011	Bullock County	Alabama	9.4	18.7	17.0
1	2018	2068	Denali Borough	Alaska	6.8	18.9	16.2
2	2018	2105	Hoonah-Angoon Census Area	Alaska	7.3	19.4	15.0
3	2018	2195	Petersburg Census Area	Alaska	9.2	17.2	17.8
4	2018	2230	Skagway Municipality	Alaska	6.6	18.3	15.8
...	...	...	...	...	...	...	...
349	2018	51820	Waynesboro City	Virginia	8.6	19.5	16.6
350	2018	51830	Williamsburg City	Virginia	8.5	18.0	15.7
351	2018	51840	Winchester City	Virginia	6.9	19.4	16.1
352	2018	53055	San Juan County	Washington	4.5	19.3	11.9
353	2018	56039	Teton County	Wyoming	3.8	10.5	10.7

354 rows × 7 columns

In [8]:
```
# Calculate skewness and kurtosis for % DIABETIC
skewness_diabetic = skew(df['% DIABETIC'])
kurtosis_diabetic = kurtosis(df['% DIABETIC'])

# Calculate skewness and kurtosis for % OBESE
skewness_obese = skew(df['% OBESE'])
kurtosis_obese = kurtosis(df['% OBESE'])

# Calculate skewness and kurtosis for % INACTIVE
skewness_inactive = skew(df['% INACTIVE'])
kurtosis_inactive = kurtosis(df['% INACTIVE'])

# Print the results
print(f'Skewness of % DIABETIC: {skewness_diabetic:.2f}')
print(f'Kurtosis of % DIABETIC: {kurtosis_diabetic:.2f}')

print(f'Skewness of % OBESE: {skewness_obese:.2f}')
print(f'Kurtosis of % OBESE: {kurtosis_obese:.2f}')

print(f'Skewness of % INACTIVE: {skewness_inactive:.2f}')
print(f'Kurtosis of % INACTIVE: {kurtosis_inactive:.2f}')
```

```
Skewness of % DIABETIC: -0.05
Kurtosis of % DIABETIC: 2.79
Skewness of % OBESE: -2.75
Kurtosis of % OBESE: 12.93
Skewness of % INACTIVE: 0.43
Kurtosis of % INACTIVE: 1.61
```

Inactivity vs. Obesity predicting Diabetes
Independent Variables: % INACTIVE and % OBESE
Dependent Variable: % DIABETIC

In [9]:
```
# Define IVs and DV
X1 = df[['% INACTIVE', '% OBESE']]
y1 = df['% DIABETIC']

# Use statsmodels for the regression to get detailed statistics
X1_sm = sm.add_constant(X1)  # Adding a constant for the intercept
model1 = sm.OLS(y1, X1_sm).fit()
print(model1.summary())

# For cross-validation (Optional)
cross_val_scores1 = cross_val_score(LinearRegression(), X1, y1, cv=5)
print("\nCross-validation scores:", cross_val_scores1)
print("Average R^2:", np.mean(cross_val_scores1))

# Check for multicollinearity using VIF
vif_data1 = pd.DataFrame()
vif_data1["Variable"] = X1.columns
vif_data1["VIF"] = [variance_inflation_factor(X1.values, i) for i in range(X1.shape[1])]
print("\nVIF values:")
print(vif_data1)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:            % DIABETIC   R-squared:                       0.341
Model:                           OLS   Adj. R-squared:                  0.337
Method:                Least Squares   F-statistic:                     90.71
Date:               Sun, 03 Dec 2023   Prob (F-statistic):           1.76e-32
Time:                       20:59:26   Log-Likelihood:                -315.89
No. Observations:                354   AIC:                             637.8
Df Residuals:                    351   BIC:                             649.4
Df Model:                          2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          1.6536      0.562      2.941      0.003       0.548       2.759
% INACTIVE     0.2325      0.023     10.023      0.000       0.187       0.278
% OBESE        0.1111      0.035      3.192      0.002       0.043       0.180
==============================================================================
Omnibus:                       17.281   Durbin-Watson:                   1.673
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               45.622
Skew:                          -0.042   Prob(JB):                     1.24e-10
Kurtosis:                       4.757   Cond. No.                         421.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Cross-validation scores: [ 0.46174692  0.02028448 -0.05981153 -0.05914246  0.41097258]
Average R^2: 0.1548099964274586

VIF values:
     Variable         VIF
0  % INACTIVE  118.789883
1     % OBESE  118.789883
```

Inactivity vs. Diabetes predicting Obesity
Independent Variables: % INACTIVE and % DIABETIC
Dependent Variable: % OBESE

In [10]:

```
# Define IVs and DV
X2 = df[['% INACTIVE', '% DIABETIC']]
y2 = df['% OBESE']

# Use statsmodels for the regression
X2_sm = sm.add_constant(X2)
model2 = sm.OLS(y2, X2_sm).fit()
print(model2.summary())

# For cross-validation (Optional)
cross_val_scores2 = cross_val_score(LinearRegression(), X2, y2, cv=5)
print("\nCross-validation scores:", cross_val_scores2)
print("Average R^2:", np.mean(cross_val_scores2))

# Check for multicollinearity using VIF
vif_data2 = pd.DataFrame()
vif_data2["Variable"] = X2.columns
vif_data2["VIF"] = [variance_inflation_factor(X2.values, i) for i in range(X2.shape[1])]
print("\nVIF values:")
print(vif_data2)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:               % OBESE   R-squared:                       0.245
Model:                           OLS   Adj. R-squared:                  0.241
Method:                Least Squares   F-statistic:                     57.04
Date:               Sun, 03 Dec 2023   Prob (F-statistic):           3.54e-22
Time:                       20:59:26   Log-Likelihood:                -462.27
No. Observations:                354   AIC:                             930.5
Df Residuals:                    351   BIC:                             942.2
Df Model:                          2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         12.7940      0.524     24.433      0.000      11.764      13.824
% INACTIVE     0.2471      0.038      6.586      0.000       0.173       0.321
% DIABETIC     0.2539      0.080      3.192      0.002       0.097       0.410
==============================================================================
Omnibus:                      144.747   Durbin-Watson:                   1.956
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              837.488
Skew:                          -1.618   Prob(JB):                    1.39e-182
Kurtosis:                       9.805   Cond. No.                         182.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Cross-validation scores: [ 0.2519533  -0.1294807   0.07745399  0.01265881  0.25160269]
Average R^2: 0.09283761590837947

VIF values:
     Variable         VIF
0  % INACTIVE  109.596899
1  % DIABETIC  109.596899
```

Obese vs. Diabetes predicting Inactivity
Independent Variables: % OBESE and % DIABETIC
Dependent Variable: % INACTIVE

In [11]:

```
# Define IVs and DV
X3 = df[['% OBESE', '% DIABETIC']]
y3 = df['% INACTIVE']

# Use statsmodels for the regression
X3_sm = sm.add_constant(X3)
model3 = sm.OLS(y3, X3_sm).fit()
print(model3.summary())

# For cross-validation (Optional)
cross_val_scores3 = cross_val_score(LinearRegression(), X3, y3, cv=5)
print("\nCross-validation scores:", cross_val_scores3)
print("Average R^2:", np.mean(cross_val_scores3))

# Check for multicollinearity using VIF
vif_data3 = pd.DataFrame()
vif_data3["Variable"] = X3.columns
vif_data3["VIF"] = [variance_inflation_factor(X3.values, i) for i in range(X3.shape[1])]
print("\nVIF values:")
print(vif_data3)
```

```
                            OLS Regression Results
==============================================================================
Dep. Variable:             % INACTIVE   R-squared:                       0.396
Model:                            OLS   Adj. R-squared:                  0.393
Method:                 Least Squares   F-statistic:                     115.2
Date:                Sun, 03 Dec 2023   Prob (F-statistic):           3.51e-39
Time:                        20:59:26   Log-Likelihood:                -566.39
No. Observations:                 354   AIC:                             1139.
Df Residuals:                     351   BIC:                             1150.
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         -0.1578      1.155     -0.137      0.891      -2.429       2.113
% OBESE        0.4450      0.068      6.586      0.000       0.312       0.578
% DIABETIC     0.9572      0.096     10.023      0.000       0.769       1.145
==============================================================================
Omnibus:                       63.469   Durbin-Watson:                   1.930
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              122.388
Skew:                           0.969   Prob(JB):                     2.65e-27
Kurtosis:                       5.130   Cond. No.                         355.
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Cross-validation scores: [ 0.56099915  0.1089885   0.20130815 -0.49398354  0.24574599]
Average R^2: 0.124611649481013

VIF values:
     Variable         VIF
0     % OBESE  110.661027
1  % DIABETIC  110.661027
```

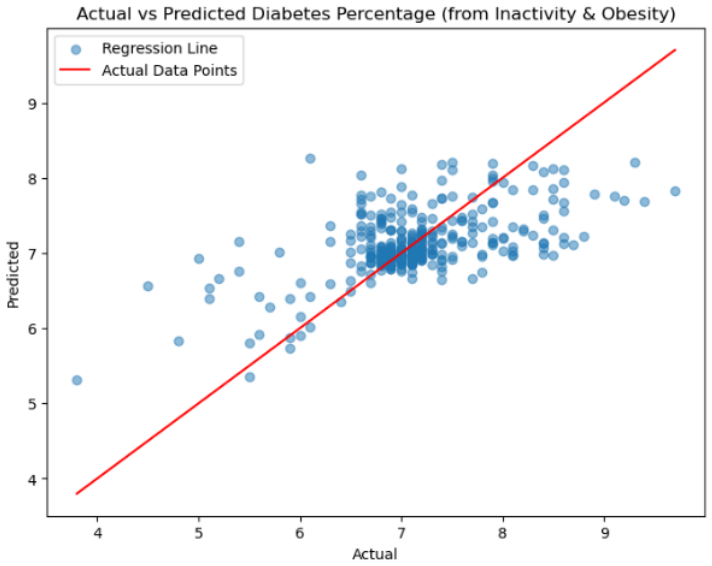
In [12]:

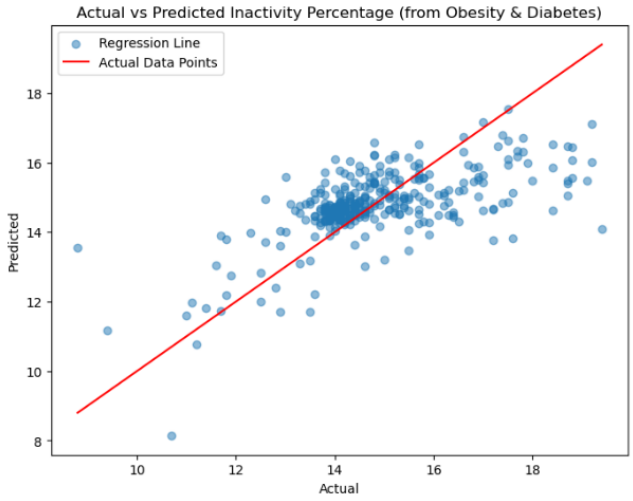
```
# Function to plot actual vs predicted values
def plot_actual_vs_predicted(y_actual, y_predicted, title):
    plt.figure(figsize=(8, 6))
    plt.scatter(y_actual, y_predicted, alpha=0.5)
    plt.plot([min(y_actual), max(y_actual)], [min(y_actual), max(y_actual)], color='red')  # regression line
    plt.xlabel('Actual')
    plt.ylabel('Predicted')
    plt.title(title)
    plt.legend(['Regression Line', 'Actual Data Points'])
    plt.show()

# 1. Inactivity vs. Obesity predicting Diabetes
y1_pred = model1.predict(X1_sm)
plot_actual_vs_predicted(y1, y1_pred, 'Actual vs Predicted Diabetes Percentage (from Inactivity & Obesity)')

# 2. Inactivity vs. Diabetes predicting Obesity
y2_pred = model2.predict(X2_sm)
plot_actual_vs_predicted(y2, y2_pred, 'Actual vs Predicted Obesity Percentage (from Inactivity & Diabetes)')

# 3. Obesity vs. Diabetes predicting Inactivity
y3_pred = model3.predict(X3_sm)
plot_actual_vs_predicted(y3, y3_pred, 'Actual vs Predicted Inactivity Percentage (from Obesity & Diabetes)')
```





Type *Markdown* and LaTeX: α^2