

# Final

December 2, 2025

```
[1]: #imported packages
import numpy as np
import matplotlib.pyplot as plt
import pandas as pd
import statsmodels.api as sm
%matplotlib inline

print("packages imported")
```

packages imported

## 1 Problem Statement

Model the total sleep duration of 374 people using linear regression, given the number of minutes engaged in physical activity, self-reported stress level (1-10), and self-reported sleep quality (1-10).

## 2 Variables and Parameters

Symbol	Description	Type	Dimension	Units
$Y$	The total number of hours slept	dependent variable	$T$	hours
$B_0$	The default sleeping duration $d$	parameter	$T$	hours
$B_1$	Regression coefficient for physical activity	parameter	$T$	hours/minutes
$X_1$	The number of minutes (minutes) engaged in physical activity during the day	independent variable	$T$	minutes
$B_2$	Regression coefficient for stress level	parameter	1	hours

Symbol	Description	Type	Dimension	Units
$X_2$	The (self-reported) stress level experienced	independent variable	1	
$B_3$	Regression coefficient for quality of sleep	parameter	$T$	hours
$X_3$	The (self-reported) quality of sleep reported on a scale of 1 to 10	independent variable	1	
$\epsilon$	residual	parameter	$T$	hours

### 3 Assumptions and Constraints:

- The subjective formulation of a score from 1 to 10 for stress and sleep quality is uniform across all participants.
- Participants do not lie about the times or scores.
- Participants in the data reflect an average sample from the general population this model will be applied to.
- The average value of the error is 0.
- No other variables affect sleep duration.
- The variance of the error is constant
- Error is a random sample from normal distribution of zero mean and variance  $\sigma^2$ .
- Errors from each observation are independent of one another

### 4 Building Solutions

We first fit a baseline linear regression model using `statsmodels.api` to understand how well the raw data satisfies the 3 core assumptions of linear regression and to diagnose any other structural issues in the untransformed data. Below is our analysis of that model.

#### 4.1 Summary Table Analysis(Untransformed Data)

Upon building the linear regression model with no feature transformation, we now analyze the summary information. The first piece of information of note is that there is a warning that the condition number is large. We know from class that this indicates there might be linearly dependent features. Additionally the  $R^2$  and  $R^2_{adj}$  are both relatively high and this indicates that most of the variance in the target variable (sleep duration) can be explained by the features of the model which is good. Finally we see that the regression coefficient value for physical activity is near 0, this could indicate that either physical activity is not a good predictor of our target variable or that the physical activity is not on the same scale as the rest of the data. From these facts alone it is not enough to conclude whether or not this is a good linear model, to further assess its validity, we now examine whether the three core assumptions of linear regression are satisfied.

## 4.2 Analysis of Linear Regression Assumptions(Untransformed Data)

### 4.2.1 Analysis of the Regression Plots

As we know from class the regression plots are used to evaluate if the first assumption of linear regression is satisfied or not, namely the average value of the error is 0 ( $\mathbb{E}(\epsilon_i) = 0, \forall i$ ).

**Residual Plots Analysis** Starting with the analysis of the plot, residuals vs physical activity. We noticed the data violates the condition of being evenly distributed over the  $y = 0$  line (with the possible exception of Physical Activity = 90). Next when we looked at residuals vs quality we again noticed some columns of data points not being evenly spread across the  $y = 0$  line, however this violation wasn't as extreme as in the physical activity plot. Finally we looked at residuals vs stress, we determined this plot actually did satisfy the condition of being evenly distributed over the horizontal axis as needed. However, with two out of the three variables not satisfying the needed condition, we conclude that the residual plots do not satisfy the condition of being evenly distributed over the horizontal axis.

**Partial Regression Plots Analysis** Because the variables take on integer values (from the responses to the survey), the scatter plots show vertical columns of points. This makes it harder to judge whether the covariates exhibit a linear trend, but the overall pattern is still relatively clear. The two covariates that best exhibit a linear trend are "Quality" and the constant term, though even these relations are still pretty weak. We've concluded that the two covariates namely "Physical\_Activity" and "Stress" don't follow a linear trend. Thus with two out of four of the plots being flat out not a linear trend and the other two being somewhat inconclusive we've determined that we have violated this condition as well.

**Conclusion of Regression Plot Analysis** Since in both cases we have violated our conditions needed for the average value of the error's is 0. Therefore, applying linear regression to the untransformed data violates the first assumption of linear regression.

### 4.2.2 Analysis of the Residuals vs Fitted Y's Plot

From class we know that if the residuals vs fitted y's produce a plot that's a random scatter about horizontal axis, then we've satisfied the second assumption of linear regression. Namely that the variance of the error is constant ( $Var(\epsilon_i) = \sigma^2, \forall i$ ). From our analysis of the plot we observe a clustering of data below the horizontal axis around fitted values of approximately 6.25 to 7.5, and the overall spread of the residuals is too narrow. This non random structure indicates that the variance is not constant, thus violating the variance of errors is a constant assumption.

### 4.2.3 Analysis of the Normal vs Residuals(QQ) plot

From class we know that if we plot our residuals vs a normal distribution and our points lie along the line  $y = x$ , then we have satisfied the third assumption of linear regression. Namely that the error  $\vec{\epsilon}$  is a random sample from the normal distribution of zero mean and variance  $\sigma^2$ . We observe that the residuals do not lie on the  $y = x$  line but instead they seem to lie on a tilted parabola instead. Hence we conclude that we have violated the third assumption of linear regression as well.

#### 4.2.4 Conclusion of Linear Regression Assumptions(Untransformed Data)

The untransformed data fails to meet the assumptions required for a valid linear regression model. As a result, we proceed to transforming the data in search of model that satisfies the three assumptions of linear regression.

### 4.3 Transforming Data and Model Fitting

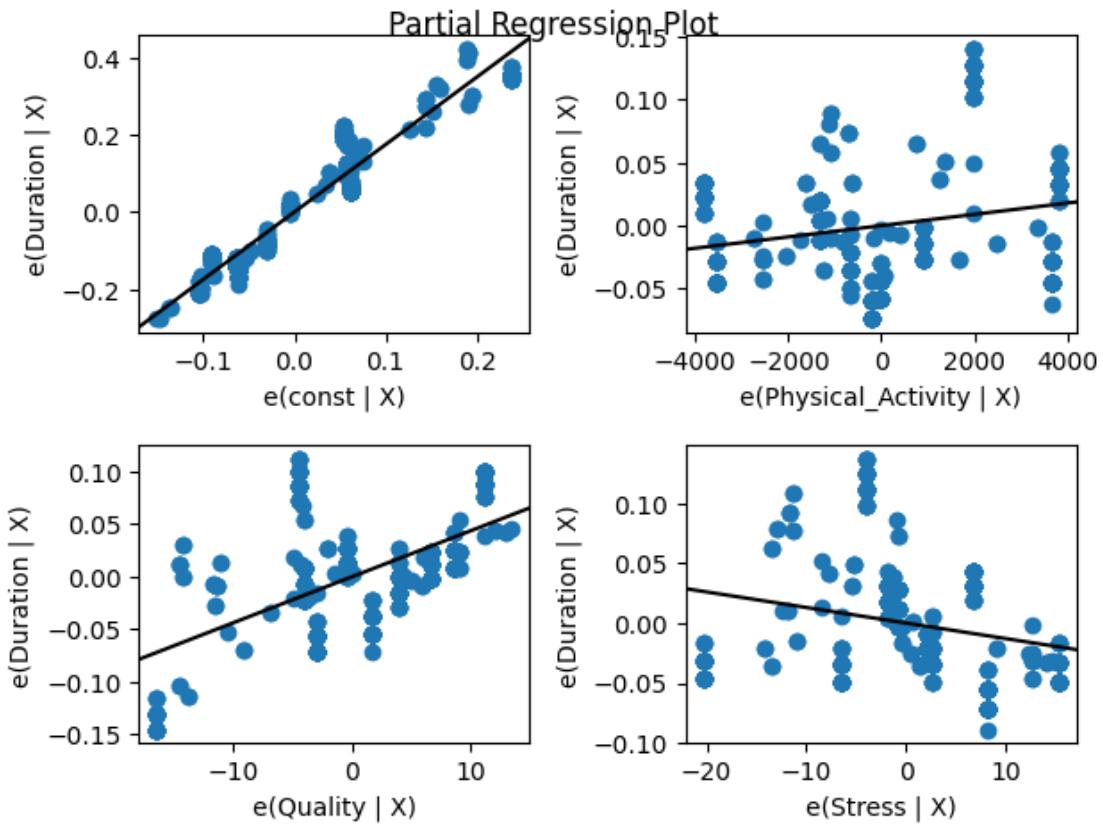
#### 4.3.1 EXPLANATION OF TRANSFORMATIONS

```
[5]: sleep_data = pd.read_csv('Sleep.csv')
sleep = pd.DataFrame(sleep_data)
```

```
[6]: sleepTransformed = sleep.copy()
sleepTransformed['Physical_Activity'] = np.sqrt(sleepTransformed['Physical_Activity'])**2
sleepTransformed['Quality'] = (sleepTransformed['Quality'])**2
sleepTransformed['Stress'] = (sleepTransformed['Stress'])**2
sleepTransformed['Duration'] = np.log(sleepTransformed['Duration'])
X_transformed = sleepTransformed[['Physical_Activity', 'Quality', 'Stress']]
Y_transformed = sleepTransformed['Duration']
X_transformed = sm.add_constant(X_transformed)
reg_transformed = sm.OLS(Y_transformed, X_transformed).fit()
print(reg_transformed.rsquared_adj)
```

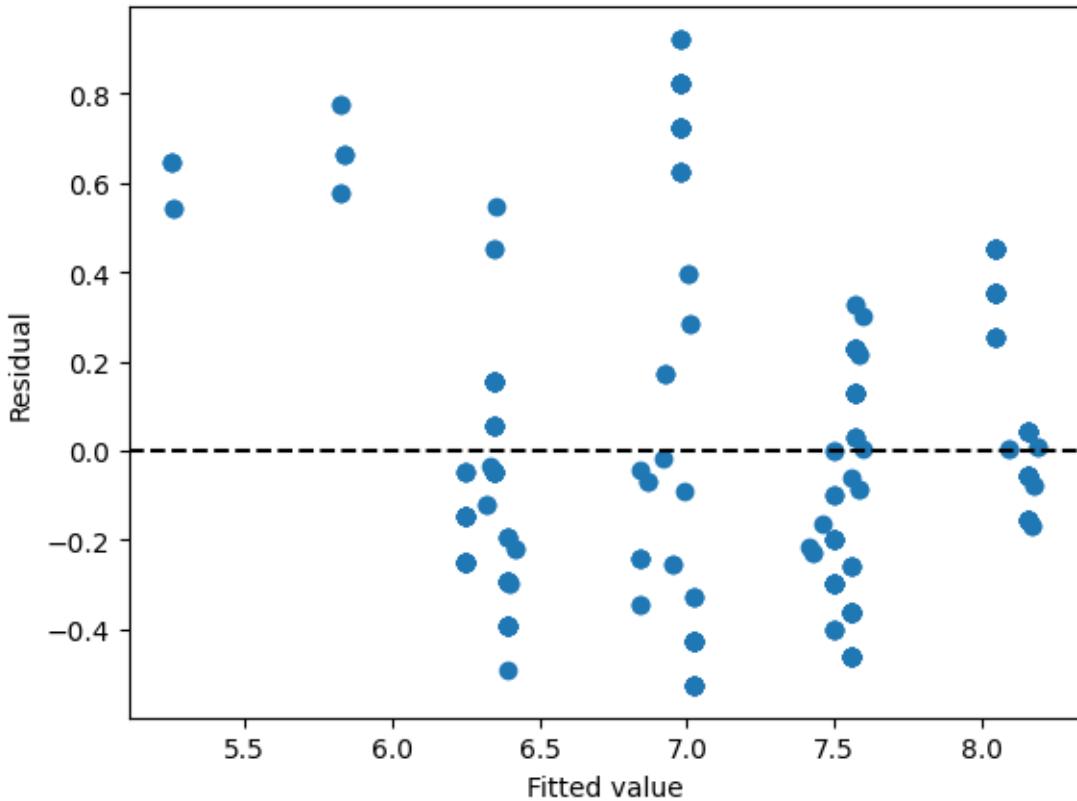
0.8047269776324298

```
[7]: # Plot 1 (Partial Regression Plots)
sm.graphics.plot_partregress_grid(reg_transformed)
plt.show()
```

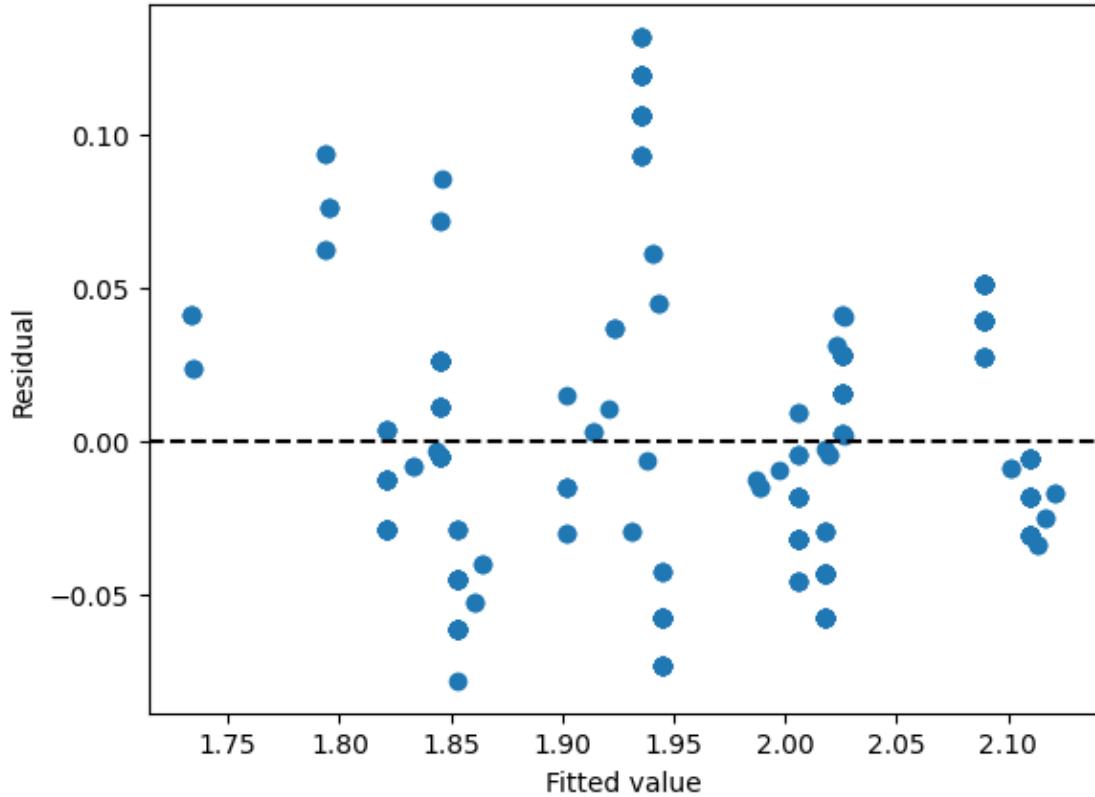


```
[10]: # Plot 2 (Orginal Residual-Fitted Value Plot)
X = sleep[['Physical_Activity', 'Stress', 'Quality']]
Y = sleep['Duration']
X = sm.add_constant(X)
reg_original = sm.OLS(Y, X).fit()
```

```
fitted_y_original = reg_original.fittedvalues
residuals_original = reg_original.resid
plt.scatter(fitted_y_original,residuals_original)
plt.axhline(y=0,color='k',linestyle = '--')
plt.xlabel('Fitted value')
plt.ylabel('Residual')
plt.show()
```



```
[11]: # Plot 3 (Transformed Residual-Fitted Value Plot)
fitted_y_transformed = reg_transformed.fittedvalues
residuals_transformed = reg_transformed.resid
plt.scatter(fitted_y_transformed,residuals_transformed)
plt.axhline(y=0,color='k',linestyle = '--')
plt.xlabel('Fitted value')
plt.ylabel('Residual')
plt.show()
```

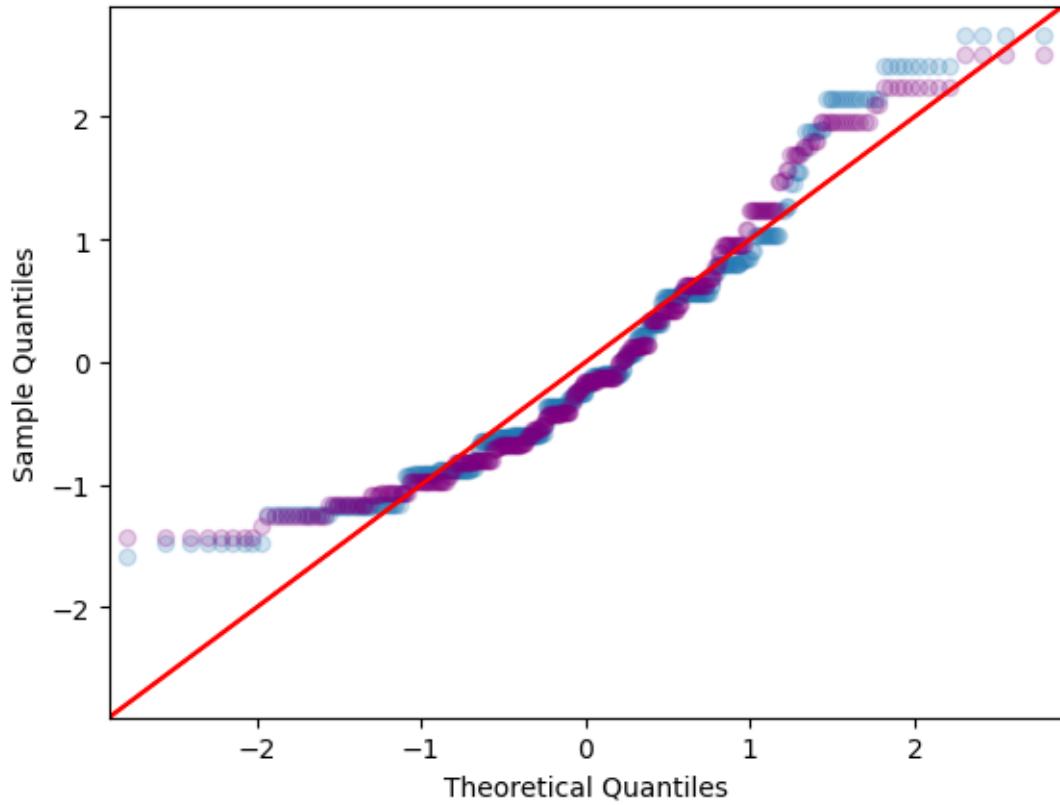


```
[12]: # Plot 4 (Overlapping QQplot)
residuals_original = reg_original.resid

fig, ax = plt.subplots()

sm.qqplot(residuals_transformed, fit = True, line='45', alpha = 0.2, ax=ax)
sm.qqplot(residuals_original, fit = True, line='45', markerfacecolor='purple', markeredgecolor='purple', alpha = 0.2, ax=ax)

plt.show()
```



## 5 Analyze and Assess