Predicting Parkinson's Disease Severity from Voice Acoustics

By: Vanshika Gupta & Ankushi Dutta

STAT 525 Final Project

December 17, 2025

Contents

1 Abstract

1.1 Project Motivation

Parkinson's Disease (PD) is a progressive disorder that significantly affects motor skills and speech. Usually, tracking the progression of PD involves frequent physical exams where doctors assign a Unified Parkinson's Disease Rating Scale (UPDRS) score. However, these visits are not cost effective and difficult for patients who already struggle with mobility. Due to the fact that doctor visits are infrequent, they only provide a few data points about the patient's history rather than a continuous picture.

Telemonitoring offers a alternative by using remote voice recordings to estimate disease severity from home. As vocal impairment is often an early symptom of PD, telemonitoring has emerged as a potential tool for tracking disease progression. In this project, we explore the statistical relationship between biomedical voice measurements and Parkinson's severity.

1.2 Objective

The objective of this study is to develop a statistical model to predict the `total_UPDRS` score based on a suite of 16 biomedical voice measures and demographic factors.

Specifically, we aim to:

- Evaluate limitations of Standard Linear Regression: Demonstrate why aggregating repeated measures to satisfy independence assumptions results in a loss of statistical power.
- Implement a Mixed Model Approach: Utilize a linear mixed model (LMM) to account for the repeated measures structure of the data, modeling `subject_id` as a random effect to control for baseline heterogeneity.
- Identify key acoustic biomarkers: Determine which specific voice features remain statistically significant predictors of disease severity after adjusting for age and subject-specific random effects.

1.3 Dataset Description

This project utilizes the Oxford Parkinson's Disease Telemonitoring Data Set, obtained from the UCI Machine Learning Repository. The data was originally collected by Tsanas et al. (2010) to investigate the feasibility of remote, non-invasive monitoring of Parkinson's progression [1].

The dataset contains 5,875 voice recordings gathered from 42 individuals with early-stage Parkinson's disease during a six-month clinical trial. Patients used a remote telemonitoring device to record phonations (sustained vowel sounds) at home, providing a high-frequency longitudinal view of disease progression.

- **Response Variable:** `total_UPDRS` (Unified Parkinson's Disease Rating Scale), a standard clinical score measuring disease severity (range: 7–55).
- **Subject Identifier:** `subject#` (integer ID for each patient).
- **Demographics:** Age and Sex.
- **Acoustic Features (16 total):**
  - **Frequency Perturbation:** `Jitter (%, Abs, RAP, PPQ5, DDP)`
  - **Amplitude Perturbation:** `Shimmer (dB, APQ3, APQ5, APQ11, DDA)`
  - **Harmonic-to-Noise Ratios:** `NHR, HNR`
  - **Nonlinear Measures:** `RPDE` (Recurrence Period Density Entropy), `DFA` (Detrended Fluctuation Analysis), and `PPE` (Pitch Period Entropy).

2 Data Preproccessing Exploratory Analysis

The raw dataset is composed of 5,875 observations across 22 variables, representing longitudinal voice recordings from 42 subjects. We first confirmed that the dataset contained no missing values. The response variable, `total_UPDRS`, ranges from 7 to 55 and follows a slightly right-skewed distribution, which approximates normality sufficiently for linear modeling (Figure 1).

Following the univariate distribution check, we explored bivariate relationships between key demographic factors and disease severity to identify potential predictors. We visualized the relationship between biological age and the `total_UPDRS` score using a scatter plot (Figure 2). This plot indicates a general positive association, suggesting that older patients tend to present with higher disease severity. Additionally, we examined potential differences in severity based on biological sex (Figure 3). The box plots depict comparable median UPDRS scores between male and female patients, although the male group shows a wider overall range of symptom severity. The overlap between the distributions suggests that sex does not appear to meaningfully impact the total UPDRS score.

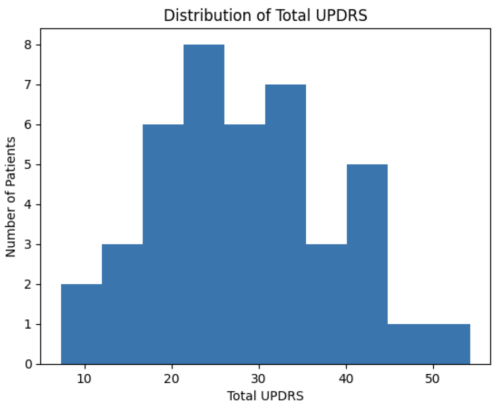


Figure 1: Histogram showing the distribution of Parkinson's disease severity scores (total_UPDRS) across all recordings. The distribution is slightly right-skewed but sufficiently close to normal to justify linear modeling.

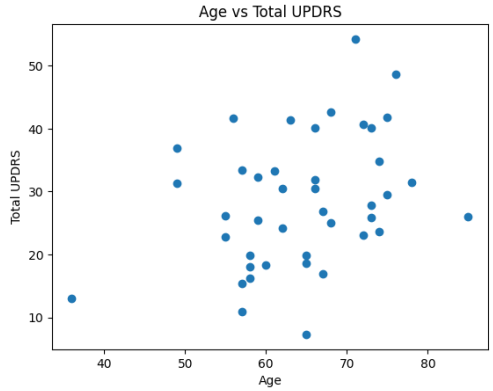


Figure 2: Scatter plot illustrating the relationship between patient age and Parkinson's disease severity (total_UPDRS). An overall increasing trend is observed, though substantial variability remains.

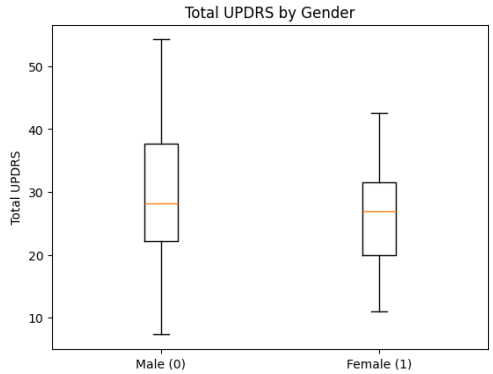


Figure 3: Boxplot comparing the distribution of Parkinson's disease severity scores (total_UPDRS) across gender groups. While median severity differs slightly between groups, considerable overlap indicates that gender alone does not strongly explain variability in disease severity.

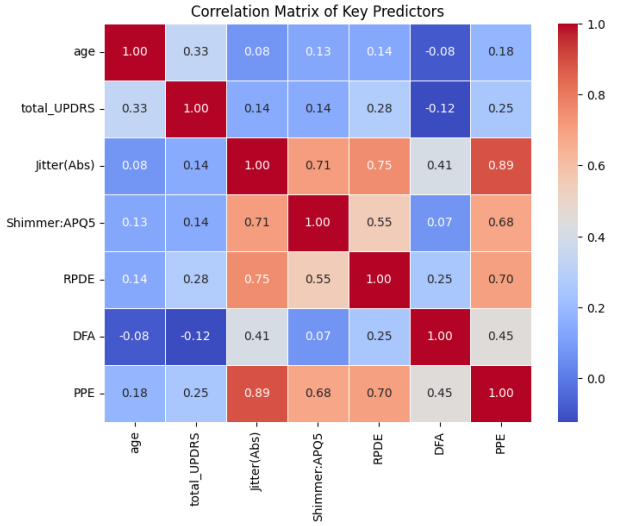


Figure 4: Correlation matrix of key predictors and total UPDRS. Strong correlations are observed among several voice-based measures (e.g., Jitter, Shimmer, PPE, RPDE), indicating substantial multicollinearity and motivating subsequent variance inflation factor (VIF) analysis and predictor reduction.

2.1 Multicollinearity Check:

To ensure stable coefficient estimates, we assessed multicollinearity among the 16 acoustic feature predictors. Initial inspection of the correlation matrix revealed severe redundancy, particularly among the measures of frequency and amplitude perturbation. Specifically, the five variations of Jitter (e.g., `Jitter:RAP`, `Jitter:PPQ5`) were highly collinear, with variance inflation factors (VIF) exceeding extreme thresholds (VIF > 100).

To address this, we applied a feature selection strategy to retain only the most distinct representative from each acoustic category: `Jitter:Abs` (frequency), `Shimmer:APQ5` (amplitude), and selected nonlinear measures. We dropped the redundant measures and removed `motor_UPDRS` to prevent data leakage, as it is a subset of the target variable. The post-cleaning correlation matrix heatmap (Figure 4) showed reduced collinearity, confirming the suitability of the remaining predictors for multivariate modeling. The moderate correlation ($r = 0.33$) between `age` and `total_UPDRS` supports that age is a key biological driver of severity, while the varying correlations among acoustic features suggest they capture different aspects of vocal impairment.

2.2 Preliminary Analysis: Aggregated Data

Standard linear regression assumes that residuals are independent and identically distributed. However, our dataset contains repeated measures ($N \approx 140$ recordings per subject), which violates this assumption. To address this initially, we followed standard aggregation by calculating the mean of all recordings for each `subject_id`. This process condensed the dataset from $N = 5,875$ observations to $N = 42$ independent data points, representing the average disease state for each patient.

Following that, we fitted a multiple linear regression (MLR) model to this aggregated dataset using `total_UPDRS` as the response and the selected acoustic features as predictors. While diagnostic tests indicated that residuals satisfied the normality assumption (Shapiro-Wilk $p > 0.05$), the model suffered from a critical loss of statistical power due to the drastic reduction in sample size.

The results of this preliminary model (Figure 5) revealed that only biological age was a statistically significant predictor ($p = 0.038$). None of the biomedical voice features, including Jitter, Shimmer, or RPDE, reached statistical significance in this regression model. This limitation provided the primary motivation for adopting a linear mixed model (LMM), which utilizes the full dataset to preserve statistical power.

OLS Regression Results

Dep. Variable:	total_UPDRS	R-squared:	0.473
Model:	OLS	Adj. R-squared:	0.019
Method:	Least Squares	F-statistic:	1.041
Date:	Sat, 13 Dec 2025	Prob (F-statistic):	0.460
Time:	07:24:29	Log-Likelihood:	−144.15
No. Observations:	42	AIC:	328.3
Df Residuals:	22	BIC:	363.1
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	38.3275	83.913	0.457	0.652	−135.697	212.352
age	0.2199	0.230	0.956	0.349	−0.257	0.697
sex	−9.3919	5.759	−1.631	0.117	−21.334	2.551
test_time	−0.3453	0.268	−1.287	0.211	−0.902	0.211
Jitter(%)	1944.2415	1.09e+04	0.178	0.860	−2.07e+04	2.45e+04
Jitter(Abs)	−6.34e+05	3.32e+05	−1.907	0.070	−1.32e+06	5.54e+04
Jitter:RAP	−4.438e+06	7.51e+06	−0.591	0.560	−2e+07	1.11e+07
Jitter:PPQ5	−1.421e+04	1.77e+04	−0.802	0.431	−5.09e+04	2.25e+04
Jitter:DDP	1.486e+06	2.5e+06	0.594	0.559	−3.7e+06	6.67e+06
Shimmer	−3505.1812	6153.933	−0.570	0.575	−1.63e+04	9257.295
Shimmer(dB)	168.8679	297.036	0.569	0.575	−447.147	784.882
Shimmer:APQ3	−3.286e+06	9.44e+06	−0.348	0.731	−2.29e+07	1.63e+07
Shimmer:APQ5	4527.7928	4764.078	0.950	0.352	−5352.299	1.44e+04
Shimmer:APQ11	970.3133	2958.790	0.328	0.746	−5165.842	7106.469
Shimmer:DDA	1.094e+06	3.15e+06	0.348	0.731	−5.43e+06	7.62e+06
NHR	−2.9643	368.919	−0.008	0.994	−768.056	762.128
HNR	−0.0559	2.098	−0.027	0.979	−4.407	4.295
RPDE	66.9012	55.162	1.213	0.238	−47.497	181.300
DFA	−45.1092	55.358	−0.815	0.424	−159.914	69.695
PPE	45.8379	112.735	0.407	0.688	−187.960	279.636

Omnibus:	2.034	Durbin-Watson:	2.240
Prob(Omnibus):	0.362	Jarque-Bera (JB):	1.115
Skew:	−0.084	Prob(JB):	0.573
Kurtosis:	3.780	Cond. No.	7.23e+08

Figure 5: Ordinary Least Squares (OLS) regression results using the full set of predictors prior to multicollinearity correction. Large standard errors and unstable coefficient estimates highlight the impact of severe multicollinearity among acoustic features.

3 Methodology

3.1 Justification for Linear Mixed Model

As demonstrated in the preliminary analysis, aggregating data to the subject level resulted in a severe loss of information and statistical power. While aggregation satisfied the independence assumption of OLS regression, it discarded the within subject variability that characterizes the progression of Parkinson's disease. To address this limitation, we adopted a linear mixed model (LMM) approach. This is specifically designed for longitudinal data where observations are nested within subjects. By modeling the correlation structure directly, the LMM allows us to utilize the full dataset ($N = 5,875$) rather than the averaged subset ($N = 42$). This approach provides two important advantages:

- Increased Power: Utilizing every recording significantly lowers standard errors, allowing us to detect subtle effects of acoustic features that were not visible in the aggregated model.
- Subject-Specific Baselines: It accounts for patients enter the study at different severity levels (heterogeneity) by assigning a unique intercept to each subject.

3.2 Model Specification

We used a LMM with `total_UPDRS` as the continuous response variable. The model includes both fixed effects (population level predictors) and random effects (subject specific deviations).

The mathematical formulation is:

$$Y_{ij} = \beta_0 + \beta_1 X_{1ij} + \cdots + \beta_k X_{kij} + u_j + \epsilon_{ij}$$

Where:

- Y_{ij} is the `total_UPDRS` score for subject j at time i.β represents the fixed effects for the population (the average effect of `Jitter` or `Age`)
- X represents the matrix of fixed predictors: Age, Sex, and the cleaned acoustic features (`Jitter:Abs`, `Shimmer:APQ5, RPDE, DFA, PPE`)
- u_j is the random intercept for subject j, which follows a normal distribution $u_j \sim N(0, \sigma_u^2)$. This term captures the unobserved baseline severity unique to each patient
- ϵ_{ij} is the residual error

3.3 Refinement

We initiated the modeling process with a full model containing all demographic and selected acoustic predictors. To derive our final model, we employed backward elimination based on Wald tests (z-statistics), where we iteratively removed predictors that failed to reach statistical significance ($\alpha = 0.05$). Specifically:

- `Sex` ($p = 0.498$): The preliminary fit indicated no significant difference in disease severity between men and women after adjusting for other factors
- `PPE` ($p = 0.627$) `DFA` ($p = 0.084$): These non-linear measures were not significant by the inclusion of stronger predictors in the model.

The final refined model retained `Age`, `Jitter:Abs`, `Shimmer:APQ5`, and `RPDE` as fixed effects, alongside the random intercept for `subject_id`.

4 Results and Diagnostics

4.1 Final Model Summary

The primary justification for utilizing a mixed model was the hypothesis that baseline disease severity varies significantly across individuals. Examination of the random effects from our final model confirms this. The estimated variance of the random intercept (τ^2) is approximately 99.64, indicating substantial variability in the starting `total_UPDRS` scores among patients. This heterogeneity is visualized in Figure 6, where the estimated random intercepts for each subject are plotted as deviations from the overall mean. Some patients begin the study with severity scores substantially below the population average, while others start well above it. This finding validates the use of a random intercept model; a standard regression would have failed to account for these massive differences, masking the effects of the voice predictors.

The results of the refined linear mixed model are summarized in Figure 7. In contrast to the aggregated model, where voice features were statistically insignificant, the mixed model reveals that acoustic biomarkers are significant predictors of disease severity when the correlation among repeated measures is properly modeled.

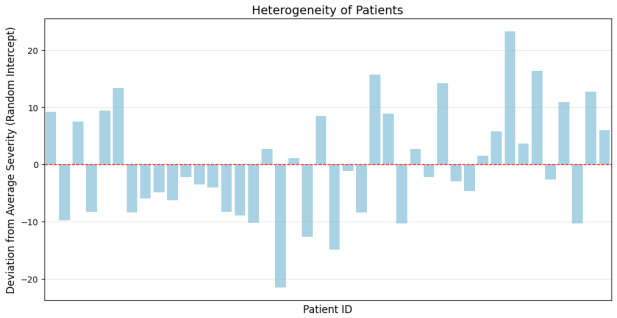


Figure 6: Estimated subject-specific random intercepts from the linear mixed-effects model. Each bar represents the deviation of an individual patient's baseline Parkinson's severity from the population average, illustrating substantial heterogeneity across subjects.

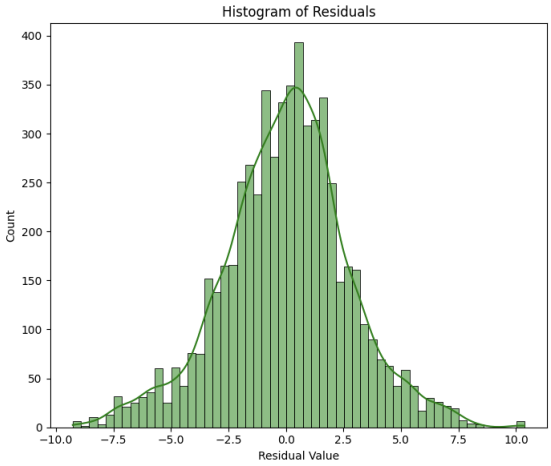


Figure 8: Residual diagnostics for the linear mixed-effects model. The histogram of residuals shows an approximately symmetric, bell-shaped distribution centered near zero.

```
======================================
OPTIMIZED MODEL RESULTS
======================================
                            OLS Regression Results
==============================================================================
Dep. Variable:            total_UPDRS   R-squared:                       0.112
Model:                            OLS   Adj. R-squared:                  0.090
Method:                 Least Squares   F-statistic:                     5.048
Date:                Tue, 16 Dec 2025   Prob (F-statistic):             0.0302
Time:                        16:42:27   Log-Likelihood:                -155.12
No. Observations:                  42   AIC:                             314.2
Df Residuals:                      40   BIC:                             317.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.1679     10.955      0.380      0.706     -17.972      26.308
age            0.3784      0.168      2.247      0.030       0.038       0.719
==============================================================================
Omnibus:                        0.902   Durbin-Watson:                   2.207
Prob(Omnibus):                  0.637   Jarque-Bera (JB):                0.958
Skew:                           0.254   Prob(JB):                        0.619
Kurtosis:                       2.462   Cond. No.                         464.
==============================================================================
```

Figure 7: Final optimized OLS regression model after stepwise backward elimination. Age emerges as the only statistically significant predictor of total UPDRS, though the model explains limited variance and does not account for subject-level dependence.

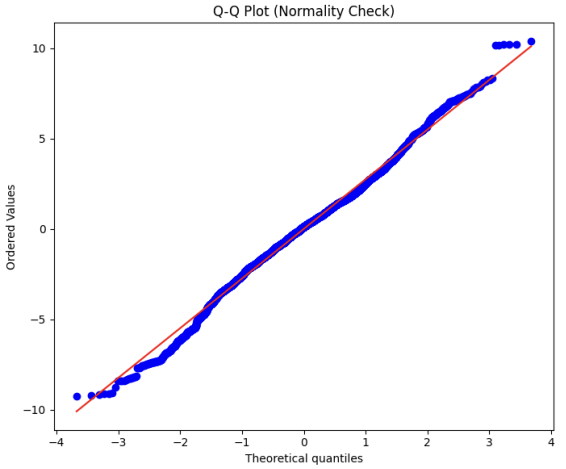


Figure 9: Normal Q-Q plot of the model residuals. The data points closely align with the reference line, indicating that the normality assumption is reasonably satisfied.

4.2 Diagnostics

Valid statistical inference requires that the model residuals follow a normal distribution. We assessed this using a histogram and a Q-Q plot. The histogram of residuals (Figure 8, left) shows a bell-shaped distribution centered at zero. The Q-Q Plot (Figure 9) demonstrates that the residuals closely hug the theoretical normal line, with only minor deviations at the extreme tails. Additionally, the Shapiro-Wilk test returned a significant result ($p < 0.001$). However, this test is known to be hypersensitive in large samples ($N = 5,875$), often flagging even negligible deviations from normality as statistically significant. Given the large sample size and the strong visual evidence from the Q-Q plot, we can conclude that the normality assumption is sufficiently met for the validity of the linear mixed model estimates.

4.3 Homoscedasticity Check

To validate the assumption of constant variance (homoscedasticity) of the errors, we examined the plot of Residuals vs. Fitted Values (Figure 9).

In a well-fitted model, this plot should display a random "cloud" of points scattered around the horizontal zero line, with no discernible patterns such as a "funnel" shape. As shown in Figure 10, the residuals appear randomly distributed with a constant bandwidth across the range of predicted values. This confirms that the linear mixed model has adequately captured the variance in the data structure, and the assumption of homoscedasticity has been satisfied.

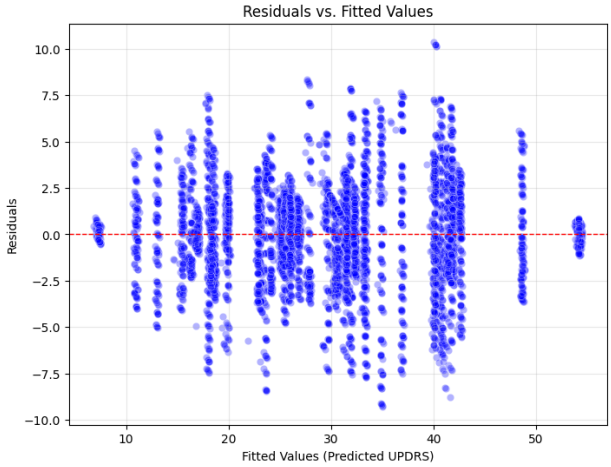


Figure 10: Plot of residuals versus fitted values from the linear mixed-effects model. The residuals display a random scatter around zero with no systematic pattern or funnel shape, indicating that the assumption of homoscedasticity is reasonably satisfied.

5 Conclusion & Future Work

This study aimed to evaluate the efficacy of vocal acoustic markers in predicting the severity of Parkinson's disease. Our analysis highlights a critical insight: standard statistical approaches that aggregate longitudinal data discard essential information regarding patient heterogeneity. As demonstrated in our preliminary analysis, an aggregated MLR model failed to identify any significant acoustic predictors, leading to an incorrect conclusion that voice features are irrelevant for tracking disease severity. By utilizing a linear mixed model, we successfully isolated the within subject variability. The results confirmed that Jitter (frequency instability), Shimmer (amplitude instability), and RPDE (non-linear complexity) are significant predictors of the UPDRS score after adjusting for subject specific baselines. The large variance found in the random intercepts ($\tau^2 \approx 99.6$) confirms that patients begin the study at vastly different severity levels. If we ignore these individual baselines, as the standard regression did, the noise overwhelms the true relationship between voice and disease progression. Ultimately, this study demonstrates that voice telemonitoring is a valid tracking tool, provided that the statistical method explicitly handles the within-subject correlation of repeated recordings.

While the linear mixed model improved upon standard regression, there are avenues for future investigation:

- Subject Expansion: Although our dataset contained nearly 6,000 observations, these were derived from only 42 subjects. To ensure these findings generalize to the broader Parkinson's population, validation on a larger, independent cohort is necessary to stabilize the estimates of population-level fixed effects.
- Non-Linear Trajectories: This study assumed a linear relationship between acoustic features and UPDRS scores. However, neurodegenerative progression is often complex and non-linear. Future analysis could involve using generalized additive mixed models (GAMMs) to capture potential plateauing or acceleration effects in the disease trajectory.

References

[1] Tsanas, A., & Little, M. (2010). *Parkinson's Telemonitoring Data Set*. UCI Machine Learning Repository. https://archive.ics.uci.edu/dataset/189/parkinsons+telemonitoring

[2] Tsanas, A., Little, M. A., McSharry, P. E., & Ramig, L. O. (2010). *Accurate telemonitoring of Parkinson's disease progression by noninvasive speech tests*. IEEE Transactions on Biomedical Engineering, 57(4), 884–893.

Data Loading and Initial Inspection of Data

In [2]:
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import statsmodels.api as sm
import statsmodels.formula.api as smf
```

In [3]:
```
file_path = 'parkinsons_updrs.csv'
df = pd.read_csv(file_path)
df.head()
```

Out[3]:

	subject#	age	sex	test_time	motor_UPDRS	total_UPDRS	Jitter(%)	Jitter(Al
0	1	72	0	5.6431	28.199	34.398	0.00662	0.0000
1	1	72	0	12.6660	28.447	34.894	0.00300	0.0000
2	1	72	0	19.6810	28.695	35.389	0.00481	0.0000
3	1	72	0	25.6470	28.905	35.810	0.00528	0.0000
4	1	72	0	33.6420	29.187	36.375	0.00335	0.0000

5 rows × 22 columns

In [4]:
```
df.shape
df.info()
df.isna().sum()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 5875 entries, 0 to 5874
Data columns (total 22 columns):
 #   Column         Non-Null Count  Dtype  
---  ------         --------------  -----  
 0   subject#       5875 non-null   int64  
 1   age            5875 non-null   int64  
 2   sex            5875 non-null   int64  
 3   test_time      5875 non-null   float64
 4   motor_UPDRS    5875 non-null   float64
 5   total_UPDRS    5875 non-null   float64
 6   Jitter(%)      5875 non-null   float64
 7   Jitter(Abs)    5875 non-null   float64
 8   Jitter:RAP     5875 non-null   float64
 9   Jitter:PPQ5    5875 non-null   float64
 10  Jitter:DDP     5875 non-null   float64
 11  Shimmer        5875 non-null   float64
 12  Shimmer(dB)    5875 non-null   float64
 13  Shimmer:APQ3   5875 non-null   float64
 14  Shimmer:APQ5   5875 non-null   float64
 15  Shimmer:APQ11  5875 non-null   float64
 16  Shimmer:DDA    5875 non-null   float64
 17  NHR            5875 non-null   float64
 18  HNR            5875 non-null   float64
 19  RPDE           5875 non-null   float64
 20  DFA            5875 non-null   float64
 21  PPE            5875 non-null   float64
dtypes: float64(19), int64(3)
memory usage: 1009.9 KB
```

Out[4]:

```
subject#         0
age              0
sex              0
test_time        0
motor_UPDRS      0
total_UPDRS      0
Jitter(%)        0
Jitter(Abs)      0
Jitter:RAP       0
Jitter:PPQ5      0
Jitter:DDP       0
Shimmer          0
Shimmer(dB)      0
Shimmer:APQ3     0
Shimmer:APQ5     0
Shimmer:APQ11    0
Shimmer:DDA      0
NHR              0
HNR              0
RPDE             0
DFA              0
PPE              0
dtype: int64
```

In [5]:
```
df['subject#'].nunique()
```

Out[5]: 42

Patient-Level Aggregation (Baseline Dataset)

To satisfy the independence assumption for linear regression, repeated recordings were averaged at the patient level.

In [6]:
```
df_patient = (
    df.groupby('subject#')
      .mean(numeric_only=True)
      .reset_index()
)

df_patient.shape
```

Out[6]: (42, 22)

In [7]:
```
df_patient.head()
```

Out[7]:

	subject#	age	sex	test_time	motor_UPDRS	total_UPDRS	Jitter(%)	Jitter(A
0	1	72.0	0.0	89.176971	31.898933	40.733201	0.004284	0.000
1	2	58.0	0.0	90.364266	13.812538	16.284014	0.006721	0.000
2	3	57.0	0.0	87.829658	27.124785	33.359701	0.003318	0.000
3	4	74.0	0.0	91.029428	15.790825	23.587321	0.004967	0.000
4	5	75.0	0.0	85.623752	31.632603	41.853987	0.004937	0.000

5 rows × 22 columns

In [8]:
```
df_patient.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 42 entries, 0 to 41
Data columns (total 22 columns):
 #   Column          Non-Null Count  Dtype  
---  ------          --------------  -----  
 0   subject#        42 non-null     int64  
 1   age             42 non-null     float64
 2   sex             42 non-null     float64
 3   test_time       42 non-null     float64
 4   motor_UPDRS     42 non-null     float64
 5   total_UPDRS     42 non-null     float64
 6   Jitter(%)       42 non-null     float64
 7   Jitter(Abs)     42 non-null     float64
 8   Jitter:RAP      42 non-null     float64
 9   Jitter:PPQ5     42 non-null     float64
 10  Jitter:DDP      42 non-null     float64
 11  Shimmer         42 non-null     float64
 12  Shimmer(dB)     42 non-null     float64
 13  Shimmer:APQ3    42 non-null     float64
 14  Shimmer:APQ5    42 non-null     float64
 15  Shimmer:APQ11   42 non-null     float64
 16  Shimmer:DDA     42 non-null     float64
 17  NHR             42 non-null     float64
 18  HNR             42 non-null     float64
 19  RPDE            42 non-null     float64
 20  DFA             42 non-null     float64
 21  PPE             42 non-null     float64
dtypes: float64(21), int64(1)
memory usage: 7.3 KB
```

Exploratory Data Analysis

Summary statistics and correlations are examined to understand the distribution of disease severity and its relationship with demographic and voice features.

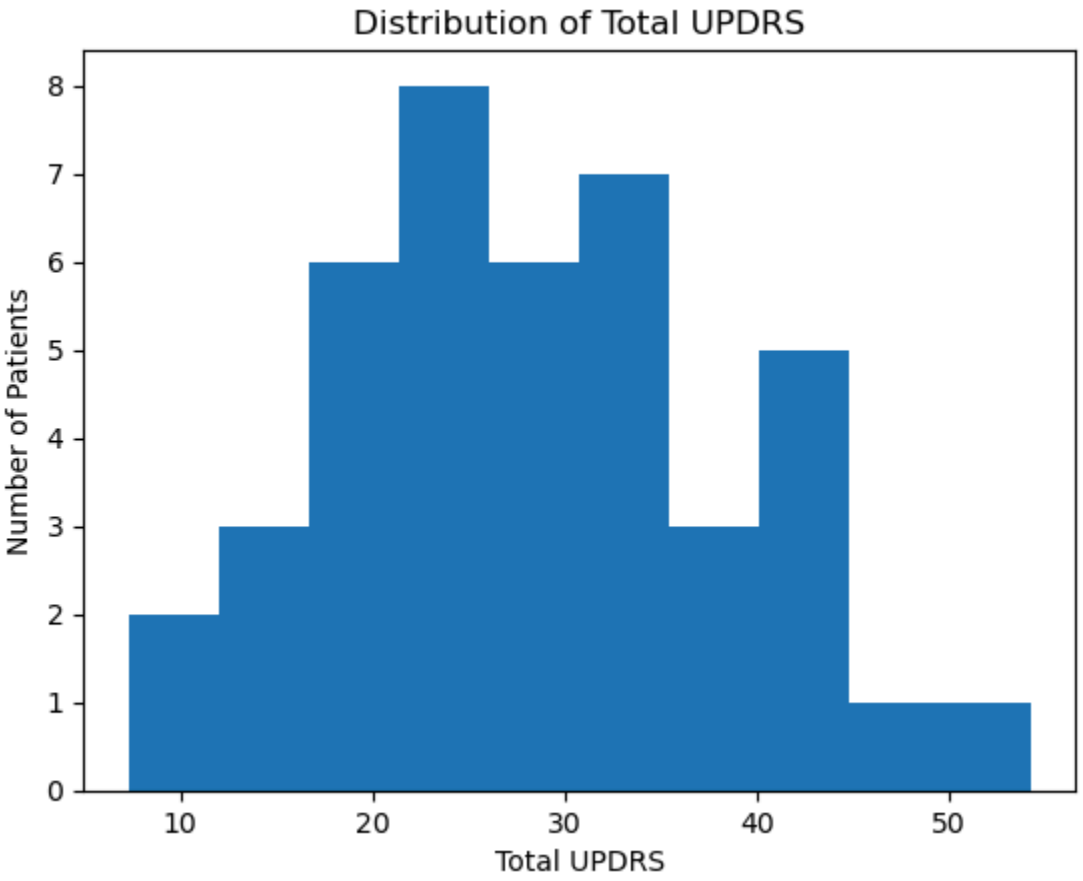
In [9]:
```
df_patient = df_patient.drop(columns=['subject#'])
```

In [10]:
```
df_patient['total_UPDRS'].describe()
```

Out[10]:
```
count    42.000000
mean     28.537761
std      10.443513
min       7.300971
25%      20.626375
50%      27.322406
75%      34.505466
max      54.253109
Name: total_UPDRS, dtype: float64
```

In [11]:
```
plt.hist(df_patient['total_UPDRS'], bins=10)
plt.xlabel("Total UPDRS")
```

```
plt.ylabel("Number of Patients")
plt.title("Distribution of Total UPDRS")
plt.show()
```



In [12]:
```
corr_with_updrs = df_patient.corr(numeric_only=True)['total_UPDRS']
corr_with_updrs.sort_values(ascending=False)
```

Out[12]:
```
total_UPDRS      1.000000
motor_UPDRS      0.950280
age              0.334764
RPDE             0.277052
PPE              0.247779
Shimmer:APQ11    0.196247
Shimmer(dB)      0.161098
Jitter(%)        0.159445
Shimmer          0.153024
Jitter:DDP       0.150366
Jitter:RAP       0.150347
Shimmer:DDA      0.146023
Shimmer:APQ3     0.146020
Jitter(Abs)      0.140860
Shimmer:APQ5     0.140557
Jitter:PPQ5      0.123734
NHR              0.104837
test_time       -0.089201
sex             -0.122419
DFA             -0.123420
HNR             -0.254166
Name: total_UPDRS, dtype: float64
```

In [13]:
```
corr_matrix = df_patient.corr(numeric_only=True)
corr_matrix
```

Out[13]:

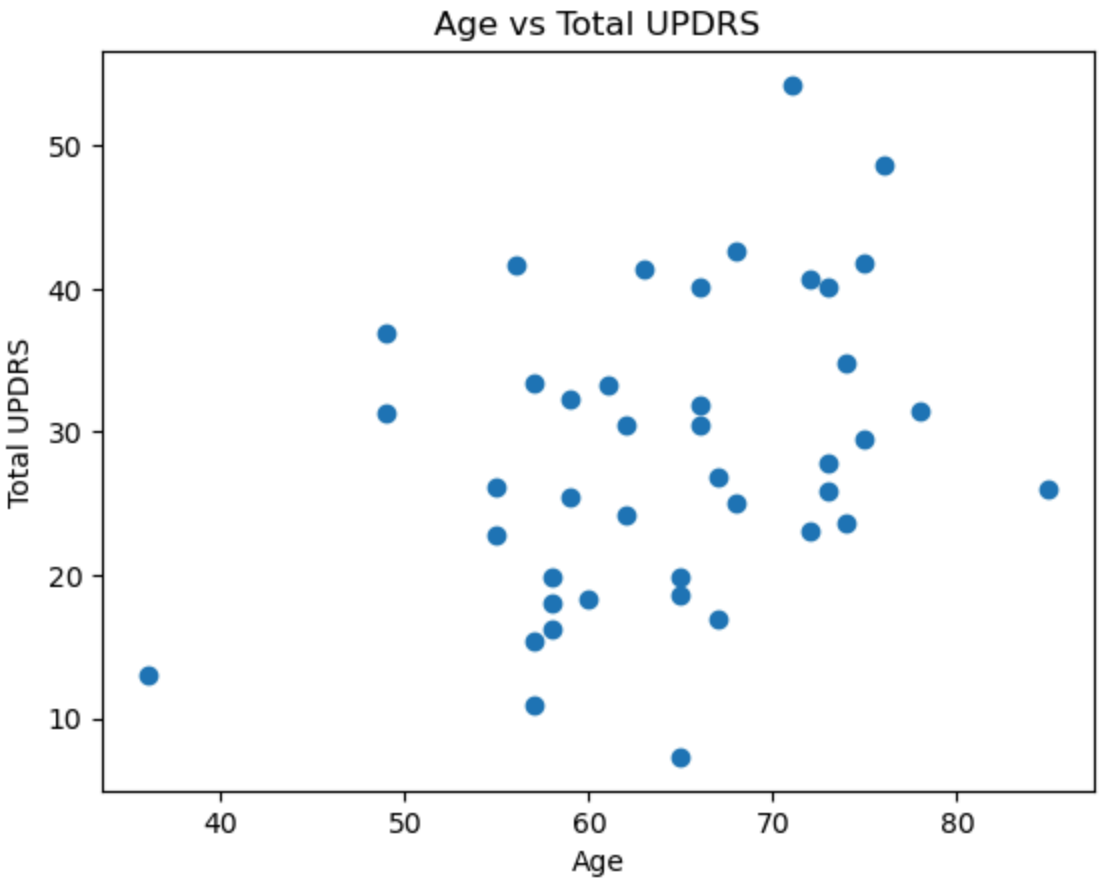
	age	sex	test_time	motor_UPDRS	total_UPDRS	Ji
age	1.000000	-0.064549	0.133709	0.298982	0.334764	0.
sex	-0.064549	1.000000	-0.070086	-0.056976	-0.122419	0.
test_time	0.133709	-0.070086	1.000000	-0.043527	-0.089201	-0.
motor_UPDRS	0.298982	-0.056976	-0.043527	1.000000	0.950280	0.
total_UPDRS	0.334764	-0.122419	-0.089201	0.950280	1.000000	0.
Jitter(%)	0.055375	0.071129	-0.027940	0.181275	0.159445	1.
Jitter(Abs)	0.075591	-0.247747	-0.012126	0.119782	0.140860	0.
Jitter:RAP	0.036285	0.124511	-0.058913	0.168658	0.150347	0.
Jitter:PPQ5	0.033397	0.128117	-0.076000	0.147917	0.123734	0.
Jitter:DDP	0.036294	0.124476	-0.058875	0.168684	0.150366	0.
Shimmer	0.146283	0.071146	-0.053642	0.171867	0.153024	0.
Shimmer(dB)	0.157692	0.068260	-0.043971	0.181015	0.161098	0.
Shimmer:APQ3	0.154665	0.055973	-0.061671	0.158008	0.146020	0.
Shimmer:APQ5	0.129722	0.079905	-0.051689	0.157423	0.140557	0.
Shimmer:APQ11	0.193419	0.023993	0.027285	0.224141	0.196247	0.
Shimmer:DDA	0.154668	0.055974	-0.061667	0.158010	0.146023	0.
NHR	0.017992	0.214789	-0.167564	0.126266	0.104837	0.
HNR	-0.150907	0.015510	0.098105	-0.251115	-0.254166	-0.
RPDE	0.135006	-0.243198	-0.130064	0.238430	0.277052	0.
DFA	-0.084598	-0.203558	0.306107	-0.122543	-0.123420	0.
PPE	0.182906	-0.163616	0.087133	0.260960	0.247779	0.

21 rows × 21 columns

We observe high correlations here, which suggests we might have multicollinearity issues to check later.

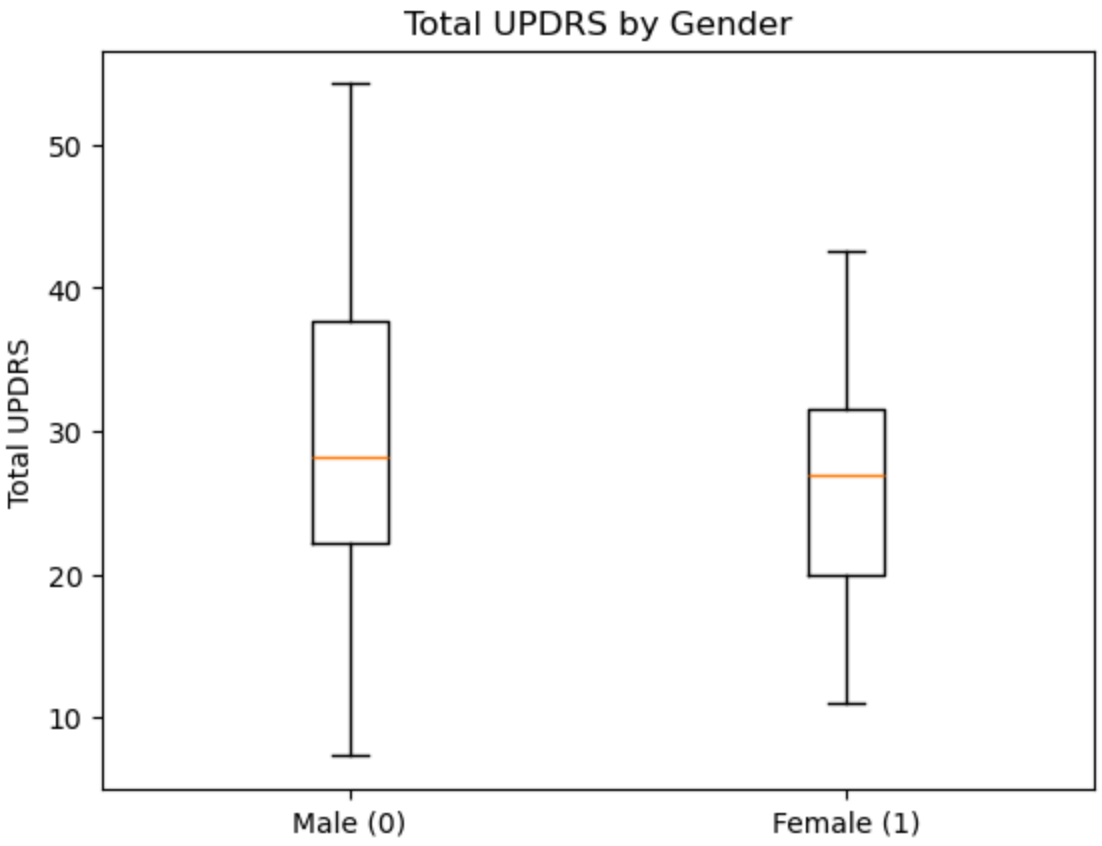
In [14]:

```
plt.scatter(
    df_patient['age'],
    df_patient['total_UPDRS']
)
plt.xlabel("Age")
plt.ylabel("Total UPDRS")
plt.title("Age vs Total UPDRS")
plt.show()
```



In [15]:
```
plt.boxplot(
    [df_patient[df_patient['sex'] == 0]['total_UPDRS'],
     df_patient[df_patient['sex'] == 1]['total_UPDRS']],
    labels=['Male (0)', 'Female (1)']
)
plt.ylabel("Total UPDRS")
plt.title("Total UPDRS by Gender")
plt.show()
```

```
C:\Users\Dell\AppData\Local\Temp\ipykernel_29200\4292131720.py:1: MatplotlibDeprecationWarning: The 'labels' parameter of boxplot() has been renamed 'tick_labels' since Matplotlib 3.9; support for the old name will be dropped in 3.11.
  plt.boxplot(
```



Baseline Model Preparation

In [16]:
```
df_model = df_patient.drop(columns=['motor_UPDRS'])
df_model.columns
```

Out[16]:
```
Index(['age', 'sex', 'test_time', 'total_UPDRS', 'Jitter(%)', 'Jitter(Abs)',
       'Jitter:RAP', 'Jitter:PPQ5', 'Jitter:DDP', 'Shimmer', 'Shimmer(dB)',
       'Shimmer:APQ3', 'Shimmer:APQ5', 'Shimmer:APQ11', 'Shimmer:DDA', 'NHR',
       'HNR', 'RPDE', 'DFA', 'PPE'],
      dtype='object')
```

The motor_UPDRS variable is removed to avoid redundancy with the total_UPDRS outcome.

In [17]:
```
df_model.head()
```

Out[17]:

	age	sex	test_time	total_UPDRS	Jitter(%)	Jitter(Abs)	Jitter:RAP	Jitter:PPQ
0	72.0	0.0	89.176971	40.733201	0.004284	0.000024	0.001885	0.00200
1	58.0	0.0	90.364266	16.284014	0.006721	0.000051	0.003441	0.00356
2	57.0	0.0	87.829658	33.359701	0.003318	0.000020	0.001588	0.00179
3	74.0	0.0	91.029428	23.587321	0.004967	0.000039	0.002306	0.00259
4	75.0	0.0	85.623752	41.853987	0.004937	0.000043	0.002248	0.00264

Patient-Level Linear Regression (Baseline OLS)

In [18]:

```
X = df_model.drop(columns=['total_UPDRS', 'subject#'], errors='ignore')
y = df_model['total_UPDRS']
X = sm.add_constant(X)
model_full = sm.OLS(y, X).fit()
print(model_full.summary())
```

OLS Regression Results

Dep. Variable:	total_UPDRS	R-squared:	0.473
Model:	OLS	Adj. R-squared:	0.019
Method:	Least Squares	F-statistic:	1.041
Date:	Wed, 17 Dec 2025	Prob (F-statistic):	0.460
Time:	19:05:36	Log-Likelihood:	-144.15
No. Observations:	42	AIC:	328.3
Df Residuals:	22	BIC:	363.1
Df Model:	19		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	38.3340	83.911	0.457	0.652	-135.687	212.355
age	0.2200	0.230	0.956	0.349	-0.257	0.697
sex	-9.3936	5.759	-1.631	0.117	-21.337	2.550
test_time	-0.3453	0.268	-1.287	0.212	-0.902	0.211
Jitter(%)	1946.5412	1.09e+04	0.179	0.860	-2.07e+04	2.45e+04
Jitter(Abs)	-6.341e+05	3.32e+05	-1.907	0.070	-1.32e+06	5.54e+04
Jitter:RAP	-4.438e+06	7.51e+06	-0.591	0.560	-2e+07	1.11e+07
Jitter:PPQ5	-1.421e+04	1.77e+04	-0.802	0.431	-5.09e+04	2.25e+04
Jitter:DDP	1.486e+06	2.5e+06	0.594	0.559	-3.7e+06	6.67e+06
Shimmer	-3505.6023	6153.923	-0.570	0.575	-1.63e+04	9256.853
Shimmer(dB)	168.8608	297.035	0.568	0.575	-447.153	784.875
Shimmer:APQ3	-3.284e+06	9.44e+06	-0.348	0.731	-2.29e+07	1.63e+07
Shimmer:APQ5	4528.5016	4764.226	0.951	0.352	-5351.899	1.44e+04
Shimmer:APQ11	970.0653	2958.809	0.328	0.746	-5166.128	7106.259
Shimmer:DDA	1.094e+06	3.15e+06	0.348	0.731	-5.43e+06	7.62e+06
NHR	-2.9054	368.920	-0.008	0.994	-767.998	762.187
HNR	-0.0563	2.098	-0.027	0.979	-4.407	4.294
RPDE	66.9014	55.162	1.213	0.238	-47.497	181.299
DFA	-45.0992	55.359	-0.815	0.424	-159.906	69.7

```
08
PPE              45.8282    112.733      0.407      0.688    -187.966     279.6
22
==============================================================================
Omnibus:                        2.034   Durbin-Watson:                   2.240
Prob(Omnibus):                  0.362   Jarque-Bera (JB):                1.115
Skew:                          -0.084   Prob(JB):                        0.573
Kurtosis:                       3.780   Cond. No.                     7.23e+08
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctl
y specified.
[2] The smallest eigenvalue is 1.07e-12. This might indicate that there are
strong multicollinearity problems or that the design matrix is singular.
```

Coefficients are massively large, detecting strong multicollinearity.

Multicollinearity Diagnosis (VIF Analysis)

In [19]:
```
from statsmodels.stats.outliers_influence import variance_inflation_factor

vif_data = pd.DataFrame()
vif_data["Variable"] = X.columns
vif_data["VIF"] = [variance_inflation_factor(X.values, i) for i in range(X.sha

print("VIF Scores (The Evidence for Dropping Columns):")
print(vif_data.sort_values(by="VIF", ascending=False))
```

```
VIF Scores (The Evidence for Dropping Columns):
         Variable           VIF
14    Shimmer:DDA  3.185232e+09
11   Shimmer:APQ3  3.185003e+09
6      Jitter:RAP  6.140895e+07
8      Jitter:DDP  6.137933e+07
9         Shimmer  5.866521e+03
0           const  2.763057e+03
12   Shimmer:APQ5  1.500538e+03
10    Shimmer(dB)  1.087144e+03
13  Shimmer:APQ11  7.719849e+02
7     Jitter:PPQ5  6.919692e+02
4       Jitter(%)  5.012674e+02
15            NHR  1.153421e+02
5     Jitter(Abs)  2.519892e+01
19            PPE  2.126571e+01
16            HNR  2.031925e+01
17           RPDE  6.266786e+00
18            DFA  4.651190e+00
2             sex  2.892267e+00
3       test_time  1.951236e+00
1             age  1.730137e+00
```

Using VIF values to determine which variables to drop

In [20]:
```
cols_to_drop = [
    'Jitter(%)', 'Jitter:RAP', 'Jitter:PPQ5', 'Jitter:DDP',  # Keep Jitter:Abs
    'Shimmer', 'Shimmer(dB)', 'Shimmer:APQ3', 'Shimmer:APQ11', 'Shimmer:DDA' #
]

df_refined = df_model.drop(columns=cols_to_drop, errors='ignore')
```

Model Refinement via VIF and Backward Elimination

Highly collinear predictors are iteratively removed using VIF thresholds, followed by backward elimination based on statistical significance.

In [21]:
```
X_refined = df_refined.drop(columns=['total_UPDRS', 'subject#'], errors='ignor
y_refined = df_refined['total_UPDRS']

X_refined = sm.add_constant(X_refined)

vif_data_refined = pd.DataFrame()
vif_data_refined["Variable"] = X_refined.columns
vif_data_refined["VIF"] = [variance_inflation_factor(X_refined.values, i) for

print("Refined VIF Scores (Goal: All < 10):")
print(vif_data_refined.sort_values(by="VIF", ascending=False))
```

```
Refined VIF Scores (Goal: All < 10):
        Variable          VIF
0          const  1862.198812
5   Shimmer:APQ5    16.376289
6            NHR    14.708225
7            HNR    12.501572
4    Jitter(Abs)    12.032674
10           PPE     8.341166
8           RPDE     4.214249
9            DFA     2.517176
2            sex     1.822520
3      test_time     1.356948
1            age     1.247753
```

Feature Removal Based on VIF Threshold

Dropping NHR

In [22]:
```
df_final = df_refined.drop(columns=['NHR'], errors='ignore')

X_final = df_final.drop(columns=['total_UPDRS', 'subject#'], errors='ignore')
y_final = df_final['total_UPDRS']
```

```
X_final = sm.add_constant(X_final)
```

Recalculation of VIF After Feature Removal

Recalculating VIF to verify

In [23]:
```
vif_data_final = pd.DataFrame()
vif_data_final["Variable"] = X_final.columns
vif_data_final["VIF"] = [variance_inflation_factor(X_final.values, i) for i in

print("Final VIF Scores (Goal: All < 10):")
print(vif_data_final.sort_values(by="VIF", ascending=False))
```

```
Final VIF Scores (Goal: All < 10):
       Variable          VIF
0         const  1830.724798
6           HNR    12.230257
4   Jitter(Abs)     8.494466
9           PPE     8.084106
5  Shimmer:APQ5     7.025490
7          RPDE     4.200811
8           DFA     1.816850
3     test_time     1.337887
2           sex     1.293180
1           age     1.164667
```

Final Predictor Set After Multicollinearity Control

Drop HNR

In [24]:
```
df_final_clean = df_final.drop(columns=['HNR'], errors='ignore')
X_clean = df_final_clean.drop(columns=['total_UPDRS', 'subject#'], errors='igr
y_clean = df_final_clean['total_UPDRS']
X_clean = sm.add_constant(X_clean)
```

Final VIF Verification

In [25]:
```
vif_data_clean = pd.DataFrame()
vif_data_clean["Variable"] = X_clean.columns
vif_data_clean["VIF"] = [variance_inflation_factor(X_clean.values, i) for i in
print("--- Final VIF Scores (Must be < 10) ---")
print(vif_data_clean.sort_values(by="VIF", ascending=False))
```

```
--- Final VIF Scores (Must be < 10) ---
       Variable         VIF
0         const  392.515115
4   Jitter(Abs)    7.192037
8           PPE    5.940421
5  Shimmer:APQ5    2.730137
6          RPDE    2.532850
7           DFA    1.701005
2           sex    1.283640
3     test_time    1.245533
1           age    1.164215
```

Final OLS Model Fitting

In [26]:
```
final_model = sm.OLS(y_clean, X_clean).fit()
print(final_model.summary())
```

OLS Regression Results

Dep. Variable:	total_UPDRS	R-squared:	0.264
Model:	OLS	Adj. R-squared:	0.086
Method:	Least Squares	F-statistic:	1.481
Date:	Wed, 17 Dec 2025	Prob (F-statistic):	0.202
Time:	19:05:37	Log-Likelihood:	-151.18
No. Observations:	42	AIC:	320.4
Df Residuals:	33	BIC:	336.0
Df Model:	8		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	18.9687	30.527	0.621	0.539	-43.138	81.076
age	0.2506	0.182	1.376	0.178	-0.120	0.621
sex	-2.5224	3.703	-0.681	0.501	-10.057	5.012
test_time	-0.1077	0.207	-0.520	0.606	-0.529	0.313
Jitter(Abs)	-2.037e+05	1.71e+05	-1.189	0.243	-5.53e+05	1.45e+05
Shimmer:APQ5	-60.9416	196.149	-0.311	0.758	-460.011	338.127
RPDE	34.5330	33.849	1.020	0.315	-34.333	103.399
DFA	-38.5646	32.313	-1.193	0.241	-104.307	27.177
PPE	95.6976	57.510	1.664	0.106	-21.308	212.703

Omnibus:	1.247	Durbin-Watson:	2.572
Prob(Omnibus):	0.536	Jarque-Bera (JB):	1.054
Skew:	0.377	Prob(JB):	0.590
Kurtosis:	2.819	Cond. No.	1.26e+07

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 1.26e+07. This might indicate that there are strong multicollinearity or other numerical problems.

Model is quite noisy, as p value of 0.202 > 0.05, indicating several weak predictors.

Must perform stepwise selection -> backward elimination

Stepwise Backward Elimination Based on Statistical Significance

In [27]:
```
#Backward Elimination based on P-values
def backward_elimination(data, target, significance_level=0.05):
    features = data.columns.tolist()
    while len(features) > 0:
        features_with_const = sm.add_constant(data[features])
        p_values = sm.OLS(target, features_with_const).fit().pvalues[1:]

        max_p_value = p_values.max()
        if max_p_value >= significance_level:
            excluded_feature = p_values.idxmax()
            print(f"Dropping: {excluded_feature} (p-value={max_p_value:.4f})")
            features.remove(excluded_feature)
        else:
            break

    return features

print("Starting Stepwise Backward Elimination")

X_input = df_final_clean.drop(columns=['total_UPDRS', 'subject#'], errors='igr
y_input = df_final_clean['total_UPDRS']

selected_features = backward_elimination(X_input, y_input)

print(f"\nBest Features Selected: {selected_features}")
```

```
Starting Stepwise Backward Elimination
Dropping: Shimmer:APQ5 (p-value=0.7580)
Dropping: test_time (p-value=0.5919)
Dropping: sex (p-value=0.4207)
Dropping: Jitter(Abs) (p-value=0.2093)
Dropping: PPE (p-value=0.4267)
Dropping: DFA (p-value=0.2715)
Dropping: RPDE (p-value=0.1175)

Best Features Selected: ['age']
```

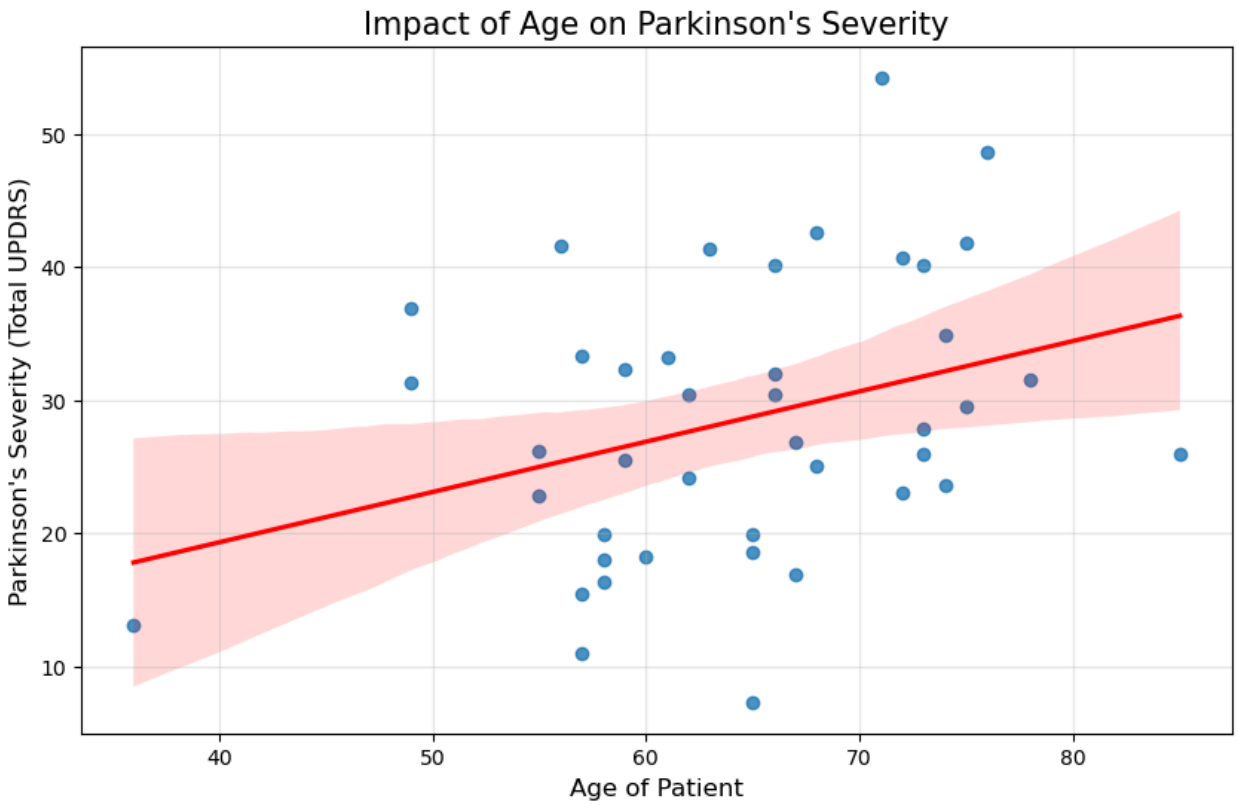
Optimized Regression Model

In [28]:
```
X_selected = sm.add_constant(X_input[selected_features])
final_model_optimized = sm.OLS(y_input, X_selected).fit()

print("\n" + "="*40)
print("OPTIMIZED MODEL RESULTS")
print("="*40)
print(final_model_optimized.summary())
```

```
========================================
OPTIMIZED MODEL RESULTS
========================================
                            OLS Regression Results
==============================================================================
Dep. Variable:            total_UPDRS   R-squared:                       0.112
Model:                            OLS   Adj. R-squared:                  0.090
Method:                 Least Squares   F-statistic:                     5.048
Date:                Wed, 17 Dec 2025   Prob (F-statistic):             0.0302
Time:                        19:05:38   Log-Likelihood:                -155.12
No. Observations:                  42   AIC:                             314.2
Df Residuals:                      40   BIC:                             317.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const          4.1679     10.955      0.380      0.706     -17.972      26.308
age            0.3784      0.168      2.247      0.030       0.038       0.719
==============================================================================
Omnibus:                        0.902   Durbin-Watson:                   2.207
Prob(Omnibus):                  0.637   Jarque-Bera (JB):                0.958
Skew:                           0.254   Prob(JB):                        0.619
Kurtosis:                       2.462   Cond. No.                         464.
==============================================================================

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

Visualization: Effect of Age on Parkinson’s Severity

In [29]:
```
plt.figure(figsize=(10, 6))
sns.regplot(x='age', y='total_UPDRS', data=df_final_clean, line_kws={"color":

plt.title('Impact of Age on Parkinson\'s Severity', fontsize=15)
plt.xlabel('Age of Patient', fontsize=12)
plt.ylabel('Parkinson\'s Severity (Total UPDRS)', fontsize=12)
plt.grid(True, alpha=0.3)
plt.show()
```



Mixed-Effects Modeling

In [30]:

```
df_full = pd.read_csv('parkinsons_updrs.csv')

cols_to_drop = [
    'motor_UPDRS',
    'Jitter(%)', 'Jitter:RAP', 'Jitter:PPQ5', 'Jitter:DDP',
    'Shimmer', 'Shimmer(dB)', 'Shimmer:APQ3', 'Shimmer:APQ11', 'Shimmer:DDA',
    'NHR', 'HNR'
]
df_full_clean = df_full.drop(columns=cols_to_drop, errors='ignore')

df_full_clean = df_full_clean.rename(columns={'subject#': 'subject_id'})

df_full_clean.columns = (df_full_clean.columns
                         .str.replace('(', '_', regex=False)
                         .str.replace(')', '', regex=False)
                         .str.replace(':', '_', regex=False))

predictors = df_full_clean.columns.drop(['subject_id', 'total_UPDRS', 'test_ti
formula = "total_UPDRS ~ " + " + ".join(predictors)

print(f"Formula: {formula}")

model_mixed = smf.mixedlm(formula, df_full_clean, groups=df_full_clean["subjec
result_mixed = model_mixed.fit()
```

```
print(result_mixed.summary())
```

```
Formula: total_UPDRS ~ age + sex + Jitter_Abs + Shimmer_APQ5 + RPDE + DFA + PPE
           Mixed Linear Model Regression Results
============================================================
Model:             MixedLM  Dependent Variable:  total_UPDRS
No. Observations:  5875     Method:              REML
No. Groups:        42       Scale:               7.6436
Min. group size:   101      Log-Likelihood:      -14449.4056
Max. group size:   168      Converged:           Yes
Mean group size:   139.9
------------------------------------------------------------
               Coef.   Std.Err.    z    P>|z|  [0.025  0.975]
------------------------------------------------------------
Intercept       7.297   11.216  0.651 0.515 -14.686   29.280
age             0.373    0.170  2.193 0.028   0.040    0.706
sex            -2.233    3.293 -0.678 0.498  -8.688    4.222
Jitter_Abs   4670.689 2020.782  2.311 0.021 710.029 8631.349
Shimmer_APQ5  -12.057    4.121 -2.926 0.003 -20.135   -3.980
RPDE           -1.163    0.575 -2.023 0.043  -2.289   -0.036
DFA            -1.953    1.131 -1.726 0.084  -4.171    0.265
PPE            -0.409    0.842 -0.486 0.627  -2.059    1.241
Group Var     100.738    8.266
============================================================
```

In [33]:
```
model_comparison = pd.DataFrame({
    "Model": [
        "OLS (After VIF)",
        "OLS (Backward Elimination)",
        "Mixed Linear Model"
    ],
    "Num_Predictors": [
        int(final_model.df_model),
        int(final_model_optimized.df_model),
        len(result_refined.fe_params) - 1
    ],
    "R_squared / Pseudo_R2": [
        final_model.rsquared,
        final_model_optimized.rsquared,
        None  # MixedLM does not have classical R²
    ],
    "AIC": [
        final_model.aic,
        final_model_optimized.aic,
        result_refined.aic
    ],
    "BIC": [
        final_model.bic,
        final_model_optimized.bic,
        result_refined.bic
    ],
    "Accounts_for_Subject_Effects": [
        "No",
```

```
        "No",
        "Yes"
    ]
})

model_comparison
```

Out[33]:

	Model	Num_Predictors	R_squared / Pseudo_R2	AIC	BIC	Accounts_for
0	OLS (After VIF)	8	0.264155	320.358218	335.997245	
1	OLS (Backward Elimination)	1	0.112067	314.249082	317.724421	
2	Mixed Linear Model	4	NaN	NaN	NaN	

Mixed-Effects Model Refinement and Predictor Selection

In [32]:

```
formula_refined = "total_UPDRS ~ age + Jitter_Abs + Shimmer_APQ5 + RPDE"

print(f"Final Formula: {formula_refined}")

model_refined = smf.mixedlm(formula_refined, df_full_clean, groups=df_full_cle
result_refined = model_refined.fit()

print(result_refined.summary())

print("\n--- P-Values for Final Predictors ---")
print(result_refined.pvalues)
```

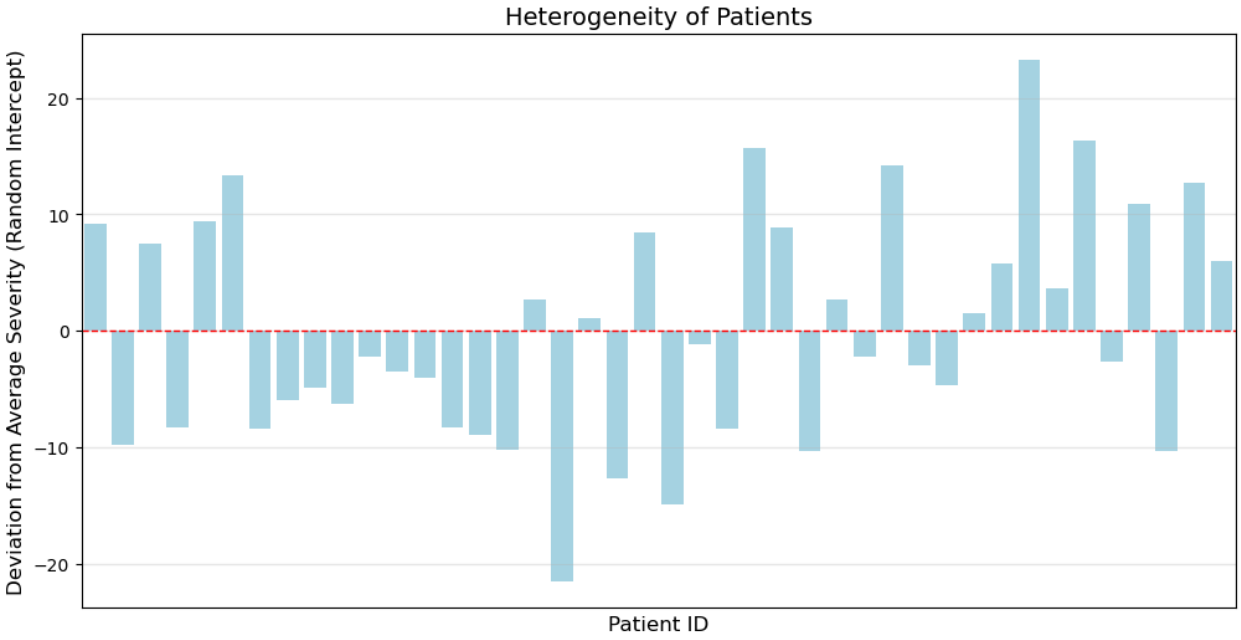
```
Final Formula: total_UPDRS ~ age + Jitter_Abs + Shimmer_APQ5 + RPDE
           Mixed Linear Model Regression Results
============================================================
Model:             MixedLM  Dependent Variable:  total_UPDRS
No. Observations:  5875     Method:              REML
No. Groups:        42       Scale:               7.6455
Min. group size:   101      Log-Likelihood:      -14455.2990
Max. group size:   168      Converged:           Yes
Mean group size:   139.9
------------------------------------------------------------
               Coef.   Std.Err.    z    P>|z|  [0.025  0.975]
------------------------------------------------------------
Intercept       4.735   10.981  0.431 0.666 -16.788   26.257
age             0.381    0.169  2.260 0.024   0.051    0.712
Jitter_Abs   3450.392 1653.170  2.087 0.037 210.239 6690.545
Shimmer_APQ5  -12.940    4.073 -3.177 0.001 -20.924   -4.956
RPDE           -1.200    0.556 -2.159 0.031  -2.289   -0.111
Group Var      99.635    8.074
============================================================


--- P-Values for Final Predictors ---
Intercept       0.666357
age             0.023842
Jitter_Abs      0.036876
Shimmer_APQ5    0.001490
RPDE            0.030864
Group Var       0.000008
dtype: float64
```

Final Mixed-Effects Model Results

In [34]:

```python
random_effects = result_refined.random_effects
re_data = {k: v.iloc[0] for k, v in random_effects.items()}

re_df = pd.DataFrame.from_dict(re_data, orient='index', columns=['Subject_Effe
re_df = re_df.sort_values('Subject_Effect')

print("Preview of Subject Effects:")
print(re_df.head())

plt.figure(figsize=(12, 6))
sns.barplot(x=re_df.index, y=re_df['Subject_Effect'], color='skyblue', alpha=0
plt.axhline(0, color='red', linestyle='--', linewidth=1)

plt.title('Heterogeneity of Patients', fontsize=14)
plt.xlabel('Patient ID', fontsize=12)
plt.ylabel('Deviation from Average Severity (Random Intercept)', fontsize=12)
plt.xticks([]) # Hide messy x-labels
plt.grid(axis='y', alpha=0.3)
plt.show()
```

```
Preview of Subject Effects:
    Subject_Effect
18      -21.583821
22      -14.953442
20      -12.693588
40      -10.382419
27      -10.304575
```



Random Effects Diagnostics and Subject Heterogeneity

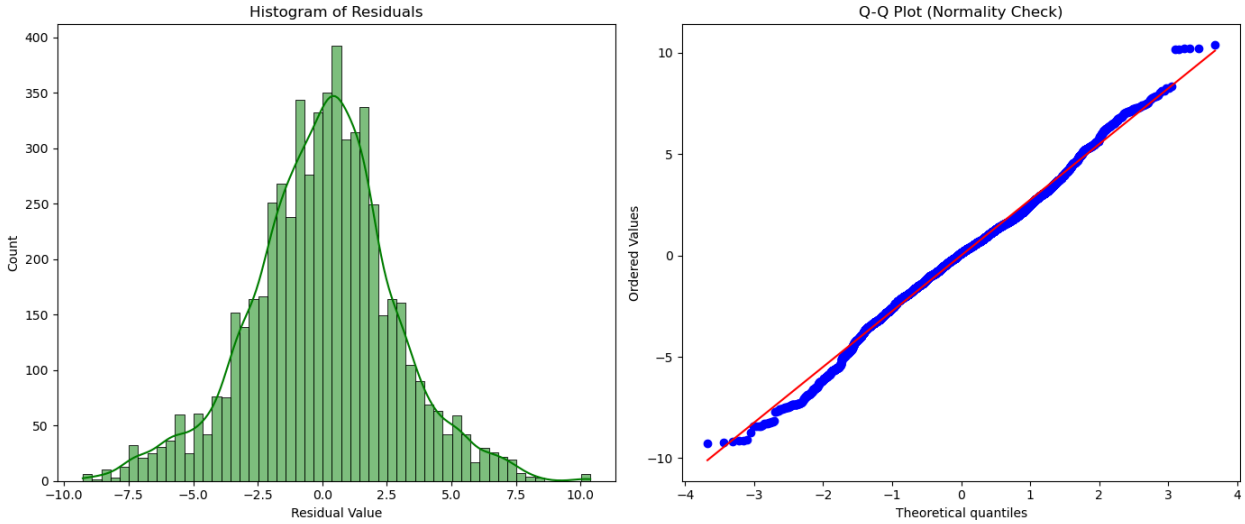
In [35]:
```
group_variance = result_refined.cov_re.iloc[0, 0]
residual_variance = result_refined.scale

# ICC calculation
icc = group_variance / (group_variance + residual_variance)

print(f"Intra-Class Correlation (ICC): {icc:.3f}")
```

```
Intra-Class Correlation (ICC): 0.929
```

In [36]:
```
from scipy import stats
```

In [37]:
```
residuals = result_refined.resid

fig, ax = plt.subplots(1, 2, figsize=(14, 6))

sns.histplot(residuals, kde=True, ax=ax[0], color='green')
ax[0].set_title('Histogram of Residuals')
ax[0].set_xlabel('Residual Value')

stats.probplot(residuals, dist="norm", plot=ax[1])
ax[1].set_title('Q-Q Plot (Normality Check)')
```

```
ax[1].get_lines()[0].set_color('blue') # dots
ax[1].get_lines()[1].set_color('red') # line

plt.tight_layout()
plt.show()

shapiro_p = stats.shapiro(residuals).pvalue
print(f"Shapiro-Wilk P-Value: {shapiro_p:.4f}")
```




```
Shapiro-Wilk P-Value: 0.0000
```

```
C:\Users\Dell\anaconda3\Lib\site-packages\scipy\stats\_axis_nan_policy.py:586:
UserWarning: scipy.stats.shapiro: For N > 5000, computed p-value may not be acc
urate. Current N is 5875.
  res = hypotest_fun_out(*samples, **kwds)
```

In [38]:
```
import matplotlib.pyplot as plt
import seaborn as sns

fitted_vals = result_refined.fittedvalues
residuals = result_refined.resid

plt.figure(figsize=(8, 6))
sns.scatterplot(x=fitted_vals, y=residuals, alpha=0.3, color='blue')
plt.axhline(0, color='red', linestyle='--', linewidth=1)
plt.title('Residuals vs. Fitted Values')
plt.xlabel('Fitted Values (Predicted UPDRS)')
plt.ylabel('Residuals')
plt.grid(True, alpha=0.3)
plt.show()
```


Residuals vs. Fitted Values
Residuals
Fitted Values (Predicted UPDRS)
10.0
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
−10.0
10
20
30
40
50