Linear regression - HW2

David Redek (MM1), Filip Kulla (SN)

December 2020

Introduction

A telemonitoring involves a remote tracking of certain patients—usually those who can not be present at the same location as their health care provider. Each patient has a number of monitoring devices which record measurements regarding the patient's health conditions. The recordings are captured, stored, and transmitted automatically.

In this homework assignment, we are interested in some subset of a telemetry data on patients with an early stage Parkinson's disease. The patients were recruited to a trial for remote symptom progression monitoring. Each patient's record consists of several biomedical measurements of the patient's voice and the voice analysis is used to determine the progress of the disease—measured by two unified Parkinson's disease rating scores (UPDRS). Moreover, for each observation there are some additional, patient specific data provided, such as the patient's gender and age.

Our primary interest is to infer whether the UPDRS scores can be (somehow) predicted from the voice recordings and the patient's specific data. In particular, we are primarily interested in the relationship between the expected ratio of the two UPDRS scores and the recorded noise-to-harmonics ratio of the patient's voice (NHR).

Our dataset contains 850 observations and 7 covariates:

- age - patient's age;
- sex - two level factor variable: 0 – male; 1 – female;
- motor_UPDRS - patient's motor UPDRS score;
- total_UPDRS - patient's total UPDRS score;
- Shimmer - variation measure for the amplitude of the patient's voice;
- NHR - noise-to-harmonics ratio of the patient's voice;
- fDFA - four-level factor variable for the signal fractal scaling exponent; (1 for low scaling – 4 for high scaling)

We will denote proportion of motor_UPDRS and total_UPDRS by proportion_UPDRS (proportion_UPDRS = motor_UPDRS / total_UPDRS). In addition, lNHR will denote logarithm of NHR (lNHR = log(NHR)).

Throughout the whole document significance level of $\alpha = 0.05$ will be considered.

Model building

Part A

We will start with a simple linear regression model where the expected proportion of the two UPDRS scores is modeled with respect to the noise-to-harmonics ratio (NHR). We will consider an optional logarithmic transformation for both

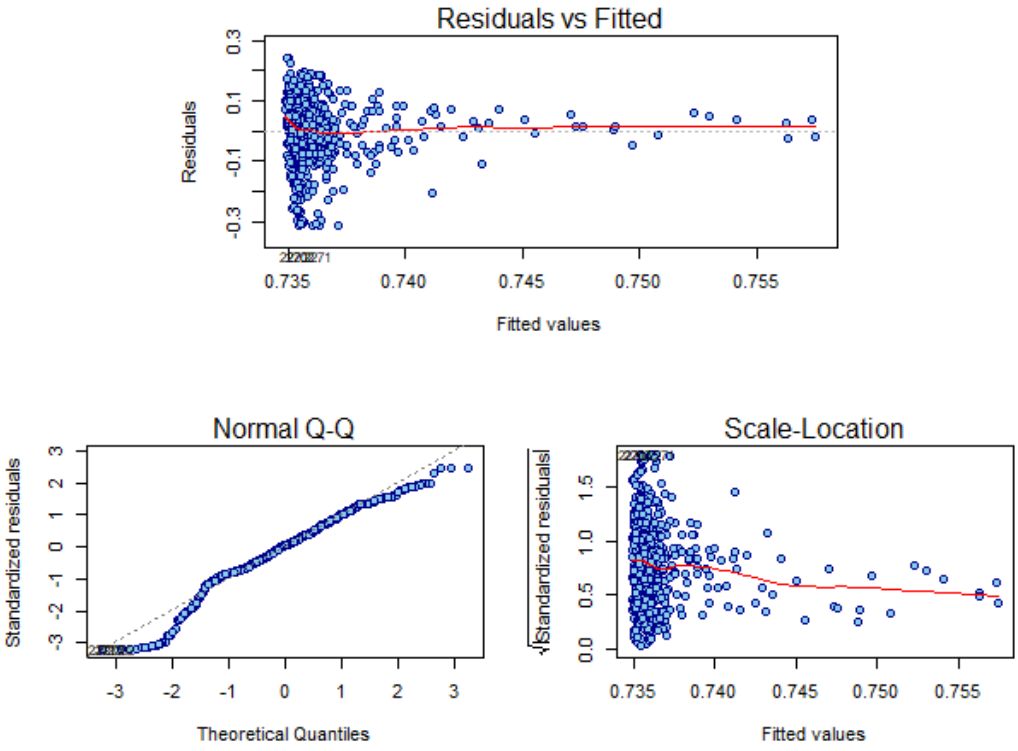

Figure 1: Diagnostic plots for model (1).

the response variable and also for the noise-to-harmonic ratio (NHR), which leads to 4 simple linear regression models:

$$\text{proportion\_UPDRS} \sim \text{NHR}, \tag{1}$$

$$\log(\text{proportion\_UPDRS}) \sim \text{NHR}, \tag{2}$$

$$\text{proportion\_UPDRS} \sim \text{lNHR}, \tag{3}$$

$$\log(\text{proportion\_UPDRS}) \sim \text{lNHR}. \tag{4}$$

Out of these models we will choose just one, basing our decision primarily on basic diagnostic plots but also on R^2.

In Figure 1 and Figure 2 we can see that both models which use NHR as an explanatory variable seem to have problems with homoscedasticity, which can be seen on both Residuals vs Fitted as well as Scale-Location graphs. Furthermore, regarding correctness of the regression function, it might be a little problematic that in Residuals vs Fitted graphs, lowess smoother increases as region where most fitted values lie is approached.

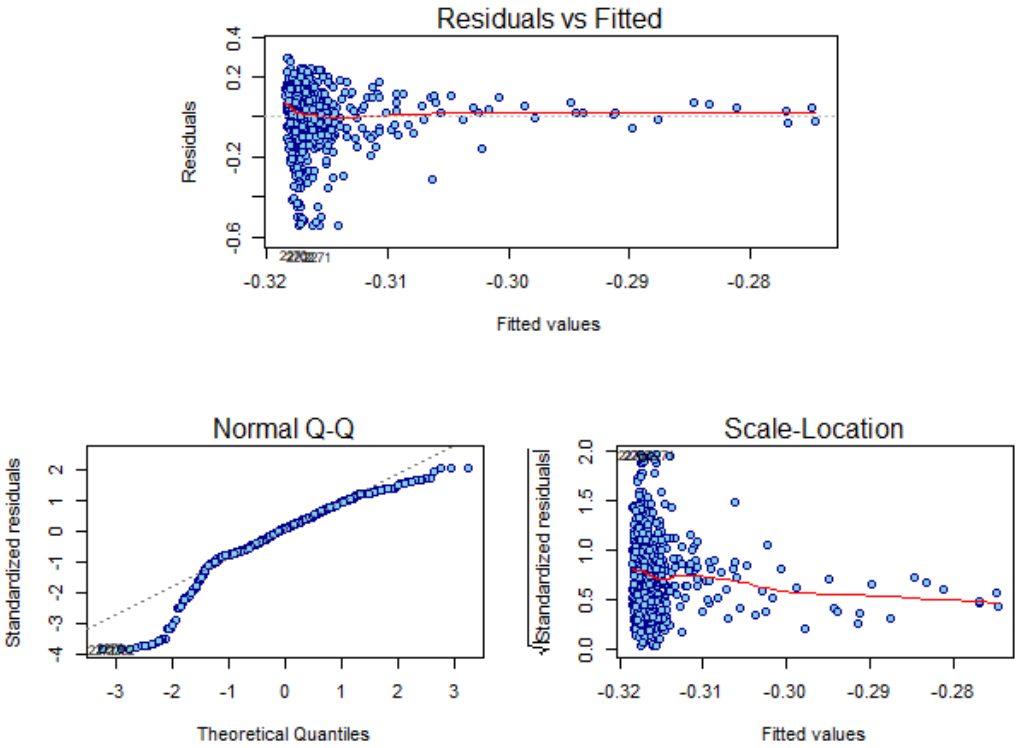

Figure 2: Diagnostic plots for model (2).

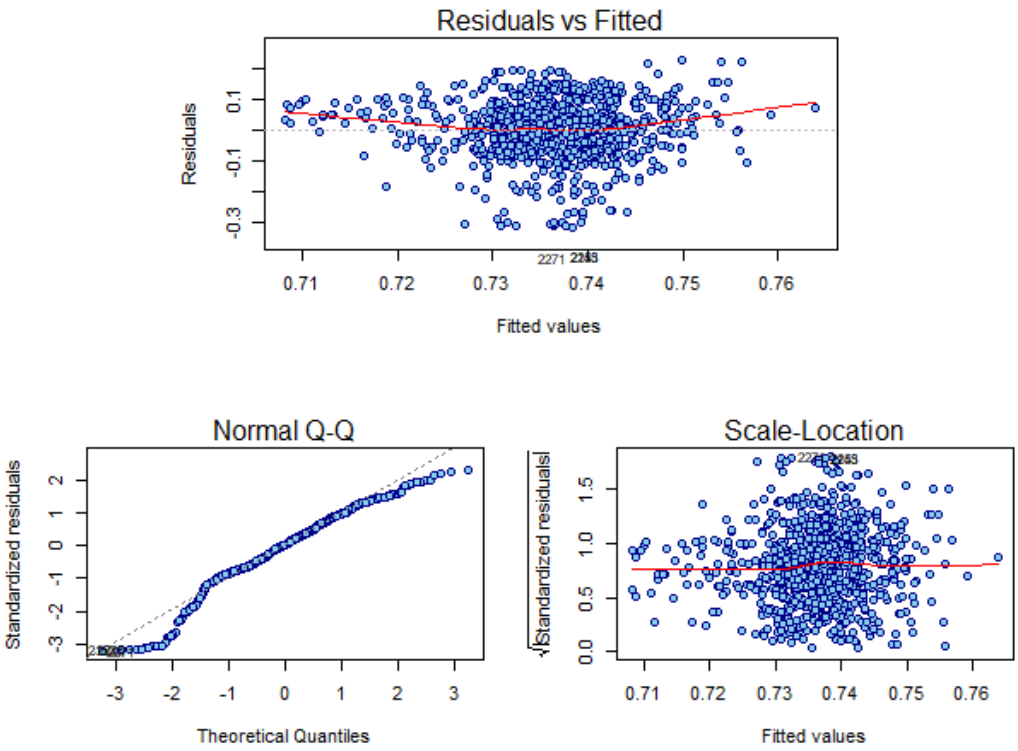

Figure 3: Diagnostic plots for model (3).

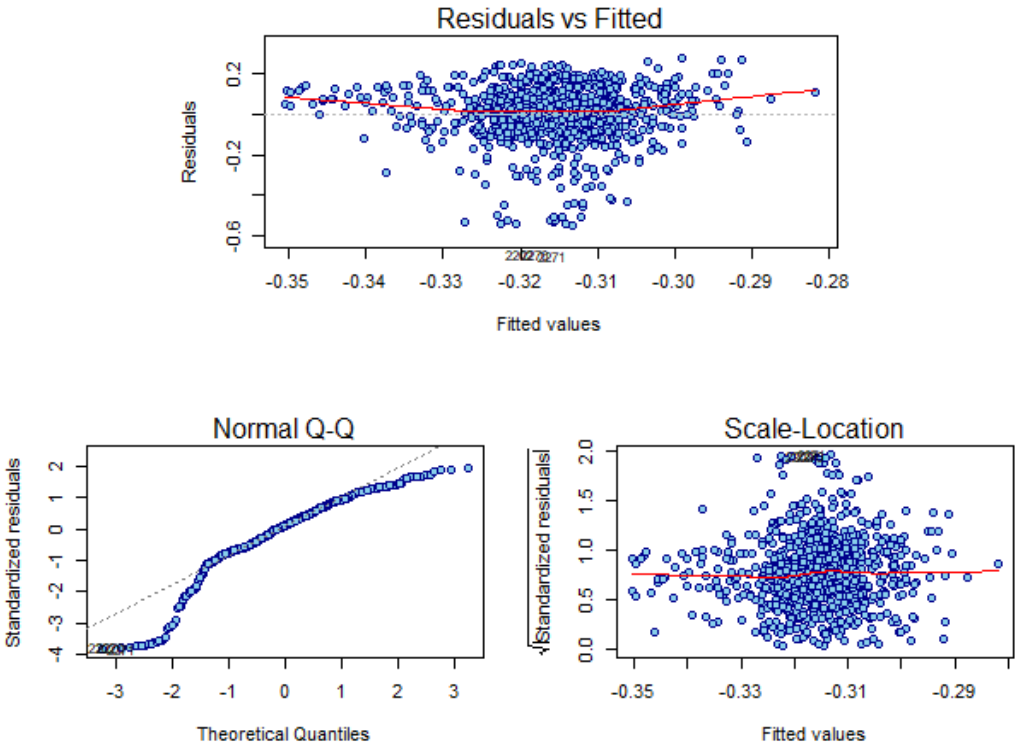

Figure 4: Diagnostic plots for model (4).

When we try to model the expected proportion of the two UPDRS scores by logarithm of NHR (lNHR), both Residuals vs Fitted as well as Scale-Location graphs seem to improve (Figure 3 and Figure 4). This improvement is visible especially for the Scale-Location graphs. Moreover, Residuals vs Fitted graphs look very good especially in regions where most fitted values are located, some minor deviations from ideal (increase of lowess smoother) are visible for observations with small as well as high fitted values, of which there are few.

From the Q-Q graphs it seems that the normality assumption seems to be violated in all 4 models, although Q-Q graphs for model (1) and model (3) look a little nicer than Q-Q graphs for model (2) and model (4).

From what has been observed so far, it seems that if we want to choose one model out of 4 proposed, especially due to Residuals vs Fitted and Scale-Location graphs, either model (3) or model (4) should be chosen. Furthermore, to decide whether to opt for logarithmic transformation of response or not, instead of looking only at diagnostic plots, which seem to be very similar, we can look at multiple R^2. In the model (3) without transformed response $R^2 = 0.006$ while in the model (4) with the logarithmic transformation $R^2 = 0.004$. Moreover, $R^2 = 0.006$ is highest from all 4 considered models. Taking it into consideration we choose model (3) as our basic model.

The fact that R^2 is so small is not surprising, since already when we look at the scatterplot of our data (Figure 5) we can see that the response and the explanatory variable are not highly correlated.

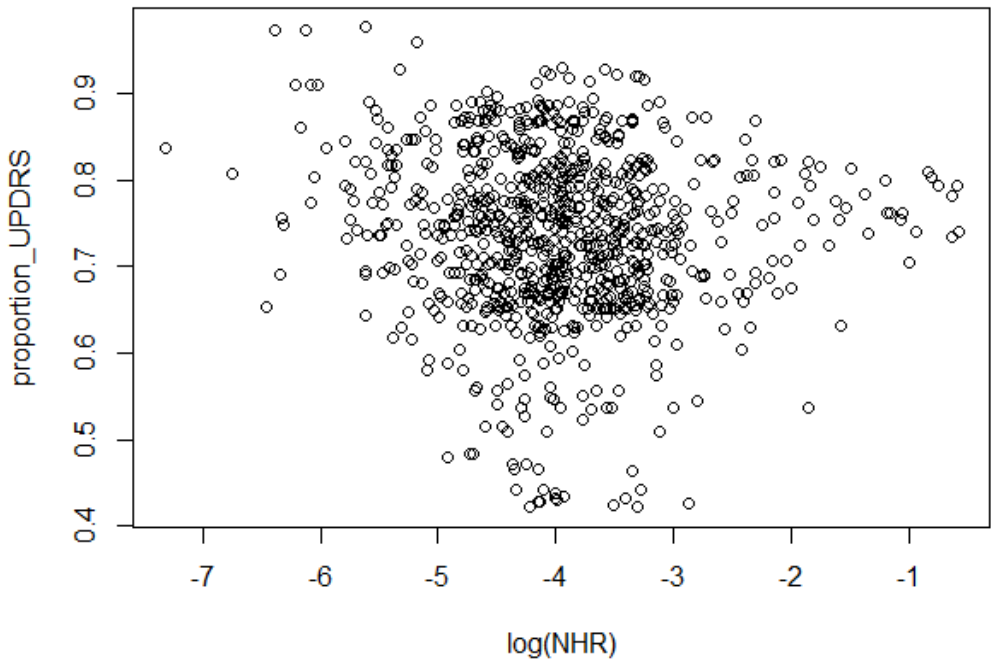

Figure 5: Scatterplot of proportion_UPDRS $\sim$ log(NHR).

Part B

Further, we will try to extend our model (3) so that it can be used to answer whether

(a) the effect of the noise-to-harmonics ratio (NHR) depends on the signal scaling component (fDFA),

(b) the effects of the patient's age and the amplitude variation (Shimmer) generally differ for male and female patients,

(c) the signal scaling component (fDFA) is a significant modifier of the effect of the patient's age.

Due to (a) our model will have to include interaction of the noise-to-harmonics ratio with the signal scaling component, due to (b) it will have to contain interaction of the amplitude variation with gender as well as interaction of age with gender and due to (c) it will have to contain interaction of age with the signal scaling component.

Let us consider a linear model in which we explain the expected proportion of the two UPDRS scores by all available covariates and all possible two-way interactions:

$$\text{proportion\_UPDRS} \sim (\text{lNHR} + \text{age} + \text{sex} + \text{Shimmer} + \text{fDFA})^2. \qquad (5)$$

To simplify model (5) we would like to perform a series of submodel F-tests to assess which interaction terms are not significant and thus can be excluded

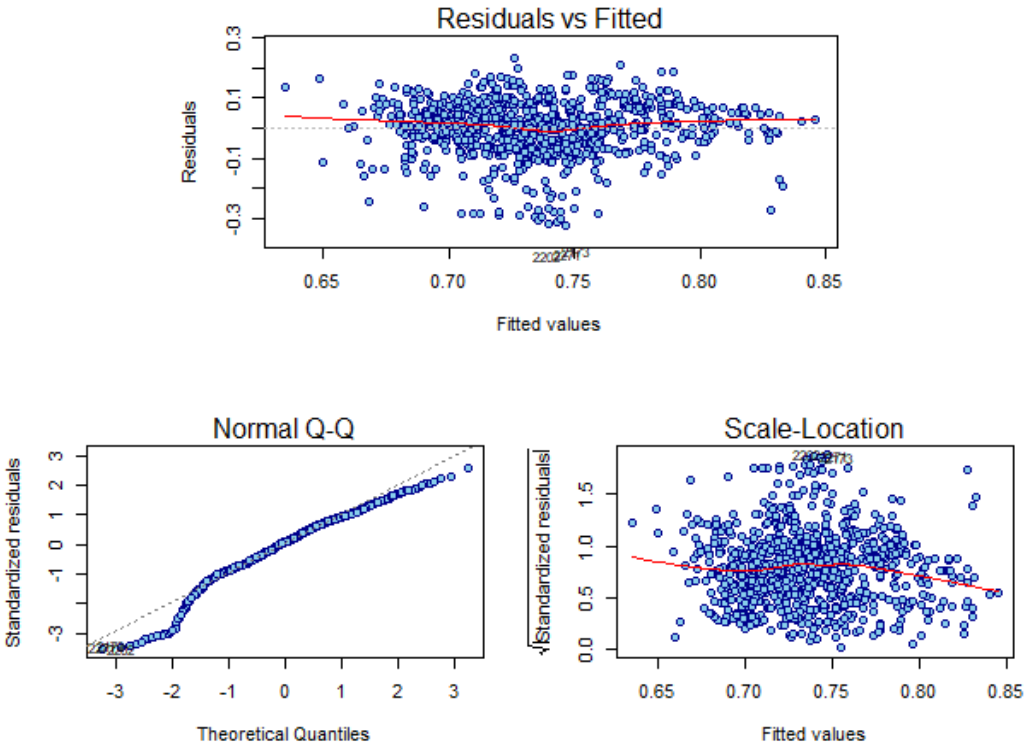

Figure 6: Diagnostic plots for model (5).

(keeping interaction terms lNHR:fDFA, sex:Shimmer, age:sex and age:fDFA in the model irrespective of their significance). Our problem here, however, is that the assumption of normality of model (5) seems to be violated as seen from the Q-Q graph in Figure 6. Fortunately, the number of observations is 850, which is sufficiently large to rely on asymptotics. Full justification of using submodel F-tests asymptotically will be provided when answering questions (a), (b) and (c).

From the Anova type II table for model (5) (Table 1) we can see that the interaction terms which are significant are only lNHR:fDFA, age:sex and age:fDFA, which we want to include in our model anyway. To make sure that we can indeed simplify model (5) to model (6):

$$\begin{aligned} \text{proportion\_UPDRS} \sim\ & \text{lNHR} + \text{age} + \text{sex} + \text{Shimmer} + \text{fDFA} \\ & + \text{lNHR:fDFA} + \text{sex:Shimmer} + \text{age:sex} + \text{age:fDFA}, \end{aligned} \tag{6}$$

excluding all unwanted interaction terms at the same time, we will perform another submodel F-test on models (6) and (5). Test statistic is equal to 0.891 and the corresponding p-value is equal to 0.5411, which means that we can indeed simplify model (5) to model (6), in which the proportion of UPDRS scores is explained by lNHR, age, sex, Shimmer, fDFA and two-way interaction terms lNHR:fDFA, sex:Shimmer, age:sex and age:fDFA. (The same model could have been obtained by excluding insignificant interaction terms sequentially, one by one.) Now we will proceed to questions (a), (b) and (c).

All of the questions (a), (b) and (c) can be answered using the Anova type II (or type III) table for model (6) (Table 2). Particularly, using rows corresponding

	Sum Sq	Df	F value	Pr(>F)
lNHR	0.100	1	11.700	0.001
age	0.037	1	4.299	0.038
sex	0.074	1	8.677	0.003
Shimmer	0.059	1	6.908	0.009
fDFA	0.121	3	4.704	0.003
lNHR:age	0.001	1	0.093	0.760
lNHR:sex	0.012	1	1.395	0.238
lNHR:Shimmer	0.001	1	0.130	0.719
lNHR:fDFA	0.110	3	4.264	0.005
age:sex	0.099	1	11.541	0.001
age:Shimmer	0.018	1	2.141	0.144
age:fDFA	0.358	3	13.894	0.000
sex:Shimmer	0.0001	1	0.009	0.924
sex:fDFA	0.027	3	1.064	0.364
Shimmer:fDFA	0.021	3	0.812	0.487
Residuals	7.072	824		

Table 1: Anova type II table for model (5)

	Sum Sq	Df	F value	Pr(>F)
lNHR	0.103	1	11.982	0.001
age	0.028	1	3.247	0.072
sex	0.087	1	10.112	0.002
Shimmer	0.089	1	10.410	0.001
fDFA	0.123	3	4.788	0.003
lNHR:fDFA	0.095	3	3.699	0.012
sex:Shimmer	0.010	1	1.143	0.285
age:sex	0.128	1	14.946	0.000
age:fDFA	0.369	3	14.358	0.000
Residuals	7.148	834		

Table 2: Anova type II table for model (6)

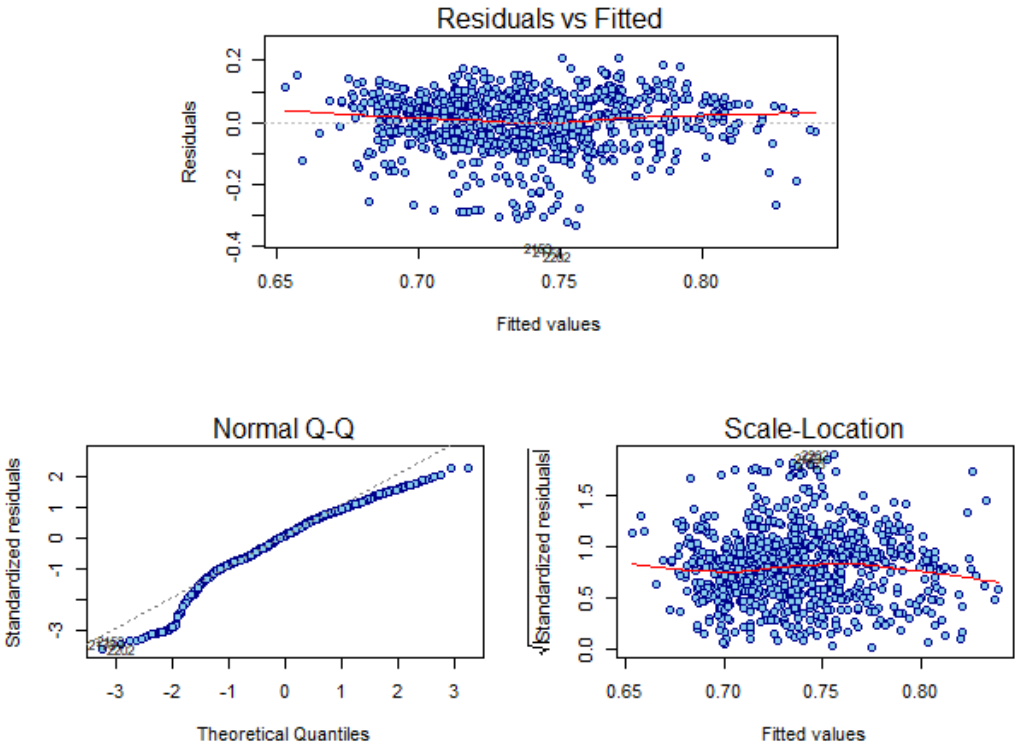

Figure 7: Diagnostic plots for model (6).

to the interaction terms lNHR:fDFA, sex:Shimmer, age:sex and age:fDFA. On each of these lines it is tested whether a given interaction term is significant in model (6) or not. As it was already stated we cannot believe in normality and we therefore have to rely on asymptotics. Let us list all necessary theoretical assumptions.

For our asymptotics we assume that our data (Y_i, X_i) are i.i.d. from distribution given by a generic random vector (Y, X). Further we assume that $\mathsf{E}[Y|X] = X^\mathsf{T}\beta$, $\mathsf{E}[XX^\mathsf{T}]$ is a positive definite matrix, $\mathsf{E}|X_iX_j| < \infty$, $\mathrm{var}(Y|X) = \sigma^2$, $\mathsf{E}|\epsilon^2 X_iX_j| < \infty$, where $\epsilon = Y - X^\mathsf{T}\beta$. We believe that in our situation all of these assumptions are satisfied because covariates in our data are bounded and it seems that we can also believe in homoscedasticity (Figure 7). Therefore, we can use submodel F-tests and Anova type II table for model (6) as we normally would under normality, only taking into consideration that now we are performing asymptotic tests.

Firstly, to answer whether the effect of the noise-to-harmonics ratio (NHR) depends on the signal scaling component (fDFA) we can use submodel F-test with bigger model (6) and its submodel M_0 which is obtained from model (6) when interaction of logarithm of NHR with fDFA (lNHR:fDFA) is excluded.

$$\begin{aligned}(M_0 : \text{proportion\_UPDRS} \sim \text{lNHR} + \text{age} + \text{sex} + \text{Shimmer} + \text{fDFA} \\ + \text{sex:Shimmer} + \text{sex:age} + \text{age:fDFA})\end{aligned}$$

Null hypothesis: In model (6) interaction term lNHR:fDFA is significant.

Alternative hypothesis: In model (6) interaction term lNHR:fDFA is not significant.

Test statistic:

$$F = \frac{\frac{SS_e^0 - SS_e}{r - r_0}}{\frac{SS_e}{n-r}},$$

where SS_e^0 is residual sum of squares in model M_0, SS_e is residual sum of squares in model (6), r_0 and r are ranks of models M_0 and (6) and n is number of observations.

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow F \geq F_{r-r_0,n-r}(1-\alpha),$$

where $F_{r-r_0,n-r}(1-\alpha)$ is $(1-\alpha)$-quantile of $F_{r-r_0,n-r}$ distribution.

P-value (asymptotic): $1 - F_n(f)$, where f is observed value of F and F_n is cumulative distribution function of $F_{r-r_0,n-r}$ distribution.

For our data the observed value of the test statistic F is 3.699 and the corresponding p-value is 0.012. Thus, we can reject the null hypothesis and therefore we have proved that the effect of the noise-to-harmonics ratio (NHR) does depend on the signal scaling component (fDFA).

Next, we will address the question, whether the signal scaling component (fDFA) is a significant modifier of the effect of the patient's age. This is analogous to what has just been done, only, instead of excluding the interaction of logarithm of NHR with fDFA (lNHR:fDFA) to obtain submodel M_0, we exclude interaction of age with fDFA (age:fDFA). Observed value of F statistic is then equal to 14.358 and the corresponding p-value (rounded to 3 decimal places) is equal to 0.000. Again, we can reject the null hypothesis and we have thus proved that the signal scaling component (fDFA) is a significant modifier of the effect of the patient's age.

Finally, we want to answer whether the effects of the patient's age and the amplitude variation (Shimmer) are generally different for male and female patients. Again, we can proceed in the same way. Firstly, we define submodel M_0, by excluding the interaction term sex:Shimmer from model (6). Observed value of F statistic is then equal to 1.143 and the corresponding p-value is equal to 0.285. As the p-value is greater than the chosen significance level of 0.05, we cannot claim that the bigger model (6) is significantly better than its submodel M_0. Therefore we cannot claim that the effect of age is modified by the gender of the patient.

Secondly, we define submodel M_0, by excluding the interaction term age:sex from model (6). Observed value of F statistic is then equal to 14.946 and the corresponding p-value (rounded to 3 decimal places) is equal to 0.000. P-value is smaller than the chosen significance level of 0.05, thus we have proved that the effect of age depends on the gender of the patient.

If we want to perform both of these tests (testing significance of sex:Shimmer and age:sex) at the same time, however, then we face a multiple comparison problem. By Bonferroni correction, we can multiply already calculated p-values by 2 to obtain adjusted p-values. The first adjusted p-value is still greater than 0.05 and the second adjusted p-value is still smaller than 0.05 and therefore our conclusions remain valid. Further, we can say, that sex is an effect modifier of either age or Shimmer.

Part C

Furthermore, since the interaction term of sex:Shimmer is not significant (p-value is 0.2854) we can further simplify model (6) by excluding this interaction term, obtaining model:

$$\begin{aligned} \text{proportion\_UPDRS} \sim \text{lNHR} + \text{age} + \text{sex} + \text{Shimmer} + \text{fDFA} \\ + \text{lNHR:fDFA} + \text{age:sex} + \text{age:fDFA}. \end{aligned} \tag{7}$$

After this exclusion all other terms are statistically significant as seen in Table 3. Moreover, let us remark that, being primarily interested in modeling expected proportion_UPDRS by lNHR, we have also considered more general parametrizations of lNHR - either including also quadratic term $(\text{lNHR})^2$ or including both the quadratic term $(\text{lNHR})^2$ as well as the cubic term $(\text{lNHR})^3$ to model (7). Model with added quadratic as well as cubic term was not significantly better than model to which only quadratic term was added (p-value was 0.149). Model with added quadratic term was, however, significantly better than model (7) (p-value was 0.036). In spite of this we have decided to stick with model (7) because a) we have already performed several tests, inflating probability of type I error which might potentially lead to unnecessarily complicated model (p-value 0.036 also isn't extremely small) b) lNHR and $(\text{lNHR})^2$ are highly correlated (their Pearson correlation coefficient is -0.97), which causes multicollinearity problems (by adding $(\text{lNHR})^2$ term, VIF for lNHR increases from 4.56 to 54.94) c) interpretation of model (7) is much simpler d) inclusion of the quadratic term does not really improve diagnostic plots of model (7) (Figure 8).

	Sum Sq	Df	F value	Pr(>F)
lNHR	0.102	1	11.925	0.001
age	0.037	1	4.281	0.039
sex	0.087	1	10.110	0.002
Shimmer	0.089	1	10.408	0.001
fDFA	0.129	3	5.021	0.002
lNHR:fDFA	0.089	3	3.467	0.016
age:sex	0.150	1	17.523	0.000
age:fDFA	0.374	3	14.539	0.000
Residuals	7.158	835		

Table 3: Anova type II table for model (7)

Model (7) is basically our final model, only, for better interpretation, let us slightly modify it by shifting lNHR, age and Shimmer by their median values, introducing $\text{lNHR4} = \text{lNHR} + 4$, $\text{age65} = \text{age} - 65$ and $\text{Shimmer0.03} = \text{Shimmer} - 0.03$.

Our final model is then:

$$\begin{aligned} \text{proportion\_UPDRS} \sim \text{lNHR4} + \text{age65} + \text{sex} + \text{Shimmer0.03} + \text{fDFA} \\ + \text{lNHR4:fDFA} + \text{age65:sex} + \text{age65:fDFA}. \end{aligned} \tag{8}$$

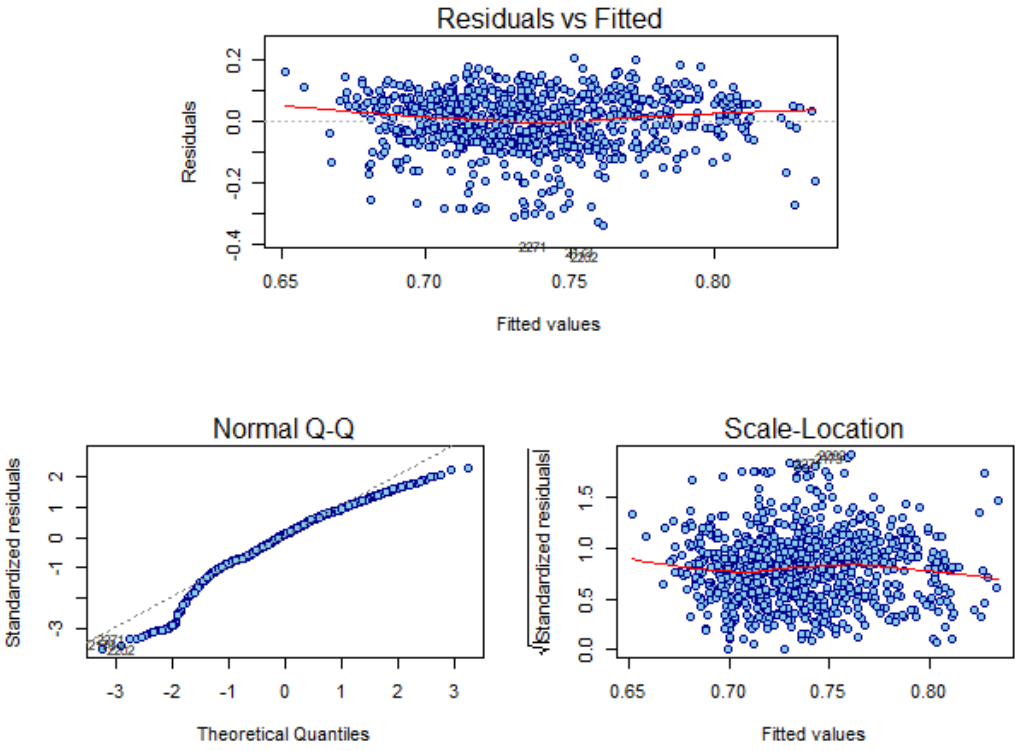

Figure 8: Diagnostic plots for model (7).

The most striking disadvantage of our final model is that it has very small $R^2 = 0.121$ and therefore it probably won't be exceptionally useful for prediction. On the other hand, nice thing about our model is that it is relatively simple and thus can be easily interpreted. What follows is interpretation of regression coefficients of our final model (8) (their estimates in Table 4).

- (Intercept) – represents expected proportion_UPDRS for a male patient who is 65 years old, his lNHR is -4 (NHR = e^{-4}), his Shimmer is 0.03 and his signal fractal scaling exponent is low (fDFA = 1).
- lNHR4 – represents effect of lNHR on expected proportion_UPDRS for patients with fDFA = 1. When lNHR increases by one (NHR is multiplied by e), expected proportion_UPDRS decreases by 0.014.
- age65 – represents effect of age on expected proportion_UPDRS for male patients with fDFA = 1. When age increases by one, expected proportion_UPDRS decreases by 0.005.
- sex1 – represents correction of the intercept for female patients. Expected proportion_UPDRS for a female patient who is 65 years old, her lNHR is -4, her Shimmer is 0.03 and her signal fractal scaling exponent is low (fDFA = 1) is $0.749 + 0.022$.
- Shimmer0.03 – represents effect of Shimmer on expected proportion_UPDRS. When Shimmer increases by one, expected proportion_UPDRS increases by 0.567.

- fDFA2 (fDFA3, fDFA4) – represents correction of the intercept when fDFA = 2. Expected proportion_UPDRS for a male patient who is 65 years old, his lNHR is -4, his Shimmer is 0.03 and his fDFA = 2 is $0.749 - 0.029$.
- lNHR4:fDFA2 (lNHR4:fDFA3, lNHR4:fDFA4) – is the modifier of the effect of lNHR. When lNHR increases by one, expected proportion_UPDRS for a patient with fDFA = 2 decreases by $(0.014 + 0.007)$.
- age65:sex1 – is the modifier of the effect of age for females. When age of a female patient with fDFA = 1 increases by 1, expected proportion_UPDRS increases by $-0.005 + 0.003$, thus decreases by 0.002.
- age65:fDFA2 (age65:fDFA3, age65:fDFA4) – is the modifier of the effect of age for fDFA. When age of a male patient with fDFA = 2 increases by 1, expected proportion_UPDRS increases by $-0.005 + 0.003$, thus decreases by 0.002.

	Estimate
(Intercept)	0.749
lNHR4	−0.014
age65	−0.005
sex1	0.022
Shimmer0.03	0.567
fDFA2	−0.029
fDFA3	−0.034
fDFA4	−0.014
lNHR4:fDFA2	−0.007
lNHR4:fDFA3	−0.017
lNHR4:fDFA4	0.020
age65:sex1	0.003
age65:fDFA2	0.003
age65:fDFA3	0.003
age65:fDFA4	0.008

Table 4: Estimates of regression coefficients from summary table of model (8)

Model assumptions

In the previous section we have decided to model expected proportion_UPDRS by model (8). In this section we will discuss the model in terms of the imposed assumptions.

Let us look at basic diagnostic plots for model (8) (Figure 9). (These are exactly the same as the ones for model (7) because the only difference between these two models is that we shifted some covariates by their median values, which naturally did not change the original regression space.)

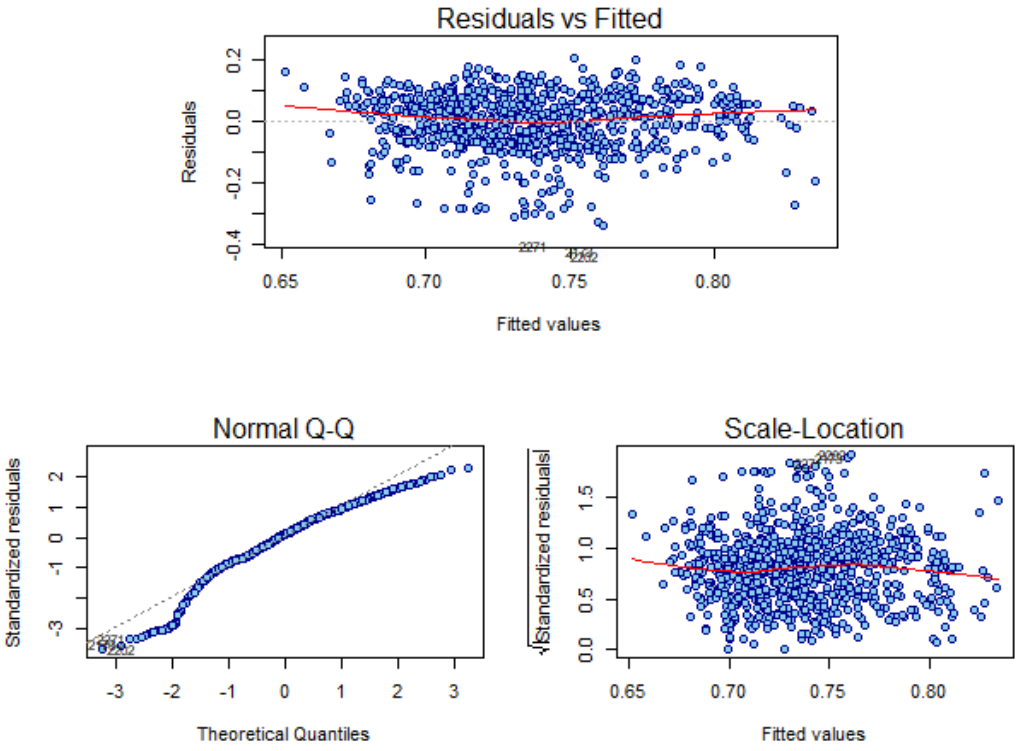

Figure 9: Diagnostic plots for model (8).

In the Residuals vs Fitted graph we can see that residuals are quite evenly scattered around horizontal line passing through 0, indicating that our regression function is correct. To check this more thoroughly we can look at plots of partial residuals (Figure 10, Figure 11 and Figure 12). The idea behind these plots is that in a linear model (partial residuals $\sim$ regressor), the estimated slope coincides with the estimated coefficient for the regressor in the original model. From these plots it seems that our covariates are included in the model in the right way. Plots for age65 and Shimmer0.03 are especially nice. Plot for lNHR4 also seems to be fine, although there is visible some slight deviation in the left part of the figure. This deviation, however, is caused by only a few datapoints. In conclusion, we believe that the assumption of correct regression function is satisfied.

Part A

Moving to homoscedasticity, in the Scale-Location graph (Figure 9) we can see that square roots from absolute values of standardized residuals are quite evenly scattered around horizontal line passing through 1. To check the assumption of homoscedasticity formally, we will use Breusch-Pagan test. Moreover, because from the Q-Q graph it seems that we cannot believe in normality of our data, we will use Koenker's studentized version of the Breusch-Pagan test.

Model: We consider 2 full rank models M_{homo} and M_{hetero}

$$M_{homo} : \mathbf{Y}|\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2 I_n),$$

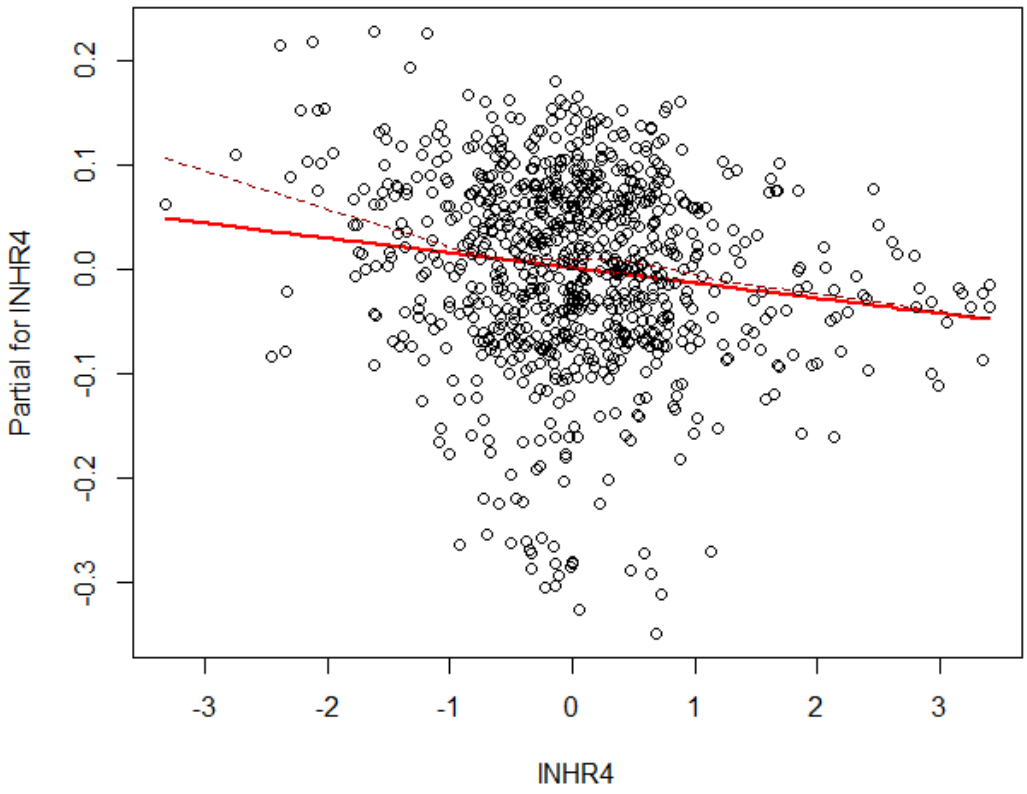

Figure 10: Partial residuals for lNHR4.

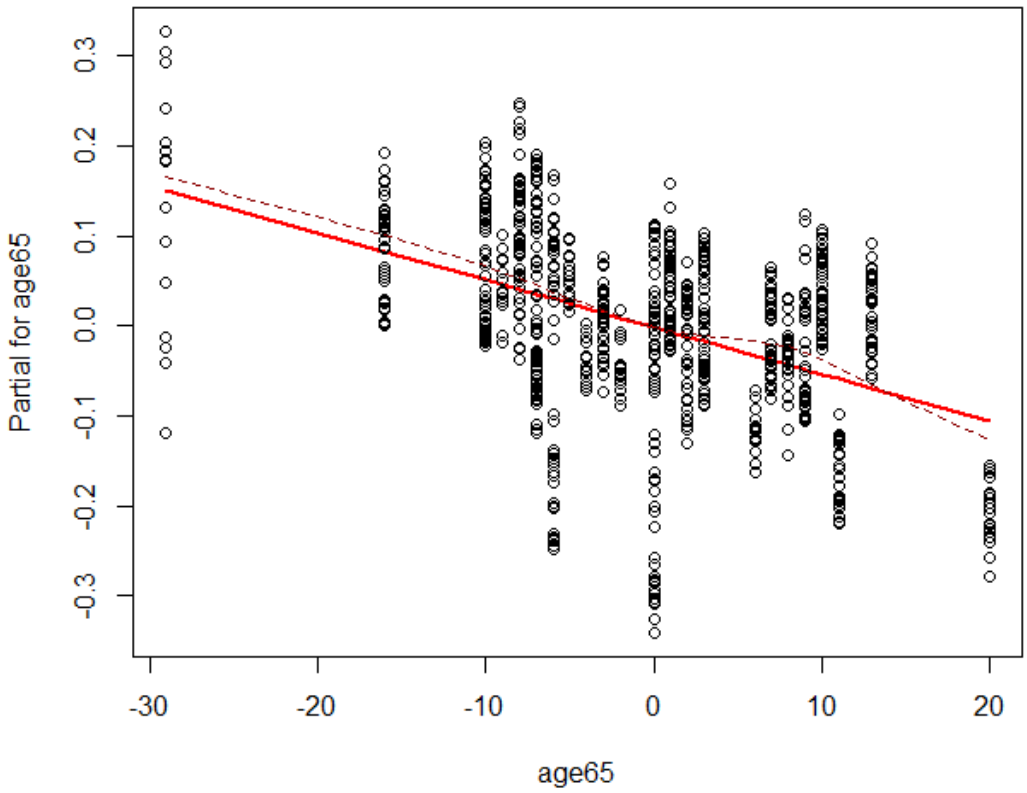

Figure 11: Partial residuals for age65.

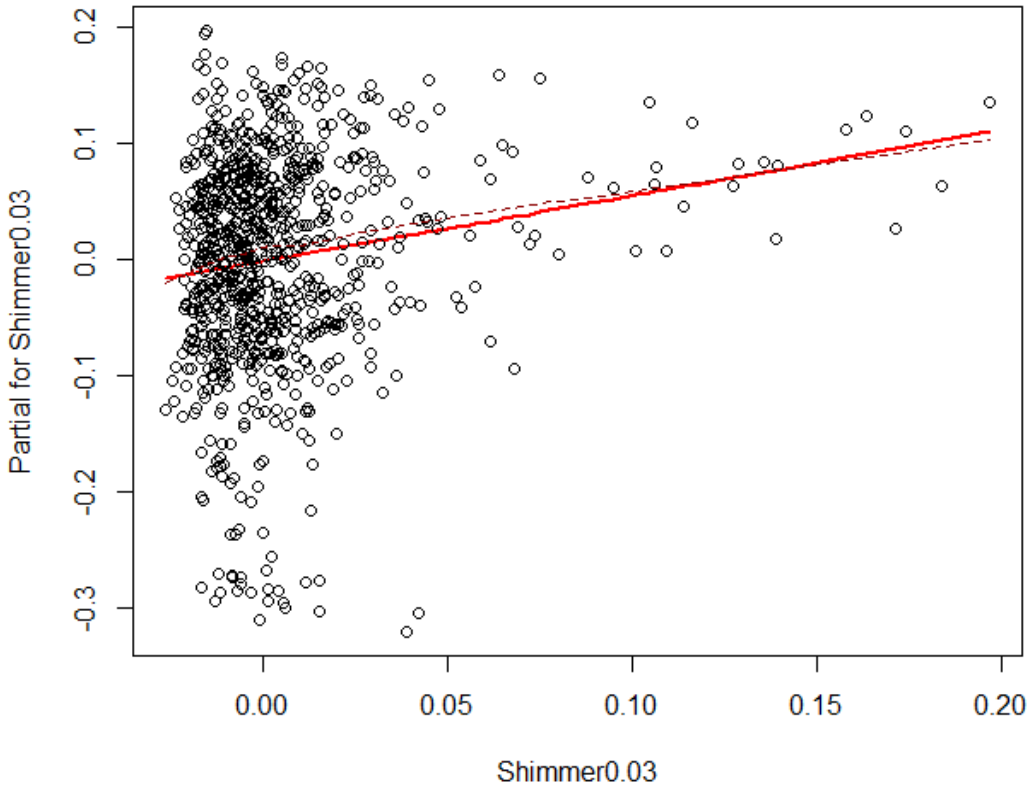

Figure 12: Partial residuals for Shimmer0.03.

$$M_{hetero} : \mathbf{Y}|\mathbb{X} \sim (\mathbb{X}\beta, \sigma^2\mathbb{W}^{-1}),$$

where $\mathbf{Y}$ denotes our response, $\mathbb{X}$ denotes model matrix, $\mathbb{W} = \text{diag}(w_1, \ldots, w_n)$, where $w_i^{-1} = e^{\lambda X_i^\mathsf{T}\beta}$, $i = 1, \ldots, n$, where X_i^T is i-th row of our model matrix. In this setting, test whether $\lambda = 0$ corresponds to test of homoscedasticity.

Null hypothesis: $\lambda = 0$.

Alternative hypothesis: $\lambda \neq 0$.

Test statistic:

$$T = n\Big(cor(U^2, \widehat{Y})\Big)^2,$$

where $cor(U^2, \widehat{Y})$ denotes sample correlation of squared residuals and fitted values. Under the null hypothesis T has asymptotically χ_1^2 distribution.

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow T \geq \chi_1^2(1-\alpha)$$

where $\chi_1^2(1-\alpha)$ is $(1-\alpha)$-quantile of χ_1^2 distribution.

P-value (asymptotic): $1 - F_n(t)$, where t is observed value of T and F_n is cumulative distribution function of χ_1^2 distribution.

For our data, the observed value of the test statistic T is 0.3104, and the corresponding p-value is 0.5774. The p–value is much higher than prescribed significance level of 0.05, which means that we cannot reject the null hypothesis and therefore it seems that it is reasonable to believe in homoscedasticity of our data.

One problem with this procedure is that our test is sensitive when variance increases/decreases with conditional expectation of our response (that is how we specified w_i). So we have chosen just one possible dependence, while it is possible to come up with many other dependencies leading to different $\mathbb{W}$ matrices. Because it is not realistic to check all of them we will believe in the test that we have performed and in homoscedasticity of our data.

Part B

Further, we want to assess whether we can assume normal linear model. From the Q-Q graph in Figure 9 it seems that this is not the case. We will check this formally using Shapiro-Wilk test. The assumption of normal linear model ($\mathbf{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2 I_n)$) implies normality of raw residuals $\mathbf{U}$ ($\mathbf{U}|\mathbb{X} \sim N_n(\mathbf{0}, \sigma^2\mathbb{M})$) which, however, are not i.i.d. since matrix $\mathbb{M}$ is neither diagonal nor has it constant diagonal. Nevertheless, it is recommended to check the assumption of normal linear model by checking normality of raw residuals, keeping in mind limitations of such approximate approach.

Null hypothesis: Raw residuals $\mathbf{U}$ are normally distributed.

Alternative hypothesis: Raw residuals $\mathbf{U}$ are not normally distributed.

Test statistic:

$$W = \frac{\left(\sum_{i=1}^{n} a_i U_{(i)}\right)^2}{\sum_{i=1}^{n} \left(U_i - \overline{U_n}\right)^2},$$

where $\overline{U_n}$ denotes sample mean of $\mathbf{U}$, $U_{(i)}$ is the i-th order statistic and the coefficients a_i are given by

$$(a_1, \ldots, a_n) = \frac{\mathbf{m}^{\mathsf{T}}\mathbb{V}^{-1}}{(\mathbf{m}^{\mathsf{T}}\mathbb{V}^{-1}\mathbb{V}^{-1}\mathbf{m})^{1/2}},$$

where vector $\mathbf{m} = (m_1, \ldots, m_n)^{\mathsf{T}}$ is made of the expected values of the order statistics of independent and identically distributed random variables sampled from the standard normal distribution and $\mathbb{V}$ is the covariance matrix of those normal order statistics.

There is no name for the distribution of W. The cutoff values for the statistic are calculated through Monte-Carlo simulations. Observed value of test statistic W and the corresponding p-value (calculated by R and rounded to 4 decimal places) are $W = 0.9590$, $p = 0.0000$. P–value is much smaller than 0.05, therefore we reject the null hypothesis in favour of the alternative, and thus we will not assume that the assumption of the normal linear model is satisfied.

Now we will comment on possibly outlying observations. We will use studentized residuals to test whether some observations are outliers of model (8). The test for observation number t is described.

Model: $\mathbf{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta + e_t\gamma_t^{out}, \sigma^2 I_n)$, where e_t is a vector whose t-th element is 1 and all other elements are 0. Let us remark that although this test was derived under normality, in our situation we do not assume normality, but can rely on asymptotics, as was already described, and thus we can use it asymptotically.

Null hypothesis: $\gamma_t^{out} = 0$ (t-th observation is not an outlier of model (8)).

Alternative hypothesis: $\gamma_t^{out} \neq 0$ (t-th observation is an outlier of model (8)).

Test statistic:

$$T_t = \frac{U_t}{\sqrt{MS_{e,(-t)}\ m_{t,t}}} \text{ (t-th studentized residual)},$$

where U_t is t-th element of vector of residuals, $m_{t,t}$ is an element of matrix $\mathbb{M} = I_n - \mathbb{X}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}\mathbb{X}^\mathsf{T}$ and $MS_{e,(-t)}$ is residual mean square in leave-one-out model (model obtained upon exclusion of observation number t).

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow |T_t| \geq t_{n-r-1}(1-\alpha/2),$$

where $t_{n-r-1}(1-\alpha/2)$ is $(1-\alpha/2)$-quantile of t_{n-r-1} distribution, where n denotes number of observations (850) and r denotes rank (15) of model (8).

P-value (asymptotic): $2(1 - F_n(|t|))$, where t is observed value of T_t and F_n is cumulative distribution function of t_{n-r-1} distribution.

We have to remember that we are performing 850 tests. Therefore we will use Bonferroni correction to deal with this multiple testing problem. In our data, observation with the smallest p-value satisfies that its observed value of T_t is -3.7137, and its adjusted p-value is 0.1852, which is greater than 0.05. Therefore it seems that in our data there are no outlying observations with respect to model (8).

In terms of leverage points, we know that observations with high leverages might be potentially dangerous because corresponding fitted values are forced to be close to the observed response values. In our data, however, all observations have relatively small leverages. The highest recorded leverage is 0.1352 and all other leverages are below 0.1.

Furthermore, considering Cook's distances for our observations, it holds that no observation from our dataset has Cook's distance higher than the rule of thumb cut-off value of 0.5-quantile of $F_{r,n-r}$ distribution.

In conclusion, we believe that in terms of possibly outlying/influential observations we do not have any major problems.

Part C

Finally, we will address problem of multicollinearity. We have already briefly mentioned it when we were deciding whether to include $(\text{lNHR})^2$ into our model or not. One of the reasons for not including this term was that its inclusion inflated VIFs quite a lot. In addition, shifting our numerical covariates by their median values also helped as seen in Table 5 and Table 6. In Table 6 we can see that there are no particularly large values among either GVIF's or GVIF's scaled by the corresponding degrees of freedom (last column of the table). We will, nevertheless, examine relations between our covariates.

We will start by considering correlations between numerical covariates in model (8) (Table 7). We can see that correlation between lNHR4 and Shimmer0.03 is 0.699, which is quite high. Let us also look at scatter plots for each pair of numerical covariates (Figure 13, Figure 14 and Figure 15). Because correlation between lNHR4 and Shimmer0.03 is so high, exclusion of Shimmer0.03 might further decrease VIFs, especially the one for lNHR4. Before trying it out, let us also consider categorical covariates. In Figure 16 we can see boxplots

	GVIF	Df	GVIF^(1/(2*Df))
lNHR	4.564	1	2.136
age	6.232	1	2.496
sex	53.354	1	7.304
Shimmer	2.134	1	1.461
fDFA	7.662×10^5	3	9.566
lNHR:fDFA	1.141×10^4	3	4.745
age:sex	52.800	1	7.266
age:fDFA	2.287×10^5	3	7.820

Table 5: Variance inflation factors in model (7)

	GVIF	Df	GVIF^(1/(2*Df))
lNHR4	4.564	1	2.136
age65	6.232	1	2.496
sex	1.073	1	1.036
Shimmer0.03	2.134	1	1.461
fDFA	1.394	3	1.057
lNHR4:fDFA	4.553	3	1.287
age65:sex	2.090	1	1.446
age65:fDFA	5.562	3	1.331

Table 6: Variance inflation factors in model (8)

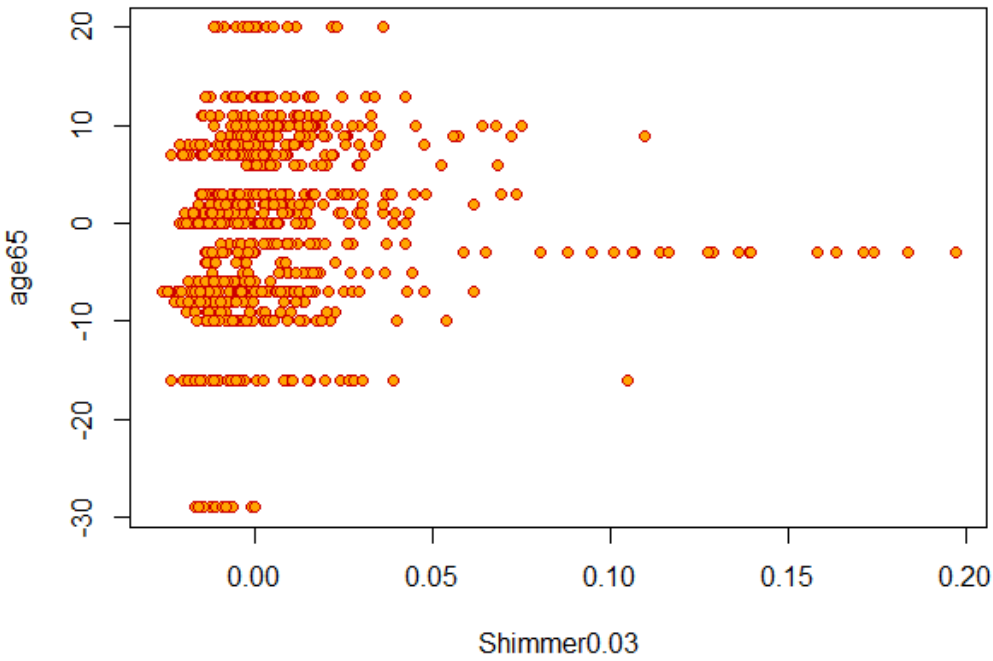

Figure 13: Scatter plot of age65 $\sim$ Shimmer0.03.

of all numerical covariates vs all categorical covariates. These boxplots suggest that there are no major problems with categorical covariates as we can see that there does not seem to be any linear relationship between any numerical and any categorical covariate.

	age65	Shimmer0.03	lNHR4
age65	1	0.105	0.117
Shimmer0.03	0.105	1	0.699
lNHR4	0.117	0.699	1

Table 7: Correlations of numerical covariates in model (8)

Now we will examine whether exclusion of some additional covariates improves the model. From our explanatory analysis it seems that exclusion of Shimmer0.03 might be the most sensible idea. After excluding Shimmer0.03 VIFs decrease (Table 8), however this decrease is not so large and by excluding Shimmer0.03 we obtain significantly worse model, since Shimmer0.03 is in model (8) significant (p-value is 0.0013). Since VIFs in model (8) are already relatively fine and exclusion of Shimmer0.03 (highly correlated with lNHR4) did not improve VIFs so much we suggest sticking to model (8). Let us remark, however, that we have tried excluding all other covariates, (age65, sex and fDFA) each with all higher order terms to keep our model hierarchically well formulated and VIFs were never as good as they were when only Shimmer0.03 was excluded. In addition, obtained

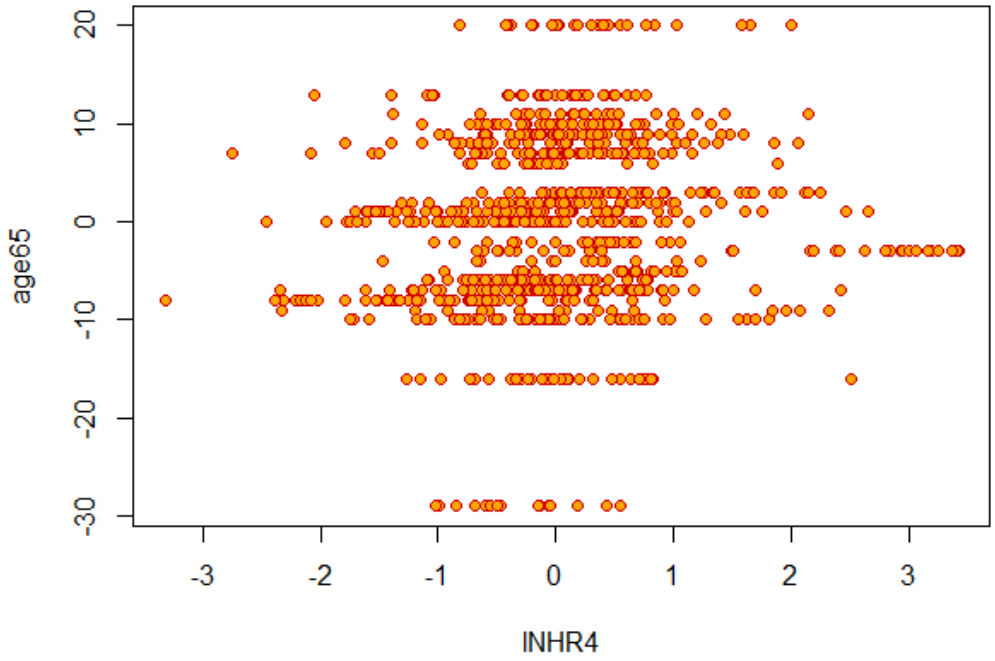

Figure 14: Scatter plot of age65 $\sim$ lNHR4.

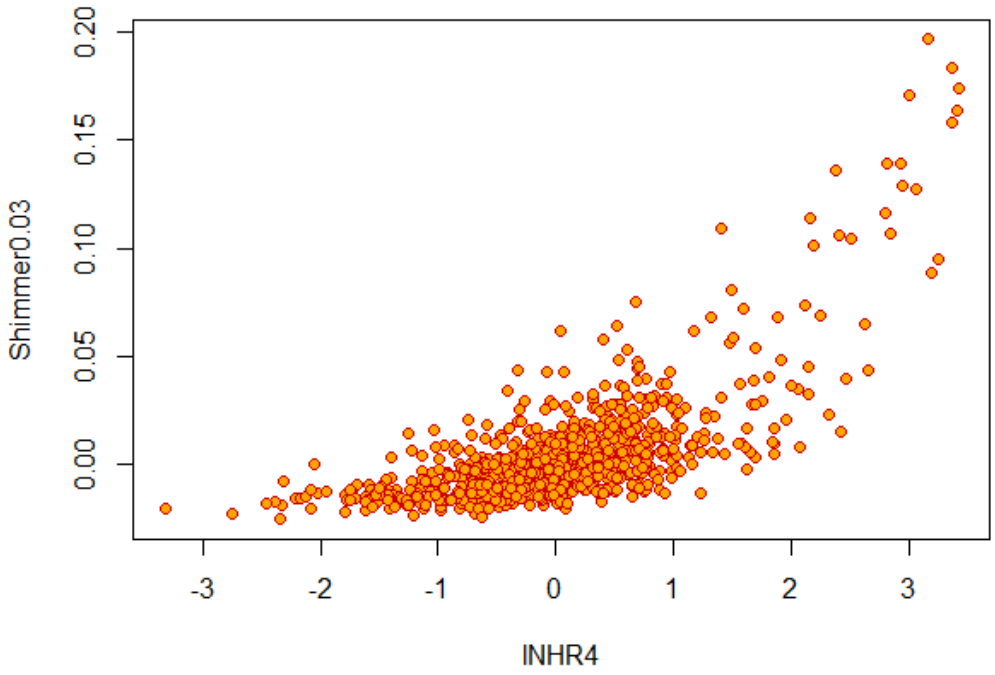

Figure 15: Scatter plot of Shimmer0.03 $\sim$ lNHR4.

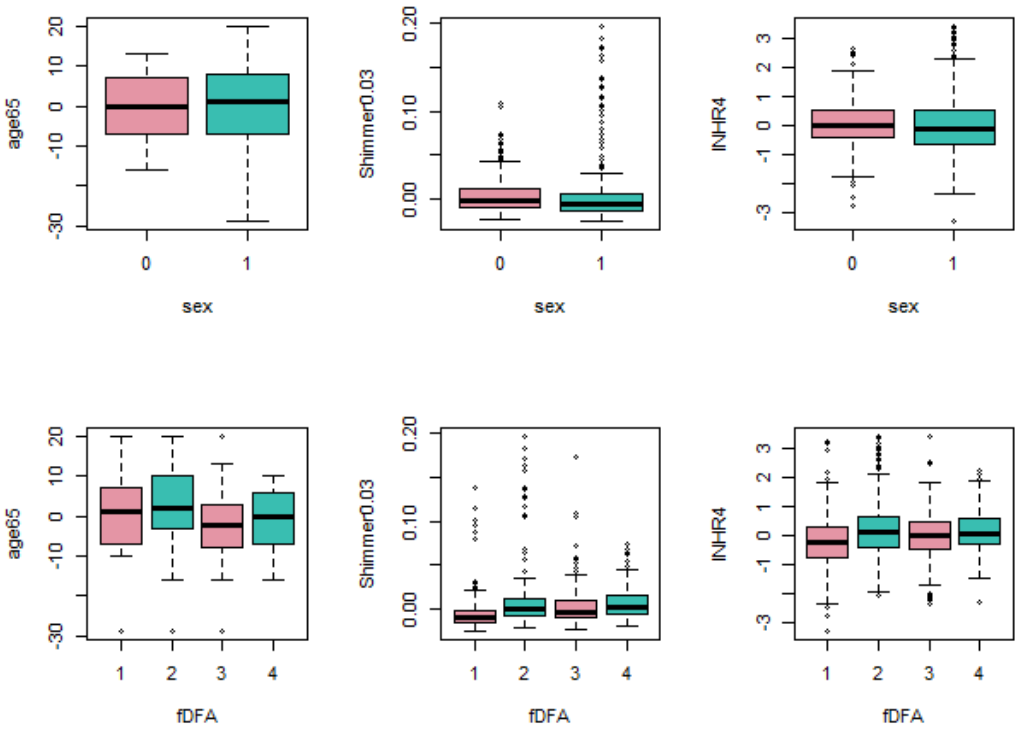

Figure 16: Boxplots of numerical vs categorical covariates.

models were already significantly worse not with p-value of 0.0013, but with p-values of 2.764×10^{-10}, 1.249×10^{-6} and 2.539×10^{-11}.

	GVIF	Df	GVIF^(1/(2*Df))
lNHR4	3.921	1	1.980
age65	6.209	1	2.492
sex	1.068	1	1.033
fDFA	1.371	3	1.054
lNHR4:fDFA	4.345	3	1.277
age65:sex	2.067	1	1.438
age65:fDFA	5.540	3	1.330

Table 8: VIFs when Shimmer0.03 is excluded from model (8)

Model inference

Part A

We are supposed to provide all pairwise comparisons between male and female patients for all groups defined by the categorical variables in our final model. Since the only categorical covariate in our model (apart from sex) is fDFA, we only need

to examine difference in expected proportion_UPDRS for males and females for each level of fDFA fixed. Let us denote regression coefficient corresponding to sex1 (indicator of a female) by β_3. (Table 4). Further, we will denote regression coefficient corresponding to age65:sex1 by β_{11}. Then, regardless of which level of fDFA is fixed, difference of expected proportion_UPDRS for female and male patients (keeping all covariate values except for sex same) is parametrized by $\beta_3 + \beta_{11}(\text{age} - 65)$, where age is age of patients we want to compare. If we choose age to be equal to 65, then our parameter of interest is directly β_3. We will test whether $\beta_3 = 0$, thus testing whether expected proportions_UPDRS for female and males patient who are 65 years of age are the same (in each level of fDFA fixed). Further we will construct a confidence interval for β_3.

As we have already discussed, we cannot rely on the assumption of the normal linear model. In spite of that we can rely on asymptotics. We will once again write down all necessary assumptions. We assume that our data (Y_i, X_i) are i.i.d. from distribution given by a generic random vector (Y, X). Further we assume that $\mathsf{E}[Y|X] = X^\mathsf{T}\beta$, $\mathsf{E}[XX^\mathsf{T}]$ is a positive definite matrix, $\mathsf{E}|X_iX_j| < \infty$, $\text{var}(Y|X) = \sigma^2$, $\mathsf{E}|\epsilon^2 X_i X_j| < \infty$, where $\epsilon = Y - X^\mathsf{T}\beta$. We believe that in our situation all of these assumptions are satisfied because covariates in our data are bounded and we also believe in homoscedasticity as was already justified. Now we can proceed to testing β_3, keeping in mind that our test is asymptotic.

Tested parameter: β_3.

Null hypothesis: $\beta_3 = 0$.

Alternative hypothesis: $\beta_3 \neq 0$.

Test statistic:

$$T = \frac{\widehat{\beta_3}}{\sqrt{MS_e \; v_{3,3}}},$$

where $v_{3,3}$ is diagonal element of matrix $(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}$, corresponding to β_3 and where $\mathbb{X}$ denotes the model matrix, MS_e denotes residual mean square and $\widehat{\beta_3}$ denotes LSE estimator of β_3 (everything under model (8).

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow |T| \geq t_{n-r}(1 - \alpha/2),$$

where $n = 850$ denotes number of observations, $r = 15$ rank of the model and $t_{n-r}(1 - \alpha/2)$ denotes $(1 - \alpha/2)$-quantile of t_{n-r} distribution.

P-value (asymptotic): $2(1 - F_n(|t|))$, where t is observed value of T and F_n is cumulative distribution function of t_{n-r} distribution.

For our data, the observed value of the test statistic T is 3.060 and the corresponding p-value is 0.002. P-value is smaller than 0.05 and thus we can reject the null hypothesis. We have proved that expected proportions_UPDRS for 65 years old females and 65 years old males are different (in each level of fDFA).

The corresponding confidence interval for β_3 is given by

$$\left(\widehat{\beta_3} \mp \sqrt{MS_e \; v_{3,3}} \; t_{n-r}(1 - \alpha/2)\right). \tag{9}$$

For our data the 95% confidence interval for β_3 is $(0.008, 0.036)$.

Part B

Now, we we will try to answer:

(a) Is the effect of the noise-to-harmonics ratio (NHR) different for male and female patients given high fractal scaling exponent (fDFA = 4)? What can be said, in general, regarding the patient's age playing the role of a modifier of the effect of NHR?

(b) Can we say, that the effect of NHR is the same for all patients with the scaling exponent fDFA = 2 and fDFA = 3?

(c) Given the model, what is the expected ratio of the UPDRS scores for a 65 years old male patient with a relatively standard noise-to-harmonic ratio (thus, NHR = 0.02) and a rather high scaling factor (fDFA = 4)? Use reasonable values for the the variables which are left unspecified. What is the corresponding prediction interval?

In question (a) we are we are interested in the interaction terms lNHR4:sex and age65:lNHR4. These interaction terms, however, are not included in our final model (8). For the sake of question (a), we will include them to our model. Firstly, we will focus on the first part of question (a) asking about the effect of the noise-to-harmonics ratio for male and female patients.

Let us include the interaction term lNHR4:sex to our model and upon this inclusion let us denote regression coefficients corresponding to lNHR4, lNHR4:fDFA4 and lNHR4:sex1 by β_1, β_{10} and β_{15}. Then the effect of lNHR4 on females with fDFA=4 (lNHR4 increases by 1 which means that NHR is multiplied by e) is parametrized as $\beta_1 + \beta_{10} + \beta_{15}$. For males it is parametrized as $\beta_1 + \beta_{10}$. Therefore, the question whether the effect of the noise-to-harmonics ratio is different for male and female patients with fDFA = 4 is reduced to testing whether $\beta_{15} = 0$. This can be tested by an asymptotic T-test as we have already described. One difference is that now rank of the tested model is 16. Observed value of the test statistic is 1.400 and the corresponding p-value is 0.162. The 95% confidence interval (given by (9)) for β_{15} is $(-0.004, 0.026)$. Since the p-value is higher than 0.05 we cannot reject that $\beta_{15} = 0$, thus we cannot reject that the effects of the noise-to-harmonics ratio for male and female patients with fDFA = 4 are the same.

Next, we will focus on the second part of question (a) asking about role of a patient's age as a modifier of the effect of NHR. To answer this question we will proceed exactly as we did when answering first part of question (a). We add the interaction term age65:lNHR4 to model (8). Upon this inclusion we can test whether the regression coefficient corresponding to this interaction term (let us denote it β_{16}) is equal to 0 or not. Again, we can test this using a T-test on a regression coefficient (again asymptotically). Observed value of the test statistic is -0.743 and the corresponding p-value is 0.458. The 95% confidence interval (given by (9)) for β_{16} is $(-0.001, 0.001)$. Since the p-value is higher than 0.05 we cannot reject that $\beta_{16} = 0$, thus we cannot reject that the age of a patient does not modify the effect of NHR.

Now we will proceed to question (b). If we denote regression coefficients in model (8) corresponding to terms lNHR4, lNHR4:fDFA2 and lNHR4:fDFA3

by β_1, β_8 and β_9 then the effect of lNHR4 for patients with fDFA=2 (lNHR4 increases by 1 which means that NHR is multiplied by e) is parametrized by $\beta_1 + \beta_8$. Similarly, the effect of lNHR4 for patients with fDFA=3 is parametrized by $\beta_1 + \beta_9$. To answer whether these 2 effects are the same we want to test whether $\beta_8 - \beta_9 = 0$. Thus we need to make inference on a linear combination of the regression coefficients from model (8). As we have already discussed, we cannot rely on normality of our model, but can rely on asymptotics, therefore we will again perform an asymptotic test.

We will denote $l = (0,0,0,0,0,0,0,0,1,-1,0,0,0,0,0)^\mathsf{T}$, $\theta = l^\mathsf{T}\beta = \beta_8 - \beta_9$, where β denotes vector of regression coefficients in model (8). As we already stated we want to test whether $\theta = 0$ or not.

Tested parameter: $\theta = \beta_8 - \beta_9$.

Null hypothesis: $\theta = 0$.

Alternative hypothesis: $\theta \neq 0$.

Test statistic:

$$T = \frac{\widehat{\theta}}{\sqrt{MS_e \; l^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}l}},$$

where $\mathbb{X}$ denotes the model matrix, MS_e denotes residual mean square and $\widehat{\theta}$ denotes LSE estimator of θ (everything under model (8).

Critical region:

$$H_0 \text{ is rejected} \Leftrightarrow |T| \geq t_{n-r}(1-\alpha/2),$$

where $n = 850$ denotes number of observations, $r = 15$ rank of the model and $t_{n-r}(1-\alpha/2)$ denotes $(1-\alpha/2)$-quantile of t_{n-r} distribution.

P-value (asymptotic): $2(1 - F_n(|t|))$, where t is observed value of T and F_n is cumulative distribution function of t_{n-r} distribution.

The corresponding confidence interval for θ is given by

$$\left(\widehat{\theta} \mp \sqrt{MS_e \; l^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}l} \; t_{n-r}(1-\alpha/2)\right).$$

Observed value of test statistic T for our data is 1.019 and the corresponding p-value is 0.308. Furthermore, the 95% confidence interval for parameter θ is $(-0.009, 0.029)$. P-value is greater than 0.05 and thus we cannot reject the null hypothesis that $\theta = 0$. Therefore, we cannot reject that the effect of NHR for patients with fDFA=2 is the same as the effect of NHR for patients with fDFA=3.

Finally, we can proceed to question (c). We are supposed to provide expected proportion_UPDRS for a male patient who is 65 years of age, whose NHR is 0.02 (thus his lNHR4 = $log(0.02) + 4$) and whose fDFA = 4. Further, we are supposed to calculate corresponding prediction interval. Since Shimmer was left unspecified we will consider a patient with relatively standard value of Shimmer (0.03), which is, approximately, median value of Shimmer in our dataset.

We will denote $x_{new} = (1,log(0.02)+4,0,0,0,0,0,1,0,0,log(0.02)+4,0,0,0,0)^\mathsf{T}$ a vector specifying the record of our patient. (Order corresponds to the order of model terms as shown in Table 4.) Then the expected proportion_UPDRS for our patient is $x_{new}^\mathsf{T}\widehat{\beta} = 0.736$, where $\widehat{\beta}$ denotes LSE estimator of vector of regression coefficients β. Now, we will calculate corresponding prediction interval.

Prediction interval was derived under the assumption of a full rank normal linear model: $\mathbf{Y}|\mathbb{X} \sim N_n(\mathbb{X}\beta, \sigma^2 I_n)$. Under that assumption, prediction interval for proportion_UPDRS (Y_{new}) of our patient is given by

$$\left(x_{new}^{\mathsf{T}}\widehat{\beta} \mp \sqrt{MS_e\ (1 + x_{new}^{\mathsf{T}}(\mathbb{X}^{\mathsf{T}}\mathbb{X})^{-1}x_{new})}\ t_{n-r}(1-\alpha/2)\right),$$

where MS_e and $\mathbb{X}$ denote residual mean square and model matrix of model (8) ($n = 850$ and $r = 15$). (Let us remark that in our theoretical derivations $Y_{new} = x_{new}^{\mathsf{T}}\beta + \epsilon_{new}$, where $\epsilon_{new} \sim N(0, \sigma^2)$ is independent with $\epsilon = \mathbf{Y} - \mathbb{X}\beta$.) For our data, prediction interval for Y_{new} is (0.554, 0.918). One problem with this procedure is that the prediction interval was derived under the assumption of a normal linear model. In our situation we do not believe that this assumption is satisfied and thus we have to view calculated prediction interval only as an approximation.

Part C

Now, we will reparametrize model (8) using the contrast sum parametrization for both sex as well as fDFA (contrast matrices in Table 9 and Table 10). We are interested in interpretation of regression coefficients of this reparametrized model (their estimates in Table 11).

	sex1
male	1
female	−1

Table 9: Contrast matrix for sex.

	fDFA1	fDFA2	fDFA3
fDFA = 1	1	0	0
fDFA = 2	0	1	0
fDFA = 3	0	0	1
fDFA = 4	−1	−1	−1

Table 10: Contrast matrix for fDFA.

- (Intercept) – represents mean of expected proportions_UPDRS in all groups defined by categorical covariates sex and fDFA, when we consider patients who are 65 years of age, their lNHR is -4 (NHR = e^{-4}) and their Shimmer is 0.03.
- lNHR4 – represents the average effect of lNHR (increasing lNHR by 1 or multiplying NHR by e) on expected proportion_UPDRS across all 4 groups defined by fDFA.

	Estimate
(Intercept)	0.741
lNHR4	−0.015
age65	−0.000
sex1	−0.011
Shimmer0.03	0.567
fDFA1	0.019
fDFA2	−0.010
fDFA3	−0.015
lNHR4:fDFA1	0.001
lNHR4:fDFA2	−0.006
lNHR4:fDFA3	−0.016
age65:sex1	−0.002
age65:fDFA1	−0.003
age65:fDFA2	−0.000
age65:fDFA3	−0.001

Table 11: Estimates of regression coefficients from summary table of model (8) reparametrized using the contrast sum parametrization.

- age65 – represents the average effect of age (increasing age by 1) on expected proportion_UPDRS across all (8) groups defined by categorical covariates sex and fDFA.
- sex1 – represents correction of the intercept (for males). (Intercept) + sex1 represents mean of expected proportions_UPDRS in all groups defined by fDFA, when we consider male patients who are 65 years of age, their lNHR is -4 (NHR = e^{-4}) and their Shimmer is 0.03.
- Shimmer0.03 – represents effect of Shimmer on expected proportion_UPDRS. (When Shimmer increases by one, expected proportion_UPDRS increases by 0.567.)
- fDFA1 (fDFA2, fDFA3) – represents correction of the intercept (for fDFA = 1). (Intercept) + fDFA1 represents mean of expected proportions_UPDRS for males and females, when we consider patients who are 65 years of age, their lNHR is -4 (NHR = e^{-4}), their Shimmer is 0.03 and their fDFA = 1. (Interpretation of fDFA2 and fDFA3 is analogous.)
- lNHR4:fDFA1 (lNHR4:fDFA2, lNHR4:fDFA3) – is the modifier of the effect of lNHR. lNHR4 + lNHR4:fDFA1 represents the effect of lNHR (increasing lNHR by 1 or multiplying NHR by e) on the expected proportion_UPDRS for patients whose fDFA = 1. (Interpretation of lNHR4:fDFA2 and lNHR4:fDFA3 is analogous.)
- age65:sex1 – is the modifier of the effect of age. age65 + age65:sex1 represents the average effect of age (increasing age by 1) on the expected proportion_UPDRS across all fDFA groups for males.

- age65:fDFA1 (age65:fDFA2, age65:fDFA3) – is the modifier of the effect of age for fDFA. age65 + age65:fDFA1 represents the average effect of age on the expected proportion_UPDRS for male patients whose fDFA = 1 and for female patients whose fDFA = 1. (Interpretation of age65:fDFA2 and age65:fDFA3 is analogous.)

Now, we will visualize the dependence of proportion_UPDRS on NHR for different levels of fDFA. We will be interested in regression functions, in confidence bands for regression functions and also in corresponding prediction bands. Let us remark that we will keep considering model (8) reparametrized using the contrast sum parametrization (estimates of regression coefficients in Table 11).

When we choose fDFA = 1, and consider 65 years old patients whose Shimmer is 0.03, then our regression function (averaged over males and females) is:

$$\begin{aligned} m_1(\text{NHR}) &= (\text{Intercept}) + \text{fDFA1} + (\text{lNHR4} + \text{lNHR4:fDFA1})(log(\text{NHR}) + 4) \\ &= 0.741 + 0.019 + (-0.015 + 0.001)(log(\text{NHR}) + 4). \end{aligned}$$

Similarly, for fDFA = 2, fDFA = 3 and fDFA = 4, we obtain regression functions m_2, m_3 and m_4:

$$\begin{aligned} m_2(\text{NHR}) &= (\text{Intercept}) + \text{fDFA2} + (\text{lNHR4} + \text{lNHR4:fDFA2})(log(\text{NHR}) + 4) \\ &= 0.741 - 0.010 + (-0.015 - 0.006)(log(\text{NHR}) + 4), \end{aligned}$$

$$\begin{aligned} m_3(\text{NHR}) &= (\text{Intercept}) + \text{fDFA3} + (\text{lNHR4} + \text{lNHR4:fDFA3})(log(\text{NHR}) + 4) \\ &= 0.741 - 0.015 + (-0.015 - 0.016)(log(\text{NHR}) + 4), \end{aligned}$$

$$\begin{aligned} m_4(\text{NHR}) &= (\text{Intercept}) - \text{fDFA1} - \text{fDFA2} - \text{fDFA3} \\ &\quad + (\text{lNHR4} - \text{lNHR4:fDFA1} - \text{lNHR4:fDFA2} - \text{lNHR4:fDFA3})(log(\text{NHR}) + 4) \\ &= 0.741 - 0.019 + 0.010 + 0.015 \\ &\quad + (-0.015 - 0.001 + 0.006 + 0.016)(log(\text{NHR}) + 4). \end{aligned}$$

These regression functions, along with confidence bands for these regression functions as well as prediction bands are plotted in Figure 17, Figure 18, Figure 19 and Figure 20 (each figure contains only datapoints corresponding to a given level of fDFA).

For fDFA = 1 confidence band for the regression line and the corresponding prediction band are constructed as follows. For a fixed value of NHR we denote $x_{new} = (1, log(\text{NHR}) + 4, 0, 0, 0, 1, 0, 0, log(\text{NHR}) + 4, 0, 0, 0, 0, 0, 0)^\mathsf{T}$ a vector specifying m_1. (Order corresponds to the order of model terms as shown in Table 11.) Then, we calculate the corresponding prediction interval:

$$\left(x_{new}^\mathsf{T}\widehat{\beta} \mp \sqrt{MS_e\ (1 + x_{new}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}x_{new})}\ t_{n-r}(1-\alpha/2)\right),$$

where MS_e and $\mathbb{X}$ denote residual mean square and model matrix of model (8) reparametrized using the contrast sum parametrization, where $n = 850$ denotes number of observations and $r = 15$ denotes rank of the model and where $\widehat{\beta}$ denotes LSE estimator of vector of regression coefficients β (Table 11). Further we calculate the confidence band for the regression line:

$$\left(x_{new}^\mathsf{T}\widehat{\beta} \mp \sqrt{r\ F_{r,n-r}(1-\alpha)\ MS_e\ x_{new}^\mathsf{T}(\mathbb{X}^\mathsf{T}\mathbb{X})^{-1}x_{new}}\right),$$

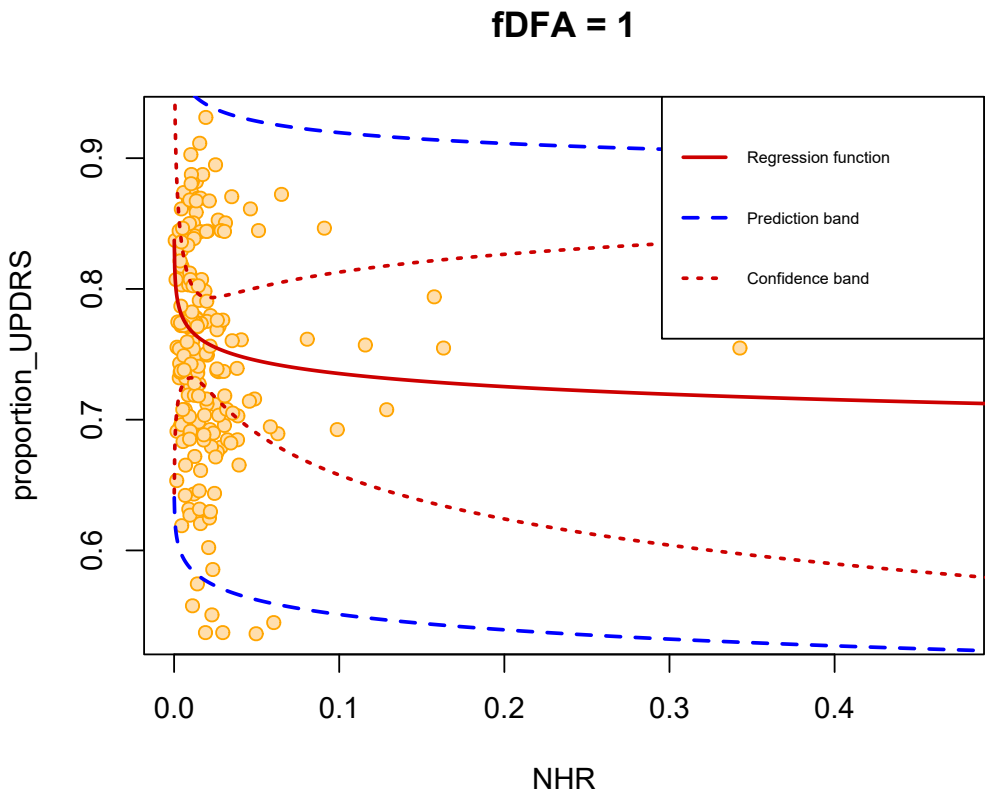

Figure 17: Regression function, confidence band for the regression function and corresponding prediction band based on NHR for fDFA = 1.

where $F_{r,n-r}(1-\alpha)$ denotes $(1-\alpha)$-quantile of $F_{r,n-r}$ distribution. Repeating this procedure for many values of NHR from interval $(0, 0.5)$ yields Figure 17.

For fDFA = 2, fDFA = 3 and fDFA = 4 we proceed in the same way, with the exception that vector x_{new} is chosen differently.

For fDFA = 2, $x_{new} = (1,log(\text{NHR})+4,0,0,0,0,1,0,0,log(\text{NHR})+4,0,0,0,0,0)^\intercal$, for fDFA = 3, $x_{new} = (1,log(\text{NHR})+4,0,0,0,0,0,1,0,0,log(\text{NHR})+4,0,0,0,0)^\intercal$ and for fDFA = 4, $x_{new} = (1,log(\text{NHR})+4,0,0,0,-1,-1,-1,-log(\text{NHR})-4,-log(\text{NHR})-4,-log(\text{NHR})-4,0,0,0,0)^\intercal$.

As we have already mentioned when answering question (c), problem with our prediction bands is that the underlying prediction interval was derived under the assumption of a normal linear model. In our situation we do not believe that this assumption is satisfied and thus we have to view calculated prediction bands only as an approximation. Similarly, for the confidence band to cover the regression function with prespecified probability of $(1-\alpha)$, the assumption of normality is also needed. Therefore, also our calculated confidence bands have to be viewed as approximative.

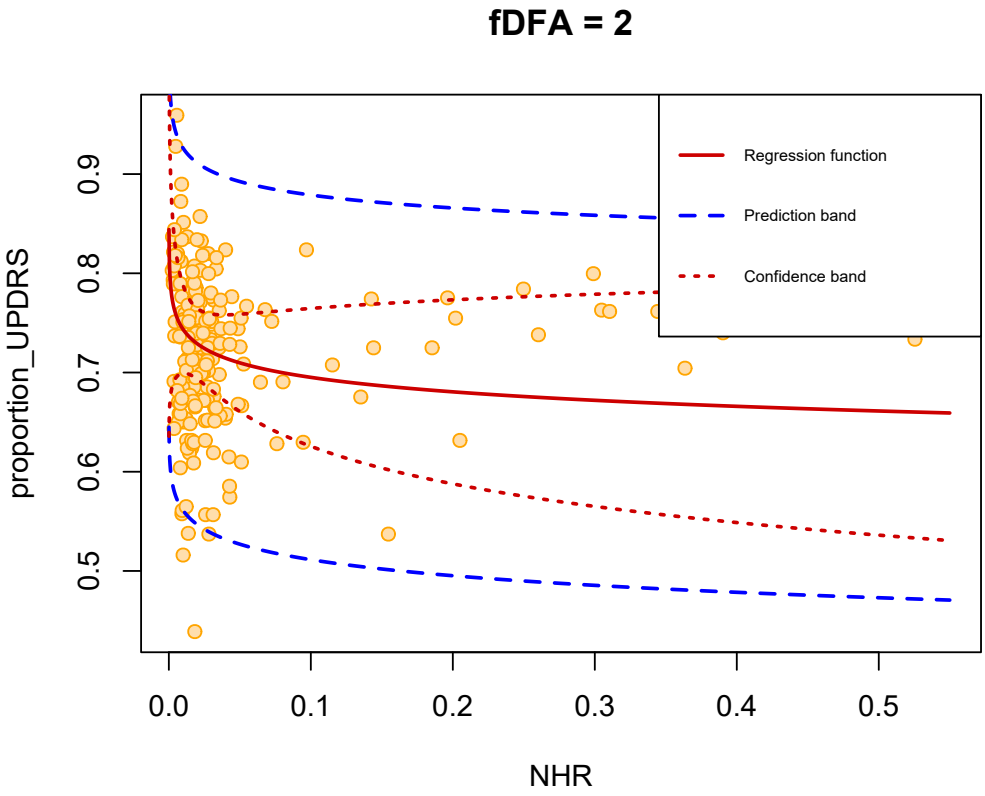

Figure 18: Regression function, confidence band for the regression function and corresponding prediction band based on NHR for fDFA = 2.

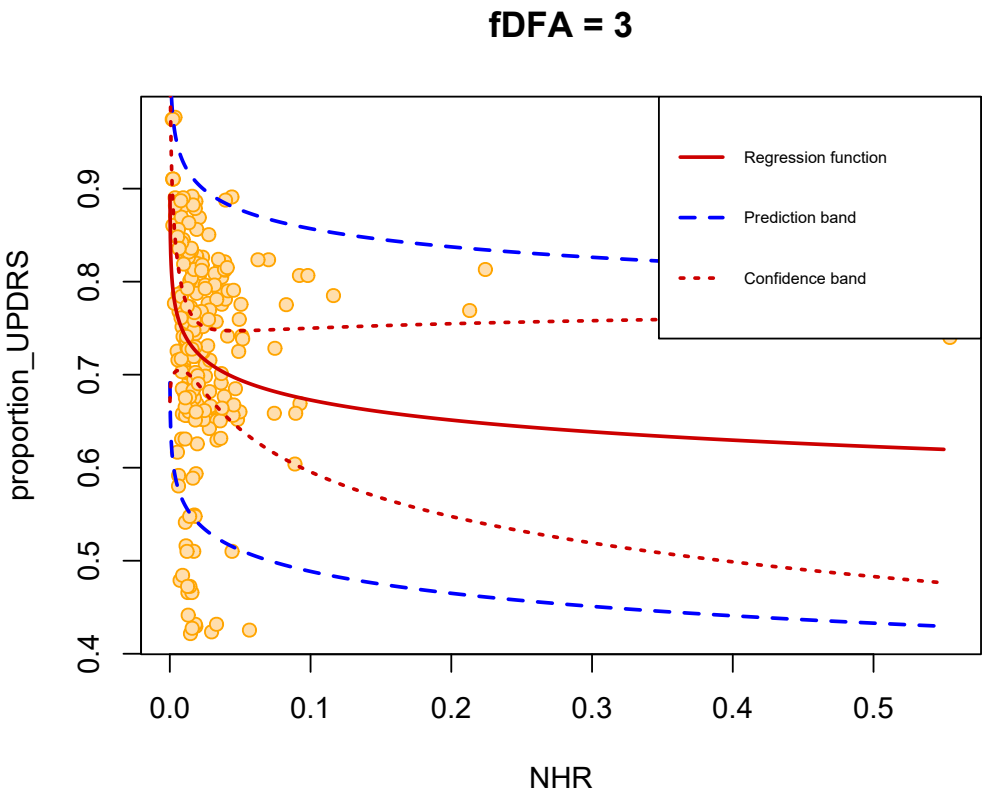

Figure 19: Regression function, confidence band for the regression function and corresponding prediction band based on NHR for fDFA = 3.

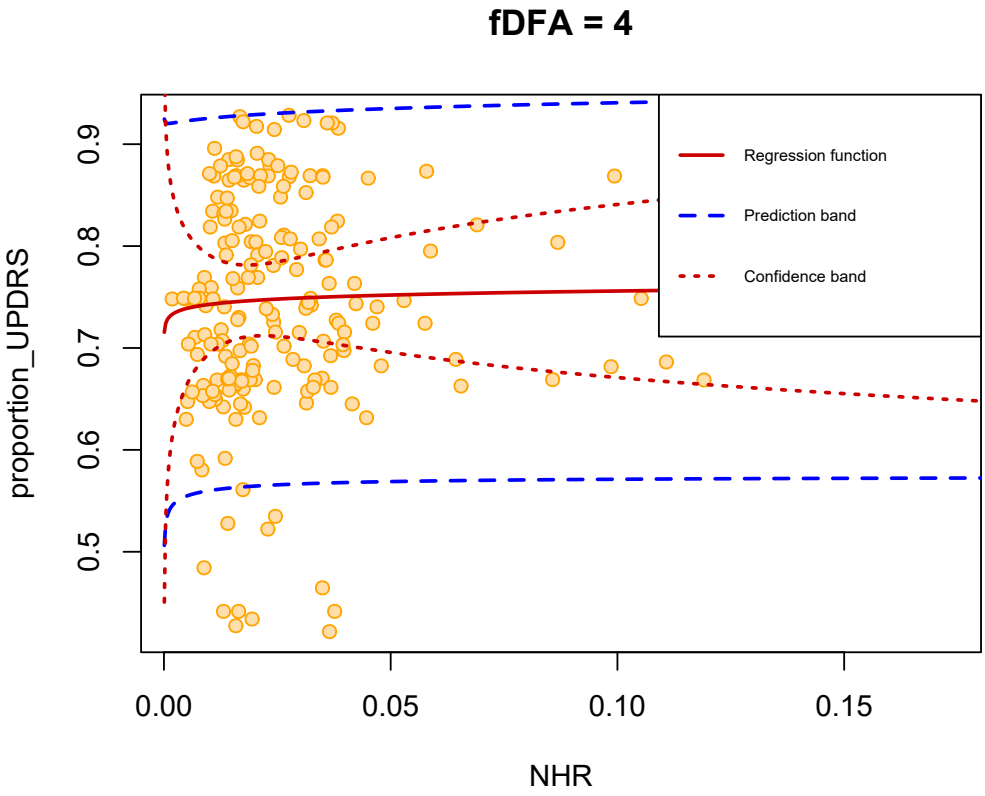

Figure 20: Regression function, confidence band for the regression function and corresponding prediction band based on NHR for fDFA = 4.