Modeling & Simulation Decisions

Моделирование динамики размера опухоли на основании данных о биомаркерах с использованием обобщенных аддитивных моделей

Кондрашина Анна

Сколтех, Life science

Научный руководитель: Яковлева Татьяна, M&S Decisions

1 Ноября 2022

Цели и задачи исследования

Цель: Исследовать взаимосвязь динамики биомаркеров с динамикой размера опухоли (SLD) у пациентов с немелкоклеточным раком легкого

Задачи:

- Описать динамику размера опухоли с помощью модели, позволяющей учитывать динамику изменений других биомаркеров
- Проанализировать, какие из биомаркеров значимы для прогноза размера опухоли
- Предсказать динамику размера опухоли для тестовой выборки, на основании модели, полученной на обучающей выборке

План Доклада

Подготовка данных для анализа

Теория GAM

Выбор модели

Результаты

Данные исследования NCT00364351

Efficacy Trial Comparing ZD6474 With Erlotinib in NSCLC After Failure of at Least One Prior Chemotherapy

https://clinicaltrials.gov/ct2/show/NCT00364351

Phase III, Sanofi Company

Вандетаниб vs **Эрлотиниб (контроль)**

Индикация: Non-small cell lung cancer (NSCLC)

Критерии включения:

- 1-2 предшествующих курсов химиотерапии не привели к достаточному улучшению состояния
- WHO Performance status <= 2

Прекращение лечения: прогрессия > 20%

Всего: 416 пациентов в контрольной руке

Маркер	Частота измерения	Значение
SLD	Раз в 4 недели до 16 недели, далее – раз в 8 недель	Размер опухоли
ALT	На 1, 2, 4, 8, 12 неделе, далее - раз в четыре недели	Повреждение печени
ALP		Повреждение печени
AST		Повреждение печени
CREAT		Повреждение почек/кахексия организма
LDH		Рост опухоли
NLR		Воспаление
NEUT		Воспаление
WBC		Воспаление

Статистика начальных значений биомаркеров (n = 416)

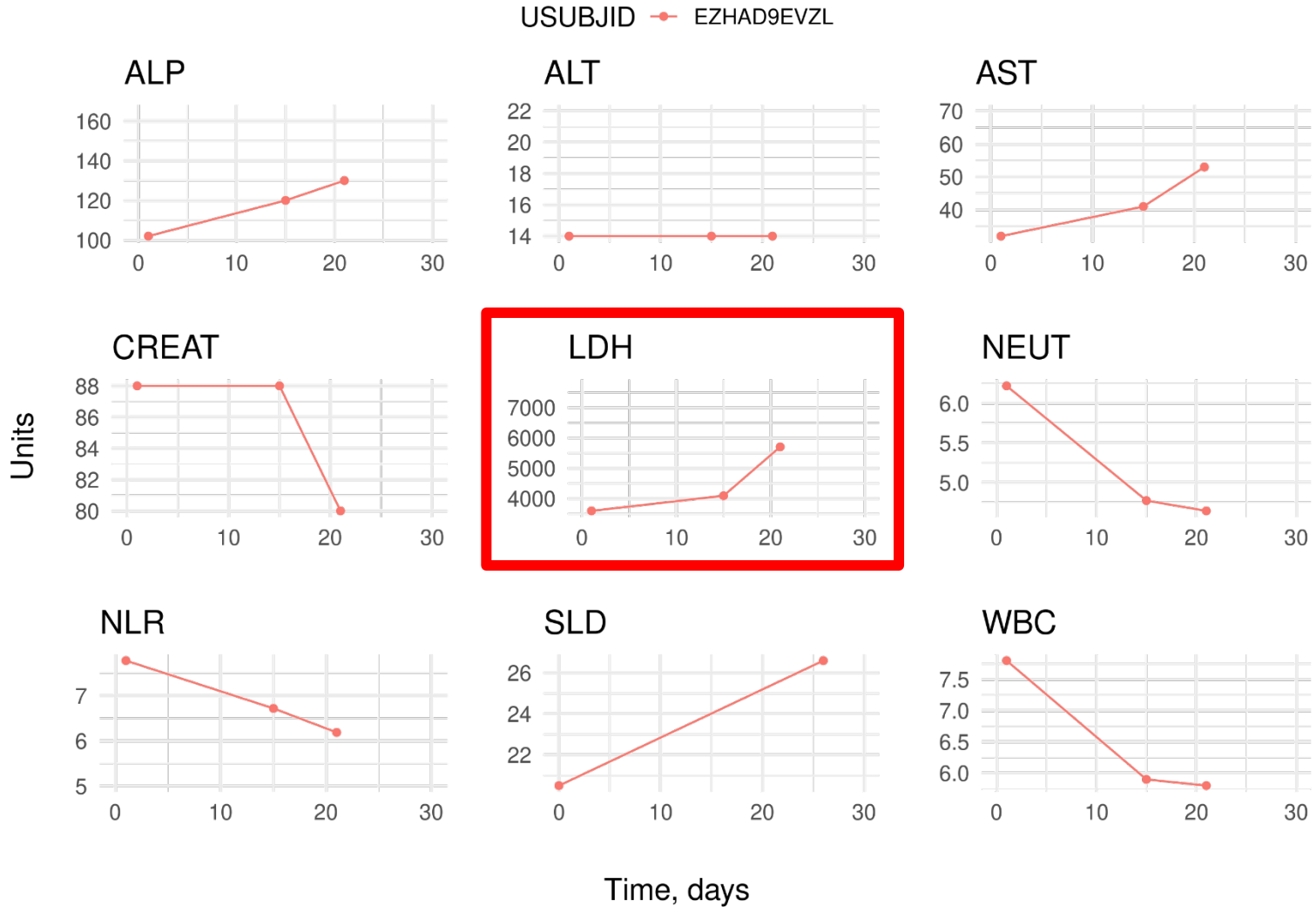
Маркер	Среднее	Медиана	Доверительный интервал, 95%	Физиологические значения	Единицы Измерения
ALP	101.8	91	(100.3; 103.4)	42 - 141[1]	U/L
ALT	18.2	15	(17.7; 18.6)	7 - 45[1]	U/L
AST	21.8	19	(21.4; 22.2)	5 - 35[1]	U/L
CREAT	82.4	80	(81.6; 83.1)	52.2 -119.3[4]	uM
LDH	**227.5**	178	(218.5; 236.5)	**125 - 220**[1]	U/L
NEUT	**6.4**	5.3	(6.3; 6.6)	**2.6 – 6.0**[2]	cells/uL
NLR	**5.8**	**3.9**	(5.6; 6.2)	**1.1 – 3.5**[2]	-
WBC	8.6	7.7	(8.5; 8.8)	4.5 – 11.0[3]	10^9 cells /L
SLD	8.7	7.3	(8.4; 9.0)	0 – 44.4*	sm

** Минимальное и максимальное значение показателя в выборке*

References:

1. Johnson-Davis KL. Enzymes. In: Bishop ML, Fody EP, Schoeff LE, eds. Clinical Chemistry: Principles, Techniques, and Correlations. 7th ed. Philadelphia: Lippincott Williams & Wilkins. 2013;262-289
2. Song, M., Graubard, B.I., Rabkin, C.S. *et al.* Neutrophil-to-lymphocyte ratio and mortality in the United States general population. *Sci Rep* **11,** 464 (2021);
3. Dean L. Blood Groups and Red Cell Antigens [Internet]. Bethesda (MD): National Center for Biotechnology Information (US); 2005. Table 1, Complete blood count.
4. Landry DW, Bazari H. Approach to the patient with renal disease. In: Goldman L, Schafer AI, eds. *Goldman-Cecil Medicine*. 26th ed. Philadelphia, PA: Elsevier; 2020:chap 106

Анализ выбросов



Для пациента EZHAD9EVZL наблюдались аномально высокие значения LDH, в том числе до начала лечения

Исключаем

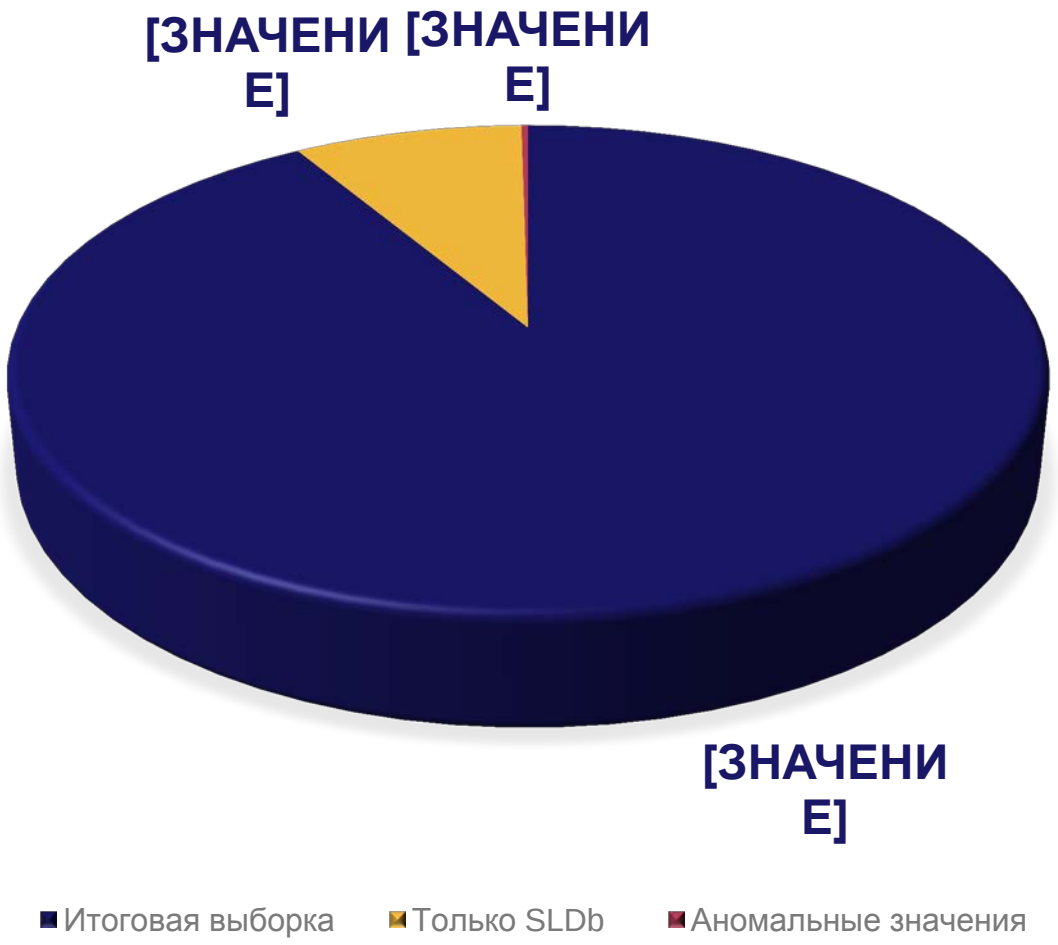
Анализ аутлаеров

Исключаем :

- 1 пациента из-за аномально высоких значений LDH
- 35 пациентов, у которых измерены только начальные значения SLD

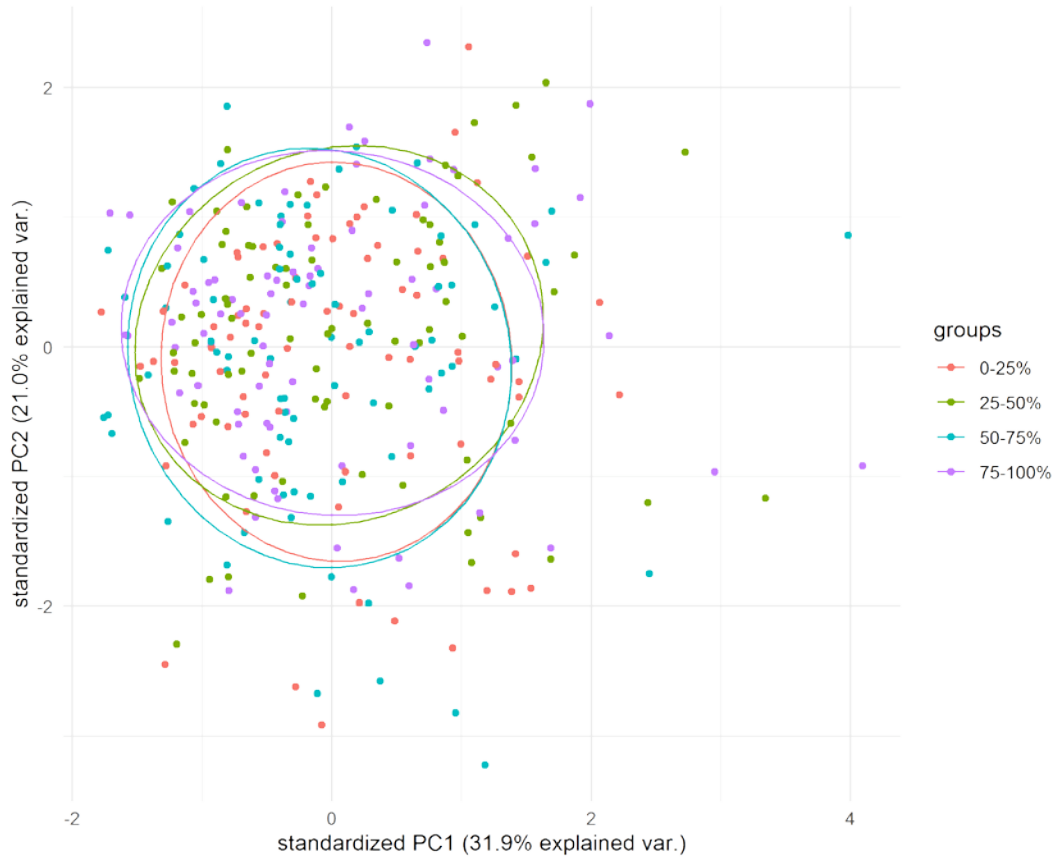
Итоговая выборка для анализа:

380 пациентов, 25382 измерения

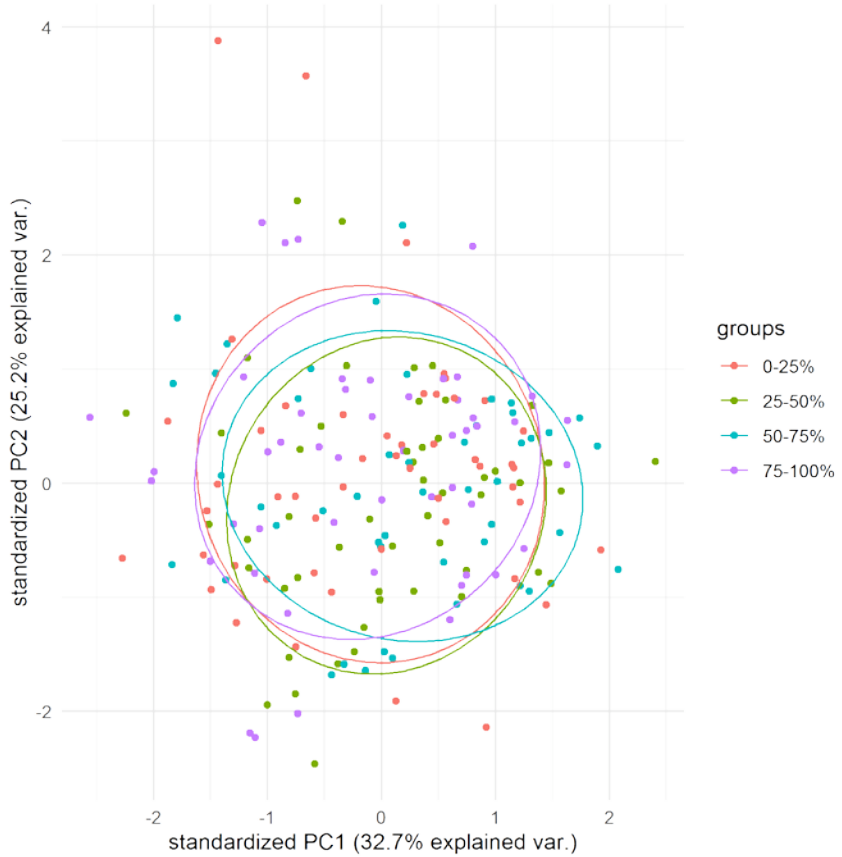


Анализ Главных Компонент

0 день

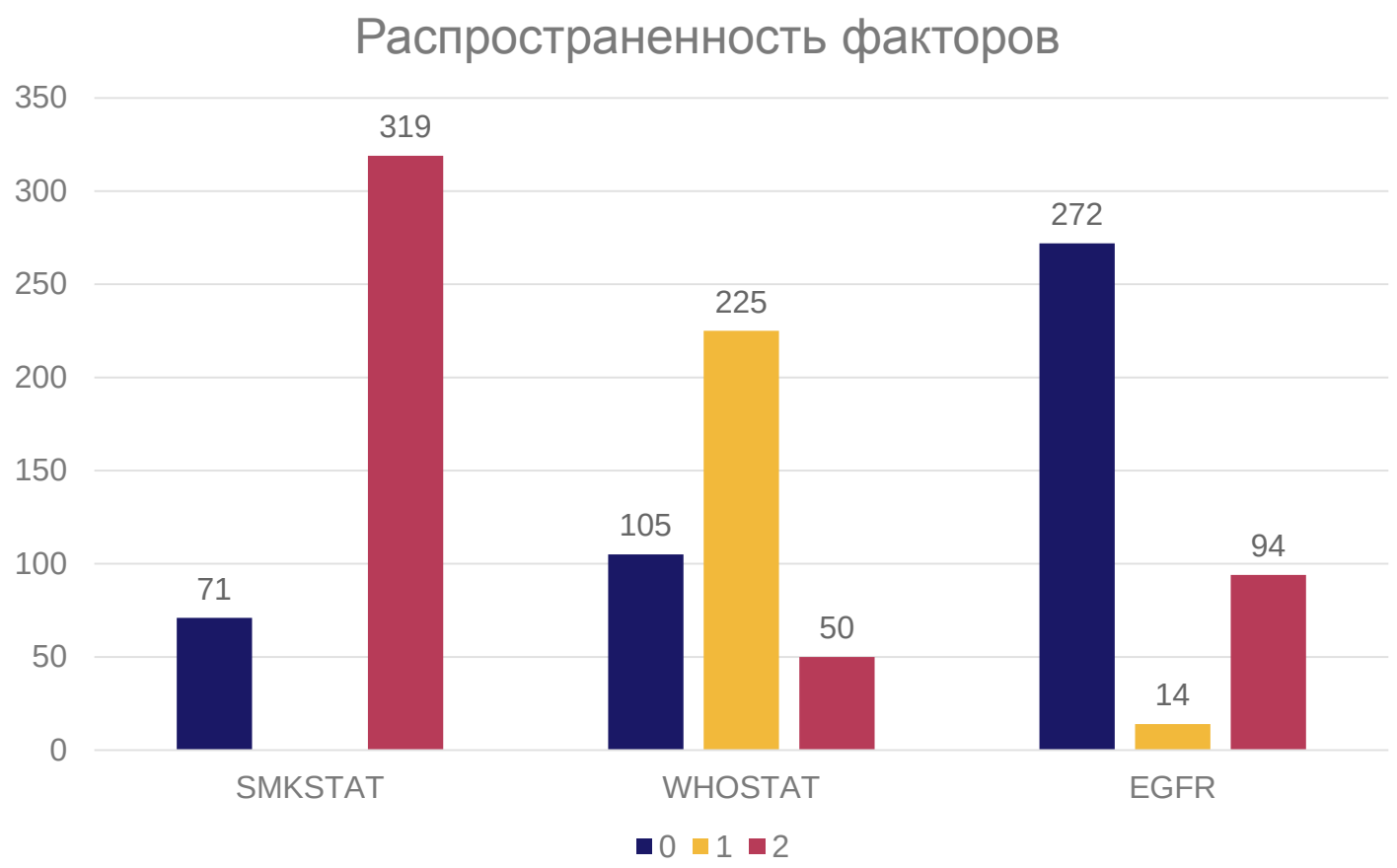


28ой день



В разные моменты времени не наблюдалось формирования кластеров пациентов

Категориальные биомаркеры



SMKSTAT:

0 – не курит; 2 – курит

WHOSTAT:

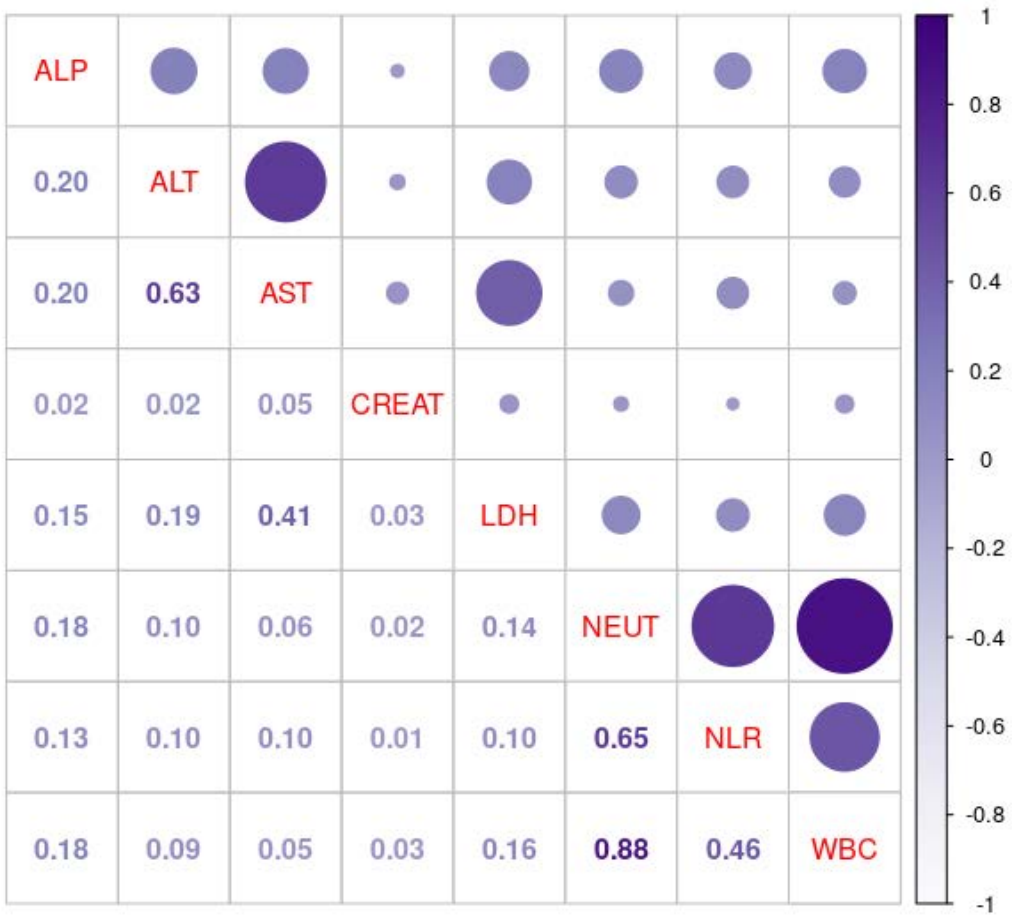
0 – способен работать;
1 – не может выполнять тяжелую работу;
2 – не способен работать

EGFRMUTN:
0 – не известно; 1 – есть мутация;
2 – нет мутации

Для некоторых из категориальных биомаркеров характерна несбалансированность

Анализ корреляций биомаркеров

Spearman correlation coefficients



- Отмечен значительный коэффициент корреляции между ALT и AST (0.63): оба биомаркера являются показателями состояния печени
- NEUT, NLR, WBC также существенно коррелируют между собой как элементы лейкоцитарной формулы крови

План доклада

Подготовка данных для анализа

Теория GAM

Выбор модели

Результаты

От LM к GAM

Linear Model (LM)

Generalized Linear Model (GLM)

Generalized Additive Model (GAM)

	LM	GLM	GAM
Basic structure	$Y_i = \beta_i X_i + \varepsilon_i$	$g(E(Y_i)) = \beta X_i + Z_i \varepsilon$	$g(E(Y_i)) = f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}, x_{4i}) + \ldots$ f_1, f_2, f_3.. - unknown
Description	The sum of linear predictor with depends linearly on the parameters	Smooth monotonic function of the linear predictor	GLM with a linear predictors involving a sum of smooth functions of covariates
Y distribution	*Y ~ Normal*	*Y ~ Exponential Family (normal, Poisson, Bernouli, gamma, etc)*	*Y ~ Exponential Family (normal, Poisson, Bernouli, gamma, etc)*
ε distribution	$\varepsilon \sim Norm(0, \sigma)$	$\varepsilon \sim Norm(0, \sigma)$	$\varepsilon \sim Norm(0, \sigma)$

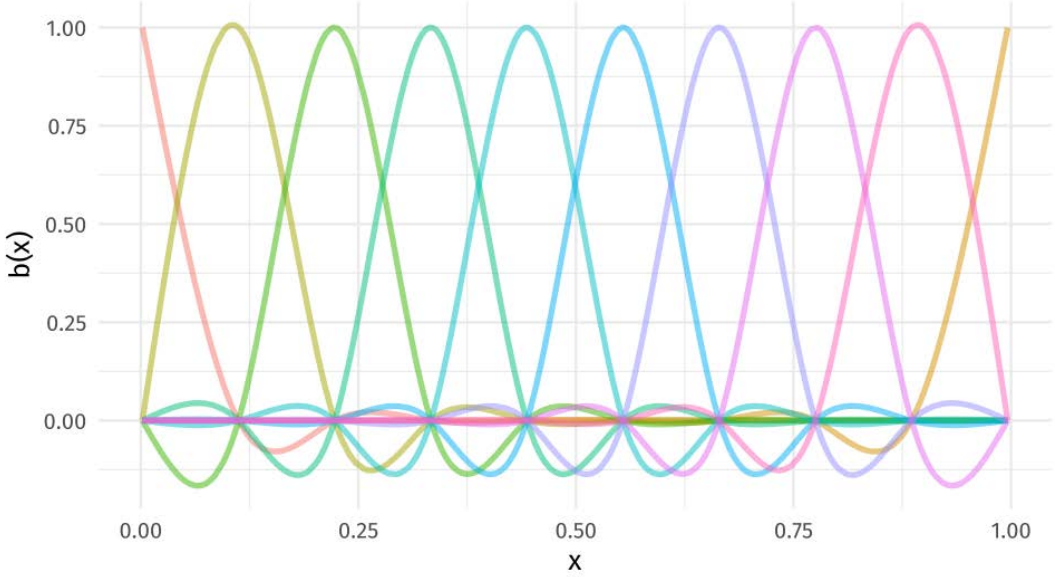
Теория GAM

$$y_i = \beta_0 + \sum_j s_j(x_{ji}) + \epsilon_i$$

Теория GAM

$$y_i = \beta_0 + \sum_j s_j(x_{ji}) + \epsilon_i$$

$$s(x) = \sum_{k=1}^{K} \beta_k b_k(x)$$

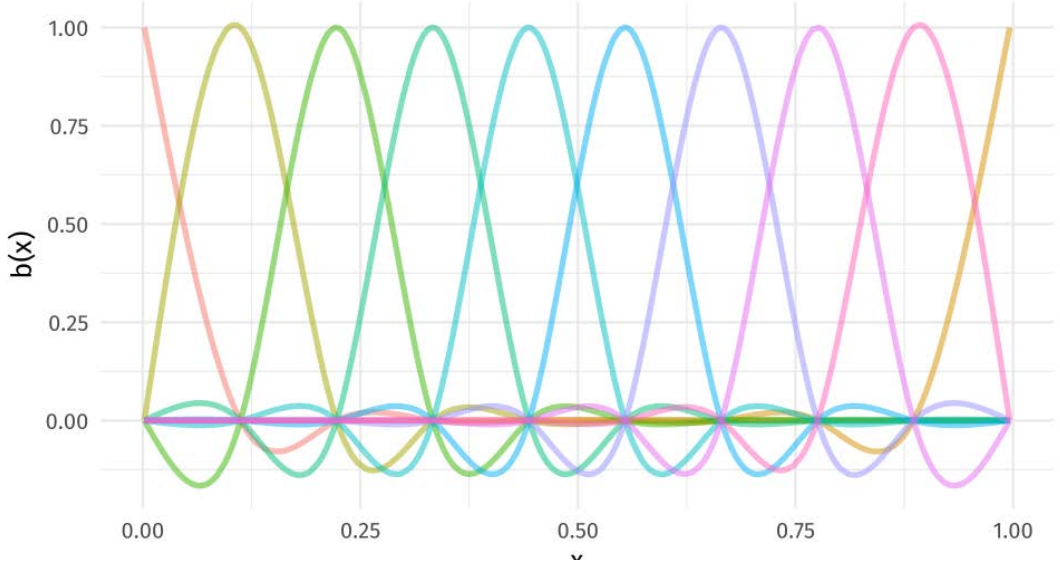


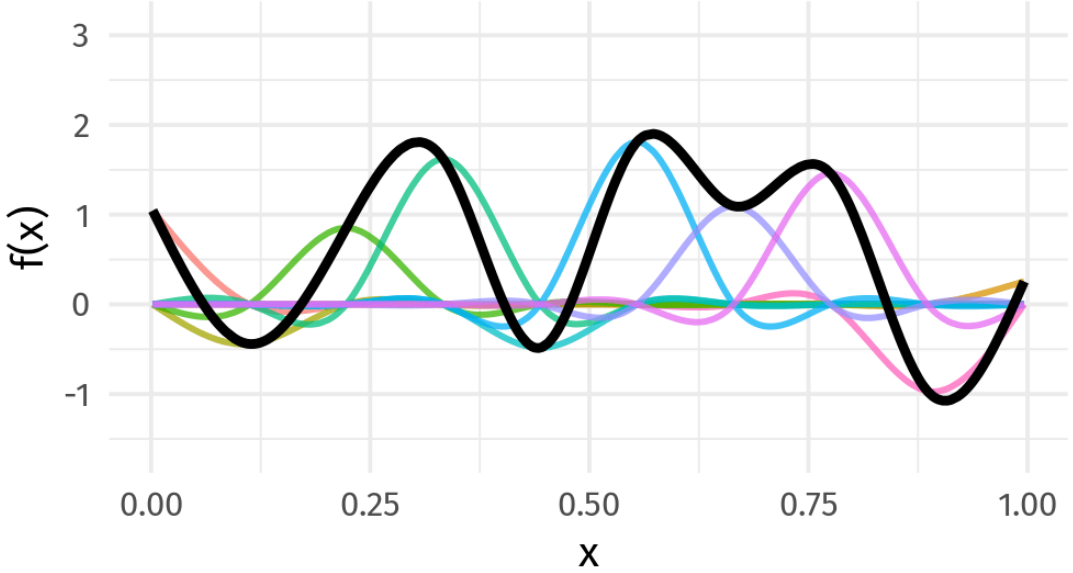
Теория GAM

$$y_i = \beta_0 + \sum_j s_j(x_{ji}) + \epsilon_i$$

Сплайн

$$s(x) = \sum_{k=1}^{K} \beta_k b_k(x)$$



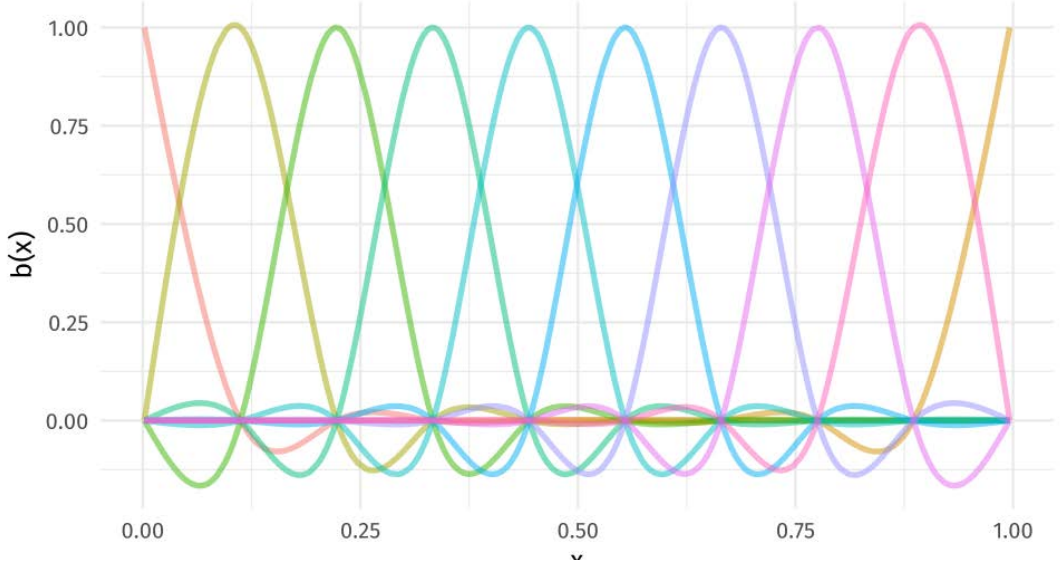


Теория GAM

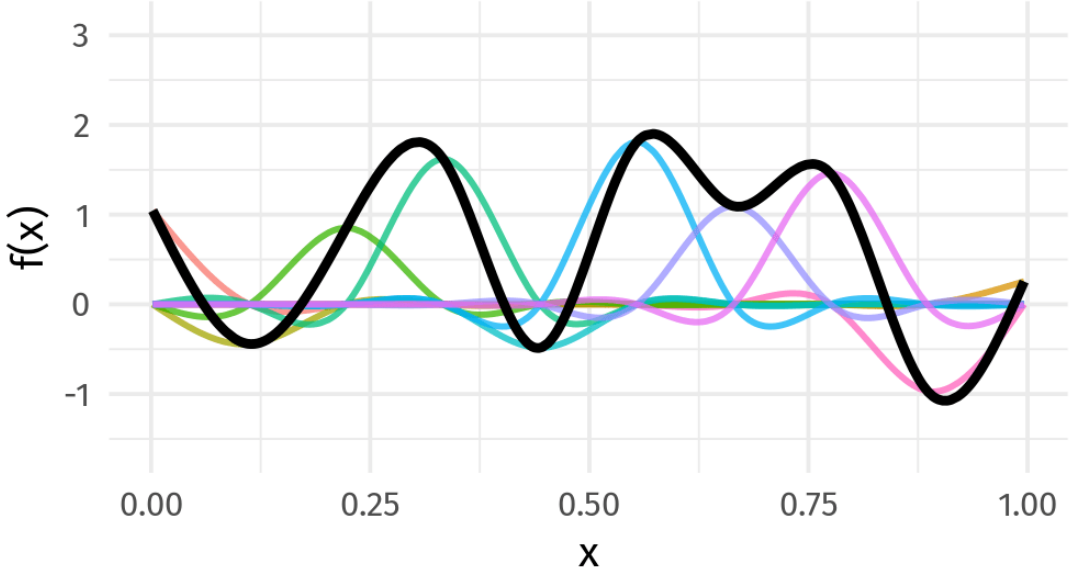
$$y_i = \beta_0 + \sum_j s_j(x_{ji}) + \epsilon_i$$

Сплайн

$$s(x) = \sum_{k=1}^{K} \beta_k b_k(x)$$



Сплайнами можно описать любые, даже самые незначительные изменения, что может привести к **переобучению**. Как избежать?

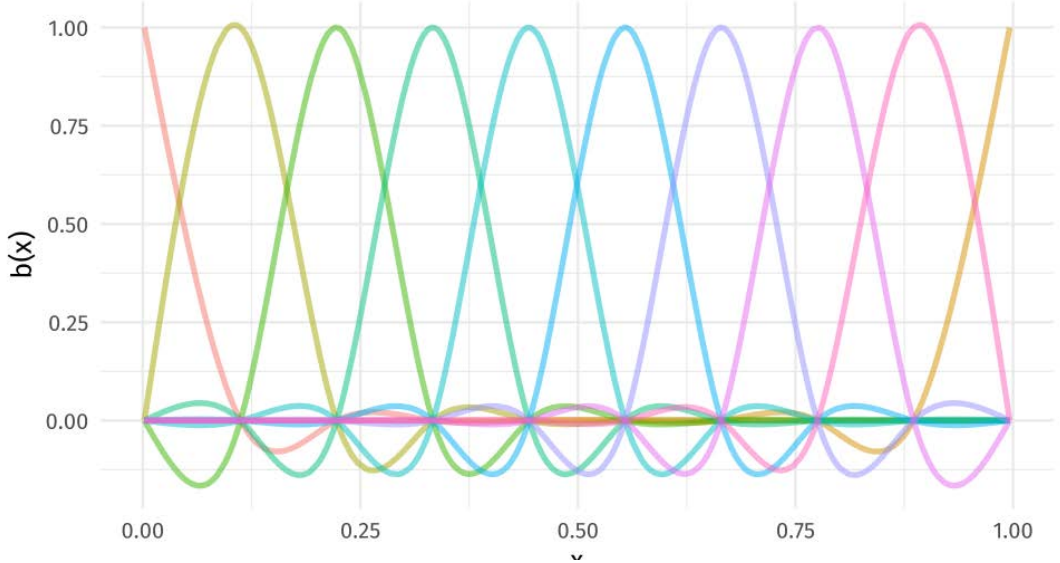


Теория GAM

$$y_i = \beta_0 + \sum_j s_j(x_{ji}) + \epsilon_i$$

Сплайн

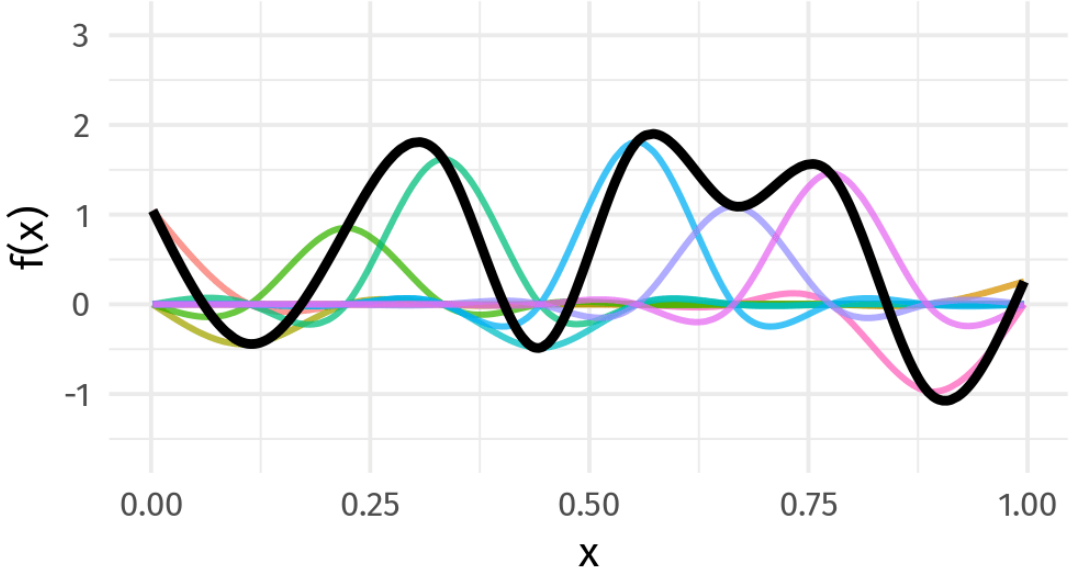
$$s(x) = \sum_{k=1}^{K} \beta_k b_k(x)$$



Сплайнами можно описать любые, даже самые незначительные изменения, что может привести к **переобучению**. Как избежать?

Регуляризация: $$\int_{\mathbb{R}} [f'']^2 dx = \boldsymbol{\beta}^{\mathsf{T}} \mathbf{S} \boldsymbol{\beta} = W$$

$$\mathcal{L}_p = \log(\text{Likelihood}) - \lambda W$$



План доклада

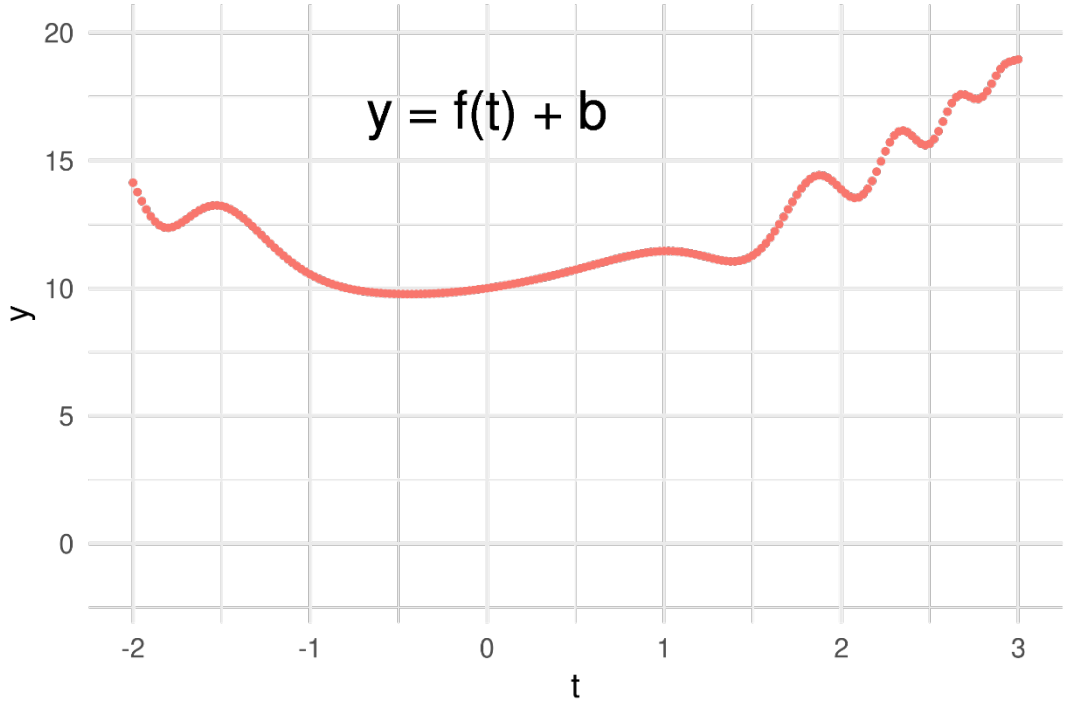
Подготовка данных для анализа

Теория GAM

Выбор модели

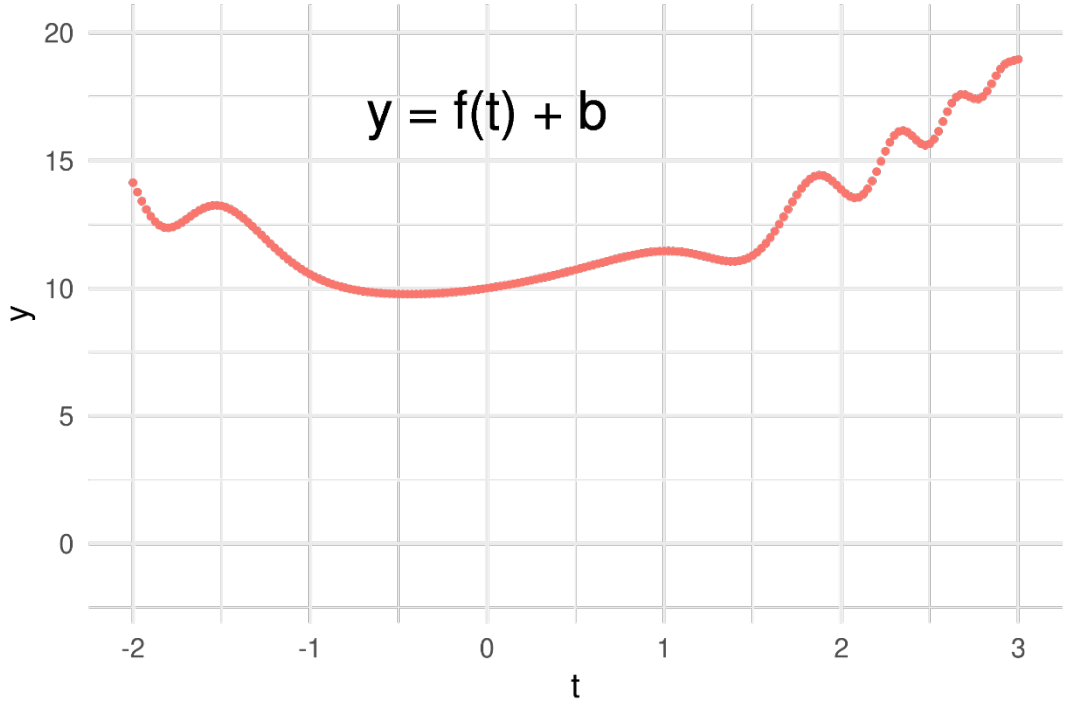
Результаты

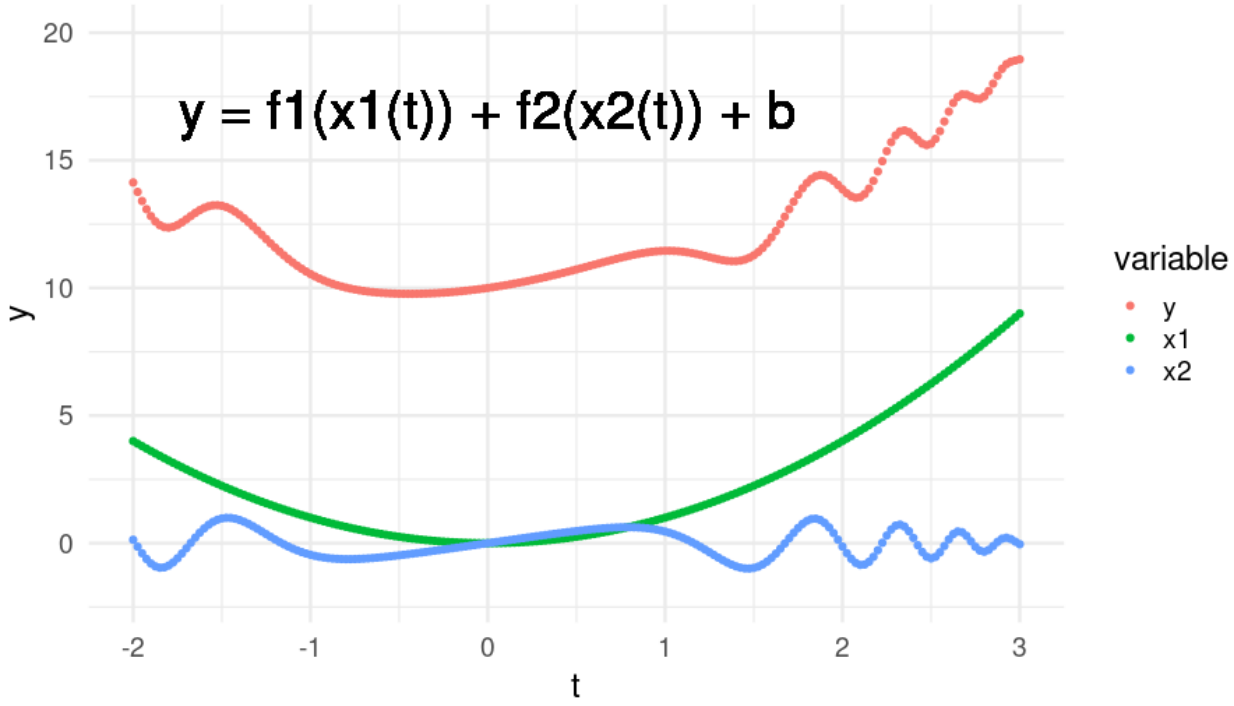
Работа со временем



С помощью сплайнов можем описывать функции от независимых переменных

Зависимость от времени и описание динамики биомаркеров

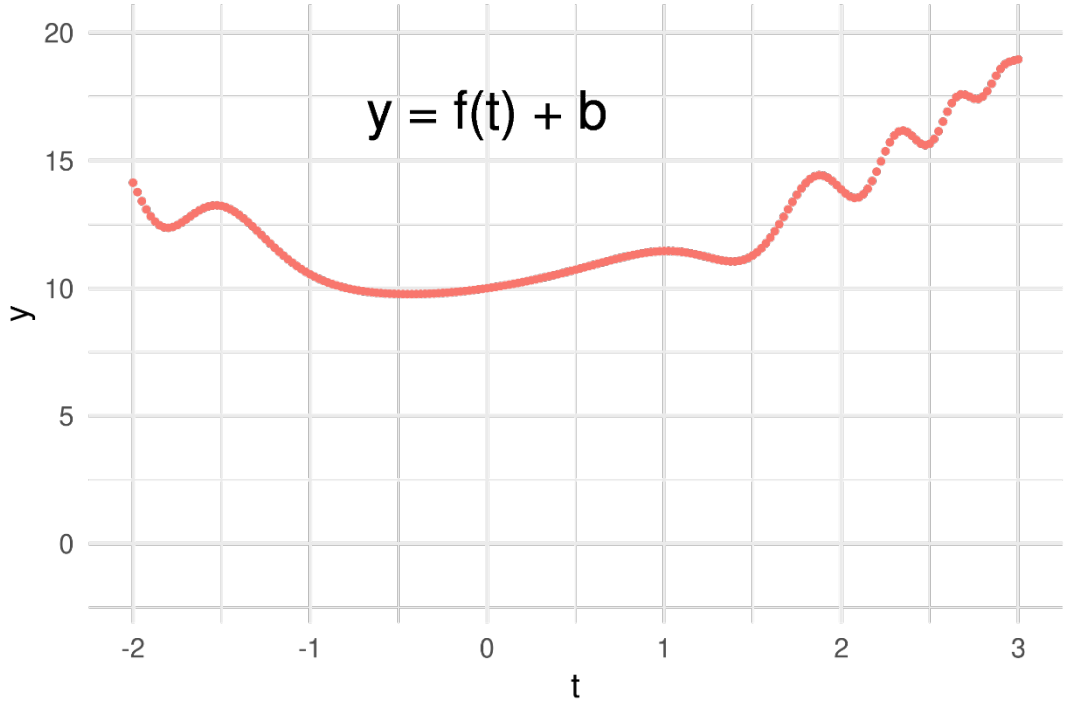




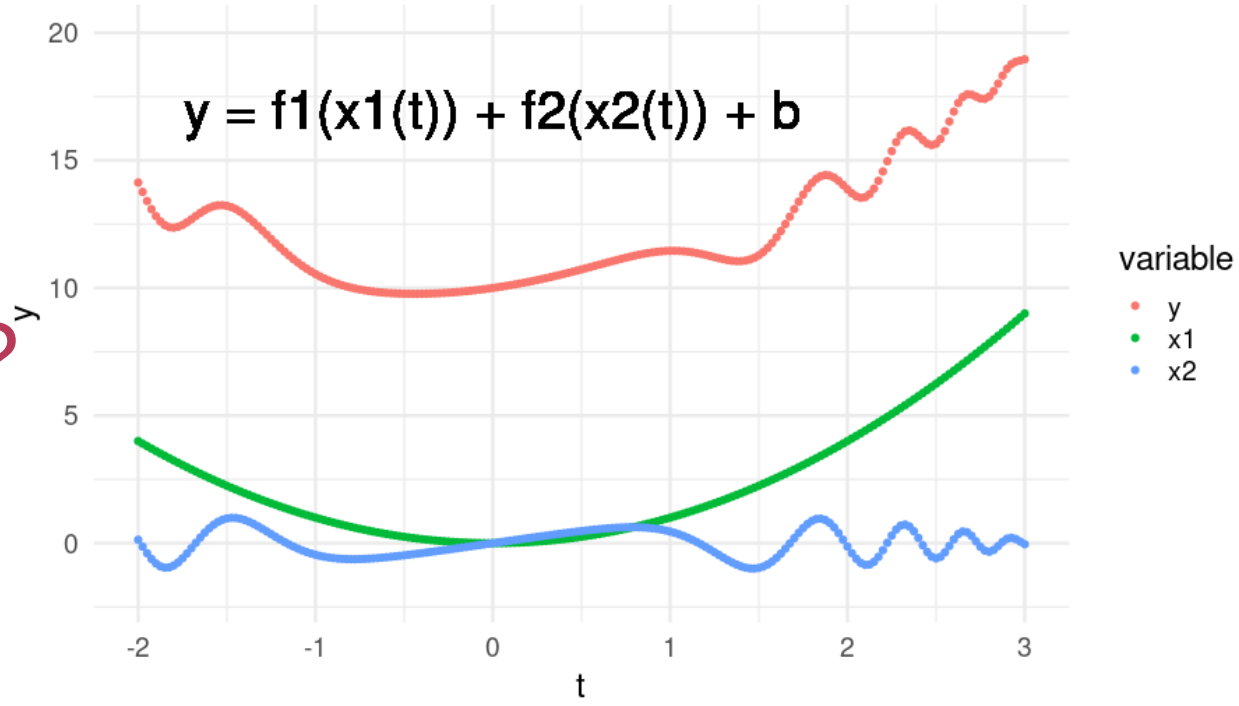
С помощью сплайнов можем описывать функции от независимых переменных

Как описать функцию от функции?

Зависимость от времени и описание динамики биомаркеров



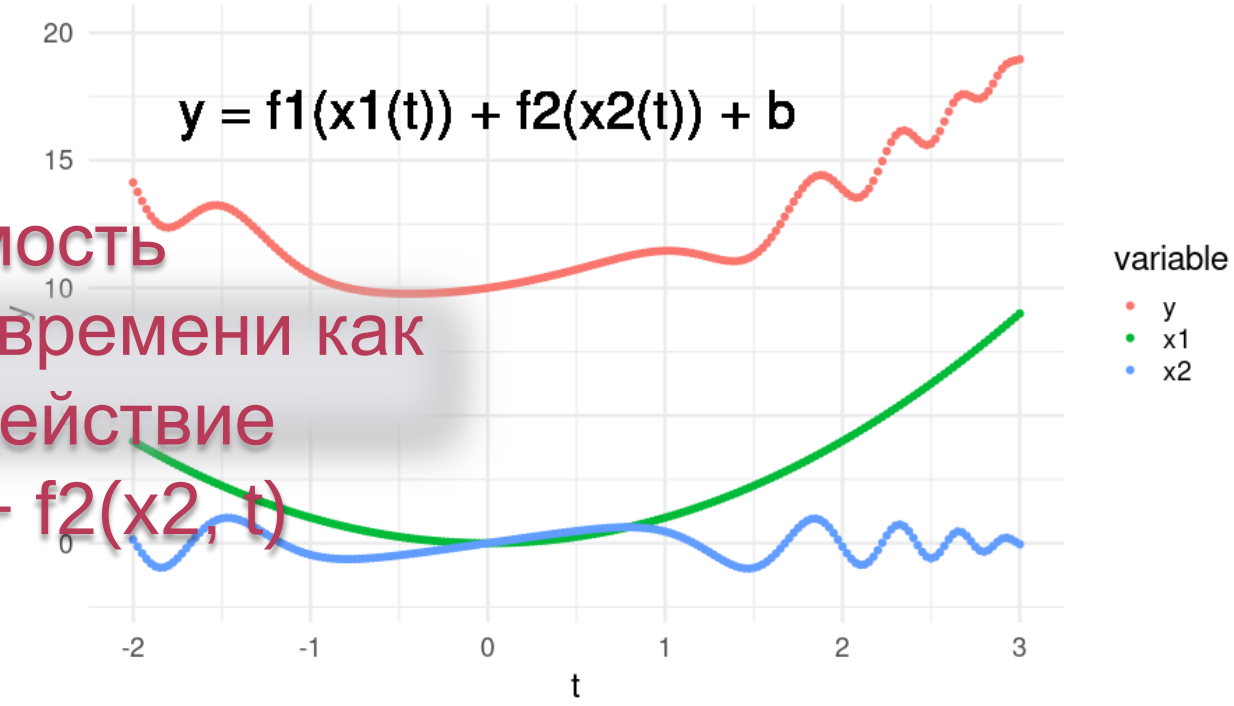
???



С помощью сплайнов можем описывать функции от независимых переменных

Как описать функцию от функции?

Зависимость от времени и описание динамики биомаркеров



Зависимость биомаркера от времени как их взаимодействие

$y = f1(x1, t) + f2(x2, t)$

С помощью сплайнов можем описывать функции от независимых переменных

Как описать функцию от функции?

Согласование времен измерения биомаркеров

	ALP	ALT	AST	CREAT	LDH	NEUT	NLR	SLD	WBC
0	1	1	1	1	1	1	1	1	1
8	1	1	1	1	1	1	1	0	1
15	1	1	1	1	1	1	1	0	1
28	0	0	0	0	0	0	0	1	0
29	1	1	1	1	1	1	1	0	1
56	0	0	0	0	0	0	0	1	0
57	1	1	1	1	1	1	1	0	1
84	0	0	0	0	0	0	0	1	0
85	1	1	1	1	1	1	1	0	1
112	0	0	0	0	0	0	0	1	0
113	1	1	1	1	1	1	1	0	1
139	1	1	1	1	1	1	1	0	1
168	0	0	0	0	0	0	0	1	0
169	1	1	1	1	1	1	1	0	1
197	1	1	1	1	1	1	1	0	1
224	0	0	0	0	0	0	0	1	0
225	1	1	1	1	1	1	1	0	1
253	1	1	1	1	1	1	1	0	1

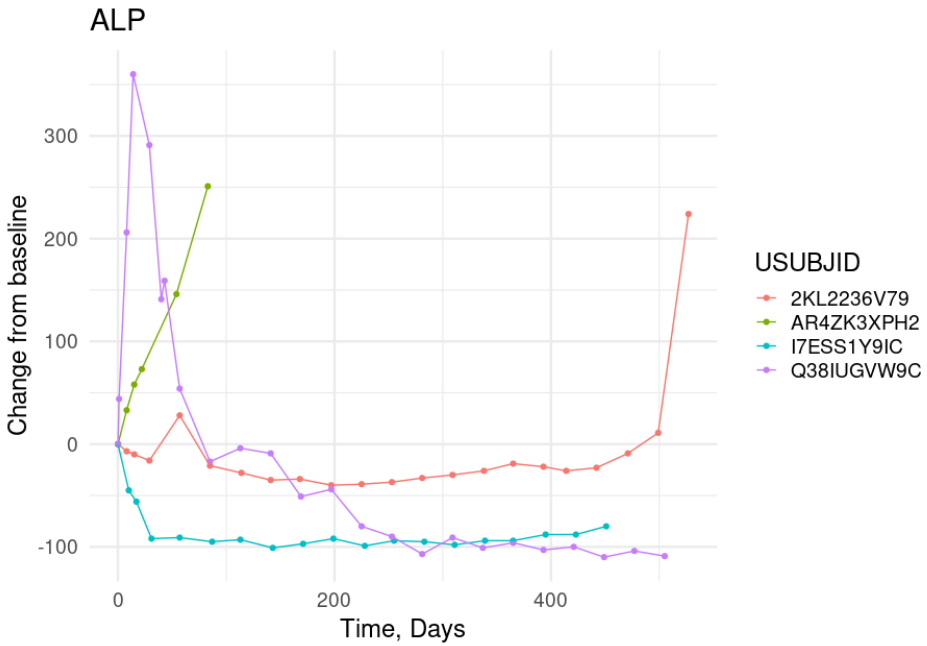
- Размер опухоли не всегда измерялся в один и тот же день с другими биомаркерами
- Согласно протоколу исследования пациенты могли посещать исследовательский центры для измерения биомаркеров с отклонением ± 3 дня от запланированной даты визита

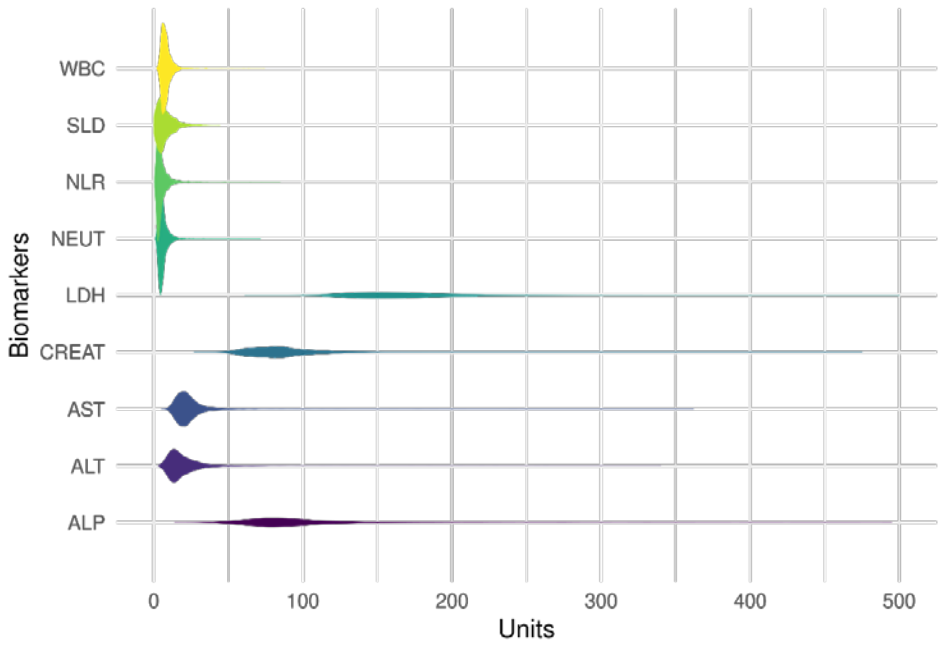
Решение:

Для синхронизации времени считаем измерения в пределах 3х дневного периода сделанными в рамках одного визита

Нормализация данных

Данные до нормализации:



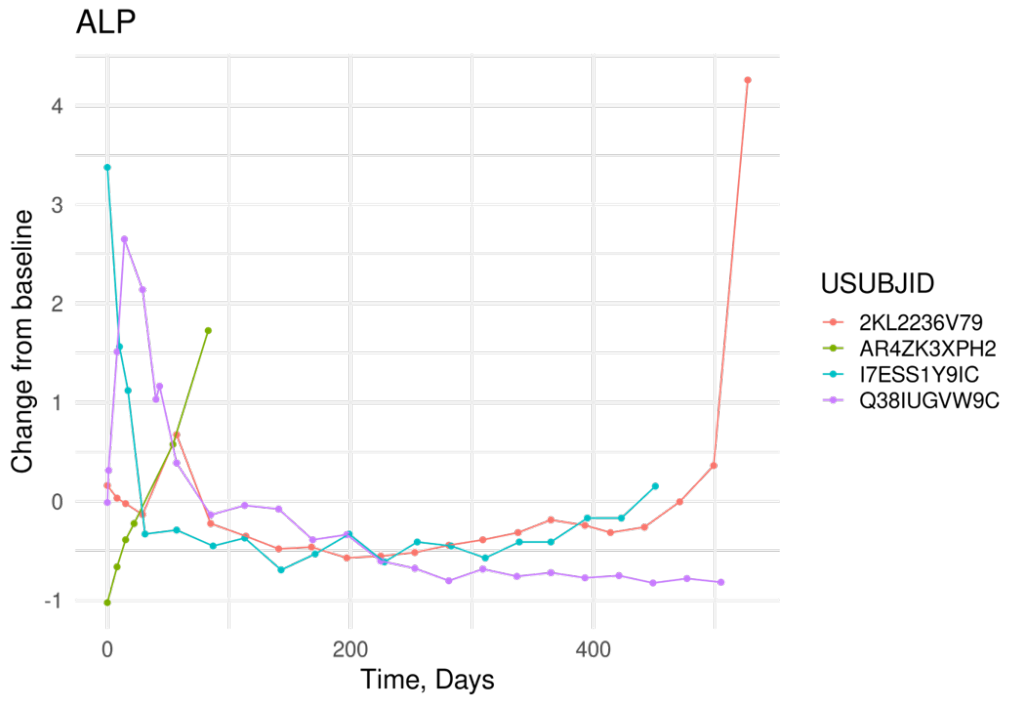


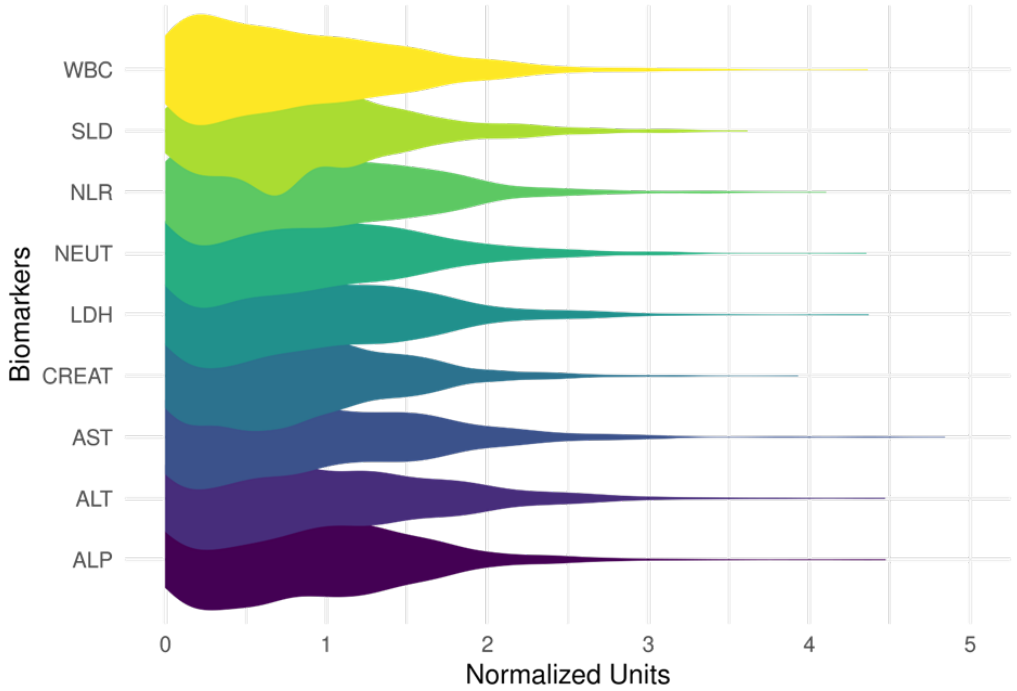
- По умолчанию GAM моделирование предполагает создание единой модели, описывающей зависимость динамики SLD от всех выбранных биомаркеров **с общим параметром пенализации на выпуклость**.
- Так как абсолютные значения биомаркеров и их единицы измерения существенно различаются это может негативно повлиять на сходимость модели.

Необходима нормализация данных

Нормализация данных

Данные после нормализации:





y = Value - Baseline

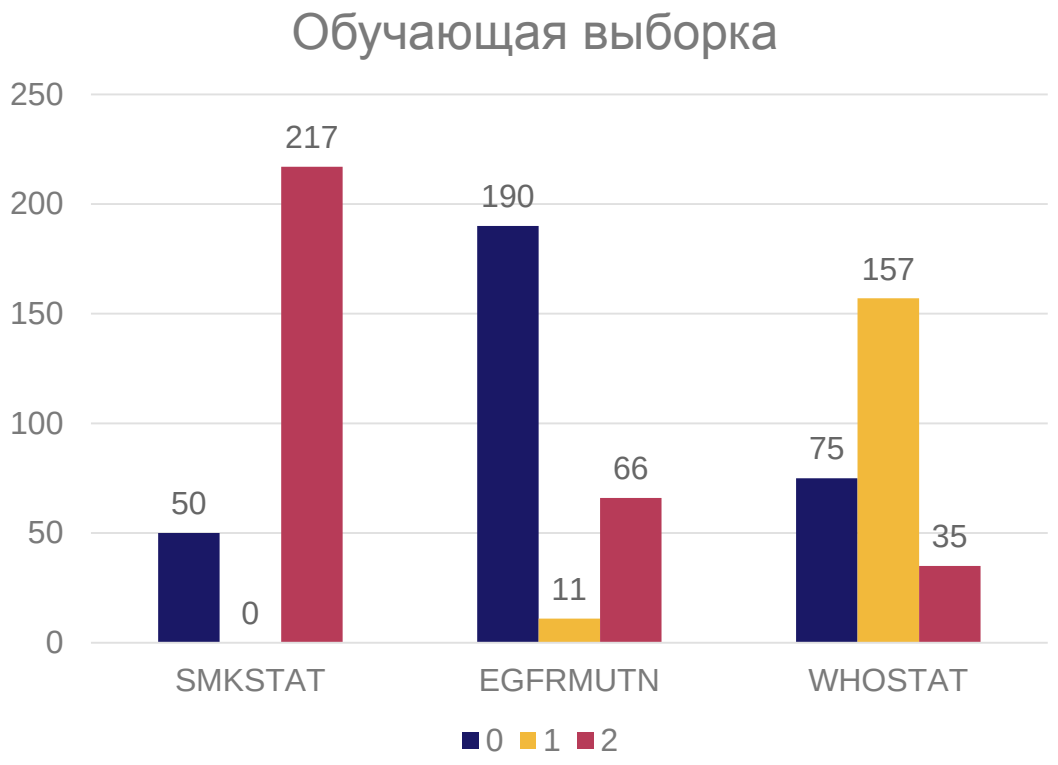
norm = (y - mean(y)) / sd(y)

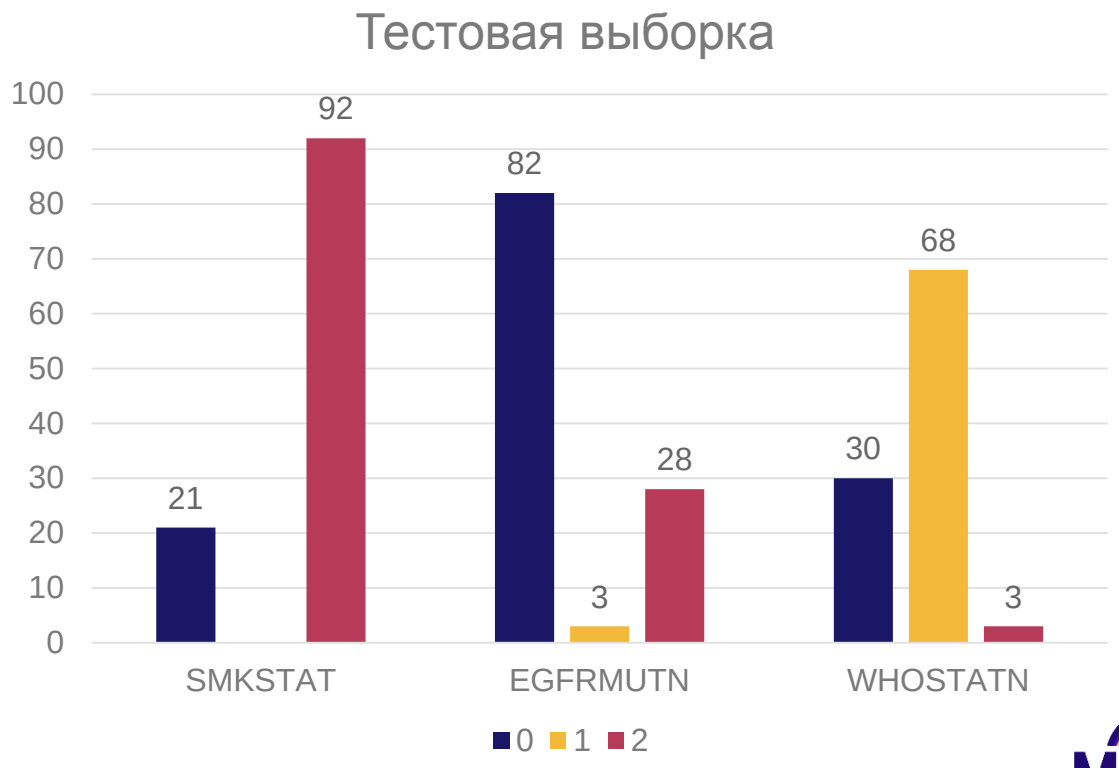
Karvanen, Juha. (2003). The Statistical Basis of Laboratory Data Normalization. Drug Information Journal - DRUG INF J. 37. 101-10
Chuang-Stein C. Summarizing Laboratory Data with Different Reference Ranges in Multi-Center Clinical Trials. Drug Information Journal

Разбиение на обучающую и тестовую выборки

Так как исходная выборка несбалансированная по факторным биомаркерам разбиение было сделано с учетом стратифицикации с помощью пакета *splitstackshape*, функция *stratified**

https://cran.r-project.org/web/packages/splitstackshape/index.html





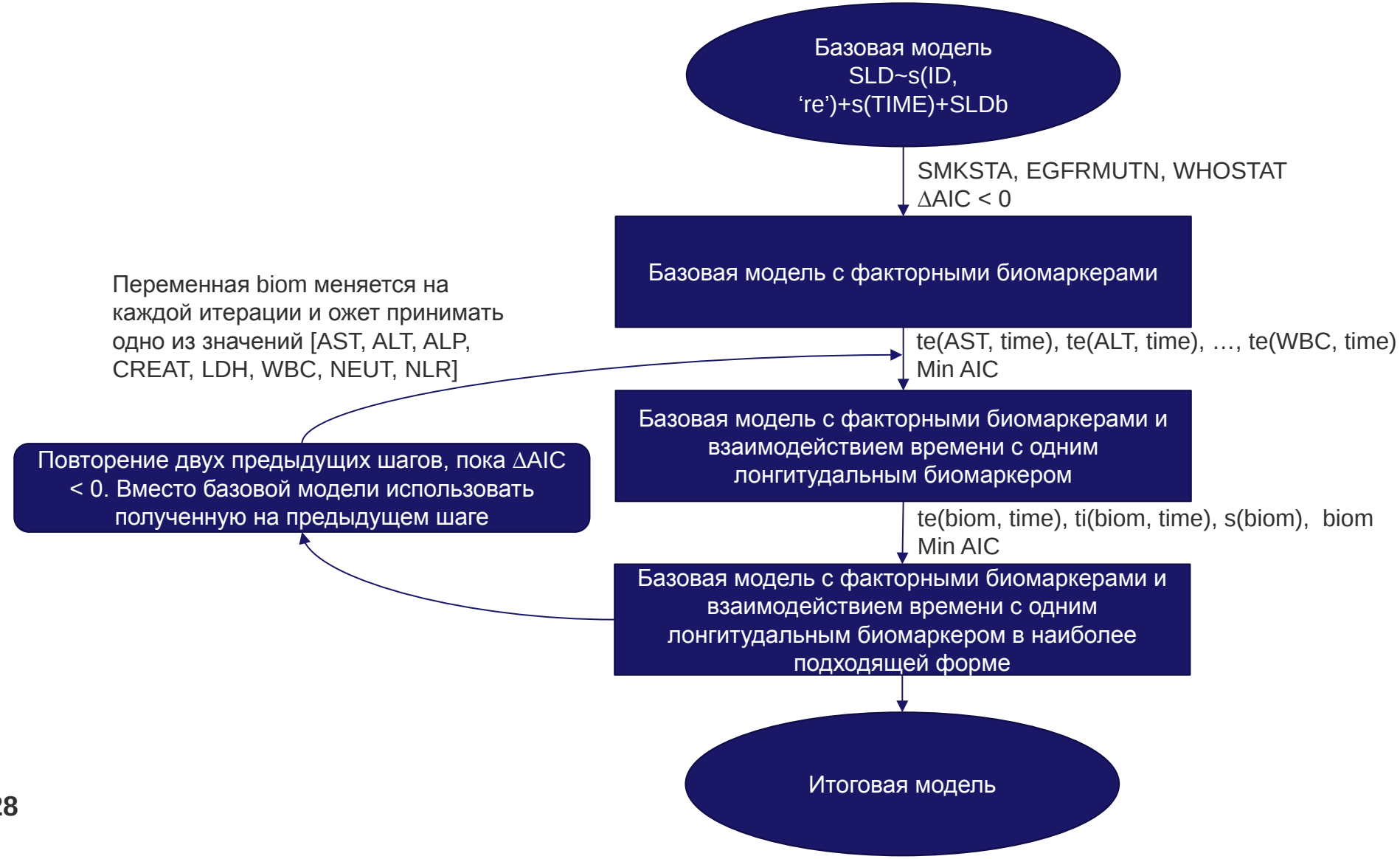
*данная функция позволяет эффективно разбивать датасеты на эквивалентно сбалансированные

Возможные гиперпараметры

Тестирование ковариат и их взаимодействия	AST, ALT, ALP, CREAT, LDH, NEUT, NLR, WBC s(), ti(), te(), линейно
Подбор вида сплайнов и количества базисных функций k	ad, tp, cs, etc.
Тестирование различных алгоритмов нормализации	$\boxed{\frac{x - mean(x)}{sd(x)}}$ $\frac{x}{mean(x)}$ $\frac{x - min(x)}{max(x) - min(x)}$
Подбор распределения зависимой переменной	$\boxed{Y \sim gaulss,}$ ***gauss, gamm, tw, etc.***

Построение GAM модели – сложный итерационный процесс
В рамках анализа были протестированы более 100 моделей

Алгоритм подбора ковариат



План доклада

Подготовка данных для анализа

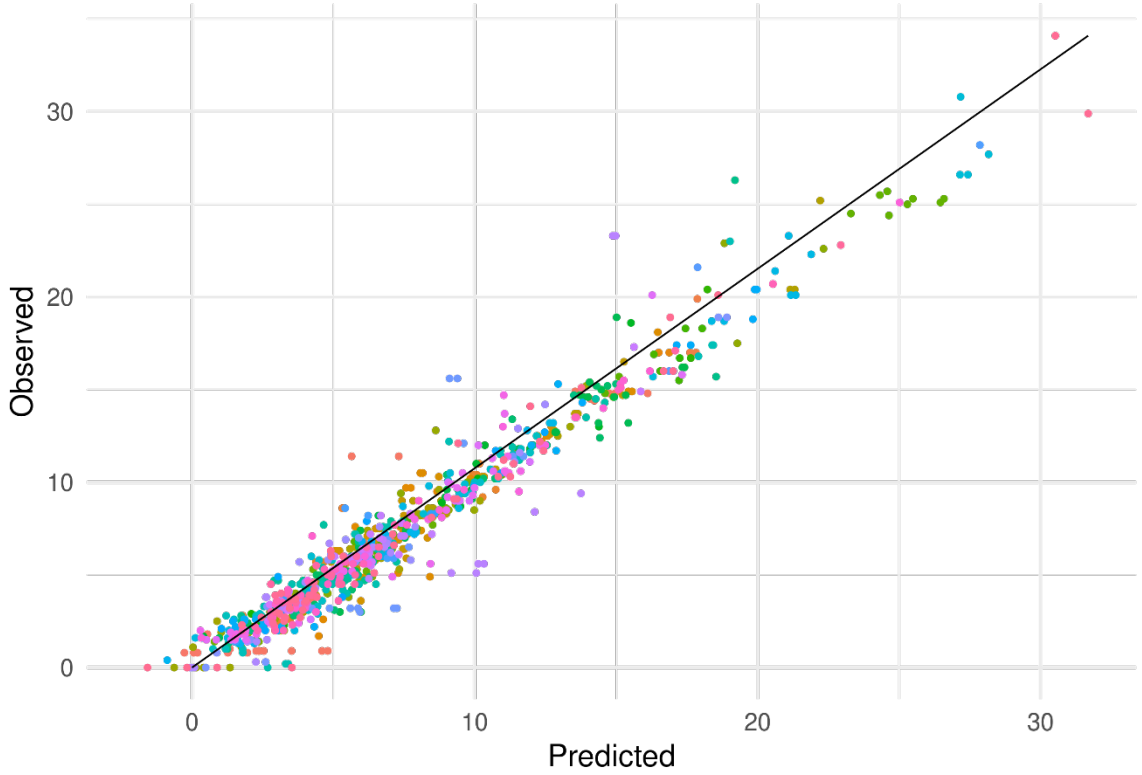
Теория GAM

Выбор модели

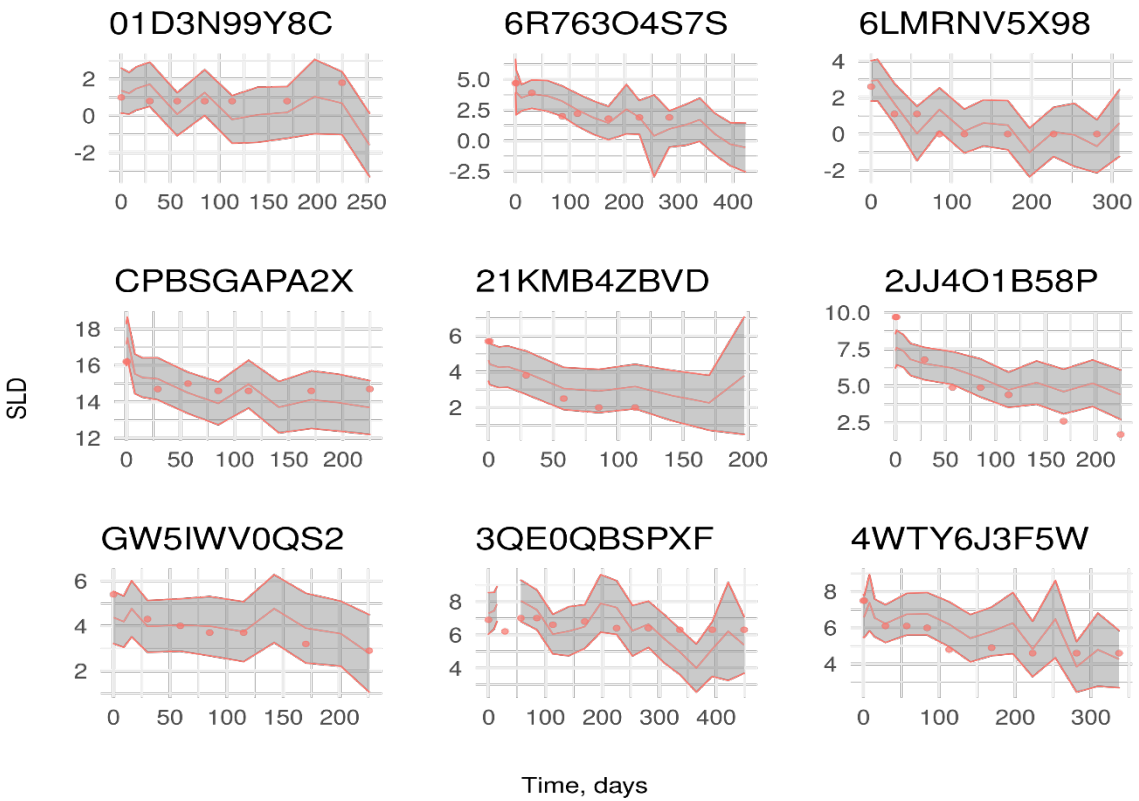
Результаты

Результаты на обучающей выборке

Observed vs Predicted



Individual time profiles



Модель адекватно описывает динамику размера опухоли для обучающей выборки и может быть использована для предсказания на тестовой выборке

Линейные компоненты модели

Финальная модель:

```
SLD ~ TIME + SLDb + s(USUBJID, bs = "re", k = 400) + ti(AST,
    TIME, bs = "tp", k = 10) + ti(CREAT, TIME, bs = "tp", k = 10) +
    te(LDH, TIME, bs = "tp", k = 10) + ti(NLR, TIME, bs = "tp",
    k = 10) + SMKSTAT + EGFRMUTN
```

Линейные компоненты модели:

- Время (TIME)
- Начальный размер опухоли (SLDb)
- Статус курения (SMKSTAT)
- Статус мутации (EGFRMUTN)

```
Parametric coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.881652   0.277682  -3.175  0.00157 **
TIME        -0.006025   0.001172  -5.142 3.60e-07 ***
SLDb         0.977603   0.017952  54.458  < 2e-16 ***
SMKSTAT2     1.212970   0.254350   4.769 2.29e-06 ***
EGFRMUTN1   -1.053460   0.471049  -2.236  0.02567 *
EGFRMUTN2    0.448822   0.230290   1.949  0.05174 .
```

Нелинейные компоненты модели

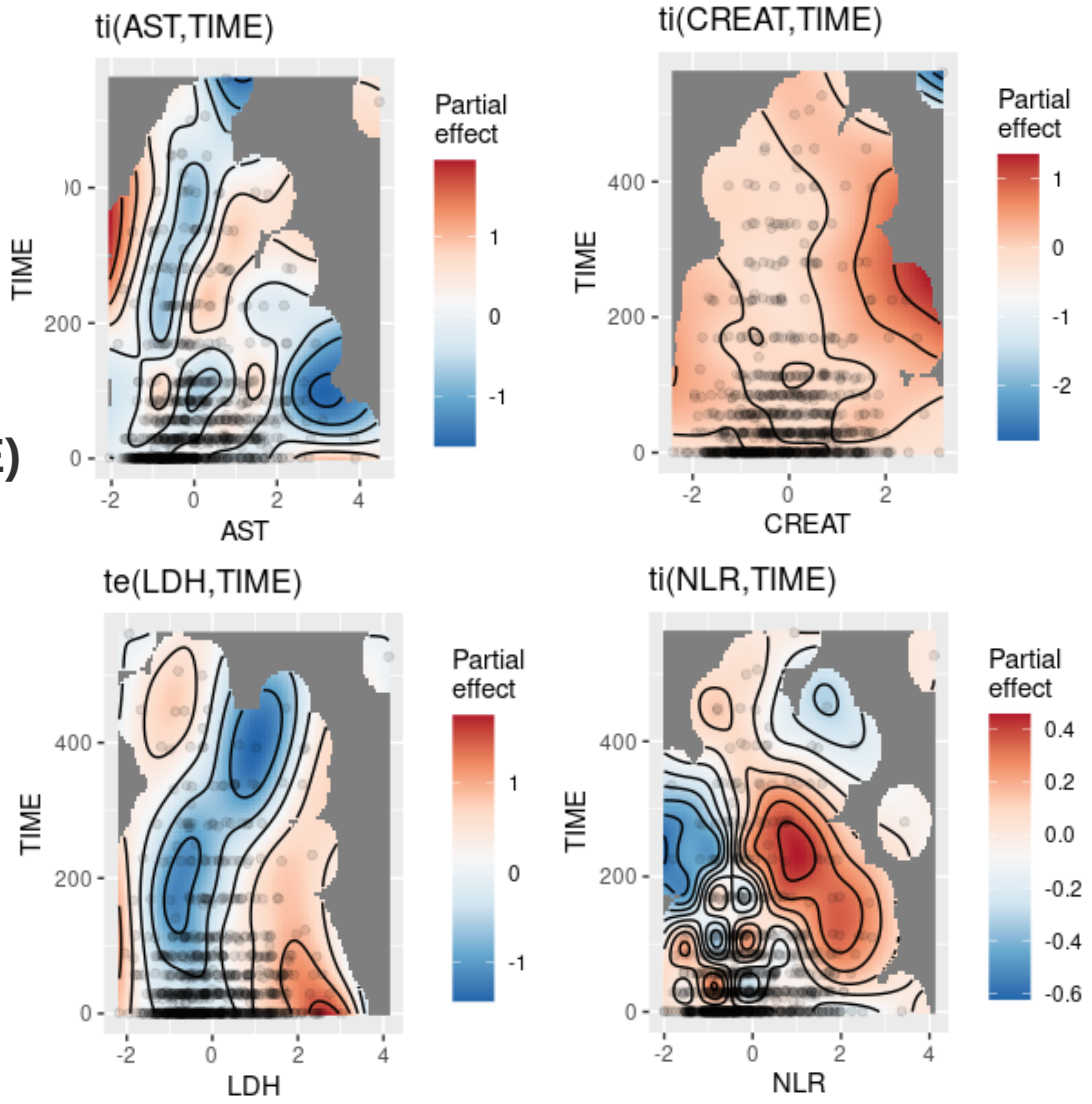
```
SLD ~ TIME + SLDb + s(USUBJID, bs = "re", k = 400) + ti(AST,
    TIME, bs = "tp", k = 10) + ti(CREAT, TIME, bs = "tp", k = 10) +
    te(LDH, TIME, bs = "tp", k = 10) + ti(NLR, TIME, bs = "tp",
    k = 10) + SMKSTAT + EGFRMUTN
```

Нелинейные компоненты модели:

- **динамика AST** **ti(AST, TIME)**
- **динамика креатинина** **ti(CREAT, TIME)**
- **динамика отношения клеток крови** **ti(NLR, TIME)**
- **полное тензорное произведение LDH te(LDH, TIME)**

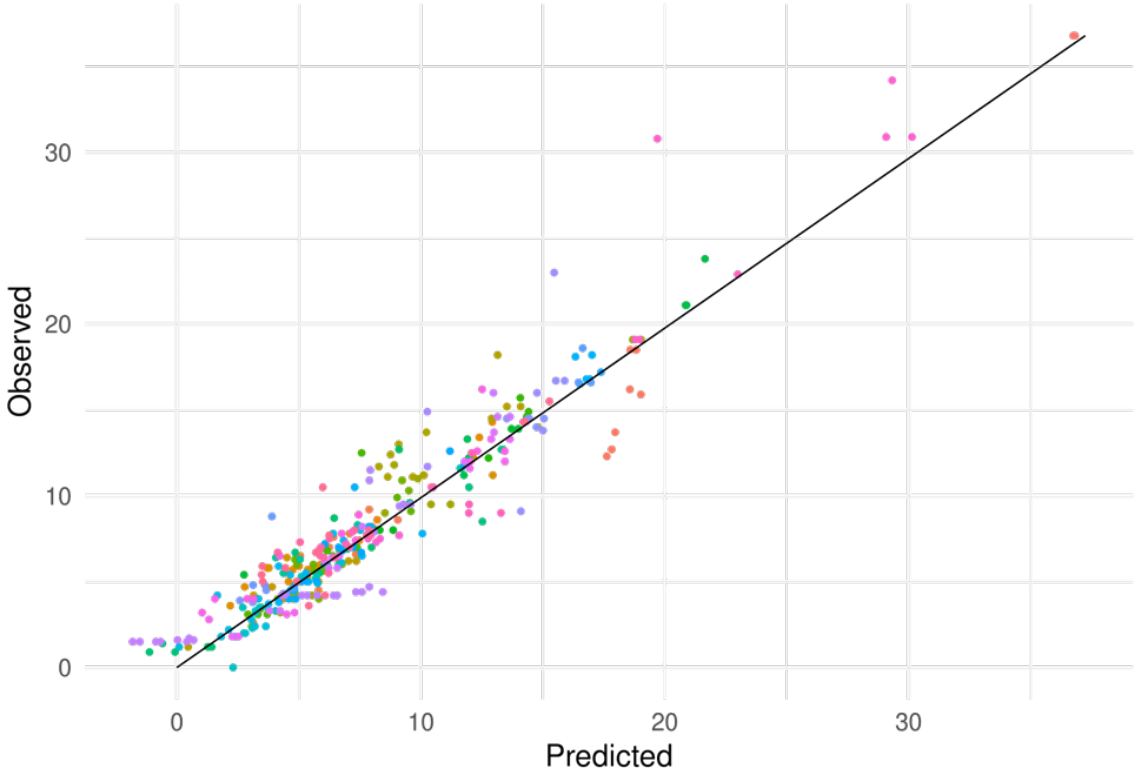
```
Approximate significance of smooth terms:
                   edf Ref.df     F  p-value
s(USUBJID)     162.866    255 2.919  < 2e-16 ***
ti(AST,TIME)    18.232     81 0.937 2.45e-05 ***
ti(CREAT,TIME)   8.770     81 0.272   0.0434 *
te(LDH,TIME)    17.813     88 1.667  < 2e-16 ***
ti(NLR,TIME)     5.814     69 0.236   0.0279 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

R-sq.(adj) =  0.933   Deviance explained =   95%
-REML = 1758.4  Scale est. = 2.125     n = 865
```

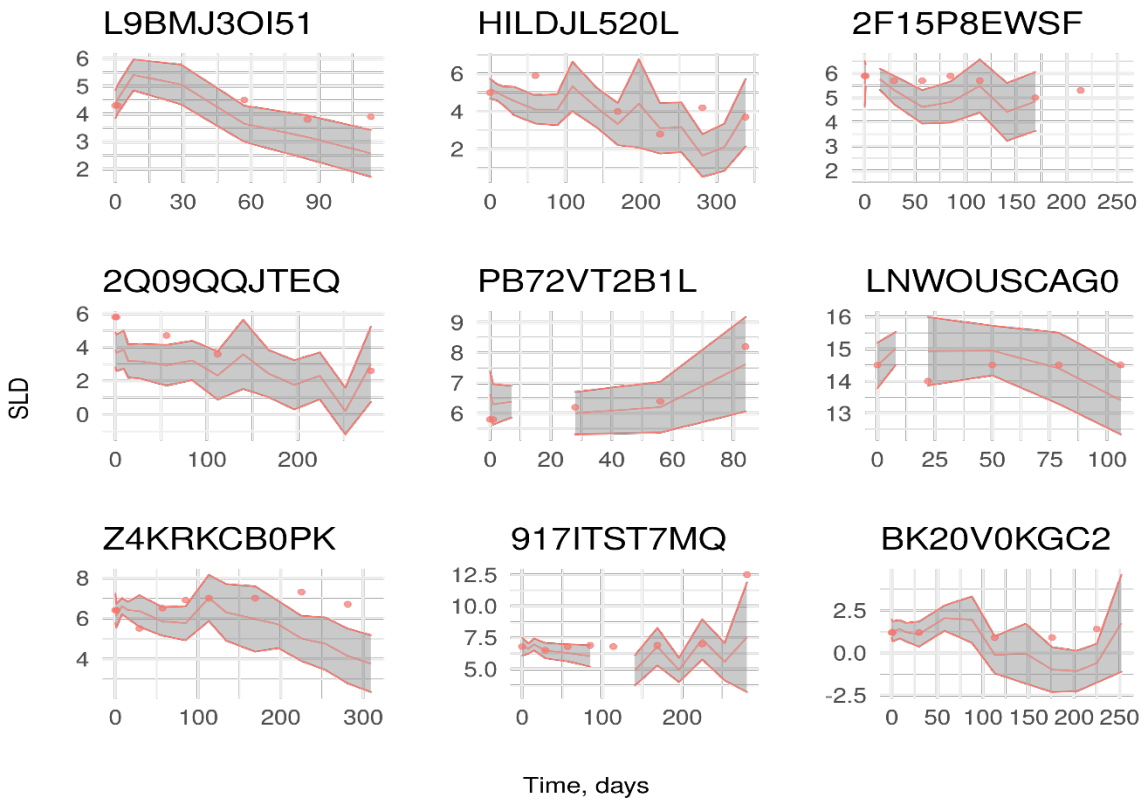


Валидация модели на тестовой выборке

Observed vs Predicted



Individual time profiles



Модель адекватно предсказывает динамику размера опухоли для тестовой выборки, используя данные о динамике других биомаркеров, времени и начальном значении SLD

Выводы

- Использование GAM подхода к моделированию позволило проанализировать факторы, позволяющие предсказывать динамику размера опухоли у пациентов с немелкоклеточным раком легкого

- В результате анализа выявлено, что для прогнозирования динамики размера опухоли значимыми оказались: начальный размер опухоли, статус курения и EGFR мутации, динамика изменений AST (повреждение печени), креатинина (повреждение почек), лактат дегидрогеназы (LDH, рост опухоли), а также изменения в отношении количества нейтрофилов к числу лейкоцитов (NLR, воспалительный процесс)

- Полученные результаты могут быть использованы для восстановления пропущенных измерений размера опухоли по измерениям биомаркеров, а также для оптимизации дизайна клинических исследований

Спасибо за внимание!