Теоретические предпосылки

Определение: Многофакторным линейным регрессионным уравнением называется статистическая связь между зависимой переменной y и независимыми факторами (регрессорами) $x_1, x_2, \dots, x_k$, представленная в виде линейной зависимости:

$$y = b_0 + b_1 x_1 + b_2 x_2 + \dots + +b_k x_k + \varepsilon$$

Определение: Мультиколлинеарность – это наличие тесной линейной зависимости между независимыми факторами.

Признаки мультиколлинеарности:

1.Оценки параметров имеют большие стандартные ошибки, малую значимость, хотя регрессия в целом является значимой (завышены значения F-статистики, R^2).

2. Небольшое изменение исходных данных приводит к существенному изменению оценок параметров модели.

3. Оценки параметров имеют неправильные с точки зрения теории знаки или неоправданно большие значения.

На практике о наличии мультиколлинеарности судят по корреляционной матрице, состоящей из коэффициентов корреляции всех переменных, включенных в модель:

$$\begin{pmatrix} 1 & r_{y,x_1} & r_{y,x_2} & \dots & r_{y,x_k} \\ r_{x_1,y} & 1 & r_{x_1,x_2} & \dots & r_{x_1,x_k} \\ \dots & \cdots & \dots & \cdots & \dots \\ r_{x_{k-1},y} & r_{x_{k-1},x_1} & \dots & 1 & r_{x_{k-1},x_k} \\ r_{x_k,y} & r_{x_k,x_1} & r_{x_k,x_2} & \cdots & 1 \end{pmatrix}$$

Матрица является симметричной относительно главной диагонали, причем на главной диагонали матрице стоят единицы. Если $r_{x_i,x_j} > 0.5 (i \neq j)$, то мультиколлинеарность считается не установленной.

Наличие мультиколлинеарности приводит к получению неэффективных и несостоятельных оценок в уравнении регрессии.

Пути устранения мультиколлинеарности:

- исключение из модели одного или нескольких факторов (при этом следует учитывать значения t-статистики);
- преобразование факторов, при котором уменьшается корреляция между ними (например, переход от первоначальных данных к их разностям).
- Применение в качестве метода оценивания параметров регрессии метода главных компонент (МГК).
- Применение альтернативных методов оценивания коэффициентов: гребневой регрессии и метода Lasso.

VIF-критерий для определения мультиколлинеарности.

Для определения наличия мультиколлинеарности регрессионной модели используется VIF-критерий.

Находят его согласно алгоритму:

1. Строят линейную регрессию одного объясняющего фактора X_j на оставшиеся регрессоры (факторы, признаки) X_s, где $s \neq j$.
2. По полученной регрессии определяют коэффициент детерминации R^2.
3. Строят линейную регрессию одного объясняющего фактора X_j на оставшиеся регрессоры (факторы, признаки) X_s, где $s \neq j$.

4. Определяют VIF_j – критерий по формуле: $VIF_j = \frac{1}{1-R_j^2}$

5. Если $VIF_j > 10$, то мультиколлинеарность независимых регрессоров по фактору X_j точно есть.

Методы регрессии Ридж и Lasso осуществляют регуляризацию параметров и позволяют преодолеть некоторые недостатки метода наименьших квадратов.

Гребнева регрессия (ридж-оценки).

Гребневая регрессия полностью не устраняет проблему мультиколлинеарности, кроме того приводит к незначительно смещенным оценкам коэффициентам регрессии, но повышает надежность модели в целом. Как правило, мультиколлинарность может давать «неправильные» с точки зрения теории знаки в регрессии, гребневая регрессия призвана устранить этот недостаток.

Если $\hat{B} = (X^t X)^{-1} X^t Y$ – несмещенная оценка коэффициентов МНК, то Ридж-оценка неизвестного параметра B сводится к тому, что в МНК-оценки добавляем некоторый параметр регляризации λ, как правило этот параметр из интервала 0,1- 0,3.

В итоге получаем оценки: $\widetilde{B_\lambda} := (W^t W + \lambda I)^{-1} W^t V$, где I – единичная матрица, W – стандартизированная матрица для матрицы X, а V – стандартизированный вектор столбец Y.

Ридж-оценка является МНК-оценкой с ограничением нормы возможных решений. Иногда для получения оценок методом гребневой регрессии данные не стандартизуют:

$$\hat{b} = (X^T X + \lambda I)^{-1} X^T y$$

Метод LASSO

Метод Lasso не сильно отличается от метода гребневой регрессии, суть в том, что на параметр регуляризации накладывается ряд ограничений.

Метод регрессии «Лассо» (LASSO, Least Absolute Shrinkage and Selection Operator) заключается во введении дополнительного слагаемого регуляризации в функционал оптимизации модели, что часто позволяет получать более устойчивое решение.

В методе Lasso параметр регуляризации вводится в функцию потерь.

$$L(w) = \frac{1}{2} \sum_{i=1}^{N} (f(x_i, w) - y_i)^2 + \lambda \sum_{j=0}^{p} |w_j|$$

Результатом этого может стать снижение признакового пространства, так как при части факторов коэффициенты «зануляются» (при больших значениях λ). Это позволяет лучше интерпретировать результаты, полученные регрессией по методу Lasso.

Выбор параметра регуляризации.

Следует так подобрать λ, чтобы это позволило прогнозировать результат с наибольшей точностью.

Слишком малые значения λ могут приводить к переобучению. Слишком большие λ могут помешать определению основных зависимостей.

Выбирать параметр регуляризации можно итерационными методами, (перебирая λ с шагом 0,1), исходя из минимума ошибок MAE (средняя абсолютная ошибка) или RMSE (квадратный корень из среднеквадратичной ошибки). Как правило, λ выбирается из интервала [0,1; 0,3].

Работа в R Studio

Создадим рабочую область в R Studio в виде отчета через команду *new file – new R Markdown.*

Загрузим следующие пакеты (библиотеки) в R:

```
library(readxl)
library(caTools)
library(DescTools)
library(dplyr)
library(ggplot2)
library(glmnet)
library(tidyverse)
```

Загрузим набор данных.

```
gem <- read_excel("ВыборкаПоГемодиализу.xlsx")
```

Проведем преобразования исходных данных.

```
gem2 <- gem
%>% select(-'№', -`Гепатит B`, -`На ГД с`, -`Дата анализа`, -starts_with("Анализ"), -starts_with("Ед изм")) %>%
  mutate(Moch_delta = Значение2 - Значение3, Пол = as.factor(Пол)) %>%
  rename(Креатинин       = Значение,
        "Мочевина до"    = Значение2,
        "Мочевина после" = Значение3,
        Альбумин         = Значение4,
        ktv              = Значение5) %>%
  select(-"Мочевина до", -"Мочевина после")
```

Для оценки качества регресионной модели разделим выборку на тестовую и обучающую.

```
set.seed(1821)
split <- sample.split(gem2$Moch_delta, SplitRatio = 0.75)
train <- subset(gem2, split == TRUE)
test  <- subset(gem2, split == FALSE)

train_y <- train$Moch_delta
train_x <- train %>% select(-Moch_delta) %>% data.matrix()
test_x  <- test %>% select(-Moch_delta) %>% data.matrix()
```

Построим модель линейной регрессии (для оценки параметров применяется метод наименьших квадратов).

```
m_lm <- lm(Moch_delta ~ ., data = train)
summary(m_lm)
```

```
Call:
lm(formula = Moch_delta ~ ., data = train)

Residuals:
    Min      1Q  Median      3Q     Max 
-6.6462 -0.5791  0.0778  0.7670  4.5727 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.109e+02  1.202e+01  17.556   <2e-16 ***
ПолМужской   3.841e-01  5.397e-01   0.712   0.4792    
Вес          2.975e-04  1.652e-02   0.018   0.9857    
Возраст     -2.148e-03  1.907e-02  -0.113   0.9106    
`Гепатит С`  1.470e+00  1.165e+00   1.262   0.2112    
Креатинин   -2.550e-03  1.136e-03  -2.244   0.0282 *  
Альбумин    -4.371e+00  3.666e-01 -11.922   <2e-16 ***
ktv         -5.902e+00  4.250e+00  -1.389   0.1696    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.954 on 66 degrees of freedom
Multiple R-squared:  0.825,	Adjusted R-squared:  0.8065 
F-statistic: 44.46 on 7 and 66 DF,  p-value: < 2.2e-16
```

Построим прогноз по полученной модели на тестовой выборке (рис. 1, рис. 2).

test$predict <- predict(m_lm, test)

	Пол	Вес	Возраст	Гепатит С	Креатинин	Альбумин	ktv	Moch_delta	predict
1	Женский	39.0	30	0	1049.0	41.0	1.47	20.20	20.33126
2	Женский	41.0	35	0	729.0	41.0	1.47	20.20	21.13726
3	Женский	65.3	65	0	673.0	39.3	1.30	28.91	29.65686
4	Женский	102.0	61	0	675.0	40.7	1.45	23.46	22.66660
5	Женский	62.0	69	0	841.0	41.0	1.47	20.20	20.78482
6	Женский	63.0	33	0	699.0	42.0	1.40	15.70	17.26686
7	Мужской	79.0	69	0	849.0	39.3	1.30	28.91	29.58757
8	Мужской	61.0	47	0	1375.0	41.0	1.47	20.20	19.85394
9	Женский	54.0	43	0	1184.0	39.3	1.30	28.91	28.39747
10	Женский	96.4	56	0	1017.0	40.7	1.45	23.46	21.80342
11	Мужской	65.0	73	0	943.0	40.7	1.45	23.46	22.33041
12	Женский	61.5	58	0	1130.0	41.0	1.47	20.20	20.07122

Рис. 1 Тестовая выборка с добавленным прогнозным значением

ggplot(test, aes(x = Moch_delta, y = predict)) + geom_point() + geom_abline() +
scale_x_continuous(limits = c(0, 35), expand = c(0, 0)) +
scale_y_continuous(limits = c(0, 35), expand = c(0, 0))

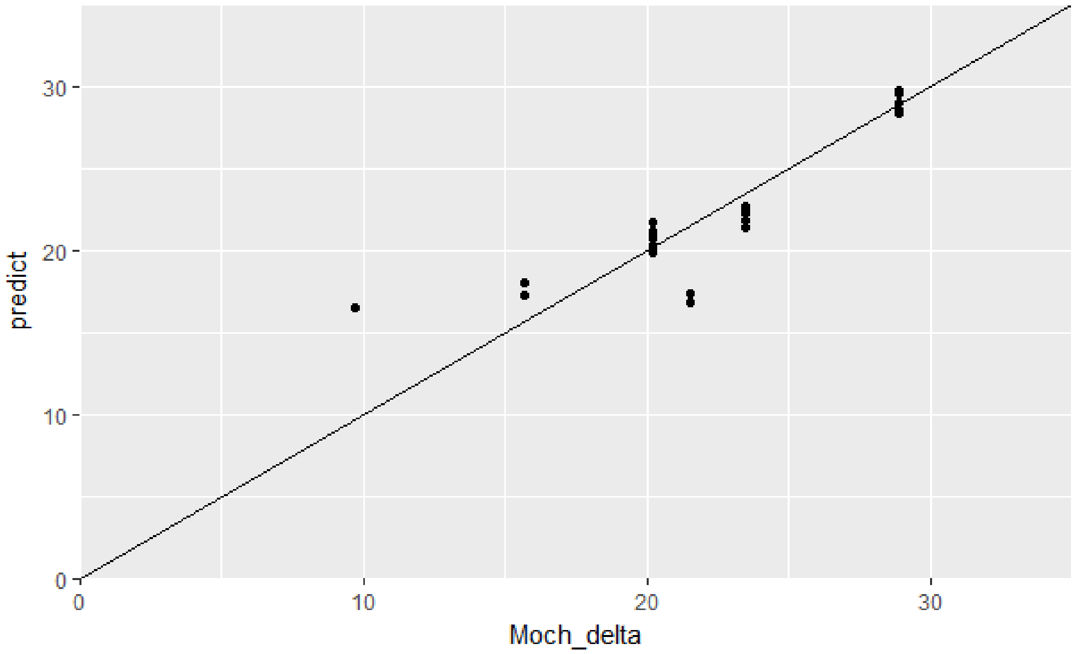

Рис. 2 График прогноза

Для анализа мультиколлениарности вычислим значения коэффициентов увеличения дисперсии VIF_j.

vif(m_lm)

```
     Пол          Вес      Возраст `Гепатит С`  Креатинин    Альбумин         ktv
1.395554     1.275359     1.431094     1.022846     1.506962     1.915328    1.982800
```

Несмотря на то, что мультиколлинеарность отсутствует, применим методы Ридж и Lasso для оценки параметров регрессии с целью сравнения моделей.

Оценим параметры с помощью метода Ридж. Для Ридж-регрессии различные значения λ будут генерировать различные наборы оценок параметров.

Сначала выберем некоторый диапазон значений λ, далее сгенерируем модель для всех λ из этого диапазона.

lambdas <- seq(0, 150, by = 0.025)

set.seed(1877)
cv_ridge <- cv.glmnet(train_x, train_y, alpha = 0, lambda = lambdas)
plot(cv_ridge)

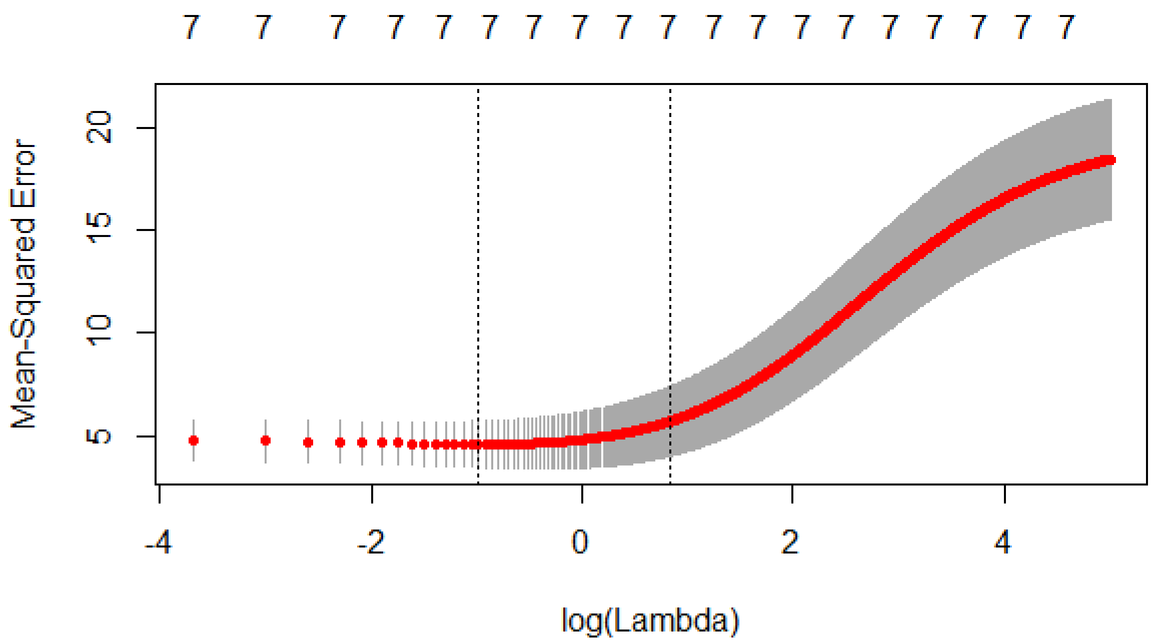

Рис. 3 Зависимость значения средней квадратичной ошибки от λ для модели, полученной методом Ридж

На графике показана зависимость средней квадратической ошибки предсказания MSE от λ. Одна из вертикальных линий пунктирных линий показывает положение минимума MSE. Вторая пунктирная линия обозначает точку, выбранную по «правилу одной стандартной ошибки».

Определим подходящие значения λ исходя из минимума ошибок.

```
cv_ridge$lambda.min
[1] 0.375
```

Построим модель линейной регрессии с помощью метода Ридж.

```
set.seed(1877)
m_ridge <- glmnet(train_x, train_y, alpha = 0, lambda = cv_ridge$lambda.min)
coef(m_ridge)

8 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept)  1.926001e+02
Пол          1.866634e-01
Вес         -1.340553e-03
Возраст      6.796165e-04
Гепатит С    1.386807e+00
Креатинин   -2.038765e-03
Альбумин    -3.824349e+00
ktv         -9.075734e+00
```

Построим прогноз для полученной модели на тестовой выборке (рис. 4).

```
test$ridge <- predict(m_ridge, s = cv_ridge$lambda.min, newx = test_x)
```

Filter

	Пол	Вес	Возраст	Гепатит С	Креатинин	Альбумин	ktv	Moch_delta	predict	ridge
1	Женский	39.0	30	0	1049.0	41.0	1.47	20.20	20.33126	20.47659
2	Женский	41.0	35	0	729.0	41.0	1.47	20.20	21.13726	21.12971
3	Женский	65.3	65	0	673.0	39.3	1.30	28.91	29.65686	29.27597
4	Женский	102.0	61	0	675.0	40.7	1.45	23.46	22.66660	22.50452
5	Женский	62.0	69	0	841.0	41.0	1.47	20.20	20.78482	20.89633
6	Женский	63.0	33	0	699.0	42.0	1.40	15.70	17.26686	17.97098
7	Мужской	79.0	69	0	849.0	39.3	1.30	28.91	29.58757	29.08816
8	Мужской	61.0	47	0	1375.0	41.0	1.47	20.20	19.85394	19.98068
9	Женский	54.0	43	0	1184.0	39.3	1.30	28.91	28.39747	28.23435
0	Женский	96.4	56	0	1017.0	40.7	1.45	23.46	21.80342	21.81137
1	Мужской	65.0	73	0	943.0	40.7	1.45	23.46	22.33041	22.20255

Showing 1 to 12 of 26 entries

Рис. 4 Тестовая выборка с добавленным прогнозным значением для модели, полученной методом Ридж

Оценим параметры модели с помощью метода Lasso.

set.seed(1886)
cv_lasso <- cv.glmnet(train_x, train_y, alpha = 1, lambda = lambdas)
plot(cv_lasso)

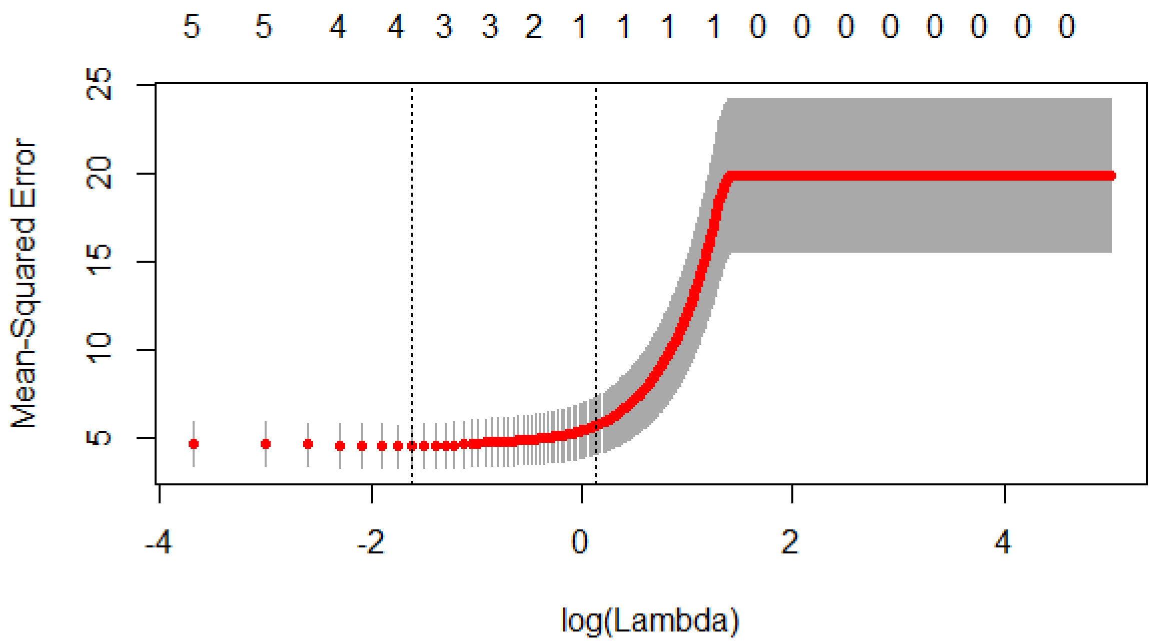

Рис. 5 Зависимость значения средней квадратичной ошибки от λ для модели, полученной методом Lasso

Определим подходящие значения λ исходя из минимума ошибок.

cv_lasso$lambda.min

```
[1] 0.2
```

Построим модель линейной регрессии с помощью метода Lasso.

set.seed(1886)
m_lasso <- glmnet(train_x, train_y, alpha = 1, lambda = cv_lasso$lambda.min)
coef(m_lasso)

```
8 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) 199.392696216
Пол           .
Вес           .
Возраст       .
Гепатит С     0.269588638
Креатинин    -0.001285386
Альбумин     -4.185544485
ktv          -3.826898074
```

Построим прогноз по полученной модели на тестовой выборке (рис. 6).

test$lasso <- predict(m_lasso, s = cv_lasso$lambda.min, newx = test_x)

	Вес	Возраст	Гепатит С	Креатинин	Альбумин	ktv	Moch_delta	predict	ridge	lasso
й	39.0	30	0	1049.0	41.0	1.47	20.20	20.33126	20.47659	20.81146
й	41.0	35	0	729.0	41.0	1.47	20.20	21.13726	21.12971	21.22279
й	65.3	65	0	673.0	39.3	1.30	28.91	29.65686	29.27597	29.06077
й	102.0	61	0	675.0	40.7	1.45	23.46	22.66660	22.50452	22.62440
й	62.0	69	0	841.0	41.0	1.47	20.20	20.78482	20.89633	21.07882
й	63.0	33	0	699.0	42.0	1.40	15.70	17.26686	17.97098	17.34369
й	79.0	69	0	849.0	39.3	1.30	28.91	29.58757	29.08816	28.83454
й	61.0	47	0	1375.0	41.0	1.47	20.20	19.85394	19.98068	20.39243
й	54.0	43	0	1184.0	39.3	1.30	28.91	28.39747	28.23435	28.40393
й	96.4	56	0	1017.0	40.7	1.45	23.46	21.80342	21.81137	22.18480
й	65.0	73	0	943.0	40.7	1.45	23.46	22.33041	22.20255	22.27991

Рис. 6 Тестовая выборка с добавленным прогнозным значением для модели, полученной методом Lasso

Рассчитаем метрики MAE (средняя абсолютная ошибка), RMSE (квадратный корень из среднеквадратичной ошибки), MAPE (средняя абсолютная ошибка в процентах) для моделей, полученных с использованием различных методов для оценки параметров.

MAPE(x = test$predict, ref = test$Moch_delta)
RMSE(x = test$predict, ref = test$Moch_delta)
MAE(x = test$predict, ref = test$Moch_delta)
MAPE(x = test$ridge, ref = test$Moch_delta)
RMSE(x = test$ridge, ref = test$Moch_delta)
MAE(x = test$ridge, ref = test$Moch_delta)
MAPE(x = test$lasso, ref = test$Moch_delta)
RMSE(x = test$lasso, ref = test$Moch_delta)
MAE(x = test$lasso, ref = test$Moch_delta)

Результаты сравнения метрик для различных моделей представлены в таблице.

	MAPE	RMSE	MAE
Метод МНК	0.07958247	2.070362	1.384027
Метод Ридж	0.08433556	2.151514	1.417138
Метод Lasso	0.07851048	2.052017	1.329962