Лабораторная работа № 4

Модели множественного выбора
ROC-анализ

Задание:

Базовый уровень:

1. Провести предварительный анализ исходных данных. Исключить аномальные наблюдения (если такие есть), заполнить пропуски (если они имеются). Провести корреляционный анализ независимых переменных, исключив переменные, значительно коррелирующие с другими переменными (>0,9).
2. Построить статистически значимую модель бинарной регрессии, оценив параметры методом максимального правдоподобия, применяя метод пошагового исключения, в которой все переменные будут статистически значимы. Подобрать функцию распределения, описывающую вероятность положительной альтернативы между нормальным распределением (пробит), логистическим (логит) и экстремальным (гомпит) на основе минимума информационных критериев.
3. Проверить качество отобранной модели, подтвердив его значениями коэффициентов R^2 МакФаддена, тестом отношения правдоподобия (LR-тестом), результатами теста Хосмера-Лемешоу и любым тестом на нормальность распределения остатков (например, Колмогорова-Смирнова или Бера-Жарка).
4. Рассчитать маржинальные эффекты и провести интерпретацию коэффициентов модели.
5. Оформить отчет о выполнении задания с приведением условия задачи, результатов решения и выводов.

В качестве информационных средств выполнения задания рекомендуется использовать Eviews, R.

Повышенный уровень: Проверка статистической значимости и условий ограничения на коэффициенты бинарной модели с помощью теста Вальда

Результатом выполнения задания является отчет по лабораторной работе № 4. К отчету предъявляются следующие требования:

1. Четкое формулирование поставленной цели исследования
2. Формулирование задач, решение которых необходимо для достижения поставленной цели.
3. Описание в виде пунктов, тех действий, которые требуются для решения поставленных задач. Все рисунки и таблицы последовательно нумеруются и описываются. Каждый пункт решения поставленных задач сопровождается анализом принятого решения. При проведении статистических тестов, обязательно выписывается нулевая и альтернативная гипотеза, формулируется принятие решения на обосновано выбранном уровне значимости, указывается критическая область отказа от нулевой гипотезы в пользу альтернативной.
4. В заключении выписывается отобранная адекватная модель с оцененными коэффициентами с указанием под оценками коэффициентов значений t-статистик в скобках или стандартных ошибок коэффициентов. Также приводятся значения маржинальных эффектов и дается их интерпретация.

Осваивается умение строить адекватные модели бинарной регрессии и проводить интерпретацию результатов моделирования на основе маржинальных эффектов влияния факторов на результат.

Теоретические предпосылки:

Построение регрессионных моделей с бинарной зависимой переменной

Цель бинарного регрессионного анализа— описание зависимости между объектом наблюдения (зависимой или результирующей переменной, имеющей только две неупорядоченные альтернативы) и факторами, воздействующими на него (независимыми переменными, предикторами, регрессорами), с тем чтобы построить модель, позволяющую по значениям регрессоров получить оценки значений зависимой переменной.

Применительно к анализу риска в медицине чаще всего используется метод бинарной логистической регрессии, когда исследуется зависимость дихотомической результирующей переменной (т.е. принимающей только два значения, например — это статус выживаемости, подразумевающий два класса: выживет или умрет) от переменных с любым типом шкалы (пол, возраст, наличие осложнений, инфаркт миокарда в анамнезе и др.).

Для оценки и построения модели риска применяются модели бинарного выбора – пробит, логит, гомпит.

1) Логит-модель.

Если бинарная модель имеет в качестве функции распределения функцию вида (1), то эта модель называется Логит-моделью.

Функция стандартного логистического распределения:

$$F(u) = \Lambda(u) = \frac{1}{1+e^{-u}} \qquad \textbf{(1)}$$

Для оценки параметров используется метод максимального правдоподобия.

2) Пробит-модель.

Если бинарная модель имеет в качестве функции распределения функцию вида (2), то эта модель называется Пробит-моделью.

Функция стандартного нормального распределения:

$$F(u) = \Phi(u) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2}\, dz \qquad \textbf{(2)}$$

Стандартное нормальное распределение подразумевает, что математическое ожидание равно M=0, а среднее квадратичное отклонение σ=1.

3) Гомпит-модель.

Если бинарная модель имеет в качестве функции распределения функцию вида (3), то эта модель называется экстрим-моделью или гомпит-моделью.

Функция экстремального (или Гомперца) распределения:

$$F(u) = E(u) = e^{-e^{-u}} \qquad \textbf{(3)}$$

Селекция моделей, проводится исходя из критериев Акайке, Шварца и Ханнана-Куинна, т.е. выбиралась модель, где наименьшие значения критериев.

Оценка качества модели (или мониторинг модели).

Если необходимо сравнить нескольких альтернативных моделей бинарного выбора с разным количеством объясняющих переменных, то, как и в случае обычных линейных моделей, сравнивать качество альтернативных моделей можно, опираясь на значения информационных критериев Акайке (4) и Шварца (5):

$$AC = \ln(\sigma^2) + \frac{2k}{n} \qquad \textbf{(4)}$$

$$SC = \ln(\sigma^2) + \frac{k \cdot \ln(n)}{n} \qquad (5)$$

а также информационного критерия Ханнана-Куинна (6):

$$HQ = \ln(\sigma^2) + 2\frac{k \cdot \ln(\ln(n))}{n} \qquad (6)$$

Здесь L_k – максимальное значение функции правдоподобия для k –й из альтернативных моделей, а p – количество объясняющих переменных в этой модели, n – общее число наблюдений ряда данных. При этом среди нескольких альтернативных моделей выбирается та, которая минимизирует значение статистики критерия.

Метод максимального правдоподобия или метод наибольшего правдоподобия в математической статистике — это метод оценивания неизвестного параметра путём максимизации функции правдоподобия.

Для оценки параметров бинарных моделей применяют метод максимального правдоподобия с функцией правдоподобия:

$$L = L(y_1, \dots, y_n) = \begin{cases} y_i = 0; P(y_i = 0) = 1 - P(y_i = 1) = 1 - F(x^T b) \\ y_i = 1; P(y_i = 1) = F(x^T b) \end{cases}$$

y_i рассмотрим как n случайных величин Y_i с одним возможным значением y_i. Эти случайные величины независимы. Их совместная вероятность = произведению их вероятности:

$$L = \prod_{y_i=0}(1 - F(x^T b)) \prod_{y_i=1} F(x^T b) = \prod_i (1 - F(x^T b))^{y_i} F(x^T b)^{1-y_i}$$

Прологарифмируем выражение. Логарифмическая функция правдоподобия имеет вид:

$$l = \ln L = \sum_{y_i=0} y_i \ln F(x_i^T b) + \sum_{y_i=1}(1 - y_i) \ln(1 - F(x_i^T b))$$

Функция правдоподобия в математической статистике — это совместное распределение выборки из параметрического распределения, рассматриваемое как функция параметра. Для нахождения максимума функции правдоподобия необходимо найти частные производные по параметрам и приравнять их к «0». Решаем дифференциальное **уравнение правдоподобия:**

$$\frac{\partial l}{\partial b} = 0 \text{ или } \sum_i \left(\frac{y_i f\left(x_i^T b\right)}{F\left(x_i^T b\right)} - \frac{(1 - y_i) f\left(x_i^T b\right)}{1 - F\left(x_i^T b\right)} \right) x_i = 0.$$

Гипотеза относительно значимости построенной модели бинарного выбора: тест отношения правдоподобия Likelihood ratio test (*LR*), высчитывается в статистике, которые сравниваются с табличным значением $\chi 2(n)$, где n – число степеней свобод, равное числу ограничений в гипотезе. Для LR-теста LR- статистика в случае значимости построенной модели близка к 1.

1) Показатели качества подгонки:

1.1) Псевдо коэффициент детерминации $R_{ps}^2 = 1 - \frac{1}{1 + \frac{2(l - \bar{l})}{n}}$,

где n– количество наблюдений,

l – логарифмическая функция правдоподобия,

l со штрихом – ограниченная логарифмическая функция правдоподобия, в которой все параметры кроме свободного члена равно нулю.

1.2) Коэффициент Макфаддена $\frac{R^2}{R^2_{MF}} = 1 - \frac{l}{\bar{l}}$.

Чем ближе показатели качества к единице, тем сильнее «объясняющая сила» модели.

Для проверки адекватности подобранной модели имеющимся данным имеется ряд статистических критериев согласия; одним из них является критерий Хосмера–Лемешоу.

Критерий согласия **Хосмера–Лемешоу** исследует расстояние между наблюдаемыми и ожидаемыми распределениями частот «плохих» и «хороших» заемщиков. Если уровень значимости является большим, то модель хорошо откалибрована и достаточно точно описывает реальные данные. Значение статистики Хосмера–Лемешова не должно быть меньше уровня значимости 0,05. Оптимальными считаются значения не меньше 0,5–0,6.

Интерпретация коэффициентов модели. Маржинальные эффекты.

Найденные коэффициенты модели множественного выбора достаточно сложно интерпретировать с практической точки зрения, т.к. они не объясняют предельный эффект влияния объясняющих факторов на зависимую переменную. В этом случае обычно используют предельные эффекты каждого фактора (маржинальные эффекты).

Предельный коэффициент каждого объясняющего фактора x_j , j=1,..,k является непрерывным и зависит от значения остальных факторов и определяется:

$$\frac{\partial P(y=1)}{\partial x} = b \cdot F'(x^T b) = b \cdot f(x^T b),$$

где f – плотность вероятности.

Для логит-модели:

$\frac{\partial P(y=1)}{\partial x} = b \cdot \Lambda'(x^T b) = b \cdot \lambda(x^T b)$, где $\lambda(u) = \frac{e^{-u}}{(1+e^{-u})^2}$.

Для пробит-модели:

$\frac{\partial P(y=1)}{\partial x} = b \cdot \Phi'(x^T b) = b \cdot \varphi(x^T b)$, где $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$.

Для гомпит-модели:

$\frac{\partial P(y=1)}{\partial x} = b \cdot E(x^T b) = b \cdot e^{-e^{-(x^T b)}} \cdot e^{-(x^T b)}$.

Направление изменений эффекта зависит только от знака коэффициента регрессии.

ROC-анализ

Для оценки качества построенной модели бинарного выбора используется ROC-кривая.

Для построения ROC-кривой необходимо выбрать оптимальный порог разделения на подгруппы по признаку. Для определения оптимального порога необходимо задать критерий его определения, т.к. в разных задачах существует своя оптимальная стратегия. Критериями выбора порога отсечения могут выступать: максимум чувствительности и специфичности (Se+Sp), баланс между чувствительностью и специфичностью (|Se-Sp|) и другие подходы.

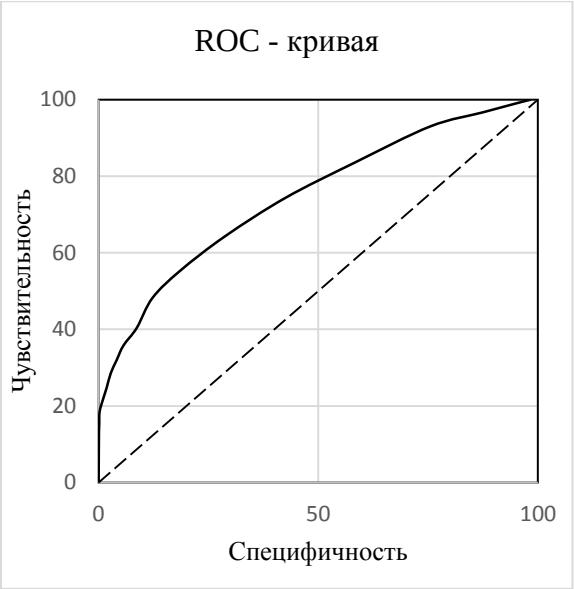

Рисунок 1 - ROC - кривая

Возможны ошибки I и II рода (таблица 1):

- Ошибка I рода имеет место тогда, когда отвергается правильная гипотеза H_0;
- Ошибка II рода – принимается неправильная гипотеза H_0.

Таблица 1

Ошибки I и II рода

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (Ошибка II рода)
	H_1	H_0 неверно отвергнута (Ошибка I рода)	H_0 верно отвергнута

Для понимания сути ошибок I и II рода необходимо рассмотреть четырехпольную таблицу сопряженности (таблица 2), которая строится на основе результатов классификации моделью и фактической (объективной) принадлежностью примеров к классам.

Таблица 2

Таблица сопряженности

	Фактически	
Модель	Положительно	Отрицательно
Положительно	TP	FP
Отрицательно	FN	TN

- TP (True Positives) – верно классифицированные положительные примеры (так называемые истинно положительные случаи);
- TN (True Negatives) – верно классифицированные отрицательные примеры (истинно отрицательные случаи);
- FN (False Negatives) – положительные примеры, классифицированные как отрицательные (ошибка I рода). Это так называемый "ложный пропуск" – когда

интересующее нас событие ошибочно не обнаруживается (ложно отрицательные примеры);

- FP (False Positives) – отрицательные примеры, классифицированные как положительные (ошибка II рода); Это ложное обнаружение, т.к. при отсутствии события ошибочно выносится решение о его присутствии (ложно положительные случаи).

Что является положительным событием, а что – отрицательным, зависит от конкретной задачи.

При анализе чаще оперируют не абсолютными показателями, а относительными – долями (rates), выраженными в процентах:

- Доля истинно положительных примеров (True Positives Rate):

$$TPR = \frac{TP}{TP + FN} \cdot 100\%$$

- Доля ложно положительных примеров (False Positives Rate):

$$FPR = \frac{FP}{TN + FP} \cdot 100\%$$

Введем еще два определения: чувствительность и специфичность модели. Ими определяется объективная ценность любого бинарного классификатора.

Чувствительность (Sensitivity) – это и есть доля истинно положительных случаев:

$$Se = TPR = \frac{TP}{TP + FN} \cdot 100\%$$

Специфичность (Specificity) – доля истинно отрицательных случаев, которые были правильно идентифицированы моделью:

$$Sp = \frac{TN}{TN + FP} \cdot 100\% = 100 - FPR$$

Модель с высокой чувствительностью часто дает истинный результат при наличии положительного исхода (обнаруживает положительные примеры). Наоборот, модель с высокой специфичностью чаще дает истинный результат при наличии отрицательного исхода (обнаруживает отрицательные примеры).

ROC-кривая получается следующим образом:

- Для каждого значения порога отсечения, которое меняется от 0 до 1 с шагом *dx* (например, 0.01) рассчитываются значения чувствительности *Se* и специфичности *Sp*. В качестве альтернативы порогом может являться каждое последующее значение примера в выборке.

- Строится график зависимости: по оси *Y* откладывается чувствительность *Se*, по оси X – 100%–*Sp* (сто процентов минус специфичность), или, что то же самое, *FPR* – доля ложно положительных случаев.

В результате вырисовывается некоторая кривая, график часто дополняют прямой *y=x*.

Визуальное сравнение кривых ROC не всегда позволяет выявить наиболее эффективную модель. Своеобразным методом сравнения ROC-кривых является оценка площади под кривыми. Теоретически она изменяется от 0 до 1, но, поскольку модель всегда характеризуются кривой, расположенной выше положительной диагонали, то обычно говорят об изменениях от 0,5 ("бесполезный" классификатор) до 1 ("идеальная" модель). Эта оценка может быть получена непосредственно вычислением площади под многогранником, ограниченным справа и снизу осями координат и слева вверху –

экспериментально полученными точками. Численный показатель площади под кривой называется AUC (Area Under Curve):

$$AUC = \int f(x)dx = \sum_i \left[\frac{x_{i+1} + x_i}{2}\right] \cdot (y_{i+1} - y_i)$$

С большими допущениями можно считать, что чем больше показатель AUC, тем лучшей прогностической силой обладает модель. Однако следует знать, что:

- показатель AUC предназначен скорее для сравнительного анализа нескольких моделей;

- AUC не содержит никакой информации о чувствительности и специфичности модели.

В литературе иногда приводится следующая экспертная шкала для значений AUC, по которой можно судить о качестве модели (таблица 3):

Таблица 3

Экспертная шкала AUC

Интервал AUC	**Качество модели**
0.9-1.0	Отличное
0.8-0.9	Очень хорошее
0.7-0.8	Хорошее
0.6-0.7	Среднее
0.5-0.6	Неудовлетворительно

Коэффициент Джини — статистический показатель степени расслоения общества данной страны или региона по отношению к какому-либо изучаемому признаку.

Коэффициент Джини изменяется от 0 до 1. Чем больше его значение отклоняется от нуля и приближается к единице, тем в большей степени доходы сконцентрированы в руках отдельных групп населения.

Показатель оценки коэффициент Джини иногда используется как показатель оценки, альтернативный AUC; эти две меры тесно связаны. Коэффициент Джини вычисляется как площадь между кривой ROC и диагональю. Коэффициент Джини всегда находится между 0 и 1, и чем он больше, тем лучше классификатор. При маловероятном условии, что кривая ROC находится ниже диагонали, коэффициент Джини будет отрицательным.

$$Коэф.\,Джини = 2 * AUC * (AUC - 0{,}5)$$

Порядок выполнения работы в R Studio:

Исходные данные: ответы граждан США на анкеты после выборов 2017 года.

Источник: https://www.kaggle.com/daliaresearch/trump-effect

Выбранные поля для анализа:

- Кандидат, за которого проголосовал гражданин (1 – Дональд Трамп, 0 – Хиллари Клинтон);
- Наличие высшего образования (1 – есть высшее образование, 0 – нет высшего образования);
- Политические предпочтения (2 – правые, 1 – центристская позиция, 0 – левые);
- Пол (1 – М, 0 – Ж);
- Возраст;

• Позиция по поводу внешней политики США (1 – положительное отношение к изоляционной политике, 0 – отрицательное).

Зависимая бинарная переменная – голос за кандидата. Цель исследования – выяснить, какие из перечисленных выше факторов влияют на выбор кандидата у голосовавших. Если есть связь – установить ее характер и силу.

Clinton - 0 Trump - 1	Education	Political view (2 -	Gender (1 - male, 0 -	Age	Isolation (1 - yes, 0
1	1	2	1	46	1
1	0	2	1	49	1
1	0	2	0	40	1
0	1	0	1	40	0
0	0	0	1	62	1
1	0	2	1	38	1
0	0	2	0	19	0
0	1	0	1	44	0
1	0	0	1	40	0

Рисунок 2 – Фрагмент исходной таблицы данных

Шаг 1. Создадим рабочую область в R Studio в виде отчета через команду *new file – new R Markdown.*

Шаг 2. Загрузим следующие пакеты (библиотеки) в R:

```
library(ggplot2)
library(memisc)
library(DescTools)
library(lmtest)
library(caTools)
library(dplyr)
library(readxl)
library(knitr)
library(kernlab)
library(caret)
library(mfx)
library(pROC)
library(ResourceSelection)
library(ROCR)
library(nortest)
```

Шаг 3. Загрузим набор данных.

```
d<- read_excel("Data_lab4.xlsx")
```

Шаг 4. Проведем предварительный анализ исходных данных.

Корреляционный анализ

Для проведения корреляционного анализа между независимыми переменными попарно используем коэффициент Спирмена.

cor.test (d$ Education, d$PolView, method = "spearman")
cor.test (d$ Education, d$ Gender, method = "spearman")
cor.test (d$ Education, d$ Age, method = "spearman")
cor.test (d$ Education, d$ Isolation, method = "spearman")

---

cor.test (d$ Age, d$ Isolation, method = "spearman")

Необходимо исключить переменные, значительно коррелирующие с другими переменными, при условии, что коэффициент корреляции >0,9 и является значимым.

Шаг 5. Преобразуем зависимую переменную как целочисленную.

d$Choice <- as.integer(d$Choice)

Шаг 5. Разделим выборку на тестовую и обучающую.

set.seed(1)
split <- sample.split(d$Choice, SplitRatio = 0.7)
train <- subset(d, split == TRUE)
test <- subset(d, split == FALSE)

Шаг 6. Построим регрессионные модели (пробит (**Probit**), логит (**Logit**) или гомпит (**Extreme Value)**) с бинарной зависимой переменной. В качестве зависимой переменной выступает – *Choice,* все остальные являются независимыми.

<u>Логит-модель</u>

```
model_1  <- glm(Choice ~ Education + PolView + Gender + Age +
            Isolation,
            train,
            family =binomial(link = "logit"))
summary(model_1)
```

```
Call:
glm(formula = Choice ~ Education + PolView + Gender + Age + Isolation,
    family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9870  -0.9273   0.4987   0.8283   2.1050

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.172152   0.424811  -5.113 3.17e-07 ***
Education   -0.488071   0.237321  -2.057   0.0397 *
PolView      0.511337   0.114703   4.458 8.28e-06 ***
Gender       0.357003   0.231029   1.545   0.1223
Age          0.020754   0.008605   2.412   0.0159 *
Isolation    1.662383   0.233669   7.114 1.13e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 565.13  on 407  degrees of freedom
Residual deviance: 454.67  on 402  degrees of freedom
AIC: 466.67

Number of Fisher Scoring iterations: 4
```

Пробит – модель

model_2 <- glm(Choice ~ Education + PolView + Gender + Age +
Isolation,
train,
family =binomial(link = "probit"))
summary(model_2)

```
Call:
glm(formula = Choice ~ Education + PolView + Gender + Age + Isolation,
    family = binomial(link = "probit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0155  -0.9295   0.4764   0.8323   2.1472

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.312814   0.248107  -5.291 1.21e-07 ***
Education   -0.312621   0.141179  -2.214   0.0268 *
PolView      0.308998   0.068272   4.526 6.01e-06 ***
Gender       0.221535   0.137376   1.613   0.1068
Age          0.012680   0.005114   2.480   0.0132 *
Isolation    1.010811   0.139643   7.239 4.54e-13 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 565.13  on 407  degrees of freedom
Residual deviance: 453.98  on 402  degrees of freedom
AIC: 465.98

Number of Fisher Scoring iterations: 4
```

Гомпит – модель

model_3 <- glm(Choice ~ Education + PolView + Gender + Age +
Isolation,
train,
family =binomial(link = "cloglog"))
summary(model_3)

```
Call:
glm(formula = Choice ~ Education + PolView + Gender + Age + Isolation,
    family = binomial(link = "cloglog"), data = train)

Deviance Residuals:
```

```
    Min       1Q   Median       3Q      Max
-2.1296  -0.8924   0.3909   0.8402   1.9959

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.991228   0.293882  -6.776 1.24e-11 ***
Education   -0.327546   0.156866  -2.088  0.03679 *
PolView      0.346187   0.077095   4.490 7.11e-06 ***
Gender       0.235792   0.150646   1.565  0.11754
Age          0.014791   0.005636   2.624  0.00868 **
Isolation    1.201345   0.169809   7.075 1.50e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 565.13  on 407  degrees of freedom
Residual deviance: 452.93  on 402  degrees of freedom
AIC: 464.93

Number of Fisher Scoring iterations: 6
```

Необходимо построить все три типа: логит-, пробит-, гомпит- модели. В моделях должны остаться только статистически значимые переменные (см. столбец **Pr(>|z|)**). Значения должны быть ≤α (α – уровень значимости или вероятность ошибки отклонения нулевой гипотезы о том, что коэффициент при данном факторе равен нулю, принимается как правило равным 0,05-0,1). Если значения *Pr*≥α , то данный фактор является не значимым и не оказывает влияния на зависимую переменную, следовательно этот фактор должен быть исключен из модели. Таким образом, на каждом этапе исключается переменная с наибольшим значением *Pr* и модель строится заново до тех пор, пока в модели не останутся только статистически значимые факторы.

В конечном итоге, остаются следующие логит-, пробит-, гомпит- модели.

Логит - модель

```
Call:
glm(formula = Choice ~ Education + PolView + Age + Isolation,
    family = binomial(link = "logit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0238  -0.9614   0.5234   0.8341   2.0503

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.896510   0.418892  -4.527 5.97e-06 ***
Education   -0.563003   0.230965  -2.438   0.0148 *
PolView      0.601024   0.115086   5.222 1.77e-07 ***
Age          0.018074   0.008869   2.038   0.0416 *
Isolation    1.555934   0.229300   6.786 1.16e-11 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 565.13  on 407  degrees of freedom
Residual deviance: 457.01  on 403  degrees of freedom
AIC: 467.01
```

```
Number of Fisher Scoring iterations: 4
```

Пробит - модель

```
Call:
glm(formula = Choice ~ Education + PolView + Age + Isolation,
    family = binomial(link = "probit"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0572  -0.9649   0.5043   0.8404   2.0806

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.141381   0.246326  -4.634 3.59e-06 ***
Education   -0.355874   0.137690  -2.585  0.00975 **
PolView      0.365284   0.068564   5.328 9.95e-08 ***
Age          0.010962   0.005282   2.075  0.03796 *
Isolation    0.947541   0.137035   6.915 4.69e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 565.13  on 407  degrees of freedom
Residual deviance: 456.28  on 403  degrees of freedom
AIC: 466.28

Number of Fisher Scoring iterations: 4
```

Гомпит - модель

```
Call:
glm(formula = Choice ~ Education + PolView + Age + Isolation,
    family = binomial(link = "cloglog"), data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2192  -0.9437   0.4185   0.8340   1.9607

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.836787   0.292766  -6.274 3.52e-10 ***
Education   -0.359259   0.152404  -2.357   0.0184 *
PolView      0.435026   0.080185   5.425 5.79e-08 ***
Age          0.013035   0.005844   2.230   0.0257 *
Isolation    1.111857   0.161997   6.863 6.72e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 565.13  on 407  degrees of freedom
Residual deviance: 454.27  on 403  degrees of freedom
AIC: 464.27

Number of Fisher Scoring iterations: 5
```

Шаг 7. Выбор среди моделей логит-, пробит-, гомпит- осуществляется по информационным критериям Акайке (AIC) и Шварца (BIC) - выбирается модель, имеющая наименьшие значения критериев.

Для сравнения трех моделей по информационным критериям воспользуемся следующей командой.

mtable(model_1, model_2, model_3)

```
Calls:
model_1: glm(formula = Choice ~ Education + PolView + Age + Isolation,
    family = binomial(link = "logit"), data = train)
model_2: glm(formula = Choice ~ Education + PolView + Age + Isolation,
    family = binomial(link = "probit"), data = train)
model_3: glm(formula = Choice ~ Education + PolView + Age + Isolation,
    family = binomial(link = "cloglog"), data = train)

=============================================================
                          model_1      model_2      model_3
-------------------------------------------------------------
  (Intercept)             -1.897***    -1.141***    -1.837***
                          (0.419)      (0.246)      (0.293)
  Education               -0.563*      -0.356**     -0.359*
                          (0.231)      (0.138)      (0.152)
  PolView                  0.601***     0.365***     0.435***
                          (0.115)      (0.069)      (0.080)
  Age                      0.018*       0.011*       0.013*
                          (0.009)      (0.005)      (0.006)
  Isolation                1.556***     0.948***     1.112***
                          (0.229)      (0.137)      (0.162)
-------------------------------------------------------------
  Aldrich-Nelson R-sq.     0.209        0.211        0.214
  McFadden R-sq.           0.191        0.193        0.196
  Cox-Snell R-sq.          0.233        0.234        0.238
  Nagelkerke R-sq.         0.311        0.312        0.317
  phi                      1.000        1.000        1.000
  Likelihood-ratio       108.118      108.849      110.860
  p                        0.000        0.000        0.000
  Log-likelihood        -228.505     -228.139     -227.134
  Deviance               457.009      456.278      454.268
  AIC                    467.009      466.278      464.268
  BIC                    487.066      486.335      484.324
  N                      408          408          408
=============================================================
```

Наименьшие информационные критерии у гомпит-модели.

Шаг 8. При построении моделей не следует забывать обращать внимание на показатели качества модели. Необходимо построить качественную статистически значимую модель. Для этого следует рассматривать **LR**-статистику и соответствующую ей вероятность ошибки отклонения нулевой гипотезы, о том что в модели все коэффициенты равны нулю.

Оценка качества построенной модели проводится на основании коэффициента R^2 Мак-Фаддена, Prob (LR statistic), log likelihood и теста Хосмера-Лемешоу.

1) Коэффициент детерминации R^2 Макфаддена (McFadden R-sq).

Коэффициент показывает, насколько изменения зависимой переменной (в процентах) объясняются изменениями совокупности независимых переменных. То есть это доля дисперсии зависимой переменной (признака), объясняемая влиянием независимых переменных (предикторов). Если значение R^2 близко к единице, это означает, что построенная модель объясняет почти всю изменчивость зависимой

переменной изменчивостью предикторов. И наоборот, значение R^2, близкое к нулю, означает, что колебания зависимой переменной не обусловлены колебаниями предикторов.

Если коэффициент близок хотя бы к 60%, это уже хорошо (оказывают влияние или нет, но при этом могут не объяснять).

2) Критическая статистика для теста отношения правдоподобия p(Likelihood-ratio).

Если p(Likelihood-ratio) мала и меньше уровня значимости α, модель является значимой.

3) Критерий правдоподобия (Log-likelihood)

Логарифмическое правдоподобие показывает, насколько хорошо модель соответствует исходным данным. Снижение его величины означает улучшение качества модели (чем меньше значение, тем выше качество модели).

4) Тест Хосмера-Лемешоу

Для проведения теста воспользуемся следующей функцией

```
Preds1 <- predict(model_1, test, type = 'response')
h <- hoslem.test(test$Choice, Preds1, g=10) #g - число интервалов, используемых для вычисления квантилей
```

```
        Hosmer and Lemeshow goodness of fit (GOF) test

data:  test$Choice, Preds1
X-squared = 146, df = 8, p-value < 2.2e-16
```

```
Preds2 <- predict(model_2, test, type = 'response')
h <- hoslem.test(test$Choice, Preds2, g=10)
Hosmer and Lemeshow goodness of fit (GOF) test

data:  test$Choice, Preds1
X-squared = 146, df = 8, p-value < 2.2e-16
```

```
Preds3 <- predict(model_3, test, type = 'response')
h <- hoslem.test(test$Choice, Preds3, g=10)
Hosmer and Lemeshow goodness of fit (GOF) test

data:  test$Choice, Preds3
X-squared = 146, df = 8, p-value < 2.2e-16
```

Тест определяет степень соответствия между оцененными вероятностями, спрогнозированными моделью, и реальными вероятностями. Тем самым калибровочный тест модели позволяет установить, насколько хорошо построенная модель согласуется с исходными данными и может быть измерена с помощью критерия согласия модели.

Критерий согласия Хосмера–Лемешоу, исследует расстояние между наблюдаемыми и ожидаемыми распределениями частот объектов. Если уровень значимости является большим, то модель хорошо откалибрована и достаточно точно описывает реальные данные. Значение статистики Хосмера–Лемешоу не должно быть меньше уровня значимости 0,05. Оптимальными считаются значения не меньше 0,5–0,6.

Шаг 9. Оценку построенной модели можно также провести на основании графика исходных данных, смоделированных данных и остатков модели. Для этого рассмотрим следующие тесты.

Рассмотрим графические тесты:

qqnorm(model_3$residuals)

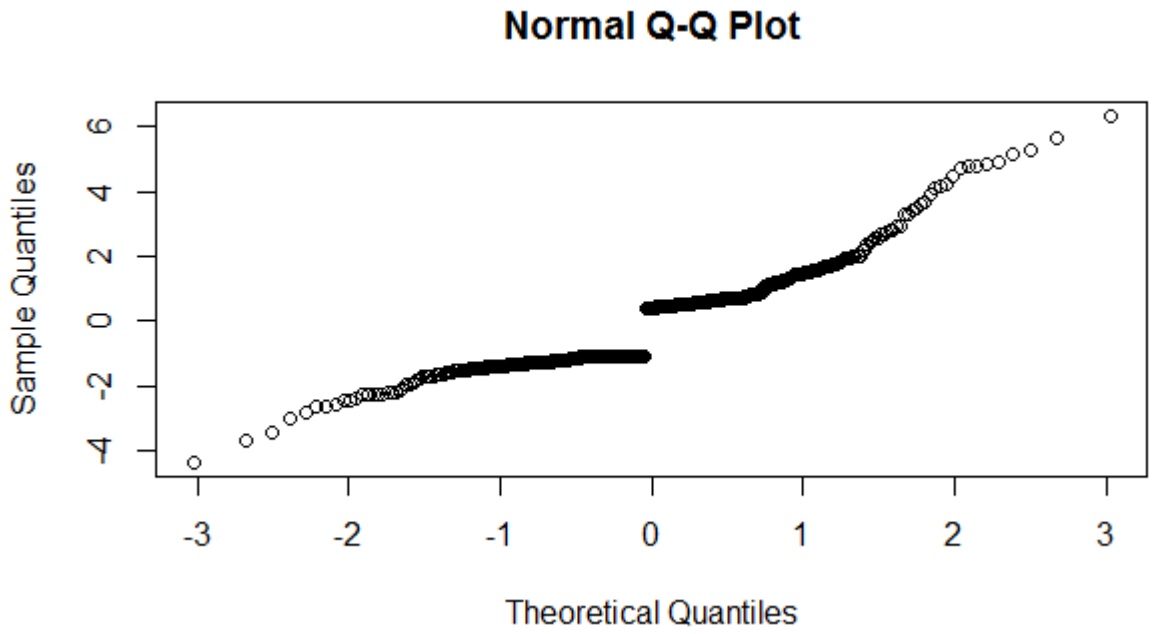

Рисунок 3

Рассмотрим параметрические тесты:

library(nortest)

lillie.test(model_3$residuals)

```
Lilliefors (Kolmogorov-Smirnov) normality test

data:  model_3$residuals
D = 0.22572, p-value < 2.2e-16
```

Поскольку значение p-value меньше 0.05 нулевая гипотеза о согласие распределения остатков с нормальным законом распределения отвергается.

Шаг 10. Для выбранной гомпит-модели интерпретация результатов моделирования проводится на основе маржинальных эффектов. Для расчета маржинального эффекта для гомпит-модели необходимо воспользоваться формулами из лекций.

Для расчета маржинального эффекта в логит-модели используется следующая функция:

logitmfx(Choice ~ Education + PolView + Age + Isolation, data = test)

Для расчета маржинального эффекта в пробит-модели используется следующая функция:

probitmfx(Choice ~ Education + PolView + Age + Isolation, data = test)

Маржинальные эффекты умножаются на 100% и интерпретируются как предсказательный эффект влияния независимого фактора на вероятность положительной альтернативы.

Шаг 11. Для построения ROC-кривой необходимо воспользоваться следующими командами

```
library(ROCR)
pr <- prediction(Preds3, test$Choice)
auc <- AUC(test$Choice, Preds3)
prf <- performance(pr, measure = "tpr", x.measure = "fpr")
plot(prf)
lines(c(0,1),c(0,1))
text(0.6,0.2,paste("AUC=", round(auc,4)), cex=1.4)
title("ROC-кривая")
```

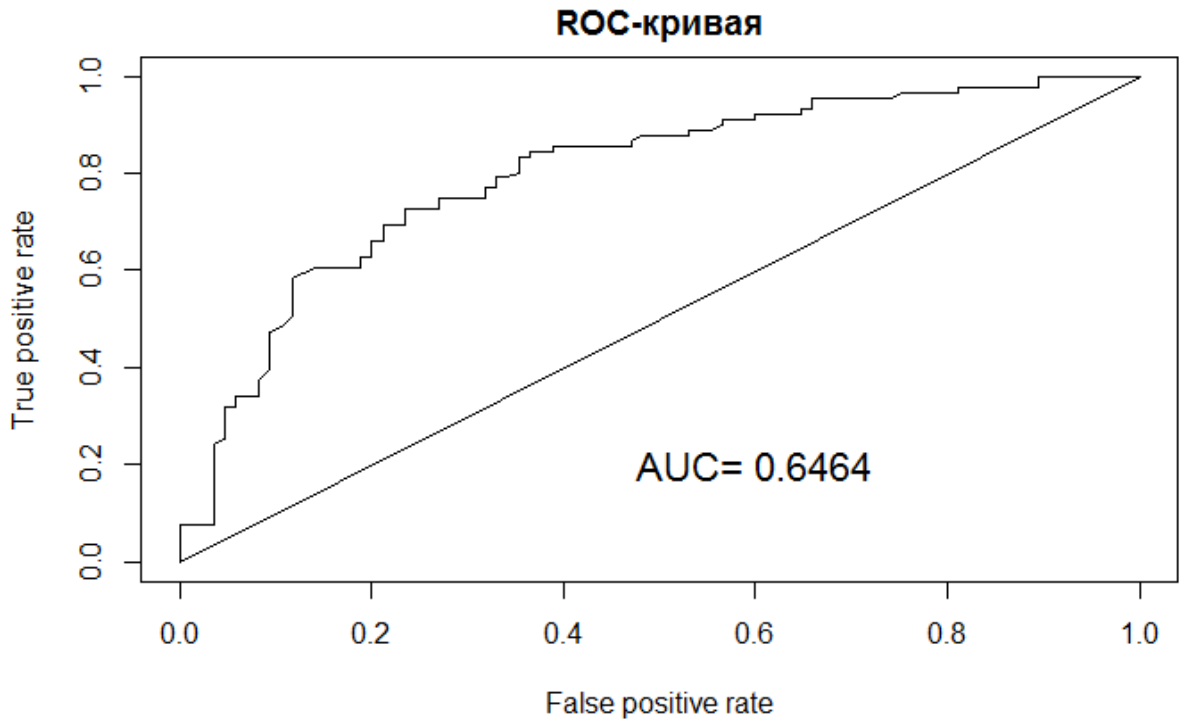

Рисунок 4

Шаг 12. Для расчета AUC воспользуемся следующей командой

```
a <- AUC(test$Choice, Preds3)
dj<-2*a*(a-0.5)
```

Памятка для выбора данных

- Систематическое завышение оценки коэффициентов регрессии при размере выборки – менее 500. Поэтому следует выбирать не менее 300 наблюдений.
- Зависимую переменную можно разделить на две группы. Независимые переменные можно брать любого типа (не только бинарные)**.**

  Например, применительно к анализу риска в медицине исследуется зависимость дихотомической результирующей переменной (т.е. принимающей только два значения, например — это статус выживаемости, подразумевающий два класса: выживет или умрет) от переменных с любым типом шкалы (пол, возраст, наличие осложнений, инфаркт миокарда в анамнезе и др.).
- При построении модели нужно минимально 10 исходов на каждую независимую переменную (рекомендованное значение 30-50). Например, интересующий исход – смерть пациента. Если 50 пациентов из 100 умирают – максимальное число независимых переменных в модели = 50/10=5.
- Данные могут быть из любой интересной для студента научной и практической области.