Санкт-Петербургский государственный университет
Кафедра управления медико-биологическими системами

Динмухаметова Дина Ринатовна

Выпускная квалификационная работа бакалавра

Прогнозирование исхода беременности

Направление 010302

Прикладная математика, фундаментальная информатика

и основы программирования

Научный руководитель,
кандидат физ.-мат. наук,
доцент
Панкратова Я. Б.

Санкт-Петербург

2020

Содержание

Введение

В настоящее время все большую актуальность обретает проблема аденомиоза при планировании беременности. Частота возникновения аденомиоза в последнее время увеличивается. Все чаще данный диагноз ставится молодым девушкам, у которых ранее было выявлено бесплодие. Однако потенциальное влияние аденомиоза на исход беременности до сих пор неясно, поскольку лишь немногие исследователи выявляли связь между заболеванием и исходом планирования ребенка. [3]. Риск осложнений беременности у женщин с аденомиозом является дискуссионным. У беременных с аденомиозом повышен риск осложнений беременности, включая невынашивание беременности, преждевременные роды, гипертонию, ограничение роста плода, инфекцию матки и прочее.

Зачастую аденомиоз протекает бессимптомно. Примерно 20 процентов случаев заболевания аденомиозом приходится на девушек моложе 40 лет, остальные 80 процентов – от 40 до 50 лет. [3] Наиболее распространенная гипотеза этиологии аденомиоза связана с инвагинацией базального слоя эндометрия в миометрий. Аденомиоз может иметь диффузное, случайное распределение или более очаговые области, известные как аденомиомы. Поэтому женщины с таким заболеванием имеют увеличенную, покрытую опухолью матку [4]. Большинство доказательств, которые связывают аденомиоз с бесплодием, ограничиваются случаями, когда у пациента наблюдалось наличие обоих диагнозов. В этих исследованиях также существует вероятность ошибочных результатов, так как аденомиоз обычно сосуществует с другими паталогическими процессами, связанными с бесплодием (например, эндометриоз, полипы, леймиома). Известно, что эндометриоз является причиной бесплодия, поэтому

существуют опасения, что результаты бесплодия были связаны с сопутствующим эндометриозом, а не с аденомиозом.

Недавние исследования показали, что аденомиоз оказывает негативное влияние на клинические результаты экстракорпорального оплодотворения (ЭКО). У женщин с диагнозом «аденомиоз», перенесших ЭКО [4], клиническая беременность за цикл, текущая беременность и живорождение значительно ниже, чем у тех, которым не диагностирован аденомиоз. Уровень выкидышей у женщин с аденомиозом также выше, чем у здоровых.

Данный вопрос до сих пор является спорным. Проведено не мало исследований в этой области. В данной работе также проводится исследование зависимости исхода беременности от аденомиоза с помощью регрессионной модели, а именно модели бинарной регрессии.

Постановка задачи

Целью выпускной квалификационной работы является изучение связи аденомиоза и исхода беременности, построение бинарной регрессионной логит-модели, исследование коэффициентов и различных параметров регрессии, выявление влияния или его отсутствия аденомиоза на беременность. Для выполнения поставленных целей необходимо решить ряд задач:

- рассмотреть понятие бинарной регрессии;
- изучить соответствующие показатели логит-модели;
- на основе базы рожениц построить регрессионную модель;
- дать интерпретацию коэффициентам и модели, оценить ее качество;
- изучить определенные гормоны: как изменяется их количество при аденомиозе.

Обзор литературы

При написании данной работы была использована научная, учебно-методическая литература, а также публикации из научных изданий и интернет-источники.

Основные медицинские определения и понятия были взяты из книги «Dorland's Illustrated Medical Dictionary», автора Dorland. [1], а также из книги «Endometriosis: etiology, pathology, diagnosis, management», Lobo RA. [2]. Дополнительным источником также являлась книга «Биологической химии» авторов Северина Е. С., Алейниковой Т. Л., Осипова Е. В., Силаевой С. А. [5]

Основными источниками для изучения регрессионного анализа были книга «Методы прикладной статистики в R и Excel» авторов Буре В.М., Парилиной Е. М., Седакова А. А. [12] и курс лекций по машинному обучению автора Воронцова К. В. [10]. Для изучения построения бинарной регрессии в медицине была изучена публикация «Построение логистической регрессии в медицине» авторов Мироновой П. Н., Владимировой Л. В. [14].

Для изучения дополнительных вспомогательных характеристик регрессионного анализа были изучены следующие материалы: «A modern introduction to probability and statistics: understanding why and how» автора Dekking F.M., «Logistic Regression Relating Patient Characteristics to Outcomes» авторов Tolles J., Meurer William J. [11]

Для изучения множественной регрессии (многофакторного анализа) был использован интернет-ресурс Studme («Многофакторный регрессионный анализ») [15]. Для более глубокой оценки качества построенных моделей был изучен ROC-анализ с помощью статьи автора Fawcett

Tom «An Introduction to ROC-analysis» в сборнике «Pattern Recognition Letters» [16].

На основе изученной литературы были построены регрессионные модели и дана их интерпретация на языке медицины. Также была дана оценка построенным моделям и разбор вспомогательных характеристик.

Глава 1. Основные понятия и определения

1.1. Аденомиоз

Определение 1.10. [1] *Аденомиоз* – гинекологическое заболевание, характеризующееся прорастанием ткани эндометрия (внутренней оболочки матки) в миометрии (толстом мышечном слое матки). В месте прорастания образуются воспаленные узлы.

Определение 1.11. [2] *Эндометриоз* – заболевание, при котором внутренний слой стенки матки распространяется за пределы его нормального расположения.

Аденомиоз и эндометриоз могут развиваться как вместе, так и по отдельности, как самостоятельные заболевания. Аденомиоз чаще встречается в детородном возрасте (25-30 лет), однако, с симптомами аденомиоза могут быть знакомы и девушки-подростки, и женщины в климаксе. Патология опасна появлением опухолевых образований на месте очагов воспалений и затрудняет, а иногда и препятствует вынашиванию беременности. У аденомиоза обычно нет специфических симптомов, однако некоторые признаки можно выявить, среди которых:

- боли внизу живота до, во время и после менструации;
- изменение менструального цикла из-за недостатка; прогестерона и повышенного эстрогена;
- анемия в связи с обильными кровотечениями;
- нарушение репродуктивной функции.

При аденомиозе принято различать четыре степени патологии в зависимости от глубины поражения мышечной ткани матки:

1. патологический процесс ограничен слизистой оболочкой тела матки, симптомы не выражены;

2. переход патологического процесса на мышечные слои, клиническая картина маловыраженная — могут наблюдаться увеличения объема кровяных выделений во время менструации, патологические выделения в середине цикла, слабый болевой синдром;

3. распространение патологического процесса на всю толщу мышечной стенки матки до ее серозного покрова, симптомы усиливаются;

4. абсолютное поражение мышечного слоя на всю его толщину с возможным распространением на рядом лежащие ткани и органы, с ярко выраженной клинической картиной.

Рассмотрим также определения некоторых гормонов, анализ влияния количества которых на разрастание эндометрия будем проводить впоследствии.

Определение 1.12. [5] *Фолликулостимулирующий гормон (ФСГ)* – это гликопротеиновый гормон, который вырабатывается и накапливается в передней доле гипофиза и влияет на функционирование половых желез. Вырабатывается базофильными клетками второго типа наряду с лютеинизирующим гормоном (ЛГ).

Определение 1.13. [5] *Лютеинизиркющий гормон (ЛГ)* – пептидный гормон, секретируемый гонадотропными клетками передней доли гипофиза. Совместно с другим гипофизарным гонадотропином – фолликулостимулирующим гормоном (ФСГ), – ЛГ необходим для нормальной работы репродуктивной системы. В женском организме ЛГ стимулирует секрецию яичниками эстрогенов, а пиковое повышение его уровня ини-

циирует овуляцию. В мужском организме ЛГ стимулирует интерстициальные клетки Лейдига, вырабатывающие тестостерон.

Определение 1.14. [5] *Прогестерон (ПРГ)* – эндогенный стероид и половой гормон прогестагена, участвующий в менструальном цикле, беременности, а также в эмбриональном развитии у человека и других видов.

Определение 1.15. [6] *Эстрадиол (ЭСТР)* – эстрогенный стероидный гормон и основной женский половой гормон. Он участвует в регуляции менструальных репродуктивных циклов, отвечает за развитие вторичных половых признаков у женщин, таких, как грудь, бедра, распределение жировой прослойки, а также играет важную роль в развитии и поддержании женских репродуктивных тканей, таких, как молочные железы, матка в период полового созревания, зрелости и беременности.

Определение 1.16 [7] *Пролактин (ПРЛ)* , также известный как лютеотропный гормон, представляет собой белок, отвечающий за производство молока (лактацию) у млекопитающих (а именно, самок). Он влияет более, чем на 300 различных процессов у позвоночных, в том числе и людей. Пролактин выделяется из гипофиза в ответ на прием пищи, спаривание, овуляцию и кормление грудью. Пролактин играет важную роль в обмене веществ, регуляции иммунной системы, развитии поджелудочной железы.

Определение 1.17 [8] *Миометрий* – средний слой стенки метки, образованный гладко-мышечными клетками матки. Его основная функция – вызывать сокращение матки.

В журнале «Акушерства и гинекологических исследований» («Obstetrics and Gynaecology Research») опубликована статья по исследованию связи между аденомиозом и беременности. В исследовании наблюдались 36

женщин с диагнозом «аденомиоз», поставленным в период с 2002 по 2012 год, которые стояли на учете в центре третичной помощи (третичная помощь оказывается врачом или группой врачей с соответствующей подготовкой для диагностики и лечения сложных заболеваний, которые требуют специальных методов диагностики и лечения). Контрольная группа состояла из 144 женщин без маточных аномалий, родивших в течение того же периода и чей возраст при родах был скорректирован путем применения оценок склонности. В итоге были сравнены исходы беременности контрольной группы и группы женщин с аденомиозом. Группа женщин с аденомиозом имела значительно более высокий уровень преждевременных родов (41.7% к 12.5%), неправильного формирования плода (27.8% к 8.3%), родов посредством кесарева сечения (58.3% к 24.3%). В результате этих исследований было зафиксировано, что при наличии аденомиоза беременность редко заканчивалась родами без нарушений и патологий.

Также были проведены исследования осложнений и исходов беременности у женщин с аденомиозом в Японии, результаты которых показали, что неблагоприятный исход беременности связан с увеличенным размером и диффузным типом аденомиоза.

Проблема аденомиоза и по сей день актуальна, поскольку до сих пор подвергается множеству дискуссий. Все чаще аденомиозом болеют молодые девушки, у которых ранее было диагностировано бесплодие. Аденомиоз может протекать бессимптомно. Стоит также отметить, что не всегда данное заболевание препятствует беременности, однако прямая связь невозможности зачатия и диагноза все же есть: присутствует выраженное воспаление миометрия, нарушена его структура, что и препятствует выходу яйцеклетки.

Рассмотрим эту проблему с математической точки зрения, применяя регрессионный анализ.

1.2. Регрессионный анализ. Бинарные логистические модели.

Перейдем к рассмотрению математической составляющей данного исследования. Основой анализа являются регрессионные модели. Почему регрессионный анализ? Регрессионный анализ активно применяется во многих сферах жизни. Например, важность регрессионного анализа для малого бизнеса заключается в том, что он помогает определить, какие факторы оказывают наибольшее влияние, как факторы взаимодействуют друг с другом. Он предоставляет мощный статистический аппарат, позволяет исследовать отношения между двумя или более интересующими переменными.

Аналогично в медицине. Выявить влияние тех или иных факторов на конкретный процесс позволяет именно корреляционно-регрессионный анализ. Этот метод используется для прогнозирования и нахождения причинно-следственной связи между переменными.

Определение 1.20. [9] *Линейная регрессия* – основной и чаще всего используемый метод прогнозного анализа. Основная идея регрессионного анализа состоит в том, чтобы изучить две вещи:

- Выполняет ли набор переменных-предикторов (факторов) при прогнозировании исходной переменной (объясняемой)?
- Какие переменные в частности являются наиболее значимыми при прогнозировании зависимой переменной? И также каким образом объясняющие переменные, характеризующиеся величиной и знаком коэффициентов β, влияют на итоговую переменную?

Такая оценка регрессии позволяет объяснить связь между одной зависимой переменной и одной или несколькими независимыми переменными.

Регрессионный анализ широко применяется в предсказывании и прогнозировании событий, его использование существенно пересекается с областью машинного обучения. В стандартной линейной регрессионной модели зависимая (объясняемая) переменная может принимать различные значения, как положительные, так и отрицательный.

В данной работе зависимая переменная дает нам понять, какой исход беременности следует ожидать. Поскольку рассматриваемых исходов 2, следовательно, зависимая переменная принимает двоичные значения, то есть линейная регрессия в данном случае не подходит.

По этой причине здесь используется бинарная логистическая регрессию. Логит-модель используется для моделирования вероятности существования определенного класса или события, такого, как выигрыш/проигрыш, живой/мертвый, здоровый/больной и т.д.

Из названия самой модели можно догадаться, что она строится на логистической функции для моделирования бинарной зависимой переменной, хотя существуют и более сложные ее расширения.

В логистической модели лог-шансы для значения-индикатора «1» представляют собой линейную комбинацию одной или нескольких независимых переменных. Каждая независимая переменная может быть бинарной (два класса, закодированные индикаторной переменной) или непрерывной переменной (любое вещественное число). [11]

Определение 1.21. [11] *Логистическая регрессия (логит-модель)* – тип регрессии, позволяющий оценивать апостериорные вероятности принадлежности объектов классам. Эта регрессия выдает ответ в виде вероятности бинарного события (1 или 0). [12]

Как мы уже выяснили, в бинарном случае левая часть регрессионной

модели может принимать только два числа: 0 и 1, которые оценивают вероятность наступления того или иного события. Однако правая часть этой модели принимает значения всей вещественной оси. Для того, чтобы превратить правую часть также в бинарную, необходимо ввести дополнительную функцию. Используется следующее уравнение регрессии - вероятность наступления события для некоторого случая:

$$f(z) = \frac{e^z}{1 + e^z},$$

где $z = \beta_0 + \beta_1 \times x$, β_i - коэффициенты, расчёт которых является задачей бинарной логистической регрессии.

Для сравнения также рассматривается пробит-модель.

Определение 1.22. [12] *Пробит-модель* – это тип регрессии, когда зависимая переменная также может принимать только два значения, например, замужем или нет. Название происходит от английских probability (вероятность) + unit(единица измерения/целое). Целью модели является оценка вероятности того, что наблюдение с конкретными характеристиками попадет в конкретную категорию; кроме того, классификация наблюдений на основе предсказанных вероятностей является типом бинарной модели классификации. В пробит-модели используются методы, аналогичные тем, которые используются в логит-модели, поскольку оба данных типа регрессии направлены на решение одного и того же набора проблем.

Практическое использование логистического распределения происходит от его гораздо более простой интегральной функции распределения.

Логит-модель рассматривает тот же набор проблем, что и пробит-модель, используя аналогичные методы, причем последний использует

интегральную кривую нормального распределения. Логистическая регрессия может рассматриваться как частный случай обобщенной линейной модели и, таким образом, аналогична линейной регрессии. Однако логистическая регрессия основана на совершенно иных предположениях о связи между зависимыми и независимыми переменными в отличие от предположений в линейной регрессии. В частности, ключевые различия между этими двумя моделями можно увидеть в следующих особенностях логистической регрессии. Во-первых, условное распределение $y|x$ является распределением Бернулли, а не распределением Гаусса, потому что зависимая переменная является бинарной. Во-вторых, прогнозируемые значения являются вероятностями и поэтому ограничены $(0, 1)$ посредством функции логистического распределения, поскольку логистическая регрессия прогнозирует вероятность конкретных результатов, а не сами результаты.

Глава 2. Построение логит-модели, анализ показателей

2.1. Построение логит-модели

Построим бинарную логистическую регрессию по данным 58 наблюдаемых пациенток. В данном исследовании рассматривается два исхода планирования беременности: 0 – «беременность прошла успешно, закончилась родами» и 1 – «беременность не состоялась». Под последним понимаем следующее: бесплодие, выкидыш, беременность с угрозой. То есть, другими словами, исследуемый признак может принимать два значения.

Поскольку аденомиоз матки — это прорастание клеток эндометрия вглубь промежуточного и мышечного слоя, при котором в местах прорастания появляются воспаленные узлы, нас интересуют два фактора: толщина соединительной зоны матки (Jzmax), так как толщина переходной зоны более 12 мм позволяет определить диагноз, и отношение толщины соединительной зоны матки к толщине миометрия (Jzmax/myo).

Задана выборка $X^2 = (x_1, x_2)$, $X \in R$, где $x_1 = Jzmax$, $x_2 = Jzmaxmyo$. Выдвигается гипотеза

$$H_0 : \beta_i = 0,$$

то есть отсутствует влияние переменных-признаков на зависимую переменную, связь аденомиоза с исходом планирования беременности отсутствует. Уровень значимости $\alpha = 0,05$ - допустимая для данной задачи вероятность ошибки первого рода.

```
Logistic regression                                     Number of obs   =         58
                                                        LR chi2(1)      =      23.80
                                                        Prob > chi2     =     0.0000
Log likelihood = -18.963056                             Pseudo R2       =     0.3855

------------------------------------------------------------------------------
       ishod |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Jzmax |   .5955551   .1801333     3.31   0.001     .2425004    .9486099
       _cons |  -3.736081   1.337849    -2.79   0.005    -6.358217   -1.113944
------------------------------------------------------------------------------
```

Рис. 1: Логистическая регрессия по Jzmax

Логит-модель, построенная с помощью статистического пакета Stata (Рис. 1), оказалась значимой, поскольку вероятность получить статистику хи-квадрат при выполнении гипотезы H_0 стремится к нулю. Это говорит о том, что хотя бы один предиктор не равен нулю и несет влияние на зависимую переменную. Коэффициенты построенной модели также значимы.

Интерпретируем полученные результаты, которые видим на рис.1.

Обратим внимание на то, что значение $Prob > chi2$ стремится к нулю. Это вероятность получить статистику хи-квадрат, если нулевая гипотеза верна; нулевая гипотеза состоит в том, что все коэффициенты регрессии модели равны нулю и влияние предикторов на зависимую переменную отсутствует. Другими словами, данное значение характеризует значимость модели. Это значение сравнивается с указанным альфа-уровнем, нашей готовностью принять ошибку первого рода, которая обычно устанавливается 0,05 или 0,01. В данном исследуемом случае была выбрана $\alpha = 0,05$. Именно за счет этого построенная модель считается значимой, поскольку вероятность ошибки меньше заданного уровня α.

Введем определение доверительного интервала.

Определение 2.10. Доверительный интервал – интервал, который по-

крывает неизвестный параметр с заданной вероятностью. В общих чертах доверительный интервал для неизвестного параметра распределения случайной величины основан на выборке с уровнем доверия p. [17]

Доверительным интервалом для заданного параметра Jzmax является интервал с границами $[0.2425004; 0.9486099]$. Для данной переменной-признака мы бы сказали, что мы на 95% уверены, что коэффициент регрессии находится между нижней и верхней границами интервала.

Коэффициент при параметре Jzmax равен:

$$\beta_1 = 0.59555551.$$

Это говорит о том, что при увеличении соединительной зоны матки на 1 мм вероятность отрицательного исхода («беременность не состоялась») увеличивается на 0.5952724. В доверительный интервал значение попадает.

LR chi2(1) – это критерий Хи-квадрат отношения правдоподобия, согласно которому хотя бы один из коэффициентов регрессии не равен нулю в модели. Число в скобках указывает степень свободы распределения хи-квадрат, используемого для проверки статистики хи-квадрат, и определяется числом предикторов в модели.

Pseudo R2 – псевдо R^2 Макфаддена или индекс отношения правдоподобия. Значение этого параметра варьируется от 0 до 1. Логистическая регрессия не имеет эквивалента показателю R-квадрат, в связи с чем предлагается интерпретировать его с большой осторожностью. Полагаться на значение этого параметра не стоит. При интерпретации pseudo R2 не стоит ожидать, что значение будет очень большим, полагается, что значения 0.2 - 0.4 указывают на отличную модель.

Построим бинарную логистическую модель для параметра Jzmaxmyo, отвечающего за отношение толщины соединительной зоны матки к толщине миометрия.

```
Logistic regression                                     Number of obs   =         58
                                                        LR chi2(1)      =      19.43
                                                        Prob > chi2     =     0.0000
Log likelihood = -21.146048                             Pseudo R2       =     0.3148
```

ishod	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Jzmaxmyo	.166256	.0521562	3.19	0.001	.0640318	.2684803
_cons	-3.316006	1.308866	-2.53	0.011	-5.881335	-.750676

Рис. 2: Логистическая регрессия по Jzmaxmyo

Доверительным интервалом для заданного параметра Jzmax является интервал с границами $[0.0640318; 0.2684803]$. Коэффициент при параметре Jzmaxmyo равен:

$$\beta_1 = 0.166256.$$

То есть при увеличении отношения толщины соединительной зоны матки к отношению толщины миометрия на 1% вероятность отрицательного исхода увеличивается на 0.1657507. В доверительный интервал значение также попадает.

Значение prob>chi2 также стремится к нулю, что говорит о том, что построенная модель является значимой. То есть в действительности величина отношения толщины соединительной зоны матки к толщине миометрия не имеет прямой связи с исходом беременности.

2.2. Построение сравнительной пробит-модели

Для сравнительного анализа построим по тем же данным с использованием тех же факторов пробит-модель. Как уже было отмечено выше, пробит-модель – это статистическая модель бинарного выбора так же, как и логит-модель. Функции и задачи ее построения аналогичны логит-модели. Сферы применения пробит-модели такие же, как и сферы применения логистической регрессии. Результаты и классификации также в целом очень похожи. Однако в отличие от логит-модели, где отклонение распределено логистически, при построении пробит-модели используется нормальное распределение.

Итак, здесь уравнение регрессии выглядит следующим образом:

$$f(z) = \frac{1}{2\pi} \int_{-\infty}^{z} e^{\frac{-u^2}{2}} du,$$

где $z = \beta_0 + \beta_1 \times x$, β_i - коэффициенты, расчёт которых является задачей бинарной пробит-регрессии.

```
Probit regression                                 Number of obs   =         58
                                                  LR chi2(1)      =      23.77
                                                  Prob > chi2     =     0.0000
Log likelihood = -18.978336                       Pseudo R2       =     0.3850

------------------------------------------------------------------------------
       ishod |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       Jzmax |    .3281103   .0907887     3.61   0.000     .1501678    .5060529
       _cons |   -2.058472   .7182471    -2.87   0.004    -3.46621   -.6507332
------------------------------------------------------------------------------
```

Рис. 3: Пробит-регрессия по Jzmax

Видим, что результаты очень близки к тем, которые были получены при построении бинарной логистической регрессионной модели для параметра Jzmax. Значение cons говорит о том, что при нулевом параметре Jzmax вероятность отрицательного исхода крайне мала, а именно

$$f(-2.058472) = 0.01977.$$

Полученная модель является значимой, коэффициент при параметре Jzmax равен:

$$\beta_1 = 0.3281103,$$

здесь также наблюдаем влияние аденомиоза на исход беременности.

В целом, можно заметить, что обе модели дают примерно один и тот же результат.

Рассмотрим пробит-модель для параметра Jzmaxmyo.

```
Probit regression                                 Number of obs   =         58
                                                  LR chi2(1)      =      18.96
                                                  Prob > chi2     =     0.0000
Log likelihood = -21.382951                       Pseudo R2       =     0.3071

------------------------------------------------------------------------------
       ishod |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    Jzmaxmyo |   .0883973   .0253728     3.48   0.000     .0386675    .1381271
       _cons |  -1.738071   .6853641    -2.54   0.011    -3.08136    -.394782
------------------------------------------------------------------------------
```

Рис. 4: Пробит-регрессия по Jzmaxmyo

Полученные значения также близки к результатам построенной выше логит-модели. Модель также является значимой.

Общий показатель ошибки определяется отношением числа неправильно классифицированных наблюдений к общему числу наблюдений [14]. В данном исследовании при построении логит-модели и при рас-

смотрении фактора Jzmax было выявлено 9 ошибочно классифицированных исходов, что говорит о том, что общий показатель ошибки равен 0,15517241, другими словами модель корректно классифицирует данные на 84,48%.

Logistic model for ishod

	True		
Classified	D	~D	Total
+	**40**	**4**	**44**
-	**5**	**9**	**14**
Total	**45**	**13**	**58**

Рис. 5: классификация логит-модели по Jzmax

При анализе фактора Jzmaxmyo имеется 11 ошибочно классифицированных исходов, общий показатель ошибки равен 0,18965517, модель корректно классифицирует данные на 81,03%.

Logistic model for ishod

	True		
Classified	D	~D	Total
+	**41**	**7**	**48**
-	**4**	**6**	**10**
Total	**45**	**13**	**58**

Рис. 6: классификация логит-модели по Jzmaxmyo

Проведем аналогичный анализ результатов для пробит-модели.

В случае рассмотрения параметра Jzmax получаем идентичные результаты тем, что были получены при оценке качества построенной логит-модели.

Видим, что полученные результаты также совпали с теми, что были

Probit model for ishod

	True		
Classified	D	~D	Total
+	**40**	**4**	**44**
-	**5**	**9**	**14**
Total	**45**	**13**	**58**

Рис. 7: классификация пробит-модели по Jzmax

Probit model for ishod

	True		
Classified	D	~D	Total
+	**41**	**7**	**48**
-	**4**	**6**	**10**
Total	**45**	**13**	**58**

Рис. 8: классификация пробит-модели по Jzmaxmyo

получены при анализе логит-модели для фактора Jzmaxmyo. Это позволяет нам сделать вывод, что при данном размере выборки результаты пробит-модели фактически не отличаются от результатов логит-модели. Для анализа зависимости исхода беременности от диагноза «аденомиоз» можно использовать как логит-модель, так и пробит-модель.

2.3. Многофакторные регрессионные модели

Рассмотрим многофакторные логит- и пробит-модели с использованием тех же факторов Jzmax, Jzmaxmyo, но теперь уже их совместное влияние. При анализе многофакторных моделей будем выдвигать ту же гипотезу H_0 и проверять ее на том же уровне значимости $\alpha = 0.05$.

Многофакторные модели более применимы в жизни, поскольку чаще всего на признак влияет не один фактор, а несколько.

Многофакторный анализ сводится к решению следующих задач [15]:

- построение уравнения множественной регрессии
- определение степени влияния каждого фактора на результативный признак
- количественная оценка тесноты связи между результативным признаком и факторами
- оценка надежности построенной регрессионной модели
- прогноз результативного признака

```
Logistic regression                                     Number of obs     =         58

                                                        Wald chi2(2)      =      11.17
Log likelihood = -18.813007                             Prob > chi2       =     0.0038
```

ishod	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
Jzmax	.7915875	.4144284	1.91	0.056	-.0206771	1.603852
Jzmaxmyo	-.0634429	.1144762	-0.55	0.579	-.2878121	.1609264
_cons	-3.626931	1.32144	-2.74	0.006	-6.216906	-1.036955

Рис. 9: Многофакторная логит-регрессия

Перейдем к изучению показателей построенной модели. Здесь Wald chi2(2) – это статистика Вальда хи-квадрат. Она используется для проверки гипотезы о том , что хотя бы один из коэффициентов регрессии не равен нулю. Число в скобках указывает на степень свободы распределения хи-квадрат, определяется числом переменных-признаков в модели.

cons – это регрессионная оценка модели, когда все переменные-признаки равны нулю. То есть, предполагая, что Jzmax и Jzmaxmyo не несут никакого влияния на исход, что означает, что беременность не имеет связи с аденомиозом, мы получаем функцию логистического распределения, равную $f = 0.02590128$. Это говорит о том, что при нулевых Jzmax и Jzmaxmyo с данными предложенными параметрами вероятность отрицательного исхода беременности крайне мала. Значение prob>chi2 меньше заданного уровня значимости, из чего следует сделать вывод о том, что построенная модель значима. Однако при исследовании двух факторов в совокупности можем наблюдать, что отношение толщины соединительной зоны матки к толщине миометрия несет противоположное влияние тому, которое было выявлено при анализе факторов по отдельности. Это произошло потому, что наблюдается сильная мультиколлинеарность, коэффициент корреляции равен $0,8927$, что говорит о высокой степени связи между переменными, p-value больше заданного уровня значимости. Таким образом, делаем вывод, что данной многофакторной моделью пользоваться нельзя.

	Jzmax	Jzmaxmyo
Jzmax	**1.0000**	
Jzmaxmyo	**0.8927**	**1.0000**

Рис. 10: Коэффициент корреляции

2.4. Построение ROC-кривых

Определение 2.40. [16] *ROC-кривая* – график, который иллюстрирует качество бинарной классификации (receiver operating characteristic), это график зависимости чувствительности от 1-специфичности. Чувствительность определяет долю положительных случаев, которые были правильно классифицированы построенной регрессионной моделью.

1-специфичность — это доля отрицательных случаев, которые также правильно классифицированы. Таким образом, чувствительность является истинно-положительным показателем, а специфичность — истинно-отрицательным. ROC-анализ количественно оценивает точность диагностических тестов или других методов оценки, используемых для выявления различий между двумя состояниями (контрольными и случайными, например). Точность диагностического теста определяется его способностью правильно классифицировать переменные.

Построим ROC-кривую для исследуемых данных.

Лучше всего проведенный тест характеризует площадь под ROC-кривой — площадь, ограниченная ROC-кривой и осью доли ложных положительных классификаций, которая сокращенно называется AUC (area under curve). Эту область можно интерпретировать как вероятность того, что результат проведенной диагностики случайно выбранного верно классифицированного исхода будет больше, чем результат той же проверки случайно выбранного ошибочно классифицированного исхода. Чем больше AUC, тем лучше общая производительность теста.

На рисунке 11 можем видеть, что площадь под ROC-кривой для нашей построенной регрессионной модели равна 0.8923, это значит, что данная модель правильно разделит классы с вероятностью 0.8923.

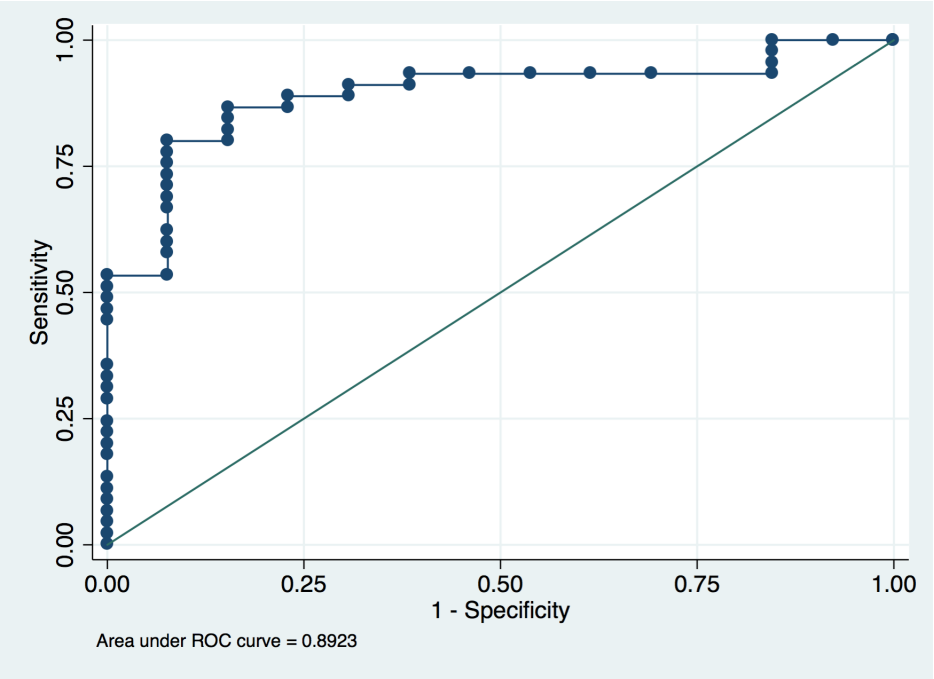

Рис. 11: ROC-кривая по параметру Jzmax

Рассмотрим также ROC-кривую для анализа параметра Jzmaxmyo.

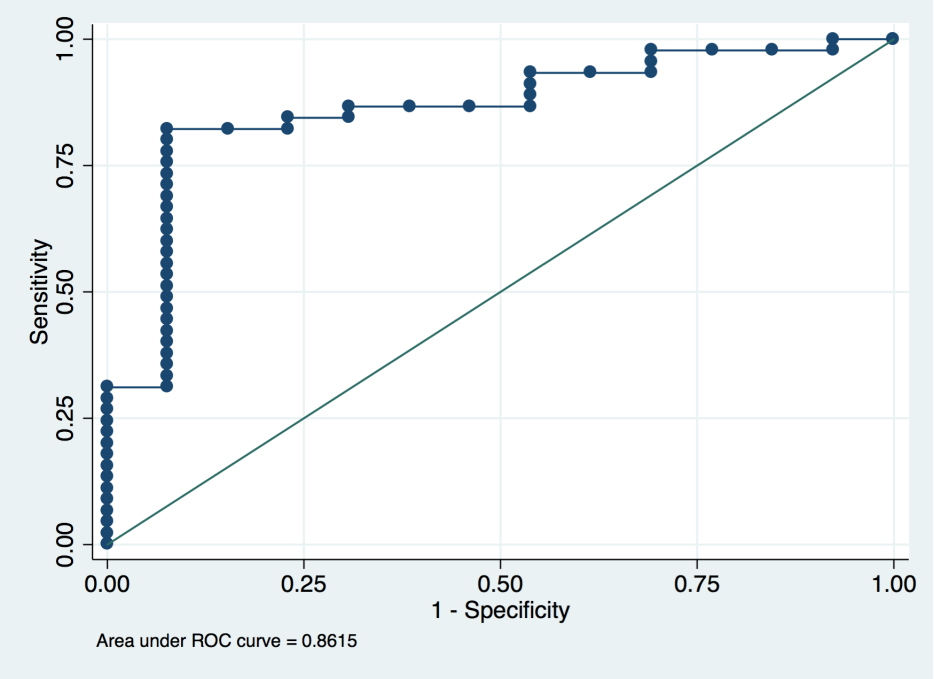

Рис. 12: ROC-кривая по параметру Jzmaxmyo

Видим, что площадь под кривой равна

$$AUC = 0.8615,$$

что говорит нам о высоком качестве построенной модели, модель можно использовать для анализа.

2.5. Анализ связи гормонов и толщины соединительной зоны матки

Как уже было отмечено ранее, ФСГ, ЛГ, ПРГ, ПРЛ, ЭСТР играют немаловажную роль в работе репродуктивной системы, обмене веществ, влияют на функционирование желез. Рассмотрим взаимосвязь толщины соединительной зоны матки и количества каждого из гормонов.

```
. regress Jzmax prg
```

Source	SS	df	MS		
Model	39.554252	1	39.554252	Number of obs	= 58
Residual	853.2781	56	15.2371089	F(1, 56)	= 2.60
				Prob > F	= 0.1128
				R-squared	= 0.0443
				Adj R-squared	= 0.0272
Total	892.832352	57	15.6637255	Root MSE	= 3.9035

Jzmax	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
prg	-.0322845	.0200378	-1.61	0.113	-.072425	.0078559
_cons	11.58481	.9334106	12.41	0.000	9.71496	13.45465

Рис. 13: Прогестерон; ss – сумма квадратов ошибок в прогнозировании; df – степень свободы минус единица; MS – сумма квадратов, деленная на df

Прогестерон вырабатывается в яичниках. Этот гормон иногда называют «гормоном беременности». Очень часто невозможность беременности связана с недостаточностью ПРГ. При эндометриозе и в частности при аденомиозе наблюдается снижение ПРГ, играющего важную защитную роль.

Значение prob>F больше заданного уровня значимости $\alpha = 0.05$, что говорит о том, что независимая переменная не показывает статистически значимой связи с зависимой переменной, независимая переменная не может надежно прогнозировать зависимую переменную.

R-squared – это та доля зависимой переменной, которая объясняется

независимой переменной. В данном случае лишь 4,43% Jzmax объясняются переменной prg. То есть однозначного вывода о влиянии количества прогестерона на толщину соединительной зоны матки и, как следствие, на наличие аденомиоза сделать нельзя. Знаем только то, что при аденомиозе наблюдается недостаток прогестерона.

Adj R-squared – скорректированный R-квадрат (коэффициент детерминации). Он вычисляется по формуле:

$$1 - (1 - R^2) \cdot \frac{N - 1}{N - k - 1}$$

Когда число наблюдений мало, а количество переменных-признаков велико, коэффициент детерминации будет сильно отличаться от скорректированного коэффициента детерминации.

Можем видеть, что скорректированный коэффициент детерминации говорит лишь о 2,72% Jzmax , которые объясняются изменением гормона прогестерон. Однако p-value также больше заданного уровня значимости, что говорит о незначимости коэффициентов.

```
regress Jzmax prl
```

Source	SS	df	MS
Model	.779104847	1	.779104847
Residual	892.053247	56	15.9295223
Total	892.832352	57	15.6637255

Number of obs	= 58
F(1, 56)	= 0.05
Prob > F	= 0.8258
R-squared	= 0.0009
Adj R-squared	= -0.0170
Root MSE	= 3.9912

Jzmax	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
prl	-.0008376	.0037873	-0.22	0.826	-.0084244	.0067492
_cons	10.60997	1.378793	7.70	0.000	7.847919	13.37203

Рис. 14: Пролактин

Пролактин приводит к снижению уровня рецепторов эстрогенов (женских половых гормонов) в матке. При наличии аденомиоза нередко за-

мечается повышение концентрации пролактина. Однако исследование взаимосвязи толщины соединительной зоны матки и пролактина также показало отсутствие статистической значимости при прогнозировании, поскольку лишь 0.009% Jzmax объясняются переменной prl. Более того значение prob>F больше заданного уровня значимости $\alpha = 0.5$, то есть статистически значимая связь не наблюдается.

Перейдем к рассмотрению фолликулоостимулирующего гормона (ФСГ). При сильном повышении уровня пролактина наблюдается низкий уровень фолликулостимулирующего гормона. Низкий уровень ФСГ наблюдается при дисфункции яичников, новообразованиях гипофиза. Проведем такой же анализ связи ФСГ и толщины соединительной зоны матки.

```
regress Jzmax fsg
```

Source	SS	df	MS	Number of obs =	58
				F(1, 56) =	2.24
Model	34.4026097	1	34.4026097	Prob > F =	0.1397
Residual	858.429742	56	15.3291025	R-squared =	0.0385
				Adj R-squared =	0.0214
Total	892.832352	57	15.6637255	Root MSE =	3.9152

Jzmax	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
fsg	.6168127	.4117333	1.50	0.140	-.2079882	1.441614
_cons	6.691288	2.481363	2.70	0.009	1.720521	11.66206

Рис. 15: Фолликулостимулирующий гормон

$Prob > F = 0.1397$, это значительно больше уровня значимости α, следовательно, в очередной раз зависимость не является статистически значимой. Только 3.85% Jzmax объясняются переменной fsg.

Аналогичные результаты дают исследования гормона эстрадиол (ЭСТР) и лютеинизирующего гормона (ЛГ).

Эстрадиол относится к эстрогенным гормонам, он вырабатывается в яичниках и отвечает за функционирование женской половой системы.

ЛГ – гормон передней доли гипофиза (нижний мозговой придаток мозга; самый главный орган эндокринной системы).

```
regress Jzmax estr
```

Source	SS	df	MS		
Model	1.02519062	1	1.02519062	Number of obs	58
Residual	891.807161	56	15.9251279	F(1, 56)	0.06
				Prob > F	0.8006
Total	892.832352	57	15.6637255	R-squared	0.0011
				Adj R-squared	-0.0167
				Root MSE	3.9906

Jzmax	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
estr	.0015486	.0061036	0.25	0.801	-.0106784	.0137757
_cons	10.11917	.9754871	10.37	0.000	8.165031	12.0733

Рис. 16: Эстрадиол

```
regress Jzmax lg
```

Source	SS	df	MS		
Model	10.824938	1	10.824938	Number of obs	58
Residual	882.007414	56	15.7501324	F(1, 56)	0.69
				Prob > F	0.4106
Total	892.832352	57	15.6637255	R-squared	0.0121
				Adj R-squared	-0.0055
				Root MSE	3.9686

Jzmax	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lg	.1409177	.1699788	0.83	0.411	-.1995909	.4814262
_cons	9.463989	1.165139	8.12	0.000	7.129935	11.79804

Рис. 17: Лютеинизирующий гормон

Анализ показал, что 0.11% Jzmax объясняются переменной estr и 1.21% Jzmax объясняются переменной lg. При этом значение prob>F в обоих случаях больше заданного уровня значимости α. Статистическая значимость не наблюдается, следует полагать, что зависимость между переменными отсутствует.

Таким образом, можно сделать вывод, что о прямой взаимосвязи повышения/понижения того или иного гормона и аденомиоза говорить

нельзя, нужны дополнительные исследования. Статистическая значимость коэффициентов и моделей не обнаружена. В любом случае как повышение концентрации гормонов, так и ее понижение говорят о нарушениях в организме.

Аденомиоз – это одна из разновидностей эндометриоза. Исследований заболевания «аденомиоз» в настоящее время не так много, именно поэтому точный ответ на вопрос: «зависит ли исход планирования беременности от аденомиоза?» до сих пор нет. В связи с этим трудно определить зависимость наличия аденомиоза от тех или иных патологий в женском организме.

Заключение

В данной работе были выполнены все поставленные задачи. В главе 1 были изучены все необходимые медицинские и математические понятия. Были выявлены возможные причины аденомиоза, а также симптомы заболевания. Далее был рассмотрен регрессионный анализ, а именно бинарные логистические модели, сферы применения бинарного анализа. Для последующего сравнительного анализа было введено понятие пробит-модели.

В главе 2 на основе предложенной базы рожениц были построены логит- и пробит-модели, был выполнен их сравнительный анализ, который показал, что обе модели дают близкие по значениям результаты. После построения моделей был проведен анализ параметров моделей, который показал, что модели значимые. Была проведена классификация моделей, которая также показала высокий уровень корректности моделей. Построить множественную регрессию не удалось, так как исследуемые факторы сильно коррелируют между собой. Были построены ROC-кривые, изучена AUC – площадь под кривой, которая отвечает за производительность теста. Исследование показало, что логит-модели по параметрам Jzmax, Jzmaxmyo правильно разделяют классы более, чем на 80%. Был проведен анализ характерных гормонов и их связи с аденомиозом. Статистическая значимость выявлена не была, более того показатель ошибки превышал допустимый уровень значимости α. Концентрация каждого из рассматриваемых гормонов крайне мало зависит от заболевания «аденомиоз».

В результате исследования выборки из 58 пациенток было выявлено влияние аденомиоза на исход беременности. В 8 случаях из 10 наличие

аденомиоза действительно препятствовало развитию беременности. Однако распространять данный результат на случайную базу данных пока еще рано.

Список литературы

1. Dorland Dorland's Illustrated Medical Dictionary. 33 изд.
2. Lobo RA. Endometriosis: etiology, pathology, diagnosis, management. In: Comprehensive Gynecology. Philadelphia, PA: Mosby; 5th ed:2007
3. Adenomyosis and its impact on fertility // Contemporary OB/GYN https://www.contemporaryobgyn.net
4. Tomassetti C, Meuleman C, Timmerman D, D'Hooghe T. Adenomyosis and Subfertility: Evidence of Association and Causation. Semin Reprod Med. 2013; 31(02):101-108.
5. Биологическая химия / Е.С. Северин, Т.Л. Алейникова, Е.В. Осипов, С.А. Силаева, Под ред. Е.С Северина.2003.
6. Biochemistry of aromatase: significance to female reproductive physiology. // PubMed https://www.ncbi.nlm.nih.gov/pubmed/
7. Bole-Feysot C, Goffin V, Edery M, Binart N, Kelly PA Prolactin (PRL) and its receptor: actions, signal transduction athways and phenotypes observed in PRL receptor knockout mice // Endocrine Reviews 19 (3): 225–68
8. Aguilar, H. N.; Mitchell, S.; Knoll, A. H.; Yuan, X. Physiological pathways and molecular mechanisms regulating uterine contractility // Human Reproduction Update, 2010, 16 (6): 725–744
9. David A. Freedman. Statistical Models: Theory and Practice. 2009
10. Машинное обучение (курс лекций) // MachineLearning http://www.machinelearning.ru/wiki

11. Tolles J., Meurer William J. Logistic Regression Relating Patient Characteristics to Outcomes // JAMA, 2016, 316 (5): 533–4.

12. Буре В. М., Парилина Е. М., Седаков А. А. Методы прикладной статистики в R и Excel. 3 изд. Лань, 2018. 152 с..

13. Bliss, C. I. Oxford English Dictionary. The Method of Probits. 3 ed. 1934. 38-39 p.

14. Миронова П.Н., Владимирова Л.В. Построение логистической регрессии в медицине // Процессы управления и устойчивость. 2018. Т. 5(21) С. 233-239.

15. Многофакторный регрессионный анализ // Studme https://studme.org/

16. Fawcett Tom, An Introduction to ROC Analysis // Pattern Recognition Letters, 2006, 27 (8): 861–874.

17. Dekking F.M. A modern introduction to probability and statistics : understanding why and how.