МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**

Кафедра теории функций и стохастического анализа

МЕТОДЫ АППРОКСИМАЦИИ В ЗАДАЧАХ МАШИННОГО ОБУЧЕНИЯ

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

Студентки 2 курса 248 группы
направления 09.04.03 — ПРИКЛАДНАЯ ИНФОРМАТИКА

механико-математического факультета
Зайнышевой Дарьи Айратовны

Научный руководитель
профессор, д. ф.-м. н. _____ П. А. Терехин

Заведующий кафедрой
д. ф.-м. н. _____ С. П. Сидоров

Саратов 2019

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

С каждым годом растет потребность в изучении больших данных как для компаний, так и для активных энтузиастов. В таких крупных компаниях, как Яндекс или Google, всё чаще используются такие инструменты для изучения данных, как язык программирования R, или библиотеки для Python. Согласно Закону Мура, количество транзисторов на интегральной схеме удваивается каждые 24 месяца. Это значит, что с каждым годом производительность компьютеров растет, а значит и ранее недоступные границы познания снова «смещаются вправо» — открывается простор для изучения больших данных, с чем и связано в первую очередь создание «науки о больших данных», изучение которых в основном стало возможным благодаря применению алгоритмов машинного обучения.

Машинное обучение — класс методов искусственного интеллекта, характерной чертой которых является не прямое решение задачи, а обучение в процессе применения решений множества сходных задач.

Целями выполнения магистерской работы являются изучение методов аппроксимации в задачах построения регрессии и применение полученных знаний на практике, а также закрепление, углубление и расширение знаний о регрессионном анализе.

Для достижения поставленных целей были выделены следующие задачи:

- Изучение парной и множественной линейной регрессии.
- Определение трудностей, которые возникают при анализе, обусловленных особенностями изучаемых наборов данных.
- Выявление способов преодоления обнаруженных трудностей.
- Изучение регуляризации, как способа преодоления мультиколлинеарности и метода понижения размерности выборки.
- Изучение Ridge и Lasso регрессий.
- Рассмотрение Ridge и Lasso регрессии для построения инвестиционного портфеля.

Термин «регресс» придумал Фрэнсис Гальтон в XIX веке, чтобы описать биологическое явление. Суть была в том, что рост потомков от роста предков, как правило, регрессирует вниз к нормальному среднему. Для Гальтона

регрессия имела только этот биологический смысл, но позже его работа была продолжена Удни Йолей и Карлом Пирсоном и выведена к более общему статистическому контексту.

В эконометрике широко используются методы статистики. Во многих практических задачах прогнозирования, изучая различного рода связи в экономических, производственных системах, необходимо на основании экспериментальных данных выразить зависимую переменную в виде некоторой математической функции от независимых переменных - регрессоров, то есть построить регрессионную модель.

Работа состоит из 4 глав:

- В первой главе "Линейная регрессия"вводятся понятия парной и множественной линейной регрессии, приводятся методы оценки значимости коэффициентов регрессионной модели и качества модели в целом.
- В главе "Нарушение некоторых предпосылок метода наименьших квадратов"приводятся проблемы, с которыми исследователи часто сталкиваются при анализе данных, способы их преодоления. Дается толкование понятия регуляризации, описывается Ridge и Lasso регрессии.
- В третьей главе, которая называется "Язык программирования R"описывается история развития языка R, а также приводятся основные функции, которые позволят реализовать Ridge и Lasso регрессию в R.
- Четвертая глава - "Использование Lasso и Ridge регрессии в портфельном инвестировании носит практический характер. В ней дается постановки исследуемой задачи и реализуются Lasso и Ridge регрессии на языке R.

Работа прошла апробацию на различных конференциях, в частности, в XIX Международной Саратовской зимней школе «Современные проблемы теории функций и их приложения», посвященной 90-летию со дня рождения академика П. Л. Ульянова, январь 2018 года на ежегодной студенческой конференции "Актуальные проблемы математики и механики которую проводил механико-математический факультет СГУ в апреле 2019 года, в секции "Анализ данных в VII Международной молодежной научно-практической конференции «Математическое и компьютерное моделирование в экономике, страховании и управлении рисками», ноябрь 2018 года.

1 Линейная регрессия

Существует много задач, требующих изучения отношения между двумя и более переменными. Для решения таких задач используется регрессионный анализ. В настоящее время регрессия получила широкое применение, включая задачи прогнозирования и управления. Целью регрессионного анализа является определение зависимости между исходной переменной и множеством внешних факторов (регрессоров).

1.1 Парная линейная регрессия

Название данного вида регрессии говорит само за себя. Это подход для прогнозирования количественного значения переменной Y на основе одной объясняющей переменной X. Он предполагает, что отношение между X и Y близко к линейному. Математически можно записать это отношение в следующем виде:

$$Y \approx \beta_0 + \beta_1 X. \tag{1.1}$$

Данное уравнение можно описать как регрессия Y от X [1]. В уравнении 1.1 β_0 и β_1 – две неизвестные константы, которые называют коэффициентами или параметрами модели. На основе обучающих данных (набора значений X и соответствующих им значений Y) необходимо найти оценки этих параметров. Подставив эти оценки в уравнение регрессии можно прогнозировать значения Y при определенных значениях X

$$\hat{y} \approx \hat{\beta}_0 + \hat{\beta}_1 x, \tag{1.2}$$

где $\hat{y}$ прогнозируемое значение Y при $X = x$. Здесь знак циркуфлекс («крышка») используется для обозначения оценочного значения неизвестного параметра или прогнозного значения переменной.

1.2 Множественная линейная регрессия

Парная линейная регрессия – полезный подход для прогнозирования значений результирующего признака на основе одной объясняющей переменной. Однако на практике, часто существует более одной объясняющей переменной.

Вместо того, чтобы строить отдельные парные регрессионные модели

для каждого предиктора, лучше расширить простую линейную модель, чтобы она могла непосредственно учитывать влияние нескольких объясняющих переменных. Следует каждому признак-фактору сопоставить свой коэффициент, отражающий влияние данного признака на результирующую переменную. В общем случае, предположим, что существует p отдельных объясняющих переменных. Тогда модель множественной регрессии принимает вид

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon, \tag{1.3}$$

где X_j – j-я объясняющая переменная, β_j – коэффициент, отражающий связь j-ой переменной и результирующей переменной. Коэффициент β_j отражает на сколько измениться значение переменной Y при изменении переменной X_j на единицу, при неизменных значениях остальных объясняющих переменных [2].

2 Нарушение некоторых предпосылок метода наименьших квадратов и их преодоление

Для получения методом наименьших квадратов наилучших результатов необходимо, чтобы выполнялся ряд предпосылок относительно случайного отклонения, которые носят название условий Гаусса-Маркова.

При построении классических линейных регрессионных моделей делаются еще некоторые предположения. Например:

- объясняющие переменные не являются случайными величинами;
- число наблюдений существенно больше числа объясняющих переменных (числа параметров уравнения);
- отсутствуют ошибки спецификации, т. е. правильно выбран вид уравнения и в него включены все необходимые переменные.

Часто полагают, что число наблюдений должно быть как минимум в 5-6 раз больше числа параметров уравнения (числа объясняющих переменных).

Более внимательно рассмотрим пятую предпосылку МНК, а также проблему снижения размерности.

2.1 Проблема мультиколлинеарности

Одним из основных условий построения уравнения множественной регрессии является независимость факторов, включенных в модель, между собой, т.е. соблюдение пятой предпосылки МНК.

Высокая взаимная коррелированность (взаимозависимость) объясняющих (независимых) переменных называется мультиколлинеарностью. Она может проявляться в функциональной (явной, полной) и стохастической (скрытой) формах [3].

Оценки становятся очень чувствительными к незначительному изменению результатов наблюдений и объема выборки. Уравнения регрессии в этом случае, как правило, не имеют реального смысла, так как некоторые из его коэффициентов могут иметь неправильные с точки зрения экономической теории знаки и неоправданно большие значения.

2.2 Уменьшение размерности

Под уменьшением размерности в машинном обучении подразумевается уменьшение числа признаков набора данных. Наличие в нем признаков избыточных, неинформативных или слабо информативных может понизить эффективность модели, а после такого преобразования она упрощается, и соответственно уменьшается размер набора данных в памяти и ускоряется работа алгоритмов.

Сокращение размерности может потребоваться когда данные избыточны в информационном плане, т.е. задачу можно решить с тем же уровнем эффективности и точности, но используя меньший объем данных. Это позволяет урезать время и вычислительные затраты на решение задачи [4].

Другой случай связан со слишком большими вычислительными затратами, требуемыми для обработки множества данного размера. Эта ситуация типична для алгоритмов, вычислительная сложность которых экспоненциально растет с увеличением числа наблюдений (т.е. немасштабируемых). Если в первом случае достаточно просто отобрать из всего множества столько признаков (атрибутов) и записей, сколько надо, то во втором, нужно сократить исходное множество до такого объема, который обеспечил бы реализуемость его обработки, невзирая на потерю полезной информации.

2.3 Регуляризация

Регуляризация — метод добавления некоторой дополнительной информации к условию с целью решить некорректно поставленную задачу. Эта информация часто имеет вид штрафа за сложность модели.

Методы регрессии Ridge и Lasso осуществляют регуляризацию параметров и позволяют преодолеть некоторые недостатки метода наименьших квадратов [5].

2.3.1 Ridge регрессия

Ridge регрессия очень похожа на метод наименьших квадратов, за исключением добавления «гребня».

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda\sum_{j=1}^{p}\beta_j^2 = RSS + \lambda\sum_{j=1}^{p}\beta_j^p, \qquad (2.1)$$

где $\lambda \geqslant 0$ - дополнительный параметр, который определяется отдельно. Уравнение 2.1 определяет два разных критерия. Как и в случае с методом наименьших квадратов, Ridge регрессия ищет оценки коэффициентов, которые хорошо подходят для данных, минимизируя RSS. Однако второй член, называемый штрафом за сокращение, тем меньше, чем ближе к нулю β_1,... , β_p, поэтому он приводит к стремлению оценок β_j к нулю.

В отличие от метода наименьших квадратов, который генерируют только один набор оценок коэффициентов, Ridge регрессия будет производить новый набор оценок коэффициентов для каждого значения λ [6].

Заметим, что в 2.1 штраф применяется к $\beta_1, ..., \beta_p$, но не к β_0.

Таким образом, Ridge-оценка является МНК-оценкой с ограничением нормы возможных решений (сферическое ограничение на параметры).

2.3.2 Lasso регрессия

У Ridge регрессии есть один недостаток. В отличие от лучшего подмножества, прямого и обратного выбора, который обычно определяет модели, которые включают только подмножество переменных, Ridge регрессия будет включать все предикторы p в финальной модели. Штраф λ сжимает все коэффициенты до нуля, но он не будет устанавливать ни одного из них точно в ноль (кроме случая, когда $\lambda = \infty$). Это не уменьшит точность предсказания, но это может создать проблему в интерпретации модели, в случае, когда число переменных p достаточно велико.

В отличие от Ridge регрессии, Lasso регрессия имеет несколько другое ограничение [7].

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p} \|\beta_j\| = RSS + \lambda \sum_{j=1}^{p} \|\beta_j\|. \qquad (2.2)$$

Коэффициент λ умножается на l_1-норму вектора $(\beta_1, \beta_2, ...\beta_p)$, тогда как в Ridge регрессии используется l_2-норма (рис. 2.1).

Положительным (в плане интерпретируемости модели) результатом такой замены нормы является тот факт, что Lasso, в отличие от Ridge регрессии, не только осуществляет регуляризацию, но и приравнивает некоторые из коэффициентов к нулю при достаточно большом значении λ [8]. То есть до-

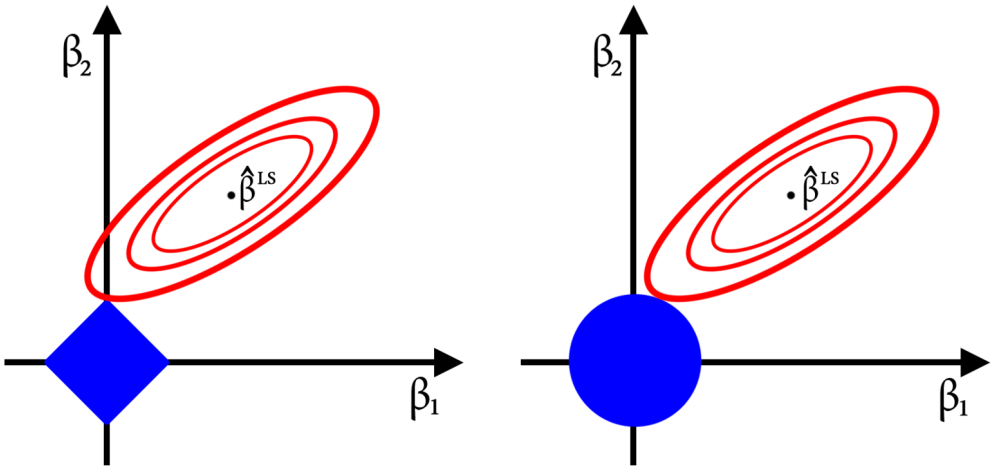

Рисунок 2.1 – Ограничения на норму весов. Случай слева соответствует l_1 норме, случай справа — l_2 норма

полнительно осуществляет выбор подмножества переменных, что позволяет легче интерпретировать модель.

2.3.3 Выбор значения λ для Ridge и Lasso

Сначала вся выборка случайно разделяется на Q блоков. Один из блоков рассматривается как контрольная выборка, а остальные $Q-1$ в совокупности составляют обучающую выборку. На практике Q обычно выбирают равным 5 или 10. Далее берется вектор $\lambda = [\lambda_s]$ с некоторым шагом, и для каждого из значений λ_s по обучающей выборке строится регрессионная модель. Для каждой модели вычисляется ошибка прогноза, то есть сумма квадратов остатков регрессии

$$RSS^q_{\lambda_x} = \sum_{i=1}^{n}(y_i - \sum_{j=0}^{p-1} |\hat{beta}_j(q,\lambda_s)x_{ij})^2, \tag{2.3}$$

где $q = \overline{1,Q}$ - номер блока, выбранного в качестве контрольной выборки. Далее вычисляется среднее значение этой ошибки по всем блокам:

$$MSE_{\lambda_s} = \frac{1}{Q}\sum_{q=1}^{Q} RSS^q_{\lambda_s}. \tag{2.4}$$

В качестве подходящего λ выбирается такое λ_s, при котором MSE_{λ_s} будет минимальной [9].

3 Использование Lasso и Ridge регрессии в портфельном инвестировании

Портфельные инвестиции (portfolio investments) это вложение средств в совокупность различных ценных бумаг с целью сохранения и извлечения прибыли. Совокупность ценных бумаг составляет портфель. Именно инвестиционный портфель позволяет получить такие характеристики при комбинации различных ценных бумаг, которые нельзя получить при инвестировании в отдельные финансовые инструменты.

3.1 Постановка задачи и исследуемые данные

Пусть задан набор 100 акций различных компаний, а также некий фондовый индекс. По каждой из акций есть данные о ее цене на момент закрытия торгов за последние 2 месяца. Задача состоит, в том, чтобы составить инвестиционный портфель (т.е. выбрать веса активов) таким образом, чтобы общая доходность портфеля была наиболее близка к доходности заданного индекса.

В качестве исходного набора возьмем 100 акций российских компаний, которые наиболее активно торговались 1 апреля 2019 года. Данные взяты с сайта https://finance.yahoo.com (Yahoo Finance). По каждой из акций возьмем статистику по цене на момент закрытия торгов за период с 1 февраля 2019 года по 29 марта 2019 включительно (40 рабочих дней). Как эталонный финансовый индекс будем использовать индекс РТС.

Данные о значениях индекса РТС за исследуемый период были взяты с сайта Московской Биржи https://www.moex.com. Все показания указаны в рублях.

3.2 Реализация Lasso и Ridge регрессии в R

Для загрузки данных из файла будем использовать функцию `read_excel` из пакета `readxl`. Из загруженных данных отбросим первый столбец, чтобы убрать дату из рассмотрения.

Для того, чтобы воспользоваться функцией `glmnet()` подготовим данные, выделив в отдельную матрицу регрессоры и отдельно в вектор зависимую переменную. В первую очередь реализуем функцию по сетке значений от $\lambda = 10^{10}$ до $\lambda = 10^{-2}$.

Следует отметить, что в нашей задаче, коэффициенты интерпретируются как веса соответствующего актива в конечном финансовом портфеле, поэтому следует добавить ограничение `lower.limits`.

```
> grid =10^ seq (10,-2, length =100)
> ridge.mod =glmnet (x,y,alpha =0, lambda =grid, lower.limits=rep(0,ncol(x)))
> lasso.mod =glmnet (x,y,alpha =1, lambda =grid, lower.limits=rep(0,ncol(x)))
```

Разделим данные на обучающий набор и тестовый. В качестве обучающей выборки, возьмем 30 более ранних значения, а в качестве тестовой оставшиеся более поздние данные (за 10 дней).

Далее используя кросс-валидацию найдем оптимальное значение λ для Ridge регрессии на обучающем наборе данных и вычислим тестовую ошибку.

```
> ridge.out =cv.glmnet (x.train,y.train,alpha =0, lower.limits=rep(0,ncol(x)))
> bestlam =ridge.out$lambda.min
[1] 155.398
> ridge.pred=predict (ridge.mod ,s=bestlam ,newx=x.test)
> mean((ridge.pred -y.test)^2)
[1] 153.1889
> ridge.coef=predict (ridge.mod,type ="coefficients",s=bestlam )[1:101 ,]
```

Для Lasso регрессии проделаем ту же процедуру.

```
> lasso.out =cv.glmnet (x.train,y.train,alpha =1, lower.limits=rep(0,ncol(x)))
> bestlam =lasso.out$lambda.min
[1] 0.2361912
> lasso.pred=predict (lasso.mod ,s=bestlam ,newx=x.test)
> mean((lasso.pred -y.test)^2)
[1] 76.03104
> lasso.coef=predict (lasso.mod,type ="coefficients",s=bestlam )[1:101 ,]
```

Для Lasso регрессии мы получили меньшее значение средней ошибки, чем для Ridge регрессии. В случае Lasso регрессии можно наблюдать, как и ожидалось, что большое число коэффициентов получилось равно 0 (75 из 100, а в случае Ridge регрессии только 42 из 100). Это облегчит процесс работы с портфелем, а также интерпретируемость модели. Однако, значение λ получилось довольно маленьким, что может говорить о том, что в модели учитываются шумы и различные побочные факторы.

ЗАКЛЮЧЕНИЕ

В данной работе были изучены такие темы, как парная и множественная линейна регрессия, проблема мультиколлинеарности и снижения размерности, а также Ridge и Lasso регрессия. Ridge и Lasso регрессии были рассмотрены на примере тестовых данных, для составления инвестиционного портфеля из 100 возможных активов, по статистическим данным за 40 рабочих дней.

Ridge регрессия – усовершенствование линейной регрессии с повышенной устойчивостью к ошибкам, налагающая ограничения на коэффициенты регрессии для получения куда более приближенного к реальности результата. Вдобавок, этот результат гораздо проще интерпретировать. Применяется метод для борьбы с переизбыточностью данных, когда независимые переменные коррелируют друг с другом (мультиколлинеарность).

Lasso регрессия сходна с Ridge, за исключением того, что коэффициенты регрессии могут равняться нулю (часть признаков при этом исключается из модели).

Оба метода успешно решают проблемы мультиколлинеарности, переобучения, и уменьшают разброс коэффициентов. Ridge регрессия использует все признаки, стараясь «выжать максимум» из всей имеющейся информации. Lasso производит отбор признаков, что предпочтительнее, когда среди признаков имеются шумовые, или измерения признаков связаны с ощутимыми затратами. Полученные методом кросс-валидации значения λ, дают наилучшие результаты при получении прогнозных значений модели.

В нашей задаче слежения за индексом, лучший результат дала регрессия Lasso. В этом случае была не только получена наименьшая средняя оценка отклонения, но и значительно сокращен размер портфеля, что делает работу с ним более легкой и мобильной. Однако, Ridge регрессия также дала приемлемые результаты. К тому же, так как из-за особенности интерпретации коэффициентов модели в нашей задаче, было наложено дополнительное ограничение, некоторые из коэффициентов также получились равными нулю.

Решение о использовании той или иной модели следует принимать индивидуально в каждой решаемой задаче, исходя из особенностей области исследования и исходных наборов данных. У каждой модели есть свои пре-

имущества и недостатки. В общем случае, когда количество наблюдений в несколько раз превосходит количество объясняющих переменных, метод наименьших ошибок дает лучшие результаты, однако использование Lasso и Ridge регрессии позволяет облегчить процесс интерпретации модели и избавиться от лишних факторов, а также преодолеть некоторые особенности набора данных, затрудняющие использование метода наименьших квадратов.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *G. Casella S. Fienberg, I. O.* An Introduction to Statistical Learning with Applications in R / I. O. G. Casella, S. Fienberg. — New York: Springer, 2015.

2 *Носков, В.* Эконометрика: учебник для студентов высш. учеб. заведений, обучающихся по экон. специальностям / В. Носков. — Москва: Дело, 2011.

3 *M.H. Kutner C.J. Nachtsheim, J. N.* Applied Linear Regression Models / J. N. M.H. Kutner, C.J. Nachtsheim. — McGraw-Hill Irwin, 2004.

4 *T. Hastie R. Tibshirani, J. F.* The elements of statistical learning: Data Mining, Inference, and Prediction / J. F. T. Hastie, R. Tibshirani. — Springer, 2009.

5 *Tibshirani, R.* The lasso problem and uniqueness / R. Tibshirani. — Electronic Journal of Statistics 7, 2013.

6 *Хардле, В.* Прикладная непараметрическая регрессия / В. Хардле. — М.: Мир, 1993.

7 *Tibshirani, R.* Sparsity and the Lasso / R. Tibshirani. — 2015.

8 *T. Hastie R. Tibshirani, M. W.* Statistical Learning with Sparsity. The Lasso and Generalizations / M. W. T. Hastie, R. Tibshirani. — McGraw-Hill Irwin, 2015.

9 *P. Cortez A. Cerdeira, F. A.* Modeling wine preferences by data mining from physicochemical properties / F. A. P. Cortez, A. Cerdeira // *Decision Support Systems*. — 2009. — Vol. 47. — Pp. 547–553.