Эконометрика: борьба за предположения

На этой неделе мы попытаемся рассмотреть несколько способов, как добиваться выполнения предположений линейной регрессии.

8.1 Ещё раз вспомним предположения

Выпишем ещё раз предположения Гаусса-Маркова, они же предположения линейной регрессии, которая в сокращённом виде записывается как

$$Y = X\theta + \varepsilon, \quad \theta \in \mathbb{R}^k, \quad X \in \mathbb{R}^{n\times k}, \quad \varepsilon \in \mathbb{R}^n.$$

1. Модель корректна: все признаки, влияющие на Y учтены;
2. X детерминирована и колонки линейно независимы;
3. ε_i независимы в совокупности и одинаково распределены, не зависят от наблюдений X_i, имеют нулевое матожидание и одинаковую дисперсию σ^2.

На прошлой неделе мы пытались детектировать нарушения этих предположений, сейчас мы попробуем привести несколько способов, позволяющих эти нарушения смягчить.

8.2 Мультиколлинеарность

Мультиколлинеарность – это коррелированность признаков или вообще их линейная зависимость. На практике это приводит прежде всего к тому, что посчитать оценку МНК

$$\hat{\theta} = (X^T X)^{-1} X^T Y$$

оказывается сложно. Может показаться, что итерационные методы этим недостатком не обладают, но это не так: число обусловленности матрицы $X^T X$ (которое может быть очень высоко в случае мультиколлинеарности) неизбежно влияет и на скорость сходимости итерационных методов. Есть фундаментально как минимум два общих подхода для избавления от мультиколлинеарности и каждый из них не является безусловно хорошим во всех случаях.

1. *Выбрасывание части признаков.* Это чаще делают исходя из практических соображений (например, нет смысла учитывать в модели площадь квартиры, если она уже

учитывает в отдельности площади всех комнат), но есть и другие эвристики вроде FeatureImportance и взаимной иноформации (имплементировано в `sklearn`). Ещё одним способом для определения важных признаков может выступить сингулярное разложение и анализ главных компонент (Principal Component Analysis, PCA), позволяющий аггрегировать линейно несколько признаков, уменьшая размерность.

2. *Техники регуляризации.* Это, к примеру, те самые хорошо известные Lasso- и Ridge-регрессия. В целевую функцию добавляется дополнительный член, который определённым образом штрафует за большие веса. В случае Ridge-регрессии даже можно несложно выписать аналитическое решение

$$\hat{\theta}_R = (X^T X + \lambda I)^{-1} X^T Y$$

с параметром регуляризации λ.

Нас интересует два вопроса: насколько мы способны решить проблему мультиколлинеарности и что мы теряем в плане статистики и применимых критериев.

Первый подход при удачном стечении обстоятельств способен хорошо адресовать оба вопроса, однако если мы заранее не знаем очевидных линейных зависимостей, это может привести к тому, что мы выбросим важную информацию. В этом смысле PCA является компромиссным подходом. Его идея состоит в том, что нужно построить сингулярное разложение матрицы плана $X = U\Sigma V^*$, а дальше урезать строки в V^*, отвечающие компонентам с малыми сингулярными числами, получив V^*_{trunc}. Преобразование данных (проекция) на первые p главных компонент осуществляется умножением на V_{trunc}: $X_{new} = XV_{trunc}$. Так, это просто замена переменных (реализована в `sklearn.decompositions.PCA`) и если её детально исследовать, можно попробовать сохранить интерпретируемость. Кроме того, из-за линейности замены остаётся открытой возможность тестирования гипотез в новой модели линейной регрессии:

$$Y = XV_{trunc}\theta_{trunc} + \varepsilon.$$

Регуляризация отлично адресует вычислительные проблемы на практике. Вместо задачи наименьших квадратов

$$min_\theta \|X\theta - Y\|_2^2$$

рассматриваем регуляризованную задачу

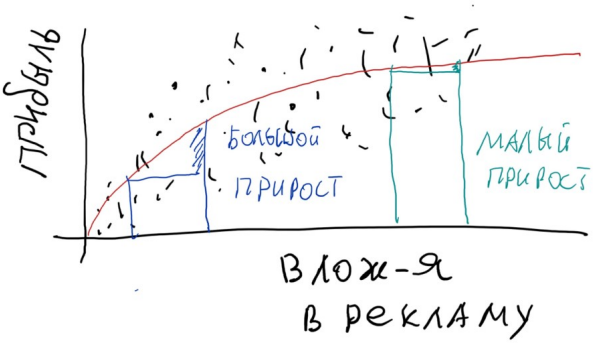
$$min_\theta \|X\theta - Y\|_2^2 + \lambda\|\theta\|_p^p,$$

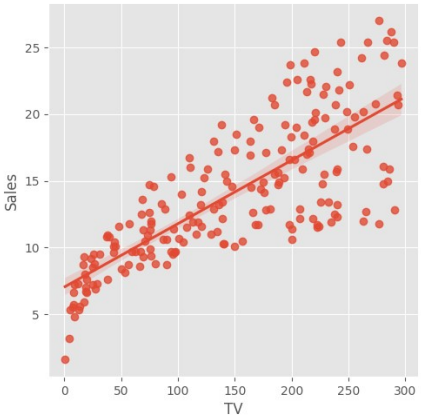
которая хорошо известна в области машинного обучения под именами LASSO- (при $p = 1$) и Ridge-регрессии (при $p = 2$). Однако в такой задаче возможность классической проверки гипотез теряется, так как оценка параметров будет смещена, хотя есть некоторые исследования, которые пытаются сохранить что-то похожее на t-тесты [3, 8].

8.3 Нелинейность модели

Модель, конечно, может быть в реальности сильно нелинейной. Некоторые нелинейные модели можно представить как линейные, просто добавив новых признаков, но есть и другие популярные приёмы.

Пример 8.1. *Например, тем, кто изучал микро- и макроэкономику, знакомы эффекты убывающей отдачи и эффект убывающей производной полезности.*





Первый связан с тем, что чем больше денег инвестируется в какой-то продукт, тем всё меньший эффект (отдачу) можно будет получить за счёт новых инвестиций. Второй очень похож и адресует одного потребителя: если вы едите очень вкусную пиццу каждый день в течение полугода, то через некоторое время она надоедает и кажется уже не такой вкусной, как в начале. Подобные эффекты насыщения объясняются чисто экономическими или физическими/химическими/инженерными предпосылками. Например, аудитория продукта конечна, и в какой-то момент все уже и так знают про ваш новый продукт и продажи не растут.

Можно предположить степенной закон $y(x) = x^\alpha$, тогда если мы преобразуем данные

$$y \to \ln y, \;\; x \to \ln x,$$

то модель

$$\ln y = \alpha \ln x + \varepsilon$$

уже будет линейной, а при условии соблюдения других предпосылок можно будет даже проверять гипотезы.

Пример 8.2. *Многие финансовые ряды имеют нелинейные тренды. Например, вспомним классическую модель геометрического Броуновского движения:*

$$X_k = X_0 e^{(\mu - \sigma^2/2)k + \sigma W_k},$$

где $W_k = \sum_{i=1}^{n} \varepsilon_i$, сумма независимых одинаково гауссовски распределённых шумов с нулевым средним.

В таком модельном предположении мы можем прологарифмировать обе части и увидеть, что

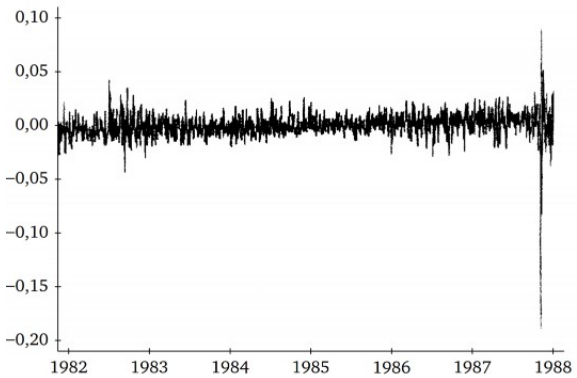
$$\ln X_k = \ln X_0 + (\mu - \sigma^2/2)k + \sigma W_k.$$

Это уже линейная модель с гауссовским шумом, но изначально гетероскедастичная и с авткоррелированными остатками. Скорректированную модель можно получить, если действовать, как мы ранее:

$$\ln X_k = \ln X_{k-1} + (\mu - \sigma^2/2) + \sigma\varepsilon_k.$$

Здесь в результате получается линейная модель с константой и неизвестной дисперсией шума. Многие реальные данные почти похожи на эту модель, но оказывается, что в реальности шумы ведут себя не так просто: они коррелированы, гетероскедастичны и имеют распределение с хвостами тяжелее, чем у нормального. Для их описания используются более сложные модели стохастической волатильности, где сама волатильность σ зависит от времени или даже сама является случайным процессом[13].

Индекс SandP-500 X_t



Доходности $\ln X_t/X_{t-1}$.

8.4 Автокорреляция

Обычно автокорреляция возникает вполне естественным образом, когда мы имеем дело с временными рядами. Временной ряд – это счётный набор случайных величин $..., X_0, X_1, X_2, ..$, проиндексированных по условному *времени* и как-то зависимых между собой. Эта последовательность в теории может продолжаться в обе стороны. Представьте, если бы мы наблюдали численность населения Земли, то X_0 могла бы быть численность населения в Рождество Христово, X_{-2} – во 2м году до н.э., а X_{2023} – численность населения сейчас, в 2023м году; точку отсчёта, конечно, можно изменить.

Пример 8.3. *Самой первой и самой простой моделью роста населения принято считать модель Мальтуса, которая задаётся как*

$$P_t = P_0 e^{rt}$$

и записана в непрерывном времени. Если бы мы хотели как-то учесть случайности, то мы могли бы предложить модель в дискретном времени вроде авторегрессии

$$P_t = \alpha P_{t-1} + X_t^T \theta + \varepsilon_t,$$

где X_t принимает значения в $\mathbb{R}^d$ и отражает прочие факторы. Если бы мы не подумали про первый член αP_{t-1}, когда составляли модель, то мы получили бы коррелированные остатки, о чём, вероятно, нам бы просигнализировали критерии.

Если мы видим коррелированные остатки, то естественным образом нужно пытаться подгонять модель некоторого временного ряда. Про это можно узнать подробнее позже в курсе временных рядов. Общая рекомендация может быть такой: если есть понятная структура времени, то нужно пытаться её учесть.

В примере выше есть два пути:

- Изначально рассмотреть модель $P_t = X_t^T\theta + \varepsilon_t$, а после оценки θ остатки $\hat{\varepsilon}_t = P_t - X_t^T\hat{\theta}$ пытаться описать временным рядом (например, авторегрессией, ARiMA или более сложными моделями);
- Изначально рассмотреть модель $P_t = \alpha P_{t-1} + X_t^T\theta + \varepsilon_t$, а после проверять предположения линейной регрессии.

Автокорреляция – одно из тех явлений, которым интересуется статистика временных рядов и мы подробно не затрагиваем эту тему.

8.5 Гетероскедастичность

Гетероскедастичность – это наличие разных дисперсий шумов в модели линейной регрессии. В самом сложном случае шумы могут иметь ковариационную матрицу V (тогда возможна и автокорреляция), если же V диагональна, а шумы гауссовские, то они независимы и мы имеем только разную дисперсию, такую матрицу для n наблюдений можно записать так:

$$V = \begin{bmatrix} \sigma_1^2 & 0 & 0 & \dots & 0 \\ 0 & \sigma_2^2 & 0 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & 0 & \sigma_n^2 \end{bmatrix}.$$

Однако, как мы замечали ранее, и тут у нас проблема: нам нужно оценить n дисперсий по n наблюдениям, что сделать непросто. Если бы мы знали матрицу V, то можно было бы выписать оценку взвешенного МНК

$$\hat{\theta} = (X^T V X)^{-1} X^T V Y.$$

Наша задача ниже построить оценку V, которую можно было бы использовать для коррекции и оценки дисперсии оценки коэффициентов.

8.5.1 Кластеры

Первый случай, который мы рассматривали был посвящён тому, что есть некоторое конечное число различных дисперсий и мы можем из практических или эвристических соображений поделить выборку на кластеры. Далее мы строим одну общую регрессию

$$Y = X\theta + \varepsilon,$$

оцениваем остатки

$$e_i = Y_i - X_{i,\cdot}\hat{\theta},$$

а далее строим матрицу $\hat{V}$ для коррекции, ставя на диагонали выборочные оценки дисперсии $\hat{\sigma}_i$, посчитанные в каждом кластере. Потом можно строить регрессию с взвешенным методом наименьших квадратов, которая будет смещённая, но всё ещё состоятельная.

8.5.2 Зависимость дисперсии от регрессора

В случае, когда есть предположение, что, к примеру, дисперсия шума $\sigma_i^2 = \sigma^2 X_{i,j}$, то можно пробовать корректировать задачу наименьших квадратов как

$$\min_{\theta} \sum_{i=1}^{n} \frac{(X_{i,\cdot}\theta - Y_i)^2}{X_{i,j}}.$$

Подобные зависимости, к примеру, характерны для данных продаж, где при больших вложениях в разных случаях изменение в продажах может сильно разниться.

8.5.3 Оценка ковариационной матрицы оценок МНК

Предположим, что мы применили метод наименьших квадратов и хотим понять дисперсию оценок коэффициентов регрессии. Если бы мы изначально знали V, то могли бы после некоторого количества формул выписать ковариационную матрицу вектора оценок

$$Var(\hat{\theta}) = (X^T X)^{-1} X^T V X (X^T X)^{-1},$$

при $V = \sigma^2 I$ это вырождается в $\sigma^2 (X^T X)^{-1}$. Знание такой дисперсии позволяет нам делать тесты на значимость коэффициентов вроде $t-$ и $z-$тестов. Вопрос состоит в следующем:

как можно состоятельно оценить $Var(\hat{\theta})$, не зная наперёд V.

Некоторое количество коррекций было предложено на основе исследования остатков метода наименьших квадратов $e_i = Y_i - X_{i,\cdot}\hat{\theta}$. Уайт в 1980 [10] предложил способ состоятельной оценки, который стал называться HC0 (Heteroscedasticity consistent):

$$HC0(\hat{\theta}) = (X^T X)^{-1} X^T diag(e_i^2) X (X^T X)^{-1},$$

а немного позже появилась небольшая коррекция

$$HC1(\hat{\theta}) = \frac{n}{n-k} HC0(\hat{\theta}),$$

которая учла количество степеней свободы и сделала метод полезнее для маленьких выборок, при росте n два метода эквивалентны.

Оказывается, правда, что с остатками не всё так просто и можно пойти ещё дальше. Даже когда ошибки гомоскедастичны с $V = \sigma^2 I$, остатки регрессии *не являются гомоскедастичными*:

$$Var(e_i) = \sigma^2(1 - h_i) \neq \sigma^2,$$

где, если точнее,

$$h_i = X_{i,\cdot}(X^T X)^{-1} X_{i,\cdot}^T.$$

Поэтому Уайт с соавторами предложили

$$HC2(\hat{\theta}) = (X^T X)^{-1} X^T diag(e_i^2/(1 - h_i)) X (X^T X)^{-1}$$

и

$$HC3(\hat{\theta}) = (X^T X)^{-1} X^T diag(e_i^2/(1 - h_i)^2) X (X^T X)^{-1},$$

последний метод сейчас самый популярный и реализован, например, в `statsmodels`. Все эти оценки являются состоятельными [9].

8.6 Эндогенность

Эндогенность – это зависимость шумов от регрессоров. Она может проявляться по-разному, например,

1. Гетероскедастичность – частный случай эндогенности;
2. Неправильная спецификация модели.

Первый случай для нас понятен, второй мы пока не затрагивали. Почему неправильная спецификация может приводить к эндогенности? Насколько это критично?

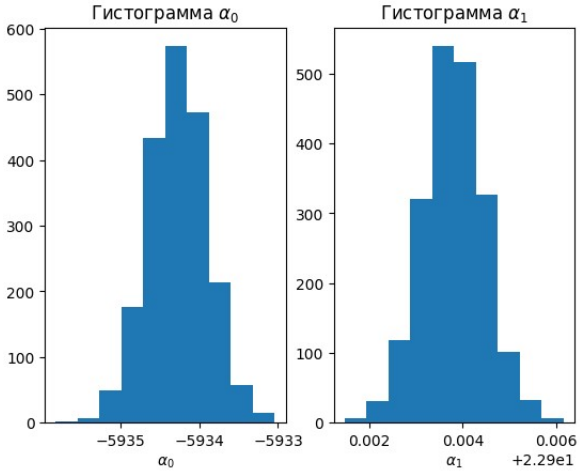
Рассмотрим пример (на семинаре можно найти код). Положим, что настоящая модель – это

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2.} + u_i,$$

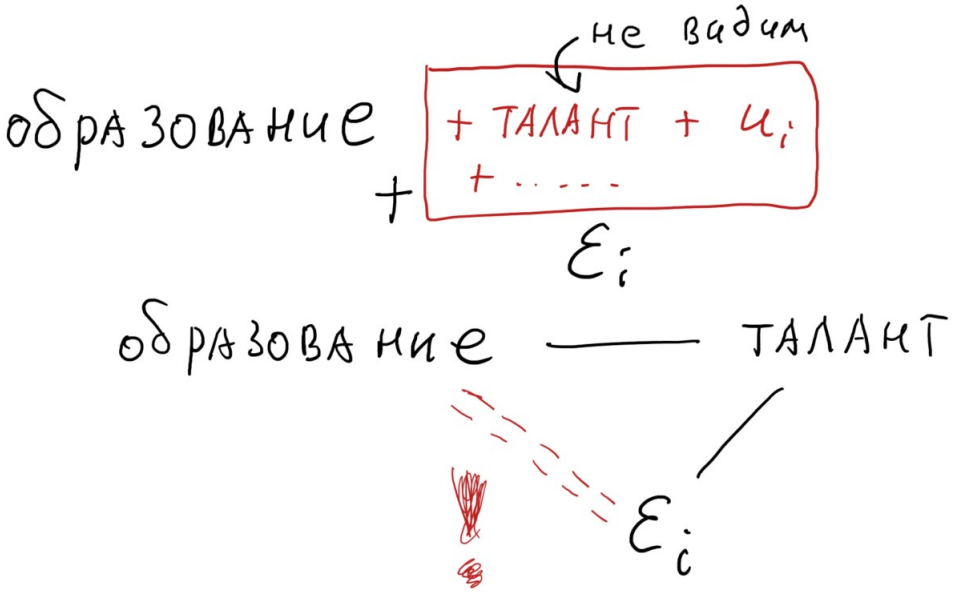
где ε_i – гомоскедастичные некоррелированные шумы. Допустим, нам никто не дал данных по X_2, но коэффициент β_2 значимо отличен от нуля. Попробуем оценить регрессию по X_1 и посмотрим на оценки $\hat{\alpha}_1, \hat{\alpha}_2$ в модели

$$Y_i = \alpha_0 + \alpha_1 X_{i1} + \varepsilon_i.$$

Зададим, например, $\beta_0 = 1, \beta_1 = 3, \beta_2 = 5$ и попробуем много раз сгенерировать датасет и попробовать посчитать оценку α. Гистограммы полученных оценок удручают:



Но почему так плохо? Причина в том, что из-за того, что X_2 был значимым и неучтённым, шумы u_i выражаются через X_1, что приводит к большим смещениям в оценках. Для того, чтобы понять как исследовать проблему зависимости ε от X, нам придётся отказаться от детерминированности X_1, X_2 и положить их случайными. Тогда мы можем, в частности, использовать факт корреляции между X_1, X_2 и корреляции между X_2, u.



Для парной регрессии мы можем явно посчитать оценки

$$\hat{\alpha}_0 = \overline{Y} - \hat{\alpha}_1 \overline{X}_1, \quad \hat{\alpha}_1 = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_{i1} - \overline{X}_1)}{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2}.$$

Попробуем взять матожидание, например, от $\hat{\alpha}_1$, чтобы понять, что происходит со смещением. Нам будет удобно использовать формулу полного матожидания, заморозив X_1:

$$\mathbb{E}\left[\hat{\alpha}_1\right] = \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{E}\left[Y_i - \overline{Y} \mid X_{i1}\right](X_{i1} - \overline{X}_1)}{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2}\right].$$

Рассмотрим внутреннее матожидание и подставим вместо Y_i и в $\overline{Y}$ истинную модель :

$$\begin{aligned}
&\mathbb{E}\left[Y_i - \overline{Y} \mid X_{i1}\right] = \\
&= \mathbb{E}\left[\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i - \beta_0 - \beta_1 \overline{X}_1 - \beta_2 \overline{X}_2 - \overline{\varepsilon} \mid X_{i1}\right] = \\
&= \beta_1 (X_{i1} - \overline{X}_1) + \beta_2 \mathbb{E}\left[(X_{i2} - \overline{X}_2) \mid X_{i1}\right] + \mathbb{E}\left[\varepsilon_i - \overline{\varepsilon} \mid X_{i1}\right].
\end{aligned}$$

Подставим аккуратным образом наверх:

$$\mathbb{E}\left[\hat{\alpha}_1\right] = \beta_1 + \beta_2 \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{E}\left[X_{i2} - \overline{X}_2 \mid X_{i1}\right](X_{i1} - \overline{X}_1)}{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2}\right] + \mathbb{E}\left[\frac{\sum_{i=1}^n \mathbb{E}\left[\varepsilon_i - \overline{\varepsilon} \mid X_{i1}\right](X_{i1} - \overline{X}_1)}{\sum_{i=1}^n (X_{i1} - \overline{X}_1)^2}\right].$$

Воспользовавшись несмещённостью ковариации и дисперсии, мы можем показать

$$\mathbb{E}\left[\hat{\alpha}_1\right] = \beta_1 + \beta_2 \frac{\mathrm{Cov}\left(X_2, X_1\right)}{\mathrm{Var}\left[X_1\right]} + \frac{\mathrm{Cov}\left(\varepsilon, X_1\right)}{\mathrm{Var}\left[X_1\right]},$$

Поэтому член с β_2 называют *covariance bias* (он возникает из зависимости X_2 и X_1), а последний – *stochastic bias* – возникает из зависимости шумов от регрессора X_1. Важно, что смещение может быть огромным, даже если шумы от X_1 явно не зависят (так было в модели миссспецификации). Рост выборки это смещение не убирает и с ним нужно бороться другими методами: с помощью инструментов, о которых можно узнать в семинаре.

Литература

[1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit"Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.

[2] Peter C Austin, Muhammad M Mamdani, David N Juurlink, and Janet E Hux. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, 59(9):964–969, July 2006.

[3] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, Sep 2011.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[6] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[7] S.J. Levitt, S.D. Dubner. *Freakonomics*. NY: Harper Trophy, 2006.

[8] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[9] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[10] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[11] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных*. Изд. дом Высшей школы экономики, 2012.

[12] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования*. М.: МЦНМО, 2013.

[13] А.Н. Ширяев. *Основы стохастической финансовой математики*. МЦНМО, 2016.