

# Эконометрика: предположения

*В этой лекции мы разберём в общем, как можно проверять предположения линейной регрессии.*

## 9.1 Вспомним предположения

Давайте вспомним 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.  $\varepsilon_i$  независимы в совокупности и одинаково распределены, не зависят от наблюдений  $X_i$ , имеют нулевое матожидание и одинаковую дисперсию  $\sigma^2$ .

При верных предположениях оценка методом наименьших квадратов является несмещённой и эффективной (в классе всех несмещённых оценок, Best Linear Unbiased Estimate). Об этом говорит теорема Гаусса-Маркова. Если есть эти предположения и гауссовость остатков, то мы получаем большой арсенал критериев на проверку самых разных модельных гипотез.

Но эти предположения могут нарушаться, в реальности так происходит почти всегда.

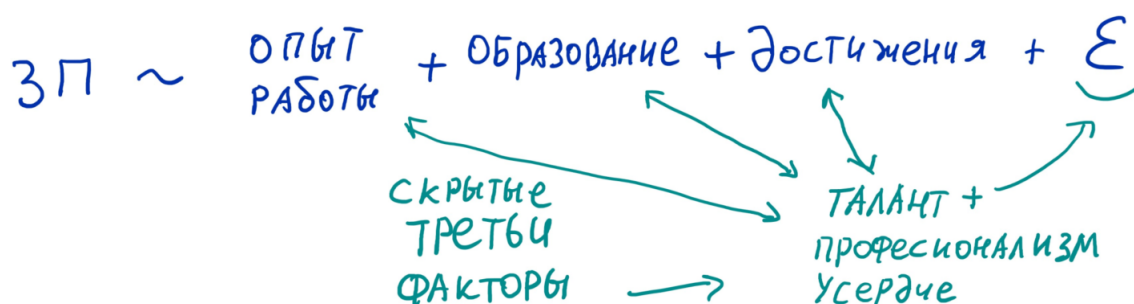
1. Модель *не идентифицирована* (не все факторы учтены) – это крайне сложно проверить и обычно модель строится из каких-то практических соображений, опыта и экспертных мыслей (например, со знанием макроэкономики или финансовой математики). Часто нужно обратиться к теории за идеями о том, какие признаки включать в модель. Для проверки полноты модели можно использовать  $R^2$ , если выполнены другие предположения регрессии.
2. Если колонки в  $X$  линейно зависимы или число обусловленности матрицы  $X^T X$  очень большое (столбцы примерно линейно зависимы), то это называют проблемой *мультиколлинеарности*.
3. Остатки  $\varepsilon_i$  могут быть коррелированы (*автокорреляция*).

АВТОКОРРЕЛЯЦИЯ:

$\rightarrow \varepsilon_t \rightarrow \varepsilon_{t+1} \rightarrow \varepsilon_{t+2} \rightarrow \dots$   
временные ряды?

4. Остатки  $\varepsilon_i$  могут быть зависимы от  $X_i$ . (эндогенность) в силу различных факторов. Например, дисперсия зарплат выше у людей с более высоким уровнем образования, цены на квартиры по-разному изменчивы изменяются в зависимости от площади и так далее.

эндогенность повсюду:



5. Очень близкая и пересекающаяся с предыдущей – проблема *гетероскедастичности* (разной дисперсии остатков), случай с одинаковой дисперсией остатков называется *гомоскедастичностью* и для проверки этой проблемы существуют специальные тесты на гомоскедастичность.



Про то, как эти проблемы можно решать, это отдельная большая тема. Давайте пока посмотрим, как можно детектировать эти проблемы с помощью статистических критериев.

## 9.2 Проверка некоррелированности остатков

В эконометрике много внимания уделяется автокорреляции, здесь для краткости мы рассмотрим один классический критерий, но который применяется почти всегда, и один более

общий.

### 9.2.1 Критерий Дарбина-Уотсона

Предположим, что мы применили метод наименьших квадратов и получили оценку  $\hat{\theta}$  и оценки остатков

$$\hat{\varepsilon}_t = Y_t - X_t \hat{\theta}.$$

Наша задача теперь состоит в том, чтобы понять, что за последовательность остатков получилась. Один из вариантов – попробовать смоделировать остатки с помощью авторегрессии

$$\varepsilon_t = \rho \varepsilon_{t-1} + \eta_t,$$

где  $\eta_t$  – последовательность независимых и одинаково распределённых нормальных шумов с нулевым средним и одинаковой дисперсией. Оказывается, что мы можем по имеющимся данным проверить гипотезу

$$H_0 : \rho = 0, \quad H_A : \rho \neq 0.$$

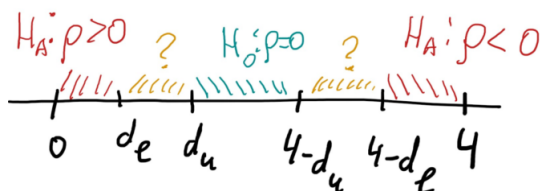
Если принимается гипотеза, то зависимости шумов типа авторегрессии нет, а если гипотеза отвергается, то в шумах есть автокорреляция и её можно убрать, совершив преобразование исходных данных. Данную гипотезу проверяет *критерий Дарбина-Уотсона*, имеющий статистику

$$T = \frac{\sum_{t=2}^n (\hat{\varepsilon}_t - \hat{\varepsilon}_{t-1})^2}{\sum_{t=1}^n \hat{\varepsilon}_t^2}$$

с табличным распределением при верной гипотезе. Эта статистика, как оказывается, принимает значения от 0 до 4 и критерий имеет одну особенность: у него есть зоны неопределённости, то есть, те значения статистики, где невозможно склониться ни к  $H_0$ , ни к  $H_A$ .

Подробнее, решение после выбора двух квантилей  $d_l, d_u$  принимается так:

1.  $T \in (0, d_l)$ , гипотеза отвергается, принимается  $H_A : \rho > 0$ ;
2.  $T \in (d_l, d_u)$ , гипотеза не отвергается и не принимается;
3.  $T \in (d_u, 4 - d_u)$ , гипотеза принимается;
4.  $T \in (4 - d_u, 4 - d_l)$ , гипотеза не отвергается и не принимается;
5.  $T \in (4 - d_l, 4)$ , гипотеза отвергается, принимается  $H_A : \rho < 0$ .



Может показаться, что тест проверяет корреляционную структуру слишком простого вида и что зависит сильно от порядка наблюдений, но это не совсем так. Если *настоящий* порядок наблюдений в структуре корреляций отличается от того, который получился, то в тестовой статистике всё равно соседние наблюдения будут скоррелированы в силу того, как устроен процесс авторегрессии.

**Пример 9.1.** Для процесса авторегрессии  $\varepsilon_t = \rho\varepsilon_{t-1} + \eta_t$ , где  $\eta_t$  – независимые и одинаково распределённые с дисперсией  $\sigma^2$  и нулевым матожиданием, вычислите корреляцию между  $\varepsilon_t$  и  $\varepsilon_{t-k}$ . Она убывает с ростом  $k$  со скоростью геометрической прогрессии. Если же зависимость более общая

$$\varepsilon_t = \sum_{i=1}^p \rho_i \varepsilon_{t-i} + \eta_t,$$

то критерием Дарбина-Уотсона она тоже уловится. Есть, правда, тонкость с зависимостями вроде

$$\varepsilon_t = \rho\varepsilon_{t-k} + \eta_t,$$

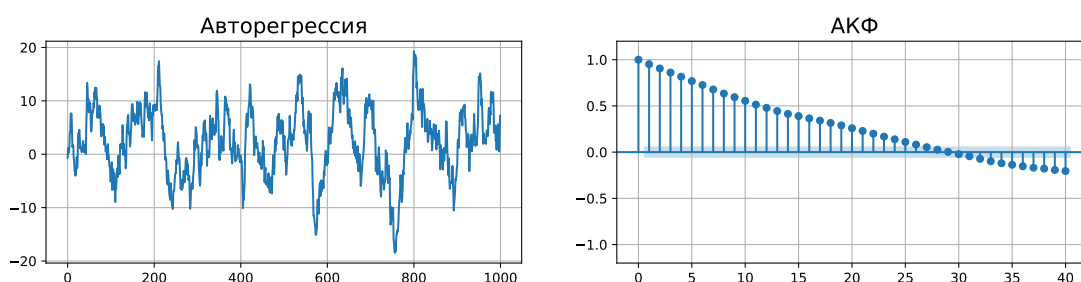
но такие случаи обычно обозначают некоторую сезонность в данных, которую можно отследить другими средствами (например, с помощью автокорреляционной функции).

Если  $\rho = 1$ , то значение статистики будет близко к нулю; если же  $\rho = -1$ , то статистика близка к 4. Поэтому критерий отвергает гипотезу, когда значения слишком маленькие или слишком большие. Наличие областей неопределённости можно неформально связать с тем, что сама статистика положительна, а в зонах неопределённости неясно, положительное  $\rho$ , отрицательное, или же вообще 0. Таблицу проверки выше можно использовать и для односторонних альтернатив  $H_A : \rho > 0$  и  $H_A : \rho < 0$ , если скорректировать уровень значимости квантилей.

### 9.2.2 Критерий Льюнга-Бокса и Бокса-Пирса

Вместо того, чтобы рассматривать корреляцию на один шаг, можно добавлять больше, но в авторегрессии это технически сложнее. Можно подойти со стороны автокорреляционной функции (АКФ)

$$\rho_p = \text{corr}(\varepsilon_{t+p}, \varepsilon_t).$$



Критерий *Льюнга-Бокса* использует выборочную АКФ  $\hat{\rho}_k$ , где просто вместо точной корреляции подставляется выборочная, подсчитанная по данным, и проверяет

$$H_0 : \rho_1 = \rho_2 = \dots = \rho_p = 0, \quad H_A : \exists \rho_j \neq 0.$$

Статистика критерия

$$T = n(n+2) \sum_{j=1}^p \frac{\hat{\rho}_j^2}{n-j}.$$

имеет асимптотическое распределение  $\chi^2(p)$ . Приближённой версией этого критерия является критерий *Бокса-Пирса*, который имеет похожую статистику

$$T = n \sum_{j=1}^k \hat{\rho}_j^2.$$

с тем же асимптотическим распределением, но критерий Льюнга-Бокса оказывается ближе к распределению хи-квадрат.

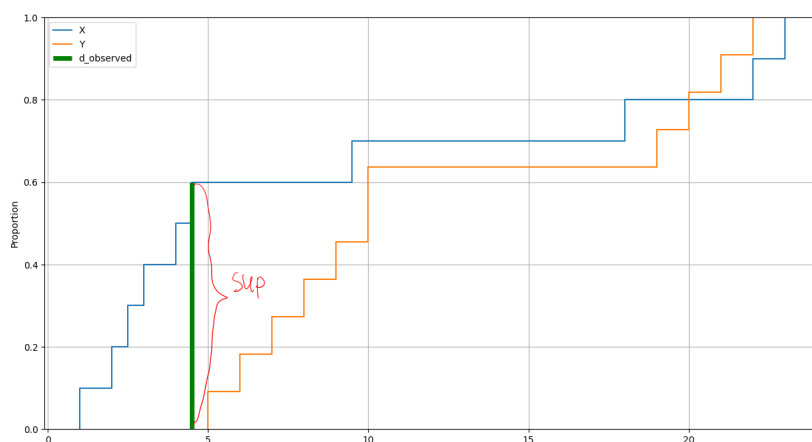
Важно заметить, что в отличие от критерия Дарбина-Уотсона эти критерии асимптотические и требуют много данных. Обычно не проверяют корреляций очень высоких порядков, но нужно понимать также, что в высоких порядках выборочные оценки корреляции сильно затрудняются из-за ограниченности выборки. Данные критерии и многие похожие особенно развиты в области временных рядов, где подобные модели автокорреляции занимают особое место.

### 9.3 Проверка гауссовости

Если мы уже приняли решение о том, что, вероятно, остатки  $\varepsilon_1, \dots, \varepsilon_n$  некоррелированы, то чтобы пользоваться классическими критериями для параметров линейной регрессии нужно либо много данных, чтобы использовать асимптотические критерии, либо проверить гауссовость остатков, чтобы убедиться, что можно использовать критерии для регрессии с гауссовскими остатками.

Важно помнить, что, как мы в какой-то момент упоминали, критерии согласия проверяют гипотезу против очень общей альтернативы и для получения достаточной мощности критерия нужно набрать достаточно большую выборку.

### 9.3.1 Критерий Колмогорова



Для проверки гипотезы о распределении у нас есть критерии согласия и главный из них – это критерий Колмогорова и похожие на него. Критерий Колмогорова (его называют также *k-тестом*) используется для проверки гипотезы о распределении для выборки из непрерывных случайных величин, конкретнее, проверяем

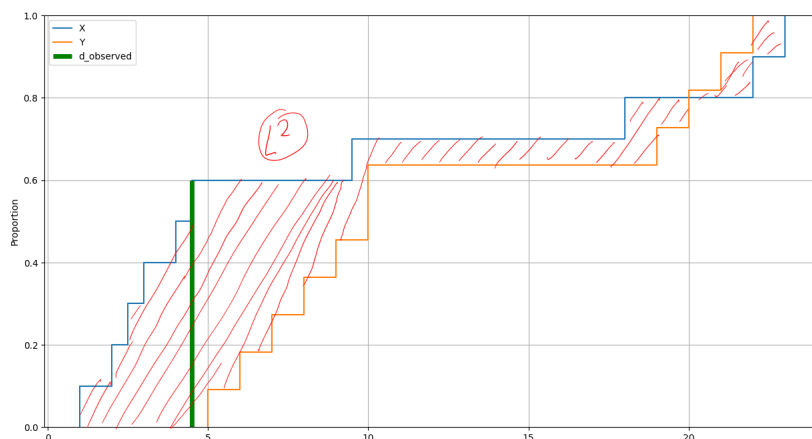
$$H_0 : \forall x \ F(x) = F_0(x), \quad H_A : F(x) \neq F_0(x),$$

альтернатива здесь максимально общая. Критерий Колмогорова смотрит на максимальное расстояние между гипотетической и эмпирической функцией распределения, его статистика

$$T = \sqrt{n} \sup_x |F_0(x) - \hat{F}(x)|$$

имеет распределение Колмогорова и критическая область расположена правее квантили  $k_{1-\alpha}$  – большая разность говорит скорее о разных распределениях.

### 9.3.2 Критерии Крамера-Мизеса и Андерсона-Дарлинга



Другой интересной идеей оказывается посмотреть не на супремум, а на  $L^2$  расстояние, то есть, на площадь разности  $\hat{F}(x) - F_0(x)$ . Общий подход к построению таких критериев –

построить статистику типа

$$\int_{\mathbb{R}} \psi(F_0(x)) \left( F_0(x) - \hat{F}(x) \right)^2 f_0(x) dx,$$

где  $\psi : \mathbb{R} \rightarrow \mathbb{R}$  – это весовая функция, которая помогает лучше учесть особенности распределения. Например, можно детальнее смотреть в середину, можно смотреть в целом на всё одинаково, а можно основное внимание уделять хвостам распределения.

Мы всё так же проверяем

$$H_0 : \forall x \ F(x) = F_0(x), \quad H_A : F(x) \neq F_0(x),$$

и первый критерий основан на идее статистики

$$T = \int_{\mathbb{R}} 1 \cdot \left( F_0(x) - \hat{F}(x) \right)^2 f_0(x) dx,$$

и называется критерием *Крамера-Мизеса*. Как видно, весовая функция у этого критерия равна единице, то есть, он в равной степени учитывает разницу по всей прямой. У этой статистики, которая выражается суммой

$$T = \frac{1}{12n} + \sum_{i=1}^n \left( \frac{2i-1}{2n} - F_0(X_i) \right)^2.$$

Критерий Андерсона-Дарлинга похожий и имеет статистику

$$T = \int_{\mathbb{R}} \frac{1}{F_0(x)(1-F_0(x))} \left( F_0(x) - \hat{F}(x) \right)^2 f_0(x) dx,$$

увеличивающую вес ошибки по мере приближения к краям, тем самым лучше анализируя разницу на хвостах распределения. Значение статистики вычисляют как

$$T = -n - \sum_{i=1}^n \frac{2i-1}{n} \ln \frac{F(X_i)}{1-F(X_{n+1-i})}.$$

Оба критерия имеют критическую область справа от квантили  $q_{1-\alpha}$ , сами квантили берутся из достаточно сложных распределений, подсчитанных исходя из того, что  $F_0^{-1}(X_i)$  имеет равномерное распределение. Там задействована даже теория случайных процессов. Подробнее можно узнать, например, в [1]. К счастью, оба критерия уже имеют устоявшуюся имплементацию и есть в готовом виде, например, в statsmodels.

## 9.4 Проверка гомоскедастичности

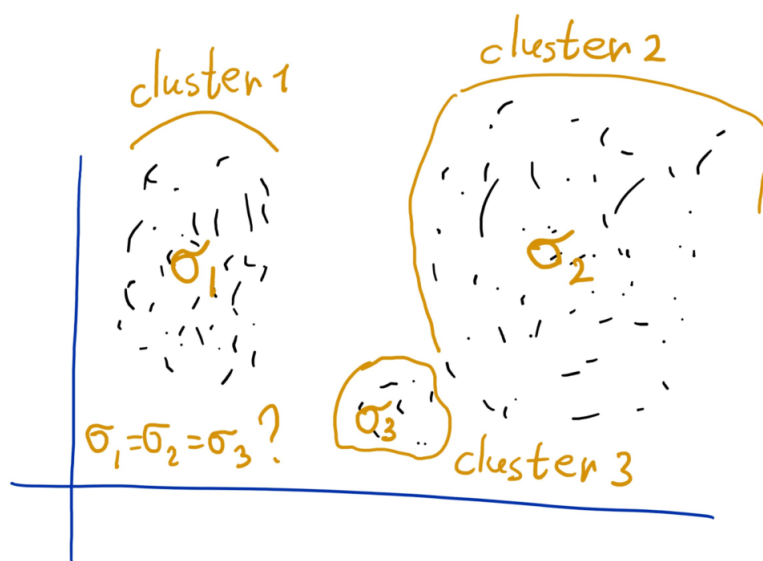
Прежде всего заметим, что просто проверить общую гипотезу о ковариационной матрице остатков невозможно без дополнительных предположений о её структуре – остатков  $n$ ,

а различных элементов в матрице  $n(n+1)/2$ . Всё же, если мы проверили некоррелированность, то задача становится проще, теперь там интересны ковариационные матрицы остатков вида

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

Однако и здесь мы не можем оценить  $\sigma_i^2$ , так как их  $n$  штук, столько же, сколько и данных. При дополнительных предположениях, многие из которых пришли из практических приложений эконометрики, возможно построить частично такую оценку и скорректировать оценку метода наименьших квадратов. Об этом позже.

Для проверки гипотез сложности похожи, но если мы приняли предположение о некоррелированности, то для сравнения  $\sigma_i$  между собой уже можно построить статистический критерий при дополнительных структурных предположениях. Такие критерии называются тестами на гетероскедастичность или на гомоскедастичность.



Проверяем гипотезу

$$H_0 : \sigma_1^2 = \sigma_2^2 = \dots = \sigma_r^2, \quad H_A : \exists i, j \sigma_i^2 \neq \sigma_j^2.$$

для построения теста нужно предположить дополнительную структуру. В *критерии Бартлетта* предполагается, что есть  $r$  различных дисперсий в различных областях пространства объясняющих переменных (например, разная дисперсия цены квартиры для разных категорий площади). Исходя из этого выборку делят на  $r$  кластеров (часто из чисто практических соображений), наследуя кластерную структуру из пространства объясняющих переменных, затем внутри каждого оценивают дисперсию остатков  $\hat{\sigma}_i^2$  далее оценивают



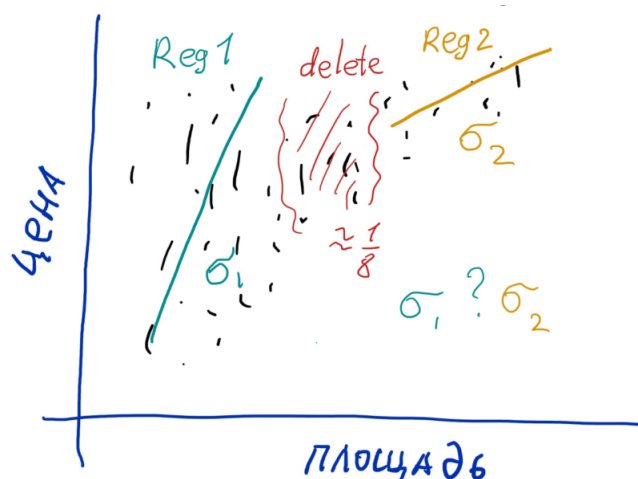
$Q$ -статистику Бартлетта  $Q/l$ , вычисляя

$$Q = n \ln \left( \sum_{i=1}^r \frac{n_i}{n} \hat{\sigma}_i^2 \right) - \sum_{i=1}^r n_i \ln \hat{\sigma}_i^2, \quad l = 1 + \frac{1}{3(r-1)} \left( \sum_{i=1}^r \frac{1}{n_i} - \frac{1}{n} \right).$$

Известно, что в предположении близости выборки к нормальному распределению и верной гипотезы  $Q/l \sim \chi^2(r-1)$  и критическая область находится справа от  $\chi_{1-\alpha}^2$ . Этот критерий очень чувствителен к отклонению выборки от нормальной, поэтому к этому тесту нужно относиться с осторожностью: нахождение статистики в критической области может также говорить о нарушении нормальности.

Другой критерий можно построить из предположения о зависимости дисперсии от одной из объясняющих переменных. Таков критерий *Голдфелда-Куандта*, основная идея которого состоит в том, что

1. Надо отсортировать данные по одной из объясняющих переменных (например, по площади квартиры);
2. Из середины отсортированного ряда нужно исключить  $d$  средних элементов (причём  $(n-d)/2 > k$ , числа параметров);
3. Надо построить две регрессии в каждой из половин и сравнить их по сумме квадратов остатков.



Смысл второго пункта в том, чтобы максимально разделить две группы наблюдений и получить по ним разные регрессии, которые можно сравнить. По этой причине нет смысла выкидывать мало наблюдений, так как разница будет невелика; но также нельзя выкидывать слишком много. Практическая рекомендация: мощность критерия максимальна, если оставлять в выборках  $n_1 = n_2 \approx 3n/8$ , то есть, выбрасывать примерно четверть наблюдений.

Когда мы оценили две регрессии и посчитали  $S_1, S_2$ , суммы квадратов остатков, для их сравнения мы можем применить асимптотический (или точный) тест Фишера. Мы проверяем гипотезу

$$H_0 : \frac{\sigma_2^2}{\sigma_1^2} = 1, \quad H_A : \frac{\sigma_2^2}{\sigma_1^2} \neq 1.$$

Статистика критерия

$$F = \frac{S_2/(n_2 - p)}{S_1/(n_1 - p)},$$

где  $p$  – число параметров, имеет (как минимум, асимптотическое) распределение Фишера  $F(n_2 - p, n_1 - p)$  и критическую область можно задать в зависимости от альтернативы.

# 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] Herbert E Rauch, F Tung, and Charlotte T Striebel. Maximum likelihood estimates of linear dynamic systems. *AIAA journal*, 3(8):1445–1450, 1965.
- [11] O. A. Stepanov. Kalman filtering: Past and present. an outlook from russia. (on the occasion of the 80th birthday of rudolf emil kalman). *Gyroscopy and Navigation*, 2(2):99–110, Apr 2011.
- [12] R. L. Stratonovich. Conditional markov processes. *Theory of Probability & Its Applications*, 5(2):156–178, 1960.
- [13] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

- [14] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных*. Изд. дом Высшей школы экономики, 2012.
- [15] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования*. М.: МЦНМО, 2013.
- [16] Р.Л. Стратонович. *Условные марковские процессы и их применение к теории оптимального управления*. Московский государственный университет, 1966.
- [17] А.Н. Ширяев. *Основы стохастической финансовой математики*. МЦНМО, 2016.