**Варианты лабораторной работы №1**

**по курсу «Введение в машинное обучение»**

**Файл с данными**

data\_vN-NN.csv, N-NN – номер варианта

**Шаблон отчета**

<https://docs.google.com/presentation/d/1SaG1JYkH_X5eFrq6VleHDd757jmKna8wxb3YyGyK7h0/edit?usp=sharing>

**Отчет**

Файл otchet\_vN-NN\_GroupFIO.pdf и **исходные коды**, написанные для выполнения заданий, высылать на [mlearninglabs@gmail.com](mailto:mlearninglabs@gmail.com)

**Указания по оформлению графиков**

<http://datalearning.ru/study/Courses/methodic/lections/graphs.pdf>

К каждому графику в отчете должны быть даны все необходимые пояснения для его понимания (название, подписи осей, легенда, условия эксперимента, параметры обработки и пр.).

**Рекомендуемые средства программной реализации**

Python, MATLAB

**Литература**

1. Alpaydin, E. (2014). *Introduction to machine learning*. MIT press.

2. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

3. Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.

4. Fox J. (2015). Applied regression analysis and generalized linear models. Sage Publications.

5. Sen, Ashish, and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 1997.

**Вариант 1-01**

**Bias-Variance decomposition**

**Задания**

1. Построить графики зависимости среднеквадратичной ошибки (MSE) модели на обучающей и тестовой выборках от параметра *m* (числа регрессоров в регрессионной модели).

2. Построить графики зависимости выхода модели от *x*, наложенные на диаграмму рассеяния, для случаев: а) слишком простой модели; б) переобучения; в) хорошей обобщающей способности.

3. Построить графики зависимостей среднеквадратичной ошибки модели на обучающей и тестовой выборках от объёма обучающей выборки для нескольких фиксированных *m*.

4. Для фиксированного *m* оценить дисперсию отклика модели и смещение выхода модели (по различным обучающим выборкам) при каждом фиксированном *x*, по полученным данным построить графики зависимости дисперсии и смещения от *x*, а также гистограммы распределения дисперсий и смещений для обучающей и тестовой выборок.

5. Построить графики зависимости дисперсии отклика модели и смещения выхода модели от параметра *m* при некоторых фиксированных *x*.

6. Сделать вывод о связи дисперсии и смещения выхода модели и параметра *m*.

**Указания**

Класс регрессионных моделей: .

Кросс-валидация: Holdout (70/30).

Выводить графики MSE и дисперсий в логарифмическом масштабе.

Смещения в пп.4,5 считать относительно усредненной (по различным обучающим выборкам) функции регрессии при *m* = 4.

**Литература**

1. Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, *1*(1), 55-77.

2. Breiman, L. (1996). Bias, variance, and arcing classifiers.

**Вариант 1-02**

**Cross-validation**

**Задания**

1. Построить зависимость средней (по фолдам) среднеквадратичной ошибки (MSE) 10‑fold кросс-валидации на обучающей и тестовой выборках от параметра *m* (числа регрессоров в регрессионной модели).

2. Построить зависимость средней (по фолдам) среднеквадратичной ошибки (MSE) LOO‑кросс-валидации на обучающей и тестовой выборках от параметра *m*.

3. Построить гистограммы распределений среднеквадратичной ошибки (MSE) модели на обучающей и тестовой выборках, получаемых в результате кросс-валидации по методу Монте-Карло, привести статистические характеристики ошибок (среднее, дисперсия, с.к.о., доверительные интервалы), для случаев: а) слишком простой модели; б) переобучения; в) хорошей обобщающей способности.

4. Построить графики зависимости средней и с.к.о. (по фолдам) среднеквадратичной ошибки (MSE) k-fold кросс-валидации на обучающей и тестовой выборках от числа фолдов *k* для нескольких фиксированных *m*.

5. Построить графики зависимости средней и с.к.о. (по фолдам) среднеквадратичной ошибки (MSE) LOO-кросс-валидации на обучающей и тестовой выборках от объёма обучающей выборки для нескольких фиксированных *m*.

6. Сделать выводы по результатам сравнения различных методов кросс-валидации.

**Указания**

Класс регрессионных моделей: .

Выводить графики MSE в логарифмическом масштабе.

**Литература**

1. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40-79.

2. Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, *6*(1), 10.

**Вариант 1-03**

**Ordinary least squares**

**Задания**

1. Сгенерировать данные в соответствии с теоретической регрессионной моделью. Построить графики зависимостей значений оцененных параметров модели от объема обучающей выборки, нанести на графики границы доверительных интервалов. По графикам сделать вывод о несмещенности и состоятельности рассчитанных оценок.

2. Для нескольких фиксированных *x* построить графики зависимости оцененных значений функции регрессии от объема обучающей выборки, нанести на графики границы доверительных интервалов. По графикам сделать вывод о несмещенности и состоятельности оценок значений функции регрессии.

3. Построить график зависимости остаточной дисперсии от объема обучающей выборки. Сделать вывод о несмещенности и состоятельности остаточной дисперсии как оценки дисперсии ошибок модели.

4. Проверить, совпадают ли распределения остатков модели и ошибок модели, используя критерий «хи-квадрат» при различных объемах обучающей выборки: *а*) при *n* = 20; *б*) при *n* = 200 (привести значения статистики критерия, *p-value*, статистическое решение).

5. Для фиксированного объема обучающей выборки *n* = 50 построить гистограмму распределения оценок параметров модели (по различным обучающим выборкам) и остаточной дисперсии, сравнить распределения с теоретическими. Рассчитать среднее и с.к.о. оценок параметров модели и остаточной дисперсии, сравнить значения с теоретическими.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Теоретическая регрессионная модель: , где .

Для получения статистики в п.5 использовать не менее 1000 обучающих выборок.

**Литература**

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

2. Fox J. Applied regression analysis and generalized linear models. – Sage Publications, 2015.

**Вариант 1-04**

**Simple regression model**

**Задания**

1. Построить простейшую линейную регрессионную модель, используя метод OLS, рассчитать доверительные интервалы для параметров модели и функции регрессии, а также интервалы предсказания с доверительными вероятностями 90% и 99%. На диаграмме рассеяния вывести рассчитанные функцию регрессии и границы интервалов.

2. Проверить распределения остатков модели на обучающей и тестовой выборках на нормальность, используя критерий Харке-Бера (привести значения статистики критерия, *p-value*, статистическое решение).

3. Построить графики зависимостей значений оцененных параметров от объема обучающей выборки, нанести на графики границы доверительных интервалов.

4. Для нескольких фиксированных *x* рассчитать выборки оцененных значений функции регрессии и остатков (по различным обучающим выборкам), построить гистограммы их распределения. Проверить, подчиняются ли распределения стандартизованных остатков для этих *x* распределению Стьюдента *T*(*n*–2), используя критерий «хи-квадрат» (привести значения статистики критерия, *p-value*, статистическое решение).

5. Построить зависимость MSE регрессионной модели от ее параметров (визуализировать в виде поверхности и контурной диаграммы). Определить визуально оптимальные значения параметров и сравнить их с полученными в результате обучения.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Кросс-валидация: Holdout (70/30).

Выводить графики MSE в логарифмическом масштабе.

**Литература**

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

2. Fox J. Applied regression analysis and generalized linear models. – Sage Publications, 2015.

**Вариант 1-05**

**Multiple regression model**

**Задания**

1. Построить множественную линейную регрессионную модель, используя метод OLS, рассчитать доверительные интервалы для параметров модели и функции регрессии, а также интервалы предсказания с доверительными вероятностями 90% и 99%. На трехмерной диаграмме рассеяния вывести рассчитанные плоскость регрессии и границы интервалов.

2. Построить гистограмму распределения остатков модели на обучающей и тестовой выборках, проверить распределение остатков на нормальность, используя критерий «хи-квадрат» (привести значения статистики критерия, *p-value*, статистическое решение).

3. Рассчитать и визуализировать проекционную матрицу модели. Для каждого наблюдения рассчитать леверидж. Отметить на трехмерной диаграмме рассеяния исходных данных точки, имеющие наибольший и наименьший левериджи. Построить: *а*) график зависимости регрессионных остатков от левериджа; *б*) график зависимости левериджей наблюдений от их расстояний до центра рассеяния . Оценить визуально наличие статистической связи между рассматриваемыми величинами.

4. Рассчитать стандартизованные остатки модели на обучающей и тестовой выборках, построить гистограммы их распределения. Проверить, соответствует ли эти выборочные распределения распределению Стьюдента *T*(*n*–2), используя критерий «хи-квадрат» (привести значения статистики критерия, *p-value*, статистическое решение). Можно ли утверждать, что математическое ожидание стандартизованных остатков равно нулю? (привести статистическую гипотезу, значения статистики критерия, *p-value*, статистическое решение).

5. При фиксированном значении параметра β0 (равном оцененному значению) построить зависимость коэффициента детерминации регрессионной модели от параметров β1, β2 (визуализировать в виде поверхности и контурной диаграммы). Определить визуально оптимальные значения параметров и сравнить их с полученными в результате обучения.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Кросс-валидация: Holdout (70/30).

**Литература**

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

2. Fox J. Applied regression analysis and generalized linear models. – Sage Publications, 2015.

**Вариант 1-06**

**Weighted least squares**

**Задания**

1. Построить простейшую линейную регрессионную модель, используя метод OLS, рассчитать коэффициенты детерминации модели на обучающей и тестовой выборках и доверительные интервалы для функции регрессии с доверительными вероятностями 90% и 99%. На диаграмме рассеяния вывести рассчитанную функцию регрессии и границы интервалов.

2. Построить графики зависимости остатков модели от моделируемых значений и от входной переменной, провести качественный анализ на гетероскедастичность остатков.

3. Построить простейшую линейную регрессионную модель модулей остатков на входную переменную. Проверить значимость модели (привести статистическую гипотезу, значение статистики критерия, *p‑value*, статистическое решение). Сделать вывод о гетероскедастичности исходных данных.

4. Построить простейшую линейную регрессионную модель, используя метод WLS, в качестве весов наблюдений использовать: *а*) величины, обратные модельным значениям функции регрессии, построенной в п. 3; *б*) величины, равные 1/*x*. Рассчитать коэффициенты детерминации моделей на обучающей и тестовой выборках и границы доверительных интервалов для функции регрессии, сравнить построенные модели с моделью, полученной в п. 1.

5. Построить гистограммы распределения остатков моделей, полученных в пп. 1 и 4, на обучающей и тестовой выборках, проверить распределения остатков на нормальность, используя критерий «хи-квадрат» (привести значения статистики критерия, *p-value*, статистическое решение).

6. Сделать выводы по результатам сравнения методов OLS и WLS.

**Указания**

Кросс-валидация: Holdout (70/30).

Для качественного анализа на гетероскедастичность остатков провести группировку значений входной переменной и рассчитать дисперсии остатков модели в каждой группе. Результат представить в виде графика зависимости групповой дисперсии от номера группы.

Если в результате построения регрессионной модели в п.4. получаются близкие к нулю или отрицательные модельные значения функции регрессии, то в качестве весов использовать значения , где .

После обучения регрессионной модели методом WLS остатки модели скорректировать с учетом весов .

**Литература**

1. Sen, Ashish, and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 1997.

**Вариант 1-07**

**Least absolute residuals regression**

**Задания**

1. Построить простейшую линейную регрессионную модель, используя метод OLS, рассчитать коэффициенты детерминации модели на обучающей и тестовой выборках. На диаграмме рассеяния вывести рассчитанную функцию регрессии.

2. Обучить простейшую линейную регрессионную модель с критерием наименьших абсолютных ошибок (LAR), рассчитать коэффициенты детерминации модели на обучающей и тестовой выборках, сравнить обученную модель с моделью, полученной в п. 1.

3. Построить зависимость значения LAR-критерия от параметров модели (визуализировать в виде поверхности и контурной диаграммы). Определить визуально оптимальные значения параметров модели и сравнить их со значениями, найденными в результате обучения.

4. Добавить в обучающую выборку несколько выбросных значений и обучить OLS- и LAR-модели. Сравнить полученные параметры модели, функции регрессии и значения коэффициентов детерминации на обучающей и тестовой выборках, сделать вывод о чувствительности критерия обучения к наличию выбросов в обучающих данных.

5. Построить зависимости, аналогичные п. 3, для OLS- и LAR-моделей после добавления выбросных значений в обучающую выборку, сравнить построенные поверхности и оптимальные значения параметров модели.

6. Сделать вывод по результатам сравнения методов OLS и LAR.

**Указания**

Кросс-валидация: Holdout (70/30).

Для обучения регрессионной LAR-модели использовать один из методов оптимизации, в качестве начального приближения использовать коэффициенты, полученные с помощью метода OLS, либо случайные значения, близкие к нулю.

**Литература**

1. Bloomfield, Peter, and William L. Steiger. *Least absolute deviations: theory, applications, and algorithms*. Boston: Birkhäuser, 1983.

2. Farebrother, Richard. *L1-Norm and L∞-Norm Estimation: An Introduction to the Least Absolute Residuals, the Minimax Absolute Residual and Related Fitting Procedures*. Springer Science & Business Media, 2013.

**Вариант 1-08**

**Iteratively reweighted least-squares**

**Задания**

1. Построить простейшую линейную регрессионную модель, используя метод OLS, рассчитать коэффициенты детерминации модели на обучающей и тестовой выборках. На диаграмме рассеяния вывести рассчитанную функцию регрессии.

2. Рассчитать стандартизованные остатки и робастные стандартизованные остатки обученной регрессионной модели на обучающей и тестовой выборках. Визуализировать (в виде диаграмм рассеяния и гистограмм) и сравнить полученные выборки. Проверить визуально распределения остатков на нормальность по гистограмме и Q-Q диаграмме.

3. Рассчитать биквадратные веса наблюдений. Отметить на диаграмме рассеяния (п. 1) наблюдения, имеющие наибольшие и наименьшие веса. Построить простейшую линейную регрессионную модель, используя метод WLS, с рассчитанными весами.

4. Повторить пп. 2-3 до тех пор, пока обученная модель не перестанет изменяться. Сравнить параметры модели, функции регрессии и значения коэффициентов детерминации на обучающей и тестовой выборках для обученной модели и модели, построенной в п. 1.

5. Рассчитать стандартизованные остатки и робастные стандартизованные остатки обученной регрессионной IRLS-модели, построить график их зависимости от входной переменной. По графику идентифицировать выбросные значения в данных.

6. Сделать вывод по результатам сравнения методов OLS и IRLS.

**Указания**

Кросс-валидация: Holdout (70/30).

Идентификацию выбросов в п.5 проводить по следующему критерию: наблюдение является выбросным, если соответствующий робастный стандартизованный остаток IRLS-модели превышает робастное с.к.о. остатков.

**Литература**

1. Holland, Paul W., and Roy E. Welsch. "Robust regression using iteratively reweighted least-squares." *Communications in Statistics-theory and Methods* 6.9 (1977): 813-827.

2. Fox, John, and Sanford Weisberg. "Robust regression." *An R and S-Plus companion to applied regression* 91 (2002).

**Вариант 1-09**

**Nonlinear least-squares**

**Задания**

1. Построить полиномиальные регрессионные модели при *m*=1,2,3, используя метод OLS, рассчитать коэффициенты детерминации моделей на обучающей и тестовой выборках. На диаграмме рассеяния исходных данных вывести рассчитанные функции регрессии.

2. Обучить нелинейную экспоненциальную модель, рассчитать коэффициент детерминации модели на обучающей и тестовой выборках, сравнить результаты с полиномиальными моделями.

3. Построить зависимость значения MSE от параметров нелинейной модели (визуализировать в виде поверхности и контурной диаграммы). Определить визуально оптимальные значения параметров и сравнить их со значениями, найденными в результате обучения.

4. Применить логарифмическое преобразование результирующей переменной и построить простейшую линейную регрессионную модель. Рассчитать коэффициент детерминации модели на обучающей и тестовой выборках, визуализировать полученную функцию регрессии, сравнить с результатам пп. 1, 2.

5. Построить гистограммы распределения остатков моделей, обученных в пп. 1, 2, 4 на обучающей и тестовой выборках, проверить распределения остатков на нормальность, используя критерий «хи-квадрат» (привести значения статистики критерия, *p-value*, статистическое решение). Оценить нормальность остатков визуально по Q-Q диаграммам.

6. Сделать вывод по результатам сравнения обученных линейных и нелинейной регрессионных моделей.

**Указания**

Класс полиномиальных регрессионных моделей: .

Класс экспоненциальных регрессионных моделей: .

Кросс-валидация: Holdout (70/30).

Для обучения нелинейной регрессионной использовать один из методов оптимизации (простой градиентный метод, метод Левенберга-Маркардта или др.).

Выводить графики MSE в логарифмическом масштабе.

**Литература**

1. Sen, Ashish, and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 1997.

2. Seber, George AF, and Christopher John Wild. "Nonlinear regression." *New Jersey: John Wiley & Sons* 62 (2003): 63.

**Вариант 1-10**

**Generalized linear models**

**Задания**

1. Построить полиномиальные регрессионные модели для степеней полиномов *m* = 1,…,10, используя метод OLS, и рассчитать коэффициенты детерминации на обучающей и тестовой выборках. Выбрать наилучшую модель и рассчитать доверительные интервалы для параметров модели и функции регрессии, а также интервалы предсказания с доверительными вероятностями 90% и 99%. На диаграмме рассеяния вывести рассчитанные функцию регрессии и границы интервалов.

2. Построить обобщенные полиномиальные регрессионные модели (GLM) для степеней полиномов *m* = 1,2,3: *а*) с логит-функцией связи; *б*) с пробит-функцией связи. Рассчитать доверительные интервалы для параметров моделей и функций регрессии, а также интервалы предсказания с доверительными вероятностями 90% и 99%. Сравнить результаты с п. 1.

3. Построить гистограммы распределения остатков моделей, полученных в пп. 1 и 2 на обучающей и тестовой выборках, проверить распределения остатков на нормальность, используя критерий «хи-квадрат» (привести значения статистики критерия, *p-value*, статистическое решение), а также визуально по Q-Q диаграмме.

4. Для модели, полученной в п. 2, проверить значимость каждого параметра (привести значения статистики критерия, *p-value*, статистическое решение).

5. Предложить нелинейную (по *x*) регрессионную модель, которая могла бы описывать имеющиеся данные. Обучить модель и сравнить ошибки на обучающей и тестовой выборках с моделью, полученной в п.2.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Класс полиномиальных регрессионных моделей: .

Кросс-валидация: Holdout (70/30).

При построении GLM-модели считать, что моделируемая переменная имеет биномиальное распределение *B*(*n*, *p*) с неизменным параметром *n* (с точностью до множителя).

**Литература**

1. Agresti A. Foundations of linear and generalized linear models. – John Wiley & Sons, 2015.

2. Dunn, Peter K., and Gordon K. Smyth. *Generalized linear models with examples in R*. New York: Springer, 2018.

**Вариант 1-11**

**Simple linear regression diagnostics**

**Задания**

1. Построить графики зависимости среднеквадратичной ошибки (MSE) модели на обучающей и тестовой выборках от параметра *m* (числа регрессоров в регрессионной модели). Определить оптимальное значение параметра *m*. Построить графики зависимости выхода модели от *x*, наложенные на диаграмму рассеяния исходных данных, для найденного значения параметра *m*.

2. Построить гистограмму распределения остатков модели, проверить распределение остатков на нормальность, используя: *а*) критерий «хи-квадрат»; *б*) критерий Харке-Бера (привести значения статистики критерия, *p-value*, статистическое решение).

3. Построить lag plot для остатков модели. Проверить остатки на наличие автокорреляции, используя критерий Дарбина-Уотсона (привести значения статистики критерия, *p‑value*, статистическое решение).

4. Построить графики зависимости остатков модели от моделируемых значений и от входной переменной. Построить простейшую линейную регрессионную модель модулей остатков на входную переменную, указать границы доверительных интервалов функции регрессии. Проверить значимость полученной модели (привести значения статистики критерия, *p‑value*, статистическое решение), сделать вывод о гетероскедастичности остатков.

5. Проверить остатки на гетероскедастичность, используя тест Бройша–Пагана (привести значения статистики критерия, *p-value*, статистическое решение).

6. Сделать выводы по результатам проведенных статистических анализов.

**Указания**

Класс регрессионных моделей: .

Кросс-валидация: 5-fold.

Выводить графики MSE в логарифмическом масштабе.

**Литература**

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

2. Sen, Ashish, and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 1997.

**Вариант 1-12**

**Multiple linear regression diagnostics**

**Задания**

1. Построить множественную линейную регрессионную модель, используя метод OLS, рассчитать доверительные интервалы для параметров модели и функции регрессии, а также интервалы предсказания с доверительными вероятностями 90% и 99%. Визуализировать рассчитанную плоскость регрессии на диаграммах рассеяния в проекциях (по горизонтальным осям – отдельные регрессоры, по вертикальной оси – моделируемая величина).

2. Оценить адекватность линейной регрессионной модели по графикам зависимостей остатков модели от моделируемых значений и от входных переменных. Рассчитать все парные коэффициенты корреляции и множественный коэффициент корреляции. Сделать выводы на основе полученных значений.

3. Оценить влияние каждой из входных переменных на результирующую переменную: *а*) построить графики частичной регрессии для каждого из регрессоров, по графику определить визуально характер зависимости; *б*) построить графики зависимости результирующей переменной от каждого из регрессоров по-отдельности, определить визуально характер зависимости. Сравнить результаты пп. а) и б), предложить интерпретацию полученных результатов, сделать выводы о статистической связи регрессоров и их необходимости включения в регрессионную модель.

4. Проверить значимость каждого из коэффициентов модели (привести значения статистик критерия, *p‑value*, статистические решения). Построить Q-Q диаграмму для распределения остатков модели в сравнении с нормальным распределением, сделать качественный вывод о нормальности распределения остатков.

5. Предложить модификацию линейной регрессионной модели по результатам диагностики в пп.2, 3, оценить параметры модели, используя метод OLS, и сравнить их с полученными в п.1. Оценить адекватность построенной модели по графикам зависимостей остатков модели от моделируемых значений и от входных переменных. Сравнить Q-Q диаграмму для распределения остатков модели с построенной в п.4.

6. Сделать выводы по результатам диагностики исходной и модифицированной моделей.

**Указания**

Кросс-валидация: Holdout (70/30).

**Литература**

1. Montgomery, D. C., Peck, E. A., & Vining, G. G. (2012). *Introduction to linear regression analysis* (Vol. 821). John Wiley & Sons.

2. Fox J. Applied regression analysis and generalized linear models. – Sage Publications, 2015.

**Вариант 1-13**

**Regression outliers and influential points**

**Задания**

1. Построить простейшую линейную регрессионную модель, используя метод OLS, рассчитать коэффициент детерминации модели на обучающей и тестовой выборках. На диаграмме рассеяния вывести рассчитанную функцию регрессии.

2. Рассчитать стандартизованные остатки обученной регрессионной модели. Визуализировать зависимости стандартизованных остатков модели от моделируемых значений и от входной переменной (в виде диаграмм рассеяния). Построить диаграмму box-and-whisker для стандартизованных остатков. Отметить на диаграмме выбросные значения.

3. Рассчитать левериджи и расстояния Кука для каждого наблюдения. Визуализировать рассчитанные значения в виде диаграмм рассеяния (в зависимости от моделируемых значений, значений входной переменной и стандартизованных остатков). На диаграмме рассеяния, построенной в п.1, отметить точки, имеющие наибольшие: *а*) стандартизованные остатки; *б*) левериджи; *в*) расстояния Кука.

4. Отсеять наблюдения с расстояниями Кука, превышающими порог, и построить линейную регрессионную модель, используя метод OLS, рассчитать коэффициент детерминации модели на обучающей и тестовой выборках. Сравнить рассчитанную функцию регрессии с полученной в п.1.

5. Рассчитать стандартизованные остатки обученной модели и построить диаграмму box-and-whisker. Сравнить диаграмму и распределения остатков с полученными в п.2.

6. Сделать вывод по результатам исследования методов обнаружения регрессионных выбросов и сравнения обученных регрессионных моделей.

**Указания**

Кросс-валидация: Holdout (70/30).

В качестве порога для расстояния Кука в п.4 использовать 3\*среднее (по всем наблюдениям) расстояние Кука.

**Литература**

1. Sen, Ashish, and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 1997.

2. Fox J. Applied regression analysis and generalized linear models. – Sage Publications, 2015.

**Вариант 1-14**

**Multicollinearity diagnostics**

**Задания**

1. Построить множественную линейную регрессионную модель, используя метод OLS, рассчитать оценки дисперсии параметров модели и доверительные интервалы для параметров модели с доверительными вероятностями 90% и 99%. Проверить значимость каждого из коэффициентов на уровнях значимости 0.1 и 0.01 (привести значения статистики критерия, *p‑value*, статистическое решение).

2. Визуализировать попарные зависимости между моделируемой переменной и регрессорами. Рассчитать все парные коэффициенты корреляции и множественный коэффициент корреляции. Сделать выводы на основе полученных значений.

3. Рассчитать число обусловленности матрицы XTX и показатели VIF для каждого регрессора. Сделать выводы на основе полученных значений. Исключить из регрессионной модели регрессор, линейно зависимый от остальных регрессоров (по критерию VIF). После исключения регрессора проверить мультиколлинеарность оставшихся регрессоров. Итеративно повторять исключение регрессоров да тех пор, пока не будет устранена мультиколлинеарность.

4. Построить линейную регрессионную модель с регрессорами, полученными в п.3, рассчитать оценки дисперсии параметров модели и доверительные интервалы для параметров модели с доверительными вероятностями 90% и 99%. Проверить значимость каждого из коэффициентов на уровнях значимости 0.1 и 0.01 (привести значения статистики критерия, *p‑value*, статистическое решение). Сравнить результаты с п.1.

5. Изменить порядок исключения линейно зависимых регрессоров из исходной модели в п.3. Сравнить полученный состав регрессоров с регрессорами в п.3 и соответствующие регрессионные модели. Рассчитать коэффициенты детерминации на обучающей и тестовой выборках, сделать вывод о качестве регрессионных моделей с различными составами неколлинеарных регрессоров.

6. Сделать выводы по результатам проведенных исследований.

**Указания**

Кросс-валидация: Holdout (70/30).

Считать, что регрессор линейно зависим с остальными, если значение VIF для него более 5.

**Литература**

1. Sen, Ashish, and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 1997.

2. Fox J. Applied regression analysis and generalized linear models. – Sage Publications, 2015.

**Вариант 1-15**

**Ridge regression**

**Задания**

1. Построить графики зависимости среднеквадратичной ошибки (MSE) регрессионной модели на обучающей и тестовой выборках от параметра *m* (числа регрессоров в регрессионной модели) для случаев: а) линейной регрессии; б) гребневой регрессии при λ ≈ 0; в) гребневой регрессии при λ >> 0.

2. Построить графики зависимости выхода регрессионной модели от *x*, наложенные на диаграмму рассеяния, при фиксированном *m* для случаев: а) линейной регрессии; б) гребневой регрессии при λ ≈ 0; в) гребневой регрессии при λ >> 0.

3. Построить графики зависимости значений коэффициентов регрессионной модели от параметра гребневой регрессии при фиксированном *m* (ridge trace plot).

4. Для фиксированного *m* оценить дисперсию отклика модели (по различным обучающим выборкам) при каждом фиксированном *x*, построить графики зависимости дисперсий от *x* (для обучающей и тестовой выборок) и гистограммы распределения дисперсий для: а) линейной регрессии; б) гребневой регрессии при λ ≈ 0; в) гребневой регрессии при λ >> 0.

5. Построить графики зависимости остатков модели от моделируемых значений и входной переменной, рассчитать среднее значение и дисперсию остатков, провести качественный анализ остатков на наличие неслучайных паттернов и на нормальность (по гистограмме и Q-Q диаграмме) для: а) линейной регрессии; б) гребневой регрессии при λ ≈ 0; в) гребневой регрессии при λ >> 0.

6. Сделать выводы по результатам сравнения линейной и гребневой регрессий.

**Указания**

Класс регрессионных моделей: .

Кросс-валидация: Holdout (70/30).

Перед построением моделей все входные и выходные переменные стандартизовать: , .

MSE регрессионной модели считать без учета регуляризатора.

**Литература**

1. Gruber, M. (1998). *Improving Efficiency by Shrinkage: The James--Stein and Ridge Regression Estimators* (Vol. 156). CRC Press.

2. Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55-67.

**Вариант 1-16**

**LASSO regression**

**Задания**

1. Построить графики зависимости среднеквадратичной ошибки (MSE) регрессионной модели на обучающей и тестовой выборках от параметра *m* (числа регрессоров в регрессионной модели) для случаев: а) линейной регрессии; б) LASSO-регрессии при λ ≈ 0; в) LASSO-регрессии при λ >> 0.

2. Построить графики зависимости выхода регрессионной модели от *x*, наложенные на диаграмму рассеяния, при фиксированном *m* для случаев: а) линейной регрессии; б) LASSO-регрессии при λ ≈ 0; в) LASSO-регрессии при λ >> 0.

3. Построить графики зависимости значений коэффициентов регрессионной модели от параметра LASSO-регрессии (LASSO trace plot).

4. Для фиксированного *m* оценить дисперсию отклика модели (по различным обучающим выборкам) при каждом фиксированном *x*, построить графики зависимости дисперсий от *x* (для обучающей и тестовой выборок) и гистограммы распределения дисперсий для: а) линейной регрессии; б) LASSO-регрессии при λ ≈ 0; в) LASSO-регрессии при λ >> 0.

5. Построить графики зависимости остатков модели от моделируемых значений и входной переменной, рассчитать среднее значение и дисперсию остатков, провести качественный анализ остатков на наличие неслучайных паттернов и на нормальность (по гистограмме и Q-Q диаграмме) для: а) линейной регрессии; б) LASSO-регрессии при λ ≈ 0; в) LASSO-регрессии при λ >> 0.

6. Сделать выводы по результатам сравнения линейной и LASSO-регрессий.

**Указания**

Класс регрессионных моделей: .

Кросс-валидация: Holdout (70/30).

Перед построением моделей все входные и выходные переменные стандартизовать: , .

MSE регрессионной модели считать без учета регуляризатора.

**Литература**

1. Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 267-288.

2. Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: a retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273-282.

**Вариант 1-17**

**Elastic net regression**

**Задания**

1. Построить графики зависимости среднеквадратичной ошибки (MSE) регрессионной модели на обучающей и тестовой выборках от параметра *m* (числа регрессоров в регрессионной модели) при а) λ = 0; б) λ > 0, α = 0; в) λ > 0, α = 0.5; г) λ > 0, α =1.

2. Для некоторого фиксированного *m* построить графики зависимости значений коэффициентов elastic net-регрессии от параметра λ при: а) α = 0; б) α = 0.5; в) α = 1.

3. Построить графики зависимости среднеквадратичной ошибки (MSE) регрессионной модели на обучающей и тестовой выборках от параметра α при: а) λ = λ1 > 0; б) λ = λ2> λ1 (для некоторого фиксированного *m*).

4. Построить графики зависимости выхода регрессионной модели от *x*, наложенные на диаграмму рассеяния исходных данных, при фиксированном *m* для случаев: а) λ = 0; б) λ > 0, α = 0; в) λ > 0, α = 0.5; г) λ > 0, α =1.

5. Построить графики зависимости остатков модели от моделируемых значений и входной переменной, рассчитать среднее значение и дисперсию остатков, провести качественный анализ остатков на наличие неслучайных паттернов и на нормальность (по гистограмме и Q-Q диаграмме) для: а) λ = 0; б) λ > 0, α = 0; в) λ > 0, α = 0.5; г) λ > 0, α = 1.

6. Сделать выводы по результатам сравнения elastic net-регрессий с различными значениями параметров.

**Указания**

Класс регрессионных моделей: .

Эмпирический риск: .

Кросс-валидация: Holdout (70/30).

MSE регрессионной модели считать без учета регуляризатора.

**Литература**

1. Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *67*(2), 301-320.

**Вариант 1-18**

**Stepwise regression**

**Задания**

1. Построить множественную линейную регрессионную модель, используя метод OLS, рассчитать доверительные интервалы для параметров модели с доверительными вероятностями 90% и 99%. Проверить значимость каждого из параметров модели (привести значения статистик критерия, *p‑value*, статистические решения). Сделать вывод о влиянии регрессоров на моделируемую переменную.

2. Расширить исходный состав признаков квадратичным признаками и проанализировать влияние каждого из них (как линейных, так и квадратичных) на модель по-отдельности. Для этого рассчитать парные коэффициенты корреляции между признаками и выходной переменной, визуализировать полученные значения, определить состав признаков, оказывающих наибольшее влияние.

3. Обучить пошаговую линейную регрессию с: *а*) прямым отбором признаков; *б*) обратным отбором признаков. Построить графики зависимости ошибок модели и коэффициент детерминации (на обучающей и тестовой выборках) на каждой итерации от номера итерации. Сравнить составы признаков, полученные в результате прямого и обратного отбора.

4. Найти оптимальный состав признаков методом полного перебора. Сравнить оптимальный состав признаков с признаками, полученными в п.3.

5. Обучить регрессионную модель со всеми признаками (линейными и квадратичными) с LASSO-регуляризатором Построить графики зависимости значений коэффициентов регрессионной модели от параметра LASSO-регрессии (LASSO trace plot). По графику определить наиболее значимые признаки и сравнить их с полученными в пп.3,4.

6. Сделать выводы по результатам отбора признаков для регрессионной модели разными методами.

**Указания**

Квадратичные признаки: 

Кросс-валидация: Holdout (70/30).

Перед построением LASSO-модели все входные и выходные переменные стандартизовать: , .

**Литература**

1. Sen, Ashish, and Muni Srivastava. *Regression analysis: theory, methods, and applications*. Springer Science & Business Media, 1997.

2. Fox J. Applied regression analysis and generalized linear models. – Sage Publications, 2015.

**Вариант 1-19**

**Bias-Variance decomposition**

**Задания**

1. Построить графики зависимости среднеквадратичной ошибки (MSE) модели на обучающей и тестовой выборках от параметра *m* (числа регрессоров в регрессионной модели) при: а) σ = 1; б) σ = 0.5; в) σ = 0.1. Определить значение параметра *m*, при котором достигается наименьшая ошибка модели на тестовой выборке.

2. Построить графики зависимости выхода модели от *x*, наложенные на диаграмму рассеяния, при σ = 0.5 для случаев: а) слишком простой модели; б) переобучения; в) хорошей обобщающей способности.

3. Построить графики зависимости среднеквадратичной ошибки (MSE) модели на обучающей и тестовой выборках от параметра σ для случаев: а) *m* = 1 % от объёма обучающей выборки; б) *m* = 10 % от объёма обучающей выборки; в) *m* = объёму обучающей выборки.

4. Для фиксированного *m* оценить дисперсию отклика модели и смещение выхода модели (по различным обучающим выборкам) при каждом фиксированном *x*, по полученным данным построить графики зависимости дисперсии и смещения от *x*, а также гистограммы распределения дисперсий и смещений для обучающей и тестовой выборок (принять σ = 0.5).

5. Построить графики зависимости дисперсии отклика модели и смещения выхода модели от параметра *m* при некотором фиксированном *x* (принять σ = 0.5).

6. Сделать вывод о связи дисперсии и смещения выхода модели и параметра *m*, а также о влиянии параметра σ на переобучение.

**Указания**

Класс регрессионных моделей: .

Величины *ci*, *i* = 1,…,*m*, задать через равные промежутки в интервале между минимальным и максимальным значениями *x* в обучающей выборке.

Кросс-валидация: Holdout (70/30).

Выводить графики MSE и дисперсий в логарифмическом масштабе.

Смещения в пп.4,5 считать относительно оцененной функции регрессии при оптимальном значении параметра *m*, найденном в п.1.

**Литература**

1. Friedman, J. H. (1997). On bias, variance, 0/1—loss, and the curse-of-dimensionality. *Data mining and knowledge discovery*, *1*(1), 55-77.

2. Breiman, L. (1996). Bias, variance, and arcing classifiers.

**Вариант 1-20**

**Cross-validation**

**Задания**

1. Построить зависимость средней (по фолдам) среднеквадратичной ошибки (MSE) 10‑fold кросс-валидации на обучающей и тестовой выборках от параметра *m* (числа регрессоров в регрессионной модели) при: а) σ = 1; б) σ = 0.5; в) σ = 0.1.

2. Построить зависимость средней (по фолдам) среднеквадратичной ошибки (MSE) LOO‑кросс-валидации на обучающей и тестовой выборках от параметра *m* при: а) σ = 1; б) σ = 0.5; в) σ = 0.1.

3. Построить гистограммы распределений среднеквадратичной ошибки (MSE) модели на обучающей и тестовой выборках, получаемых в результате кросс-валидации по методу Монте-Карло, привести статистические характеристики ошибок (среднее, дисперсия, с.к.о., доверительные интервалы), при σ = 0.5 для случаев: а) слишком простой модели; б) переобучения; в) хорошей обобщающей способности.

4. Визуализировать матрицы значений средней и с.к.о. (по фолдам) среднеквадратичной ошибки (MSE) k-fold кросс-валидации на обучающей и тестовой выборках. По столбцам матриц – число фолдов *k*, по строкам – значения параметра σ (принять *m* = 10 % от объёма обучающей выборки).

5. Визуализировать матрицы значений средней и с.к.о. (по фолдам) среднеквадратичной ошибки (MSE) LOO-кросс-валидации на обучающей и тестовой выборках. По столбцам матриц – значения параметра *m*, по строкам – значения параметра σ.

6. Сделать выводы по результатам сравнения различных методов кросс-валидации и о влиянии параметров *m* и σ на переобучение.

**Указания**

Класс регрессионных моделей: .

Величины *ci*, *i* = 1,…,*m*, задать через равные промежутки в интервале между минимальным и максимальным значениями *x* в обучающей выборке.

Выводить графики MSE в логарифмическом масштабе.

**Литература**

1. Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics surveys*, *4*, 40-79.

2. Krstajic, D., Buturovic, L. J., Leahy, D. E., & Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of cheminformatics*, *6*(1), 10.