1 Домашнее задание

На лекции мы начали заниматься линейной регрессией на примере данных `iris` и `Advertising`. Напомню, что данные `Advertising` (как и все остальные данные-примеры из "Introduction to Statistical Learning") можно скачать на странице книги: `http://www-bcf.usc.edu/~gareth/ISL/data.html`.

2 Теоретическое

Задание 2.1. На слайде 16 презентации приведены примеры различных моделей. Проинтерпретируйте их и объясните результат.

Задание 2.2. Проверьте корректность формул (со слайдов) для коэффициентов регрессии.

Задание 2.3. Ответить (с обоснованием) на следующие вопросы:

1. Есть ли связь между продажами и рекламными бюджетами?
2. Если есть, то насколько сильная?
3. Влияние оказывают все виды рекламы или какие-то особенные? Какие именно?
4. Является ли связь линейной?
5. Есть ли взаимное влияние между разными видами рекламы?

Комментарии: Корректным измерением "силы связи" является коэффициент множественной корреляции R (как относительная мера уменьшения дисперсии остатков). Задание можно пока рассматривать как сугубо теоретическое (надо еще раз переосмыслить то, что было на слайдах последние два раза), но в следующую пятницу мы посмотрим на то, как именно можно повторить это исследование самостоятельно и доведем его до конца.

3 Практическое

Задание 3.1. Для данных `Advertising` повторить анализ из презентации и моей демонстрации, а также проверить эффективность линейной регрессии с помощью тестовой подвыборки (test subset), т.е. разделить данные на обучающую и тестовую подвыборки отношении 2:1, построить модель регрессии по обучающей выборке, предсказать значения на тестовой и обучающей и:

1. Изобразить скаттерплот предсказанных (predicted) и реальных значений продаж на тестовой и обучающей выборке
2. Вычислить среднюю ошибку (в смысле остаточной суммы квадратов) на обучающей и тестовой выборке. Сравнить. Проинтерпретировать результат
3. Повторить предыдущий пункт для различных моделей — к примеру, удалить незначимый (по t-критерию) признак, удалить информативный признак, удалить вообще все признаки, кроме сдвига (intercept, β_0). Проинтерпретировать результат.

4 Дополнительные задачи

Задание 4.1. Для проверки значимости регрессии в целом (т.е. гипотезы о неравенстве нулю всех коэффициентов) часто применяют так называемый критерий Фишера. Вычисляется статистика, имеющая смысл значимости регрессии:

$$F = \frac{(\mathrm{TSS} - \mathrm{RSS})/p}{\mathrm{RSS}/(n-p-1)},$$

где RSS — сумма квадратов остатков (residuals) полной модели, TSS — сумма квадратов остатков в модели, состоящей из одного только среднего (т.е. модели, которая прогнозирует все средним значением по популяции, не используя независимые признаки вовсе), p — количество признаков (без учета среднего) в полной модели, n — объем выборки.

Нулевая гипотеза — все коэффициенты $\beta_i = 0$, кроме сдвига. Альтернативная — какой-то коэффициент отличается от нуля. В случае, когда нулевая гипотеза верна, статистика F имеет (при некоторых условиях) *распределение Фишера* с $(p, n-p-1)$ степенями свободы (обозначается $F_{p,n-p-1}$). В случае, когда верна альтернатива, значение статистики стремится к бесконечности[1].

Нашей задачей будет проверить корректность критерия Фишера. Будем действовать по следующей схеме:

1. Фиксируем n, p и коэффициент β_0 (можно взять любое число, хоть 0)[2]
2. Промоделируем независимые переменные как выборки из независимых нормально распределенных случайных величин
3. Зависимую переменную будем моделировать как $y = \beta_0 + \varepsilon$, где ε — независимые одинаково нормально распределенные случайные величины (т.е. мы моделируем нулевую гипотезу)
4. Строим модели (полную и модель-среднее), вычисляем RSS, TSS и F
5. Повторяя пункты 2–4, получаем выборку из значений F
6. Используя критерий Колмогорова-Смирнова (`ks.test()`), проверяем гипотезу о распределении F. интерпретируем результат

Пункты 2–6 имеет смысл повторить несколько раз, чтобы получить выборку из значений p-value Колмогорова-Смирнова и уменьшить эффект случайности. Напомню, что p-value это универсальная статистика критерия, выражающая меру согласия с нулевой гипотезой; в случае, когда нулевая гипотеза верна, распределение p-value должно быть равномерным $U[0,1]$, если неверна — стремиться к нулю.

Задание 4.2. В предыдущем задании заменить нормальное распределение предикторов (независимых признаков) экспоненциальным, равномерным, дискретным бросанием монеты. Проверить, останется ли критерий корректным.

---

[1] Вообще, легко заметить, что значение F-статистики монотонно зависит от значения множественного коэффициента корреляции R:

$$R = \frac{\mathrm{TSS} - \mathrm{RSS}}{\mathrm{TSS}}$$

Таким образом, чем больше F, тем регрегрессия “значимее”

[2] Объясните, почему это значение ни на что не повлияет

Аналогично, заменить распределение остатков ε равномерным и проверить гипотезу для $n = 10, 100, 1000$.

Сделать выводы об условиях применимости критерия. Найти в литературе точные условия для критерия Фишера и проверить свои выводы.