# Задание. Линейная регрессия – цена автомобиля – включение категориальной переменной

(Файл kuiper.tsv. Источник данных: Shonda Kuiper. Introduction to Multiple Regression: How Much Is Your Car Worth? Journal of Statistics Education Volume 16, Number 3 (2008))

Price – цена: рекомендованная розничная цена автомобиля General Motors 2005 года выпуска в отличном состоянии в долл. Все автомобили в этих данных на момент оценки были младше одного года и в отличном состоянии.

Mileage – пробег автомобиля в милях

Make – марка (производитель автомобилей), например, Saturn, Pontiac и Chevrolet.

Model – модель (конкретная модель для каждого производителя), например, Ion, Vibe, Cavalier.

Trim – комплектация (конкретная разновидность модели автомобиля), например SE Sedan 4D, Quad Coupe 2D.

Type – тип кузова: седан, купе и т. д.

Cylinder – количество цилиндров в двигателе.

Liter – объем двигателя в литрах.

Doors – количество дверей.

Cruise – есть ли у автомобиля круиз-контроль (1 = есть)

Sound – есть ли в автомобиле обновленные динамики (1 = есть)

Leather – есть ли в автомобиле кожаные сиденья (1 = есть)

1. Загрузите данные из файла kuiper.tsv (команда read.delim("*имя файла*"), опции skip и stringsAsFactors=TRUE).
2. Постройте регрессию логарифма цены от всех количественных и индикаторных переменных. Вместо количества дверей возьмите индикаторную переменную для 2 дверей (добавить с помощью I(*выражение*)). Команда: *регрессия* <- lm(y ~ x1+x2+…).
3. Вместо количества дверей возьмите категориальную (качественную) переменную типа кузова Type. Сравните показатели точности подгонки регрессии. Стоит ли включать такую переменную?
4. Проведите тест на добавление переменной (команда anova(*модель1*, *модель2*)).
5. Найдите разность расчетных значений (fitted) для двух регрессий. Сильно ли различаются две модели по расчетным значениям?

# Задание. Линейная регрессия – прогноз интенсивности дорожного движения в США

Фирма StreetLight получила оценку интенсивности дорожного движения на дорогах США [и Канады], используя данные мобильных устройств и разные дополнительные данные.

Показатель называется AADT (Annual average daily traffic). Это стандартный показатель в транспортном планировании в США и других стран. Фактические наблюдения AADT собираются автоматическими счетчиками (traffic counters), расставленными на дорогах.

https://www.streetlightdata.com/aadt-average-annual-daily-traffic-count/

Используем данные с графика из отчета по США:

actual – фактические наблюдения,

estimated – рассчитанная фирмой оценка.

Для теоретически правильного прогноза данные должны удовлетворять модели регрессии

с коэффициентами и . (Регрессия Минцера–Зарновица для прогнозов). В частности

имеет нулевое матожидание.

* Загрузите данные из файла StreetLight2020.tsv.
* Постройте точечную диаграмму фактических actual от estimated и добавьте на нее «теоретически правильную» линию регрессии.
* Постройте регрессию actual от estimated . Просмотр результатов – команда summary(*регрессия*).
* Постройте доверительные интервалы (команда coefci(*регрессия*)). Проведите тест на равенство коэффициента при estimated единице.
* Постройте регрессию actual–estimated от estimated (и константы). Проведите тест на равенство коэффициента при estimated нулю. Найдите p-значение. Как данный тест связан с тестом предыдущего пункта?
* Постройте регрессию actual–estimated от константы. Проведите тест на равенство константы нулю. Найдите p-значение. Проведите тот же тест с помощью команды t.test().

Для теоретически правильного прогноза данные должны, в частности, удовлетворять модели регрессии

с коэффициентами , , и .

* Добавьте в регрессию actual от estimated квадрат estimated. (Переменные в правой части формулы регрессии разделяются знаком +, например lm(y ~ x1 + x2). Функции от переменных, записанные формулами, заключаются в функцию I(*формула переменной*)).
* Найдите в результатах регресии информацию для теста на значимость квадрата estimated (нулевая гипотеза – квадрат не нужен).
* Постройте регрессию actual–estimated от estimated и квадрата estimated (и константы). (Формулы в левой части регрессии можно не окружать I()).
* Найдите в результатах последней регресии информацию для одновременного теста на значимость estimated и квадрата estimated (нулевая гипотеза – обе переменные не нужны в регрессии, т. е. регрессия в целом не значима).
* Повторите предыдущий тест с помощью команды anova(*регрессия1, регрессия2*). (Регрессию с одной константой можно построить командой lm(y ~ 1).)
* Проведите в последней регрессии с помощью команды anova() тест на то, что все три коэффициента равны нулю. (Регрессию совсем без регрессоров можно построить командой lm(y ~ 0).)