# Задание. Линейная регрессия – цена автомобиля

(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 график цены от номера наблюдения и гистограмму.
3. Посмотрите на характеристики отдельных переменных (описательные статистики, таблицы частот, …). Какие значения они принимают? Какие переменные количественные, а какие качественные (категориальные)? (Команды table(), unique(), …)
4. С помощью таблиц сопряженности для пар переменных выясните, какие переменные однозначно определяются другими.
5. Выбирите количественную переменную, от которой, по вашему мнению может сильно зависеть цена. Постройте регрессию логарифма цены от выбранной переменной. Команда: *регрессия* <- lm(y ~ x) (опция data=*фрейм* для указания фрейма, константа добавляется по умолчанию). Просмотр результатов – команда summary(*регрессия*).
6. Посмотрите на показатели точности подгонки регрессии – *R*-квадрат, сумму квадратов остатков, стандартную ошибку регрессии. То же для регрессии с одной константой.
7. Постройте соответствующую точечную диаграмму с линией регрессии (команда abline(*регрессия*)).
8. Постройте точечную диаграмму фактического логарифма цены от расчетных значений (используя fitted(*регрессия*)) и добавьте туда линию с наклоном 1 из начала координат (команда abline(c(0,1))).
9. Постройте точечную диаграмму остатков (resid(*регрессия*)) от расчетных значений и добавьте туда линию нуля. Проинтерпретируйте значения остатков.
10. Постройте и проинтерпретируйте доверительный интервал для коэффициента наклона (команда confint(*регрессия*)).
11. Постройте таблицу корреляций и точечные диаграммы всех пар количественных переменных (например, используя psych::pairs.panels). Какие переменные заметно коррелируют между собой?
12. Постройте такую же регрессию без логарифмирования (зависимая перменная в долларах). Постройте и проинтерпретируйте доверительный интервал для коэффициента наклона.
13. 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+…). Проинтерпретируйте знаки коэффициентов. Значимы ли переменные?
14. Посмотрите на показатели точности подгонки регрессии.
15. Постройте точечную диаграмму фактического логарифма цены от расчетных значений и добавьте туда линию с наклоном 1 из начала координат.
16. Постройте и проинтерпретируйте доверительные интервалы для коэффициентов.
17. Постройте точечную диаграмму остатков от расчетных значений, добавьте туда линию нуля. Цвет точек задайте в зависимости от Cylinder. То же самое с Make и Type. (Легенда – команда legend, например, "bottomright", pt.bg – цвет и использовать unique())
18. Удаляя по одной «наименее значимой» переменной постройте серию регрессий со все меньшим числом перменных. Как меняется точность подгонки?