Задание 1. Линейная регрессия – прогноз интенсивности дорожного движения в США (с прошлого занятия)

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

$$\text{actual}_i = \beta_0 + \beta_1 \text{estimated}_i + \varepsilon_i,$$

с коэффициентами $\beta_0 = 0$ и $\beta_1 = 1$. (Регрессия Минцера–Зарновица для прогнозов). В частности

$$\text{actual}_i - \text{estimated}_i = 0 + \varepsilon_i$$

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

$$\text{actual}_i = \beta_0 + \beta_1 \text{estimated}_i + \beta_2 \text{estimated}_i^2 + \varepsilon_i,$$

с коэффициентами $\beta_0 = 0, \beta_1 = 1$, и $\beta_2 = 0$.

- 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)**.)

Задание 2. Плотности и частоты для actual из StreetLight

Построить гистограмму для **actual** и ее логарифма.

Построить ядерную оценку плотности для **actual** и ее логарифма (команда density с опцией **adjust** для выбора ширины полосы или **bw**, график плотности строится функцией **plot**).

Взять границы 0.1, 0.33, 1, 3.3, … и рассчитать частоты в шт. для actual. (Команда hist с опциями **breaks=*границы*** и **plot=FALSE**)

Рассчитать частоты в процентах. Распечатать таблицу с частостами.

Задание 3. Некоторые возможности пакета ggplot2

- Нарисовать оценки плотности actual.
- Нарисовать частоты (см. Задание 2) разными способами.
- Нарисовать точки estimated, actual, линию с наклоном 1, линию регрессии, эллипс.

ggplot2 – пакет для рисования графиков.

Вид команды

```
ggplot(фрейм данных, aes(x =…, y=…)) +
  добавка1()+
  добавка2()+
…
  theme_имя()
```

aes – это «эстетика», которая показывает, что именно изображать

Добавки:

– точки (**geom_point**),

– линии (**geom_line**),

– столбики (**geom_col**),

– текст (**geom_text**),

– сглаженная линия по облаку точек (**geom_smooth**),

– эллипс по облаку точек (**stat_ellipse**),

– гистограмма (**geom_histogram**),

– ядерная оценка плотности (**geom_density**),

– график функции (**stat_function**),

– «бахрома» (**geom_rug**),

– вертикальная линия (**geom_vline**),

– прямая линия (**geom_abline(slope=…, intercept=…)**)

– заголовки и надписи для осей (**labs**),

– поменять местами оси x и y (`coord_flip`),

– полярные координаты (`coord_polar`),

Темы: `theme_light, theme_bw, theme_gray, theme_void`