## Задание. Линейная регрессия – зарплата и температура

**temp** – климатическая норма для средней температуры января (°C)
**wage** – среднемесячная начисленная заработная плата работников организаций (тыс. руб.) по регионам России в 2019 году

- Загрузите данные из файла **temp_wage.tsv** (команда **read.delim("*имя файла*")**).
- Постройте точечную диаграмму зарплаты от температуры. Найдите коэффициент корреляции.
- Постройте регрессию зарплаты от температуры. Команда: ***регрессия* <- lm(y ~ x)** (опция **data=*фрейм*** для указания фрейма). Константа добавляется по умолчанию. Просмотр результатов – команда **summary(*регрессия*)**.
- Проверьте, что $R^2$=corr(y,x)$^2$.
- Добавьте на точечную диаграмму зарплаты от температуры линию регрессии (команда **abline(*регрессия*)**).
- Постройте точечную диаграмму фактической зарплаты от расчетных значений зарплаты (используя **fitted(*регрессия*)**) и добавьте туда линию с наклоном 1 из начала координат (команда **abline(c(0,1))**).
- Проверьте, что $R^2$=corr(y, $\hat{y}$)$^2$.
- Постройте регрессию температуры от зарплаты.
- Проверьте, что $R^2$ такой же.
- Проверьте, что коэффициенты другие, «перевернув» уравнение обратной регрессии и сравнив с исходной.
- Добавьте на точечную диаграмму зарплаты от температуры две линии регрессии (для обратной команда **abline(c(*константа, наклон*))**).

## 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

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

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)`**.)