НИУ ВШЭ, ОП «Политология», 2019 – 2020
Курс «Теория вероятностей и математическая статистика»

Регрессионный анализ.
Лекция 1: план, learning outcomes

1. Сравнение корреляционного анализа и регрессии. Задачи, решаемые регрессионным анализом
2. Обозначение «рабочего поля» на 2-ой курс:
   - регрессия линейная по параметрам (при этом возможно рассмотрение нелинейной функциональной взаимосвязи)
   - на 2-ом курсе интерпретация пока в терминах взаимосвязи зависимой переменной и предиктора, о выявлении эффекта (treatment effect) поговорим на 3-ем курсе
   - пока работаем только с cross-section data
3. Путевая диаграмма для визуализации регрессионной модели. Не забывайте про обозначения: наблюдаемые переменные обозначаются прямоугольниками, латентные – внутри окружности.
4. Аналитическая запись регрессионной модели. Для того, чтобы задать спецификацию регрессионной модели, предварительно определите и обоснуйте набор переменных и функциональную взаимосвязь зависимой и объясняющих переменных.
5. Интерпретация оценок коэффициентов в регрессионной модели (константа (intercept), коэффициент при предикторе))
6. Что содержательно включает в себя ошибка в регрессионной модели?
7. Разница между ошибками (errors) и остатками (residuals). Остатки – оцененные ошибки
8. Метод наименьших квадратов (МНК) / Ordinary least squares (OLS): ключевая идея. Почему минимизируется сумма именно квадратов остатков, а не исходных остатков?
9. Выведение оценок коэффициентов в парной регрессионной модели посредством МНК (OLS)

   Найдем оптимальные оценки константы ($\hat{\beta}_0$) и коэффициента при предикторе ($\hat{\beta}_1$) в парной линейной регрессии, при которых сумма квадратов остатков будет минимальна.

   Рассмотрим частную производную по $\hat{\beta}_0$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_0} = 0$$

$$(-2)\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^{n} x_i = 0$$

$$\hat{\beta}_0 = \frac{\sum_{i=1}^{n} y_i}{n} - \hat{\beta}_1 \frac{\sum_{i=1}^{n} x_i}{n}$$

   Мы получили оценку константы в парной регрессии:

$$\boxed{\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}}$$

Рассмотрим частную производную по $\hat{\beta}_1$:

$$\frac{\partial \sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\partial \hat{\beta}_1} = 0$$

$$(-2)\sum_{i=1}^{n}(x_i)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \hat{\beta}_0 x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

Вспомним, что на предыдущем шаге мы уже получили оценку константы, подставим ее в уравнение:

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n}(\bar{y} - \hat{\beta}_1 \bar{x})x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} \bar{y} x_i + \sum_{i=1}^{n} \hat{\beta}_1 \bar{x} x_i - \sum_{i=1}^{n} \hat{\beta}_1 x_i^2 = 0$$

$$\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \bar{y} + \hat{\beta}_1 \sum_{i=1}^{n} x_i \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i (y_i - \bar{y})}{\sum_{i=1}^{n} x_i (x_i - \bar{x})} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2} = \frac{Cov(x, y)}{Var(x)}$$

10. Условия Гаусса–Маркова (допущения об ошибках в регрессии)