Section 8: Review of regression

STA 35C – Statistical Data Science III

Instructor: Akira Horiguchi

Fall Quarter 2025 (Sep 24 – Dec 12)
MWF, 12:10 PM – 1:00 PM, Olson 158
University of California, Davis

Overview

Based on Chapter 3 of ISL book James et al. (2021).

1. Simple linear regression

2. Idea of multiple linear regression

3. Idea of polynomial regression

Simple linear regression

$Y = f(X) + \varepsilon$

Simple linear regression (SLR): predict a quantitative response Y using a linear relationship between X and Y.

- The relationship can be expressed mathematically as

$$\underline{Y = \beta_0 + \beta_1 X + \varepsilon,}$$

  where β_0 and β_1 are two unknown, fixed constants, that represent the *intercept* and the *slope* in the linear model, and ε is an error term.
- β_0 and β_1 are usually called *coefficients* or *parameters*.
- We sometimes say that we *regress* Y on X.

In practice, the parameters β_0 and β_1 are unknown.

- We aim to estimate these parameters using observed data

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

  (This is the *training data* discussed the previous section.)

- Aim: find estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ to predict Y based on $X = x$ by computing

$$\hat{y} = \underbrace{\hat{\beta}_0 + \hat{\beta}_1 x}.$$

- We would like the resulting line with $\hat{\beta}_0$ and $\hat{\beta}_1$ to be a good representation of the relationship between Y and X.

Our chosen estimators will be derived using the *ordinary least squares (OLS)* method.

- For each observed (x_i, y_i), let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X. Then $e_i := y_i - \hat{y}_i$ represents the ith *residual*.
- Any pair $\left(\hat{\beta}_0, \hat{\beta}_1\right)$ of estimators induces a *residual sum of squares* (RSS):

$$
\begin{aligned}
RSS\left(\hat{\beta}_0, \hat{\beta}_1\right) &:= e_1^2 + e_2^2 + \cdots + e_n^2 \\
&= \left(y_1 - \hat{y}_1\right)^2 + \left(y_2 - \hat{y}_2\right)^2 + \cdots + \left(y_n - \hat{y}_n\right)^2 \\
&= \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2.
\end{aligned}
$$

- The *OLS coefficient estimates* are defined as the minimizers of the RSS; one can show (using calculus) that these estimates are

$$
\hat{\beta}_1^{OLS} = \frac{\sum_{i=1}^n (x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^n (x_i - \overline{x})^2} \quad \text{and} \quad \hat{\beta}_0^{OLS} = \overline{y} - \hat{\beta}_1^{OLS}\overline{x} \tag{1}
$$

where $\overline{x} := \frac{1}{n}\sum_{i=1}^n x_i$ and $\overline{y} := \frac{1}{n}\sum_{i=1}^n y_i$ are the sample means.

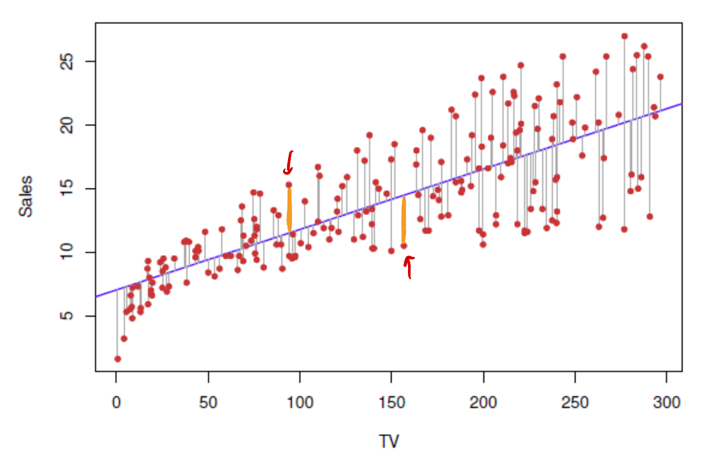

Figure 1: Image by James et al. (2021). The least squares fit for the regression of sales onto TV advertising budget. The fit is found by minimizing the residual sum of squares (RSS). Each grey line segment represents a residual.

$$\text{Bias} = \beta_0 - E(\hat{\beta}_0) \Rightarrow$$

0-Bias means $E(\hat{\beta}_0) = \beta_0$

The OLS estimators $\hat{\beta}_0^{OLS}$ and $\hat{\beta}_1^{OLS}$ are *unbiased estimators* for the unknown parameters β_0 and β_1, respectively.

- Mathematically, this means $E\left(\hat{\beta}_0^{OLS}\right) = \beta_0$ and $E\left(\hat{\beta}_1^{OLS}\right) = \beta_1$.
- Intuitively, this means that the estimator does not *systematically* over- or under-estimate the true parameter.

Idea of multiple linear regression

$Y = f(X_1, \dots, X_p) + \varepsilon$

Multiple linear regression extends SLR by allowing more than one predictor.

- In multiple linear regression, the response Y is modelled by

$$Y = \underbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \cdots + \beta_p X_p} + \varepsilon, \tag{2}$$

where $p \in \mathbb{N}$ is the number of predictors.

- We interpret β_j as the average effect on Y of a one unit increase in X_j, holding all other predictors fixed. (Math?)

$\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$ vs $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_3 + 1)$

$\left[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_3 + 1)\right] - \left[\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3\right]$

$= \beta_3$

Example

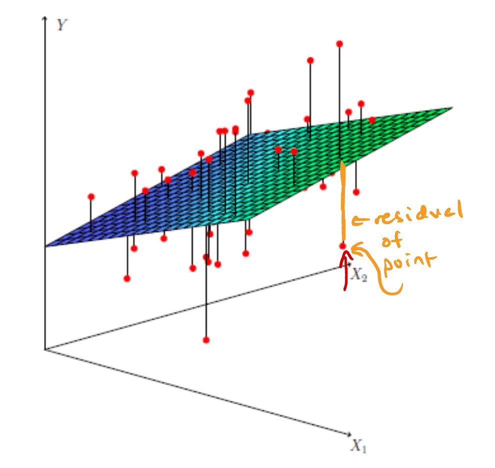

Figure 2: Image by James et al. (2021). The response depends on two predictors in a linear manner as the least squares regression line is a plane (chosen by minimizing the sum of the squared vertical distances between each observation (shown in red) and the plane).

As with SLR, our chosen estimators will be derived using the OLS method.

- For each observed $(\mathbf{x}_i, y_i)$, where $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathbb{R}^p$, let

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} \tag{3}$$

  be the prediction for Y based on the ith value of $\mathbf{X}$.

- Any vector $\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right)$ of estimators induces a *residual sum of squares* (RSS):

$$\begin{aligned} RSS\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right) &:= \sum_{i=1}^{n} e_i^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \left[\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip}\right]\right)^2 \end{aligned}$$

- The *OLS coefficient estimates* $\hat{\beta}_0^{OLS}, \hat{\beta}_1^{OLS}, \ldots, \hat{\beta}_p^{OLS}$ are defined as the minimizers of the RSS; closed-form expressions can be obtained by calculus. (Without *matrix algebra*, these expressions can be scary to look at.)

Idea of polynomial regression

Idea

Polynomial regression extends SLR by allowing sums of predictors raised by powers.

- The response Y is modelled using an order-d polynomial of the predictor X_1:

$$Y = \overbrace{\beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_1^3 + \cdots + \beta_d X_1^d} + \varepsilon. \tag{4}$$

- The order/degree d describes the flexibility of the model.
- As with SLR, our chosen estimators will be derived using the OLS method.

Example 1: A non-linear function

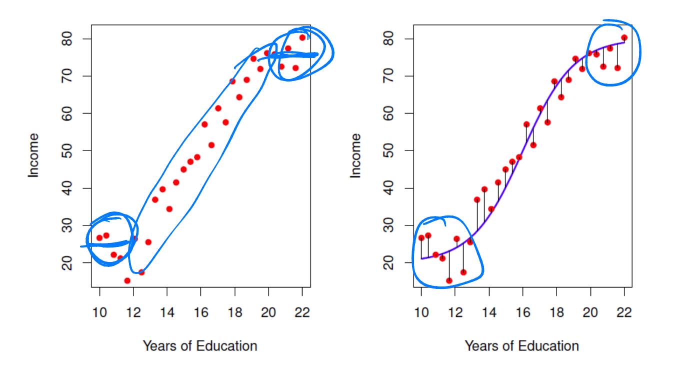

Figure 3: Image by James et al. (2021), based on the `Income` data set in R. The red dots are the observed values of `income` in tens of thousand dollars and `years of education` for 30 individuals.

Example 2: degree-4 polynomial

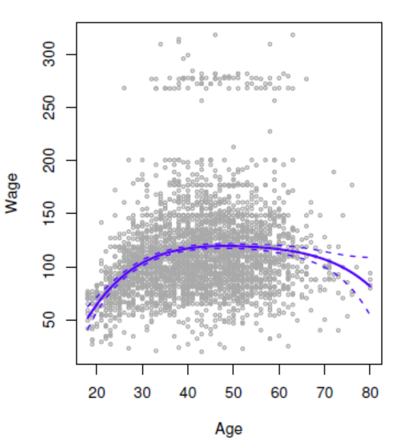

Figure 4: Image by James et al. (2021). The solid blue curve is a degree-4 polynomial of wage (in thousands of dollars) as a function of age, fit by least squares.

Example 3: Polynomial regression with two predictors

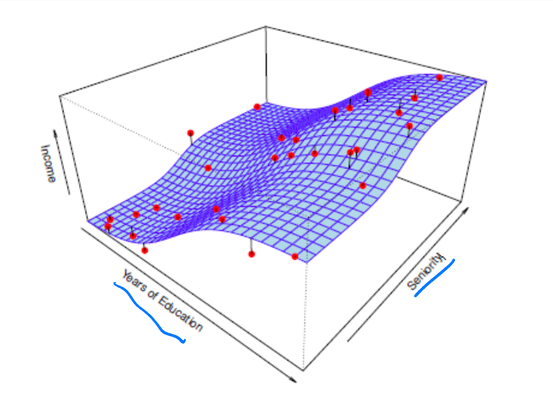

Figure 5: Image by James et al. (2021), based on the `Income` data set in R. The `income` is displayed as a function of `years of education` and `seniority`, where linearity does not seem appropriate. It might be reasonable to do polynomial regression with two predictors.