Linear Regression

Main idea

- Regression analysis examines the relation between a single dependent variable Y and one or more independent variables $x_1, \ldots, x_p$.
- Regression analysis describes the conditional distribution of Y given $x_1, \ldots, x_p$: $f(Y|x_1, \ldots, x_p)$. Usually we describe the mean of this distribution.
- It can be used for:
  - describing how Y depends on $x_1, \ldots, x_p$
  - predicting Y from $x_1, \ldots, x_p$

2 / 18

Linear regression

- Full name: Ordinary least squares multiple linear regression.
- Assumptions of linear regression (see overhead):
  - Data is representative for the population of interest.
  - $E(Y|x_1, \ldots, x_p)$ is a *linear* function of $x_1, \ldots, x_p$.
  - The variance of $f(Y|x_1, \ldots, x_p)$ does not depend on $x_1, \ldots, x_p$.
  - $f(Y|x_1, \ldots, x_p)$ is (approximately) normal.

3 / 18

Linear regression?

- Why linear regression?
  - Sometimes data (nearly) satisfy the assumptions.
  - Sometimes the assumptions can be (nearly) satisfied by transforming the data.
  - There are many useful extensions of linear regression: weighted regression, robust regression, nonparametric regression, generalized linear models, high-dimensional linear regression.
- How does linear regression work? We start with one independent variable.

Simple linear regression

Linear model

- Linear statistical model: $Y = \beta_0 + \beta_1 x + E$.
- β_0 is the intercept of the line, and β_1 is the slope of the line. One unit increase in x is associated with β_1 units increase in Y, on average.
- E is called a statistical error. It accounts for the fact that the statistical model does not give an exact fit to the data.
- Statistical errors can have a fixed and a random component.
  - Fixed component: arises when the true relation is not linear (also called lack of fit error, bias) - we assume this component is negligible.
  - Random component: due to measurement errors in Y, variables that are not included in the model, random variation.

Statistical errors

- We assume:
  - $E(E_i) = 0$
  - $\text{Var}(E_i) = \sigma^2$ for all $i = 1, \ldots, n$
  - $\text{Cov}(E_i, E_j) = 0$ for all $i \neq j$
  - The E_i are normally distributed
- Then:
  - $E_i \sim N(0, \sigma^2)$, i.i.d.
  - $Y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$, Y_i independent

Linear model

- The *population parameters* β_0, β_1 and σ are unknown. We use lower case Greek letters for population parameters.
- Based on data $(x_1, y_1), \ldots, (x_n, y_n)$ we compute *estimates* of the population parameters: $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\sigma}$.
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ is called the *fitted value*.
- $r_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)$ is called the *residual*.
- The residuals are observable, and can be used to check assumptions on the statistical errors E_i.
- Points above the line have positive residuals, and points below the line have negative residuals.
- A line that fits the data well has small residuals.

Small residuals

- We want the residuals to be small in *magnitude*, because large negative residuals are as bad as large positive residuals.
- So we cannot simply require $\sum r_i = 0$.
- In fact, any line through the means of the variables - the point $(\bar{x}, \bar{y})$ - satisfies $\sum r_i = 0$.
- Two immediate solutions:
  - Require $\sum |r_i|$ to be small.
  - Require $\sum r_i^2$ to be small.
- We consider the second option because working with squares is mathematically easier than working with absolute values (for example, it is easier to take derivatives). However, the first option is more resistant to outliers.
- Eyeball regression line (see overhead).

9 / 18

Minimize $\sum r_i^2$

- SSE stands for *Sum of Squared Error*.
- We want to find the pair $(\hat{\beta}_0, \hat{\beta}_1)$ that minimizes $SSE(\beta_0, \beta_1) := \sum (y_i - \beta_0 - \beta_1 x_i)^2$.
- Thus, we set the partial derivatives of $SSE(\beta_0, \beta_1)$ with respect to β_0 and β_1 equal to zero:
  - Differentiate wrt β_0: $\sum (-1)(2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
    $\Rightarrow \sum (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$.
  - Differentiate wrt β_1: $\sum (-x_i)(2)(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$
    $\Rightarrow \sum x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0$.
- We now have two *normal equations* in two unknowns $\hat{\beta}_0$ and $\hat{\beta}_1$. The solution is:
  - $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}$
  - $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$

10 / 18

Regression to the mean

- Note that:

$$\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$
$$= \frac{\frac{1}{n-1}\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}\sqrt{\frac{1}{n-1}\sum(y_i - \bar{y})^2}} \frac{\sqrt{\frac{1}{n-1}\sum(y_i - \bar{y})^2}}{\sqrt{\frac{1}{n-1}\sum(x_i - \bar{x})^2}}$$
$$= r\frac{s_y}{s_x}$$

- The regression line goes through $(\bar{x}, \bar{y})$. Thus, if $x_i = \bar{x} + s_x$, we have $\hat{y}_i = \bar{y} + rs_y$. Thus, observations that are one standard deviation of x above the average of x, will be, on average, less than one standard deviation of y above/below the average of y (since $|r| \leq 1$).
- This is called "regression to the mean".

11 / 18

Comments

- Note that $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables. See simulation.
- One can construct tests and confidence intervals for β_0 and β_1.
- One can also construct confidence intervals for $E(Y_i|x_i)$ and for $Y_i|x_i$.

12 / 18

Model fit

- One should check residual plots to see if the assumptions on the E_i's are (roughly) satisfied: *diagnostic plots*.
- Compute $\hat{\sigma} = \sqrt{SSE/(n-2)} = \sqrt{\frac{\sum r_i^2}{n-2}}$. This is an estimate for σ. Interpretation:
  - the smaller $\hat{\sigma}$, the closer points are to the line.
  - if the residuals are approximately normal, then about 2/3 of the observations are in the range $\pm\hat{\sigma}$ from the line and about 95% are in the range $\pm 2\hat{\sigma}$ from the line (in vertical direction).

13 / 18

Multiple linear regression

14 / 18

≥ 2 independent variables

- $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + E.$
- This describes a plane in the 3-dimensional space $\{X_1, X_2, Y\}$ (see figure):
  - β_0 is the intercept
  - β_1 is the average increase in Y associated with a one-unit increase in x_1 when x_2 is held constant
  - β_2 is the average increase in Y for a one-unit increase in x_2 when x_1 is held constant.

15 / 18

Statistical error

- $Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + E_i$, where E_i is the statistical error for the ith case.
- We make the same assumptions about E as before:
  - $E(E_i) = 0$
  - $\text{Var}(E_i) = \sigma^2$ for all $i = 1, \ldots, n$
  - $\text{Cov}(E_i, E_j) = 0$ for all $i \neq j$
  - The E_i are normally distributed

Estimates and residuals

- The *population parameters* β_0, β_1, β_2, and σ are unknown.
- Based on data $(x_{11}, x_{12}, y_1), \ldots, (x_{n1}, x_{n2}, y_n)$, we compute *estimates* of the population parameters: $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\beta}_2$ and $\hat{\sigma}$.
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 X_{i2}$ is called the *fitted value*.
- $r_i = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2})$ is called the *residual*.
- The residuals are observable, and can be used to check assumptions on the statistical errors E_i.
- Points above the *plane* have positive residuals, and points below the plane have negative residuals.
- A plane that fits the data well has small residuals.

Computing estimates

- The triple $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ minimizes $SSE(\beta_0, \beta_1, \beta_2) = \sum r_i^2 = \sum (Y_i - \beta_0 - \beta_1 X_{i1} - \beta_2 X_{i2})^2$.
- We can again take partial derivatives and set these equal to zero.
- This gives three equations in the three unknowns β_0, β_1 and β_2. Solving these *normal equations* gives the regression coefficients $\hat{\beta}_0$, $\hat{\beta}_1$ and $\hat{\beta}_2$.
- The same procedure works for p independent variables $X_1, \ldots, X_p$. However, it is then easier to use matrix notation. Concept for model checking and model fit also go through.