

AAIL100: Introduction to Data Science

Ragesh Jaiswal, CSE, IIT Delhi

Prediction

Prediction

- Until now, we have mainly focused on the following data analysis tasks:
 1. Hypothesis Testing
 2. Estimating some parameter of the population
- Another important task is **prediction**
 - Example: Based on a person's internet profile, which websites are likely to interest them?
- We start with one-variable prediction:
 - Predicting the value of one variable based on the value of another
 - Let us build a predictor for a child's height based on the average height of their parents.
- The simplest and most commonly used predictor is a **linear predictor**, which works well if there is a **linear association** between the variables.
 - Question: Is there a measure for linear association?
 - Qualitative: How tightly clustered a scatter diagram is about a straight line.
 - Quantitative: **Correlation coefficient** or simply **correlation** r , has the following desirable properties:
 - $-1 \leq r \leq 1$
 - $r = 1$ if the scatter plot is a perfect straight line sloping upwards and $r = -1$ if the scatter plot is a perfect straight line sloping downwards.
 - Candidate measure: r is the average of the product of the two variables in standard units. This is also known as **Pearson correlation**.

- Where is the line given r ?
 - The line through $(0, 0)$ with slope r passes through the “core” of the data points in the scatter plot in the standard units.
 - The slope being ≤ 1 is called **regression** to the mean.
- How does the line give a predictor?
 - A good predictor for y based on x , would consider the points in the scatter plot with similar values of x and take the average of their y values.

$$\frac{\bar{y}_i - \mu_Y}{\sigma_Y} = r \cdot \frac{x_i - \mu_X}{\sigma_X}, \text{ for every } i.$$

• So,

$$\bar{y}_i = r \cdot \frac{\sigma_Y}{\sigma_X} \cdot x_i + \left(\mu_Y - r \cdot \mu_X \cdot \frac{\sigma_Y}{\sigma_X} \right)$$

Slope

Intercept

Prediction

- *The line through $(0, 0)$ with slope r passes through the “core” of the data points in the scatter plot in the standard units.*
- Let's define and find the “*best-fit line*” through our scatter plot and check how the regression line compares
 - Consider the line $ax + b$
 - If this were to be used to predict the y variable from x_i 's, then for each prediction, there is an associated prediction **error**.
 - *What is a reasonable quantification of prediction error?*

- $\frac{\bar{y}_i - \mu_Y}{\sigma_Y} = r \cdot \frac{x_i - \mu_X}{\sigma_X}$, for every i .
- So,
$$\bar{y}_i = \underbrace{r \cdot \frac{\sigma_Y}{\sigma_X}}_{\text{Slope}} \cdot x_i + \underbrace{(\mu_Y - r \cdot \mu_X \cdot \frac{\sigma_Y}{\sigma_X})}_{\text{Intercept}}$$
- This is called the **regression line**.

Prediction

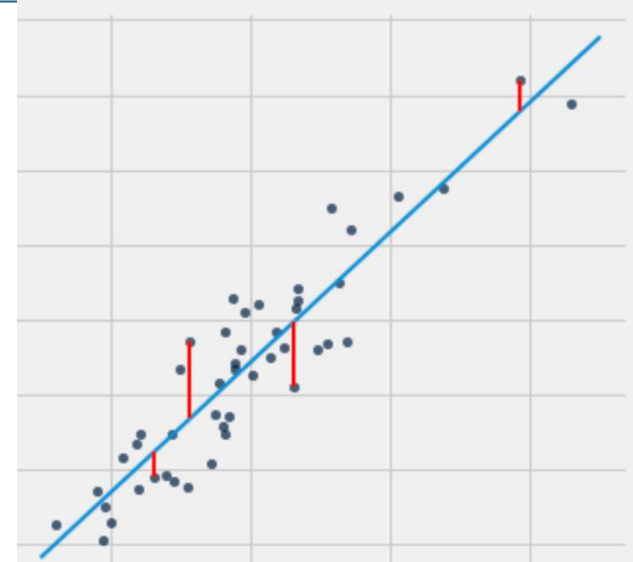
- The line through $(0, 0)$ with slope r passes through the “core” of the data points in the scatter plot in the standard units.
- Let's define and find the “best-fit line” through our scatter plot and check how the regression line compares
 - Consider the line $ax + b$
 - If this were to be used to predict the y variable from x_i 's, then for each prediction, there is an associated **prediction error**.
 - What is a reasonable quantification of prediction error?
 - **Mean Squared Error**: $\frac{1}{n} \sum_i (y_i - (ax_i + b))^2$

- $\frac{\bar{y}_i - \mu_Y}{\sigma_Y} = r \cdot \frac{x_i - \mu_X}{\sigma_X}$, for every i .
- So,
$$\bar{y}_i = \underbrace{r \cdot \frac{\sigma_Y}{\sigma_X}}_{\text{Slope}} \cdot x_i + \underbrace{(\mu_Y - r \cdot \mu_X \cdot \frac{\sigma_Y}{\sigma_X})}_{\text{Intercept}}$$
- This is called the **regression line**.

Prediction

- The line through $(0, 0)$ with slope r passes through the “core” of the data points in the scatter plot in the standard units.
- Let's define and find the “best-fit line” through our scatter plot and check how the regression line compares
 - Consider the line $ax + b$
 - If this were to be used to predict the y variable from x_i 's, then for each prediction, there is an associated prediction **error**.
 - **Mean Squared Error**: $\frac{1}{n} \sum_i (y_i - (ax_i + b))^2$
 - What are the values of a, b that minimizes MSE?

- $\frac{\bar{y}_i - \mu_Y}{\sigma_Y} = r \cdot \frac{x_i - \mu_X}{\sigma_X}$, for every i .
- So,
$$\bar{y}_i = \underbrace{r \cdot \frac{\sigma_Y}{\sigma_X}}_{\text{Slope}} \cdot x_i + \underbrace{(\mu_Y - r \cdot \mu_X \cdot \frac{\sigma_Y}{\sigma_X})}_{\text{Intercept}}$$
- This is called the **regression line**.



Prediction

- The line through $(0, 0)$ with slope r passes through the “core” of the data points in the scatter plot in the standard units.
- Let's define and find the “best-fit line” through our scatter plot and check how the regression line compares
 - Consider the line $ax + b$
 - If this were to be used to predict the y variable from x_i 's, then for each prediction, there is an associated prediction **error**.
 - **Mean Squared Error**: $\frac{1}{n} \sum_i (y_i - (ax_i + b))^2$
 - What are the values of a, b that minimizes MSE?
 - **The regression line is the best-fit line!**

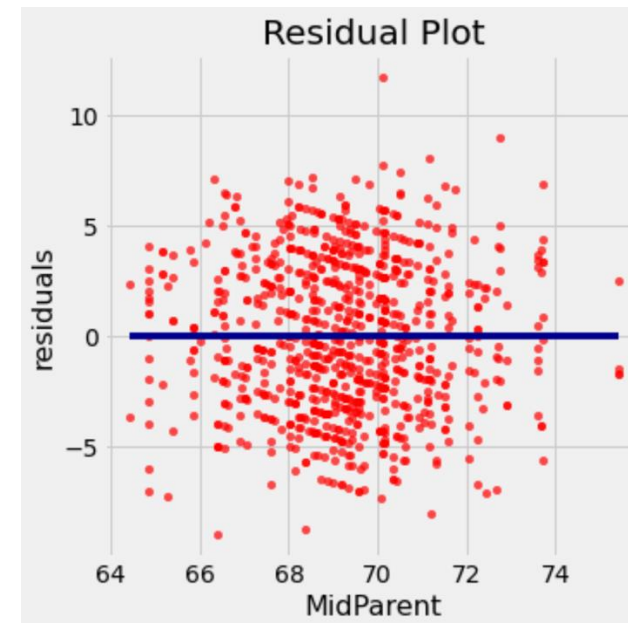
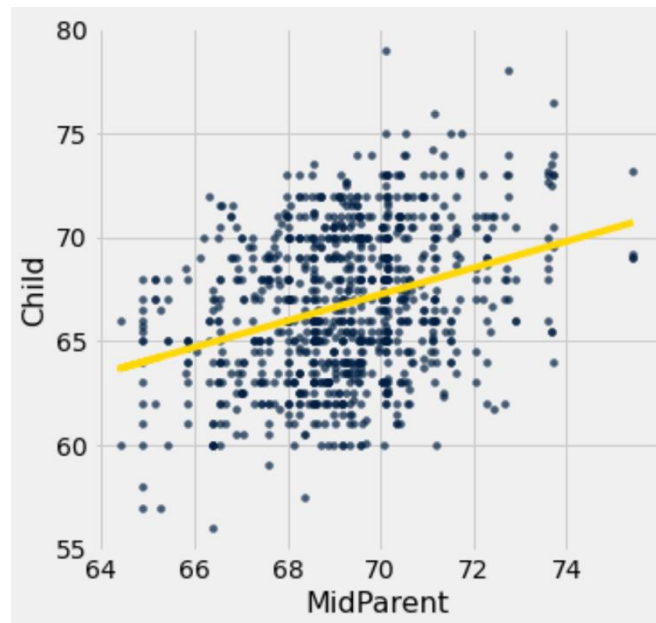
- $\frac{\bar{y}_i - \mu_Y}{\sigma_Y} = r \cdot \frac{x_i - \mu_X}{\sigma_X}$, for every i .
- So,
$$\bar{y}_i = \underset{\substack{\uparrow \\ \text{Slope}}}{r} \cdot \frac{\sigma_Y}{\sigma_X} \cdot x_i + \left(\mu_Y - r \cdot \mu_X \cdot \frac{\sigma_Y}{\sigma_X} \right) \underset{\substack{\uparrow \\ \text{Intercept}}}{}$$
- This is called the **regression line**.

- *The regression line is the best-fit line!*
- The best fit curve through the scatter plot may not be a line. In that case, we may want to consider fitting a higher degree curve. This is known as **non-linear regression**.
 - Example: Instead of fitting a line $ax + b$, we may try to fit a quadratic curve $ax^2 + bx + c$.

- *The regression line is the best-fit line.*
- How do we evaluate the method of linear regression for predicting the **response variable** using the **predictor variable**.
 - **Visual diagnostics**: Compute the **residuals** and check for patterns
$$\begin{aligned} \text{residual} &= \text{observed value} - \text{regression estimate} \\ &= y - \text{fitted value of } y \end{aligned}$$

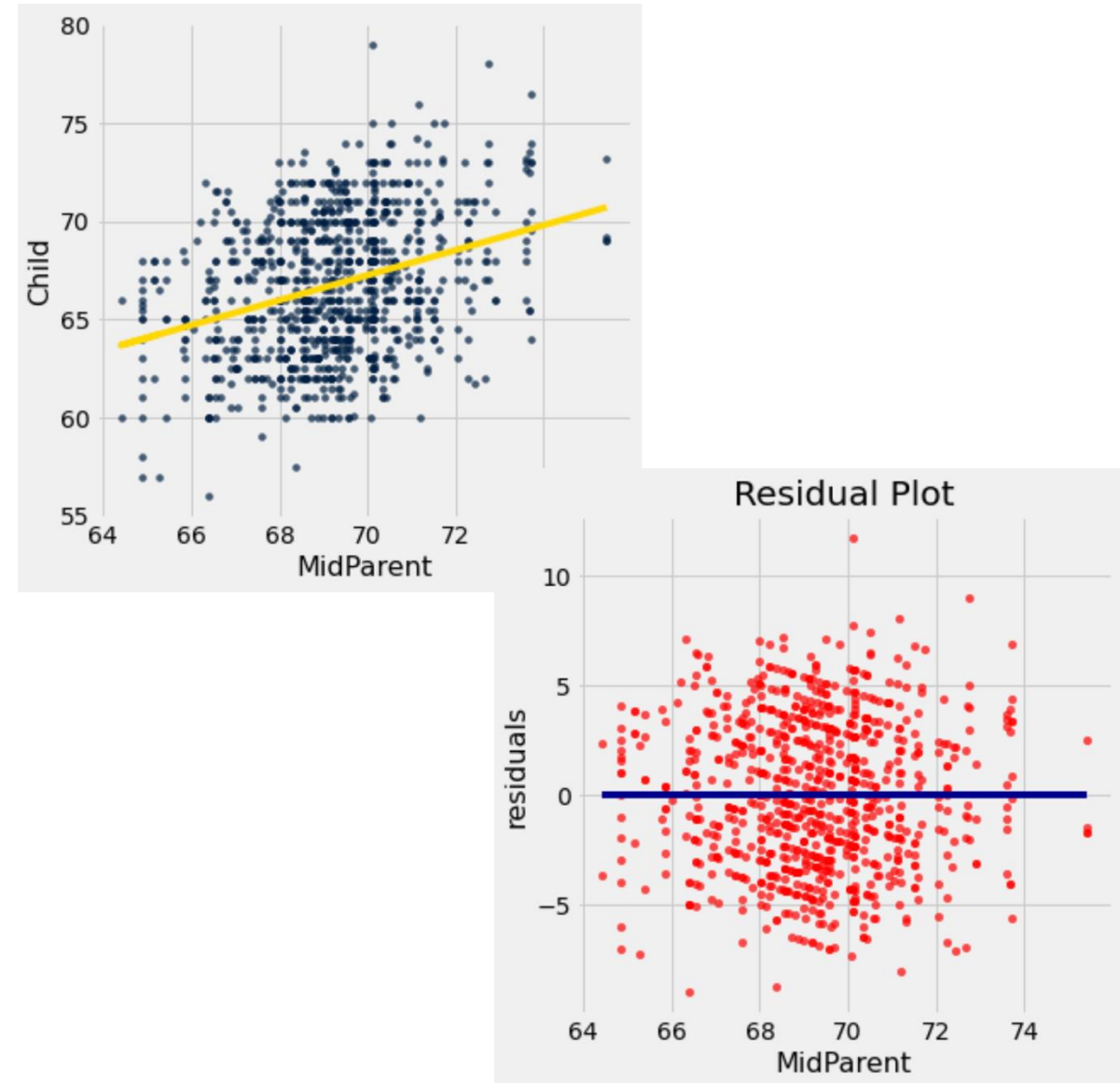
Prediction

- *The regression line is the best-fit line.*
- How do we evaluate the method of linear regression for predicting the **response variable** using the **predictor variable**.
 - **Visual diagnostics**: Compute the **residuals** and check for patterns
$$\text{residual} = \text{observed value} - \text{regression estimate}$$
$$= y - \text{fitted value of } y$$



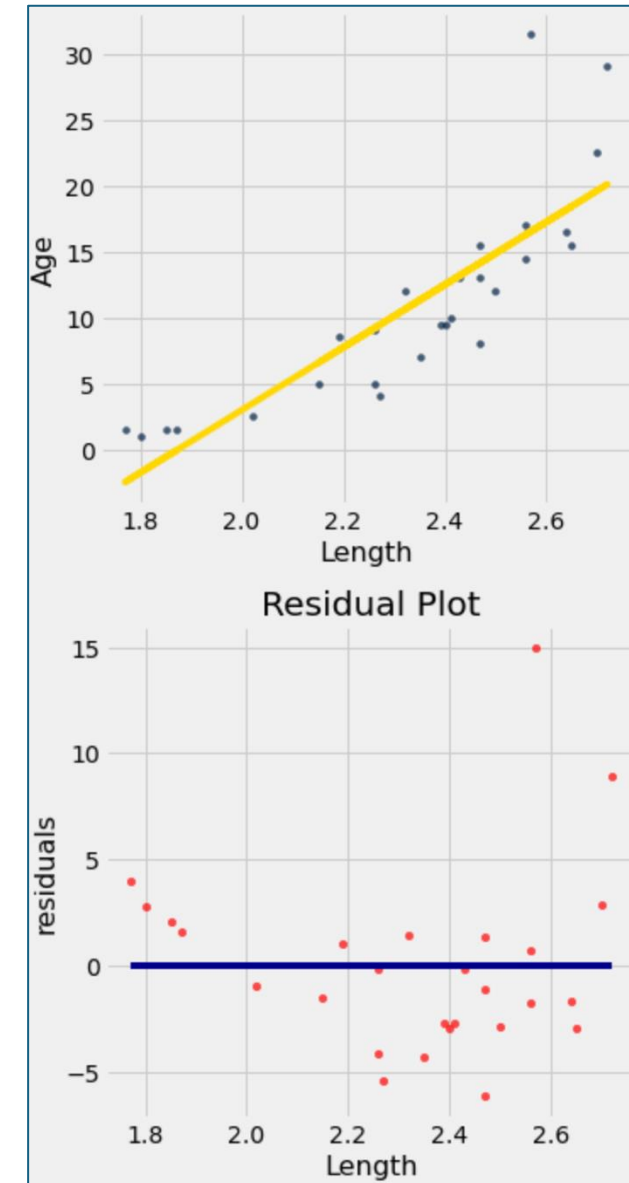
Prediction

- *The regression line is the best-fit line!*
- How do we evaluate the method of linear regression for predicting the **response variable** using the **predictor variable**.
 - **Visual diagnostics:** Compute the **residuals** and check for patterns
 $\text{residual} = \text{observed value} - \text{regression estimate}$
 $= y - \text{fitted value of } y$
 - *The residual plot of a good regression shows no pattern. The residuals look about the same, above and below the horizontal line at 0, across the range of the predictor variable.*



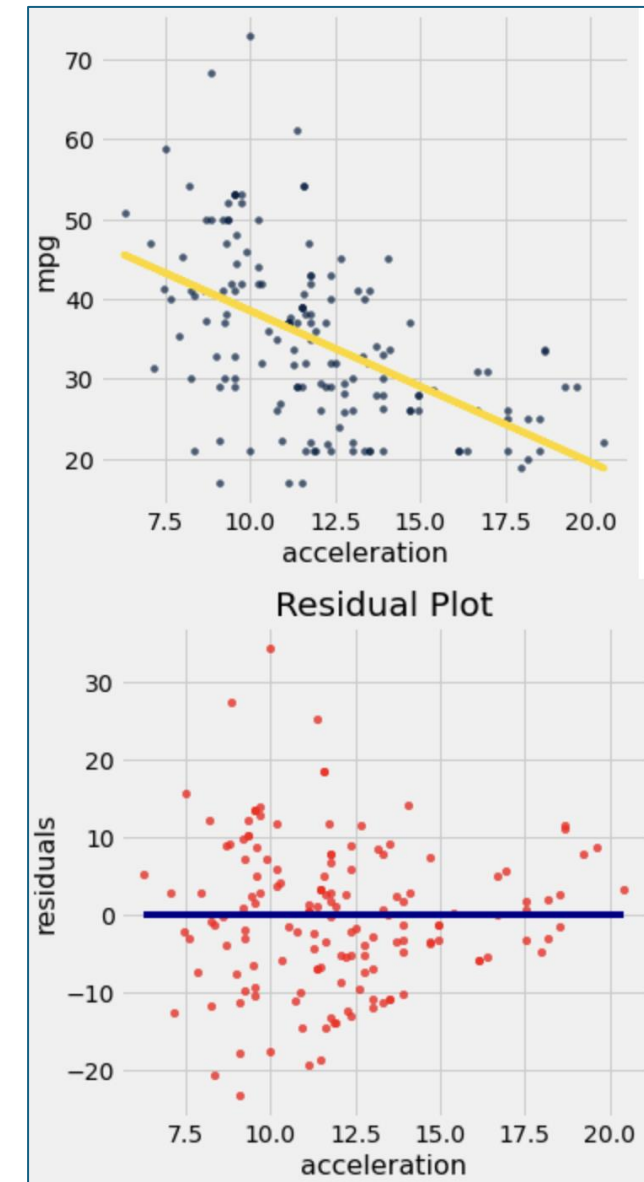
Prediction

- *The regression line is the best-fit line!*
- How do we evaluate the method of linear regression for predicting the **response variable** using the **predictor variable**.
 - **Visual diagnostics:** Compute the **residuals** and check for patterns
 $\text{residual} = \text{observed value} - \text{regression estimate}$
 $= y - \text{fitted value of } y$
 - *The residual plot of a good regression shows no pattern. The residuals look about the same, above and below the horizontal line at 0, across the range of the predictor variable.*
 - Non-linearity: *When a residual plot shows a pattern, there may be a non-linear relation between the variables.*



Prediction

- *The regression line is the best-fit line!*
- How do we evaluate the method of linear regression for predicting the **response variable** using the **predictor variable**.
 - **Visual diagnostics:** Compute the **residuals** and check for patterns
$$\text{residual} = \text{observed value} - \text{regression estimate}$$
$$= y - \text{fitted value of } y$$
 - *The residual plot of a good regression shows no pattern. The residuals look about the same, above and below the horizontal line at 0, across the range of the predictor variable.*
 - Non-linearity: When a residual plot shows a pattern, there may be a non-linear relation between the variables.
 - Heteroscedasticity: If the residual plot shows uneven variation about the horizontal line at 0, the regression estimates are not equally accurate across the range of the predictor variable.



- *The regression line is the best-fit line!*
- How do we evaluate the method of linear regression for predicting the **response variable** using the **predictor variable**.
 - **Visual diagnostics**: Compute the **residuals** and check for patterns
$$\begin{aligned} \text{residual} &= \text{observed value} - \text{regression estimate} \\ &= y - \text{fitted value of } y \end{aligned}$$
 - **Numerical measures**:
 - The average of residuals is 0.
 - The residuals and predictor has no correlation.
 - $\frac{SD \text{ of residuals}}{SD \text{ of } y} = \sqrt{1 - r^2}$ or $r^2 = \frac{\text{Variance of fitted values}}{\text{Variance of } y}$

End