Regression Methods

Understanding Regression

- Regression is concerned with specifying the relationship between a single numeric **dependent variable** (the value to be predicted) and one or more numeric **independent variables** (the predictors).
- As the name implies, the dependent variable depends upon the value of the independent variable or variables.
- The simplest forms of regression assume that the relationship between the independent and dependent variables follows a straight line.

Understanding regression

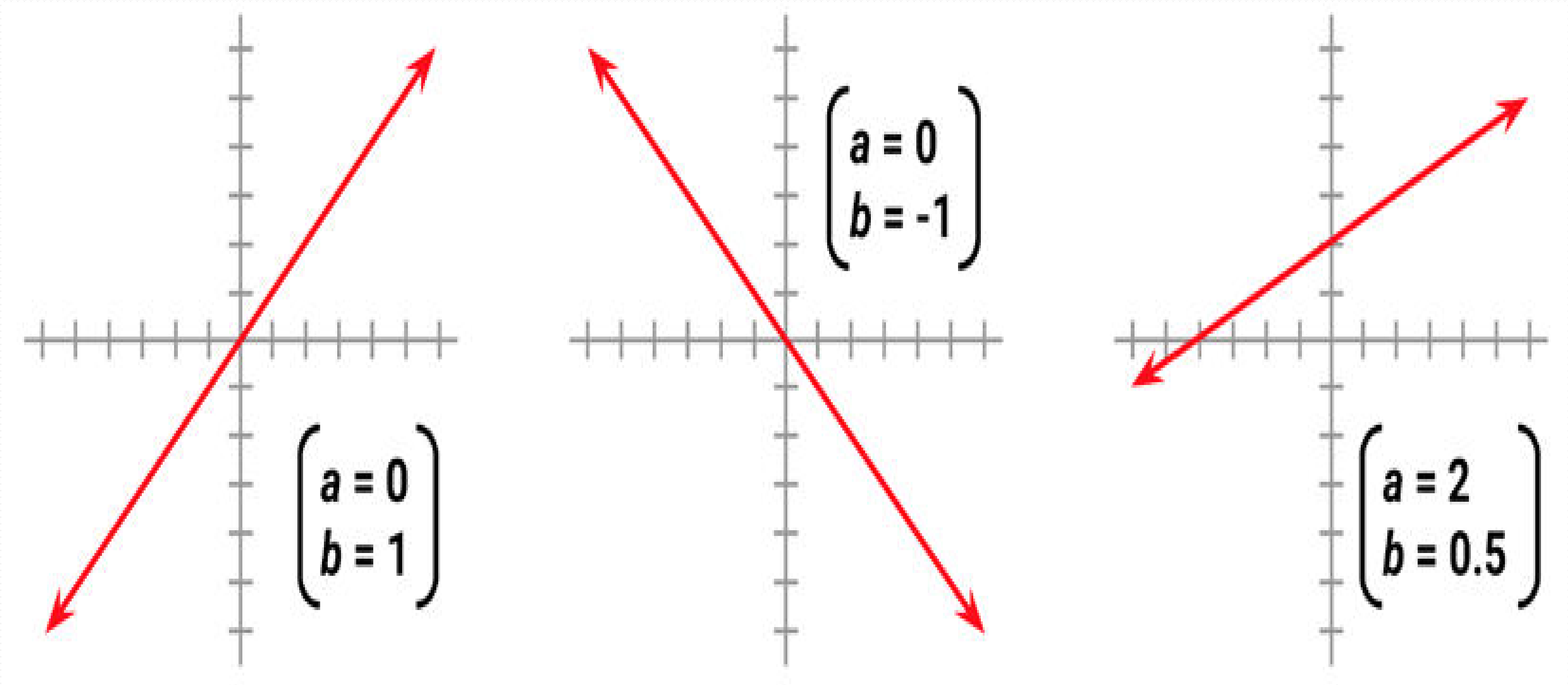
- The origin of the term "regression" to describe the process of fitting lines to data is rooted in a study of genetics by Sir Francis Galton in the late 19th century.
- He discovered that fathers who were extremely short or extremely tall tended to have sons whose heights were closer to the average height.
- He called this phenomenon "regression to the mean".

Understanding regression

- Recall from basic algebra that lines can be defined in a **slope-intercept form** similar to $y = a + bx$.
- In this form, the letter y indicates the dependent variable and x indicates the independent variable.
- The **slope** term b specifies how much the line rises for each increase in x.

Understanding regression

- Positive values define lines that slope upward while negative values define lines that slope downward.
- The term *a* is known as the **intercept** because it specifies the point where the line crosses, or intercepts, the vertical *y* axis. It indicates the value of *y* when $x = 0$.
- Regression equations model data using a similar slope-intercept format.


a = 0
b = 1
a = 0
b = -1
a = 2
b = 0.5

Understanding regression

- The machine's job is to identify values of *a* and *b* so that the specified line is best able to relate the supplied *x* values to the values of *y*.
- Regression analysis is commonly used for modeling
  - complex relationships among data elements,
  - estimating the impact of a treatment on an outcome,
  - and extrapolating into the future.

Understanding regression

- Some specific use cases include:
  - Examining how populations and individuals vary by their measured characteristics, for use in scientific research across fields as diverse as economics, sociology, psychology, physics, and ecology.
  - Quantifying the causal relationship between an event and the response, such as those in clinical drug trials, engineering safety tests, or marketing research.
  - Identifying patterns that can be used to forecast future behavior given known criteria, such as predicting insurance claims, natural disaster damage, election results, and crime rates.

Understanding regression

- Regression methods are also used for **statistical hypothesis testing**, which determines whether a premise is likely to be true or false in light of the observed data.

- The regression model's estimates of the strength and consistency of a relationship provide information that can be used to assess whether the observations are due to chance alone.

Understanding regression

- When there is only a single independent variable it is known as **simple linear regression**.
- In the case of two or more independent variables, this is known as **multiple linear regression**, or simply "multiple regression".
- Both of these techniques assume that the dependent variable is measured on a continuous scale.

Understanding regression

- Regression can also be used for other types of dependent variables and even for some classification tasks.
- **Logistic regression** is used to model a binary categorical outcome.
- **Poisson regression**—named after the French mathematician Siméon Poisson—models integer count data.
- **Multinomial logistic regression** models a categorical outcome for multiclass problems; thus, it can be used for classification.

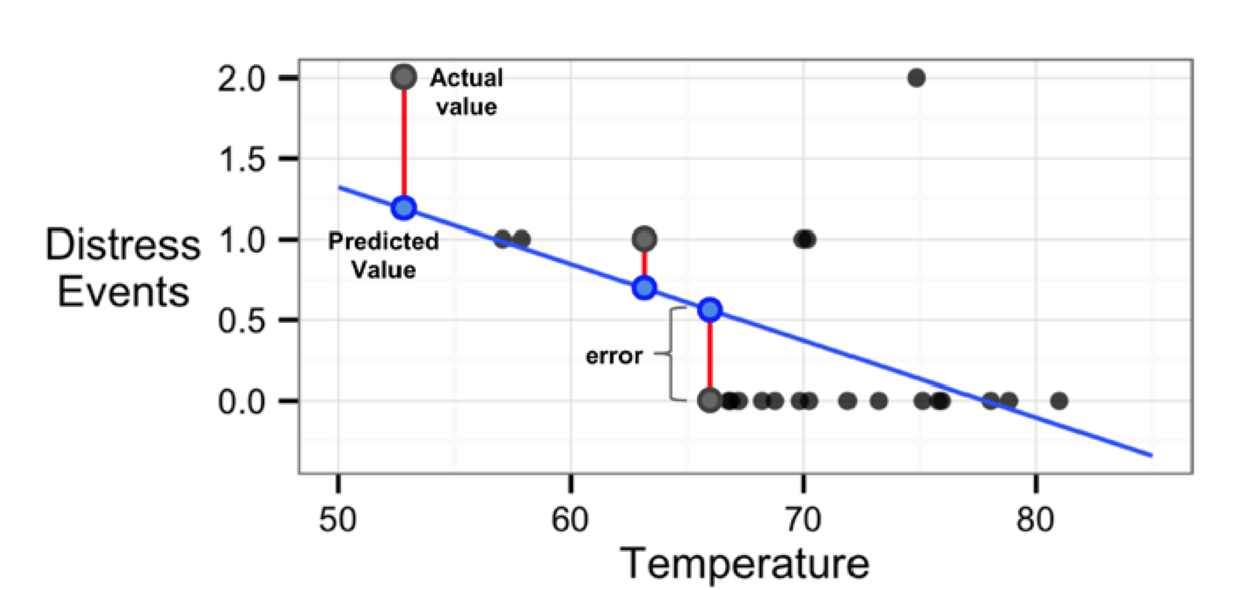
Simple linear regression

- *x* is the independent variable
- *y* is the dependent variable
- The regression model is $y=a+bx + \varepsilon$
- The model has two variables, the independent or explanatory variable, *x*, and the dependent variable *y*, the variable whose variation is to be explained.
- The relationship between *x* and *y* is a linear or straight line relationship.
- Two parameters to estimate – the slope of the line b and the *y*-intercept a (where the line crosses the vertical axis).
- ε is the unexplained, random, or error component.

Ordinary least squares estimation

- In order to determine the optimal estimates of b_0 and b_1, an estimation method known as **Ordinary Least Squares** (**OLS**) was used.
- In OLS regression, the slope and intercept are chosen so that they minimize the sum of the squared errors, that is, the vertical distance between the predicted *y* value and the actual *y* value.
- These errors are known as **residuals**, and are illustrated for several points in the following diagram:

Ordinary least squares estimation



- In mathematical terms, the goal of OLS regression can be expressed as the task of minimizing the following equation:

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

Ordinary least squares estimation

- In plain language, this equation defines *e* (the error) as the difference between the actual *y* value and the predicted *y* value.

$$\sum (y_i - \hat{y}_i)^2 = \sum e_i^2$$

- The error values are squared and summed across all the points in the data.

$$a = \bar{y} - b\bar{x}$$

$$b = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

$$\mathrm{Var}(x) = \frac{\sum(x_i - \bar{x})^2}{n} \quad \mathrm{Cov}(x, y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$b = \frac{\mathrm{Cov}(x, y)}{\mathrm{Var}(x)}$$

Correlations

- The **correlation** between two variables is a number that indicates how closely their relationship follows a straight line.
- Without additional qualification, correlation typically refers to **Pearson's correlation coefficient**, which was developed by the 20th century mathematician Karl Pearson.
- The correlation ranges between -1 and +1. The extreme values indicate a perfectly linear relationship, while a correlation close to zero indicates the absence of a linear relationship.

Correlations

- The following formula defines Pearson's correlation:

$$\rho_{x,y} = \mathrm{Corr}(x, y) = \frac{\mathrm{Cov}(x, y)}{\sigma_x \sigma_y}$$

- Using this formula, we can calculate the correlation between x and y.

Multiple linear regression

- Most real-world analyses have more than one independent variable.
- Therefore, it is likely that you will be using **multiple linear regression** for most numeric prediction tasks.
- We can understand multiple regression as an extension of simple linear regression.
- The goal in both cases is similar—find values of coefficients that minimize the prediction error of a linear equation.
- The key difference is that there are additional terms for additional independent variables.

Multiple linear regression

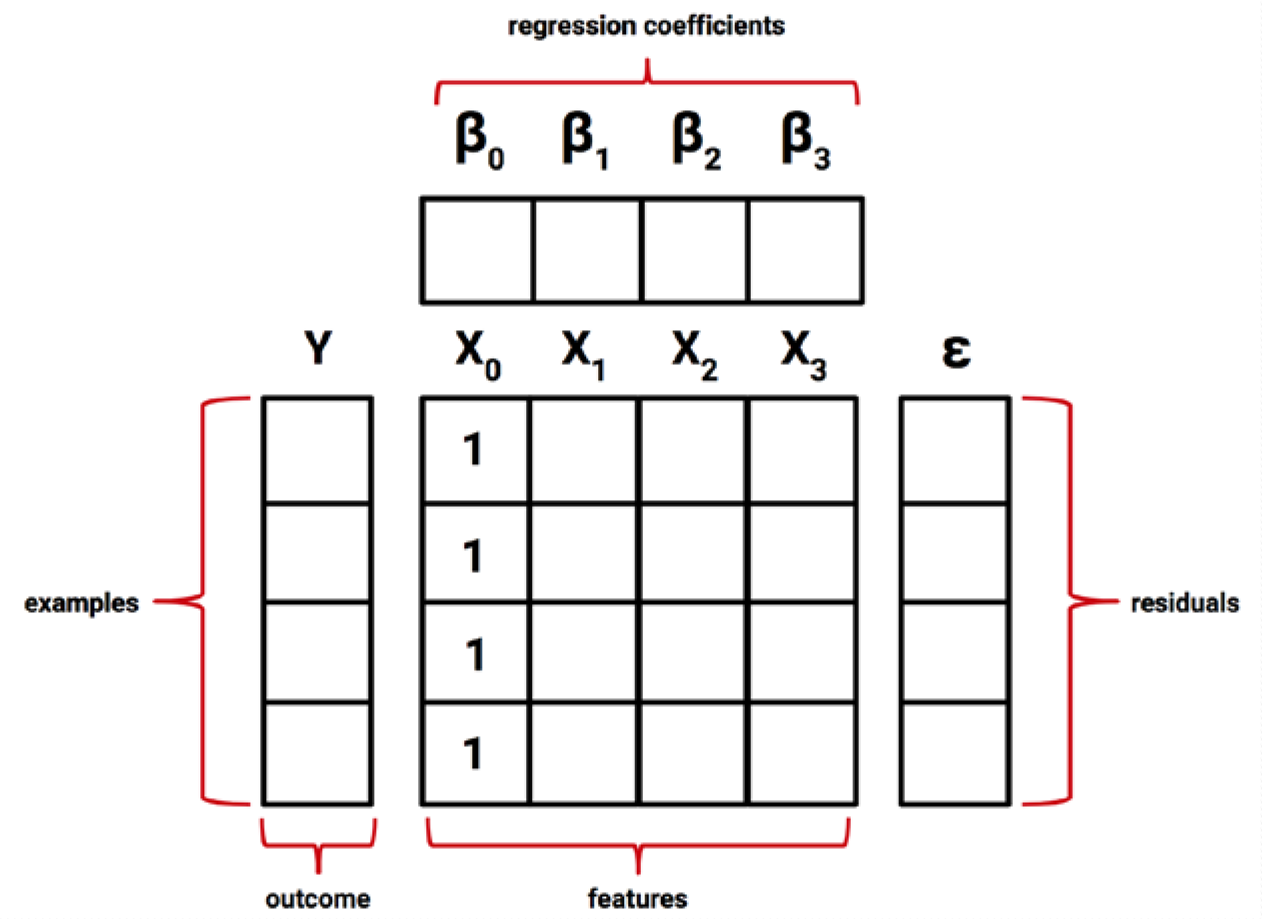
Strengths	Weaknesses
• By far the most common approach for modeling numeric data • Can be adapted to model almost any modeling task • Provides estimates of both the strength and size of the relationships among features and the outcome	• Makes strong assumptions about the data • The model's form must be specified by the user in advance • Does not handle missing data • Only works with numeric features, so categorical data requires extra processing • Requires some knowledge of statistics to understand the model

Multiple linear regression

- Multiple regression equations generally follow the form of the following equation.
- The dependent variable *y* is specified as the sum of an intercept term *α* plus the product of the estimated *β* value and the *x* values for each of the *i* features.
- An error term (denoted by the Greek letter *epsilon*) has been added here as a reminder that the predictions are not perfect. This represents the **residual** term noted previously:

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_i x_i + \varepsilon$$

- In order to estimate the values of the regression parameters, each observed value of the dependent variable *y* must be related to the observed values of the independent
- *x* variables using the regression equation in the previous form. The following figure illustrates this structure:


regression coefficients
β0 β1 β2 β3
Y
X0 X1 X2 X3
ε
1
1
1
1
examples
residuals
outcome
features

- The dependent variable is now a vector, **Y**, with a row for every example $\mathbf{Y} = \beta\mathbf{X} + \varepsilon$
- The independent variables have been combined into a matrix, **X**, with a column for each feature plus an additional column of '1' values for the intercept term.
- The regression coefficients **β** and residual errors **ε** are also now vectors.
- The goal is now to solve for **β**, the vector of regression coefficients that minimizes the sum of the squared errors between the predicted and actual **Y** values. $\hat{\beta} = (\mathbf{X}^{\mathrm{T}}\mathbf{X})^{-1}\mathbf{X}^{\mathrm{T}}\mathbf{Y}$