Regression

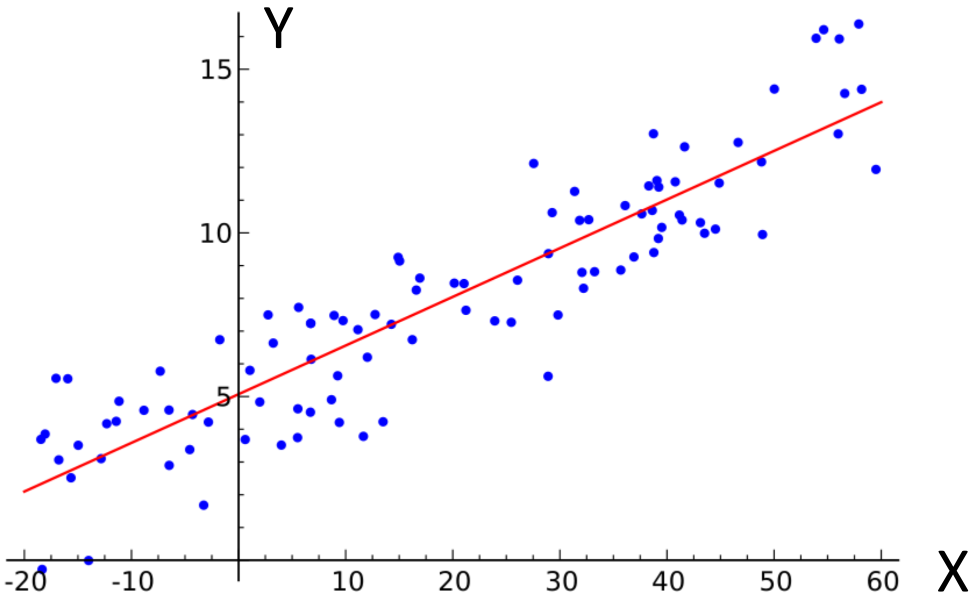
Regression

- Given a dataset containing N observations X_i, Y_i $i = 1, 2, \ldots, N$
- **Regression** is the task of learning a target function f that maps each input attribute set X into an output Y that is *continuous*.
- The goal is to find the target function that can fit the input data with minimum error.
- The error function can be expressed as
  - Absolute Error = $\sum_i |y_i - f(x_i)|$
  - Squared Error = $\sum_i (y_i - f(x_i))^2$

residuals

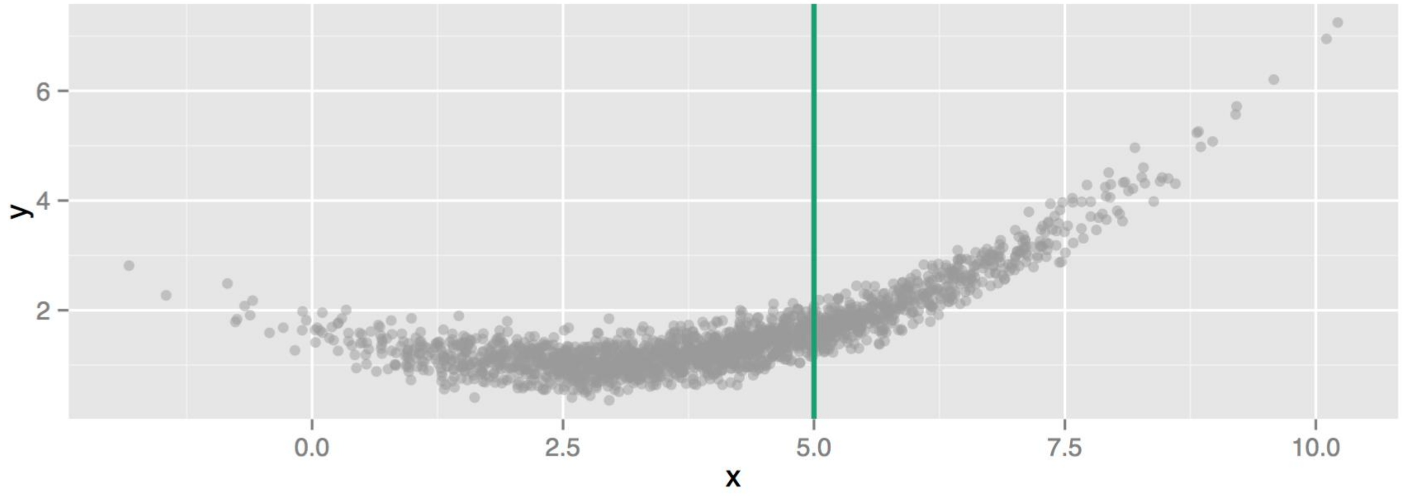
Linear Regression

- **Linear regression** is a linear approach to modeling the relationship between a *dependent variable Y* and one or more *independent* (explanatory) variables *X*.
- The case of *one* explanatory variable is called **simple linear regression**.
- For *more than one explanatory variable*, the process is called **multiple linear regression**.
- For *multiple correlated dependent variables*, the process is called **multivariate linear regression**.



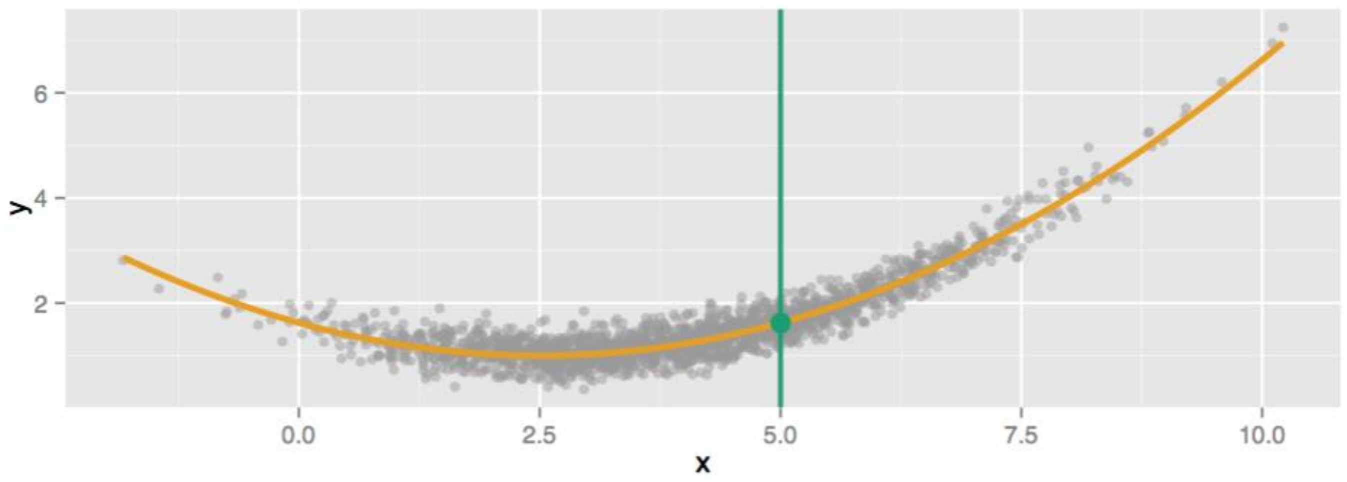
What does it mean to predict Y?

- Look at *X = 5*. There are many different *Y* values at *X=5*.
- When we say predict *Y* at *X =5*, we are really asking:
- What is the expected value (average) of *Y* at *X = 5*?



What does it mean to predict Y?

- Formally, the ***regression function*** is given by *E(Y|X=x)*. This is the expected value of Y at X=x.
- The ideal or optimal predictor of *Y* based on *X* is thus
  - f(X) = E(Y | X=x)



Simple Linear Regression

Dependent Variable

Independent Variable

Linear Model: $$f(x) = \omega_1 x + \omega_0,$$

Slope

Intercept (bias)

- In general, such a relationship may not hold exactly for the largely unobserved population
- We call the unobserved deviations from Y the errors.
- The goal is to find estimated values for the parameters ($\boldsymbol{w_1}$, $\boldsymbol{w_0}$) which would provide the "best" fit for the data points.

Least Square Method

- A standard approach for doing this is to apply the **method of least squares** which attempts to find the parameters *m*, *b* that minimizes the sum of squared error.

$$\text{SSE} = \sum_i (y_i - f(x_i))^2 = \sum_i (y_i - w_1 x_i - w_0)^2$$

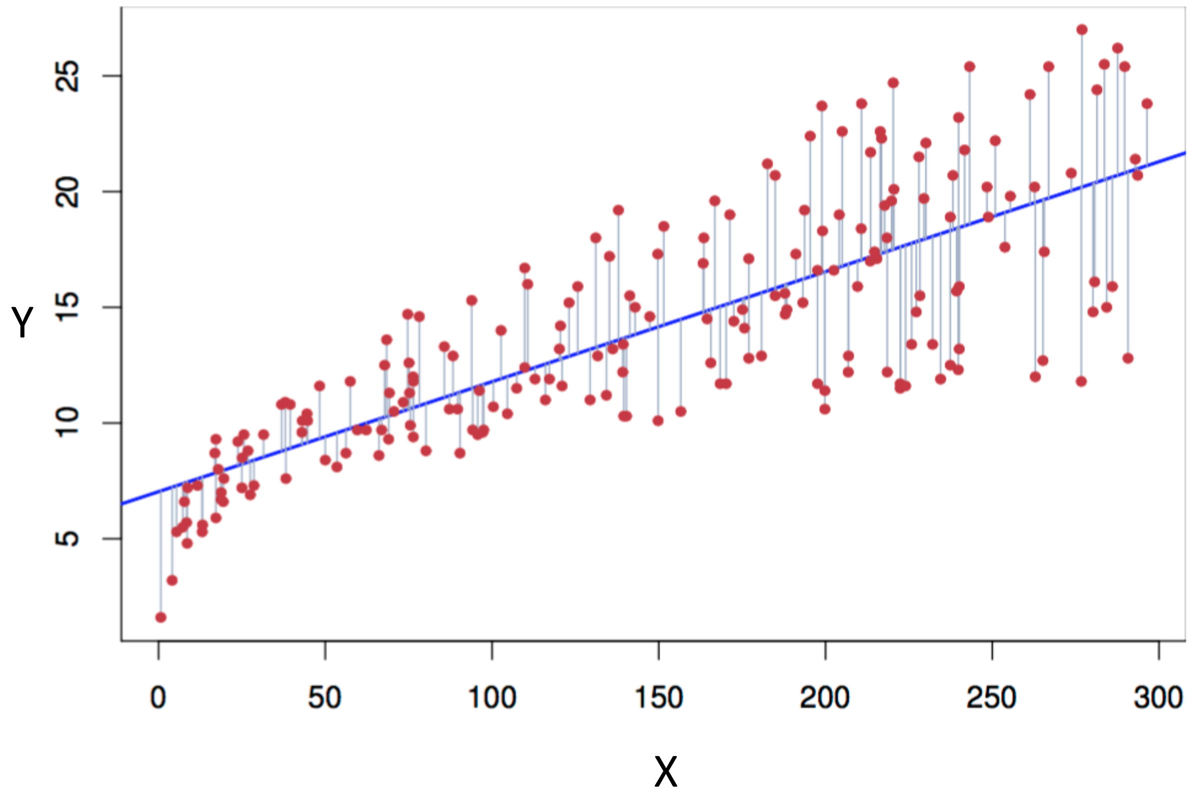
- known as the **residual sum of squares**.
- That starting from random w_0 and w_1, it changes them by setting their values as the corresponding **partial derivatives** of the equation above, **until convergence is reached**.

$$\frac{\partial E}{\partial \omega_0} = -2 \sum_{i=1}^{N} [y_i - \omega_1 x_i - \omega_0] = 0$$

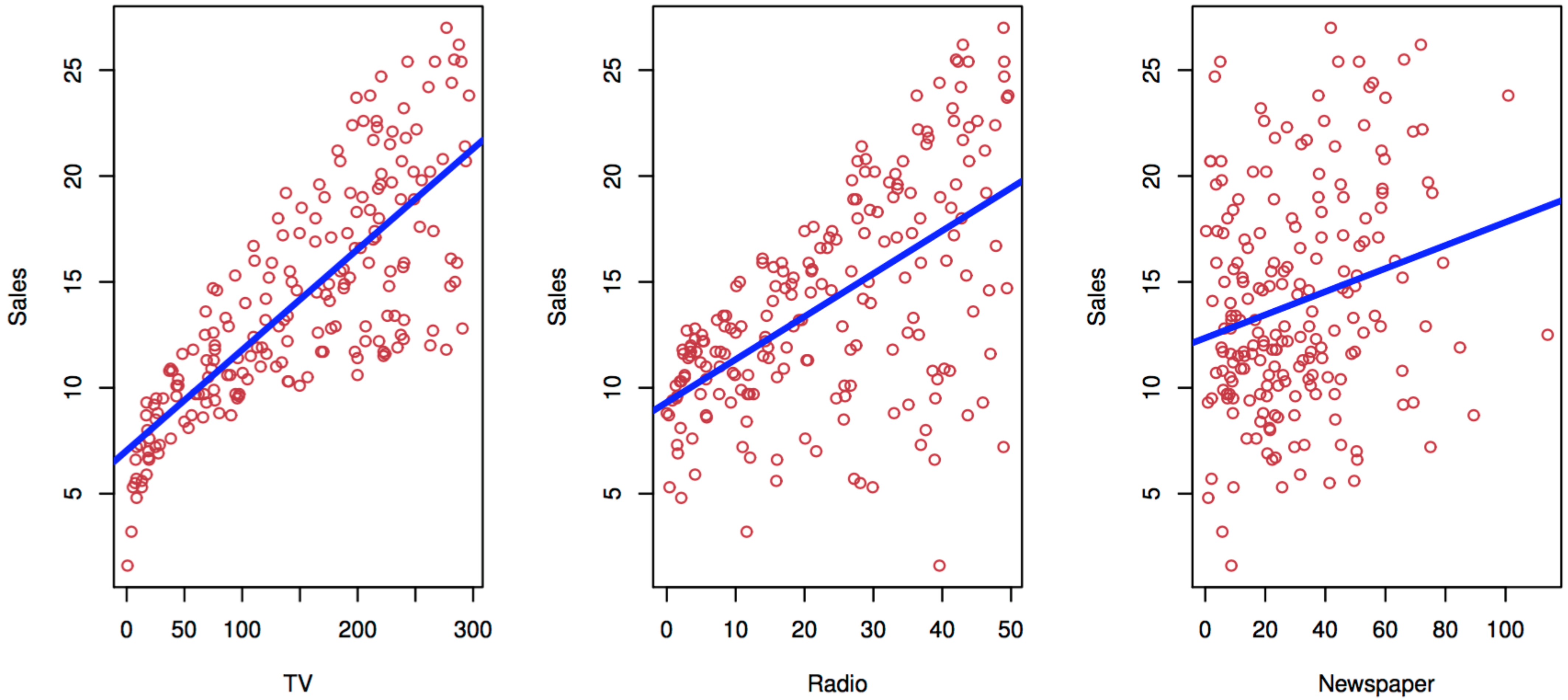
$$\frac{\partial E}{\partial \omega_1} = -2 \sum_{i=1}^{N} [y_i - \omega_1 x_i - \omega_0] x_i = 0$$

Least Square Method

- Blue line shows the least square fit. Lines from red points to the regression line illustrate the residuals.
- For any other choice of slope w_1 or intercept w_0 the SSE between that line and the observed data would be larger than the SSE of the blue line.



Examples



Sales
5
10
15
20
25
0
50
100
200
300
TV
Sales
5
10
15
20
25
0
10
20
30
40
50
Radio
Sales
5
10
15
20
25
0
20
40
60
80
100
Newspaper

Multiple Linear Regression

In case we have *m variables* $X=x_1, x_2, \dots, x_m$ the prediction model is

$$y = w_0 + \sum_{i=[1,\dots,m]} w_i x_i$$

if we extend X to $X=1, x_1, x_2, \dots, x_m$ *the prediction model* may be expressed as

$$y = \sum_{i=[0,\dots,m]} w_i x_i$$

The optimum parameter is defined as such that minimizes:

$$\sum_{j=[0,\dots,N]} (y_i - \sum_{i=[0,\dots,m]} w_i x_i)^2$$

Alternative Fitting Methods

- However, they can be fitted in other ways, such as by minimizing a penalized version of the least squares cost function as in **ridge regression** (L2-norm penalty) and **lasso** (L1-norm penalty).
- **Tikhonov** regularization, also known as *ridge regression*, is a method of regularization of ill-posed problems particularly useful to mitigate the multicollinearity, which commonly occurs in models with large numbers of parameters.
- **Lasso** (least absolute shrinkage and selection operator) performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

Alternative Fitting Methods

- However, they can be fitted in other ways, such as by minimizing a penalized version of the least squares cost function
  - **ridge regression** (L2-norm penalty)

penalty

$$\sum_{j=[0,\ldots,N]}(y_i \;-\; \sum_{i=[0,\ldots,m]} w_i x_i)^2 \;+ \lambda \sum_{i=[0,\ldots,m]} {w_i}^2$$

Constraint

$$\sum_{i=[0,\ldots,m]} {w_i}^2 \leq c$$

  - λ term regularizes the coefficients such that if the coefficients take large values the optimization function is penalized.
  - **lasso** (L1-norm penalty)

$$\sum_{j=[0,\ldots,N]}(y_i \;-\; \sum_{i=[0,\ldots,m]} w_i x_i)^2 \;+ \lambda \sum_{i=[0,\ldots,m]} |w_i|$$

Evaluating Regression

- **Coefficient of determination** R^2
  - is the proportion of the variance in the dependent variable that is predictable from the independent variable(s)

$$R^2 = \frac{SSM}{SST} = \frac{\sum_i [f(x_i) - \bar{y}]^2}{\sum_i [y_i - \bar{y}]^2} \qquad \bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

- **Mean Squared/Absolute Error** MSE/MAE
  - a risk metric corresponding to the expected value of the squared (quadratic)/absolute error or loss

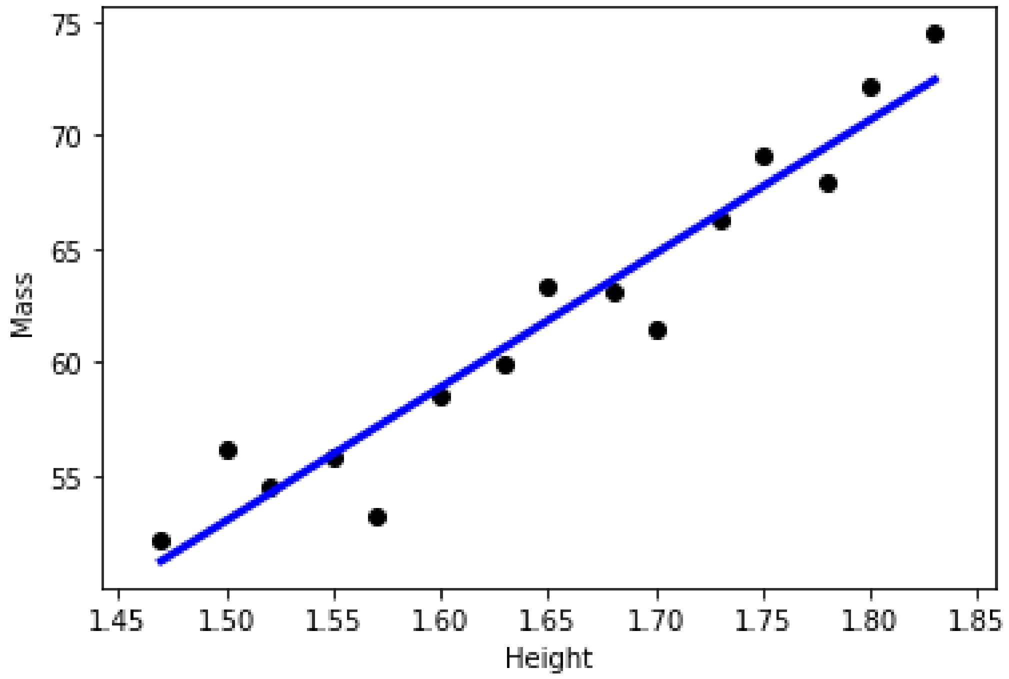
f(x)

$$\text{MSE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} (y_i - \hat{y}_i)^2 \quad \text{MAE}(y, \hat{y}) = \frac{1}{n_{\text{samples}}} \sum_{i=0}^{n_{\text{samples}}-1} |y_i - \hat{y}_i|$$

Example

- Height (m): 1.47, 1.50, 1.52, 1.55, 1.57, 1.60, 1.63, 1.65, 1.68, 1.70, 1.73, 1.75, 1.78, 1.80, 1.83
- Mass (kg): 52.21, 56.12, 54.48, 55.84, 53.20, 58.57, 59.93, 63.29, 63.11, 61.47, 66.28, 69.10, 67.92, 72.19, 74.46

- Intercept: -35.30454824113264
- Coefficient: 58.87472632
- R^2: 0.93
- MSE: 3.40
- MAE: 1.43



References

- Regression. Appendix D. Introduction to Data Mining.