Introduction to Linear Regression

金融投资学

Instructor: Haoran LEI

Hunan University

Review

- Supervised Learning:
  - Nearest neighbor, Linear regression
- Tradeoffs:

1. Prediction **accuracy** versus **interpretability**.
2. **Good fit** versus **over-fit** or **under-fit**.
3. **Parsimony** versus **black-box**

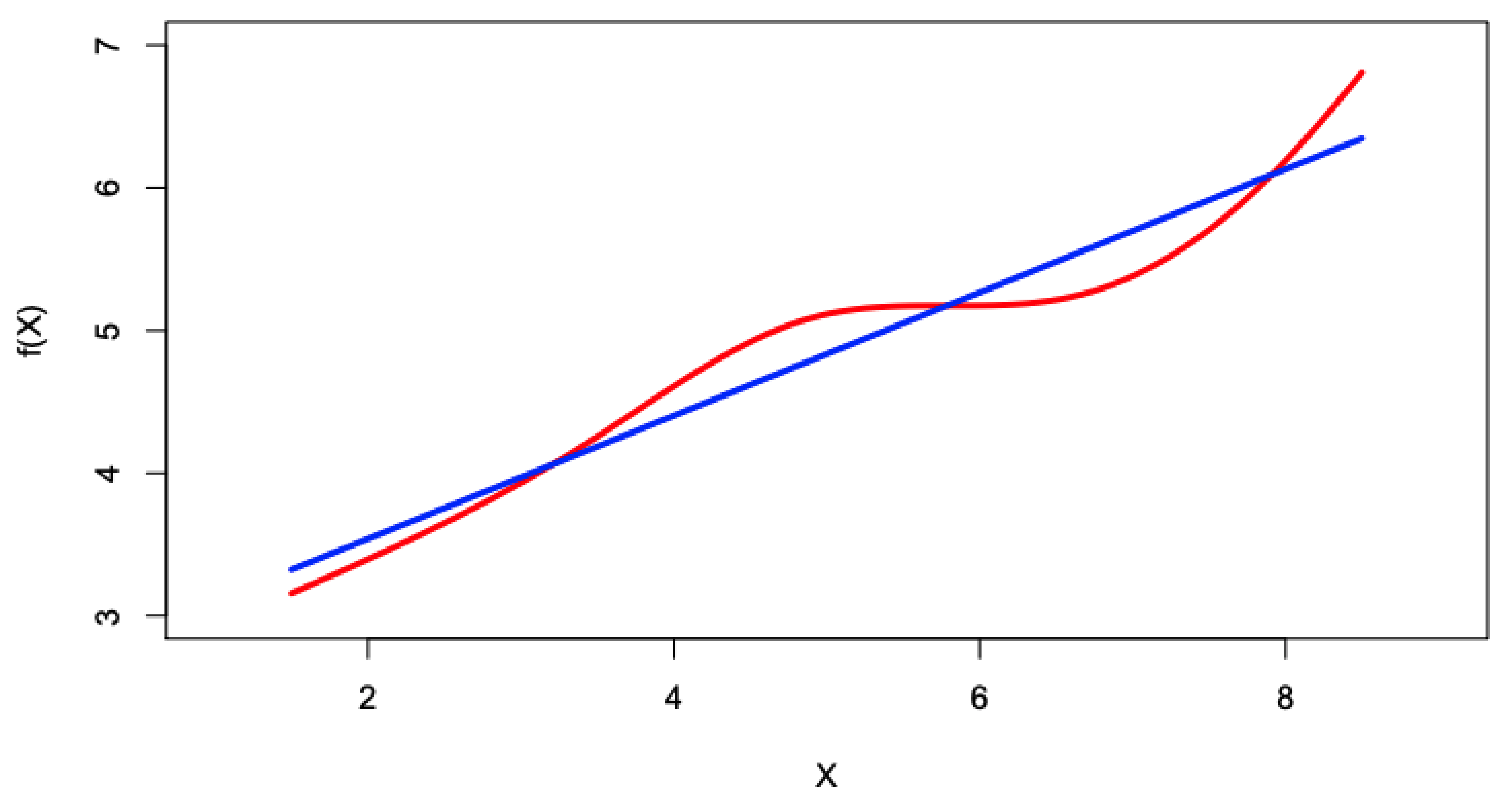
Review: Bias-variance tradeoff

- In (most) models, we can **reduce the variance of the parameter estimated across samples** by **increasing the bias** in the estimated parameters.
- Homework: Explain the three plots.

Linear regression

- Linear regression is (perhaps) the simplest approach to supervised learning.
- It assumes that the dependence of Y on $X_1, \ldots, X_p$ are linear.
- True regression functions are (almost) never linear.

Although it may seem overly simplistic, linear regression is extremely useful both conceptually and practically.



Linear regression with a single predictor X

- Model: $Y = \beta_0 + \beta_1 X + \epsilon$
- β_0 and β_1 are two unknown constants that represent the *intercept* and *slope*.
- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we make the predictions:

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

- where $\hat{y}$ indicates a prediction of Y given $X = x$. The hat symbol ^ denotes an **estimated value**.

Least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y_i.
- The i-th residual: $e_i = y_i - \hat{y}_i$.
- Define the **residual sum of squares (RSS)**:

$$
\begin{aligned}
RSS &= (e_1)^2 + \cdots + (e_n)^2 \\
&= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2
\end{aligned}
$$

Least squares

- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y_i.
- The i-th residual: $e_i = y_i - \hat{y}_i$.
- Define the **residual sum of squares (RSS)**:

$$\begin{aligned} \text{RSS} &= (e_1)^2 + \cdots + (e_n)^2 \\ &= (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2 \end{aligned}$$

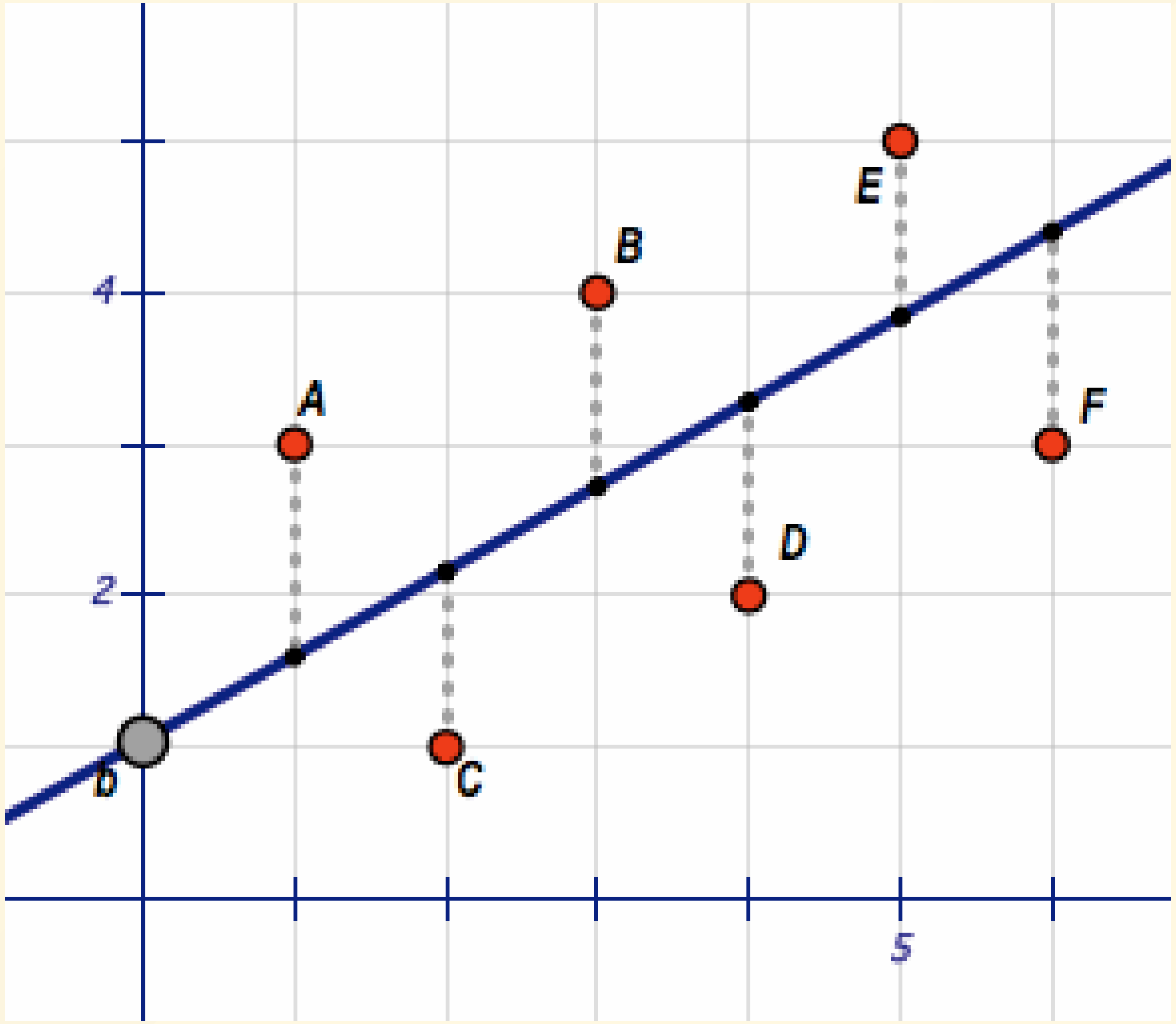
Least squares: choose $\hat{\beta}_0$, $\hat{\beta}_1$ to minimize RSS.
(Or minimizing MSE_{Tr} as we've seen in previous lecture slides)

Least squares

The estimated values that minimize RSS are:

$$\begin{cases} \hat{\beta}_1 = \dfrac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sum_i (x_i - \bar{x})^2} \\ \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \end{cases}$$

where $\bar{x} = \sum_i x_i / n$ and $\bar{y} = \sum_i y_i / n$ are the sample means.



Animation of LS regression line

Example (advertising data)

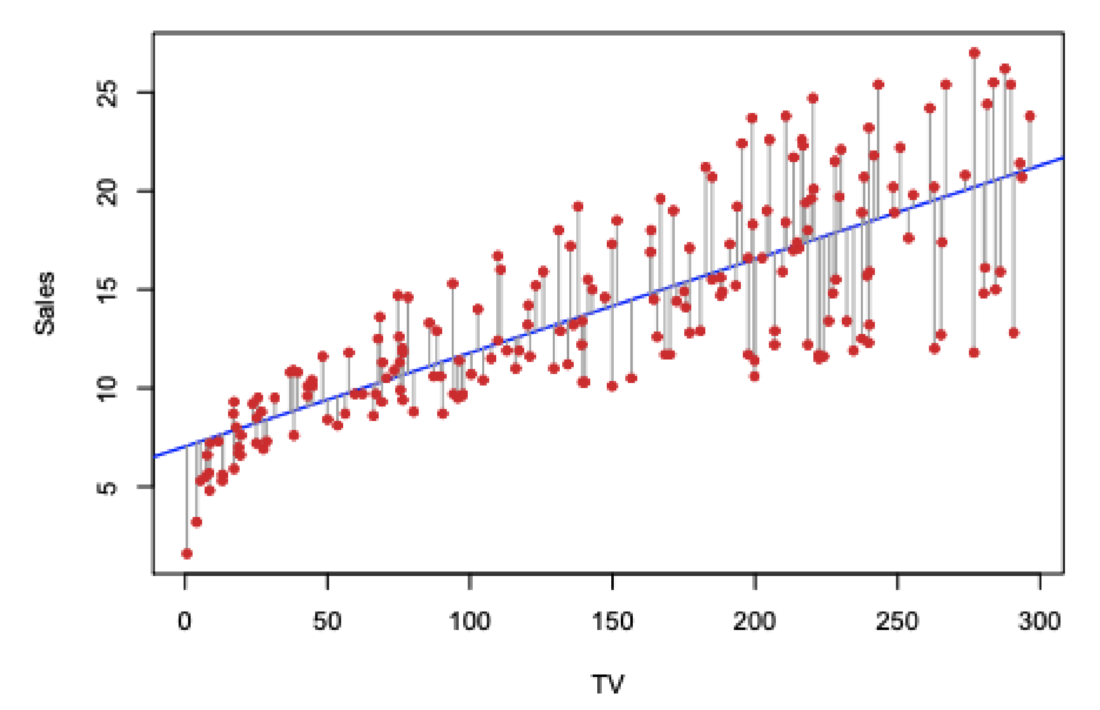


Fig: The least squares fit for the regression of sales onto TV.

- A linear fit captures the essence of the relationship, but it seems somewhat deficient in the left of the plot.

Assessing the Accuracy of the LS Estimates

- The **standard error (SE)** of an estimator reflects how it varies under repeated sampling:

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

- Standard error is not the *variance* of the LS estimator. Instead, it measures how **accurate** the LS estimator is. SE depends on:
  1. the variance of noise: σ^2
  2. how "spread" our datas are: $\sum_{i=1}^{n}(x_i - \bar{x})^2$

Assessing the Accuracy of the LS Estimates

- The **standard error (SE)** of an estimator reflects how it varies under repeated sampling:

$$SE(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, SE(\hat{\beta}_0)^2 = \sigma^2[\frac{1}{n} + \frac{\bar{x}^2}{\sum_i(x_i - \bar{x})^2}]$$

- Note: When σ^2 (variance of ϵ) is unknown, use

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_i e_i^2$$

Confidence interval

- A **95% confidence interval** is defined as a range of values such that *"with 95% probability, the range will contain the true unknown value of the parameter."*
- It has the form:

$$[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1), \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)]$$

A popular way of describing confidence intervals:

- "I am 95% confident that the interval contains the true value."

Hypothesis testing

- **Standard errors** can also be used to perform *hypothesis testing*.
- The most common *hypothesis test* involves testing the null hypothesis H_0 vs the alternative hypothesis H_A:

$$H_0 : \text{ There is no relationship between } X \text{ and } Y$$

$$H_A : \text{ There is some relationship between } X \text{ and } Y$$

Hypothesis testing

- **Standard errors** can also be used to perform *hypothesis testing*.
- The most common *hypothesis test* involves testing the null hypothesis H_0 vs the alternative hypothesis H_A:

$$H_0 : \beta_1 = 0, \quad H_A : \beta_1 \neq 0$$

- To test the null hypothesis, we compute a *t-statistic* given by:

$$t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$$

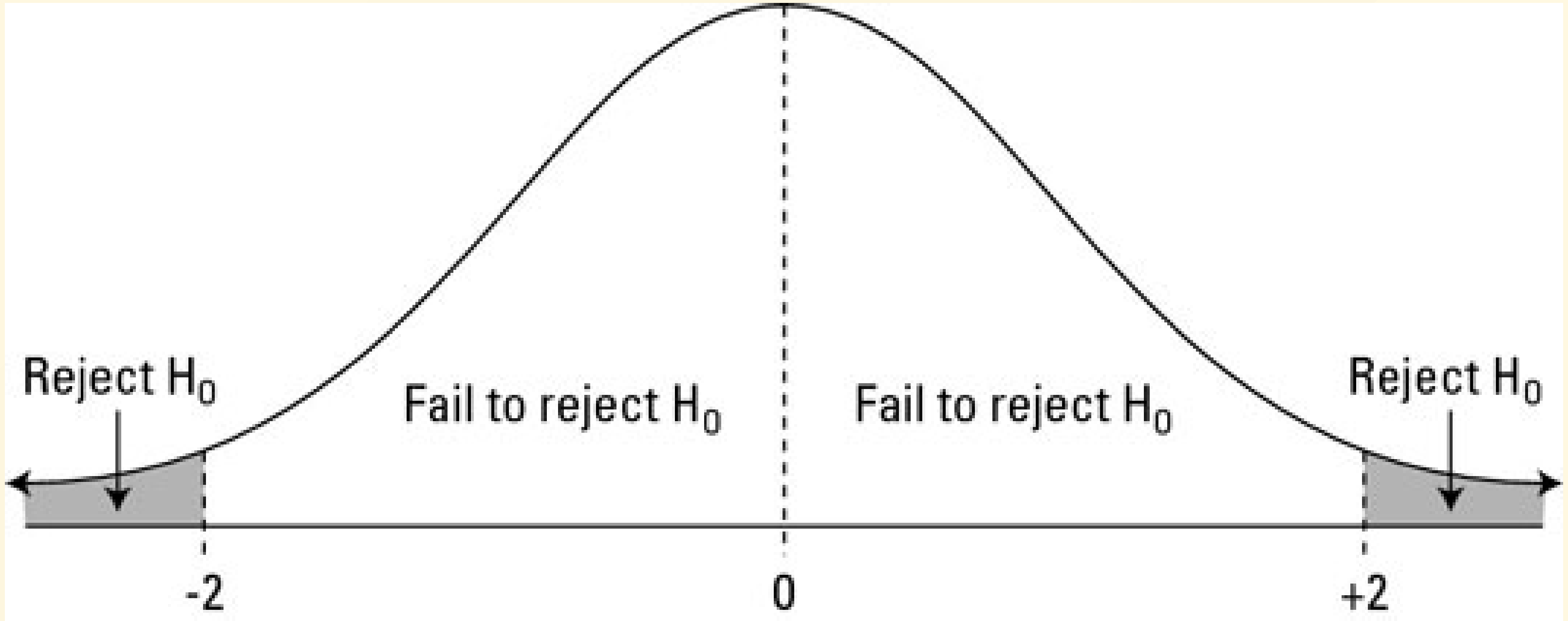
Hypothesis testing

- Assuming $\beta_1 = 0$ (ie, H_0 holds), then $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ will follow the **t-distribution with $n - 2$ degrees of freedom**.

Hypothesis testing

- Assuming $\beta_1 = 0$ (ie, H_0 holds), then $t = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$ will follow the **t-distribution with $n - 2$ degrees of freedom**.
- Using statistical software, it is easy to compute the probability of observing *any value equal or larger than* $|t|$.
  - We call this probability the *p-value*.

- In practice, usually we say that the effects of X is significant (rejecting H_0) when the p-value is less than 0.05.



- You can see from the figure that *p-values* and *confidence intervals* are just two sides of the same coin.

Results for the advertising data

	Coefficient	Std. Error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.1%
TV	0.0475	0.0027	17.67	< 0.1%

Assessing the Overall Accuracy of the Model

- We compute the **Residual Standard Error** (**RSE**):

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}}$$

- RSE is to RSS what standard error is to variance.

Assessing the Overall Accuracy of the Model

- *R-squared* is the fraction of variance explained:

$$R^2 = \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS}$$

- where TSS $= \sum_i e_i^2 = \sum_i (y_i - \bar{y})^2$ is the total sum of squares.

In the simple linear regression setting with one predictor, $R^2 = r^2$ where r is the (linear) correlation between X and Y.

Advertising data results

Quantity	Value
Residual Standard Error	3.26
R^2	0.612

Next: Multiple Linear Regression

We have focused on the simple linear regression model with one predictor.

Now let's move on to Multiple Linear Regression; aka, linear regression with multiple predictors!