CPSC 8420 Advanced Machine Learning
Week 3: Linear Regression

Dr. Kai Liu

September 1, 2020

Learning Outcomes

Our goal for today's lecture is to understand

- the use of linear regression with one predictor as a simple supervised learning model,
- how the linear regression model is estimated and used for prediction,
- using linear regression with multiple predictors as a simple supervised learning model,
- how to derive optimized regression,
- how we select the "best" predictors from a set of p predictors.

Linear Regression

Linear Regression

- Linear regression is a simple supervised learning method that assumes a linear relationship between the predictor Y and the features $X_1, X_2, \ldots X_p$.
- Linear regression is a useful tool for predicting a quantitative response.
- Although it is a very simple model, linear regression is extremely useful both conceptually and practically.

Simple Linear Regression

- We assume the following model:

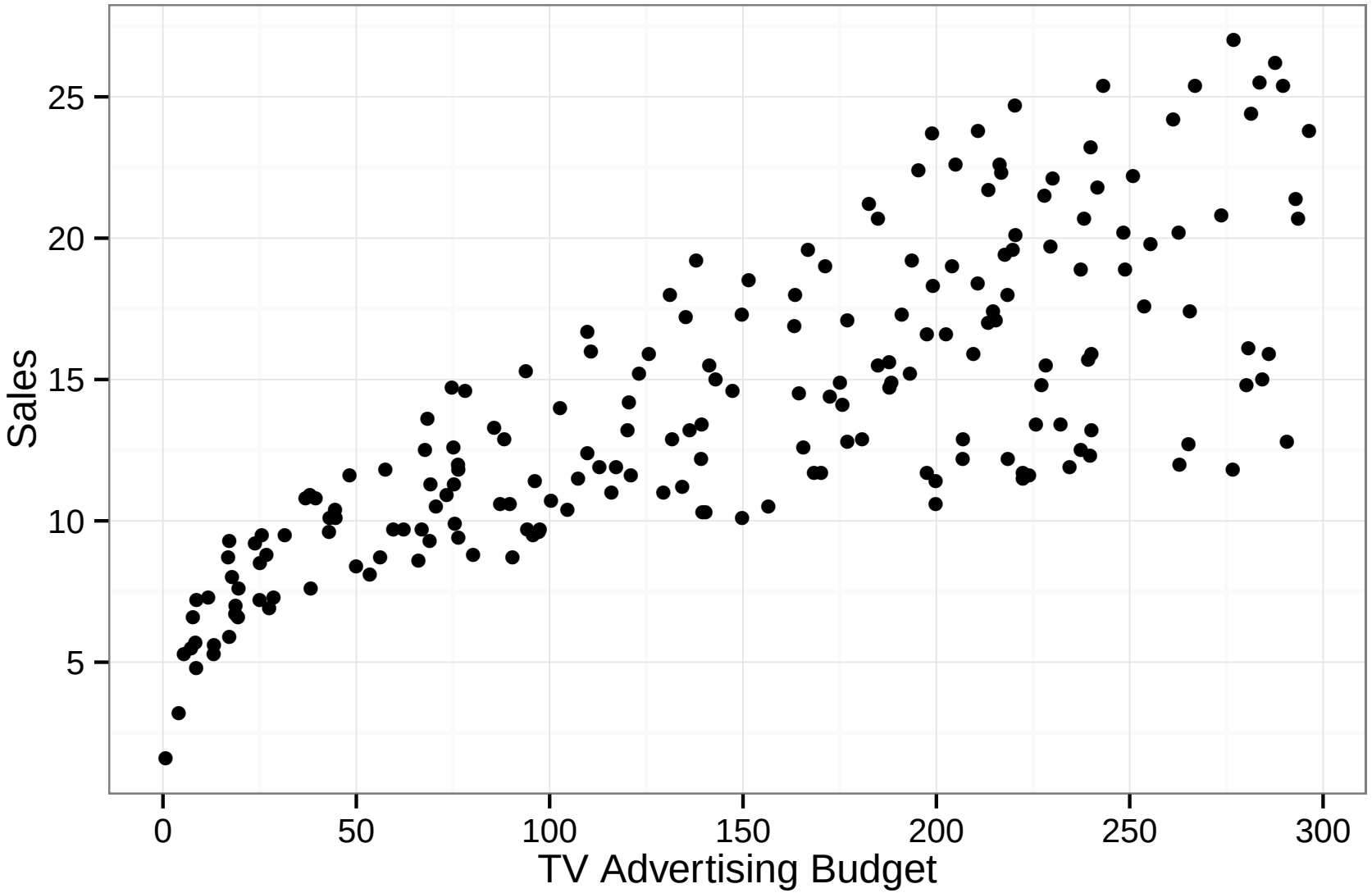
$$Y = f(X) + \epsilon \tag{1}$$
$$= \beta_0 + \beta_1 X + \epsilon \tag{2}$$

- β_0 and β_1 are two unknown parameters (or coefficients) that represent the intercept and slope of the regression line
- ϵ is the irreducible error
- Given some estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we predict outcomes of Y as

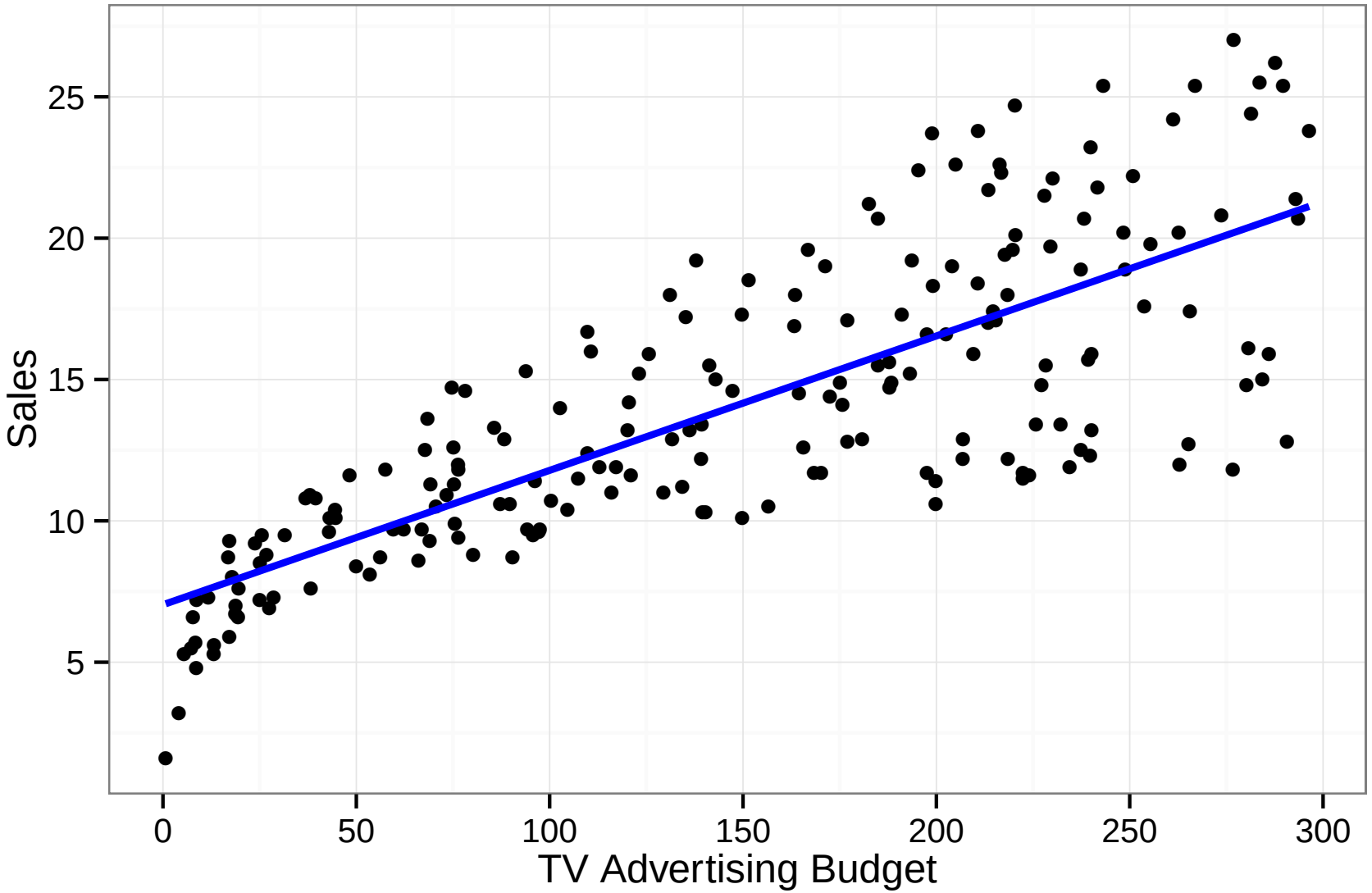
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{3}$$

where $\hat{y}$ indicates a prediction of Y on the basis of $X = x$. (The hat symbol denotes an estimated value.)

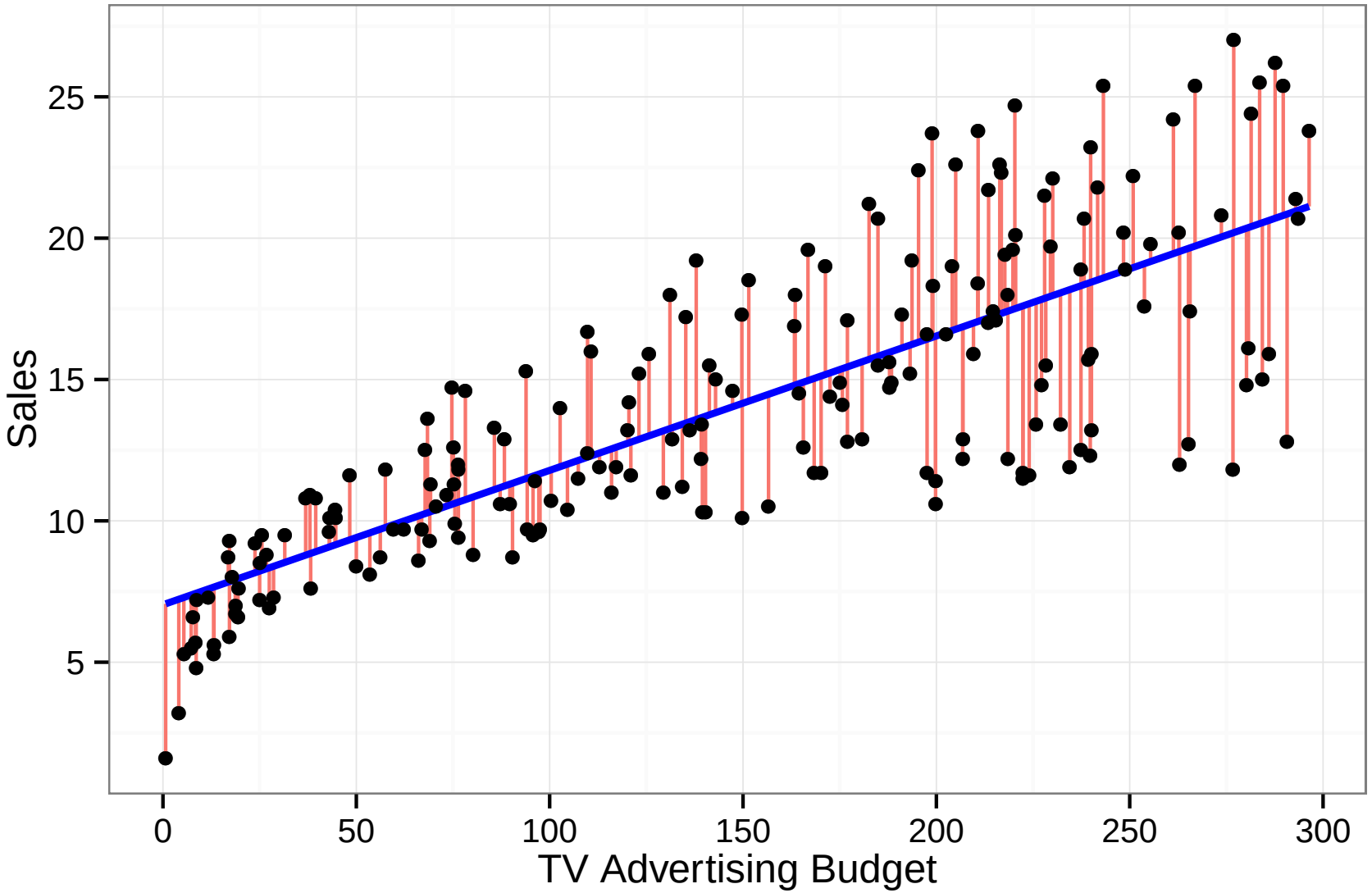
Example: TV Advertising Budget and Sales



Example: TV Advertising Budget and Sales



Example: TV Advertising Budget and Sales



Estimation

Prediction for a single observation:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, \qquad \text{where} \tag{4}$$

$$\epsilon_i = y_i - \hat{y}_i \tag{5}$$

Residual sum of squares (RSS):

$$RSS = \epsilon_1^2 + \epsilon_2^2 + \cdots + \epsilon_n^2 \tag{6}$$

$$= \sum_{i=1}^{n} \epsilon_i^2 = \sum_{i=1}^{n} (y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_i))^2 \tag{7}$$

Estimation

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. The minimizing values can be shown to be:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{8}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \tag{9}$$

where $\bar{y}$ and $\bar{x}$ are the sample means.

Interpretation

Results: $\hat{\beta}_0 = 7.03$, $\hat{\beta}_1 = 0.0475$

Interpretation:

- An increase of TV advertising by \$1,000 (=one unit on the measurement scale of X) is associated with an average increase in sales by 0.0475 units on the measurement scale of Y, which is in thousands of units. Hence, an increase by 0.0475 corresponds to approximately 47.5 additional sold units.

- General interpretation:
  "A change in X by one unit is associated with an average change in Y by $\hat{\beta}_1$ units."

Prediction

Prediction: How many items will we sell if we spend $150,000 on TV advertising?

Prediction

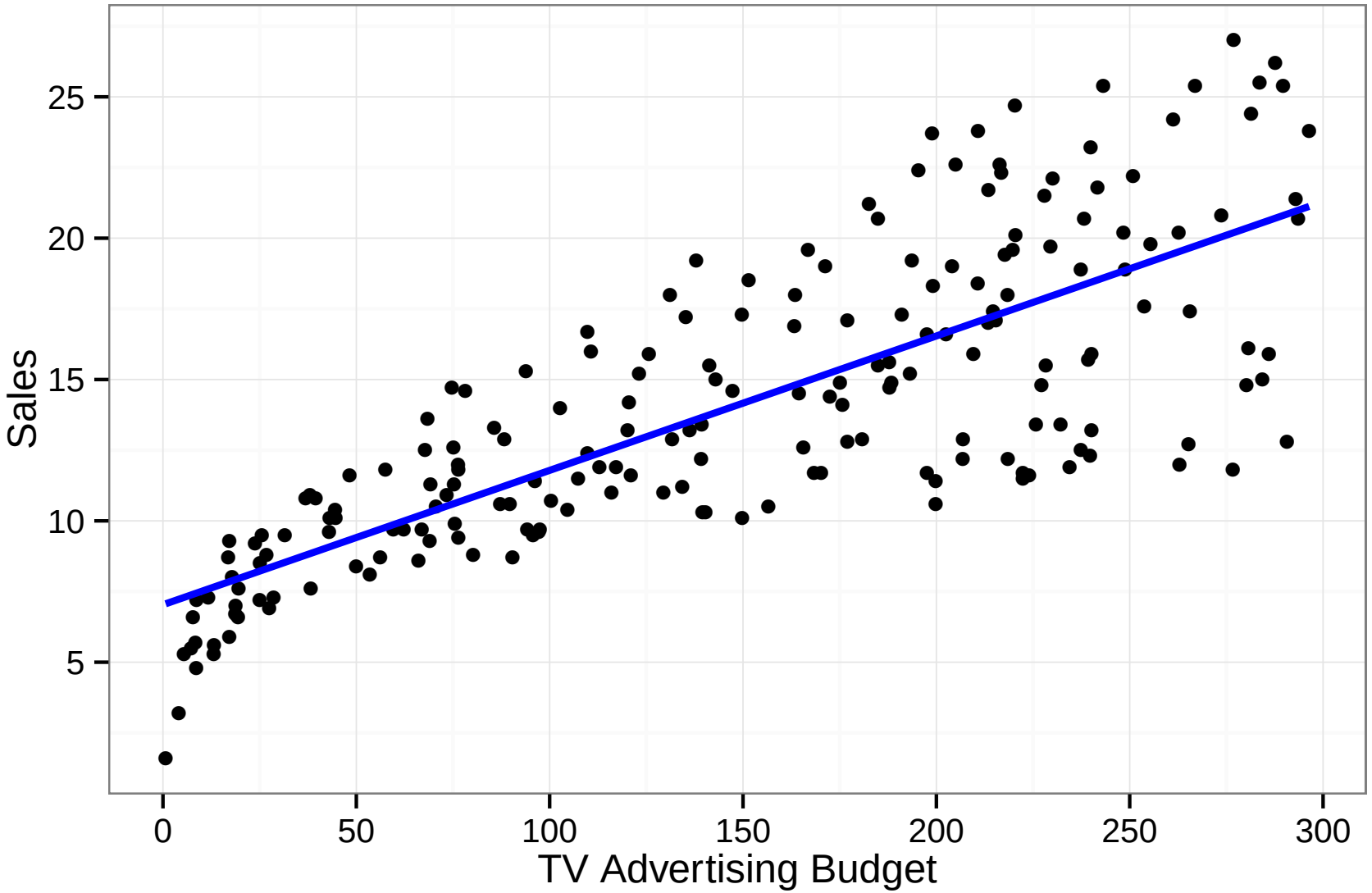
Prediction: How many items will we sell if we spend \$150,000 on TV advertising?

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{10}$$

$$= 7.03 + 0.0475 * 150 \approx 14 \tag{11}$$

Hence, if we spend \$150,000 on TV advertising, we predict to sell 14,000 units.

Example: TV Advertising Budget and Sales



Multiple Linear Regression

Multiple Linear Regression

- Instead of one predictor and one slope coefficient, we have multiple predictors and coefficients associated with them:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{12}$$

Multiple Linear Regression

- Instead of one predictor and one slope coefficient, we have multiple predictors and coefficients associated with them:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{12}$$

- We interpret β_j as the average effect on Y for a one unit increase in X_j, holding all other predictors constant.

Multiple Linear Regression

- Instead of one predictor and one slope coefficient, we have multiple predictors and coefficients associated with them:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{12}$$

- We interpret β_j as the average effect on Y for a one unit increase in X_j, holding all other predictors constant.

- In the advertising example, the model becomes:

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon \tag{13}$$

Multiple Linear Regression

- Instead of one predictor and one slope coefficient, we have multiple predictors and coefficients associated with them:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \tag{12}$$

- We interpret β_j as the average effect on Y for a one unit increase in X_j, holding all other predictors constant.

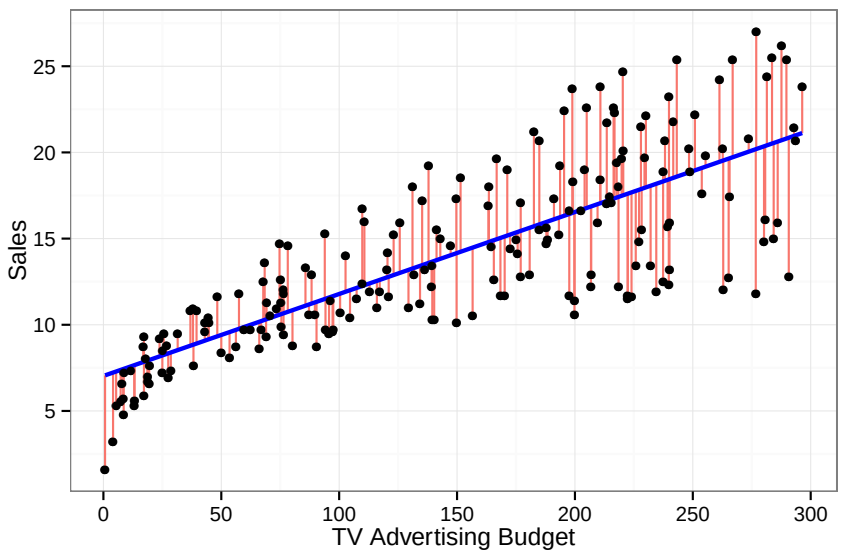
- In the advertising example, the model becomes:

$$sales = \beta_0 + \beta_1 TV + \beta_2 radio + \beta_3 newspaper + \epsilon \tag{13}$$

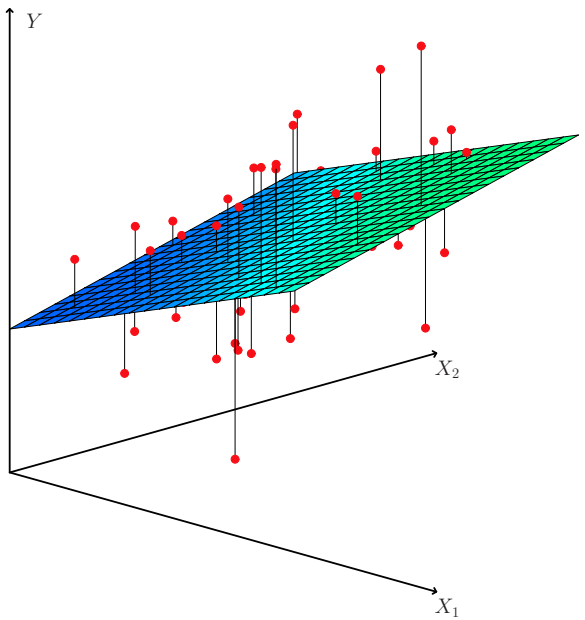
- Interpretation: "*A change in X by one unit is associated with an average change of Y by $\hat{\beta}_1$ units, holding all other predictors constant.*"

Multiple Linear Regression

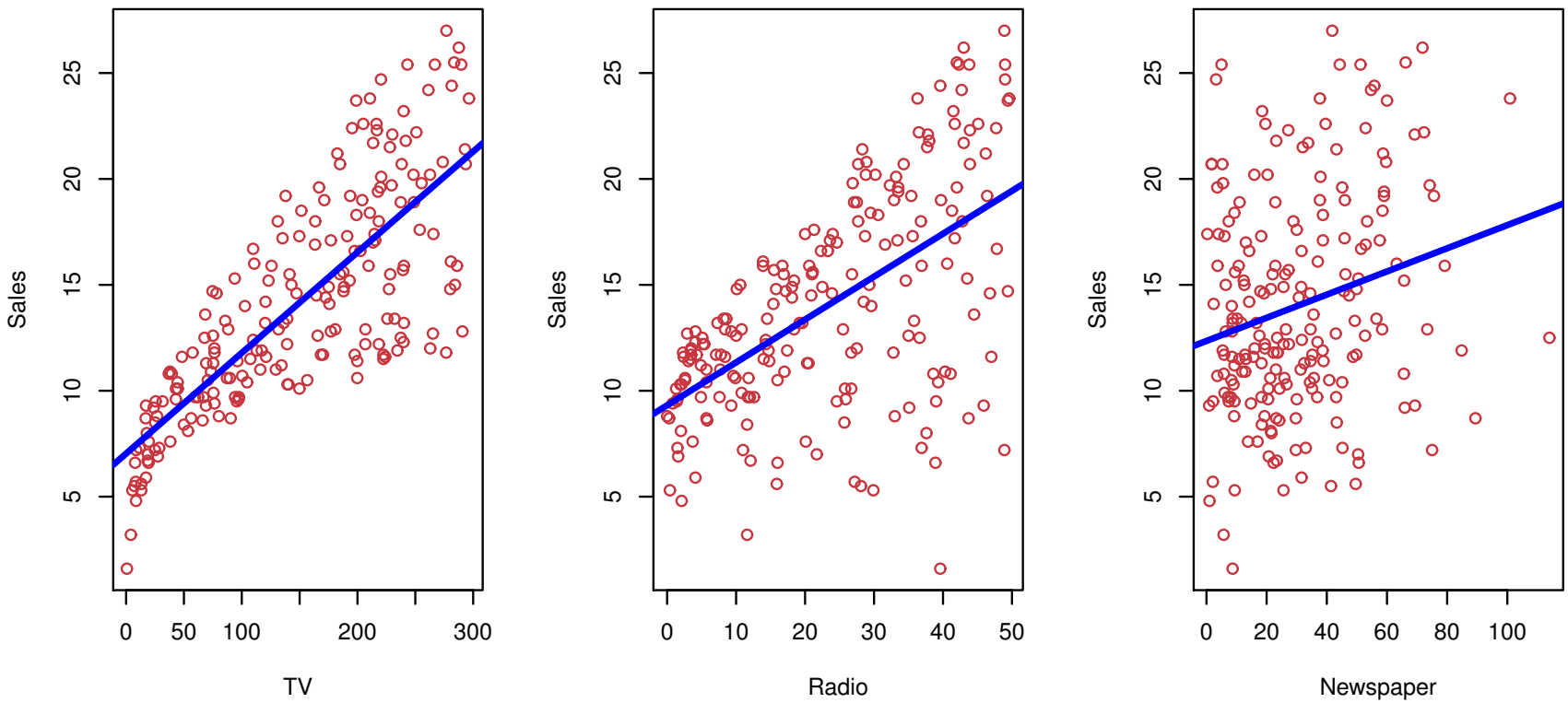
One predictor



Two predictors



Why Multiple Linear Regression?



Multiple Linear Regression

- The ideal scenario is when the predictors are uncorrelated – a balanced design:
  - Each coefficient can be estimated and tested separately.
  - Interpretations such as "a unit change in X_j is associated with a β_j change in Y, while all other predictors are held constant", are possible.
- Correlations amongst predictors cause problems:
  - The variance of all coefficients tends to increase, sometimes dramatically
  - Interpretations become hazardous – when X_j changes, everything else changes.
- Claims of causality should be avoided for observational data.

How to get the optimized coefficients?

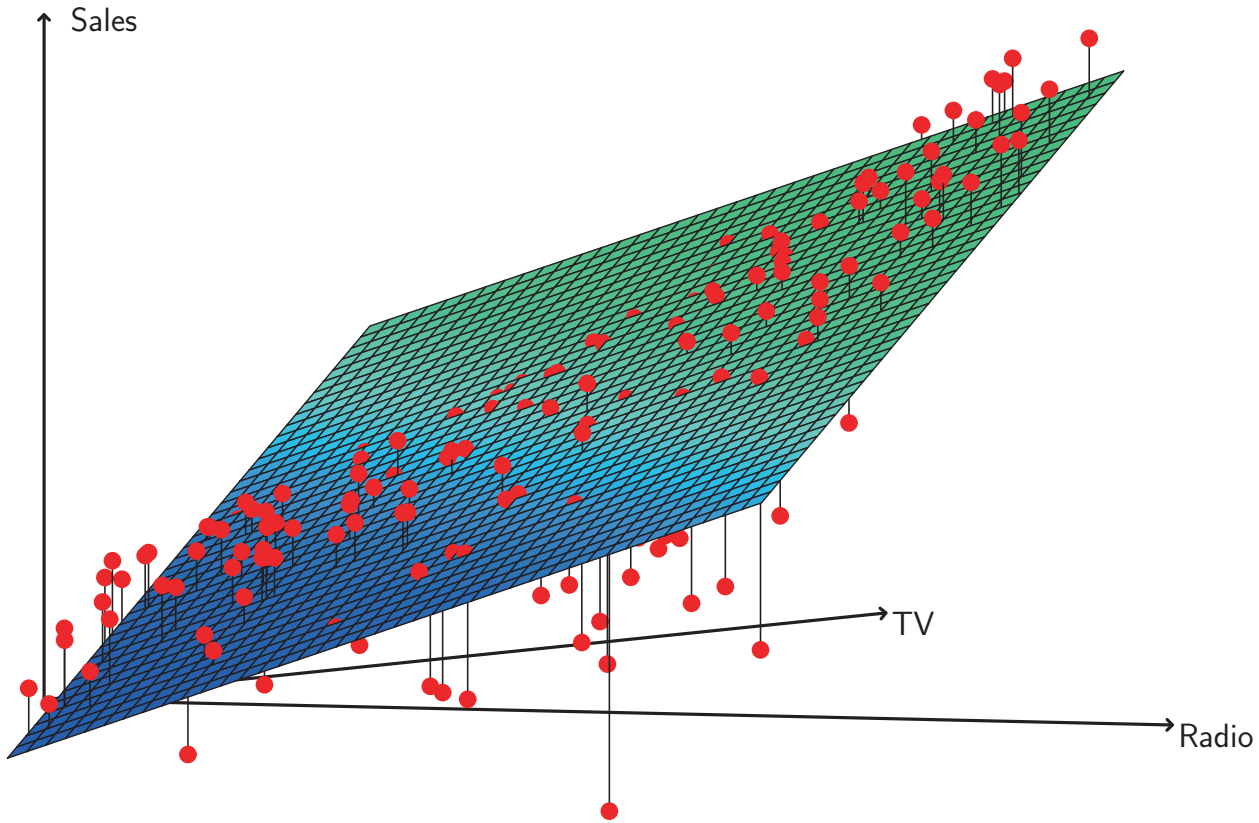


FIGURE 3.5. *For the* `Advertising` *data, a linear regression fit to* `sales` *using* `TV` *and* `radio` *as predictors. From the pattern of the residuals, we can see that there is a pronounced non-linear relationship in the data. The positive residuals (those visible above the surface), tend to lie along the 45-degree line, where TV and Radio budgets are split evenly. The negative residuals (most not visible), tend to lie away from this line, where budgets are more lopsided.*

Formulation

Our target is:

$$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \tag{14}$$

Assume we can approximate by multiple linear regression by:

$$\begin{aligned} y_1 &\approx \beta_0 + \beta_1 X_{11} + \beta_2 X_{12} + \cdots + \beta_p X_{1p} := \hat{y}_1 \\ y_2 &\approx \beta_0 + \beta_1 X_{21} + \beta_2 X_{22} + \cdots + \beta_p X_{2p} := \hat{y}_2 \\ &\cdots \\ y_n &\approx \beta_0 + \beta_1 X_{n1} + \beta_2 X_{n2} + \cdots + \beta_p X_{np} := \hat{y}_n \end{aligned} \tag{15}$$

where $\boldsymbol{X}$ and $\boldsymbol{y}$ is known, and $\boldsymbol{\beta}$ is the coefficients to optimize.

Formulation

Now let's define:

$$\boldsymbol{A} = \begin{bmatrix} 1 & X_{11} & X_{12} & \dots & X_{1p} \\ 1 & X_{21} & X_{22} & \dots & X_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{n1} & X_{n2} & \dots & X_{np} \end{bmatrix} \tag{16}$$

$$\boldsymbol{y} = [y_1, y_2, \dots, y_n]^T$$

$$\boldsymbol{\beta} = [\beta_0, \beta_1, \beta_2, \dots, \beta_p]^T$$

then we can formulate our objective function as:

$$\boldsymbol{J} = \min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}\|_2^2 \tag{17}$$

Optimization

Given $\boldsymbol{J}$ is convex *w.r.t.* $\boldsymbol{\beta}$, the function achieves the minimum iff:

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{\beta}} = 0$$

A few lemmas are in order:

- $\|\boldsymbol{z}\|^2 = trace(\boldsymbol{z}^T\boldsymbol{z}) = \langle \boldsymbol{z}, \boldsymbol{z} \rangle$
- $\frac{\partial trace(\boldsymbol{a}^T\boldsymbol{b})}{\partial \boldsymbol{a}} = \boldsymbol{b}$
- $trace(\boldsymbol{a}\boldsymbol{b}^T) = trace(\boldsymbol{b}\boldsymbol{a}^T)$

then we have:

$$\begin{aligned} \boldsymbol{J} &= \|\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta}\|_2^2 \\ &= trace[(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{A}\boldsymbol{\beta})] \\ &= trace(\boldsymbol{y}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\beta} - 2\boldsymbol{\beta}^T\boldsymbol{A}^T\boldsymbol{y}) \end{aligned} \tag{18}$$

Optimized Solution

$$\frac{\partial \boldsymbol{J}}{\partial \boldsymbol{\beta}} = 2(\boldsymbol{A}^T\boldsymbol{A}\boldsymbol{\beta} - \boldsymbol{A}^T\boldsymbol{y}) = 0$$

thus we have:

$$\boldsymbol{\beta}^* = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{y}$$

Now let's do some complexity analysis:

$$\begin{aligned} \mathcal{O}(n^2p + p^3 + p^2n + pn) &\implies \mathcal{O}(n^2p), when\ n > p \\ &\implies \mathcal{O}(p^3), when\ n < p \end{aligned} \tag{19}$$

we come to certain problems:

- heavy complexity load, especially when n or p is large.
- there could be 0 eigenvalue for $\boldsymbol{A}^T\boldsymbol{A}$, which may make this method fail
- if $\boldsymbol{A}^T\boldsymbol{A}$ is *ill-conditioned*, the algorithm would turn to be unstable

Projection

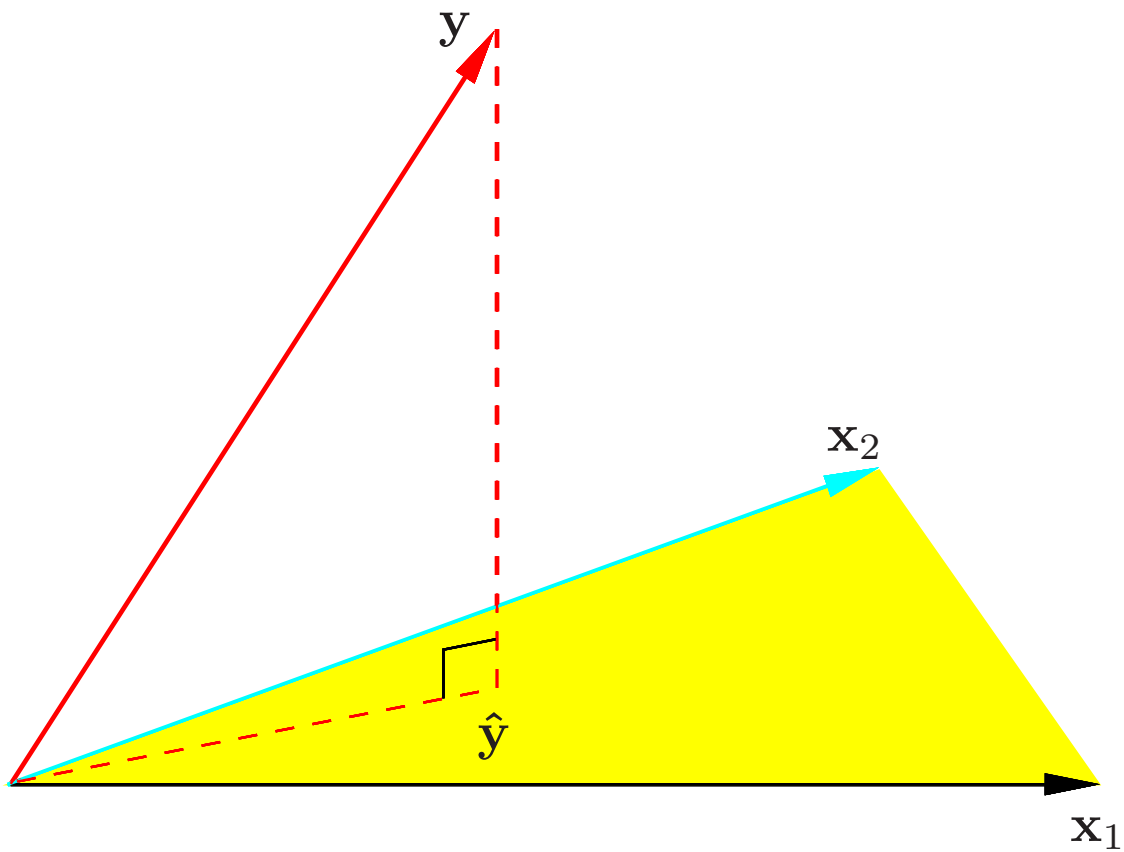
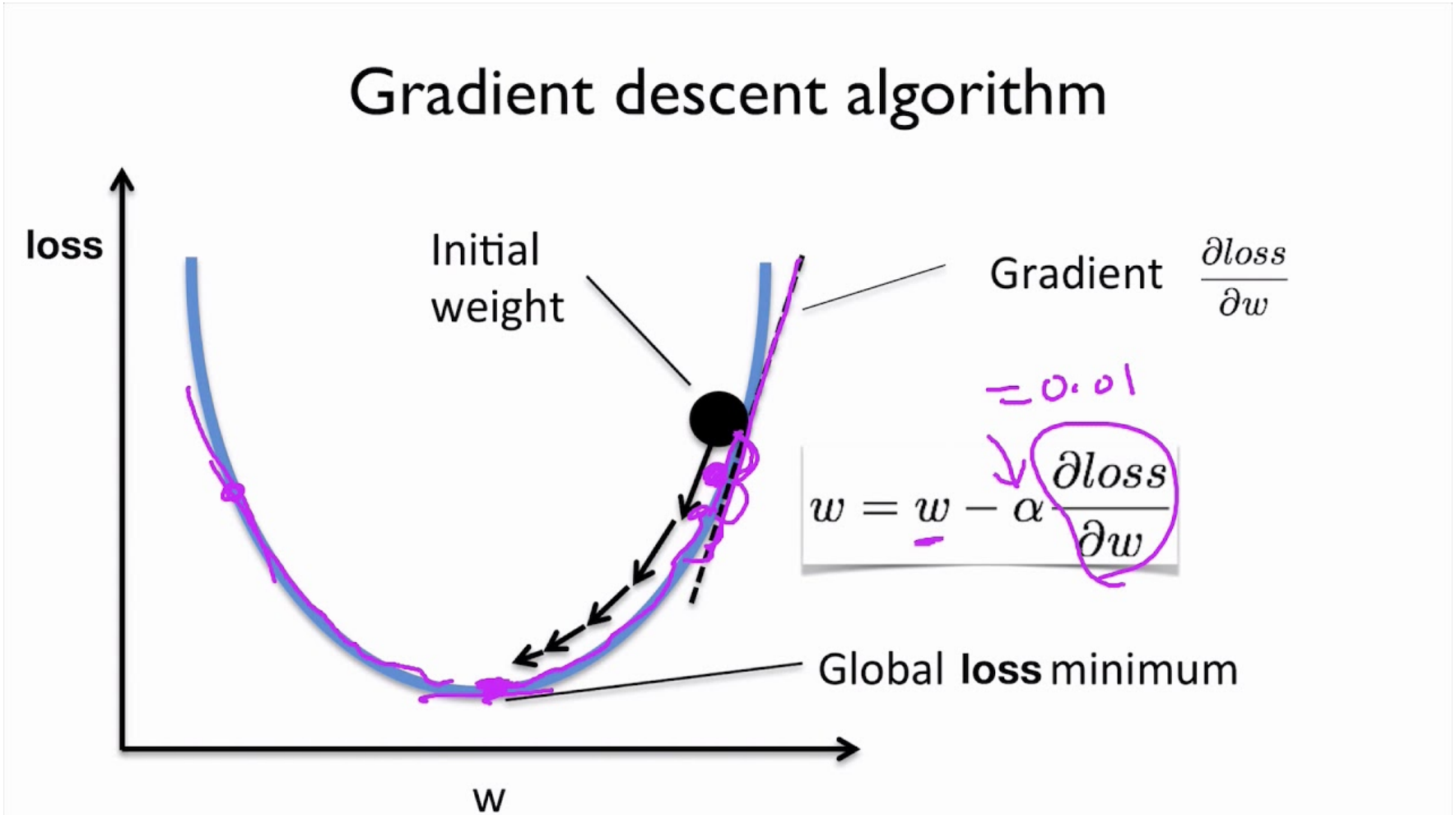


FIGURE 3.2. *The N-dimensional geometry of least squares regression with two predictors. The outcome vector $\mathbf{y}$ is orthogonally projected onto the hyperplane spanned by the input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The projection $\hat{\mathbf{y}}$ represents the vector of the least squares predictions*

Gradient Descent

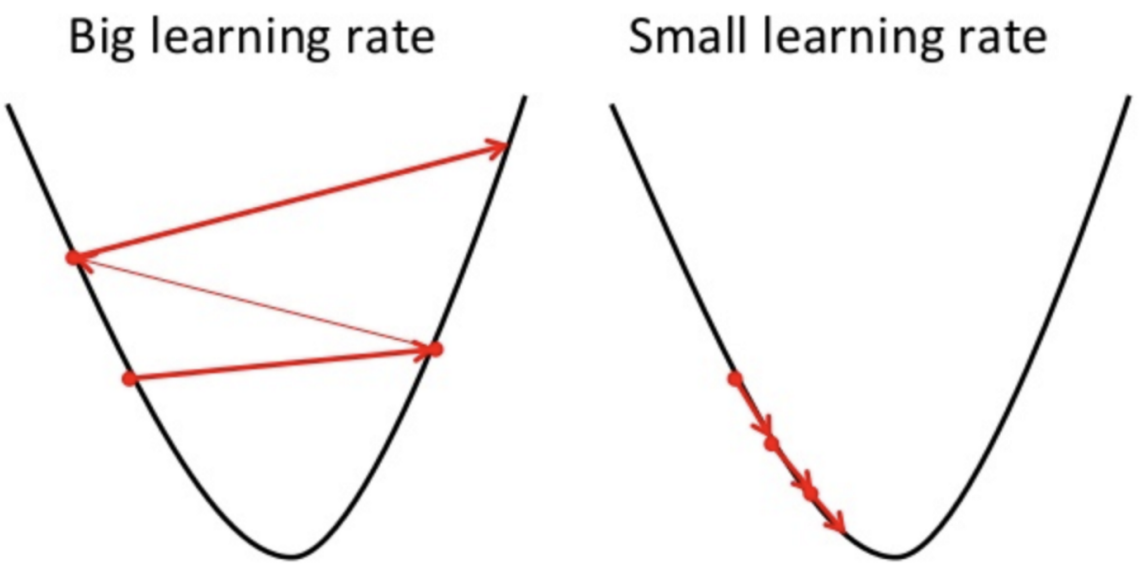
A typical gradient descent algorithm works as following:



GD to solve Least Square

Algorithm 1 GD for Multiple Linear Regression

1: *Initialize* $\boldsymbol{\beta}_0$ *and learning rate* $\alpha = .001$
2: **for** $i = 1$ to K **do**
3: $\quad \boldsymbol{\beta}_i = \boldsymbol{\beta}_{i-1} - \alpha \left(\boldsymbol{A}^T \boldsymbol{A} \boldsymbol{\beta}_{i-1} - \boldsymbol{A}^T \boldsymbol{y} \right)$
4: **end for**



Subset Selecting

Subset Selecting

- The most direct approach is called all subsets or best subsets regression: we compute the least squares fit for all possible subsets and then choose between them based on some criterion that balances training error with model size.

Subset Selecting

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

However, there are 2^p possible models. For example, when $p = 40$ there are over a billion models!

Forward Selection

- Begin with the null model – a model that contains an intercept but no predictors.
- Fit p simple linear regressions and add to the null model the predictor that results in the lowest RSS.
- Add to that model the predictor that results in the lowest RSS amongst all two-predictor models.
- Continue until some stopping rule is satisfied, for example when all remaining variables have a p-value above some threshold.

Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

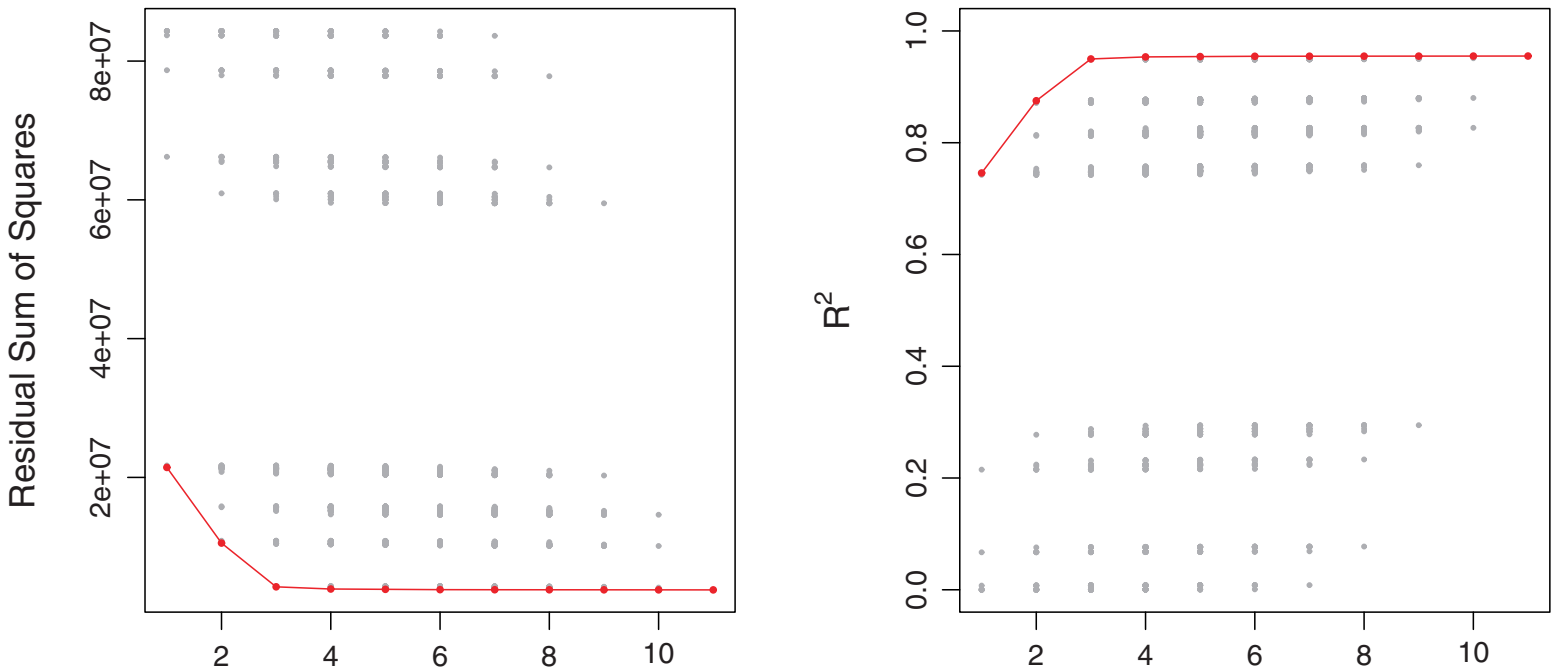
1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

There will be around $p(p+1)/2$ models!

Forward Stepwise Selection

# Variables	Best subset	Forward stepwise
One	`rating`	`rating`
Two	`rating`, `income`	`rating`, `income`
Three	`rating`, `income`, `student`	`rating`, `income`, `student`
Four	`cards`, `income`, `student`, `limit`	`rating`, `income`, `student`, `limit`

TABLE 6.1. *The first four selected models for best subset selection and forward stepwise selection on the* `Credit` *data set. The first three models are identical but the fourth models differ.*



Backward Selection

- Start with all predictors in the model.
- Remove the predictor with the largest p-value – that is, the predictor that is the least statistically significant.
- The new $(p-1)$-predictor model is fit, and the predictor with the largest p-value is removed.
- Continue until a stopping rule is reached. For instance, we may stop when all remaining predictors have a significant p-value defined by some significance threshold.

Backward Stepwise Selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

There will be around $p(p+1)/2$ models!