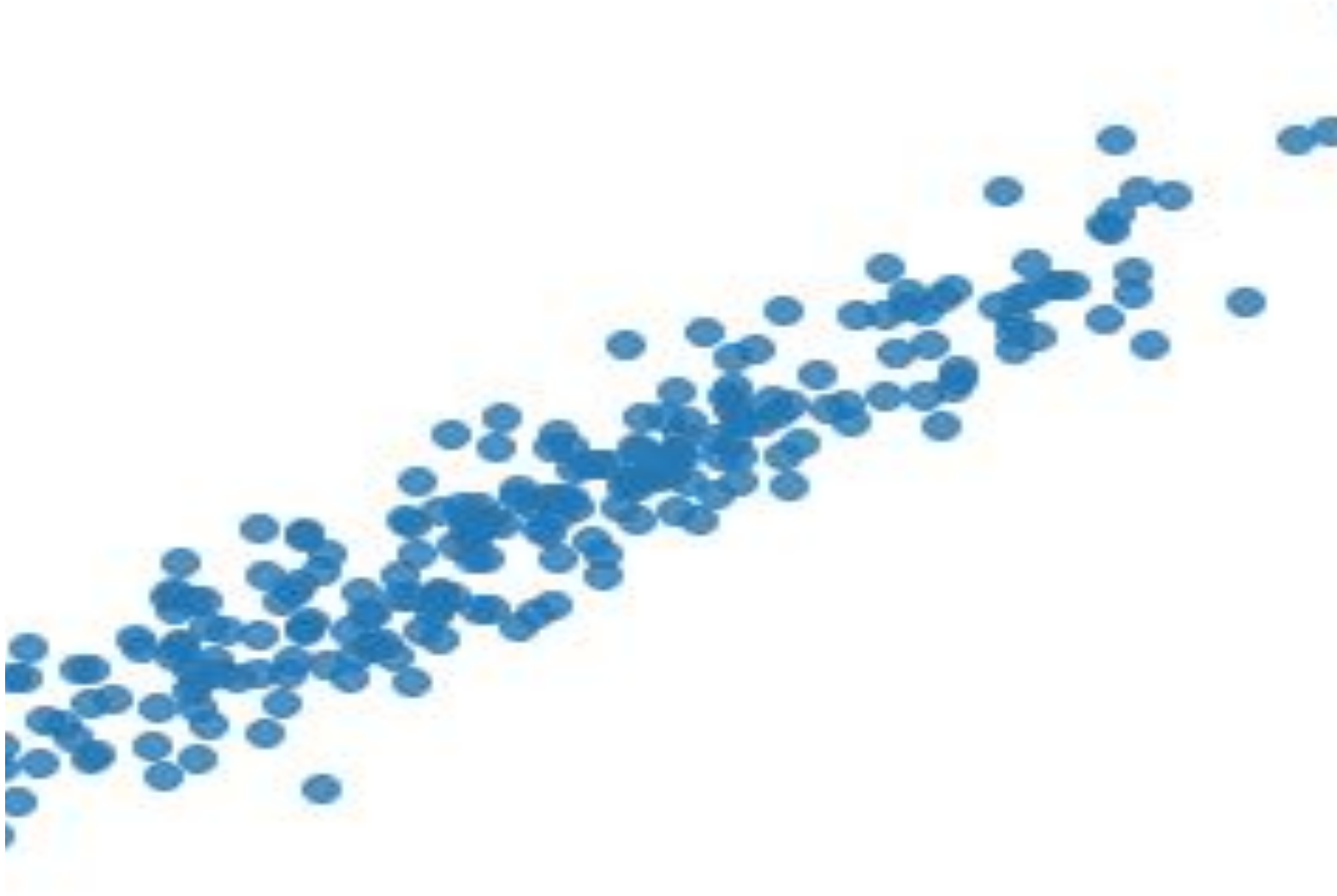
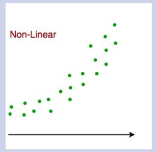
Module 2: STATISTICAL MACHINE LEARNING



Reference: James, G., Witten, D., Hastie, T., Tibshirani, R. (2017). An Introduction to Statistical Learning: with Applications in R., Springer.)

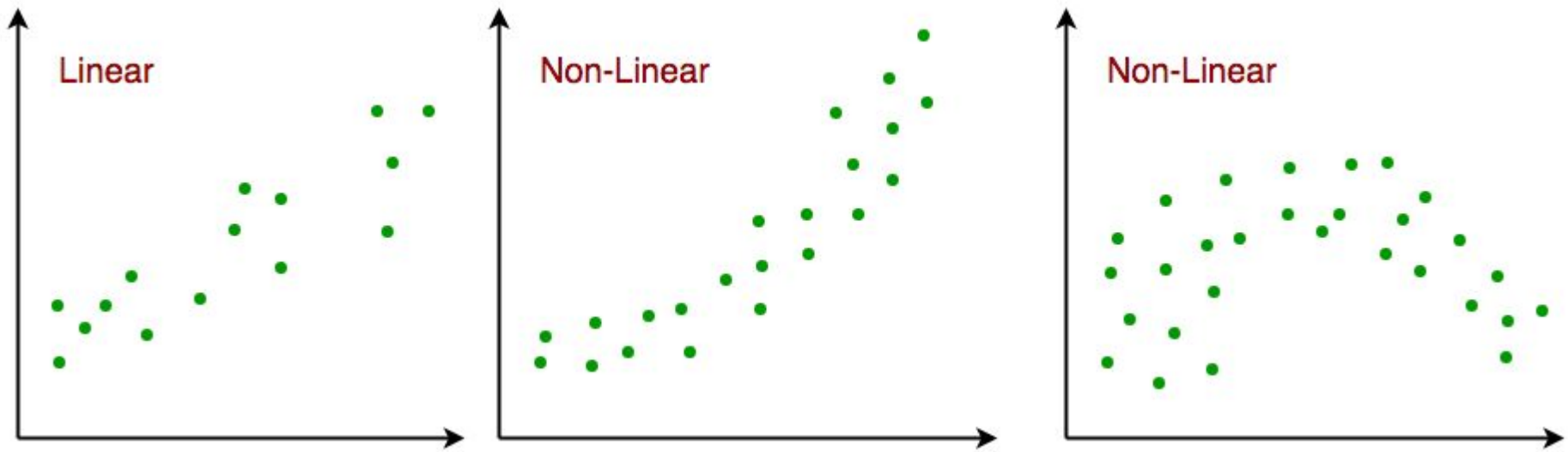
Linear regression

Linear regression is a statistical method that is used to predict a continuous dependent variable(target variable) based on one or more independent variables(predictor variables).



This technique assumes a linear relationship between the dependent and independent variables, which implies that the dependent variable changes proportionally with changes in the independent variables.

In other words, linear regression is used to determine the extent to which one or more variables can predict the value of the dependent variable.

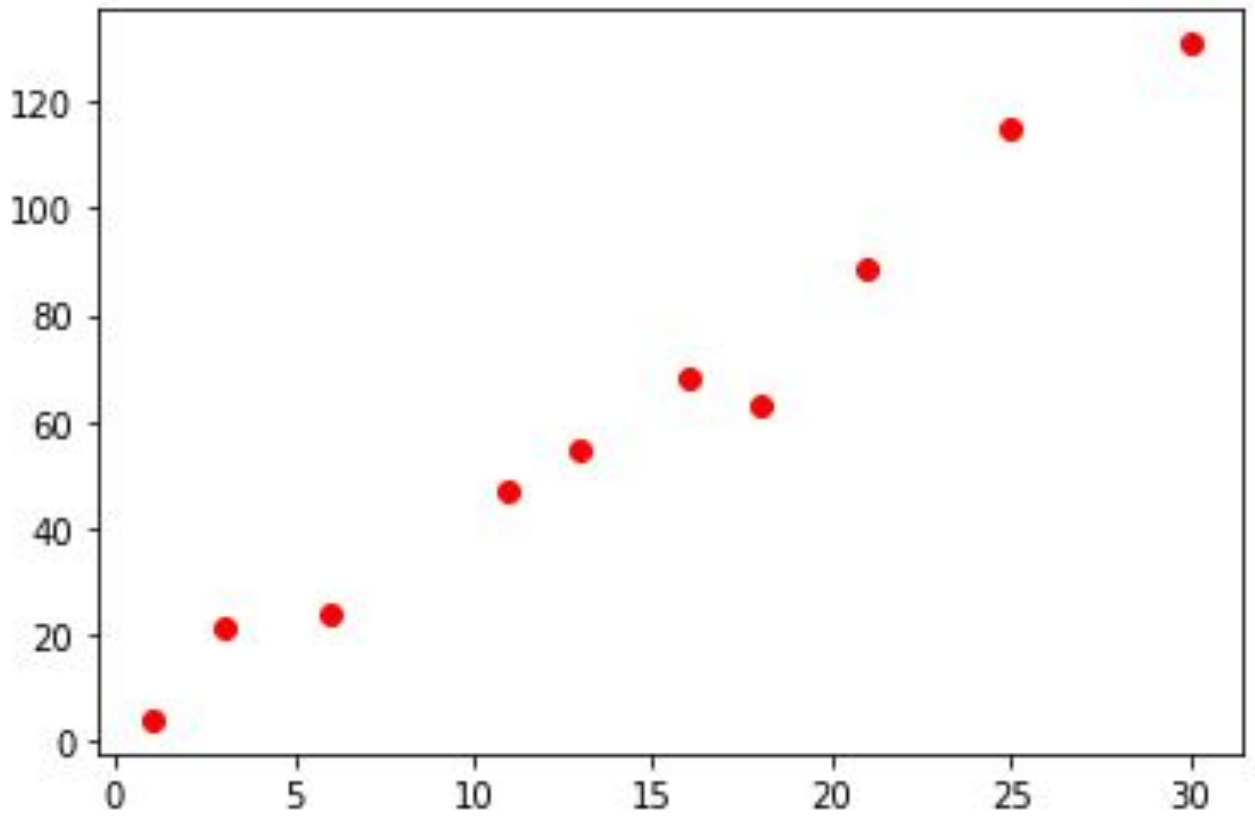

Linear
Non-Linear
Non-Linear

Types of Linear Regression

- **Simple linear regression:** This involves predicting a dependent variable based on a single independent variable.
- **Multiple linear regression:** This involves predicting a dependent variable based on multiple independent variables.

Simple linear regression

- Simple linear regression is an approach for predicting a response using a single feature.
- In linear regression, we assume that the two variables i.e. dependent and independent variables are linearly related.
- Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).
- Predictor variable (feature vector) **x**: [1, 3, 6, 11, 13, 16, 18, 21, 25, 30]
- Response variable **y**: [4, 21, 24, 47, 55, 68, 63, 89, 115, 131]



- It assumes that there is approximately a linear relationship between X and Y

$$Y \approx \beta_0 + \beta_1 X.$$

- β_0 and β_1 represent the *intercept* and *slope* terms in the linear model

- Using our training data, we estimate the coefficients $\hat{\beta}_0$ and $\hat{\beta}_1$

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

- $\hat{y}$ indicates a prediction of Y on the basis of X = x

Estimating the Coefficients

- Minimize least squared criterion
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ be the prediction for Y based on the ith value of X.
- Then $e_i = y_i - \hat{y}_i$ represents the ith residual
- We define the *residual sum of squares* (RSS) as

$$RSS = e_1^2 + e_2^2 + \cdots + e_i^2 + \cdots + e_n^2$$

$$RSS = \left(y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1\right)^2 + \left(y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2\right)^2 + \cdots + \left(y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n\right)^2$$

The least squares approach chooses $\hat{\beta}_0$ and $\hat{\beta}_1$ to minimize the RSS. Using some calculus, one can show that the minimizers are

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x}, \end{aligned} \tag{3.4}$$

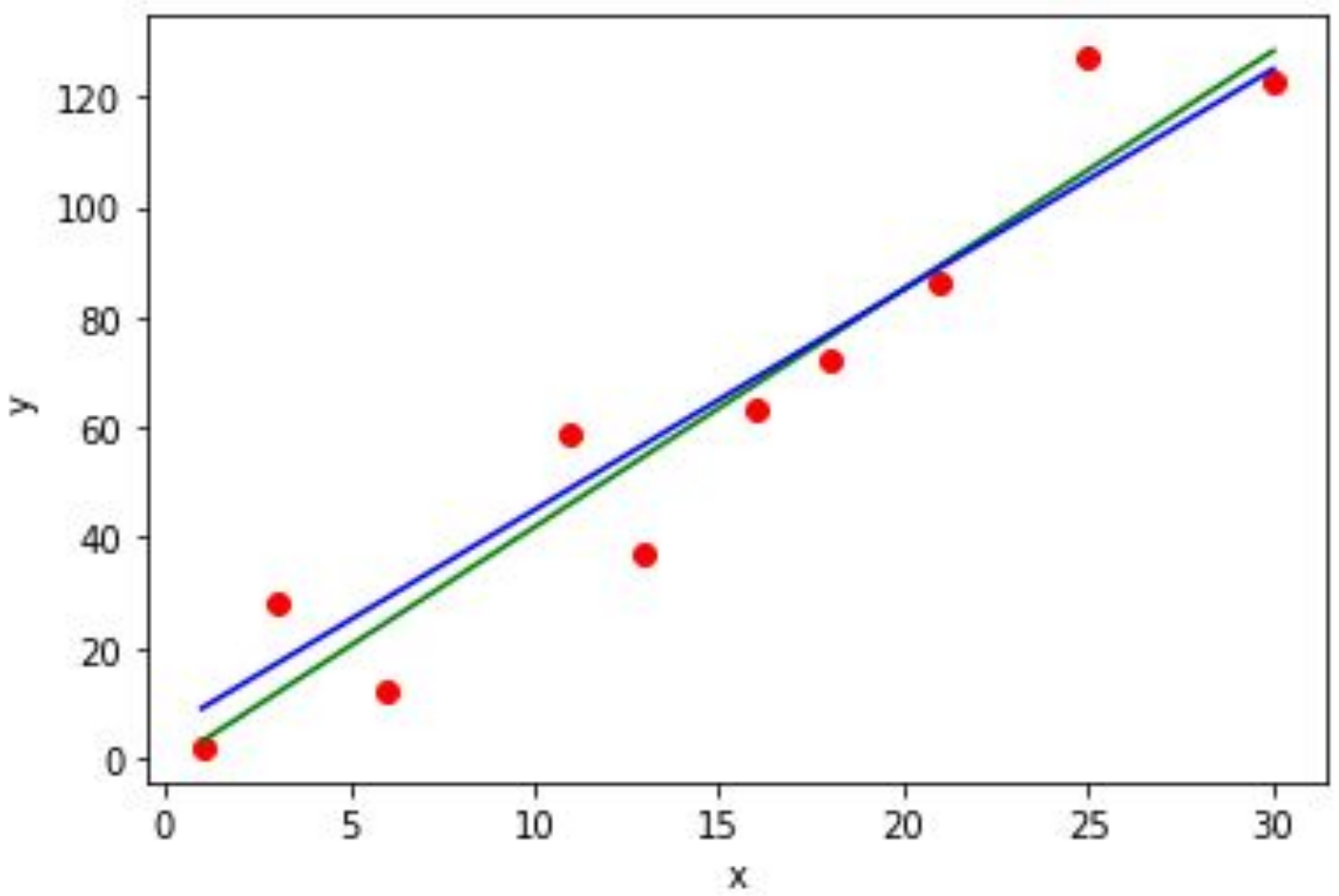
where $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$ are the sample means. In other words, (3.4) defines the *least squares coefficient estimates* for simple linear regression.

Assessing the model accuracy

- We assume that the true relationship between X and Y takes the form $Y = f(X) + \varepsilon$ for some unknown function f, where ε is a mean-zero random error term.
- If f is to be approximated by a linear function, then we can write this relationship as

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

- Population regression line
- least squares line



Population regression line

Least squares line

x = [1, 3, 6, 11, 13, 16, 18, 21, 25, 30]
y = [2, 28, 12, 59, 37, 63, 72, 86, 127, 123]

Standard errors can also be used to perform *hypothesis tests* on the coefficients. The most common hypothesis test involves testing the *null hypothesis* of

$$H_0 : \text{There is no relationship between } X \text{ and } Y \tag{3.12}$$

versus the *alternative hypothesis*

$$H_a : \text{There is some relationship between } X \text{ and } Y. \tag{3.13}$$

Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_a : \beta_1 \neq 0,$$

Assessing the Accuracy of the Model

- The quality of a linear regression fit is typically assessed using two related quantities:
- the *Residual Standard Error* (RSE) and
- the R^2 statistic.

$$Y = \beta_0 + \beta_1 X + \epsilon.$$

Residual Standard Error

- The RSE is an estimate of the standard deviation of ε
- RSS

$$RSS = e_1^2 + e_2^2 + \cdots + e_i^2 + \cdots + e_n^2$$

$$RSS = (y_1 - \hat{\beta}_0 - \hat{\beta}_1 x_1)^2 + (y_2 - \hat{\beta}_0 - \hat{\beta}_1 x_2)^2 + \cdots + (y_n - \hat{\beta}_0 - \hat{\beta}_1 x_n)^2$$

$$\text{RSE} = \sqrt{\frac{1}{n-2}\text{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

The odel (3.5)
to

the data

Quantity	Value
Residual standard error	3.26
R^2	0.612
F-statistic	312.1

TABLE 3.2. *For the* `Advertising` *data, more information about the least squares model for the regression of number of units sold on TV advertising budget.*

In the case of the advertising data, we see from the linear regression output in Table 3.2 that the RSE is 3.26. In other words, actual sales in each market deviate from the true regression line by approximately 3,260 units, on average. Another way to think about this is that even if the model were correct and the true values of the unknown coefficients β_0 and β_1 were known exactly, any prediction of sales on the basis of TV advertising would still be off by about 3,260 units on average. Of course, whether or not 3,260 units is an acceptable prediction error depends on the problem context. In the advertising data set, the mean value of `sales` over all markets is approximately 14,000 units, and so the percentage error is $3.260/14.000 = 23\,\%$.

R^2 Statistic

- To calculate R^2, we use the formula

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

- where $TSS = \sum(y_i - \bar{y})^2$ is the total sum of squares, and RSS is defined in (3.16).

- TSS measures the total variance in the response Y, and can be thought of as the amount of variability inherent in the response before the regression is performed.
- TSS−RSS measures the amount of variability in the response that is explained (or removed) by performing the regression.
- R^2 measures the proportion of variability in Y that can be explained using X.

The R^2 statistic is a measure of the linear relationship between X and Y. Recall that *correlation*, defined as

$$\text{Cor}(X, Y) = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}, \tag{3.18}$$

is also a measure of the linear relationship between X and Y.[5] This suggests that we might be able to use $r = \text{Cor}(X, Y)$ instead of R^2 in order to assess the fit of the linear model. In fact, it can be shown that in the simple linear regression setting, $R^2 = r^2$. In other words, the squared correlation

MULTIPLE LINEAR REGRESSION

Multiple linear regression

- Multiple Linear Regression attempts to model the relationship between two or more features and a response by fitting a linear equation to observed data.
- The steps to perform multiple linear Regression are almost similar to that of simple linear Regression. The Difference Lies in the evaluation.
- We can use it to find out which factor has the highest impact on the predicted output and how different variables relate to each other.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

Estimating the Regression Coefficients

As was the case in the simple linear regression setting, the regression coefficients $\beta_0, \beta_1, \ldots, \beta_p$ in (3.19) are unknown, and must be estimated. Given estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$, we can make predictions using the formula

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p. \tag{3.21}$$

The parameters are estimated using the same least squares approach that we saw in the context of simple linear regression. We choose $\beta_0, \beta_1, \ldots, \beta_p$ to minimize the sum of squared residuals

$$\begin{aligned} \text{RSS} &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \end{aligned} \tag{3.22}$$

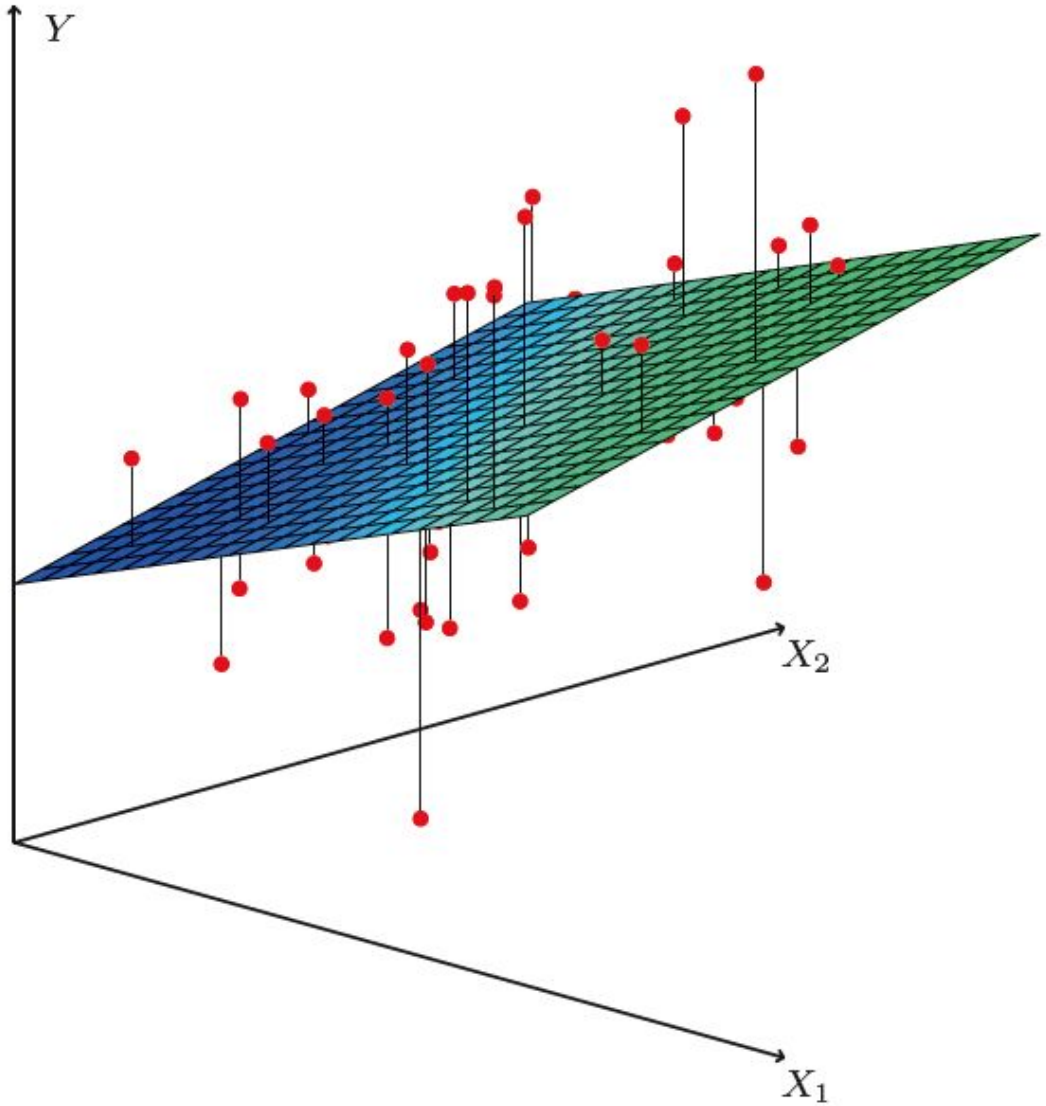


FIGURE 3.4. *In a three-dimensional setting, with two predictors and one response, the least squares regression line becomes a plane. The plane is chosen to minimize the sum of the squared vertical distances between each observation (shown in red) and the plane.*

Some Important Questions

1. *Is at least one of the predictors* $X_1, X_2, \ldots, X_p$ *useful in predicting the response?*

2. *Do all the predictors help to explain* Y*, or is only a subset of the predictors useful?*

3. *How well does the model fit the data?*

4. *Given a set of predictor values, what response value should we predict, and how accurate is our prediction?*

Is There a Relationship Between the Response and Predictors?

Recall that in the simple linear regression setting, in order to determine whether there is a relationship between the response and the predictor we can simply check whether $\beta_1 = 0$. In the multiple regression setting with p predictors, we need to ask whether all of the regression coefficients are zero, i.e. whether $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. As in the simple linear regression setting, we use a hypothesis test to answer this question. We test the null hypothesis,

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus the alternative

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

This hypothesis test is performed by computing the *F-statistic*,

$$F = \frac{(\text{TSS} - \text{RSS})/p}{\text{RSS}/(n-p-1)}, \tag{3.23}$$

where, as with simple linear regression, $\text{TSS} = \sum(y_i - \bar{y})^2$ and $\text{RSS} = \sum(y_i - \hat{y}_i)^2$. If the linear model assumptions are correct, one can show that

$$E\{\text{RSS}/(n-p-1)\} = \sigma^2$$

and that, provided H_0 is true,

$$E\{(\text{TSS} - \text{RSS})/p\} = \sigma^2.$$

Hence, when there is no relationship between the response and predictors, one would expect the F-statistic to take on a value close to 1. On the other hand, if H_a is true, then $E\{(\text{TSS} - \text{RSS})/p\} > \sigma^2$, so we expect F to be greater than 1.

In (3.23) we are testing H_0 that all the coefficients are zero. Sometimes we want to test that a particular subset of q of the coefficients are zero. This corresponds to a null hypothesis

$$H_0 : \quad \beta_{p-q+1} = \beta_{p-q+2} = \ldots = \beta_p = 0,$$

where for convenience we have put the variables chosen for omission at the end of the list. In this case we fit a second model that uses all the variables *except* those last q. Suppose that the residual sum of squares for that model is RSS_0. Then the appropriate F-statistic is

$$F = \frac{(\text{RSS}_0 - \text{RSS})/q}{\text{RSS}/(n-p-1)}. \tag{3.24}$$

Two: Deciding on Important Variables

- The task of determining which predictors are associated with the response, in order to fit a single model involving only those predictors, is referred to as **variable selection**.
- **Forward selection**.
  - We begin with the null model—a model that contains an intercept but no predictors.
  - We then fit p simple linear regressions and add to the null model the variable that results in the lowest RSS.
  - We then add to that model the variable that results in the lowest RSS for the new two-variable model.
  - This approach is continued until some stopping rule is satisfied.

Two: Deciding on Important Variables

- **Backward selection**
  - We start with all variables in the model, and remove the variable with the largest p-value—that is, the variable that is the least statistically significant.
  - The new $(p - 1)$-variable model is fit, and the variable with the largest p-value is removed.
  - This procedure continues until a stopping rule is reached.
- **Mixed selection**
  - We start with no variables in the model, and as with forward selection, we add the variable that provides the best fit.
  - We continue to add variables one-by-one.
  - If at any point the p-value for one of the variables in the model rises above a certain threshold, then we remove that variable from the model.
  - We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Three: Model Fit

- Two of the most common numerical measures of model fit are the **RSE** and $\mathbf{R^2}$, the fraction of variance explained.
- $\mathbf{R^2}$ is the square of the correlation of the response and the variable.
- In multiple linear regression, it turns out that it equals $Cor(Y, \hat{Y})^2$, the square of the correlation between the response and the fitted linear model; in fact one property of the fitted linear model is that it maximizes this correlation among all possible linear models.

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

Four: Predictions

1. The coefficient estimates $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$ are estimates for $\beta_0, \beta_1, \ldots, \beta_p$. That is, the *least squares plane*

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \cdots + \hat{\beta}_p X_p$$

is only an estimate for the *true population regression plane*

$$f(X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p.$$

Four: Predictions

2. Of course, in practice assuming a linear model for $f(X)$ is almost always an approximation of reality, so there is an additional source of potentially reducible error which we call *model bias*. So when we use a linear model, we are in fact estimating the best linear approximation to the true surface. However, here we will ignore this discrepancy, and operate as if the linear model were correct.

Four: Predictions

3. Even if we knew $f(X)$—that is, even if we knew the true values for $\beta_0, \beta_1, \ldots, \beta_p$—the response value cannot be predicted perfectly because of the random error ϵ in the model (3.21). In Chapter 2, we referred to this as the *irreducible error*. How much will Y vary from $\hat{Y}$? We use *prediction intervals* to answer this question. Prediction intervals are always wider than confidence intervals, because they incorporate both the error in the estimate for $f(X)$ (the reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (the irreducible error).