

CA MAS-I Chapter 3

3.1.0 Overview

5m

Statistical analyses are usually performed on a big dataset that records multiple variables on many observations. This section presents techniques in **statistical learning**, which aims to find possible and meaningful relationships between these variables. Among a wide scope of statistical learning methods, the focus is on linear models and other related techniques.

Computer softwares are the primary means of performing these statistical analyses; one that is commonly used in the field is the programming language R. While you are not expected to be familiar with reading or writing code, you will need to know how to interpret statistical outputs. Thus, we include snippets of R output where relevant.

To help motivate many of the key concepts, we begin with a scenario that recurs throughout Section 3.

As a consultant, Chris enjoys having flexible working hours. The commute from his home to his office covers much of the city where he lives. He considers several things related to his commute:

1. Is there an ideal time of day to start the commute in order to minimize time on the road?
2. If there is an accident on the trip to work, how much time is added to the commute, on average?
3. Are there meaningful differences/similarities in the days that the commutes are recorded? If yes, how are they characterized?

To answer these questions, Chris records the following variables for 100 days:

- **Commute:** time taken to travel from home to work (minutes)
- **Departure:** hour of departure from home (real number); for example, 14.25 represents 2:15pm
- **Temp:** temperature recorded at the time of departure (°F)
- **Precip Chance:** chance of precipitation – rain or snow – recorded at the time of departure (%)
- **Precip:** presence of precipitation during any point of the commute (Yes/No)
- **Season:** calendar season (Winter/Spring/Summer/Fall)
- **Accident:** presence of at least one accident along the route of commute (Yes/No)
- **Police:** number of police vehicles seen along the route of commute

Understanding a statistical method requires knowing key terminology used in statistical learning. This subsection covers an introduction to data terminology, fundamental statistical concepts, and basic data summaries of the numerical and graphical variety often referred to as ***descriptive statistics***.

3.1.1 Data Terminology and Notation

Let's discuss the common terms and notation used in this manual in relation to a dataset under investigation.

Types of Variables

RESPONSE VS. EXPLANATORY

The **response variable** is a variable of primary concern to an analyst. Typically, we hope to predict the response variable of a future observation and to see whether the response variable can be understood better using other variables. In the Commuting Chris setup, his first two out of three questions have the variable Commute as the response variable. Other names for "response variable" include **output variable**, **dependent variable**, **target variable**, and **outcome**.

An **explanatory variable** is any variable used to study the response variable. In other words, we aim to discover and exploit potential relationships that exist between the response variable and an explanatory variable. In the Commuting Chris setup, his first question pertains to using the variable Departure as an explanatory variable. Other names for "explanatory variable" include **input variable**, **independent variable**, **predictor**, and **feature**.

Coach's Remarks

The terms "dependent variable" and "independent variable" here should not be confused with the concept of independent random variables from probability.

QUANTITATIVE VS. QUALITATIVE

There are two types of quantitative variables of interest: **count variables** and **continuous variables**. The distinction between these two types of variables is the same as the distinction between discrete and continuous random variables. Specifically, count variables are assumed to take on non-negative integers, and continuous variables are assumed to take on values from an interval. A quantitative explanatory variable is called a **covariate**.

Qualitative variables are commonly called **categorical variables**. These variables take on a relatively small number of possible outcomes or categories. Other names for "category" include **class** and

level. When a categorical variable has only two classes, it is often referred to as a ***binary variable***.

Moreover, a qualitative explanatory variable is called a ***factor***.

Here is a breakdown of the Commuting Chris variables:

Count Variables	Continuous Variables	Categorical Variables
Police	Commute Departure Temp Precip Chance	Precip Season Accident

If a categorical variable has categories without a meaningful or logical order, it is also called a ***nominal variable***. If there is a meaningful order to the categories, then it is called an ***ordinal variable*** instead. Furthermore, it is common practice to assign numbers to the categories of a categorical variable.

Using the variable Season as an example, we may assign 1 to Winter, 2 to Spring, 3 to Summer, and 4 to Fall. Since this particular assignment follows the calendar seasons in sequence, Season is an ordinal variable here.

On the other hand, if the numbers are assigned based on the alphabetical order of the four seasons, i.e. 1 to Fall, 2 to Spring, 3 to Summer, and 4 to Winter, then the numbers do not follow a meaningful numerical sequence. This would make Season a nominal variable instead.

Notation

In general, a variable is denoted as x , along with a subscript if it is necessary to distinguish between variables. We use the subscript j for this purpose. However, y is the generally accepted symbol for a response variable.

Let p be the number of variables in a dataset, **excluding** the response variable if there is one. Thus, $j = 1, 2, \dots, p$.

OBSERVATION SUBSCRIPTS

Sometimes, it is also important to reference specific observations of a variable. We use the subscript i to achieve this. Let n be the total number of observations in a dataset. So, $i = 1, 2, \dots, n$.

For example, y_i represents the response variable data point recorded on the i^{th} observation, and $x_{i,j}$ represents the j^{th} (explanatory) variable data point recorded on the i^{th} observation. To further illustrate, the values $\{y_5, x_{5,1}, x_{5,2}, \dots, x_{5,p}\}$ are recorded from the **same** observation, i.e. the

5th one. In the Commuting Chris context, $\{y_5, x_{5,1}, x_{5,2}, \dots, x_{5,p}\}$ are the values taken from the 5th recorded day of commuting.

Coach's Remarks

When x has **two** numbers in its subscript, the subscript denotes " i, j ".

When x has **one** number in its subscript, read the context carefully. In general, if only one x variable is defined, then the subscript likely denotes " i " (because there is no purpose for j when $p = 1$); if two or more x variables are defined, then the subscript likely denotes " j ".

When considering the variables as **random variables**, we will use uppercase letters instead, such as Y and X . Similarly, subscripts may be employed.

Although introducing subscript i adds clarity, using it in an equation or expression could result in something messy and hard to read. To address this, vector and matrix notations help express things more succinctly. When using a matrix to reference a dataset, the rows of the matrix correspond to the observations, while the columns correspond to the variables.

The notation \mathbf{A}^T is the **transpose** of a matrix \mathbf{A} . This swaps the rows and columns, so the k^{th} column of \mathbf{A} is the k^{th} row of \mathbf{A}^T , and vice versa. If the dimension of \mathbf{A} is $a \times b$, then the dimension of \mathbf{A}^T is $b \times a$.

Coach's Remarks

We hope to minimize confusion by being internally consistent and clear with all notations used in Section 3. However, it is important to **not** assume that the conventions adopted here are the same in other resources or even CAS-authored problems.

It helps to train yourself to distinguish the **concept** represented by a notation from the **notation itself**. Given the lack of notation uniformity in statistics, a concept could easily be represented by several different symbols, which has occurred in previous exams. In short, always pay close attention to how notations are defined.

3.1.2 Contrasting Statistical Learning Elements

20m

The goal of statistical learning is to learn about the data at hand in order to find answers and solve problems. We start with two main ways of learning.

Supervised vs. Unsupervised

Supervised learning studies the data with a response variable. Everything centers around analyzing the response variable through the explanatory variables. There is a clear target to the investigation, so to speak. Examples of supervised learning problems are Chris's first two out of the three questions.

The official reading mentions the following supervised learning methods:

1. Least squares linear regression (SLR, MLR)
2. Subset selection
3. Shrinkage methods (e.g. lasso)
4. Generalized linear models (GLM, e.g. logistic regression)
5. Generalized additive models (GAM)
6. K-Nearest Neighbors (KNN)
7. Decision trees
8. Bagging
9. Boosting
10. Support vector machines

We will study the first five methods in detail. For the remaining methods, just know that they are used for supervised learning.

Unsupervised learning analyzes the observations or variables without a response variable. The idea is to identify patterns that may exist in the data, but there are no clear objectives or ways to verify the quality of the findings. Chris's last question is an example of an unsupervised learning problem.

The only unsupervised learning technique we will discuss is known as principal components analysis (PCA), but it also has supervised learning applications. In addition, the official reading shares that cluster analysis is another way to perform unsupervised learning.

Regression vs. Classification

Many supervised learning problems can be easily divided into two types. A **regression problem** involves a **quantitative** response variable, whereas a **classification problem** involves a **categorical** response variable. But occasionally, the distinction is a little blurred. For example, a classification problem with a "yes/no" response can be addressed by a regression that estimates the **probability** (i.e. quantitative) of a "yes".

Regression problems are the main focus in this exam. They start by assuming a relationship between the response variable and the explanatory variables. A common general form of the relationship is

$$Y = f(x_1, x_2, \dots, x_p) + \varepsilon \quad (3.1.2.1)$$

where f is some function (**not** a joint probability function) and ε is a random variable known as the **error term**. In this hypothesized relationship, we let ε have an expected value of 0, which leads to

$$\mathbb{E}[Y] = \mathbb{E}[f(x_1, x_2, \dots, x_p)] + \mathbb{E}[\varepsilon] = f(x_1, x_2, \dots, x_p)$$

In other words, an observation from the response variable is said to be made up of two components. The first is systematic, which is the expected value of the response variable. The second is random, which comes from an error term with mean 0. As a result, the mean of Y is supplied by $f(x_1, x_2, \dots, x_p)$, whereas everything else about the distribution of Y (e.g. variance and shape) is supplied by ε . A phrase often used to describe this decomposition of Y is "signal plus noise".

Notice that we assume the mean of Y is a function of the explanatory variables. Therefore, f captures how the response variable is systematically influenced by the x_j 's. For example, f may equal a constant regardless of the x_j 's. In this case, none of the explanatory variables affect $\mathbb{E}[Y]$, which means none affect Y . In other words, there is no relationship between the response and explanatory variables. Alternatively, f may equal an expression with only two of the x_j 's rather than all p of them. That would mean only those two explanatory variables have an impact on the response variable.

We do not know what f truly is in any situation, but we presume it is a relationship that does not change. In short, the goal is to estimate f to the best of our ability. We let \hat{f} denote a generic estimator or estimate of f . The process of obtaining an optimal \hat{f} is called "training", where **training data** refers to the observations in the dataset used for this purpose.

Coach's Remarks

On a technical note, all instances of "Y" thus far are more accurately denoted as

$$Y \mid X_1 = x_1, X_2 = x_2, \dots, X_p = x_p$$

However, this level of detail is not expected to be important. To give priority to more relevant content, we will tend to be less precise with probability notation.

Prediction vs. Inference

The objectives of supervised learning can be summarized into two areas:

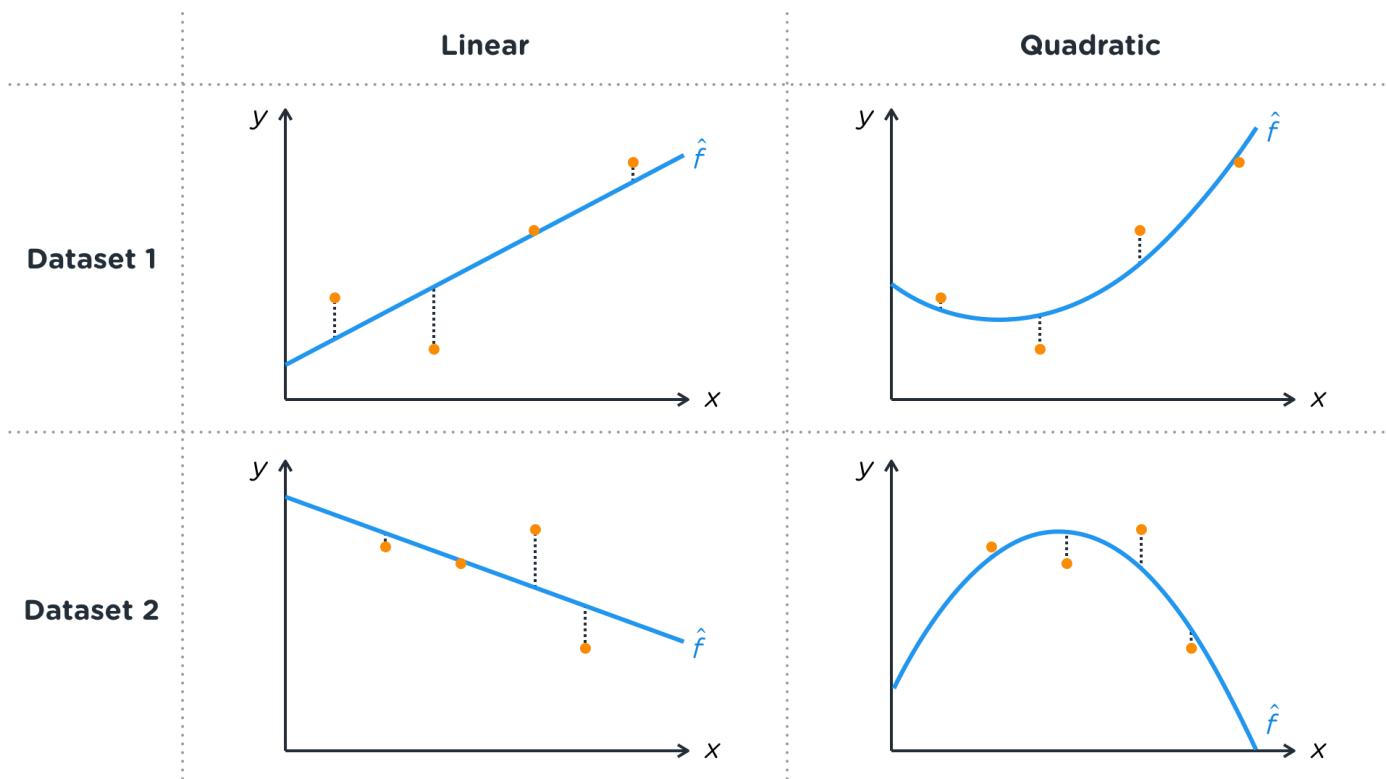
- ***Prediction*** — we want to accurately predict values of the response variable, given a set of explanatory variable values.
- ***Inference*** — we want to understand the influence that the explanatory variables have on the response variable. An example is to test whether an explanatory variable contributes significantly to f , and if so, in what way.

A method's predictive strength is highly influenced by its flexibility. ***Flexibility*** describes how closely \hat{f} is able to follow the data. A "rougher fit" describes the use of a more flexible \hat{f} ; a "smoother fit" describes the use of a less flexible \hat{f} .

Higher flexibility often comes from having more free parameters in the functional form. Here is a simple illustration:

Using one explanatory variable x , which \hat{f} is more flexible: a linear function of x or a quadratic function of x ?

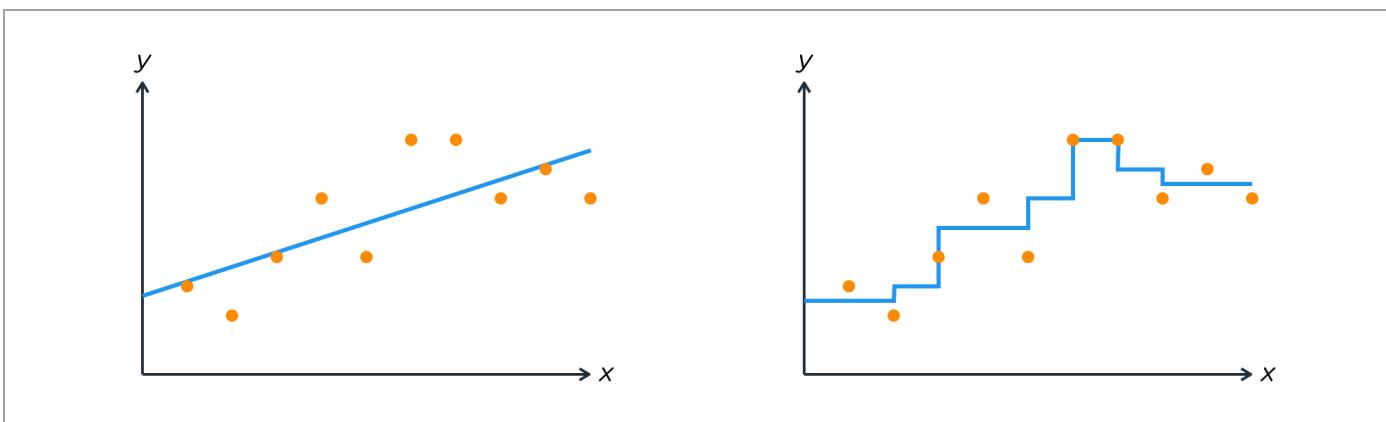
Let's consider two datasets, each with four observations. In each case, the best linear \hat{f} and the best quadratic \hat{f} are obtained using ordinary linear squares, which we will develop in detail later. This is depicted in the four plots below.



This shows how the linear \hat{f} 's are less flexible since they are not as capable of following the data compared to the quadratic \hat{f} 's. Said differently, a **quadratic \hat{f} is more flexible** because it has more free parameters than a linear \hat{f} (specifically, one more free parameter; details to come later).

Coach's Remarks

To see why a more flexible fit is also called a rougher fit, consider the plots below, both showing the same 10 observations:



In the left plot, a linear (i.e. inflexible) \hat{f} is fitted to the data. In the right plot, a very flexible \hat{f} is fitted instead. Notice that the fitted lines are

Left Plot	Right Plot
Less flexible	More flexible
Straight and "smooth"	Bumpy and "rough"

Perhaps a highly flexible \hat{f} seems like a good idea, but there is a crucial drawback:

Flexibility and accuracy do not always go hand-in-hand. High flexibility improves predictions **on the training data**. It may be possible to create an \hat{f} that makes perfect predictions on past data, but that is not the objective. An \hat{f} with good predictive ability on future data is likely one that is not extremely flexible. This idea is further developed in the next subsection.

Coach's Remarks

Many of these concepts will likely become clearer after learning about the specific methods covered on the exam. Revisiting this subsection upon completing Section 3 would help to solidify the concepts.

Example 3.1.2.1

Determine which statements below are true regarding the following relationship between a response variable Y and an explanatory variable x .

$$Y = f(x) + \varepsilon$$

- I. This relationship addresses an unsupervised learning problem.
- II. $f(x)$ captures the systematic part of Y .
- III. This is a regression problem only if x is quantitative.
- IV. Assuming a form for f with more free parameters typically leads to a more flexible fit.

Solution

I is false because an unsupervised learning problem does not have a response variable.

III is false because a regression problem involves a quantitative response variable, which is implied by the given relationship. Explanatory variables are free to be quantitative or categorical, but they do not dictate whether the problem belongs to regression or classification.

Therefore, **only II and IV are true.**



3.1.3 Model Accuracy

In a regression setting, recall that $f(x_1, \dots, x_p) = E[Y]$. Thus, let

$$\hat{Y} = \hat{f}(x_1, \dots, x_p) = \hat{E}[Y]$$

This means \hat{Y} is an estimator for $E[Y]$.

To quantify model accuracy, we establish a suitable measure for "model error" using \hat{Y} and Y . A large model error signifies low accuracy (i.e. the predicted responses differ prominently from the actual responses), and vice versa. In short, model accuracy revolves around comparing \hat{Y} against Y – the smaller the disparity, the better the accuracy.

Mean Squared Error

A typical measure for a regression model error is the **mean squared error (MSE)**:

$$MSE = E\left[\left(Y - \hat{Y}\right)^2\right] \quad (3.1.3.1)$$

Coach's Remarks

This formula looks identical to Equation 2.2.2.3. The key difference is:

- Equation 2.2.2.3 measures **the quality of an estimator** by taking the difference between a random variable and a parameter (i.e. constant).
- Equation 3.1.3.1 measures **model error** by taking the difference between two random variables.

The statistical learning method with the smallest MSE would be considered the most accurate. But in order to compute the MSE exactly, specifics on the relevant distributions are needed. An alternative is to estimate the MSE as the average of $(y - \hat{y})^2$ from a reasonably large number of observations, n .

$$\frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n} \quad (3.1.3.2)$$

where

- y_i is the **actual/observed** response for the i^{th} observation, and
- \hat{y}_i is the **predicted** response for the i^{th} observation.

One might decide to evaluate Equation 3.1.3.2 using the training data, i.e. calculate the **training MSE**. However, due to the nature of many training algorithms, the training MSE decreases as the flexibility of \hat{f} increases. This makes it tempting to conclude that a highly flexible \hat{f} is synonymous with a highly accurate \hat{f} ; as alluded to in the previous subsection, this thinking is misguided.

Although a more flexible \hat{f} narrows the difference between each pair of y_i and \hat{y}_i , it does not account for responses outside of the training data. In other words, an \hat{f} can be so flexible that it fails to accurately predict the responses of new observations. **Overfitting** occurs when \hat{f} fits the training data too closely and does not properly capture f , the true relationship between the response and explanatory variables.

Consequently, the training MSE is not a suitable indicator of model accuracy. To emphasize this, the MSE of Equation 3.1.3.1 is more properly called the **test MSE**.

$$\text{Test MSE} = E \left[(Y - \hat{Y})^2 \right] \quad (3.1.3.3)$$

The logic is that using **test data** – observations that were not used to train \hat{f} – to evaluate Equation 3.1.3.2 would be more suitable than using training data. Interestingly, across many different contexts with comparable methods, it is generally a moderately-flexible method which produces the lowest test MSE.

Let's consider a simple example to see the concepts in action.

The following is based on the Commuting Chris scenario.

- Training data:

Observation	Commute,	Departure,
1	24.283	9.250
2	21.433	13.183
3	31.483	6.683
4	23.650	14.183
5	26.633	12.433
6	28.567	11.050

Observation	Commute, y	Departure, x
7	23.750	14.000

- We fit three methods to the training data – \hat{f}_L has low flexibility, \hat{f}_M has moderate flexibility, and \hat{f}_H has high flexibility. Upon training each method, we obtain the following formulas:

- $\hat{f}_L(x) = 36.475 - 0.935x$
- $\hat{f}_M(x) = 141.837 - 33.009x + 3.111x^2 - 0.097x^3$
- $\hat{f}_H(x) = -3,633 + 1,962.53x - 410.695x^2 + 41.9784x^3 - 2.09855x^4 + 0.04110x^5$

What is the training MSE for each method?

For the method with **low flexibility**, the predicted responses on the training data are:

- $\hat{y}_1 = \hat{f}_L(x_1) = \hat{f}_L(9.25) = 36.475 - 0.935(9.25) = 27.826$
- $\hat{y}_2 = \hat{f}_L(x_2) = \hat{f}_L(13.183) = 24.149$
- ...
- $\hat{y}_7 = \hat{f}_L(x_7) = \hat{f}_L(14) = 23.385$

Therefore,

$$\begin{aligned} \text{Training MSE} &= \frac{\sum_{i=1}^7 (y_i - \hat{y}_i)^2}{7} \\ &= \frac{(24.283 - 27.826)^2 + (21.433 - 24.149)^2 + \dots + (23.750 - 23.385)^2}{7} \\ &= \mathbf{4.412} \end{aligned}$$

For the method with **moderate flexibility**,

- $\hat{y}_1 = \hat{f}_M(9.25) = 141.837 - 33.009(9.25) + 3.111(9.25^2) - 0.097(9.25^3) = 25.918$
- $\hat{y}_2 = \hat{f}_M(13.183) = 25.109$
- ...
- $\hat{y}_7 = \hat{f}_M(14) = 23.299$

$$\begin{aligned} \text{Training MSE} &= \frac{(24.283 - 25.918)^2 + (21.433 - 25.109)^2 + \dots + (23.750 - 23.299)^2}{7} \\ &= \mathbf{3.435} \end{aligned}$$

For the method with **high flexibility**,

- $\hat{y}_1 = \hat{f}_H(9.25) = -3,633 + 1,962.53(9.25) - 410.695(9.25^2) + 41.9784(9.25^3) - 2.09855(9.25^4) + 0.0411015(9.25^5) = 24.226$
- $\hat{y}_2 = \hat{f}_H(13.183) = 22.135$
- ...
- $\hat{y}_7 = \hat{f}_H(14) = 22.406$

$$\text{Training MSE} = \frac{(24.283 - 24.226)^2 + (21.433 - 22.135)^2 + \dots + (23.750 - 22.406)^2}{7} = \mathbf{0.435}$$

Resuming from the previous setup, we have the following test data:

Observation	Commute, y	Departure, x
1	32.317	13.100
2	26.017	10.400
3	26.767	7.667
4	25.450	11.350
5	28.033	8.600
6	22.017	14.467
7	24.150	9.267

Using the test data and Equation 3.1.3.2, what is the estimated test MSE for each method?

For each flexibility level, calculate the \hat{y} 's with the same formula from before, now using the x 's from the test data. Then, estimate the test MSEs.

For the method with **low flexibility**,

$$\text{Test MSE} = \frac{(32.317 - 24.227)^2 + (26.017 - 26.751)^2 + \dots + (24.150 - 27.810)^2}{7} = \mathbf{12.434}$$

For the method with **moderate flexibility**,

$$\text{Test MSE} = \frac{(32.317 - 25.233)^2 + (26.017 - 25.917)^2 + \dots + (24.150 - 25.912)^2}{7}$$

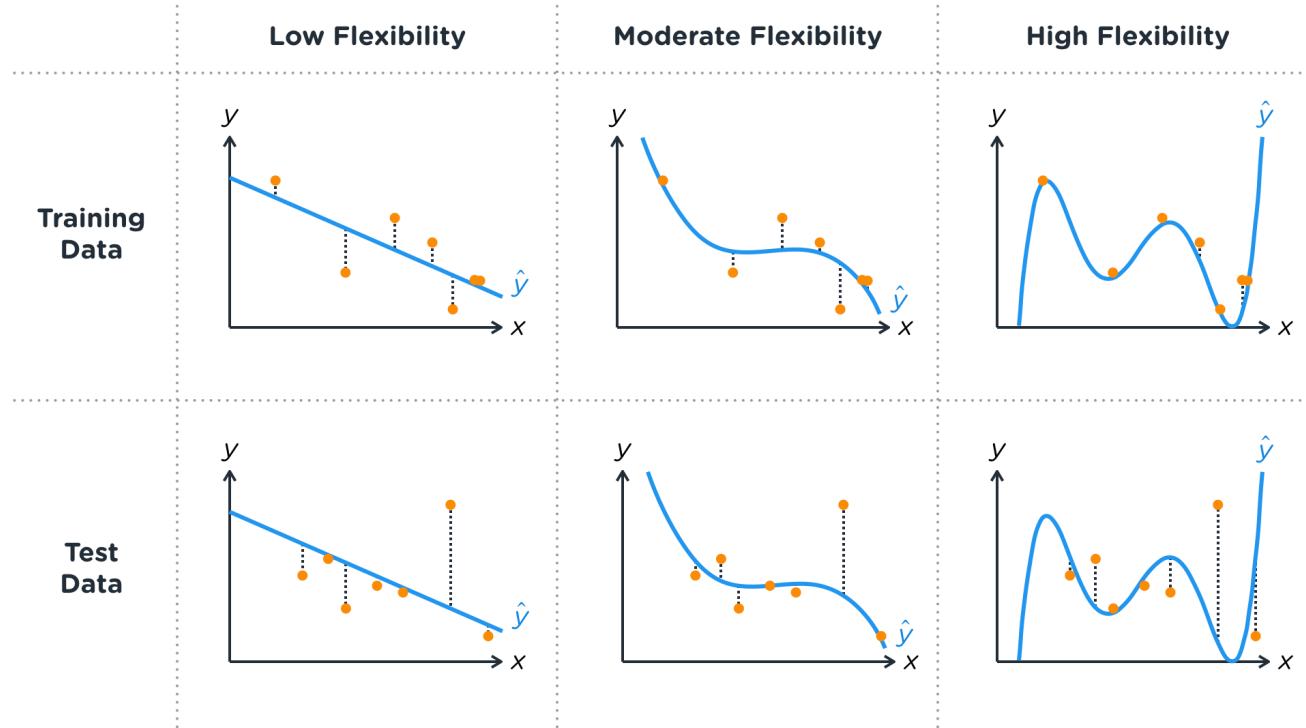
$$= 8.284$$

For the method with **high flexibility**,

$$\text{Test MSE} = \frac{(32.317 - 22.505)^2 + (26.017 - 27.091)^2 + \dots + (24.150 - 24.242)^2}{7}$$

$$= 24.874$$

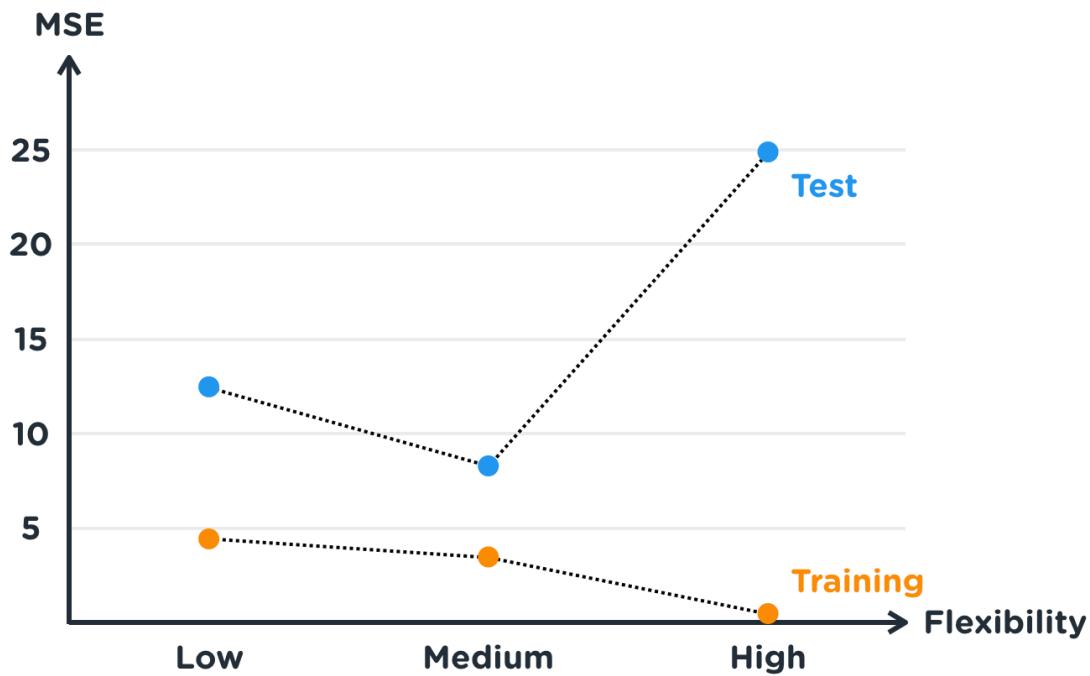
Before we consider these values, let's examine model accuracy from a more visual standpoint. The following six plots show each $\hat{y} = \hat{f}(x)$ superimposed on both the training data and the test data.



With the training data, \hat{y} gets closer to the data points as flexibility increases. However, this is not the case with the test data. Notice that the \hat{y} 's of both low and high flexibilities are not as close to the test observations, relative to the \hat{y} of moderate flexibility. In conclusion, \hat{f}_M is considered the most accurate among the three because

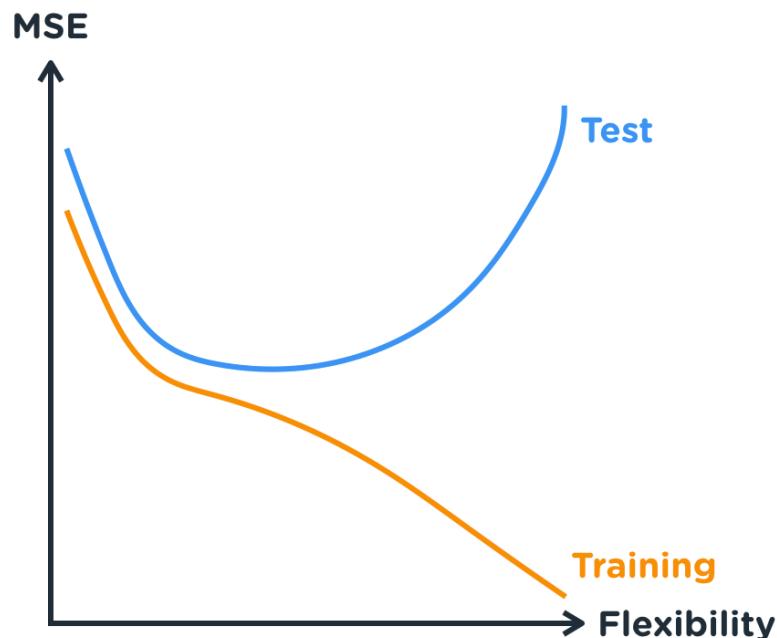
1. it best captures the relationship between x and y based on the training data, and
2. it does not excessively fit to patterns that might be unique to the training data.

For a more computational analysis, let's plot the calculated MSEs as a function of flexibility.



Again, this reveals that \hat{f}_M produces the lowest (estimated) test MSE, and thus, is considered the more accurate fit.

In general, the training and test MSEs follow this pattern:



The key points to note from this plot are:

1. **The training MSE is consistently less than the test MSE at every level of flexibility** – since \hat{f} is optimized by the training data, its predictive performance on test data is expected to be weaker.
2. **The training MSE decreases as flexibility increases** – higher flexibility makes it easier to narrow the difference between each pair of y_i and \hat{y}_i .
3. **The test MSE is u-shaped** – accuracy is worse when \hat{f} is either too inflexible or too flexible.

Bias-Variance Trade-off

The behavior of the test MSE can be further understood by the *bias-variance trade-off*. The test MSE at fixed inputs x_1, \dots, x_p can be written as the following sum:

$$\text{Var}[\hat{f}(x_1, \dots, x_p)] + (\text{Bias}[\hat{f}(x_1, \dots, x_p)])^2 + \text{Var}[\varepsilon]$$

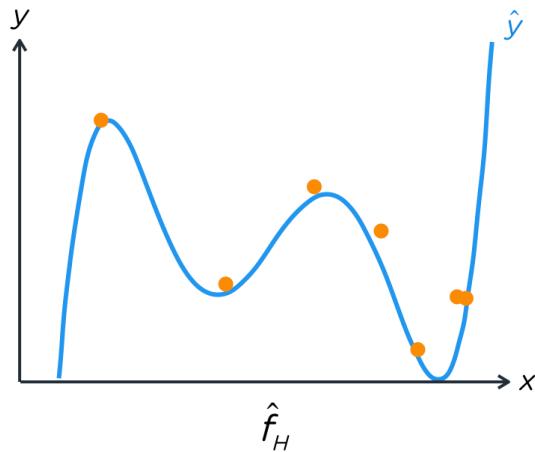
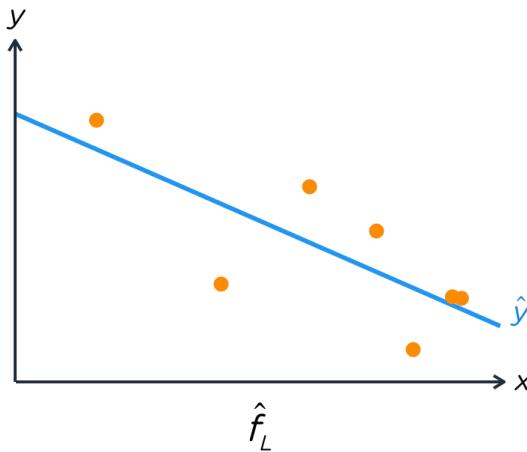
The last term is the variance of the error term, ε . It is also called the *irreducible error* – the variation in the response variable which is not captured or quantified by \hat{f} . Regardless of our choice of \hat{f} , $\text{Var}[\varepsilon]$ will not change. Therefore, the test MSE cannot be smaller than the irreducible error.

Overfitting can be described as \hat{f} being too flexible such that it captures the patterns caused by the irreducible error instead of only the patterns from f .

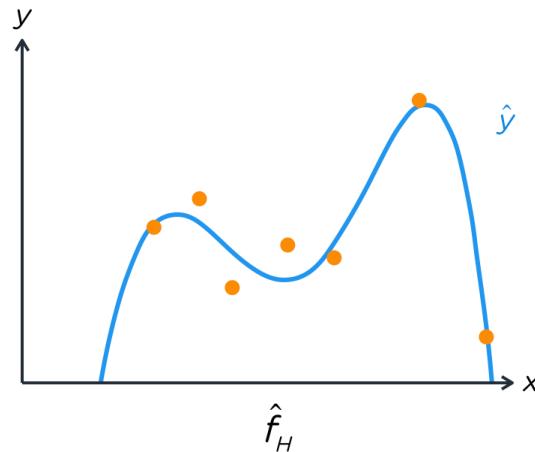
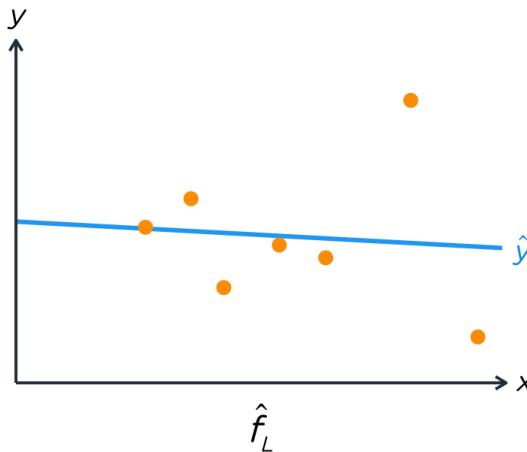
Collectively, the first two terms make up the *reducible error* since they are directly linked to the chosen statistical learning method. Said differently, our choice of \hat{f} impacts the test MSE via two sources. In seeking the lowest test MSE, we desire an $\hat{f}(x_1, \dots, x_p)$ with both low variance and low bias.

VARIANCE

The variance of $\hat{f}(x_1, \dots, x_p)$ relates to the variation in the shape of \hat{f} when different training data is used. It helps to demonstrate this concept by revisiting these two plots of the training data:



\hat{f}_L is a linear function, and \hat{f}_H is a quintic function. Both functions have parameter estimates that minimize the gaps between each pair of y_i and \hat{y}_i . If given a different set of seven observations to train on, we would expect to get different parameter estimates and for both \hat{f}_L and \hat{f}_H to change accordingly. As an example, let's use what we previously referred to as test data to train the linear and quintic functions. Here are the new plots:



A larger variance indicates that the shape of \hat{f} has a tendency to change more, depending on the training data used. As seen in the plots, the shape of \hat{f}_H changed far more than that of \hat{f}_L , which suggests that \hat{f}_H has a higher variance. Generally, methods with higher flexibility also have higher variance.

BIAS

The concept of bias was introduced in Section 2.2.2. Hence, the bias of $\hat{f}(x_1, \dots, x_p)$ measures the average closeness between \hat{f} and f .

For an \hat{f} with low flexibility, the possible predicted responses are more confined and skewed, which makes it more likely for their average to be far from f . By increasing flexibility, the average of the predicted responses is able to get closer to f , hence lowering the squared bias. Once \hat{f} is close to capturing f well, raising flexibility further would hardly lower the squared bias.

In short, bias arises when \hat{f} assumes a simpler form of f than the true f . Choosing a more complex form helps to lower the bias.

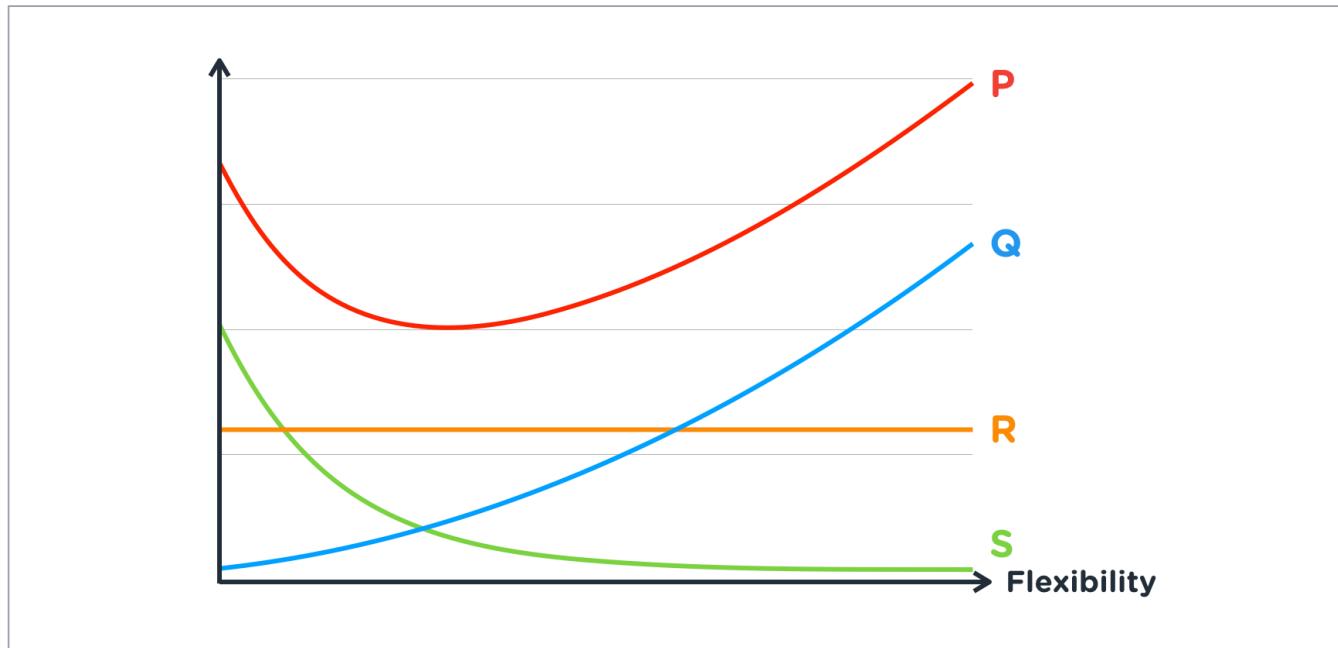
Altogether, the rule of thumb is

Flexibility	Variance	Squared Bias
Low	Low	High
High	High	Low

Let's tie things back to the test MSE. When considering methods with increasing flexibility, we simultaneously see a rising variance and a falling squared bias. Typically, the falling squared bias is initially more significant, causing the test MSE to experience a net decrease. At some point, the influence of the rising variance is matched by that of the falling squared bias. When this occurs, the test MSE is minimized. Past that level of flexibility, the rising variance dominates, and the test MSE experiences a net increase. In other words, a more accurate method is unlikely to have the lowest variance or the lowest squared bias, even though both are desirable. There is a trade-off between variance and bias to consider.

Example 3.1.3.1

Below is a plot of model accuracy quantities as functions of flexibility for comparable statistical learning methods, assuming a regression problem.



Identify the quantity that most likely belongs to each curve.

Solution

P is most likely the **test MSE** since it exhibits a u-shaped pattern. It is also the highest curve among the rest, which makes sense given the other possible quantities.

Q is most likely the **variance of $\hat{f}(x_1, \dots, x_p)$** since it increases with flexibility.

R is most likely the **irreducible error, $\text{Var}[\epsilon]$** . It is constant, regardless of flexibility.

S is most likely the **squared bias of $\hat{f}(x_1, \dots, x_p)$** since it decreases with higher flexibility. One might argue that S could be the training MSE, but notice how combining Q, R, and S results in P.



Example 3.1.3.2

Determine which statements are true.

- I. For regression, \hat{Y} typically denotes the estimator of the response variable.
- II. The training MSE is a good measure of model accuracy because it decreases as model flexibility increases.
- III. The variance of a model's error term can be influenced by the choice of statistical learning method.
- IV. In general, the lower bound of the test MSE is 0.
- V. A statistical learning method that has both the lowest variance and the lowest bias is usually attainable.

Solution

I is false because \hat{Y} denotes the estimator of the mean of the response variable in a regression setting. This can be a source of confusion because \hat{Y} is also referred to as "the predicted response". The key is that Y is not a parameter, but its mean is.

II is false because the training MSE is actually a poor measure of model accuracy for the very reason given. The statement insinuates that flexibility and accuracy go hand-in-hand.

III is false because the variance of a model's error term is the irreducible error.

IV is false because the lower bound of the test MSE is the irreducible error — it is impossible for the test MSE to be lower than it. The statement is only true if the irreducible error equals 0.

V is false because it directly contradicts the bias-variance trade-off.

Therefore, **none of the statements are true.**



3.1.4 Numerical Summaries

Next, we need to be familiar with descriptive statistics to give us a preliminary overview of the data. We begin with discussing numerical summaries, many of which have already been mentioned.

Univariate

For a variable x , we denote the sample mean and the (unbiased) sample variance as \bar{x} and s_x^2 , respectively. These are calculated as shown in Equations 2.1.5.1 and 2.3.3.3.

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad (2.1.5.1)$$

$$s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \quad (2.3.3.3)$$

Moreover, sample percentiles are also useful. As hinted in Section 2.1.2, there is no consensus on how sample percentiles should be computed. Hence, a conceptual understanding is sufficient.

The **interquartile range** is the difference between the third and first quartiles. As a result, it measures the width of the interval containing the central 50% of possible values (i.e. between the 25th and 75th percentiles).

Bivariate

For variables x and y , their (unbiased) **sample covariance** is

$$cov_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \quad (3.1.4.1)$$

and their **sample correlation** is

$$r_{x,y} = \frac{cov_{x,y}}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.1.4.2)$$

A sample correlation must be between -1 and 1. This value describes the linear strength between the two variables. Thus,

- $r_{x,y} = -1$ means x and y have a **perfect negative linear** relationship, i.e. y is a linear function of x with a negative slope.
- $-1 < r_{x,y} < 0$ means x and y have a **negative linear** relationship, i.e. as y increases, x generally decreases linearly, and vice versa. The relationship is stronger the closer $r_{x,y}$ is to -1.
- $r_{x,y} = 0$ means x and y have **no linear** relationship.
- $0 < r_{x,y} < 1$ means x and y have a **positive linear** relationship, i.e. as y increases, x generally increases linearly, and vice versa. The relationship is stronger the closer $r_{x,y}$ is to 1.
- $r_{x,y} = 1$ means x and y have a **perfect positive linear** relationship, i.e. y is a linear function of x with a positive slope.

You observe the following values for two variables.

$x :$	9	4	7	3	7
$y :$	5	7	5	4	4

Compute their

- sample means
- sample variances
- sample covariance
- sample correlation

The sample means are

$$\bar{x} = \frac{9 + 4 + 7 + 3 + 7}{5} = 6$$

$$\bar{y} = \frac{5 + 7 + 5 + 4 + 4}{5} = 5$$

The sample variances are

$$s_x^2 = \frac{(9 - 6)^2 + (4 - 6)^2 + \dots + (7 - 6)^2}{5 - 1} = 6$$

$$s_y^2 = \frac{(5 - 5)^2 + (7 - 5)^2 + \dots + (4 - 5)^2}{5 - 1} = 1.5$$

The sample covariance is

$$cov_{x,y} = \frac{(9 - 6)(5 - 5) + (4 - 6)(7 - 5) + \dots + (7 - 6)(4 - 5)}{5 - 1} = -0.5$$

Using previous calculations, the sample correlation is

$$r_{x,y} = \frac{-0.5}{\sqrt{6} \cdot \sqrt{1.5}} = -\frac{1}{6}$$

Alternatively, the sample correlation can be found directly as

$$(9 - 6)(5 - 5) + (4 - 6)(7 - 5) + \dots + (7 - 6)(4 - 5) = -2$$

$$(9 - 6)^2 + (4 - 6)^2 + \dots + (7 - 6)^2 = 24$$

$$(5 - 5)^2 + (7 - 5)^2 + \dots + (4 - 5)^2 = 6$$

$$r_{x,y} = \frac{-2}{\sqrt{24 \cdot 6}} = -\frac{1}{6}$$

Example 3.1.4.1

Determine which statement is true.

- I. For variables with a sample correlation of -1, a plot of their values will form a line with a slope of -1.
- II. A sample covariance of 0 will lead to a sample correlation of 0.
- III. A sample correlation of 0 means there is no relationship between the two variables.

Solution

I is false because a sample correlation of -1 means the values will form a line with a negative slope. The slope can be any negative number, not necessarily -1.

II is true because $r_{x,y} = \frac{\text{cov}_{x,y}}{s_x \cdot s_y} = \frac{0}{s_x \cdot s_y} = 0$.

III is false because a sample correlation of 0 means there is no linear relationship. A non-linear relationship could exist while the sample correlation is 0.

Therefore, **only II is true.**



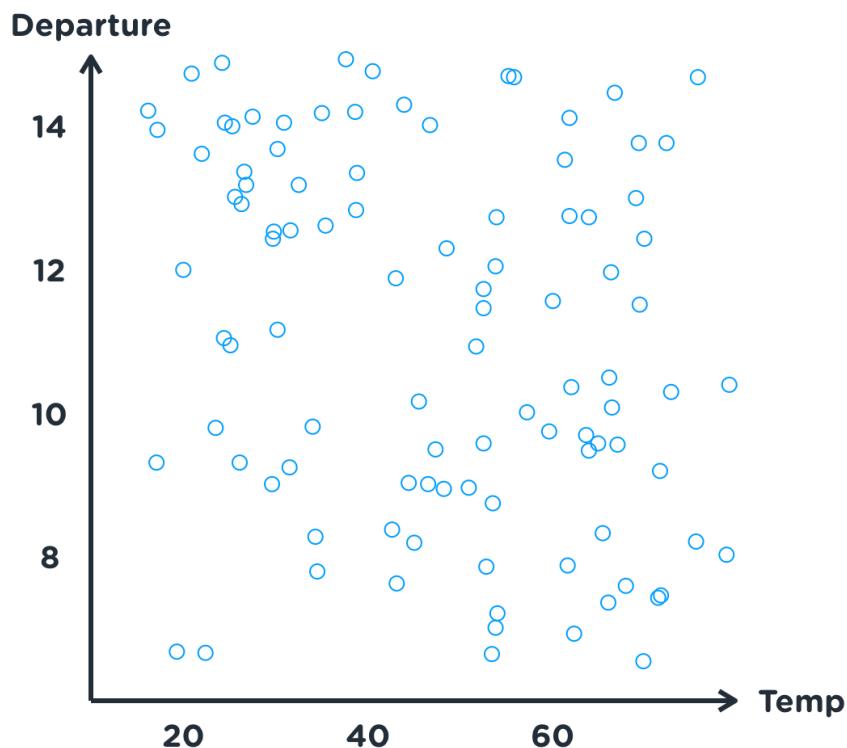
3.1.5 Scatterplots

We resume with introducing three graphical summaries in the remaining subsections, beginning with scatterplots.

A **scatterplot** is a plot of values from two variables where each point represents an observation. We can discover potential relationships between variables by noting patterns in a scatterplot.

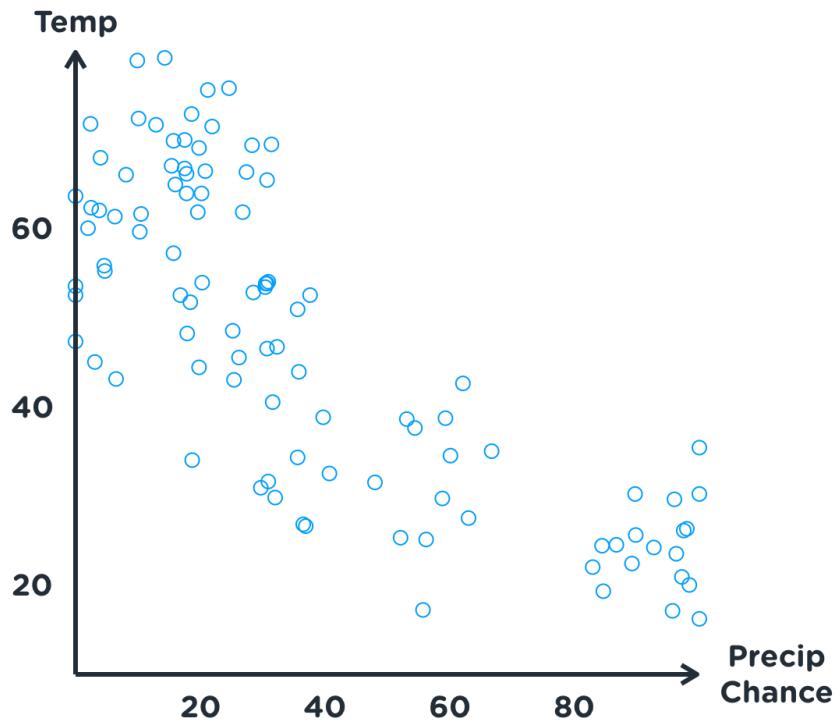
Let's dissect the Commuting Chris data with several scatterplots. We will use all 100 observations of the dataset.

First, consider the scatterplot of Departure versus Temp:



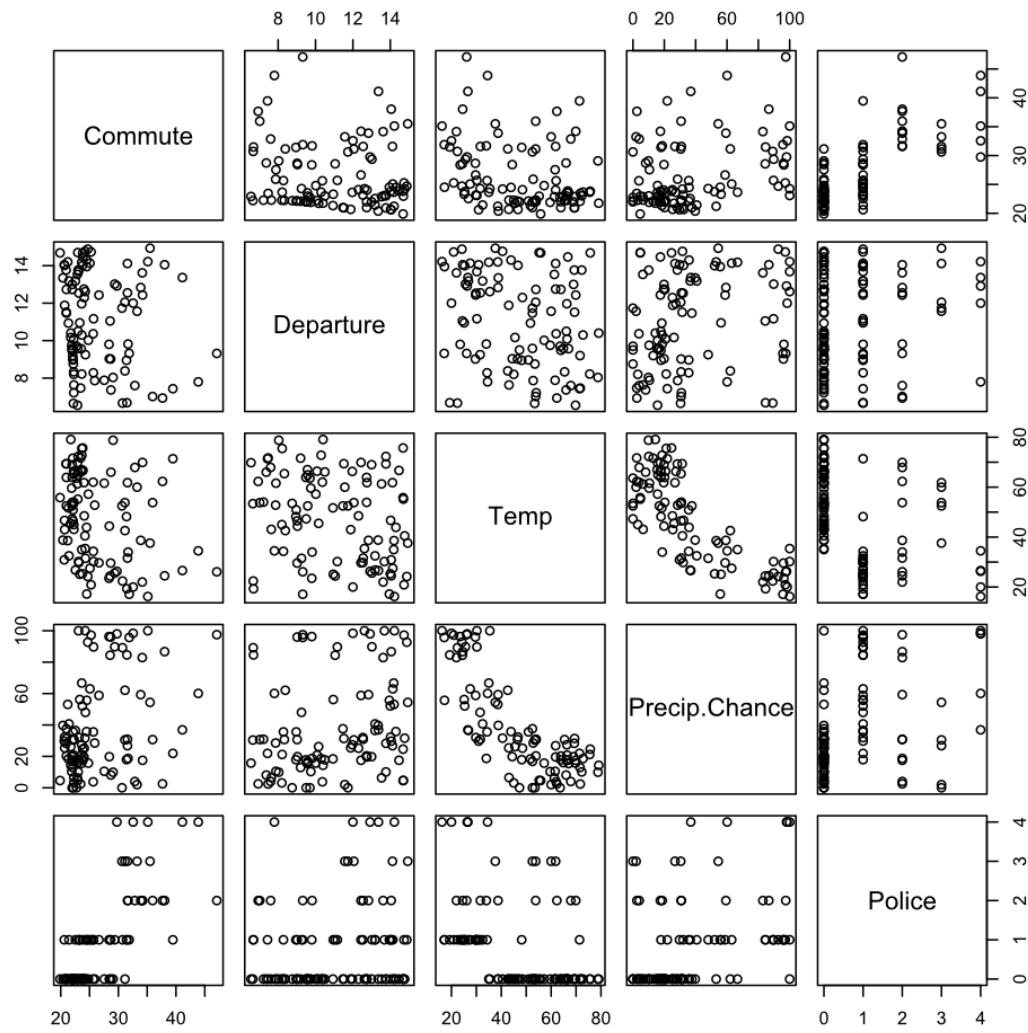
The observations appear completely patternless, with points in practically every region. One may conclude that there is no clear relationship between the times that Chris leaves for work and the temperatures at those times.

Next, consider the scatterplot of Temp versus Precip Chance:



Higher temperatures are clearly linked with a lower chance of precipitation, and vice versa. This result is expected and agrees with intuitive climate patterns. There even appears to be a faint quadratic relationship between the two variables.

A **scatterplot matrix** arranges the scatterplots for every pair of variables into a square matrix. Below is a scatterplot matrix on the five quantitative variables – Commute, Departure, Temp, Precip Chance, and Police – produced by the statistical software R:



To understand its format, note that:

- Every row of the matrix has only one box with the name of a variable. The same is true for every column of the matrix.
- For any specific scatterplot, the variable name listed in the same **row** corresponds to the **vertical axis**; the variable name listed in the same **column** corresponds to the **horizontal axis**.

For example, the scatterplot in the top-right corner has Commute for the vertical axis and Police for the horizontal axis.

- This matrix has only 10 unique scatterplots, although there are twice as many plots in total. This is because the upper-diagonal scatterplots mirror the lower-diagonal scatterplots, the only difference being the swapping of axes.

For example, the scatterplot in the bottom-left corner mirrors the one in the top-right corner, but with Commute for the horizontal axis and Police for the vertical axis.

Connection with Correlation

While correlation measures the linear strength between two variables numerically, a scatterplot may detect it visually. Here are the sample correlations among all pairs of the five variables:

	Commute	Departure	Temp	Precip Chance	Police
Commute	1.000	-0.145	-0.290	0.342	0.764
Departure		1.000	-0.266	0.242	0.162
Temp			1.000	-0.785	-0.453
Precip Chance				1.000	0.401
Police					1.000

This table is similar to the scatterplot matrix, in that the upper-diagonal is mirrored in the lower-diagonal. For this reason, the lower-diagonal is left empty.

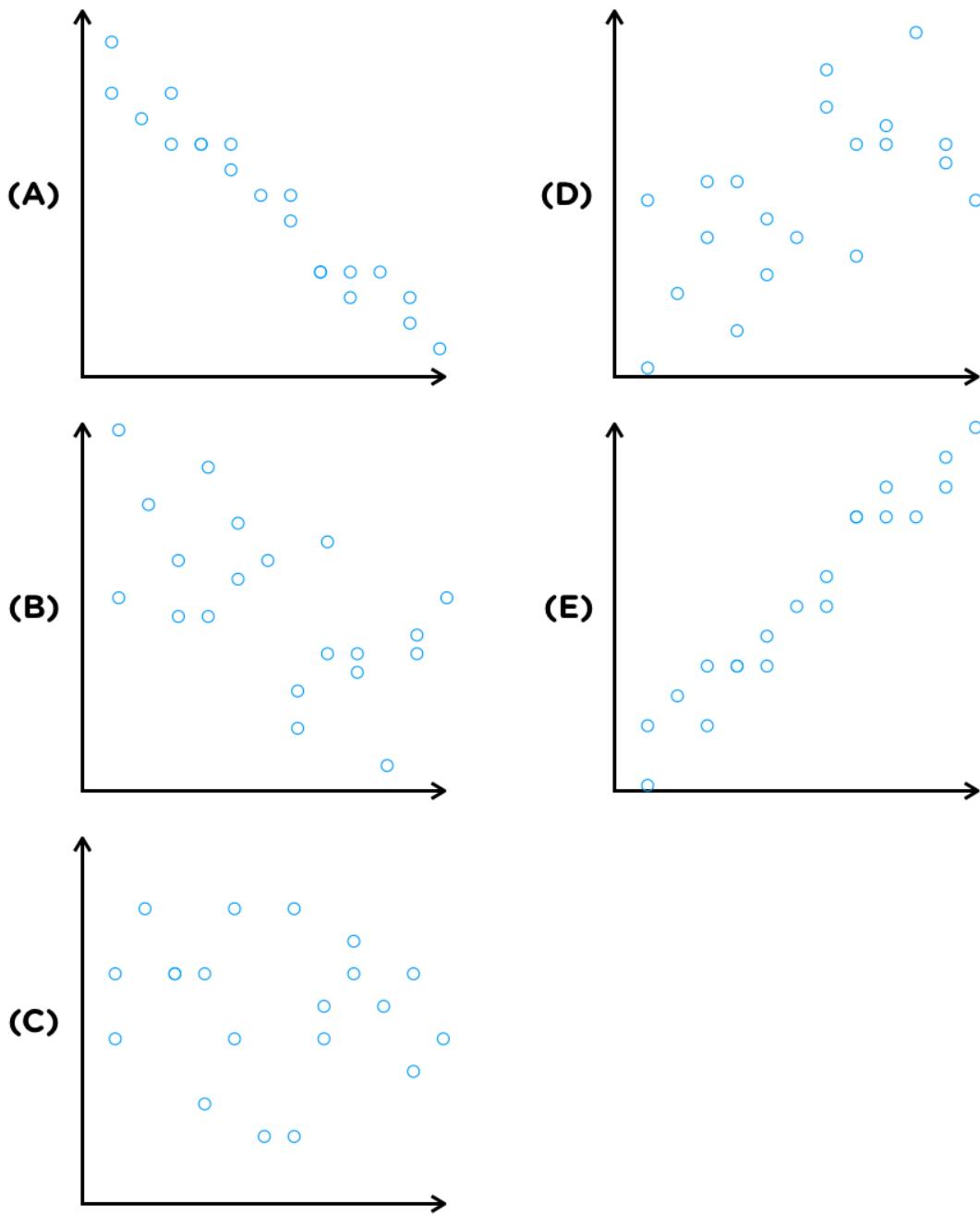
Notice how the values coincide with the scatterplots. For example, the sample correlation between Departure and Temp is -0.266, suggesting there is a weak negative linear relationship. This agrees with the scatterplot, which does not have much of a pattern.

On the other hand, the sample correlation between Temp and Precip Chance is -0.785, suggesting there is a substantial negative linear relationship. This number agrees with the prominent downward pattern of the scatterplot. However, recall that we noted the faint quadratic pattern. This is not captured by correlation which only measures linearity. In this case, the quadratic shape is not very different from a linear one, so the sample correlation is decently close to -1.

Example 3.1.5.1

The sample correlation of two variables is 0.6271.

Determine which is most likely the scatterplot of these two variables.



Solution

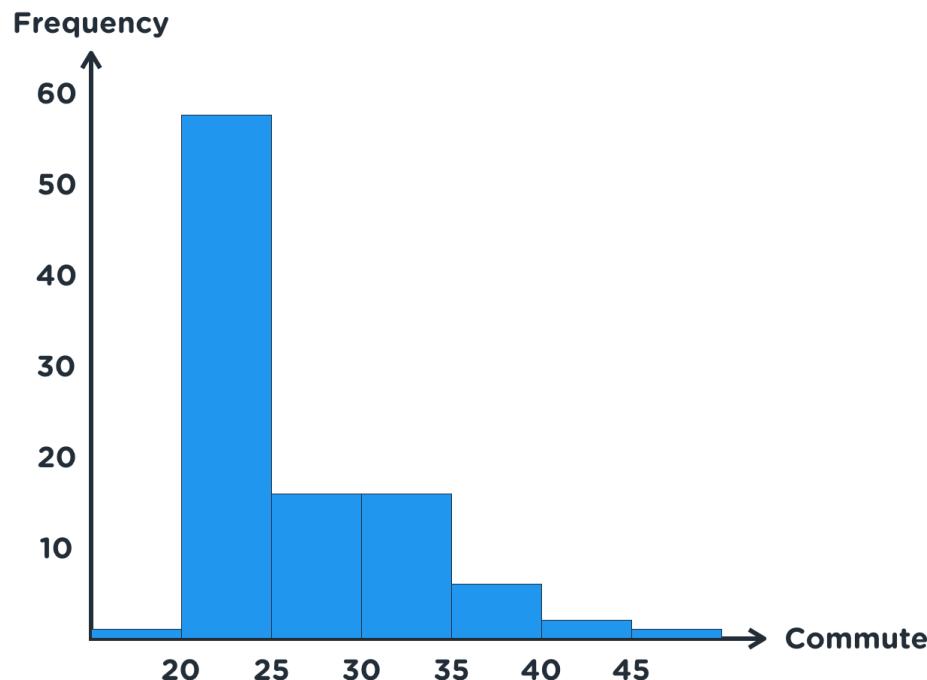
The sample correlation is positive, which eliminates options (A) and (B) since they have clear negative slopes.

Option (C) corresponds to a sample correlation very close to 0, while option (E) corresponds to a sample correlation very close to 1.

Therefore, the answer is **(D)**.

3.1.6 Box Plots

A histogram is a common way to depict the distribution of a quantitative variable. Here is the histogram of Commute:

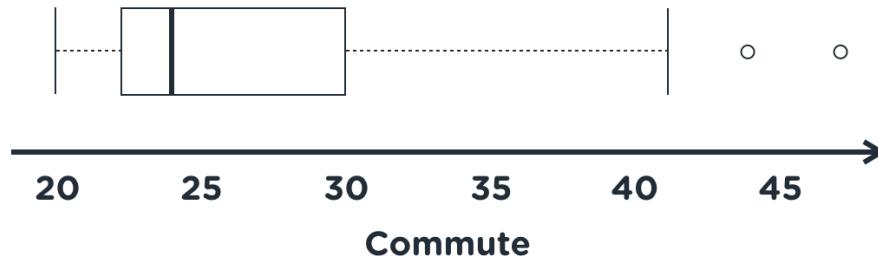


This shows a **right-skewed** distribution, as there are a few unusually large values of Commute in contrast to most of the values that are smaller.

A **box plot** has a similar purpose in capturing the distribution of a variable. This is accomplished by focusing on a few characteristics of a distribution:

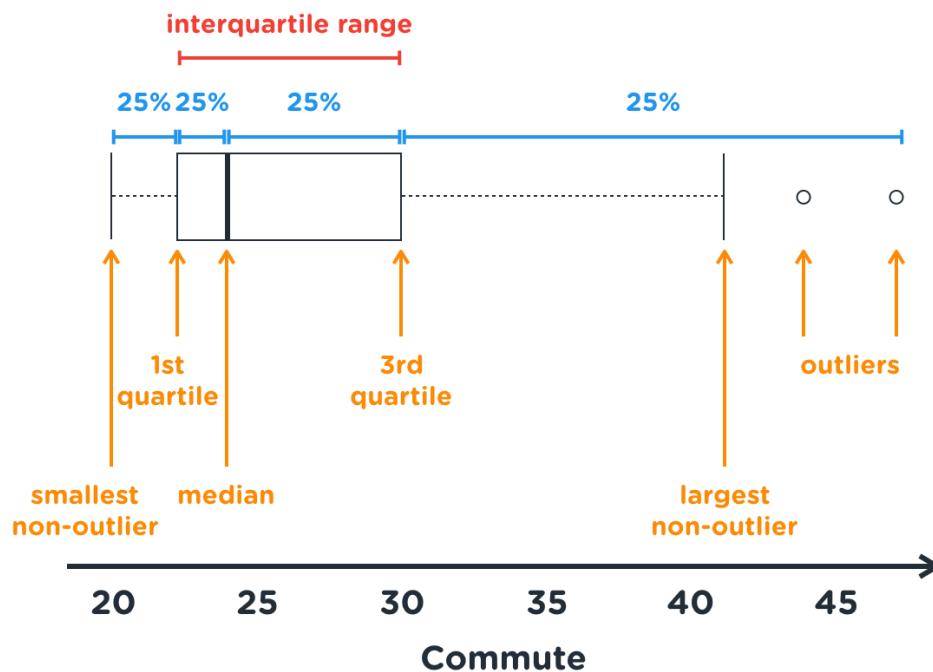
- the median
- the first and third quartiles
- the distribution tails

Here is the box plot of Commute:



The left and right edges of the box are the first and third quartiles, respectively. The line inside of the box is the median. Therefore, the box spans the interquartile range; 50% of the values are depicted in the range of the box alone. The range to the left of the box covers another 25%, as does the range to the right of the box.

The lines emerging from the box are called whiskers. The end of the left whisker is the smallest value of the variable that is not an outlier; the end of the right whisker is the largest value of the variable that is not an outlier. This means the two circles on the far right are outliers.



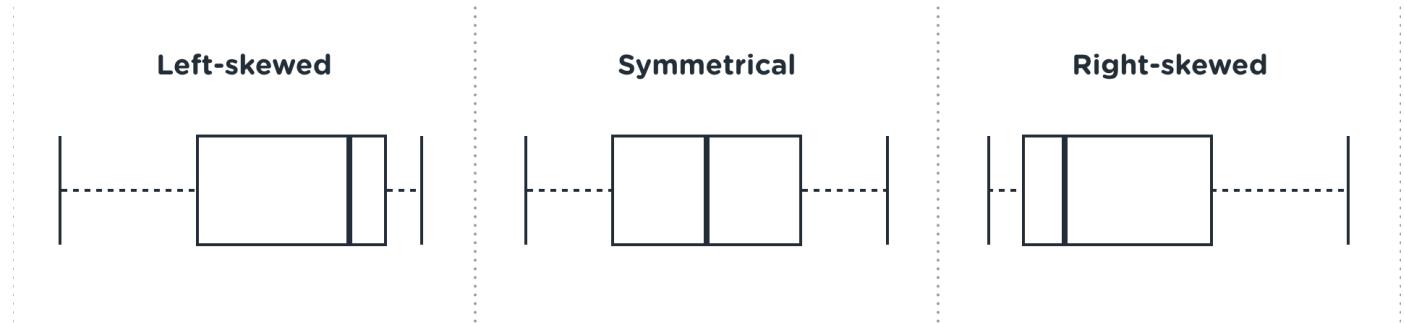
Coach's Remarks

With respect to a box plot, an outlier is a value that is outside of the following interval:

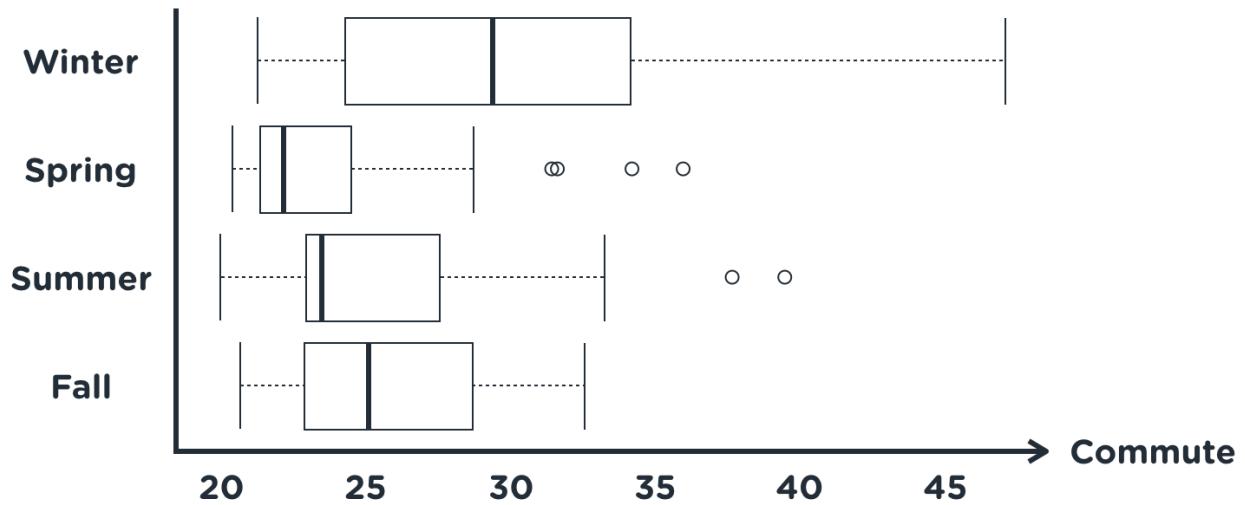
$$(1^{\text{st}} \text{ quartile} - 1.5 \times \text{interquartile range}, 3^{\text{rd}} \text{ quartile} + 1.5 \times \text{interquartile range})$$

However, we do not believe this is important for the exam. Moreover, there is in fact no concrete definition for an outlier in statistics.

In general, distributions that are left-skewed, symmetrical, and right-skewed will have the following box plots shapes:



Box plots can be drawn horizontally or vertically. In addition, drawing box plots side-by-side allows for easy comparison between variable distributions. For example, the following are box plots of Commute separated by the four calendar seasons:



Two key things that we can infer from the box plots are:

- During the winter, the commute times fluctuate a lot and tend to be longer.

- Springtime commutes are highly right-skewed. Half of the commutes take less than 22.13 minutes, which is shorter than the first quartiles of the other three seasons.

3.1.7 QQ Plots

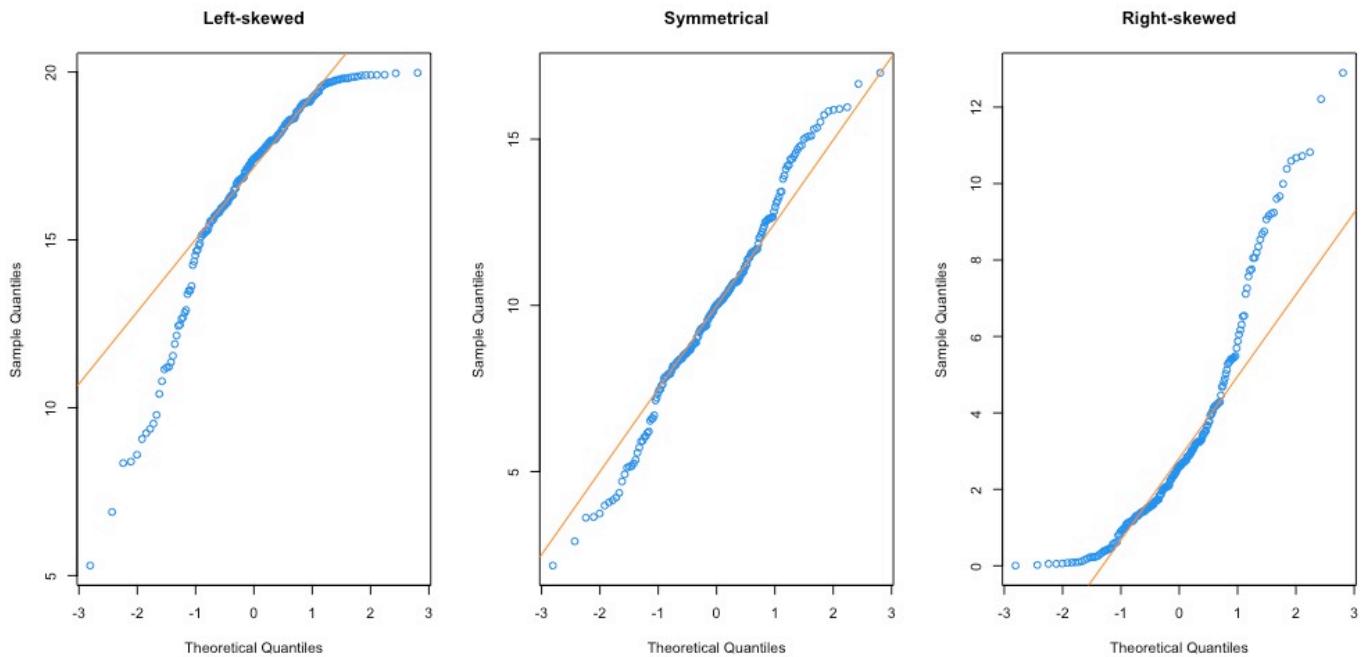
To identify whether a variable is distributed similarly to a certain theoretical distribution, we may use a **QQ plot**, also called a **quantile-quantile plot**. A quantile is simply a percentile viewed from a different scale, much like a quartile.

Here are the steps to construct the plot:

1. Using an approach to calculate sample percentiles, identify the q values for each of the n percentiles (i.e. values) of the variable.
2. Determine the percentiles of the theoretical distribution that correspond to the q values.
3. Plot the variable percentiles against the theoretical percentiles.
4. Draw a straight line through the 25th and 75th percentiles.

If the theoretical distribution is normal, then the QQ plot is more precisely referred to as a **normal QQ plot**.

In general, distributions that are left-skewed, symmetrical, and right-skewed would have the following normal QQ plots shapes:



Consider this simple demonstration.

You observe the following six values of variable y :

6 13 9 21 14 5

Sample percentiles are based on the following formula:

$$q = \frac{i - 0.5}{n}$$

such that the $100q^{\text{th}}$ sample percentile is the i^{th} observation in ascending order, and n is the number of observations.

Compare y to the standard normal distribution using a QQ plot.

First, sort the values of y in ascending order. For each value, determine its q using the formula given.

For example, since 5 is the 1st observation in ascending order, its q is $\frac{1-0.5}{6} = 0.0833$. This means the 8.33th sample percentile is 5. Here are the six q 's with their percentiles.

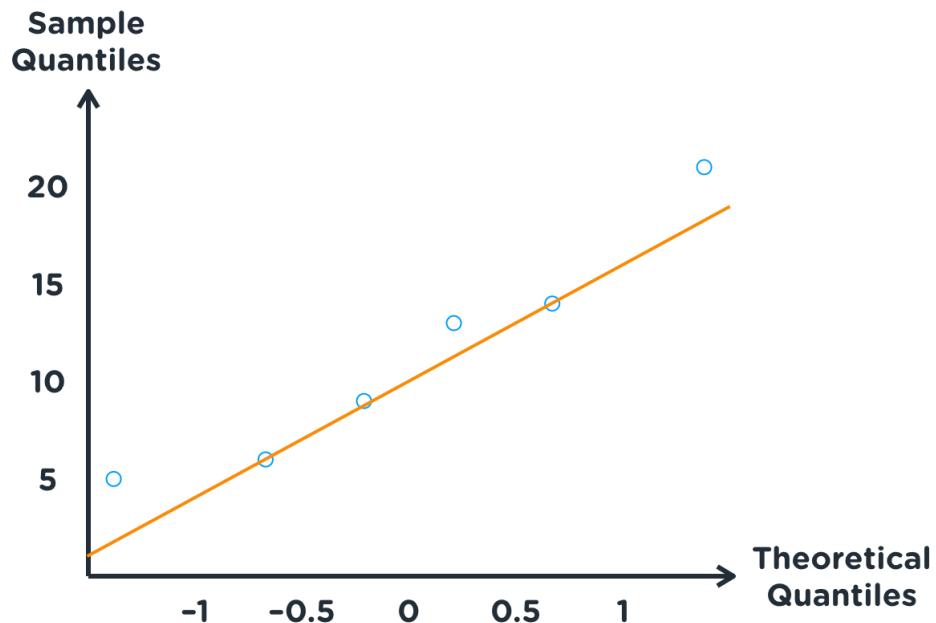
q	Percentile of y
0.0833	5
0.2500	6
0.4167	9
0.5833	13
0.7500	14
0.9167	21

Next, determine the 8.33th, 25th, 41.67th, 58.33th, 75th, and 91.67th percentiles of the standard normal distribution by using the exam table.

	Percentile of	Percentile of
0.0833	5	-1.38
0.2500	6	-0.67
0.4167	9	-0.21
0.5833	13	0.21
0.7500	14	0.67

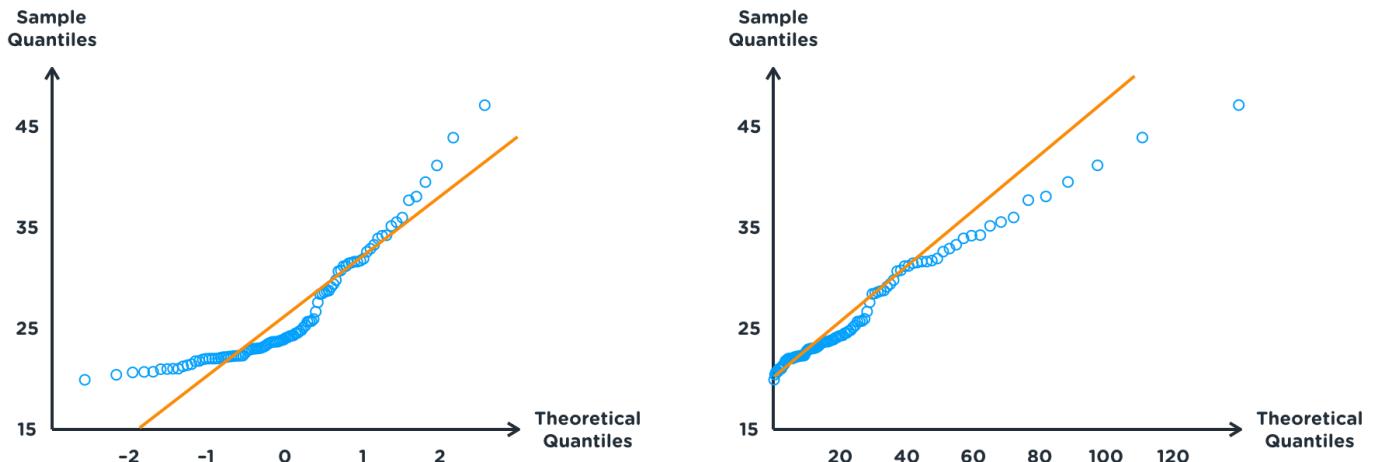
q	Percentile of y	Percentile of Z
0.9167	21	1.38

The QQ plot below has the percentiles of y for the vertical axis and the standard normal percentiles for the horizontal axis. The line drawn intersects the coordinates $(-0.67, 6)$ and $(0.67, 14)$, which represent the 25th and 75th percentiles, respectively.



If the majority of the points follow the superimposed line, we conclude that the variable's distribution and the theoretical distribution have similar shapes. If the majority of the points deviate significantly from the superimposed line, we believe that the two distributions have different shapes.

With only six points in the above QQ plot, it is challenging to comment on the pattern. Instead, let's analyze two other QQ plots involving Commute.



In the left plot, Commute is compared to the standard normal distribution. Knowing that the Commute values are right-skewed, it is not surprising that many points do not follow the line.

In the right plot, Commute is compared to an exponential distribution with mean 26.39 (i.e. the sample mean of Commute). Initially, many of the points follow the line, suggesting that the distribution of Commute is shaped like the exponential distribution. However, the points eventually deviate further right of the plot. This indicates dissimilar right-tail behavior between Commute and the exponential distribution.

3.1 Summary

5m

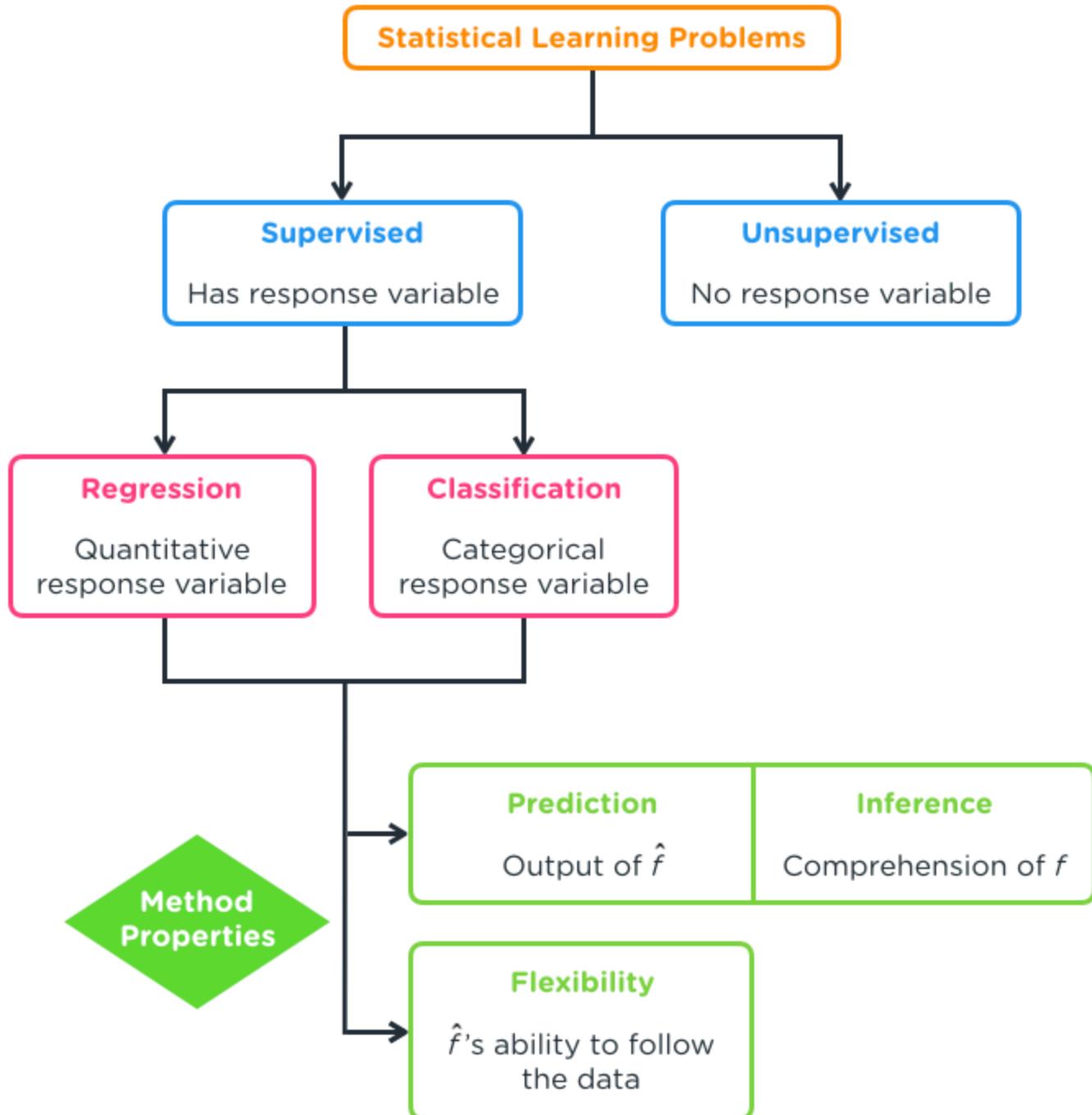
Variable Types

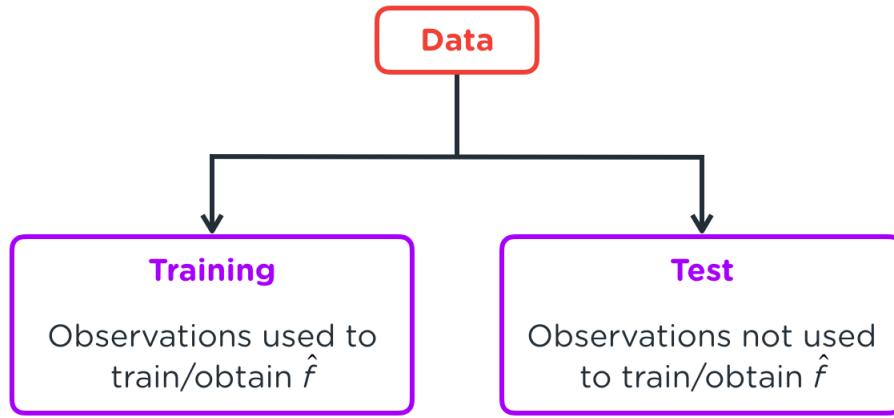
Variable	Description
Response	A variable of primary interest
Explanatory	A variable used to study the response variable
Count	A quantitative variable usually valid on non-negative integers
Continuous	A real-valued quantitative variable
Nominal	A categorical/qualitative variable having categories without a meaningful or logical order
Ordinal	A categorical/qualitative variable having categories with a meaningful or logical order

Notation

Symbol	Concept
y, Y	Response variable
x, X	Explanatory variable
Subscript i	Index for observations
n	Number of observations
Subscript j	Index for variables except response
p	Number of variables except response
\mathbf{A}^T	Transpose of matrix \mathbf{A}
$f(x)$	$E[Y]$
ε	Error term
$\hat{y}, \hat{Y}, \hat{f}(x)$	Estimate/Estimator of $f(x)$

Statistical Learning Terminology





Model Accuracy

$$Y = f(x_1, \dots, x_p) + \varepsilon, \quad \mathbb{E}[\varepsilon] = 0$$

$$\text{Test MSE} = \mathbb{E} \left[(Y - \hat{Y})^2 \right]$$

which can be estimated with a large number of observations using

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

If training data y_i 's are used, it computes the training MSE.

For fixed inputs x_1, \dots, x_p , the test MSE can be written as

$$\underbrace{\text{Var}[\hat{f}(x_1, \dots, x_p)]}_{\text{reducible error}} + \underbrace{\left(\text{Bias}[\hat{f}(x_1, \dots, x_p)] \right)^2}_{\text{irreducible error}} + \underbrace{\text{Var}[\varepsilon]}_{\text{irreducible error}}$$

The bias-variance trade-off reveals why flexibility and accuracy are not synonymous. Specifically:

- As flexibility increases, the training MSE decreases, but the test MSE follows a u-shaped pattern.
- Low flexibility leads to a method with low variance and high bias; high flexibility leads to a method with high variance and low bias.

Numerical Summaries

UNIVARIATE

- Sample mean, \bar{x}
- Sample variance, s_x^2
- Sample percentiles
- Interquartile range = 3rd quartile – 1st quartile

SAMPLE COVARIANCE

$$cov_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

SAMPLE CORRELATION

$$r_{x,y} = \frac{cov_{x,y}}{s_x \cdot s_y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}, \quad -1 \leq r_{x,y} \leq 1$$

Correlation measures the linear strength between two variables.

Graphical Summaries

SCATTERPLOT

Plots values of two variables to investigate their relationship.

BOX PLOT

Captures a variable's distribution using its median, 1st and 3rd quartiles, and distribution tails.

QQ PLOT

Plots sample percentiles against theoretical percentiles to determine whether the sample and theoretical distributions have similar shapes.

3.2.0 Overview

 5m

We begin exploring statistical learning methods that are collectively referred to as linear regression. In particular, **simple linear regression** uses one explanatory variable to predict a quantitative response variable. This topic introduces many foundational concepts that recur in linear regression, as well as other areas of statistical learning. The concepts introduced here include:

- Ordinary least squares
- Residuals
- Partitioning of variability

We also discuss how to interpret the outputs from running a regression, such as the parameter estimates, the residual standard error, and R^2 . Furthermore, additional insights are gained from hypothesis tests and confidence intervals.

3.2.1 Main Idea and Assumptions

🕒 5m

In simple linear regression, the relationship between the response and explanatory variables takes the form of

$$Y = \beta_0 + \beta_1 x + \varepsilon \quad (3.2.1.1)$$

In light of Equation 3.1.2.1, this means $\beta_0 + \beta_1 x$ is the chosen functional form for f .

β_0 and β_1 are free parameters. Their estimates are found using training data. Since the chosen functional form is a **linear equation** of x , we refer to

- β_0 as the *intercept parameter*, and
- β_1 as the *slope parameter*.

Here are the assumptions of a simple linear regression model:

1. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
2. x_i 's are non-random
3. $E[\varepsilon_i] = 0$
4. $\text{Var}[\varepsilon_i] = \sigma^2$
5. ε_i 's are independent
6. ε_i 's are normally distributed

These assumptions lead to other important conclusions:

- Y_i 's are independent normal random variables
- $E[Y_i] = \beta_0 + \beta_1 x_i$
- $\text{Var}[Y_i] = \sigma^2$

σ^2 is another parameter that requires estimation. Notice that the variance of Y_i (or similarly, ε_i) is σ^2 **without** a subscript. Thus, the variance is constant for all observations. This property is referred to as **homoscedasticity**.

In summary, this model proposes that the relationship between the response and explanatory variables is systematically linear in nature. As a result, an explanatory variable that is strongly correlated (i.e. forms a strong **linear** pattern) with the response variable is a good candidate for simple linear regression.

Example 3.2.1.1

Determine which statements are true regarding a simple linear regression model with a response variable Y and an explanatory variable x .

- I. The model equation is $Y = \beta_0 + \beta_1 x$.
- II. The mean of the response is modeled to vary based on the observations.
- III. The variance of the response is modeled to vary based on the observations.
- IV. A correlation of -0.9 between the response and explanatory values supports the use of the model.

Solution

I is false because the model equation is $Y = \beta_0 + \beta_1 x + \varepsilon$. It is $E[Y]$ that equals $\beta_0 + \beta_1 x$.

II is true because the mean of Y_i depends on x_i .

III is false because the model assumes homoscedasticity. This means the variance of the response is constant for all observations.

IV is true because a correlation of -0.9 is close to -1, suggesting a strong negative linear relationship between the response and explanatory variables.

Therefore, **only II and IV are true.**



Coach's Remarks

If the response and explanatory variables do not have a strong linear relationship, one option is to consider transforming either or both of the variables. For example, assume that $g(y)$ and x

exhibit a strong correlation for some function $g(\cdot)$. Then, it makes sense to have $g(Y)$ rather than Y in Equation 3.2.1.1.

In doing so, the regression will examine the behavior of $g(Y)$, **not** the original response. The regression results would require some tweaks in order to study the original response.

3.2.2 Parameter Estimates

As described in the previous subsection, there are three parameters that require estimation: β_0 , β_1 , and σ^2 .

Estimation of β_0 and β_1

Let

- $\hat{\beta}_0$ denote the estimate of β_0
- $\hat{\beta}_1$ denote the estimate of β_1
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are determined using *ordinary least squares*, which minimizes the expression

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

using training data. Hence, ordinary least squares is an optimization problem. We find the partial derivatives of the expression with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, then set them to equal 0. Solving the equations produces

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.2.2.1)$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (3.2.2.2)$$

If interested in the proof, see the appendix at the end of the section.

Coach's Remarks

If all the data points of the response and explanatory variables are available, the TI-30XS Multiview calculator can compute $\hat{\beta}_0$, $\hat{\beta}_1$, and $r_{x,y}$ (sample correlation - Equation 3.1.4.2) using the data function. Here are the steps:

1. Access the data table: stat data .
2. Enter the explanatory variable data in "L1" and the response variable data in "L2".
3. Access the stat functionality: 2nd stat .
4. Select "2-Var Stats".
5. Select "L1" for "xDATA" and "L2" for "yDATA".
6. On the results screen, scroll to
 - "a" for $\hat{\beta}_1$
 - "b" for $\hat{\beta}_0$
 - "r" for $r_{x,y}$

To demonstrate the formulas, consider performing a simple linear regression with data from the Commuting Chris scenario:

i	Commute, y_i	Precip Chance, x_i
1	24.283	48.0
\vdots	\vdots	\vdots
100	32.567	98.4

Their sample means are:

- $\bar{y} = 26.3875$
- $\bar{x} = 37.546$

As a result,

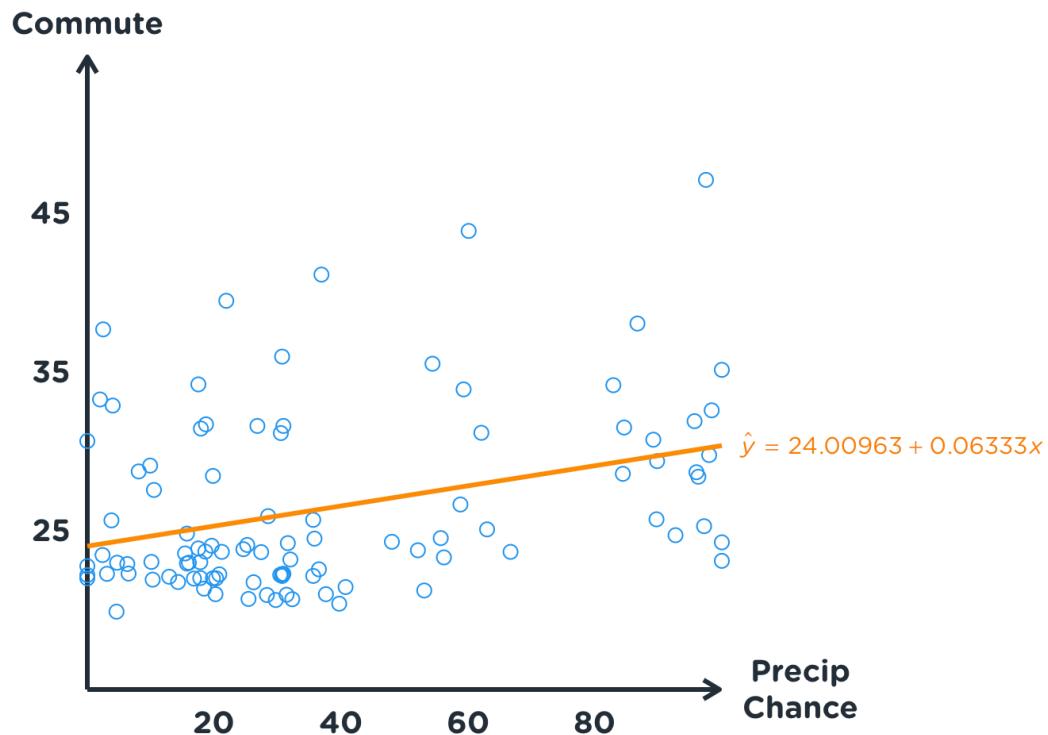
$$\begin{aligned}\hat{\beta}_1 &= \frac{(48 - 37.546)(24.283 - 26.3875) + \dots + (98.4 - 37.546)(32.567 - 26.3875)}{(48 - 37.546)^2 + \dots + (98.4 - 37.546)^2} \\ &= \frac{5,874.448}{92,756.31} \\ &= 0.06333\end{aligned}$$

$$\hat{\beta}_0 = 26.3875 - 0.06333(37.546) = 24.00963$$

Altogether, the regression line fitted to the data has the equation of

$$\hat{y} = 24.00963 + 0.06333x$$

The plot below shows the fitted line superimposed on a scatterplot of Commute against Precip Chance.



It is crucial to know how to interpret these results.

- For every increase of 100 basis points in the chance of precipitation, the expected commute time increases by 0.06333 minutes (or about 3.8 seconds).

- In general, $\hat{\beta}_1$ is the change in \hat{y} for every unit increase in x .
- 24.00963 minutes is the expected commute time when there is a 0% chance of precipitation.
- In general, $\hat{\beta}_0$ is \hat{y} when $x = 0$.

Coach's Remarks

Notice the interpretations mention "**expected** commute time" instead of just "commute time". This emphasizes the distinction between \hat{y} and y . Recall that \hat{y} is the estimated **mean** response in a regression setting. In addition, the words "**predicted**" and "**fitted**" refer to \hat{y} as well.

Furthermore, if a transformed response was modeled instead, we can predict the original response by first finding $\hat{g}(y)$ from the fitted equation, then reversing the transformation.

By running this regression in the statistical software R, we obtain the following output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.00963   0.85011  28.243 < 2e-16 ***
Precip.Chance 0.06333   0.01758   3.602 0.000499 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.355 on 98 degrees of freedom
Multiple R-squared:  0.1169, Adjusted R-squared:  0.1079 
F-statistic: 12.97 on 1 and 98 DF,  p-value: 0.0004986
```

Notice that the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are found in the table-like structure called **Coefficients** under the column header **Estimate**.

- The row **(Intercept)** lists information related to β_0 . Thus, the estimate of β_0 (i.e. $\hat{\beta}_0$) is 24.00963.
- The row **Precip.Chance** lists information related to β_1 since this is the coefficient of Precip Chance, x . Thus, the estimate of β_1 (i.e. $\hat{\beta}_1$) is 0.06333.

The rest of the output is addressed in later parts of the manual.

Example 3.2.2.1

For 30 observations of a response variable y and an explanatory variable x , you are given the following summary statistics:

- The unbiased sample covariance of x and y is -41.759.
- The sum of squared x values is 1,502.18.
- The sample mean of x is 6.33.

For a simple linear regression, calculate the estimate of the slope parameter.

Solution

The slope parameter is β_1 , and its estimate has the formula

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{30} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{30} (x_i - \bar{x})^2}$$

Use the formula for the unbiased sample covariance to solve for the numerator of $\hat{\beta}_1$.

$$cov_{x,y} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1}$$

$$\begin{aligned} \Rightarrow \sum_{i=1}^{30} (x_i - \bar{x})(y_i - \bar{y}) &= (30 - 1)cov_{x,y} \\ &= 29(-41.759) \\ &= -1,211.011 \end{aligned}$$

Next, expand and rewrite the denominator of $\hat{\beta}_1$.

$$\begin{aligned}
\sum_{i=1}^{30} (x_i - \bar{x})^2 &= \sum_{i=1}^{30} x_i^2 - 2\bar{x} \left(\sum_{i=1}^{30} x_i \right) + 30\bar{x}^2 \\
&= \sum_{i=1}^{30} x_i^2 - 2\bar{x} (30\bar{x}) + 30\bar{x}^2 \\
&= \sum_{i=1}^{30} x_i^2 - 60\bar{x}^2 + 30\bar{x}^2 \\
&= \sum_{i=1}^{30} x_i^2 - 30\bar{x}^2 \\
&= 1,502.18 - 30 (6.33^2)
\end{aligned}$$

Altogether, the answer is

$$\hat{\beta}_1 = \frac{-1,211.011}{1,502.18 - 30 (6.33^2)} = -4.035$$



Coach's Remarks

You may choose to memorize that

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\
&= \sum_{i=1}^n x_i^2 - \frac{(\sum_{i=1}^n x_i)^2}{n}
\end{aligned}$$

Example 3.2.2.2

With a sample of university students, an analyst investigates how the number of years of

enrollment relates to the time spent to complete a typical homework assignment. The following R output is the result of the analysis:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.5252    0.5387 43.672 <2e-16 ***
Years.enrolled 3.9485    0.4230  9.334 5e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.798 on 72 degrees of freedom
Multiple R-squared:  0.5475, Adjusted R-squared:  0.5413
F-statistic: 87.13 on 1 and 72 DF,  p-value: 4.999e-14
```

Predict the time spent to complete a typical homework assignment by a student who has been enrolled for 3 years.

Solution

Based on the R output, the prediction formula is

$$\hat{y} = 23.5252 + 3.9485x$$

where x is the number of years of enrollment. Thus, we seek the value of \hat{y} when $x = 3$, which is

$$\hat{y} = 23.5252 + 3.9485 (3) = \mathbf{35.3707}$$



Example 3.2.2.3

In studying the value of secondhand laptops, a regression produces the prediction formula of

$$\hat{y} = 395.174 - 9.468x$$

where

- \hat{y} is the predicted laptop price, and
- x is the laptop age in months.

Determine which statement correctly interprets the slope estimate.

- 395.174 is the predicted price of a laptop that is brand new.
- 395.174 is the predicted laptop age in months when its price is 0.
- For every one month increase in a laptop's age, its predicted price increases by 9.468.
- For every one month increase in a laptop's age, its predicted price decreases by 9.468.
- For every one month increase in a laptop's age, its predicted price decreases by -9.468.

Solution

The slope estimate is -9.468, which eliminates options (A) and (B).

Option (C) is incorrect because a negative slope estimate means the predicted price decreases as age increases.

Option (E) is incorrect because decreasing by a negative value is the same as increasing. This means options (C) and (E) are synonymous.

Therefore, the answer is (D). ■

Coach's Remarks

Option (A) is the correct interpretation of the intercept estimate.

Estimation of σ^2

The estimate of σ^2 is called the **mean squared error (MSE)**, which is calculated as

$$\text{MSE} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} \quad (3.2.2.3)$$

Coach's Remarks

While not identical, the MSE introduced in Section 3.1.3 and the MSE referenced here are much more similar than different. An obvious difference is that Equation 3.1.3.2 has a denominator of n rather than $n - 2$. Keep the following in mind:

- Here, the focus is to use an **unbiased** estimator of σ^2 . This requires a denominator of $n - 2$ because the numerator uses the estimates of β_0 and β_1 . We expand on this in Section 3.2.6.
- In Section 3.1.3, the focus was on estimating a measure of model accuracy. The use of n was sufficient for that purpose.

Recall the data and the prediction formula from the Commuting Chris setup involving Commute and Precip Chance:

i	Commute, y_i	Precip Chance, x_i
1	24.283	48.0
\vdots	\vdots	\vdots
100	32.567	98.4

$$\hat{y} = 24.00963 + 0.06333x$$

First, obtain

- $\hat{y}_1 = 24.00963 + 0.06333(48) = 27.049$
- \dots
- $\hat{y}_{100} = 24.00963 + 0.06333(98.4) = 30.241$

Using Equation 3.2.2.3 produces

$$\text{MSE} = \frac{(24.283 - 27.049)^2 + \dots + (32.567 - 30.241)^2}{100 - 2} = 28.68$$

Since the MSE estimates σ^2 , the square root of the MSE is treated as the estimate of σ . The square root of the MSE is referred to as the **residual standard error**.

$$\text{residual standard error} = \sqrt{\text{MSE}} \quad (3.2.2.4)$$

Thus, the residual standard error for this Commuting Chris setup is $\sqrt{28.68} = 5.355$, which can be found in the output from R.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.00963   0.85011  28.243 < 2e-16 ***
Precip.Chance 0.06333   0.01758   3.602 0.000499 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.355 on 98 degrees of freedom
Multiple R-squared:  0.1169, Adjusted R-squared:  0.1079
F-statistic: 12.97 on 1 and 98 DF,  p-value: 0.0004986
```

3.2.3 Matrix Notation

The model equation in relation to the training data is

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1, \dots, n$$

Another way to represent the same information is to use matrix notation.

$$\begin{aligned} \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} &= \begin{bmatrix} \beta_0 + \beta_1 x_1 + \varepsilon_1 \\ \beta_0 + \beta_1 x_2 + \varepsilon_2 \\ \vdots \\ \beta_0 + \beta_1 x_n + \varepsilon_n \end{bmatrix} \\ &= \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \vdots \\ \beta_0 + \beta_1 x_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ &= \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix} \\ \Rightarrow \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \end{aligned}$$

For a refresher on matrix multiplication, revisit Section 1.6.2.

\mathbf{X} is known as the **design matrix**. By noting that $\beta_0 + \beta_1 x_i = \beta_0(1) + \beta_1(x_i)$, it is clear that the first column of \mathbf{X} corresponds to the first element of the vector $\boldsymbol{\beta}$, and the second column of \mathbf{X} corresponds to the second element of $\boldsymbol{\beta}$.

Matrix notation is also useful in expressing the estimates $\hat{\beta}_0$ and $\hat{\beta}_1$. By defining two vectors $\hat{\boldsymbol{\beta}}$ and \mathbf{y} as

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

we can write

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \quad (3.2.3.1)$$

Its proof is not required for the exam. Furthermore,

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \vdots \\ \hat{y}_n \end{bmatrix} = \hat{\mathbf{y}} = \mathbf{X} \hat{\beta} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Then, by defining a *projection matrix* or *hat matrix* \mathbf{H} as

$$\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \quad (3.2.3.2)$$

we get the formula

$$\hat{\mathbf{y}} = \mathbf{H} \mathbf{y} \quad (3.2.3.3)$$

\mathbf{H} is called a hat matrix because the fitted values of the response can be found via multiplying \mathbf{H} by the vector of actual responses.

Example 3.2.3.1

A linear regression is given by $\mathbf{Y} = \mathbf{X} \beta + \epsilon$, where $\beta^T = [\beta_0 \quad \beta_1]$.

You are given the following matrices:

- $(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{75} \begin{bmatrix} 109 & -19 \\ -19 & 4 \end{bmatrix}$
- $\mathbf{X}^T \mathbf{y} = \begin{bmatrix} 45 \\ 192 \end{bmatrix}$

Calculate the ordinary least squares estimate of β_0 .

Solution

The goal is $\hat{\beta}_0$. Note that

$$\begin{aligned}
 \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} &= \hat{\beta} \\
 &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\
 &= \frac{1}{75} \begin{bmatrix} 109 & -19 \\ -19 & 4 \end{bmatrix} \begin{bmatrix} 45 \\ 192 \end{bmatrix} \\
 &= \frac{1}{75} \begin{bmatrix} (109)(45) + (-19)(192) \\ (-19)(45) + (4)(192) \end{bmatrix} \\
 &= \frac{1}{75} \begin{bmatrix} 1,257 \\ -87 \end{bmatrix} \\
 &= \begin{bmatrix} 16.76 \\ -1.16 \end{bmatrix}
 \end{aligned}$$

Therefore, the answer is the first element of $\hat{\beta}$, which is **16.76**.



Coach's Remarks

This problem is based on the following data:

y	x
10	5
8	8
15	2
12	4

You are encouraged to replicate the results using Equations 3.2.2.1 and 3.2.2.2 or with the TI-30XS MultiView calculator. You may also benefit from performing the matrix calculation yourself to match the given values.

3.2.4 Other Numerical Results

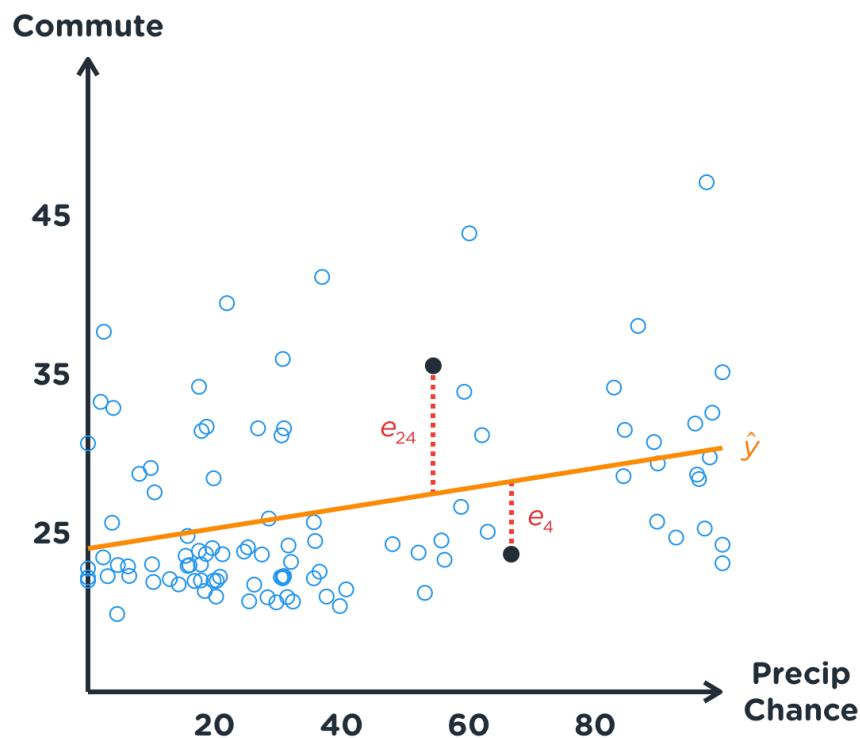
Residuals

The i^{th} **residual** is calculated as

$$e_i = y_i - \hat{y}_i \quad (3.2.4.1)$$

If \hat{f} is correctly specified, then the e_i 's are realizations of the ε_i 's. This idea is further developed in Section 3.5.

One way to visualize residuals is to revisit the scatterplot of Commute against Precip Chance.



The 4th observation is represented by the black dot below the fitted line, while the 24th observation is represented by the black dot above the fitted line.

i	Commute, y_i	Precip Chance, x_i
4	23.650	66.7
24	35.500	54.4

Consequently,

$$e_4 = 23.65 - (24.00963 + 0.06333 \cdot 66.7) = -4.584$$

$$e_{24} = 35.5 - (24.00963 + 0.06333 \cdot 54.4) = 8.045$$

A **positive** residual signifies that the actual observation is **larger** than the prediction, whereas a **negative** residual signifies that the actual observation is **smaller** than the prediction.

There are important connections between residuals and some concepts we have previously discussed. Recall that ordinary least squares seeks to minimize the expression

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

in order to estimate β_0 and β_1 . Equivalently, ordinary least squares is interested in minimizing the sum of the squared residuals.

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

Another connection is that the numerator of the MSE formula (i.e. Equation 3.2.2.3) is the same expression that ordinary least squares desires to minimize. Thus,

$$\text{MSE} = \frac{\sum_{i=1}^n e_i^2}{n - 2}$$

This implies that ordinary least squares results in the "best" linear fit, in that the MSE is minimized.

Example 3.2.4.1

Determine which statement is true regarding residuals.

- I. A residual is the predicted response minus the actual response for a certain observation.

II. A good model is characterized by having a significant amount of large residuals.

III. A positive residual indicates that the corresponding prediction is under-valued.

Solution

I is false because the formula is "actual minus predicted".

II is false because large residuals result from the actual responses being substantially different from the predicted responses. In contrast, ordinary least squares finds the best β estimates by minimizing the sum of squared residuals.

III is true because a residual is positive when the predicted response is lower than the actual response.

Therefore, **only III is true.**



Sum of Squares

To judge the usefulness of a linear regression model in a more concrete way, we partition the variability found in the response variable. Variability is measured by the sum of certain squared quantities. To make sense of this, we begin with the *null model*, which is described by the equation

$$Y = \beta_0 + \varepsilon$$

The null model assumes that there is no relationship between the response and any explanatory variable. This is a univariate analysis at its core, which means we estimate

- the mean response (i.e. β_0 in this case) as the sample mean, \bar{y} , and
- σ^2 as the unbiased sample variance, s_y^2 .

As a reminder,

$$s_y^2 = \frac{\sum_{i=1}^n (y_i - \bar{y})^2}{n - 1}$$

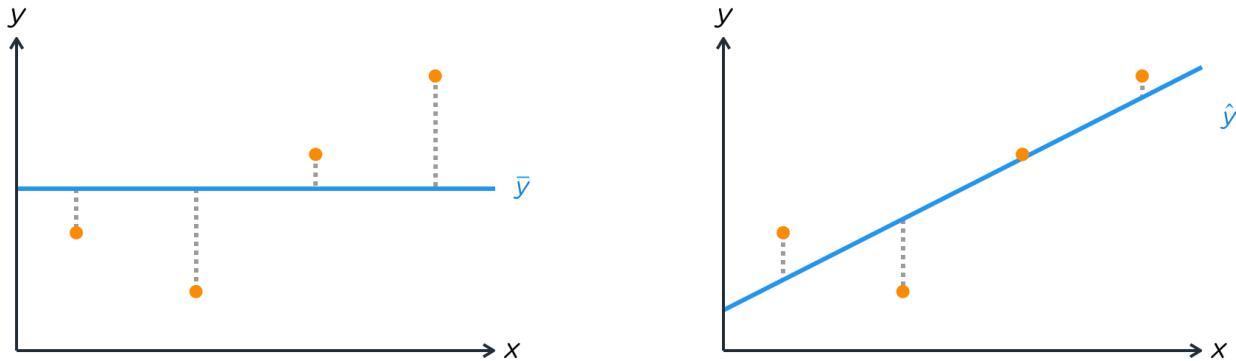
The numerator, $\sum_{i=1}^n (y_i - \bar{y})^2$, is called the **total sum of squares (SST)** because it represents the entire amount of y 's variability around its sample mean, \bar{y} .

There are many possible theories as to how y varies. One of them is that y varies due to an underlying relationship with an explanatory variable, x . So we consider upgrading from the null model to simple linear regression. This upgrade proposes to more intelligently estimate

- the mean response as \hat{y} , and
- σ^2 as the MSE.

The numerator of the MSE, $\sum_{i=1}^n (y_i - \hat{y}_i)^2$, is called the **error sum of squares (SSE)** because it represents the amount of variability in y that remains after accounting for the variability due to x . In other words, SSE is the amount of variability in y that remains **unexplained** by the simple linear regression.

We may visualize SST and SSE by the following plots of the same four observations:



In the left plot, each dashed line represents one $y_i - \bar{y}$, and these collectively depict SST. In the right plot, each dashed line represents one $y_i - \hat{y}_i = e_i$, and these collectively depict SSE. Therefore, the reduction in variability from SST to SSE captures how much of an improvement the simple linear regression is over the null model.

It can be proven that

$$\text{SST} - \text{SSE} = \sum_{i=1}^n (y_i - \bar{y})^2 - \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 \quad (3.2.4.2)$$

$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ is the **regression sum of squares (SSR)**, which represents the amount of variability in y that is **explained** by the simple linear regression.

In summary,

Sum of Squares	Math Notation	Type of Variability
SST	$\sum_{i=1}^n (y_i - \bar{y})^2$	Total variability
SSR	$\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	Explained variability
SSE	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$	Unexplained variability

$$\text{SST} = \text{SSR} + \text{SSE}$$

Coach's Remarks

Recall that SSE is also referred to as the "sum of squared residuals". Some resources have SSE as the "**residual** sum of squares" and abbreviate to RSS instead. Mistaking RSS for SSR can be a source of confusion.

Coefficient of Determination

The **coefficient of determination**, also called the **R^2 statistic**, is calculated as

$$R^2 = \frac{\text{SSR}}{\text{SST}} \tag{3.2.4.3}$$

This is the proportion of variability in the response that is explained by the simple linear regression. As a proportion, it must be between 0 and 1. Thus, it serves as a simple indicator of the regression's performance. A high R^2 indicates a regression that explains much of the variability in y , hence suggesting that the explanatory variable is highly informative.

Another way to interpret R^2 is through sample correlation. R^2 can be written as the squared sample correlation of y and \hat{y} . In the case of simple linear regression, \hat{y} is perfectly correlated to x (because \hat{y} is a linear function of x). Therefore, the linear strength between y and \hat{y} is identical to the linear strength between y and x . As a result,

$$R^2 = r_{x,y}^2 \quad (3.2.4.4)$$

Back to the Commuting Chris scenario, the R output gives an R^2 of 0.1169. This means only 11.69% of the variability in Commute is explained by this regression using Precip Chance. This regression is not a big improvement over the null model.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.00963   0.85011  28.243 < 2e-16 ***
Precip.Chance 0.06333   0.01758   3.602 0.000499 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.355 on 98 degrees of freedom
Multiple R-squared:  0.1169, Adjusted R-squared:  0.1079 
F-statistic: 12.97 on 1 and 98 DF,  p-value: 0.0004986
```

In addition, the sample correlation between Commute and Precip Chance is $\sqrt{0.1169} = 0.342$, as stated in Section 3.1.5 as well. We know to take the **positive** square root of R^2 since the slope estimate, $\hat{\beta}_1$, is positive.

Example 3.2.4.2

You are given the following information regarding a linear regression on 15 observations:

- The sum of squared residuals equals 2,867.55.
- The unbiased sample standard deviation of the response is 33.21.

Calculate R^2 .

Solution

The sum of squared residuals is SSE. Therefore, SSE equals 2,867.55.

Next, determine SST from the unbiased sample standard deviation.

$$s_y = \sqrt{\frac{\sum_{i=1}^{15} (y_i - \bar{y})^2}{15 - 1}} = \sqrt{\frac{\text{SST}}{14}} = 33.21$$

$$\Rightarrow SST = 14 \cdot 33.21^2 = 15,440.6574$$

Solve for the answer using Equations 3.2.4.2 and 3.2.4.3.

$$\begin{aligned} R^2 &= \frac{SSR}{SST} \\ &= \frac{SST - SSE}{SST} \\ &= 1 - \frac{SSE}{SST} \\ &= 1 - \frac{2,867.55}{15,440.6574} \\ &= \mathbf{0.814} \end{aligned}$$



Example 3.2.4.3

For a simple linear regression,

- The explanatory variable records the following values:

20.0 16.0 19.8 18.4 17.1

- $\hat{\beta}_1 = 4.57682$
- $R^2 = 0.91245$

Calculate the total sum of squares.

Solution

Given the value of R^2 , use Equation 3.2.4.3 to find the total sum of squares, SST.

$$\text{SST} = \frac{\text{SSR}}{R^2}$$

By rewriting the formula for SSR, we can solve it using the other details provided. First, we need to calculate $\sum_{i=1}^5 (x_i - \bar{x})^2$.

$$\bar{x} = \frac{20.0 + 16.0 + 19.8 + 18.4 + 17.1}{5} = 18.26$$

$$\sum_{i=1}^5 (x_i - \bar{x})^2 = (20.0 - 18.26)^2 + \dots + (17.1 - 18.26)^2 = 11.872$$

$$\begin{aligned}\text{SSR} &= \sum_{i=1}^5 (\hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^5 (\hat{\beta}_0 + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^5 (\bar{y} - \hat{\beta}_1 \bar{x} + \hat{\beta}_1 x_i - \bar{y})^2 \\ &= \sum_{i=1}^5 (\hat{\beta}_1 x_i - \hat{\beta}_1 \bar{x})^2 \\ &= \hat{\beta}_1^2 \sum_{i=1}^5 (x_i - \bar{x})^2 \\ &= 4.57682^2 \cdot 11.872 \\ &= 248.686\end{aligned}$$

Finally,

$$\text{SST} = \frac{248.686}{0.91245} = \mathbf{272.55}$$

Example 3.2.4.4

Determine which statements are true regarding simple linear regression.

- I. A consequence of ordinary least squares is the minimization of the coefficient of determination.
- II. The choice of explanatory variable affects the total sum of squares.
- III. The positive square root of R^2 will equal the sample correlation of the response and explanatory variables.

Solution

I is false because ordinary least squares minimizes SSE. When SSE decreases, R^2 increases.

II is false because SST is the sum of the squared deviations of y from \bar{y} . There is no connection with any explanatory variable.

III is false because the sample correlation could be either the positive or negative square root of R^2 . The sign of the sample correlation is identical to the sign of $\hat{\beta}_1$.

Therefore, **none of the statements are true.**

3.2.5 Estimators

We investigate statistical inference for the remainder of our discussion on simple linear regression. Specifically, we perform hypothesis tests and construct confidence intervals for a more thorough analysis. The parameters of primary interest are β_0 and β_1 .

The ordinary least squares **estimators** for β_0 and β_1 are:

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

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$$

Coach's Remarks

As described in Section 2.2.1, we take the same approach of denoting estimators with the same symbol used for estimates.

Most importantly, since x_1, \dots, x_n are constants, each estimator turns out to be a linear combination of the random variables Y_1, \dots, Y_n .

In addition, $E[Y] = \beta_0 + \beta_1 x$ (or more precisely, $E[Y | X = x]$) is another parameter of interest. Its estimator is unsurprisingly $\hat{\beta}_0 + \hat{\beta}_1 x$, which we will denote as \hat{Y} .

Let's consider the properties of these estimators.

Distributions

Since $\hat{\beta}_0$ and $\hat{\beta}_1$ are both linear combinations of independent and normally distributed Y_i 's, both estimators are normally distributed as well. In turn, \hat{Y} is also normally distributed.

Means

It can be proven that these estimators are unbiased.

$$\mathbb{E}[\hat{\beta}_0] = \beta_0, \quad \mathbb{E}[\hat{\beta}_1] = \beta_1, \quad \mathbb{E}[\hat{Y}] = \mathbb{E}[Y]$$

Standard Errors

The variances of the estimators are

$$\text{Var}[\hat{\beta}_0] = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{Var}[\hat{\beta}_1] = \sigma^2 \left[\frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

$$\text{Var}[\hat{Y}] = \sigma^2 \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]$$

and in turn, the standard errors of the estimators are

$$\text{SE}[\hat{\beta}_0] = \sqrt{\text{Var}[\hat{\beta}_0]}$$

$$\text{SE}[\hat{\beta}_1] = \sqrt{\text{Var}[\hat{\beta}_1]}$$

$$\text{SE}[\hat{Y}] = \sqrt{\text{Var}[\hat{Y}]}$$

If interested in the proof, see the appendix at the end of the section.

However, the value of σ^2 is unknown. We obtain estimated standard errors by substituting MSE for σ^2 .

$$se(\hat{\beta}_0) = \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \quad (3.2.5.1)$$

$$se(\hat{\beta}_1) = \sqrt{\frac{\text{MSE}}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad (3.2.5.2)$$

$$se(\hat{y}) = \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \quad (3.2.5.3)$$

Coach's Remarks

Note that Equation 3.2.5.3 is the general form of Equation 3.2.5.1. When $x = 0$,

$$se(\hat{y}) = \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} = se(\hat{\beta}_0)$$

as a consequence of

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 0 = \hat{\beta}_0$$

In addition, be aware of how the following estimates are different:

- MSE is the estimate of $\text{Var}[Y] = \sigma^2$
- $se(\hat{y})^2$ is the estimate of $\text{Var}[\hat{Y}]$

Variance-Covariance Matrix

The vector of estimators is

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

Then, the **variance-covariance matrix** of $\hat{\beta}$ is

$$\text{Var}[\hat{\beta}] = \begin{bmatrix} \text{Var}[\hat{\beta}_0] & \text{Cov}[\hat{\beta}_0, \hat{\beta}_1] \\ \text{Cov}[\hat{\beta}_0, \hat{\beta}_1] & \text{Var}[\hat{\beta}_1] \end{bmatrix} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

This means the variance-covariance matrix equals the matrix $(\mathbf{X}^T \mathbf{X})^{-1}$ whose entries are multiplied by the constant σ^2 . But since the value of σ^2 is unknown, the variance-covariance matrix is estimated by substituting MSE for σ^2 . Note that its diagonal entries are $se(\hat{\beta}_0)^2$ and $se(\hat{\beta}_1)^2$.

$$\widehat{\text{Var}}[\hat{\beta}] = \begin{bmatrix} \widehat{\text{Var}}[\hat{\beta}_0] & \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] \\ \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] & \widehat{\text{Var}}[\hat{\beta}_1] \end{bmatrix} = \text{MSE} (\mathbf{X}^T \mathbf{X})^{-1} \quad (3.2.5.4)$$

$$\widehat{\text{Var}}[\hat{\beta}_0] = se(\hat{\beta}_0)^2, \quad \widehat{\text{Var}}[\hat{\beta}_1] = se(\hat{\beta}_1)^2$$

As for the regression of Commute on Precip Chance, the estimated variance-covariance matrix of $\hat{\beta}$ is

$$\text{MSE}(\mathbf{X}^T \mathbf{X})^{-1} = \begin{bmatrix} 0.722684 & -0.011609 \\ -0.011609 & 0.000309 \end{bmatrix}$$

This means

$$se(\hat{\beta}_0) = \sqrt{0.722684} = 0.85011$$

$$se(\hat{\beta}_1) = \sqrt{0.000309} = 0.01758$$

which can also be found in the R output:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.00963  0.85011 28.243 < 2e-16 ***
Precip.Chance 0.06333  0.01758  3.602 0.000499 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.355 on 98 degrees of freedom
Multiple R-squared:  0.1169, Adjusted R-squared:  0.1079
F-statistic: 12.97 on 1 and 98 DF,  p-value: 0.0004986
```

Furthermore, by combining Equations 3.2.5.1, 3.2.5.2, and 3.2.5.4, we can conclude that

$$\text{1}^{\text{st}} \text{ diagonal entry of } (\mathbf{X}^T \mathbf{X})^{-1} = \frac{se(\hat{\beta}_0)^2}{\text{MSE}} = \frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

$$\text{2}^{\text{nd}} \text{ diagonal entry of } (\mathbf{X}^T \mathbf{X})^{-1} = \frac{se(\hat{\beta}_1)^2}{\text{MSE}} = \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2}$$

From Section 3.2.2, recall that

- $\bar{x} = 37.546$
- $\sum_{i=1}^{100} (x_i - \bar{x})^2 = 92,756.31$

- $\text{MSE} = 28.68$

Hence, the numbers and formulas are in agreement.

$$\text{1}^{\text{st}} \text{ diagonal entry of } (\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{100} + \frac{37.546^2}{92,756.31} = 0.025198$$

$$\text{2}^{\text{nd}} \text{ diagonal entry of } (\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{92,756.31} = 0.000011$$

$$(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{28.68} \begin{bmatrix} 0.722684 & -0.011609 \\ -0.011609 & 0.000309 \end{bmatrix} = \begin{bmatrix} 0.025198 & -0.000405 \\ -0.000405 & 0.000011 \end{bmatrix}$$

Bootstrapping

While we use the estimated standard errors of Equations 3.2.5.1 to 3.2.5.3 in making statistical inference, consider the alternative based on Monte Carlo simulation.

As suggested in Section 1.8.4, if we wish to estimate the standard error of these ordinary least squares estimators, we should

- simulate a large number of estimates for the parameter of interest, then
- calculate the sample standard deviation of the simulated values.

However, Monte Carlo simulation requires knowing the true distribution of the estimator. We assume a normal distribution in simple linear regression, but this may be an incorrect assumption.

We can circumvent this with **bootstrapping**; it is a statistical technique that involves random sampling of observations with replacement. The goal of bootstrapping is to create many bootstrap samples, i.e. artificial datasets created from one dataset. We run a regression on each bootstrap sample, thus producing a different parameter estimate each time. With a massive number of bootstrap samples, we can mimic a Monte Carlo simulation.

Since the sampling is done with replacement, an observation can appear in a bootstrap sample more than once. Note that each bootstrap sample must have the same size as the original dataset.

Example 3.2.5.1

The following data was used in a simple linear regression:

Observation, i	Feature, x_i	Response, y_i
1	9	3.02
2	12	1.66
3	8	2.91
4	10	2.40
5	11	2.08
6	11	2.37

You are told:

- Five bootstrap samples were drawn from the data, where

Bootstrap Sample	Estimate for β_1
1	-0.3194
2	-0.1845
3	-0.2826
4	-0.2661

- The fifth bootstrap sample consists of only the even-numbered observations, with each equally represented.

Calculate the standard error of these bootstrap estimates of β_1 .

Solution

Since the original dataset has a size of 6, each bootstrap sample has 6 observations. Therefore, the fifth bootstrap sample consists of

Feature,	Response,
12	1.66
12	1.66
10	2.40

Feature, x	Response, y
10	2.40
11	2.37
11	2.37

Calculate $\hat{\beta}_1$ for this bootstrap sample. Use Equation 3.2.2.1 or the TI-30XS MultiView calculator to obtain -0.37 as the estimate.

The answer is the unbiased sample standard deviation of the slope estimates.

$$\frac{-0.3194 + (-0.1845) + (-0.2826) + (-0.2661) + (-0.37)}{5} = -0.28452$$

$$\frac{[-0.3194 - (-0.28452)]^2 + \dots + [-0.37 - (-0.28452)]^2}{5 - 1} = 0.00472$$

$$\sqrt{0.00472} = \mathbf{0.069}$$



3.2.6 t Tests

Two-Tailed t Tests

We can learn more about a model's parameters using hypothesis tests. An important set of hypotheses to consider is

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

To see why we should test whether the slope parameter is 0, consider what happens to the model equation if H_0 is true.

$$\begin{aligned} Y &= \beta_0 + \beta_1 x + \varepsilon \\ &= \beta_0 + 0 \cdot x + \varepsilon \\ &= \beta_0 + \varepsilon \end{aligned}$$

The simple linear regression now becomes the null model. Therefore, if H_0 is true, then there is **no relationship** between the response and explanatory variables. So, to establish that a linear relationship between the variables is more plausible, we aim to reject H_0 .

Using the estimator $\hat{\beta}_1$, we can conduct a hypothesis test for its mean, β_1 . The generalization for testing means presented in Section 2.3.3 applies here. Since $\hat{\beta}_1$ is normally distributed, we have a test statistic of

$$t.s. = \frac{\hat{\beta}_1 - h}{se(\hat{\beta}_1)} \tag{3.2.6.1}$$

which comes from a t -distribution with $n - 2$ degrees of freedom. It is **not** a test statistic from the standard normal distribution because MSE is used instead of σ^2 . The degrees of freedom is $n - 2$ since this is the denominator of MSE.

Coach's Remarks

Recall the discussion on degrees of freedom from Section 2.5.2. To see how it applies in this case, we revisit Equation 3.2.3.3:

$$\text{MSE} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - 2} = \frac{\sum_{i=1}^n e_i^2}{n - 2}$$

This calculation has n data points: e_1, \dots, e_n . Not all n of them are free to take on any value. The residuals must follow two constraints:

$$\sum_{i=1}^n e_i = 0, \quad \sum_{i=1}^n x_i e_i = 0$$

These constraints are the ordinary least squares equations; see the appendix at the end of the section for the proof. So, the two estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ will force two values on the residuals once the other $n - 2$ values are specified. In other words, only $n - 2$ residuals are unconstrained in calculating MSE, i.e. there are $n - 2$ degrees of freedom. Consequently, a denominator of $n - 2$ results in MSE being an unbiased estimator of σ^2 .

We reject H_0 at the α significance level when

$$|t. s. | \geq t_{\alpha, n-2}$$

or equivalently, if

$$p\text{-value} \leq \alpha$$

By default, the R output gives us the result for testing $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 24.00963   0.85011 28.243 < 2e-16 ***
Precip.Chance 0.06333   0.01758   3.602 0.000499 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.355 on 98 degrees of freedom
Multiple R-squared:  0.1169, Adjusted R-squared:  0.1079
F-statistic: 12.97 on 1 and 98 DF,  p-value: 0.0004986
```

t value refers to the test statistic. By applying Equation 3.2.6.1,

$$t. s. = \frac{0.06333 - 0}{0.01758} = 3.602$$

The output also reveals that $df = 98$. Knowing that $df = n - 2$ for a simple linear regression, we can solve for the number of observations, i.e. $n = 98 + 2 = 100$.

The exam table does not list t percentiles for 98 degrees of freedom. So it is not convenient to compare the test statistic to the critical value without any computer software. However, the R output supplies the p -value for the two-tailed test under $\Pr(>|t|)$. We can see that the p -value is very small (i.e. 0.0499%). There is also an indicator of the test result to the right of the p -value. The symbol *** can be interpreted using the spectrum listed next to **Signif. codes**.

In the spectrum, a given symbol signifies that the p -value falls between the two numbers surrounding the symbol. Our p -value of 0.0499% falls between 0 and 0.1%, which is why *** is listed next to it. Thus, the numbers listed in **Signif. codes** are choices of significance levels.

So the symbol *** denotes a p -value smaller than a 0.001 significance level, which leads to the conclusion that the parameter is significantly different from 0. On the opposite extreme, not having a symbol denotes a p -value larger than a 0.1 significance level, which leads to the conclusion that the parameter is plausibly 0. This makes it easy to identify the test result for each row of the **Coefficients** table.

Thus, we reject H_0 at the 0.001 significance level. Equivalently, we say that the slope parameter is significantly different from 0 at the 0.001 level. We conclude that there is a meaningful relationship between Commute and Precip Chance. This does not contradict the low R^2 of 11.69%. Even though Precip Chance is unable to explain much of the variability in Commute, it is still a helpful predictor of Commute.

Typically, a hypothesis test for β_0 is of little importance. This is because the intercept of a linear model seldom addresses relevant issues or interests. For example, knowing that the intercept parameter is significantly different from 0 does not give much insight into how Commute and Precip Chance interact. Having said that, this does not mean we should ignore testing β_0 altogether. Unsurprisingly, the details to testing β_0 is analogous to that of testing β_1 .

Example 3.2.6.1

With a sample of university students, an analyst investigates how the number of years of enrollment relates to the time spent to complete a typical homework assignment. The following R output is the result of the analysis:

```

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept) 23.5252   0.5387 43.672 <2e-16 ***
Years.enrolled 3.9485   0.4230  9.334 5e-14 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.798 on 72 degrees of freedom
Multiple R-squared: 0.5475, Adjusted R-squared: 0.5413
F-statistic: 87.13 on 1 and 72 DF, p-value: 4.999e-14

```

Determine:

1. the number of students in the sample.
2. the decision of whether the slope parameter equals 0 at the 1% significance level using a t test.
3. the decision of whether the intercept parameter equals 18 at the 1% significance level using a t test.

Solution to (1)

This is a simple linear regression with $df = 72$ associated with error. Therefore,

$$df = n - 2$$

$$\Rightarrow n = 72 + 2 = 74$$



Solution to (2)

The "e" in the output reads as "10 raised to the power of". So the p -value for this t test is given as 5×10^{-14} . Since $5 \times 10^{-14} < 0.01$, we conclude that **the slope parameter is significantly different from 0 at the 0.01 level**.

Solution to (3)

First, calculate the test statistic using the form of Equation 3.2.6.1.

$$\frac{\hat{\beta}_0 - h}{se(\hat{\beta}_0)} = \frac{23.5252 - 18}{0.5387} = 10.257$$

We cannot use the exam table to find the critical value because it does not include t percentiles for 72 degrees of freedom. We also cannot calculate the p -value directly. However, we see that the test statistic for the test in Part (2) is 9.334, and that

$$|10.257| > |9.334|$$

This means the p -value for this test must be **even smaller than** the p -value for the test in Part (2). This is true because both tests are based on a t -distribution with 72 degrees of freedom.

Since the p -value for this test is even smaller than 5×10^{-14} , we can conclude that **the intercept parameter is significantly different from 18 at the 0.01 level**.

Example 3.2.6.2

A linear regression on 10 observations is given by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\beta}^T = [\beta_0 \quad \beta_1]$.

You are given:

- $(\mathbf{X}^T \mathbf{X})^{-1} = \frac{1}{222} \begin{bmatrix} 119 & -22 \\ -22 & 5 \end{bmatrix}$
- The ordinary least squares estimate of β_1 is -1.2432.
- The sum of squared residuals is 17.473.

You look for evidence that β_1 is significantly different from 0 at the 2% significance level using a t test.

Determine the test result.

Solution

To perform this test, we need $se(\hat{\beta}_1)$. Recall that the 2nd diagonal element of $MSE(\mathbf{X}^T \mathbf{X})^{-1}$ is $se(\hat{\beta}_1)^2$.

To calculate MSE, note that the sum of squared residuals, or SSE, is 17.473. Thus,

$$\begin{aligned} MSE &= \frac{SSE}{n - 2} \\ &= \frac{17.473}{10 - 2} \\ &= 2.184 \end{aligned}$$

This means

$$\begin{aligned} se(\hat{\beta}_1) &= \sqrt{MSE \cdot 2^{\text{nd}} \text{ diagonal element of } (\mathbf{X}^T \mathbf{X})^{-1}} \\ &= \sqrt{2.184 \cdot \frac{5}{222}} \\ &= 0.2218 \end{aligned}$$

Next, calculate the test statistic. The null hypothesis is that $\beta_1 = 0$.

$$\frac{-1.2432 - 0}{0.2218} = -5.605$$

This test involves $10 - 2 = 8$ degrees of freedom. From the exam table, the critical value is $t_{0.02, 8} = 2.896$.

Since $| -5.605 | > 2.896$, we conclude that β_1 is significantly different from 0 at the 0.02 significance level.



Regression through the Origin

Imagine a scenario where we fail to reject $H_0 : \beta_0 = 0$. This means the evidence supports a model equation of

$$\begin{aligned} Y &= 0 + \beta_1 x + \varepsilon \\ &= \beta_1 x + \varepsilon \end{aligned}$$

which makes the regression line \hat{y} pass through the origin of a y versus x scatterplot.

While an analyst might have good reason for running a regression through the origin, it has important consequences that might be surprising or unintuitive:

- It claims that the value of the intercept parameter is known, i.e. 0. Rather than letting the data estimate the intercept's value, it is fixed at 0, which restricts the scope of the "best" linear fit between the response and explanatory variables.
- It claims that when $x = 0$, then $\hat{y} = 0$. Depending on the context of the analysis, this result may not be sensible, such as if the response can only take on positive values.

One-Tailed t Tests

In studying Commute using Precip Chance, it is reasonable to anticipate that a higher chance of precipitation (i.e. an indication of bad weather) might result in longer commutes, even before any data is collected. In a simple linear regression context, this is looking to prove that there is a **positive** slope between the two variables. Effectively, we are considering the following set of hypotheses:

$$H_0 : \beta_1 \leq 0 \quad H_1 : \beta_1 > 0$$

As a reminder, one-tailed tests mainly differ from two-tailed tests in the following ways:

- Critical region
- Calculation of the critical value
- Calculation of the p -value

For a **left-tailed** test, reject H_0 when

$$t.s. \leq -t_{2\alpha, n-2}$$

For a **right-tailed** test, reject H_0 when

$$t.s. \geq t_{2\alpha, n-2}$$

Regardless of the type of test performed, we reject H_0 when p -value $\leq \alpha$.

To find evidence for a positive β_1 in the Commuting Chris scenario, note that:

- The two-tailed p -value of 0.0499% is the probability that the t random variable is less than -3.602 or greater than 3.602.
- The right-tailed p -value is the probability that the t random variable is greater than 3.602.

Since the p -value for this test is $0.0499\% \div 2 = 0.025\%$, we reject H_0 at the 0.001 significance level. This supports the claim that the slope parameter is positive.

Example 3.2.6.3

Determine which statements are true about hypothesis testing under a simple linear regression.

- I. To find evidence for a negative slope parameter, the null hypothesis is that the slope parameter is less than 0.
- II. For a given dataset, a two-tailed t test will have different degrees of freedom compared to a one-tailed t test.
- III. For a given dataset and significance level, the positive critical value for a two-tailed t test is the same as the critical value for a right-tailed t test.
- IV. In failing to reject the null hypothesis that the slope parameter is 0, we conclude that there is a meaningful relationship between the response and explanatory variables.

V. There are situations where a large p -value leads to rejecting the null hypothesis.

Solution

I is false because it is the alternative hypothesis that should be $\beta_1 < 0$. Typically, the alternative hypothesis is the statement that we seek to prove.

II is false because the t test degrees of freedom equals the denominator of MSE, which is fixed at $n - 2$ for a given dataset. This is the case regardless of the number of tails in the critical region.

III is false because the positive critical value of a two-tailed t test is $t_{\alpha/2, n-2}$, whereas the critical value of a right-tailed t test is $t_{\alpha, n-2}$.

IV is false because failing to reject that the slope is 0 means the null model is plausible, which indicates no relationship between the response and explanatory variables.

V is false because a large p -value always indicates that the test statistic is not extreme. Thus, the null hypothesis would never be rejected when the p -value is large.

Therefore, **none of the statements are true.**



3.2.7 Confidence and Prediction Intervals

Similar to hypothesis testing, t -distributions are used for interval estimation in simple linear regression. It should be no surprise that the confidence intervals in this subsection have this general expression:

$$\text{estimate} \pm (t \text{ percentile}) (\text{standard error})$$

Confidence Intervals

The $100k\%$ confidence interval for β_0 has the expression

$$\hat{\beta}_0 \pm t_{1-k, n-2} \cdot se(\hat{\beta}_0) \quad (3.2.7.1)$$

The $100k\%$ confidence interval for β_1 has the expression

$$\hat{\beta}_1 \pm t_{1-k, n-2} \cdot se(\hat{\beta}_1) \quad (3.2.7.2)$$

The $100k\%$ confidence interval for $E[Y]$ has the expression

$$\hat{y} \pm t_{1-k, n-2} \cdot se(\hat{y}) \quad (3.2.7.3)$$

Example 3.2.7.1

A simple linear regression was performed on the following data:

Response, y	Predictor, x
114.8	5
109.1	4
99.7	4
116.0	7

The regression's residual standard error is 6.172.

Calculate the 90% confidence interval for the slope parameter.

Solution

Start by solving for the estimate of the slope parameter. Using Equation 3.2.2.1, we require the sample means of both x and y .

$$\bar{x} = \frac{5 + 4 + 4 + 7}{4} = 5, \quad \bar{y} = \frac{114.8 + 109.1 + 99.7 + 116.0}{4} = 109.9$$

$$\begin{aligned}\hat{\beta}_1 &= \frac{\sum_{i=1}^4 (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^4 (x_i - \bar{x})^2} \\ &= \frac{(5 - 5)(114.8 - 109.9) + \dots + (7 - 5)(116.0 - 109.9)}{(5 - 5)^2 + \dots + (7 - 5)^2} \\ &= \frac{23.2}{6} \\ &= 3.867\end{aligned}$$

Next, calculate the estimated standard error required for the confidence interval with Equation 3.2.5.2. Recall that the square root of the MSE is the residual standard error.

$$\begin{aligned}se(\hat{\beta}_1) &= \sqrt{\frac{\text{MSE}}{\sum_{i=1}^4 (x_i - \bar{x})^2}} \\ &= \frac{\sqrt{\text{MSE}}}{\sqrt{(5 - 5)^2 + \dots + (7 - 5)^2}} \\ &= \frac{6.172}{\sqrt{6}} \\ &= 2.5197\end{aligned}$$

For 90% confidence and $n = 4$, we need

$$t_{1-0.9, 4-2} = t_{0.1, 2} = 2.920$$

Finally, the 90% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{0.1, 2} \cdot se(\hat{\beta}_1)$$

$$\Rightarrow 3.867 \pm 2.920 \cdot 2.5197$$

$$\Rightarrow (-3.49, 11.22)$$



Example 3.2.7.2

Running a simple linear regression in R produced the following results:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 16.9310   1.0091 16.778 2.86e-06 ***
x           -1.1897   0.2082 -5.715  0.00124 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.373 on 6 degrees of freedom
Multiple R-squared:  0.8448, Adjusted R-squared:  0.8189
F-statistic: 32.66 on 1 and 6 DF,  p-value: 0.001243
```

The sample mean and unbiased sample variance of x are 4.25 and 6.2143, respectively.

Calculate:

1. the 95% confidence interval for the intercept parameter.
2. the 95% confidence interval for the mean response when $x = 5$.

Solution to (1)

For 95% confidence and 6 degrees of freedom, we need

$$t_{1-0.95, 6} = t_{0.05, 6} = 2.447$$

The 95% confidence interval for β_0 is

$$\hat{\beta}_0 \pm t_{0.05, 6} \cdot se(\hat{\beta}_0)$$

$$\Rightarrow 16.931 \pm 2.447 \cdot 1.0091$$

$$\Rightarrow (14.462, 19.400)$$



Solution to (2)

Having already found $t_{0.05, 6} = 2.447$, calculate \hat{y} and $se(\hat{y})$ for the confidence interval expression. Given $x = 5$,

$$\hat{y} = 16.931 - 1.1897(5) = 10.9825$$

Recall Equation 3.2.5.3:

$$se(\hat{y}) = \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

Note that

- $\sqrt{\text{MSE}} = \text{residual standard error} = 1.373$
- $n = \text{df} + 2 = 6 + 2 = 8$
- $x = 5$
- $\bar{x} = 4.25$
- $s_x^2 = 6.2143 \Rightarrow \sum_{i=1}^8 (x_i - \bar{x})^2 = (8 - 1)(6.2143) = 43.5$

Thus,

$$se(\hat{y}) = 1.373 \sqrt{\frac{1}{8} + \frac{(5 - 4.25)^2}{43.5}} = 0.5099$$

Finally, the 95% confidence interval for $E[Y]$ at $x = 5$ is

$$\hat{y} \pm t_{0.05, 6} \cdot se(\hat{y})$$

$$\Rightarrow 10.9825 \pm 2.447 \cdot 0.5099$$

$$\Rightarrow (9.735, 12.230)$$



Coach's Remarks

In Part (2), the parameter of interest is more precisely $E[Y | X = 5]$. In general, the

parameter of interest is $E[Y | X = x]$. This is seen in \hat{y} and $se(\hat{y})$ since both are functions of x . In other words, a different value of x points to a different parameter and, in turn, produces a different center for the interval, as well as a different width. This is further developed when we discuss prediction intervals.

Example 3.2.7.3

A simple linear regression on 13 observations results in the following:

- The 90% confidence interval for the slope parameter is (-2.687, 4.253).
- The residual standard error is 22.058.

Determine which of the following statements are true.

- I. The slope estimate is 0.783.
- II. The appropriate t percentile used to compute the confidence interval is 1.363.
- III. The unbiased sample variance of the explanatory variable cannot be determined.
- IV. In testing whether there is a meaningful linear relationship between the response and explanatory variables, we would fail to reject H_0 at the 0.1 level.

Solution

I is true because $\hat{\beta}_1$ is the midpoint of the given confidence interval.

$$\frac{-2.687 + 4.253}{2} = 0.783$$

II is false because the appropriate percentile is

$$t_{1-0.9, 13-2} = t_{0.1, 11} = 1.796$$

III is false. The unbiased sample variance of the predictor can be determined as follows:

$$\hat{\beta}_1 + t_{0.1, 11} \cdot se(\hat{\beta}_1) = 4.253$$

$$\begin{aligned}\Rightarrow se(\hat{\beta}_1) &= \frac{4.253 - \hat{\beta}_1}{t_{0.1, 11}} \\ &= \frac{4.253 - 0.783}{1.796} \\ &= 1.9321\end{aligned}$$

Next, use Equation 3.2.5.2 to find the numerator of the unbiased sample variance of the predictor. Recall that the residual standard error is the square root of the MSE.

$$\sqrt{\frac{\text{MSE}}{\sum_{i=1}^{13} (x_i - \bar{x})^2}} = 1.9321$$

$$\begin{aligned}\Rightarrow \sum_{i=1}^{13} (x_i - \bar{x})^2 &= \frac{\text{MSE}}{1.9321^2} \\ &= \left(\frac{22.058}{1.9321} \right)^2\end{aligned}$$

$$\begin{aligned}s_x^2 &= \frac{\sum_{i=1}^{13} (x_i - \bar{x})^2}{13 - 1} \\ &= \frac{(22.058 / 1.9321)^2}{12} \\ &= 10.86\end{aligned}$$

IV is true. We are testing the hypotheses

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

at the 0.1 level. This means we can determine the test result using the $100(1 - 0.1) = 90\%$ confidence interval for β_1 . This interval is given in the problem.

Since the hypothesized value of 0 is within the 90% confidence interval, we fail to reject H_0 at the 0.1 level.

Therefore, **only I and IV are true.**



Prediction Intervals

A ***prediction interval*** is a range of values that estimates a new observation's response. We denote the new observation's response as \hat{Y}_{n+1} . Since the interval is constructed for a **random variable** instead of a parameter, there is no estimator to speak of per se. Rather than an estimator, we rely on $\hat{Y}_{n+1} - \hat{Y}_{n+1}$ being normally distributed with mean 0 and variance

$$\begin{aligned}\text{Var}[Y_{n+1} - \hat{Y}_{n+1}] &= \text{Var}[Y_{n+1}] + \text{Var}[\hat{Y}_{n+1}] \\ &= \text{Var}[\varepsilon_{n+1}] + \text{Var}[\hat{Y}_{n+1}] \\ &= \sigma^2 + \sigma^2 \left[\frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \\ &= \sigma^2 \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]\end{aligned}$$

where $\hat{Y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$. Note that x_{n+1} is a chosen value for the new observation's explanatory variable.

This leads to \hat{Y}_{n+1} having an "estimate" (i.e. the center of the prediction interval) of

$$\hat{y}_{n+1} = \hat{\beta}_0 + \hat{\beta}_1 x_{n+1}$$

and the "standard error" for the interval is

$$se(\hat{y}_{n+1}) = \sqrt{\text{MSE} \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \quad (3.2.7.4)$$

Therefore, the $100k\%$ prediction interval has the expression

$$\hat{y}_{n+1} \pm t_{1-k, n-2} \cdot se(\hat{y}_{n+1})$$

Coach's Remarks

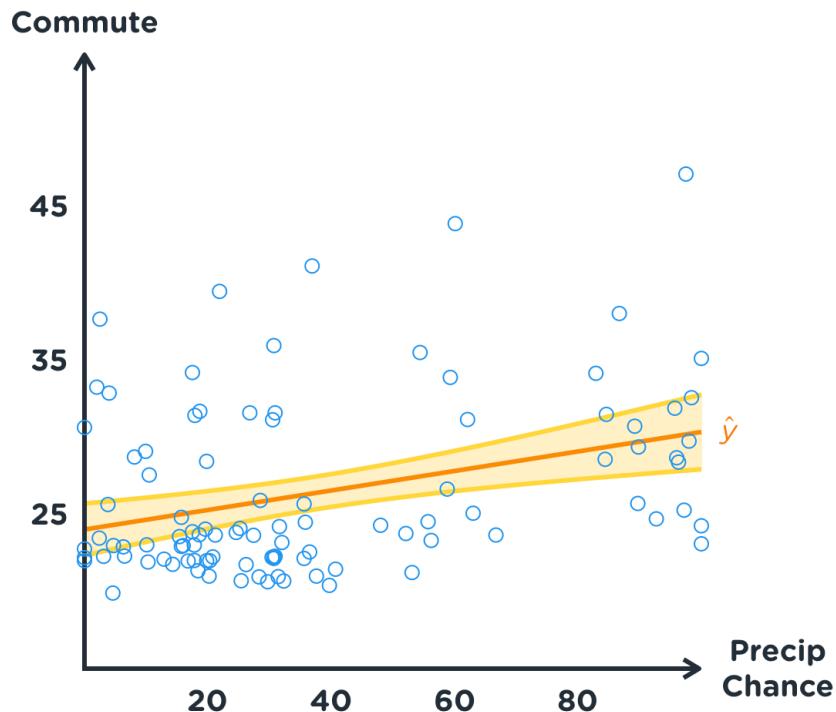
When making inferences on $E[Y]$, an explanatory variable value needs to be specified, given by the symbol x . The same is necessary when constructing a prediction interval for \hat{Y}_{n+1} , where the symbol x_{n+1} is used instead.

The symbols x and x_{n+1} have the same function and purpose. For prediction intervals, the subscript $n + 1$ is added to x simply to emphasize its association to \hat{Y}_{n+1} .

It is important to know how a prediction interval is different from a confidence interval for $E[Y]$:

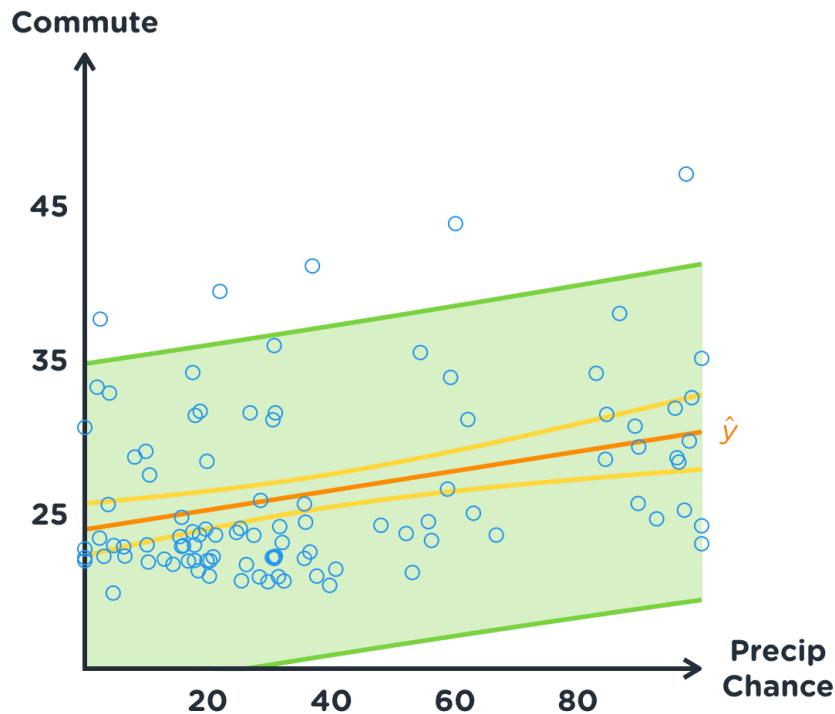
- The confidence interval is a range for the mean response, whereas a prediction interval is a range for a response value.
- At the same explanatory variable value and confidence level, the prediction interval is at least as wide as the confidence interval. This is because a prediction interval accounts for the variance of ε in addition to that of \hat{Y} . This is evident from comparing the standard error formulas (i.e. Equations 3.2.5.3 and 3.2.7.4).

We can visualize this with the Commuting Chris scenario.



The plot above shows 95% confidence intervals for the mean response superimposed on the scatterplot of Commute and Precip Chance. Specifically, a total of 101 confidence intervals were created for $\text{Precip Chance} = 0, 1, \dots, 100$. The lower bounds of all the intervals are connected to form the curve below the fitted line \hat{y} ; likewise, the upper bounds form the curve above \hat{y} . There is 95% confidence that the true mean, $E[Y] = \beta_0 + \beta_1 x$, lies within the shaded region.

Note that the yellow lower and upper bounds form curves rather than straight lines. This means the width of a confidence interval is **different** across the Precip Chance values, even though the confidence level is fixed at 95%. By examining the $se(\hat{y})$ formula, realize that the closer the chosen Precip Chance value is to its sample mean, the narrower the interval becomes.



The plot above further adds the 95% prediction intervals for the same values of Precip Chance. The shaded region depicts where the new data point is with 95% confidence. The prediction intervals are wider than the confidence intervals, which is expected since an observation has an additional source of variability compared to a mean.

Although it is hard to tell from the plot, the green lower and upper bounds form curves as well. In fact, they behave much like the yellow curves: the prediction interval widths are narrower as Precip Chance approaches its sample mean, as given by the $se(\hat{y}_{n+1})$ formula.

Example 3.2.7.4

Determine which statement is true for a simple linear regression.

- I. For a given confidence level and predictor value, it is possible for the prediction interval for the response to be narrower than the confidence interval for the mean response.
- II. For a given confidence level, the prediction interval is narrowest when the chosen predictor value is the sample mean of the predictor.
- III. A confidence interval for the mean response is unaffected by the estimate of the irreducible error.

Solution

I is false because the prediction interval will be at least as wide as the confidence interval for the mean response, given a confidence level and an explanatory variable value.

II is true. When $x_{n+1} = \bar{x}$,

$$\begin{aligned} se(\hat{y}_{n+1}) &= \sqrt{\text{MSE} \left[1 + \frac{1}{n} + \frac{(\bar{x} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \\ &= \sqrt{\text{MSE} \left[1 + \frac{1}{n} + \frac{0}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]} \\ &= \sqrt{\text{MSE} \left(1 + \frac{1}{n} \right)} \end{aligned}$$

Thus, the standard error term is minimized at $x_{n+1} = \bar{x}$, resulting in the narrowest prediction interval for a given confidence level.

III is false because the irreducible error is σ^2 , whose estimate is the MSE. The formula for $se(\hat{y})$ includes the MSE, so confidence intervals for the mean response are affected by it.

Therefore, **only II is true.**



Coach's Remarks

The claim that statement II makes regarding a prediction interval is also true regarding a confidence interval for the mean response.

3.2 Summary

Simple Linear Regression Notation

Symbol	Concept
β_0	Intercept parameter
β_1	Slope parameter
σ^2	Variance of response / Irreducible error
$\hat{\beta}_0$	Estimate/Estimator for β_0
$\hat{\beta}_1$	Estimate/Estimator for β_1
MSE	Estimate of σ^2
\mathbf{X}	Design matrix
\mathbf{H}	Hat matrix
e	Residual
SST	Total sum of squares
SSR	Regression sum of squares
SSE	Error sum of squares
\hat{Y}	Estimator for $E[Y]$
se	Estimated standard error
df	Degrees of freedom
$t_{2(1-q), df}$	100 q^{th} percentile of a t -distribution
Y_{n+1}	Response of new observation

Assumptions

1. $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$
2. x_i 's are non-random
3. $E[\varepsilon_i] = 0$
4. $\text{Var}[\varepsilon_i] = \sigma^2$ (i.e. homoscedasticity)
5. ε_i 's are independent
6. ε_i 's are normally distributed

Estimation of β_0 and β_1

Ordinary least squares solves for $\hat{\beta}_0$ and $\hat{\beta}_1$ by minimizing SSE.

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

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

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{bmatrix} = \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

PREDICTION

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

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T, \quad \hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$$

Estimation of σ^2

$$\text{MSE} = \frac{\text{SSE}}{n - 2}$$

$$\text{residual standard error} = \sqrt{\text{MSE}}$$

Residuals

$$e = y - \hat{y}$$

Sum of Squares

- $SST = \sum_{i=1}^n (y_i - \bar{y})^2$ = total variability
- $SSR = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ = explained variability
- $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$ = unexplained variability
- $SST = SSR + SSE$

Coefficient of Determination

This is the proportion of variability in the response that is explained by the regression.

$$R^2 = \frac{SSR}{SST}$$

For simple linear regression, $R^2 = r_{x,y}^2$.

Standard Errors

$$se(\hat{\beta}_0) = \sqrt{MSE \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

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

$$se(\hat{y}) = \sqrt{\text{MSE} \left[\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

$$se(\hat{y}_{n+1}) = \sqrt{\text{MSE} \left[1 + \frac{1}{n} + \frac{(x_{n+1} - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right]}$$

Estimating standard errors could also be done using bootstrapping.

Variance-Covariance Matrix

$$\widehat{\text{Var}}[\hat{\beta}] = \begin{bmatrix} \widehat{\text{Var}}[\hat{\beta}_0] & \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] \\ \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] & \widehat{\text{Var}}[\hat{\beta}_1] \end{bmatrix} = \text{MSE}(\mathbf{X}^T \mathbf{X})^{-1}$$

Hypothesis Tests

df = denominator of MSE = $n - 2$

$$t. s. = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

For an α level test, we reject H_0 if $p\text{-value} \leq \alpha$, or equivalently

Test Type	Critical Region
Left-tailed	$t. s. \leq -t_{2\alpha, n-2}$
Two-tailed	$ t. s. \geq t_{\alpha, n-2}$
Right-tailed	$t. s. \geq t_{2\alpha, n-2}$

Regression through the origin assumes $\beta_0 = 0$.

Confidence and Prediction Intervals

estimate $\pm (t \text{ percentile}) (\text{standard error})$

Quantity	Interval Expression
β_0	$\hat{\beta}_0 \pm t_{1-k, n-2} \cdot se(\hat{\beta}_0)$
β_1	$\hat{\beta}_1 \pm t_{1-k, n-2} \cdot se(\hat{\beta}_1)$
$E[Y]$	$\hat{y} \pm t_{1-k, n-2} \cdot se(\hat{y})$
Y_{n+1}	$\hat{y}_{n+1} \pm t_{1-k, n-2} \cdot se(\hat{y}_{n+1})$

Appendix

SLR Ordinary Least Squares Equations

For simple linear regression, the SSE is

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The ordinary least squares estimates minimize this expression; they are found by taking partial derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, and then setting each to equal 0.

For the intercept,

$$\begin{aligned} \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 &= \sum_{i=1}^n \frac{\partial}{\partial \hat{\beta}_0} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\ &= \sum_{i=1}^n 2(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-1) \\ &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \end{aligned}$$

$$\begin{aligned}
& -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
& \Rightarrow \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) = 0 \\
& \Rightarrow \sum_{i=1}^n y_i - \sum_{i=1}^n \hat{\beta}_0 - \sum_{i=1}^n \hat{\beta}_1 x_i = 0 \\
& \Rightarrow \sum_{i=1}^n y_i - n\hat{\beta}_0 - \hat{\beta}_1 \sum_{i=1}^n x_i = 0 \\
& \Rightarrow -n\hat{\beta}_0 = -\sum_{i=1}^n y_i + \hat{\beta}_1 \sum_{i=1}^n x_i \\
& \Rightarrow \hat{\beta}_0 = \frac{\sum_{i=1}^n y_i}{n} - \frac{\hat{\beta}_1 \sum_{i=1}^n x_i}{n} \\
& \Rightarrow \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}
\end{aligned}$$

For the slope,

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 &= \sum_{i=1}^n \frac{\partial}{\partial \hat{\beta}_1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^n 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-x_i) \\
&= -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)
\end{aligned}$$

$$\begin{aligned}
& -2 \sum_{i=1}^n x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \\
& \Rightarrow \sum_{i=1}^n x_i \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i \right) = 0 \\
& \Rightarrow \sum_{i=1}^n \left(x_i y_i - \hat{\beta}_0 x_i - \hat{\beta}_1 x_i^2 \right) = 0 \\
& \Rightarrow \sum_{i=1}^n x_i y_i - \hat{\beta}_0 \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \\
& \Rightarrow \sum_{i=1}^n x_i y_i - \left(\bar{y} - \hat{\beta}_1 \bar{x} \right) \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \\
& \Rightarrow \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i + \hat{\beta}_1 \bar{x} \sum_{i=1}^n x_i - \hat{\beta}_1 \sum_{i=1}^n x_i^2 = 0 \\
& \Rightarrow \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i + \hat{\beta}_1 \left(\bar{x} \sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2 \right) = 0 \\
& \Rightarrow \hat{\beta}_1 = \frac{-\sum_{i=1}^n x_i y_i + \bar{y} \sum_{i=1}^n x_i}{\bar{x} \sum_{i=1}^n x_i - \sum_{i=1}^n x_i^2} \\
& \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i}{\sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i} \\
& \Rightarrow \hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$

Note that:

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) &= \sum_{i=1}^n x_i y_i - x_i \bar{y} - \bar{x} y_i + \bar{x} \bar{y} \\
&= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + n \bar{x} \bar{y} \\
&= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i - \bar{x} \sum_{i=1}^n y_i + \bar{x} \sum_{i=1}^n y_i \\
&= \sum_{i=1}^n x_i y_i - \bar{y} \sum_{i=1}^n x_i
\end{aligned}$$

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + n\bar{x}^2 \\
&= \sum_{i=1}^n x_i^2 - 2\bar{x} \sum_{i=1}^n x_i + \bar{x} \sum_{i=1}^n x_i \\
&= \sum_{i=1}^n x_i^2 - \bar{x} \sum_{i=1}^n x_i
\end{aligned}$$

Therefore, the two ordinary least squares equations for a simple linear regression are

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \quad (3.2.2.1)$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (3.2.2.2)$$

Alternative Form of $\hat{\beta}_1$

$$\begin{aligned}
\hat{\beta}_1 &= r_{x,y} \cdot \frac{s_y}{s_x} \\
&= \frac{cov_{x,y}}{s_x \cdot s_y} \cdot \frac{s_y}{s_x} \\
&= \frac{cov_{x,y}}{s_x^2} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{n - 1} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n - 1} \\
&= \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned} \quad (3.2.2.2)$$

Simple Linear Regression Means of Estimators

For $\hat{\beta}_0$:

$$\begin{aligned}
 E[\hat{\beta}_0] &= E[\bar{Y} - \hat{\beta}_1 \bar{x}] \\
 &= E[\bar{Y}] - E[\hat{\beta}_1 \bar{x}] \\
 &= E\left[\frac{\sum_{i=1}^n Y_i}{n}\right] - \bar{x} E[\hat{\beta}_1] \\
 &= \frac{\sum_{i=1}^n (\beta_0 + \beta_1 x_i)}{n} - \bar{x} \beta_1 \\
 &= \frac{n\beta_0 + \beta_1 \sum_{i=1}^n x_i}{n} - \bar{x} \beta_1 \\
 &= \beta_0 + \beta_1 \bar{x} - \bar{x} \beta_1 \\
 &= \beta_0
 \end{aligned}$$

For $\hat{\beta}_1$:

$$\begin{aligned}
E[\hat{\beta}_1] &= E\left[\frac{\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2}\right] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot E\left[\sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y})\right] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot E\left[\sum_{i=1}^n Y_i(x_i - \bar{x})\right] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot E\left[\sum_{i=1}^n (\beta_0 + \beta_1 x_i + \varepsilon_i)(x_i - \bar{x})\right] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot E\left[\sum_{i=1}^n \beta_0(x_i - \bar{x}) + \sum_{i=1}^n \beta_1 x_i(x_i - \bar{x}) + \sum_{i=1}^n \varepsilon_i(x_i - \bar{x})\right] \\
&= \frac{1}{\sum_{i=1}^n (x_i - \bar{x})^2} \cdot E\left[\sum_{i=1}^n \beta_1 x_i(x_i - \bar{x})\right] \\
&= \frac{\sum_{i=1}^n \beta_1 x_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta_1 \cdot \frac{\sum_{i=1}^n x_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \beta_1
\end{aligned}$$

Note that:

- $$\begin{aligned}
\sum_{i=1}^n \beta_0(x_i - \bar{x}) &= \sum_{i=1}^n \beta_0 x_i - \sum_{i=1}^n \beta_0 \bar{x} \\
&= \beta_0 \sum_{i=1}^n x_i - n\beta_0 \bar{x} \\
&= \beta_0 n \bar{x} - n\beta_0 \bar{x} \\
&= 0
\end{aligned}$$
- $$E\left[\sum_{i=1}^n \varepsilon_i(x_i - \bar{x})\right] = 0$$

$$\begin{aligned}
 \bullet \quad \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n x_i^2 - n\bar{x}\bar{x} \\
 &= \sum_{i=1}^n x_i^2 - \sum_{i=1}^n \bar{x}x_i \\
 &= \sum_{i=1}^n x_i (x_i - \bar{x})
 \end{aligned}$$

For \hat{Y} :

$$\begin{aligned}
 E[\hat{Y}] &= E[\hat{\beta}_0 + \hat{\beta}_1 x] \\
 &= E[\hat{\beta}_0] + E[\hat{\beta}_1 x] \\
 &= \beta_0 + \beta_1 x \\
 &= E[Y]
 \end{aligned}$$

Simple Linear Regression Variances of Estimators

For $\hat{\beta}_1$:

$$\begin{aligned}
\text{Var} [\hat{\beta}_1] &= \text{Var} \left[\frac{\sum_{i=1}^n (x_i - \bar{x}) (Y_i - \bar{Y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \right] \\
&= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \cdot \text{Var} \left[\sum_{i=1}^n (x_i - \bar{x}) (Y_i - \bar{Y}) \right] \\
&= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \cdot \text{Var} \left[\sum_{i=1}^n Y_i (x_i - \bar{x}) \right] \\
&= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \cdot \sum_{i=1}^n \left\{ (x_i - \bar{x})^2 \text{Var}[Y_i] \right\} \\
&= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \cdot \sum_{i=1}^n \left\{ (x_i - \bar{x})^2 \sigma^2 \right\} \\
&= \frac{1}{\left[\sum_{i=1}^n (x_i - \bar{x})^2 \right]^2} \cdot \sigma^2 \sum_{i=1}^n (x_i - \bar{x})^2 \\
&= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
\end{aligned}$$

Note that

$$\begin{aligned}
\sum_{i=1}^n (x_i - \bar{x}) (Y_i - \bar{Y}) &= \sum_{i=1}^n (x_i Y_i - x_i \bar{Y} - \bar{x} Y_i + \bar{x} \bar{Y}) \\
&= \sum_{i=1}^n [Y_i (x_i - \bar{x}) - \bar{Y} (x_i - \bar{x})] \\
&= \sum_{i=1}^n Y_i (x_i - \bar{x}) - \sum_{i=1}^n \bar{Y} (x_i - \bar{x}) \\
&= \sum_{i=1}^n Y_i (x_i - \bar{x}) - 0
\end{aligned}$$

where

$$\begin{aligned}
\sum_{i=1}^n \bar{Y} (x_i - \bar{x}) &= \sum_{i=1}^n \bar{Y} x_i - \sum_{i=1}^n \bar{Y} \bar{x} \\
&= \bar{Y} \sum_{i=1}^n x_i - n \bar{Y} \bar{x} \\
&= \bar{Y} n \bar{x} - n \bar{Y} \bar{x} \\
&= 0
\end{aligned}$$

For $\hat{\beta}_0$:

$$\begin{aligned}
\text{Var} [\hat{\beta}_0] &= \text{Var} [\bar{Y} - \hat{\beta}_1 \bar{x}] \\
&= \text{Var} [\bar{Y}] + \bar{x}^2 \text{Var} [\hat{\beta}_1] - 2\bar{x} \text{Cov} [\bar{Y}, \hat{\beta}_1] \\
&= \text{Var} [\bar{Y}] + \bar{x}^2 \text{Var} [\hat{\beta}_1] \\
&= \frac{\sigma^2}{n} + \bar{x}^2 \cdot \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)
\end{aligned}$$

For \hat{Y} :

$$\begin{aligned}
\text{Var} [\hat{Y}] &= \text{Var} [\hat{\beta}_0 + \hat{\beta}_1 x] \\
&= \text{Var} [\hat{\beta}_0] + x^2 \text{Var} [\hat{\beta}_1] + 2x \text{Cov} [\hat{\beta}_0, \hat{\beta}_1] \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) + \frac{x^2 \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{2x \bar{x} \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \\
&= \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2} + \frac{x^2}{\sum_{i=1}^n (x_i - \bar{x})^2} - \frac{2x \bar{x}}{\sum_{i=1}^n (x_i - \bar{x})^2} \right) \\
&= \sigma^2 \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2} \right)
\end{aligned}$$

where:

$$\begin{aligned}
 \text{Cov} [\hat{\beta}_0, \hat{\beta}_1] &= \text{Cov} [\bar{Y} - \hat{\beta}_1 \bar{x}, \hat{\beta}_1] \\
 &= \text{Cov} [\bar{Y}, \hat{\beta}_1] - \bar{x} \text{Cov} [\hat{\beta}_1, \hat{\beta}_1] \\
 &= 0 - \bar{x} \text{Var} [\hat{\beta}_1] \\
 &= -\frac{\bar{x} \sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}
 \end{aligned}$$

Note that:

$$(x - \bar{x})^2 = x^2 - 2x\bar{x} + \bar{x}^2$$

Simple Linear Regression Ordinary Least Squares Equations

For simple linear regression, the SSE is

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2$$

The ordinary least squares estimates minimize this expression; they are found by taking partial derivatives with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$, and then setting each to equal 0.

For the intercept,

$$\begin{aligned}
 \frac{\partial}{\partial \hat{\beta}_0} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 &= \sum_{i=1}^n \frac{\partial}{\partial \hat{\beta}_0} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
 &= \sum_{i=1}^n 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-1) \\
 &= -2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)
 \end{aligned}$$

$$\begin{aligned}
-2 \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\Rightarrow \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\Rightarrow \sum_{i=1}^n (y_i - \hat{y}_i) &= 0 \\
\Rightarrow \sum_{i=1}^n e_i &= 0
\end{aligned}$$

For the slope,

$$\begin{aligned}
\frac{\partial}{\partial \hat{\beta}_1} \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 &= \sum_{i=1}^n \frac{\partial}{\partial \hat{\beta}_1} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \\
&= \sum_{i=1}^n 2 (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) \cdot (-x_i) \\
&= -2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)
\end{aligned}$$

$$\begin{aligned}
-2 \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\Rightarrow \sum_{i=1}^n x_i (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i) &= 0 \\
\Rightarrow \sum_{i=1}^n x_i (y_i - \hat{y}_i) &= 0 \\
\Rightarrow \sum_{i=1}^n x_i e_i &= 0
\end{aligned}$$

Therefore, the two ordinary least squares equations for a simple linear regression are

$$\sum_{i=1}^n e_i = 0$$

$$\sum_{i=1}^n x_i e_i = 0$$

3.3.0 Overview

 5m

Multiple linear regression (MLR) is the natural extension or generalization of simple linear regression (SLR). We now consider using p explanatory variables to analyze the response rather than just one explanatory variable. Consequently, many concepts overlap between SLR and MLR, but there are important distinctions.

This subsection will emphasize both similarities and differences between SLR and MLR, along with several areas unique to MLR. Beyond clarifying the minor distinctions between inference in SLR versus MLR, we will study a new type of hypothesis test known as an F test.

3.3.1 Main Idea and Assumptions

🕒 5m

In multiple linear regression, the relationship between the response and the p explanatory variables takes the form of

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon \quad (3.3.1.1)$$

In light of Equation 3.1.2.1, the chosen functional form for f is $\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p$.

As expected, $\beta_0, \beta_1, \dots, \beta_p$ require estimation using training data. The β_j 's are called **regression coefficients**. We still refer to β_0 as the intercept but will not use "slope" for β_1, \dots, β_p to avoid confusion.

Coach's Remarks

Previously, we defined j as the index for the explanatory variables, running from 1 to p . But for the sake of convenience in multiple linear regression, we will permit j to run from 0 to p to include the intercept parameter. One could argue that there are effectively $p + 1$ explanatory variables rather than just p ; the additional explanatory variable is $x_0 = 1$ for all observations, with the intercept β_0 as its coefficient.

Here are the assumptions of a multiple linear regression model:

1. $Y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} + \varepsilon_i$
2. $x_{i,j}$'s are non-random
3. $E[\varepsilon_i] = 0$
4. $\text{Var}[\varepsilon_i] = \sigma^2$
5. ε_i 's are independent
6. ε_i 's are normally distributed
7. The predictor x_j is not a linear combination of the other p predictors, for $j = 0, 1, \dots, p$

The first six assumptions mirror the simple linear regression assumptions in Section 3.2.1. They lead to similar key conclusions:

- Y_i 's are independent normal random variables

- $E[Y_i] = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}$
- $\text{Var}[Y_i] = \sigma^2$ (i.e. homoscedasticity)

The 7th assumption serves to avoid redundant explanatory variables. If an x_j is a linear combination of the other predictors, then it will not aid in understanding the response beyond what can be learned from the other predictors.

3.3.2 Parameter Estimates and Matrix Notation

Estimation of β_j

Let

- $\hat{\beta}_j$ denote the estimate of β_j , for $j = 0, 1, \dots, p$
- $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$

The estimates $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ are determined by ordinary least squares, which minimizes the SSE

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_p x_{i,p})^2$$

using training data. While ordinary least squares is the familiar optimization problem (i.e. take partial derivatives; set equal to 0; solve), the formulas for every $\hat{\beta}_j$ are challenging to express algebraically. Thus, this topic will rely heavily on interpreting statistical software outputs.

Coach's Remarks

The 7th multiple linear regression assumption ensures that the ordinary least squares estimates of the regression coefficients are unique.

Here is the R output from regressing Commute on Departure, Temp, Precip Chance, and Police from the Commuting Chris scenario:

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.98958   2.64219  9.836  3.7e-16 ***
Departure   -0.63290   0.13560 -4.667  1.0e-05 ***
Temp        0.05584   0.03117  1.791   0.0764 .
Precip.Chance 0.04208   0.01748  2.408   0.0180 *
Police      4.00287   0.32881 12.174 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.283 on 95 degrees of freedom
Multiple R-squared:  0.6782, Adjusted R-squared:  0.6646
F-statistic: 50.05 on 4 and 95 DF,  p-value: < 2.2e-16

```

Therefore, the fitted equation is

$$\hat{y} = 25.98958 - 0.6329x_1 + 0.05584x_2 + 0.04208x_3 + 4.00287x_4$$

where

- x_1 is the time of departure,
- x_2 is the temperature at the time of departure,
- x_3 is the chance of precipitation at the time of departure, and
- x_4 is the number of police vehicles along the commute route.

This equation cannot be easily drawn or visualized because it is a hyperplane in 5-dimensional space.

Here are the interpretations of the first three regression coefficient estimates for this model:

- 25.98958 minutes is the expected commute time when the departure time is midnight, the temperature is 0 °F, the chance of precipitation is 0%, and no police vehicles are found along the commute route.
 - In general, $\hat{\beta}_0$ is \hat{y} when all predictors x_1, \dots, x_p are 0.
- For every hour later in departure time, the expected commute time decreases by 0.6329 minutes (or about 38 seconds), assuming all other predictors are held constant.
- For every 1 °F increase in temperature, the expected commute time increases by 0.05584 minutes (or about 3.4 seconds), assuming all other predictors are held constant.
 - In general, $\hat{\beta}_j$ is the change in \hat{y} for every unit increase in predictor x_j , assuming all other predictors are held constant, for $j = 1, \dots, p$.

Example 3.3.2.1

Julianne studies the price of gas at 18 gas stations. For each station, she has a record of its daily average gas price, its daily average demand (a function of gallons sold), and its daily average maintenance cost. She obtains the following result from running a multiple linear regression:

	Coefficients	Standard Error
Intercept	5.4511	2.6649
Demand	-1.0308	0.5073
Maintenance Cost	0.2877	0.0369

Calculate the predicted daily average gas price at a station with a daily average demand of 3.333 and a daily average maintenance cost of 2.25.

Solution

Based on the output, the prediction formula is

$$\hat{y} = 5.4511 - 1.0308x_1 + 0.2877x_2$$

where

- x_1 is the daily average demand, and
- x_2 is the daily average maintenance cost.

When $x_1 = 3.333$ and $x_2 = 2.25$,

$$\begin{aligned}\hat{y} &= 5.4511 - 1.0308(3.333) + 0.2877(2.25) \\ &= \mathbf{2.663}\end{aligned}$$



Example 3.3.2.2

From a regression based on the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

you are told:

- The predicted response is 28.75 when $x_1 = 5$, $x_2 = -2.3$, and $x_3 = 18$.
- The predicted response is 47.11 when $x_1 = 5$, $x_2 = 1.7$, and $x_3 = 18$.
- The predicted response is 52.87 when $x_1 = 5$, $x_2 = 1.7$, and $x_3 = 13$.
- The predicted response is 30.46 when $x_1 = 6$, $x_2 = -2.3$, and $x_3 = 13$.

Determine which statements are true.

- I. The predicted response increases by 4.59 for every unit increase in x_2 with fixed inputs for x_1 and x_3 .
- II. The predicted response increases by 4.05 for every unit increase in x_1 with fixed inputs for x_2 and x_3 .
- III. If x_1 increases by 3, x_2 remains unchanged, and x_3 decreases by 10, the predicted response will decrease by 0.63.

Solution

The four bullets provide the following equations:

$$28.75 = \hat{\beta}_0 + \hat{\beta}_1(5) + \hat{\beta}_2(-2.3) + \hat{\beta}_3(18)$$

$$47.11 = \hat{\beta}_0 + \hat{\beta}_1(5) + \hat{\beta}_2(1.7) + \hat{\beta}_3(18)$$

$$52.87 = \hat{\beta}_0 + \hat{\beta}_1(5) + \hat{\beta}_2(1.7) + \hat{\beta}_3(13)$$

$$30.46 = \hat{\beta}_0 + \hat{\beta}_1(6) + \hat{\beta}_2(-2.3) + \hat{\beta}_3(13)$$

I is true. Subtracting the second equation from the first, we get

$$28.75 - 47.11 = \hat{\beta}_2(-2.3) - \hat{\beta}_2(1.7)$$

$$\Rightarrow -18.36 = -4\hat{\beta}_2$$

$$\Rightarrow \hat{\beta}_2 = 4.59$$

II is false because the predicted response would decrease by 4.05. Subtracting the fourth equation from the third, we get

$$52.87 - 30.46 = \hat{\beta}_1(5) + \hat{\beta}_2(1.7) - \hat{\beta}_1(6) - \hat{\beta}_2(-2.3)$$

$$\Rightarrow 22.41 = -\hat{\beta}_1 + 4\hat{\beta}_2$$

$$\Rightarrow \hat{\beta}_1 = -[22.41 - 4(4.59)] = -4.05$$

III is true. Subtracting the third equation from the second, we get

$$47.11 - 52.87 = \hat{\beta}_3(18) - \hat{\beta}_3(13)$$

$$\Rightarrow -5.76 = 5\hat{\beta}_3$$

$$\Rightarrow \hat{\beta}_3 = -1.152$$

Then, the change in the predicted response if x_1 increases by 3 and x_3 decreases by 10 is

$$\begin{aligned} 3\hat{\beta}_1 - 10\hat{\beta}_3 &= 3(-4.05) - 10(-1.152) \\ &= -0.63 \end{aligned}$$

Therefore, **only I and III are true.**



Estimation of σ^2

The estimate of σ^2 is the MSE, which has the formula

$$\text{MSE} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1} \quad (3.3.2.1)$$

for multiple linear regression. Note that the denominator is $n - p - 1$ (rather than Equation 3.2.2.3's denominator of $n - 2$). This leads to an unbiased estimator of σ^2 .

Coach's Remarks

Realize that the denominator of the MSE formula is

$$n - p - 1 = n - (p + 1)$$

where $p + 1$ is the number of estimated regression coefficients ($\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$). This is consistent with the MSE formula for a simple linear regression, i.e. $p = 1$; the denominator is $n - 2$ because there are 2 estimated regression coefficients ($\hat{\beta}_0$ and $\hat{\beta}_1$).

The square root of the MSE is still referred to as the residual standard error.

$$\text{residual standard error} = \sqrt{\text{MSE}}$$

From the R output, the residual standard error for this Commuting Chris setup of four predictors is 3.283. Squaring the (unrounded) residual standard error gives an MSE of 10.7812. In addition, the output gives $\text{df} = 95$, which is the denominator of the MSE. This makes sense, as this model uses 100 observations to calculate 5 $\hat{\beta}_j$'s.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.98958  2.64219  9.836  3.7e-16 ***
Departure   -0.63290  0.13560 -4.667  1.0e-05 ***
Temp        0.05584  0.03117  1.791   0.0764 .
Precip.Chance 0.04208  0.01748  2.408   0.0180 *
Police      4.00287  0.32881 12.174 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.283 on 95 degrees of freedom
Multiple R-squared:  0.6782, Adjusted R-squared:  0.6646
F-statistic: 50.05 on 4 and 95 DF,  p-value: < 2.2e-16
```

Matrix Notation

The matrix formulas from Section 3.2.3 are the same for multiple linear regression, with the expected tweaks in the definitions of some matrices. Thus, the model equation in relation to the

training data is

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where

- \mathbf{X} is the $n \times (p + 1)$ design matrix, i.e. $\begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix}$, and
- $\boldsymbol{\beta}$ is the vector of $p + 1$ regression coefficients.

Equations 3.2.3.1 to 3.2.3.3 still hold:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X} \hat{\boldsymbol{\beta}} \\ &= \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \\ &= \mathbf{H} \mathbf{y} \end{aligned}$$

Realize this means that $(\mathbf{X}^T \mathbf{X})^{-1}$ is a $(p + 1) \times (p + 1)$ matrix.

Example 3.3.2.3

A linear regression is modeled by $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Let \mathbf{x} represent a vector of feature inputs, i.e. $\mathbf{x}^T = [1 \quad x_1 \quad \cdots \quad x_p]$.

You are given

$$(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \begin{bmatrix} 254.687 \\ 0.274 \\ 3.632 \\ -4.894 \\ 11.755 \end{bmatrix}$$

Determine which statements are true.

- I. This model uses five features to study the response.
- II. 3.632 is the predicted response corresponding to the 3rd row of \mathbf{X} .
- III. $(\mathbf{X}^T \mathbf{X})^{-1}$ is a 5×5 matrix.
- IV. $\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$ is the predicted response for the inputs given by \mathbf{x} .

Solution

We have

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \\ \hat{\beta}_4 \end{bmatrix} = \begin{bmatrix} 254.687 \\ 0.274 \\ 3.632 \\ -4.894 \\ 11.755 \end{bmatrix}$$

I is false because there are four features (i.e. $p = 4$) in this model.

II is false because $\hat{\beta}_2 = 3.632$, not \hat{y}_3 .

III is true. Since $p = 4$, we know that $(\mathbf{X}^T \mathbf{X})^{-1}$ is a 5×5 matrix. Alternatively, note that $(\mathbf{X}^T \mathbf{X})^{-1}$ is the first matrix in the formula for $\hat{\boldsymbol{\beta}}$. Therefore, both $(\mathbf{X}^T \mathbf{X})^{-1}$ and $\hat{\boldsymbol{\beta}}$ must have the same number of rows, due to the nature of matrix multiplication. Knowing that $(\mathbf{X}^T \mathbf{X})^{-1}$ has 5 rows, then it must have 5 columns being a square matrix.

IV is true.

$$\begin{aligned}
 \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} &= \mathbf{x}^T \hat{\boldsymbol{\beta}} \\
 &= [1 \quad x_1 \quad \cdots \quad x_4] \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_4 \end{bmatrix} \\
 &= \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_4 x_4 \\
 &= \hat{y}
 \end{aligned}$$

Therefore, **only III and IV are true.**



Coach's Remarks

As a specific application of statement IV, if \mathbf{x}_i^T represents the i^{th} row of \mathbf{X} , then

$$\hat{y}_i = \mathbf{x}_i^T \hat{\boldsymbol{\beta}} = \mathbf{x}_i^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

This agrees with $\hat{\mathbf{y}} = \mathbf{X} \hat{\boldsymbol{\beta}} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, which performs the same calculation for every row of \mathbf{X} .

3.3.3 Other Numerical Results

The formulas for residuals, the three sums of squares, and R^2 remain the same as presented in Section 3.2.4. The only distinction is that $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$ rather than $\hat{\beta}_0 + \hat{\beta}_1 x$ from simple linear regression.

$$e_i = y_i - \hat{y}_i$$

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2, \quad \text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \quad \text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

$$\text{SST} = \text{SSR} + \text{SSE}$$

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

However, Equation 3.2.4.4 does not hold for multiple linear regression. This is because \hat{y} is not a linear function of only one x_j ; it is not perfectly correlated to any specific x_j .

$$R^2 \neq r_{x_j, y}^2, \quad j = 1, \dots, p$$

Adjusted R²

To better understand the next topic, we should be familiar with the concept of **nested models**. These are often a collection of models that share a set of predictors, where each model is a subset of every model with more predictors. As a result, the subsets follow a sequential order. To demonstrate, consider the following models and their predictors:

Model	Predictors
A	x_1
B	x_1, x_2
C	x_1, x_2, x_3

Model	Predictors
D	x_1, x_2, x_4, x_5

Models A, B, and C are nested models that share the predictors x_1, x_2 , and x_3 . This is because the predictors of the three models form ordered subsets. Similarly, Models A, B, and D are nested models that share the predictors x_1, x_2, x_4 , and x_5 . However, since Model C predictors are not a subset of Model D predictors, the four models collectively are not nested.

Now, recall from

- Section 3.1.2: higher flexibility often comes from more free parameters.
- Section 3.1.3: higher flexibility leads to a lower training MSE.

For nested multiple linear regression models, p is a flexibility measure. Increasing the number of predictors in a model leads to more β_j 's, thus creating a more flexible \hat{f} . Further note that SSE is the numerator of the training MSE. Therefore, p and SSE are inversely related, since higher flexibility leads to a lower training MSE. With every additional predictor, SSE decreases. In turn, R^2 will increase, since

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

As R^2 is the proportion of variability explained by the regression, a high R^2 is desirable. But simply adding more predictors will not lead to the best model. Doing so will increase R^2 , even when unhelpful and/or nonsensical predictors are included. This parallels how the lowest training MSE does not imply an accurate model.

Consequently, it is better to compare models (nested or otherwise) using *adjusted R^2* rather than R^2 . The formula for adjusted R^2 is

$$\begin{aligned} R_{\text{adj.}}^2 &= 1 - \frac{\text{MSE}}{s_y^2} && (3.3.3.1) \\ &= 1 - \frac{\text{SSE} \div (n - p - 1)}{\text{SST} \div (n - 1)} \\ &= 1 - \frac{\text{SSE}}{\text{SST}} \cdot \frac{n - 1}{n - p - 1} \\ &= 1 - (1 - R^2) \left(\frac{n - 1}{n - p - 1} \right) \end{aligned}$$

Here are some key facts regarding $R^2_{\text{adj.}}$:

- A high $R^2_{\text{adj.}}$ is desirable, which is similar to R^2 .
- The best model by $R^2_{\text{adj.}}$ is the model with the lowest MSE of Equation 3.3.2.1.
- A model's $R^2_{\text{adj.}}$ is **less than** its R^2 except for two cases; $R^2_{\text{adj.}} = R^2$ occurs when $R^2 = 1$ or when $p = 0$.
- $\frac{n-1}{n-p-1}$ must be greater than 1 (except for when $p = 0$), so it inflates the proportion of unexplained variability. We may say that $R^2_{\text{adj.}}$ is a shrunken value of R^2 based on the number of predictors. This makes it possible for $R^2_{\text{adj.}}$ to decrease for larger values of p .
- $R^2_{\text{adj.}}$ does not have to be between 0 and 1. This means it cannot be interpreted as a proportion.

The R output for the Commuting Chris setup with four predictors shows $R^2 = 0.6782$ and $R^2_{\text{adj.}} = 0.6646$.

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 25.98958   2.64219   9.836 3.7e-16 ***
Departure   -0.63290   0.13560  -4.667 1.0e-05 ***
Temp        0.05584   0.03117   1.791  0.0764 .  
Precip.Chance 0.04208   0.01748   2.408  0.0180 *  
Police       4.00287   0.32881  12.174 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.283 on 95 degrees of freedom
Multiple R-squared:  0.6782, Adjusted R-squared:  0.6646 
F-statistic: 50.05 on 4 and 95 DF,  p-value: < 2.2e-16
```

This means 67.82% of the variability in Commute can be explained by the model with these four predictors. However, $R^2_{\text{adj.}}$ does not have as simple of an interpretation. Even so, it is a measure of model quality that is useful for comparing models, especially with different values of p . Model comparison is a topic more extensively covered in Section 3.6.

Example 3.3.3.1

Six features are used to run a multiple linear regression on 20 observations. You are given the following results from the regression:

- The residual standard error is 9.1435.
- The coefficient of determination is 0.5833.

Find:

1. Adjusted R^2 .
2. The unbiased sample standard deviation of the response, s_y .

Solution to (1)

We are given $n = 20$ and $p = 6$. To calculate $R_{\text{adj.}}^2$, inflate the proportion of unexplained variability in the response by a factor of

$$\frac{n - 1}{n - p - 1} = \frac{20 - 1}{20 - 6 - 1} = \frac{19}{13}$$

Therefore,

$$\begin{aligned} R_{\text{adj.}}^2 &= 1 - (1 - R^2) \left(\frac{n - 1}{n - p - 1} \right) \\ &= 1 - (1 - 0.5833) \left(\frac{19}{13} \right) \\ &= 1 - 0.609 \\ &= \mathbf{0.391} \end{aligned}$$



Solution to (2)

Having solved for $R_{\text{adj.}}^2$ in Part (1), the goal can be found using Equation 3.3.3.1.

$$R_{\text{adj.}}^2 = 1 - \frac{\text{MSE}}{s_y^2}$$

$$\Rightarrow s_y = \sqrt{\frac{\text{MSE}}{1 - R_{\text{adj.}}^2}}$$

$$= \frac{9.1435}{\sqrt{1 - 0.391}}$$

$$= \mathbf{11.716}$$

■

Alternative Solution to (2)

Instead of using $R_{\text{adj.}}^2$, we can use the unbiased sample standard deviation formula.

$$s_y = \sqrt{\frac{\sum_{i=1}^{20} (y_i - \bar{y})^2}{20 - 1}} = \sqrt{\frac{\text{SST}}{19}}$$

SST can be found using R^2 as follows:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}}$$

$$\Rightarrow \text{SST} = \frac{\text{SSE}}{1 - R^2}$$

$$= \frac{\text{MSE} (n - p - 1)}{1 - R^2}$$

$$= \frac{9.1435^2 \cdot 13}{1 - 0.5833}$$

$$= 2,608.223$$

Therefore, the answer is

$$s_y = \sqrt{\frac{2,608.223}{19}} = 11.716$$



Example 3.3.3.2

A statistical software produces the following output from running two different linear models on the same data:

Model	R^2	Adjusted R^2
A	0.7254	0.6538
B	0.7631	?

Determine which statements are true.

- I. It is appropriate to conclude that Model B is preferred because it explains more of the response's variability than Model A does.
- II. It is impossible for Model B predictors to be a strictly smaller subset of Model A predictors.
- III. If Model B's adjusted R^2 is greater than 0.6538, then it is appropriate to conclude that Model B is preferred.
- IV. If Model B's adjusted R^2 is less than 0.6538, then Model B must have more predictors than Model A.
- V. If Models A and B have the same number of predictors, then it is appropriate to conclude that Model B is preferred.

Solution

I is false. Although Model B does explain more of the variability than Model A, this does not

mean that Model B is preferred. When comparing models, we should not use R^2 because it does not correct for flexibility.

II is true. SSE decreases as more predictors are added. So, if Model B predictors are a strictly smaller subset of Model A predictors, then SSE would be smaller for Model A than for Model B. However, this cannot be true based on Model B's larger R^2 , which implies that Model B has a smaller SSE.

III is true. A model with a larger $R_{\text{adj.}}^2$ is preferred.

IV is true. Since Model B's R^2 is greater than Model A's R^2 , there is only one way for Model B's $R_{\text{adj.}}^2$ to be lower than Model A's $R_{\text{adj.}}^2$. — Model B has a larger factor of $\frac{n-1}{n-p-1}$. This means Model B must then have a greater p .

V is true. When two models have the same number of predictors, $R_{\text{adj.}}^2$ becomes redundant because it inflates the proportion of unexplained variability by the same factor. This makes a comparison using $R_{\text{adj.}}^2$ equivalent to a comparison using R^2 . Furthermore, having the same p means that $R_{\text{adj.}}^2$ for Model B can actually be solved to equal 0.7013. In conclusion, Model B is preferred.

Therefore, **only II, III, IV and V are true.**



3.3.4 Special Variables Types

There are three special types of predictors that deserve attention:

- Higher-order variables
- Dummy variables
- Interactions

Higher-Order Terms

Consider the following relationship between the response and **one** explanatory variable, x_j :

$$Y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \dots + \beta_k x_j^k + \varepsilon$$

This model suggests that the response and explanatory variables are systematically related by a k^{th} order polynomial. This model is a specific instance of the multiple linear regression model (i.e. Equation 3.3.1.1), where

- $p = k$,
- $x_1 = x_j$,
- $x_2 = x_j^2$,
- ...
- $x_p = x_j^k$.

This means polynomial relationships fall under the umbrella of multiple linear regression, with regression coefficients estimated by ordinary least squares, and so on. The main difference is the interpretation of the $\hat{\beta}_j$'s. In contrast to a linear function, polynomials do not have a constant slope, meaning they do not change consistently by unit increases of its variable. It is not simple to isolate and explain the effect of a single $\hat{\beta}_j$.

By extension, higher-order variables of one predictor can certainly be added to a model with **other** predictors, with or without their own higher-order variables.

The following output shows a regression using Departure and Police from the Commuting Chris scenario:

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.04271   7.20325  5.975 3.89e-08 ***
Departure   -3.14172   1.37165 -2.290  0.0242 *
Departure^2  0.11579   0.06264  1.849  0.0676 .
Police      4.03762   0.29798 13.550 < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.306 on 96 degrees of freedom
Multiple R-squared:  0.6703, Adjusted R-squared:  0.66
F-statistic: 65.05 on 3 and 96 DF,  p-value: < 2.2e-16

```

Note that the square of Departure is included as a predictor.

The 57th observation has the following recorded values:

Commute	Departure	Police
32.867	7.6	2

Calculate the residual of the 57th observation.

The fitted equation is

$$\hat{y} = 43.04271 - 3.14172x_1 + 0.11579x_1^2 + 4.03762x_2$$

where

- x_1 is the time of departure, and
- x_2 is the number of police vehicles along the commute route.

Therefore,

$$\begin{aligned}\hat{y}_{57} &= 43.04271 - 3.14172(7.6) + 0.11579(7.6^2) + 4.03762(2) \\ &= 33.929\end{aligned}$$

$$\begin{aligned}e_{57} &= y_{57} - \hat{y}_{57} \\ &= 32.867 - 33.929 \\ &= -1.062\end{aligned}$$

Here are a few more ideas to consider:

- When constructing a model with a k^{th} order polynomial of x_j , the default is to include a term for each power of x_j from 1 to k . For example, a model with a 3rd order polynomial of x_j

should have the equation

$$Y = \beta_0 + \beta_1 x_j + \beta_2 x_j^2 + \beta_3 x_j^3 + \varepsilon$$

rather than

$$Y = \beta_0 + \beta_3 x_j^3 + \varepsilon$$

The second equation assumes that β_1 and β_2 are both 0. This restricts the possible cubic shapes, thus hindering the data from estimating the "best" possible cubic.

- Higher-order variables can be generalized as transformations of an explanatory variable. For instance, a quadratic variable comes from transforming a variable by squaring. A long list of other transformations – such as taking the natural log or the square root – are certainly possible.
- It might seem contradictory for polynomial (i.e. non-linear) relationships to be within the scope of **linear** regression. However, the "linear" here does not refer to the predictors used, but rather the regression parameters.

Dummy Variables

As the x_j 's only take on numerical values, we require a method to translate **categorical** predictors for linear regression. To address this, dummy variables are often used. A **dummy variable** takes on the values 0 or 1, making it similar to an indicator function, $I(\cdot)$. Specifically, for a categorical predictor, a dummy variable is defined to equal

- 1 when an observation is classified as category c , or
- 0 when an observation is classified as a category other than c .

Coach's Remarks

While a dummy variable can be defined using values other than 0 and 1, we will not discuss those variants in this manual. Their effects can be easily inferred from understanding the way dummy variables operate.

Realize that it takes $w - 1$ dummy variables to represent a predictor with w categories. For example, when a predictor has two categories, only one dummy variable is needed. In this situation, let x_1 be a

dummy variable which equals 1 when an observation is classified as category 1. Otherwise, x_1 equals 0. As shown in the table below, both categories are accounted for:

Category, c	x_1
1	1
2	0

In general, for a predictor with w categories, let

$$x_c = \begin{cases} 1, & \text{category } c \text{ observation} \\ 0, & \text{otherwise} \end{cases}, \quad c = 1, 2, \dots, w-1$$

As shown in the table below, all w categories are accounted for:

Category, c	x_1	x_2	...	x_{w-1}
1	1	0	...	0
2	0	1	...	0
:	:	:	:	:
$w-1$	0	0	...	1
w	0	0	...	0

As a result, there will always be one category that is referenced when all $w-1$ dummy variables equal 0. This category acts as a **baseline category** for the other categories. Choosing the baseline category is arbitrary.

The following output shows a regression using Precip Chance and Season from the Commuting Chris scenario:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.19931   1.45313 16.653 <2e-16 ***
Precip_Chance 0.03831   0.02297  1.668  0.0986 .  
SeasonSpring -1.13806   1.48961 -0.764  0.4468    
SeasonSummer  1.12868   1.70719  0.661  0.5101    
SeasonWinter  3.37147   1.56541  2.154  0.0338 *  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.208 on 95 degrees of freedom
Multiple R-squared:  0.1905, Adjusted R-squared:  0.1564 
F-statistic: 5.588 on 4 and 95 DF,  p-value: 0.0004396
```

Predict the commute time when there is a 40% chance of precipitation in the

- spring season.

- summer season.
- winter season.
- fall season.

Since there are four categories in Season, three dummy variables appear in this model. The output reveals that there is no dummy variable dedicated to Fall, which means Fall was selected as the baseline category.

The fitted equation is

$$\hat{y} = 24.19931 + 0.03831x_1 - 1.13806x_2 + 1.12868x_3 + 3.37147x_4$$

where

- x_1 is the chance of precipitation,
- x_2 is the dummy variable associated with Spring,
- x_3 is the dummy variable associated with Summer, and
- x_4 is the dummy variable associated with Winter.

With a 40% chance of precipitation in the spring,

$$\begin{aligned}\hat{y} &= 24.19931 + 0.03831(40) - 1.13806(1) + 1.12868(0) + 3.37147(0) \\ &= \mathbf{24.594}\end{aligned}$$

With a 40% chance of precipitation in the summer,

$$\begin{aligned}\hat{y} &= 24.19931 + 0.03831(40) - 1.13806(0) + 1.12868(1) + 3.37147(0) \\ &= \mathbf{26.860}\end{aligned}$$

With a 40% chance of precipitation in the winter,

$$\begin{aligned}\hat{y} &= 24.19931 + 0.03831(40) - 1.13806(0) + 1.12868(0) + 3.37147(1) \\ &= \mathbf{29.103}\end{aligned}$$

With a 40% chance of precipitation in the fall,

$$\begin{aligned}\hat{y} &= 24.19931 + 0.03831(40) - 1.13806(0) + 1.12868(0) + 3.37147(0) \\ &= \mathbf{25.732}\end{aligned}$$

In addition, let's interpret the regression coefficient estimates:

- In the fall, note that

$$\begin{aligned}\hat{y} &= 24.19931 + 0.03831x_1 - 1.13806(0) + 1.12868(0) + 3.37147(0) \\ &= 24.19931 + 0.03831x_1\end{aligned}$$

Therefore, 24.19931 is the estimated intercept (i.e. expected commute time with a 0% chance of precipitation) in the fall.

- In the spring, note that

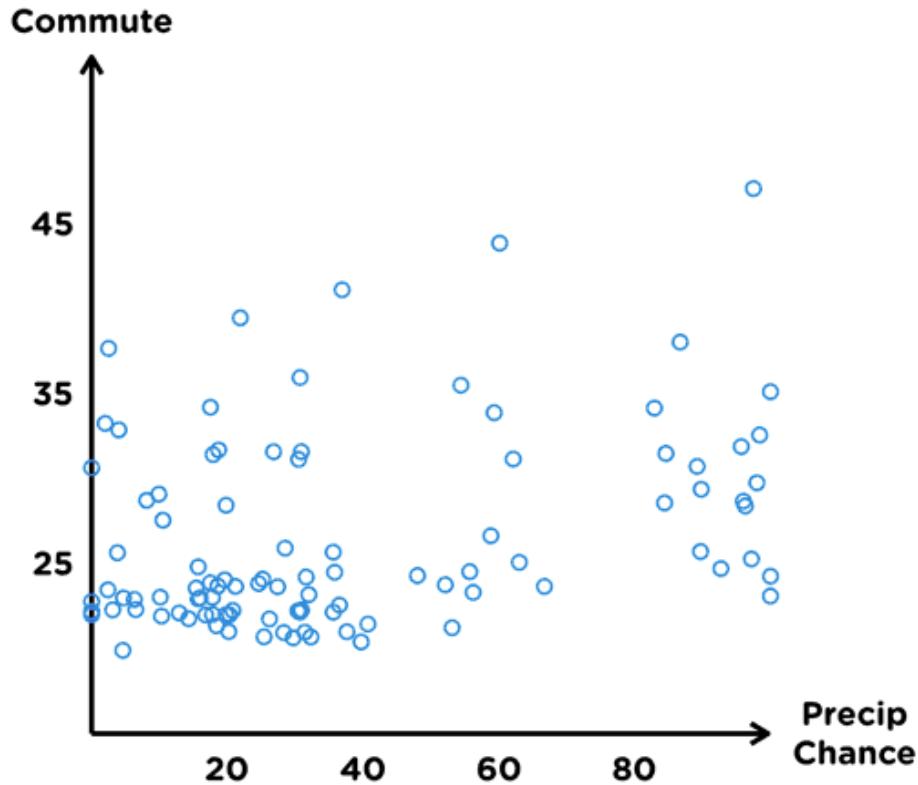
$$\begin{aligned}\hat{y} &= 24.19931 + 0.03831x_1 - 1.13806(1) + 1.12868(0) + 3.37147(0) \\ &= (24.19931 - 1.13806) + 0.03831x_1 \\ &= 23.06125 + 0.03831x_1\end{aligned}$$

Therefore, the estimated intercept decreases by 1.13806 in the spring **relative to fall**.

The same explanation applies to increases in the estimated intercept of 1.12868 for summer and 3.37147 for winter, each relative to fall.

- For every 1% increase in the chance of precipitation, the expected commute time increases by 0.03831 minutes (or about 2.3 seconds), regardless of season.

The animation below cycles through the scatterplot of Commute against Precip Chance with different observations emphasized based on Season. The fitted equation can be expressed as four distinct linear functions of Precip Chance – one for each season. Notice that all four linear functions have the same slope, but different intercepts.



In conclusion, dummy variables adjust the intercept according to the different categories. As is, they do not affect the regression coefficients of other predictors.

Coach's Remarks

To see why the choice of baseline category is arbitrary, consider the output for the model that uses Winter as the baseline rather than Fall:

Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	Signif. codes:
(Intercept)	27.57078	1.86181	14.809	<2e-16 ***	0 ****
Precip.Chance	0.03831	0.02297	1.668	0.0986 .	0.05 .
SeasonFall	-3.37147	1.56541	-2.154	0.0338 *	0.1 *.
SeasonSpring	-4.50953	1.72407	-2.616	0.0104 *	0.05 **
SeasonSummer	-2.24279	1.99320	-1.125	0.2633	0.1 .

Signif. codes: 0 **** 0.001 *** 0.01 ** 0.05 . 0.1 . 1					
Residual standard error: 5.208 on 95 degrees of freedom					
Multiple R-squared: 0.1905, Adjusted R-squared: 0.1564					
F-statistic: 5.588 on 4 and 95 DF, p-value: 0.0004396					

All of the outputs are the same, except for the details involving the intercept and the coefficients for the three dummy variables. However, these differences are expected since the four regression coefficients take on different interpretations now that Winter is the baseline category.

Even so, the four fitted equations for each season are exactly the same as the equations from the model with a Fall baseline. Thus, the choice of baseline category is arbitrary from a prediction standpoint.

Interactions

An *interaction* is a product of different explanatory variables. It allows for dependence between predictors to be included in a model. To illustrate, consider this model equation with two explanatory variables and their interaction:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

One way to rewrite the equation is

$$Y = \beta_0 + (\beta_1 + \beta_3 x_2) x_1 + \beta_2 x_2 + \varepsilon$$

As x_1 changes, the response is affected by both β_1 and $\beta_3 x_2$. Therefore, the influence of x_1 on the response is allowed to **depend on** the value of x_2 (and vice versa).

While interactions are possible for any combination of explanatory variables, it is noteworthy to study the interaction between a quantitative predictor and a dummy variable. As an application of the interaction model above, let x_2 be a dummy variable for a categorical predictor with two classes. For the class where $x_2 = 0$, the model equation is

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 \cdot 0) x_1 + \beta_2 (0) + \varepsilon \\ &= \beta_0 + \beta_1 x_1 + \varepsilon \end{aligned}$$

whereas for the class where $x_2 = 1$, the model equation is

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 \cdot 1) x_1 + \beta_2 (1) + \varepsilon \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1 + \varepsilon \end{aligned}$$

Previously, dummy variables could only affect the intercept of a model. Here, the dummy variable still influences the intercept via β_2 , but it also alters the impact of x_1 on the response via β_3 . In this case, **both** the intercept and slope of the linear f are allowed to change depending on whether the observations are in class $x_2 = 0$ or $x_2 = 1$.

Let's extend the model with Precip Chance and Season to include interaction terms. Here is that model's output in R:

```

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  24.76495  1.55054 15.972 <2e-16 ***
Precip.Chance 0.02549  0.02641  0.965  0.3370    
SeasonSpring   1.72271  3.35543  0.513  0.6089    
SeasonSummer    2.66374  2.62485  1.015  0.3129    
SeasonWinter   -3.28696  3.81271 -0.862  0.3909    
Precip.Chance:SeasonSpring -0.12590  0.11718 -1.074  0.2855    
Precip.Chance:SeasonSummer -0.15970  0.15015 -1.064  0.2903    
Precip.Chance:SeasonWinter  0.10350  0.05615  1.843  0.0685 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.116 on 92 degrees of freedom
Multiple R-squared:  0.2433, Adjusted R-squared:  0.1857 
F-statistic: 4.226 on 7 and 92 DF,  p-value: 0.0004407

```

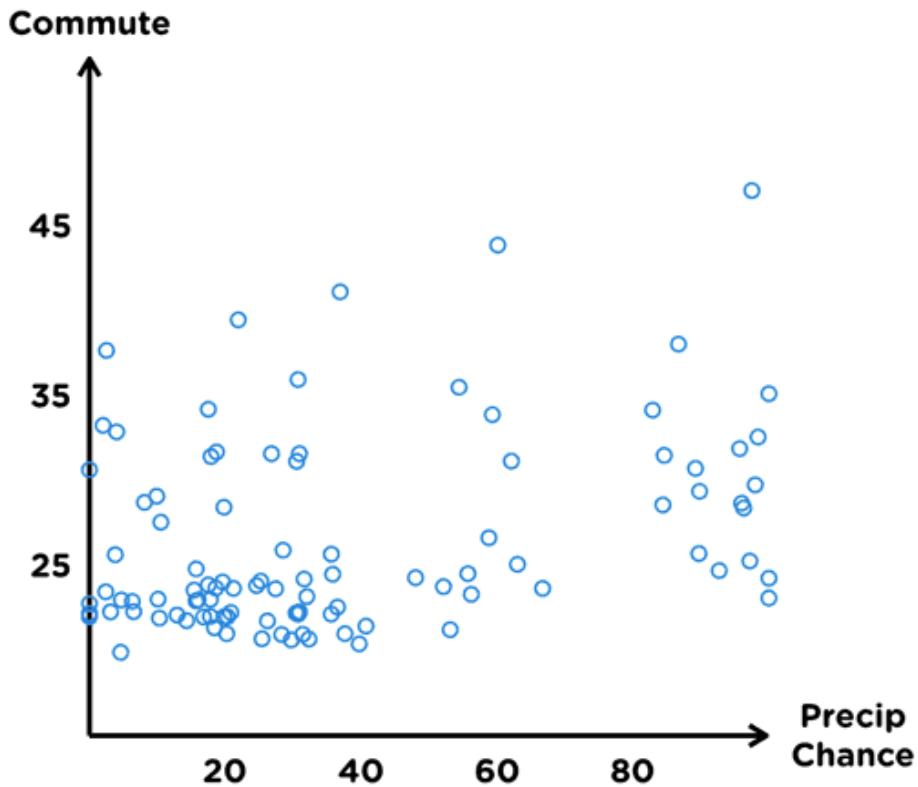
Note that R denotes an interaction by using a colon between variable names. Hence, the fitted equation is

$$\hat{y} = 24.76495 + 0.02549x_1 + 1.72271x_2 + 2.66374x_3 - 3.28696x_4 \\ -0.1259x_1x_2 - 0.1597x_1x_3 + 0.1035x_1x_4$$

where

- x_1 is the chance of precipitation,
- x_2 is the dummy variable associated with Spring,
- x_3 is the dummy variable associated with Summer, and
- x_4 is the dummy variable associated with Winter.

The animation below cycles through the scatterplot of Commute against Precip Chance with different observations emphasized based on Season. Once again, the fitted equation contains four distinct linear functions of Precip Chance. Having introduced the interaction terms, both the intercepts and slopes are allowed to vary.



Example 3.3.4.1

A researcher from Agonite Airlines wants to learn about customer luggage weight per bag (in pounds) from 60 customers. Running a multiple linear regression produces:

	Coefficients	Standard Error
Intercept	41.9603	3.463184
Age of customer (years)	-0.69064	0.143536
Squared age of customer (years^2)	0.00612	0.001431
International flight (1 = Yes, 0 = No)	16.86309	2.097807
Number of connecting flights	-2.75903	0.942170

Determine which statements are true.

- I. This model has 56 degrees of freedom associated with error.
- II. All else being equal, a 30-year-old customer is expected to fly with 2.622 pounds of luggage per bag more than a 40-year-old customer.
- III. For a customer on an international flight, the model's intercept is estimated to equal 58.823 pounds.

- IV. For a customer on an international flight, every increase in the number of connecting flights by 1 increases the predicted luggage weight per bag by 14.104 pounds, holding the customer's age constant.
- V. Every increase in a customer's age by 1 lowers the predicted luggage weight per bag by 0.6845 pounds, holding all other predictors constant.

Solution

I is false. The number of observations is 60, and the number of estimated regression coefficients is 5. The degrees of freedom is their difference, which is 55.

II is true. The goal is to calculate

$$\begin{aligned} & -0.69064(30) + 0.00612(30^2) - [-0.69064(40) + 0.00612(40^2)] \\ &= -0.69064(30 - 40) + 0.00612(30^2 - 40^2) \\ &= 2.6224 \end{aligned}$$

III is true. The estimated intercept for a customer on an international flight is

$$41.9603 + 16.86309 = 58.82339$$

IV is false. Increasing the number of connecting flights by 1 decreases the predicted luggage weight per bag by 2.759 pounds, holding the other predictors constant. This is true regardless of whether the flight is domestic or international. Without an interaction term, a dummy variable cannot affect the regression coefficient of another predictor.

V is false. Since the model assumes a quadratic relationship between the predicted luggage weight per bag and customer age, there is no constant slope between the two. This means the change in predicted luggage weight per bag is not fixed for customer age increments of 1; it depends on the ages being compared.

Therefore, **only II and III are true.**



Example 3.3.4.2

The fitted equation for a multiple linear regression model is

$$\hat{y} = -4.812 + 0.811x_1 + 2.049x_2 - 1.586x_3 - 3.143x_4 + 0.1875x_1x_4$$

where

- x_1 is a quantitative variable,
- x_2, x_3 are dummy variables for categorical variable A, and
- x_4 is a dummy variable for categorical variable B.

Determine which statements are true.

- I. The categorical variables divide the observations into one of four classes.
- II. For observations classified by $x_2 = 0, x_3 = 1$, and $x_4 = 0$, the fitted equation is a linear function of x_1 with an intercept of -6.398.
- III. For observations classified by $x_2 = 1, x_3 = 0$, and $x_4 = 1$, the fitted equation is a linear function of x_1 with a slope of 0.811.
- IV. Regardless of categorical variable A, the observations classified by $x_4 = 1$ have the same predicted response as those with $x_4 = 0$ when x_1 is 16.763.

Solution

I is false. Based on the number of dummy variables, categorical variable A has three classes, while categorical variable B has two classes. Each class from A could be paired with each class from B, leading to $3 \times 2 = 6$ classes in total.

II is true.

$$\begin{aligned}\hat{y} &= -4.812 + 0.811x_1 + 2.049(0) - 1.586(1) - 3.143(0) + 0.1875x_1(0) \\ &= -6.398 + 0.811x_1\end{aligned}$$

III is false.

$$\begin{aligned}\hat{y} &= -4.812 + 0.811x_1 + 2.049(1) - 1.586(0) - 3.143(1) + 0.1875x_1(1) \\ &= -5.906 + 0.9985x_1\end{aligned}$$

IV is true. The objective is to verify the value of x_1 when the line given by $x_4 = 1$ intersects the line given by $x_4 = 0$. Note that the categorical variable A will only affect the intercept. Changing the intercept of two lines by the same amount does not alter the

horizontal coordinate where they intersect. To simplify, we may assume $x_2 = 0$ and $x_3 = 0$. Set the two equations equal to each other, then solve for x_1 .

$$-4.812 + 0.811x_1 - 3.143(1) + 0.1875x_1(1) = -4.812 + 0.811x_1 - 3.143(0) + 0.1875x_1(0)$$

$$\Rightarrow -3.143 + 0.1875x_1 = 0$$

$$\Rightarrow x_1 = 16.76267$$

Therefore, **only II and IV are true.**



3.3.5 Estimators

We now consider statistical inference in multiple linear regression. Let

- $\hat{\beta}_j$ be the ordinary least squares **estimator** of β_j , and
- $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$ be the ordinary least squares estimator of $E[Y] = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$.

As expected, $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$, and \hat{Y} are normally distributed and unbiased. To perform inference based on a t -distribution, we use their estimated standard errors: $se(\hat{\beta}_0), se(\hat{\beta}_1), \dots, se(\hat{\beta}_p)$, and $se(\hat{y})$. Recall that these are **estimated** standard errors since σ^2 in the true standard error formulas is replaced with MSE.

Just as closed form expressions for the $\hat{\beta}_j$'s were avoided in Section 3.3.2, formulas for the se 's will not be covered here. This means the estimated standard errors will have to be supplied as regression output or determined from a matrix.

For the vector of estimators

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

its variance-covariance matrix is

$$\text{Var}[\hat{\beta}] = \begin{bmatrix} \text{Var}[\hat{\beta}_0] & \text{Cov}[\hat{\beta}_0, \hat{\beta}_1] & \dots & \text{Cov}[\hat{\beta}_0, \hat{\beta}_p] \\ \text{Cov}[\hat{\beta}_0, \hat{\beta}_1] & \text{Var}[\hat{\beta}_1] & \dots & \text{Cov}[\hat{\beta}_1, \hat{\beta}_p] \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}[\hat{\beta}_0, \hat{\beta}_p] & \text{Cov}[\hat{\beta}_1, \hat{\beta}_p] & \dots & \text{Var}[\hat{\beta}_p] \end{bmatrix} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}$$

Therefore, $\text{MSE}(\mathbf{X}^T \mathbf{X})^{-1}$ is a $(p + 1) \times (p + 1)$ matrix whose diagonal entries are $se(\hat{\beta}_0)^2$, $se(\hat{\beta}_1)^2, \dots, se(\hat{\beta}_p)^2$.

Coach's Remarks

It is possible to find $se(\hat{y})$ from a matrix. However, we do not expect this to be tested on the exam. You may skip this Coach's Remarks if you wish.

Let \mathbf{x} represent a vector of feature inputs, i.e. $\mathbf{x}^T = [1 \quad x_1 \quad \cdots \quad x_p]$. Thus, $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p = \mathbf{x}^T \hat{\beta}$. This leads to

$$\begin{aligned}\text{Var}[\hat{Y}] &= \text{Var}[\mathbf{x}^T \hat{\beta}] \\ &= \mathbf{x}^T (\text{Var}[\hat{\beta}]) \mathbf{x} \\ &= \sigma^2 \cdot \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}\end{aligned}$$

$$\Rightarrow se(\hat{y}) = \sqrt{\text{MSE} \cdot \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}$$

For the regression of Commute on Departure, Temp, Precip Chance, and Police from the Commuting Chris scenario, here is how to compute the estimated standard errors beginning with the design matrix.

Observation	Departure	Temp	Precip Chance	Police
1	9.250	31.5	48.0	1
2	13.183	32.5	40.7	1
:	:	:	:	:
100	12.000	20.0	98.4	4

$$\mathbf{X} = \begin{bmatrix} 1 & 9.25 & 31.5 & 48 & 1 \\ 1 & 13.183 & 32.5 & 40.7 & 1 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 12 & 20 & 98.4 & 4 \end{bmatrix}$$

Next, compute $(\mathbf{X}^T \mathbf{X})^{-1}$. Notice this is a 5×5 matrix since this model has five regression coefficients – one intercept and one for each predictor.

$$(\mathbf{X}^T \mathbf{X})^{-1} = \begin{bmatrix} 0.647534 & -0.020407 & -0.006330 & -0.002652 & -0.015055 \\ -0.020407 & 0.001706 & 0.000045 & -0.000011 & -0.000181 \\ -0.006330 & 0.000045 & 0.000090 & 0.000037 & 0.000226 \\ -0.002652 & -0.000011 & 0.000037 & 0.000028 & -0.000042 \\ -0.015055 & -0.000181 & 0.000226 & -0.000042 & 0.010028 \end{bmatrix}$$

Recall from Section 3.3.2 that the (rounded) MSE was 10.7812. Therefore, the estimated variance-covariance matrix is

$$\text{MSE}(\mathbf{X}^T \mathbf{X})^{-1} = \begin{bmatrix} 6.981187 & -0.220013 & -0.068245 & -0.028592 & -0.162311 \\ -0.220013 & 0.018387 & 0.000481 & -0.000120 & -0.001949 \\ -0.068245 & 0.000481 & 0.000972 & 0.000396 & 0.002434 \\ -0.028592 & -0.000120 & 0.000396 & 0.000305 & -0.000456 \\ -0.162311 & -0.001949 & 0.002434 & -0.000456 & 0.108117 \end{bmatrix}$$

The $(j+1)^{\text{st}}$ diagonal entry of $\text{MSE}(\mathbf{X}^T \mathbf{X})^{-1}$ is $se(\hat{\beta}_j)^2$. The square root of the (unrounded) diagonal entries agrees with the R output.

$$se(\hat{\beta}_0) = \sqrt{6.981187} = 2.64219$$

$$se(\hat{\beta}_1) = \sqrt{0.018387} = 0.13560$$

$$se(\hat{\beta}_2) = \sqrt{0.000972} = 0.03117$$

$$se(\hat{\beta}_3) = \sqrt{0.000305} = 0.01748$$

$$se(\hat{\beta}_4) = \sqrt{0.108117} = 0.32881$$

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	25.98958	2.64219	9.836	3.7e-16 ***
Departure	-0.63290	0.13560	-4.667	1.0e-05 ***
Temp	0.05584	0.03117	1.791	0.0764 .
Precip.Chance	0.04208	0.01748	2.408	0.0180 *
Police	4.00287	0.32881	12.174	< 2e-16 ***

Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *
	0.1 .	0.1 ' '	1	

Residual standard error: 3.283 on 95 degrees of freedom
 Multiple R-squared: 0.6782, Adjusted R-squared: 0.6646
 F-statistic: 50.05 on 4 and 95 DF, p-value: < 2.2e-16

Coach's Remarks

Let's elaborate on the previous Coach's Remarks with an example:

Find the value of $se(\hat{y}_2)$, i.e. the estimated standard error for the mean response with the inputs of the 2nd training observation. Note that

$$\mathbf{x}_2^T = [1 \quad 13.183 \quad 32.5 \quad 40.7 \quad 1]$$

which leads to

$$se(\hat{y}_2) = \sqrt{\text{MSE} \cdot \mathbf{x}_2^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_2} = 0.57412$$

If you intend to compute this by hand based on the given components, note that the result will differ slightly due to rounding error.

Note that the comments on estimating standard errors through bootstrapping in Section 3.2.5 remain unchanged in multiple linear regression. We can artificially simulate many estimates of a regression coefficient by creating many bootstrap samples and running the same multiple linear regression on each one.

3.3.6 t Tests

All concepts mentioned in Section 3.2.6 carry over to multiple linear regression. The key difference is that the test statistic of

$$t. s. = \frac{\hat{\beta}_j - h}{se(\hat{\beta}_j)} \quad (3.3.6.1)$$

now comes from a t -distribution with $n - p - 1$ degrees of freedom. This is consistent with how the MSE formula in multiple linear regression has a denominator of $n - p - 1$.

Therefore, for an α level test, we reject H_0 if $p\text{-value} \leq \alpha$, or equivalently

Test Type	Critical Region
Left-tailed	$t. s. \leq -t_{2\alpha, n-p-1}$
Two-tailed	$ t. s \geq t_{\alpha, n-p-1}$
Right-tailed	$t. s. \geq t_{2\alpha, n-p-1}$

In multiple linear regression, the main challenge with t tests is interpreting the results correctly. It helps to remember that t tests have hypotheses that only involve **one** regression coefficient at a time.

Let's practice performing t tests by revisiting the three regressions on the Commuting Chris data presented in Section 3.3.4.

QUADRATIC MODEL

The model equation is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$$

where

- x_1 is the time of departure, and
- x_2 is the number of police vehicles along the commute route.

The R output is

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 43.04271   7.20325  5.975 3.89e-08 ***
Departure   -3.14172   1.37165 -2.290  0.0242 *  
Departure^2  0.11579   0.06264  1.849  0.0676 .  
Police      4.03762   0.29798 13.550 < 2e-16 *** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.306 on 96 degrees of freedom
Multiple R-squared:  0.6703, Adjusted R-squared:  0.66 
F-statistic: 65.05 on 3 and 96 DF,  p-value: < 2.2e-16
```

The **Coefficients** table shows the results of testing the hypotheses

$$H_0 : \beta_j = 0 \quad H_1 : \beta_j \neq 0$$

for $j = 0, 1, 2, 3$. Thus, the test statistic for testing β_1 is given as -2.29 under **t value**. We may calculate the test statistic as

$$\begin{aligned} t.s. &= \frac{\hat{\beta}_1 - h}{se(\hat{\beta}_1)} \\ &= \frac{-3.14172 - 0}{1.37165} \\ &= -2.290 \end{aligned}$$

We are given $df = 96$, which agrees with the fact that 100 observations were used to calculate 4 $\hat{\beta}_j$'s in this model.

Since this test for β_1 is two-tailed, its p -value is the probability that a t random variable with 96 degrees of freedom is less than -2.29 or greater than 2.29. The output supplies this probability as 2.42% under **Pr(>|t|)**. For a significance level of 5%, we reject H_0 in favor of H_1 since 2.42% is less than 5%. In conclusion, x_1 is significant to the model because its coefficient is plausibly non-zero; x_1 seems to explain the response well.

On the other hand, notice that the p -value for testing β_2 is 6.76%. At a 5% significance level, we would fail to reject H_0 . This indicates that x_1^2 is not significant to the model because its coefficient is plausibly zero; x_1^2 does not seem to explain the response well. In other words, the model is likely to improve if the quadratic term of Departure is dropped and thus prevented from influencing the $\hat{\beta}$'s of the other predictors.

Coach's Remarks

In testing the slope parameter under simple linear regression (Section 3.2.6), we thought of the t test as a comparison between the simple linear regression model and the null model. The idea of comparing models applies in multiple linear regression as well.

Using the quadratic model to illustrate, when testing

$$H_0 : \beta_1 = 0 \quad H_1 : \beta_1 \neq 0$$

we are effectively comparing the models:

- A. $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$
- B. $Y = \beta_0 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$

In rejecting this H_0 , we conclude that this Model A is preferred over this Model B.

However, when testing

$$H_0 : \beta_2 = 0 \quad H_1 : \beta_2 \neq 0$$

we are effectively comparing the models:

- A. $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \varepsilon$
- B. $Y = \beta_0 + \beta_1 x_1 + \beta_3 x_2 + \varepsilon$

In failing to reject this H_0 , we conclude that this Model B is preferred over this Model A.

However, there are better approaches to extensive model comparison than aggregating all of the t test conclusions. So, it is best to consider each t test separately, where each performs a unique comparison between two models revolving around one β_j .

DUMMY VARIABLE MODEL

The model equation is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

where

- x_1 is the chance of precipitation,
- x_2 is the dummy variable associated with Spring,
- x_3 is the dummy variable associated with Summer, and
- x_4 is the dummy variable associated with Winter.

The R output is

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.19931   1.45313 16.653 <2e-16 ***
Precip.Chance 0.03831   0.02297  1.668  0.0986 .  
SeasonSpring -1.13806   1.48961 -0.764  0.4468  
SeasonSummer  1.12868   1.70719  0.661  0.5101  
SeasonWinter  3.37147   1.56541  2.154  0.0338 * 
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.208 on 95 degrees of freedom
Multiple R-squared:  0.1905, Adjusted R-squared:  0.1564 
F-statistic: 5.588 on 4 and 95 DF,  p-value: 0.0004396
```

At the 5% significance level, x_1 is not significant to the model, having a p -value of 9.86%. However, when Precip Chance was the only predictor in Section 3.2.6, its t test suggested that Precip Chance was helpful to the model. This is not a contradiction. More precisely, the t test in Section 3.2.6 is suggesting that Precip Chance as the only predictor is better than the null model. Furthermore, we see from the t test here that adding Season to the model lowered the importance of Precip Chance in predicting Commute.

The p -values associated with x_2 and x_3 are very large, whereas the p -value associated with x_4 is rather small. Keep in mind what the associated regression coefficients represent:

- β_2 is the change in the intercept in the spring **relative to fall**.
- β_3 is the change in the intercept in the summer **relative to fall**.
- β_4 is the change in the intercept in the winter **relative to fall**.

Therefore,

- failing to reject $\beta_2 = 0$ means that spring and fall plausibly have the same intercept at the 0.05 level.

- failing to reject $\beta_3 = 0$ means that summer and fall plausibly have the same intercept at the 0.05 level.
- rejecting $\beta_4 = 0$ means that winter and fall plausibly have different intercepts at the 0.05 level.

Realize that these t tests are unable to compare the intercepts for seasons that do not involve fall, as well as determine whether Season as a whole does well at explaining the response.

In general, a dummy variable that is "not significant to the model" should not simply be dropped. It leads to altering a categorical variable with w classes to have $w - 1$ classes instead, which may not be an intended result. Typically, further investigation is needed to discover the best way to improve the model.

INTERACTION MODEL

The ***hierarchical principle*** states that a significant interaction term implies that its individual terms should also be in the model, regardless of the t tests associated with the individual terms. This is because it is not important what the coefficients of the individual terms are when their interaction explains the response well. Moreover, the meaning of an interaction could change if any of its individual terms are removed from the model.

The model equation is

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \varepsilon$$

where

- x_1 is the chance of precipitation,
- x_2 is the dummy variable associated with Spring,
- x_3 is the dummy variable associated with Summer, and
- x_4 is the dummy variable associated with Winter.

The R output is

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 24.76495  1.55054 15.972 <2e-16 ***
Precip.Chance 0.02549  0.02641  0.965  0.3370  
SeasonSpring   1.72271  3.35543  0.513  0.6089  
SeasonSummer    2.66374  2.62485  1.015  0.3129  
SeasonWinter   -3.28696  3.81271 -0.862  0.3909  
Precip.Chance:SeasonSpring -0.12590  0.11718 -1.074  0.2855 
Precip.Chance:SeasonSummer -0.15970  0.15015 -1.064  0.2903 
Precip.Chance:SeasonWinter  0.10350  0.05615  1.843  0.0685 .  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 

Residual standard error: 5.116 on 92 degrees of freedom
Multiple R-squared:  0.2433, Adjusted R-squared:  0.1857 
F-statistic: 4.226 on 7 and 92 DF,  p-value: 0.0004407
```

Here, the regression coefficients associated with the interaction terms are interpreted as:

- β_5 is the change in the Precip Chance slope in the spring **relative to fall**.
- β_6 is the change in the Precip Chance slope in the summer **relative to fall**.
- β_7 is the change in the Precip Chance slope in the winter **relative to fall**.

Since all three interaction terms have p -values that are greater than 5%, we fail to reject the null hypotheses at the 0.05 level. The Precip Chance slopes in the spring, summer, and winter are each plausibly the same as the slope in the fall.

Due to interacting with a categorical variable, realize that these t tests are unable to compare the slopes for seasons that do not involve fall, as well as determine whether the overall interaction between Precip Chance and Season is meaningful to the model.

In addition, dropping an interaction term that contains a dummy variable is not recommended for reasons similar to those given regarding dropping a dummy variable.

Example 3.3.6.1

Julianne studies the price of gas at 18 gas stations. For each station, she has a record of its daily average gas price, its daily average demand (a function of gallons sold), and its daily average maintenance cost. She obtains the following result from running a multiple linear regression:

	Coefficient	Standard Error
Intercept	5.4511	2.6649
Demand	-1.0308	0.5073
Maintenance Cost	0.2877	0.0369

With a hypothesis test, Julianne wants to show that each additional dollar of daily average maintenance cost increases the expected price of gas by less than 0.35 for a fixed volume of demand.

Determine the test result at the 5% significance level.

Solution

Let β_2 represent the change in the expected price of gas for each additional dollar of daily average maintenance cost while holding demand constant. Since "less than 0.35" is what Julianne wants to demonstrate, this is a left-tailed test.

First, calculate the test statistic as

$$t. s. = \frac{0.2877 - 0.35}{0.0369} = -1.6883$$

Next, find the critical value from the exam table. For $\alpha = 0.05$ and $df = 18 - 3 = 15$,

$$-t_{2(0.05), 15} = -t_{0.1, 15} = -1.753$$

Since $-1.6883 > -1.753$, we conclude that **it is implausible for β_2 to be less than 0.35 at the 0.05 level.**



Example 3.3.6.2

A researcher from Agonite Airlines wants to learn about customer luggage weight per bag (in pounds) from 60 customers. Running a multiple linear regression produces:

	Coefficients	Standard Error
Intercept	41.9603	3.463184
Age of customer (years)	-0.69064	0.143536

	Coefficients	Standard Error
Squared age of customer (years ²)	0.00612	0.001431
International flight (1 = Yes, 0 = No)	16.86309	2.097807
Number of connecting flights	-2.75903	0.942170

Determine which statements are true.

- I. The "age of customer" variable should be kept in the model, using a 1% significance level.
- II. The "squared age of customer" variable should be dropped from the model, using a 1% significance level.
- III. The "international flight" variable should be kept in the model, using a 1% significance level.
- IV. Using a t test, we can determine whether the current model is better than the model with only the customer's age and its squared term as predictors.

Solution

A two-tailed t test can determine whether a predictor should be kept or dropped from a model. The null hypothesis is that the corresponding regression coefficient equals 0. With $\alpha = 0.01$ and $df = 60 - 5 = 55$, the appropriate critical value for every t test is $t_{0.01, 55} = 2.668$.

If the absolute value of a test statistic equals or exceeds 2.668, then the predictor should be kept in the model at the 0.01 level. Otherwise, the predictor should be dropped.

I is true. The test statistic is

$$t. s. = \frac{-0.69064 - 0}{0.143536} = -4.8116$$

Since $|-4.8116| > 2.668$, the "age of customer" variable should be kept.

II is false. The test statistic is

$$t. s. = \frac{0.00612 - 0}{0.001431} = 4.2767$$

Since $|4.2767| > 2.668$, the "squared age of customer" variable should be kept.

III is true. The test statistic is

$$t. s. = \frac{16.86309 - 0}{2.097807} = 8.0384$$

Since $|8.0384| > 2.668$, the "international flight" variable should be kept.

IV is false. A t test is unable to comment on whether a model should keep or drop more than one predictor at a time since each t test can only handle one regression coefficient.

Therefore, **only I and III are true.**



3.3.7 F Tests

Since t tests examine only one regression coefficient at a time, they cover a limited scope of inquiries about a model. Other kinds of inquiries can be tied to several regression coefficients at once. Considering the dummy variable model and the interaction model in the previous subsection, we may wish to know:

1. Is there a significant overall interaction between Precip Chance and Season in predicting Commute?
2. As a whole, does Season contribute significantly to the dummy variable model?

To answer these questions, we first need to be familiar with ANOVA tables and F tests.

Analysis of Variance (ANOVA)

An **ANOVA table** organizes the partitioning of variability in the response. Recall the three sums of squares introduced in Section 3.2.4:

- The regression sum of squares (SSR) is the amount of variability explained by the linear regression.
- The error sum of squares (SSE) is the amount of variability left unexplained by the linear regression.
- The total sum of squares (SST) is the total amount of variability.

In addition, recall that

$$\text{MSE} = \frac{\text{SSE}}{n - p - 1}, \quad s_y^2 = \frac{\text{SST}}{n - 1}$$

Thus far, we can populate an ANOVA table as follows:

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)
Regression	SSR	?	?
Error	SSE	$n - p - 1$	MSE
Total	SST	$n - 1$	s_y^2

This table operates on the following relationships:

- $\text{SSR} + \text{SSE} = \text{SST}$
- Regression df + Error df = Total df
- $\text{MS} = \frac{\text{SS}}{\text{df}}$

Therefore, the complete ANOVA table looks like:

Source	Sum of Squares (SS)	Degrees of Freedom (df)	Mean Square (MS)
Regression	SSR	p	MSR
Error	SSE	$n - p - 1$	MSE
Total	SST	$n - 1$	s_y^2

where

$$\text{MSR} = \frac{\text{SSR}}{p} \quad (3.3.7.1)$$

MSR is the **average** variability explained by the linear regression per degree of freedom. Keep in mind that s_y^2 is not the sum of MSR and MSE.

F Test

We use an F test to simultaneously examine all of the regression coefficients except the intercept. Specifically, we consider the set of hypotheses

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_p = 0 \quad H_1 : \text{At least one } \beta_j \neq 0 \text{ for } j = 1, \dots, p$$

This test determines whether the proposed multiple linear regression is preferred over the null model. To be precise, the MLR model equation of

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

is compared to the null model equation of

$$Y = \beta_0 + \varepsilon$$

The test statistic is

$$\begin{aligned} t.s. &= \frac{\text{MSR}}{\text{MSE}} \\ &= \frac{\text{SSR}/p}{\text{SSE}/(n-p-1)} \end{aligned} \tag{3.3.7.2}$$

which comes from an F -distribution with $\text{ndf} = p$ and $\text{ddf} = n - p - 1$. In this context, we commonly say that

- p is the number of degrees of freedom associated with the regression.
- $n - p - 1$ is the number of degrees of freedom associated with error.

These F tests are right-tailed tests, so H_0 is rejected at the α level when

$$t.s. \geq F_{\alpha, p, n-p-1}$$

Coach's Remarks

Here is the intuition behind this F test:

A multiple linear regression with at least one meaningful predictor should explain a significant amount of variability, resulting in a large MSR. Thus, H_0 is rejected when MSR is reasonably large, as quantified by its ratio to MSE.

By that reasoning, one could argue that the ratio of SSR to SSE measures a similar value. However, the two sums of squares have different degrees of freedom. Since MSR and MSE are average quantities per degree of freedom, their ratio makes more sense as the test statistic rather than the ratio of SSR to SSE.

Let's revisit the R output with four predictors from the Commuting Chris scenario:

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 25.98958   2.64219   9.836 3.7e-16 ***
Departure   -0.63290   0.13560  -4.667 1.0e-05 ***
Temp        0.05584   0.03117   1.791  0.0764 .
Precip.Chance 0.04208   0.01748   2.408  0.0180 *
Police      4.00287   0.32881  12.174 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.283 on 95 degrees of freedom
Multiple R-squared:  0.6782, Adjusted R-squared:  0.6646
F-statistic: 50.05 on 4 and 95 DF,  p-value: < 2.2e-16
```

It gives the test statistic as 50.05, along with $\text{ndf} = 4$ and $\text{ddf} = 95$. The p -value for this test is smaller than 2.2×10^{-16} . Therefore, we reject H_0 and conclude that at least one of the four predictors is significant to the model.

To calculate the test statistic, we need the MSR and MSE. Recall from Section 3.3.2 that $\text{MSE} = 10.7812$. Given $s_y^2 = 32.1486$, we can arrange the information in an ANOVA table to determine the MSR.

Source	SS	df	MS
Regression	2,158.50	4	539.625
Error	1,024.21	95	10.7812
Total	3,182.71	99	32.1486

As a result,

$$t.s. = \frac{\text{MSR}}{\text{MSE}} = \frac{539.625}{10.7812} = 50.05$$

Simply memorizing the components of an F test can easily lead to mistakes. Instead, we suggest framing the details with the following perspective:

The more flexible multiple linear regression model has p more regression coefficients than the null model. Thus, the MLR model **reduces** or **explains** some of the variability in the response, which is calculated as SSR. But to estimate the coefficients, we must "spend" one degree of freedom for each estimate.

In summary, p additional degrees of freedom are "spent" in order to reduce the variability by SSR. Therefore, the test statistic – with MSR in its numerator – captures whether the amount of variability explained is large enough to justify the degrees of freedom "spent". We may express the MSR in words, i.e.

$$\text{MSR} = \frac{\text{reduction in variability}}{\text{additional df spent}}$$

Example 3.3.7.1

A financial analyst believes that the return on a certain stock portfolio can be explained using three macroeconomic features. Running a multiple linear regression produces the following ANOVA table:

Source	SS	df	MS
Regression	?	3	0.00331
Error	0.0269	?	?
Total	?	29	?

Determine which statement is true.

- I. The estimate of the irreducible error is 0.00103.
- II. The adjusted R^2 is 0.270.
- III. The conclusion of the F test is that all three macroeconomic features are significant to the model at the 0.05 level.

Solution

From the ANOVA table, we can solve for:

- $\text{SSR} = 3 \cdot 0.00331 = 0.00993$
- $\text{SST} = 0.00993 + 0.0269 = 0.03683$
- Error df = $29 - 3 = 26$
- $\text{MSE} = \frac{0.0269}{26} = 0.0010346$
- $s_y^2 = \frac{0.03683}{29} = 0.00127$

I is true because the irreducible error is σ^2 , whose estimate is the MSE.

II is false because it is R^2 , rather than $R_{\text{adj.}}^2$, that equals 0.270.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = \frac{0.00993}{0.03683} = 0.270$$

$$R_{\text{adj.}}^2 = 1 - \frac{\text{MSE}}{s_y^2} = 1 - \frac{0.0010346}{0.00127} = 0.185$$

III is false. The statement suggests that **all** three regression coefficients associated with the features are plausibly non-zero. However, H_1 of this F test is that **at least one** of the three regression coefficients is non-zero. Hence, the correct conclusion is that at least one of the three macroeconomic features is significant to the model at the 0.05 level. This is because the test statistic of

$$\frac{\text{MSR}}{\text{MSE}} = \frac{0.00331}{0.0010346} = 3.199$$

is greater than the critical value of $F_{0.05, 3, 26} = 2.975$.

Therefore, **only I is true.**



Remember that a simple linear regression can be perceived as a multiple linear regression with $p = 1$. Consequently, the F test for a simple linear regression has the same H_0 and H_1 as the t test for its **slope parameter**. Furthermore,

- the squared test statistic for the t test equals the test statistic for the F test, and
- the p -values of both tests are the same.

You are encouraged to verify these results for yourself by revisiting the R outputs in Section 3.2.

Partial F Test

To answer the questions posed at the start of this subsection, we need to conduct partial F tests, rather than ordinary F tests.

A partial F test compares two nested multiple linear regression models. The model with **more** regression coefficients is referred to as the **full model**; the model with **fewer** regression coefficients is referred to as the **reduced model**. Let's go through our two questions to see the concepts in action.

QUESTION 1

To test whether there is a significant interaction between Precip Chance and Season, we compare the full model with the equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1 x_4 + \varepsilon$$

to the reduced model with the equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

In other words, we want to test the hypotheses

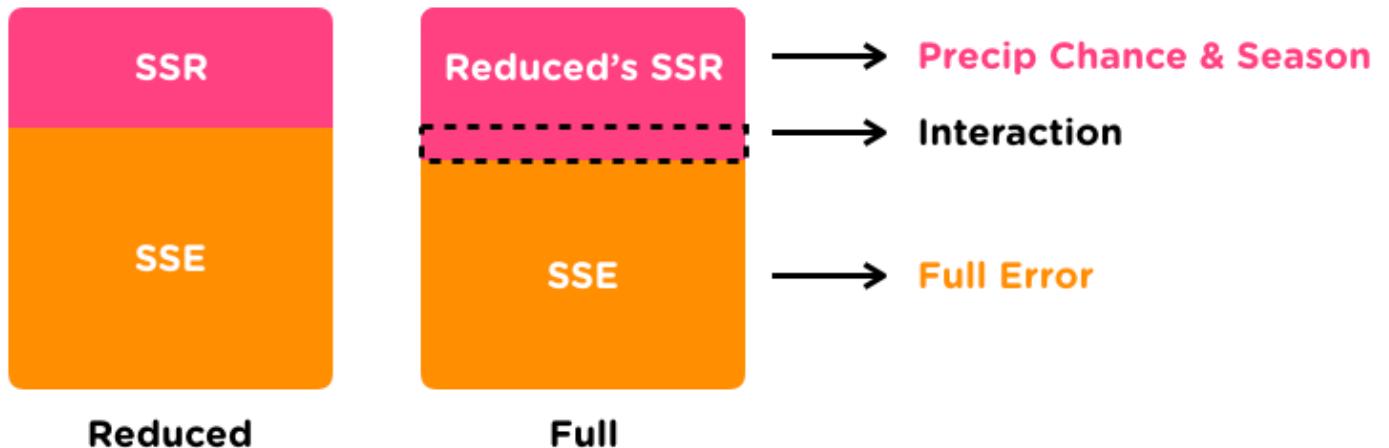
$$H_0 : \beta_5 = \beta_6 = \beta_7 = 0 \quad H_1 : \text{At least one } \beta_j \neq 0 \text{ for } j = 5, 6, 7$$

3 additional degrees of freedom are "spent" to estimate β_5 , β_6 , and β_7 for the full model. Hence, this partial F test examines whether "spending" these 3 degrees of freedom reduces a significant amount of variability or not. The next step is to calculate how much variability was dropped or explained going from the reduced model to the full model.

Consider the following diagrams depicting the variability as partitioned by the full and reduced models:



Notice the reduction in the error portion going from the reduced model to the full model. This is the amount of variability explained by the interaction terms. A different but equivalent perspective is to view the regression portion of the full model being further partitioned into two: the reduced model's contribution and the interaction terms' contribution.



Consequently, we can tabulate the partitioning of the variability as follows:

Source	SS	df
Precip Chance and Season	606.213	4
Interaction	168.146	3
Full Error	2,408.351	92
Total	3,182.71	99

To summarize, this table says that 3 degrees of freedom were "spent" in estimating the interaction coefficients to lower the variability by 168.146 units. There are two ways to arrive at 168.146. The first is to note that the reduced model (without interactions) has 2,576.497 units of unexplained

variability, which is brought down to 2,408.351 units by the full model (with interactions). The second is to note that the full model explains 774.359 units of variability, from which 606.213 units have already been explained by the reduced model.

The test statistic of the partial F test is the ratio of:

- the mean square for the Interaction source, and
- the mean square for the Full Error source.

$$t.s. = \frac{168.146/3}{2,408.351/92} = 2.141$$

This comes from an F -distribution with $ndf = 3$ and $ddf = 92$, and its 95th percentile is 2.704. Since $2.141 < 2.704$, we fail to reject H_0 at the 0.05 level. We thus conclude that there is no significant interaction between Precip Chance and Season at 5% significance; the 168.146 units of variability explained by the interaction terms was not large enough. In this case, the reduced model is preferred.

QUESTION 2

To test whether Season contributes significantly to the dummy variable model, we compare the full model with the equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

to the reduced model with the equation

$$Y = \beta_0 + \beta_1 x_1 + \varepsilon$$

In other words, we want to test the hypotheses

$$H_0 : \beta_2 = \beta_3 = \beta_4 = 0 \quad H_1 : \text{At least one } \beta_j \neq 0 \text{ for } j = 2, 3, 4$$

Clearly, the additional degrees of freedom "spent" is 3. Find the reduction in variability using a table:

Source	SS	df
Precip Chance	372.038	1

Source	SS	df
Season	234.175	3
Full Error	2,576.497	95
Total	3,182.71	99

Thus, the 3 degrees of freedom "spent" in estimating the dummy variable coefficients lowered the variability by 234.175 units, going from the reduced model to the full model.

The test statistic is

$$t.s. = \frac{234.175/3}{2,576.497/95} = 2.878$$

2.700 is the 95th percentile of an *F*-distribution with *ndf* = 3 and *ddf* = 95. Since $2.878 > 2.700$, we reject H_0 at the 0.05 level. We thus conclude that Season is significant to the dummy variable model at 5% significance; the 234.175 units of variability explained by the dummy variables is considered large enough. In this case, the full model is preferred.

GENERALIZATION

In general, we can organize the components of a partial *F* test in a table as follows:

Source	SS	df
Reduced Regression	SSR_r	p_r
Difference	$\text{SSE}_r - \text{SSE}_f$ or $\text{SSR}_f - \text{SSR}_r$	$p_f - p_r$
Full Error	SSE_f	$n - p_f - 1$
Total	SST	$n - 1$

where subscripts *f* and *r* stand for full and reduced, respectively.

Hence, the test statistic has the formula

$$t.s. = \frac{(\text{SSE}_r - \text{SSE}_f) / (p_f - p_r)}{\text{SSE}_f / (n - p_f - 1)} \quad (3.3.7.3)$$

Rather than memorizing the numerator of the test statistic in math notation, we can remember it using words:

$$\frac{\text{SSE}_r - \text{SSE}_f}{p_f - p_r} = \frac{\text{reduction in variability}}{\text{additional df spent}}$$

In addition, when a partial F test examines only one regression coefficient (i.e. $p_f - p_r = 1$), it is connected to the corresponding multiple linear regression t test such that

- the squared test statistic for the t test equals the test statistic for the F test, and
- the p -values of both tests are the same.

This is essentially the same connection between a simple linear regression's ordinary F test and the t test for its slope parameter.

Coach's Remarks

Notice that an ordinary F test is equivalent to a partial F test with

- the null model as the reduced model,
- $\text{SSE}_r = \text{SST}$,
- $\text{SSR}_r = 0$,
- $p_r = 0$,
- $\text{SSE}_f = \text{SSE}$, and
- $p_f = p$.

Example 3.3.7.2

A regression model has the equation

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

Using ordinary least squares and 25 observations, you are given:

- The unbiased sample variance of the response is 3,589.
- For the given model, $R^2 = 0.7344$.
- Without x_1 and x_3 in the model, $R^2 = 0.5439$.
- Without x_2 and x_4 in the model, $R^2 = 0.6127$.

The null hypothesis $H_0 : \beta_2 = \beta_4 = 0$ is tested at the 5% significance level.

Determine the correct conclusion for this test.

- A. The test statistic is 4.582 and H_0 should not be rejected.
- B. The test statistic is 4.582 and H_0 should be rejected.
- C. The test statistic is 7.172 and H_0 should not be rejected.
- D. The test statistic is 7.172 and H_0 should be rejected.
- E. None of the above.

Solution

Based on H_0 , this is a partial F test that compares the model with all four predictors against the model without x_2 and x_4 . Thus, the third bullet point can be ignored.

Given $n = 25$ and $s_y^2 = 3,589$, we find $\text{SST} = (25 - 1)3,589 = 86,136$.

Let

- SSR_f be the explained variability for the model with all four predictors, and
- SSR_r be the explained variability for the model without x_2 and x_4 .

Solve for the two SSR's using their respective coefficients of determination and SST.

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

$$\Rightarrow \text{SSR}_f = 0.7344 \cdot 86,136 = 63,258.2784$$

$$\Rightarrow \text{SSR}_r = 0.6127 \cdot 86,136 = 52,775.5272$$

The 2 additional degrees of freedom "spent" contribute to the explained variability by $63,258.2784 - 52,775.5272 = 10,482.7512$ units. This leads to the test statistic of

$$\frac{10,482.7512/2}{(86,136 - 63,258.2784)/(25 - 4 - 1)} = 4.582$$

Since the test statistic is greater than the critical value of $F_{0.05, 2, 20} = 3.493$, we should reject H_0 at the 0.05 level.

Therefore, the answer is **(B)**.



3.3.8 Confidence Intervals

Aside from the content in Section 3.2.7 with sensible adjustments, there are no new concepts for constructing confidence intervals in multiple linear regression. As expected, t -distributions are still used, resulting in the familiar general expression of

$$\text{estimate} \pm (t \text{ percentile}) (\text{standard error})$$

The $100k\%$ confidence interval for β_j has the expression

$$\hat{\beta}_j \pm t_{1-k, n-p-1} \cdot se(\hat{\beta}_j) \quad (3.3.8.1)$$

The $100k\%$ confidence interval for $E[Y]$ has the expression

$$\hat{y} \pm t_{1-k, n-p-1} \cdot se(\hat{y}) \quad (3.3.8.2)$$

Example 3.3.8.1

Julianne studies the price of gas at 18 gas stations. For each station, she has a record of its daily average gas price, its daily average demand (a function of gallons sold), and its daily average maintenance cost. She obtains the following result from running a multiple linear regression:

	Coefficients	Standard Error
Intercept	5.4511	2.6649
Demand	-1.0308	0.5073
Maintenance Cost	0.2877	0.0369

Calculate:

- the 99% confidence interval for β_1 , the regression coefficient of Demand.
- the estimated standard error for the mean response when Demand is 3.333, Maintenance Cost is 2.25, and the lower bound of the 95% confidence interval for the mean response is 0.5119.

Solution to (1)

With 99% confidence and $18 - 3 = 15$ degrees of freedom, we need

$$t_{1-0.99, 15} = t_{0.01, 15} = 2.947$$

as given by the exam table.

The 99% confidence interval for β_1 is

$$\hat{\beta}_1 \pm t_{0.01, 15} \cdot se(\hat{\beta}_1)$$

$$\Rightarrow -1.0308 \pm 2.947 \cdot 0.5073$$

$$\Rightarrow (-2.526, 0.464)$$



Solution to (2)

With 95% confidence and $18 - 3 = 15$ degrees of freedom, we need

$$t_{1-0.95, 15} = t_{0.05, 15} = 2.131$$

as given by the exam table.

Next, calculate the predicted response at the specified values of Demand and Maintenance Cost.

$$\begin{aligned}\hat{y} &= 5.4511 - 1.0308(3.333) + 0.2877(2.25) \\ &= 2.663\end{aligned}$$

Finally, solve for the estimated standard error for the mean response using the formula for the confidence interval's lower bound.

$$\hat{y} - t_{0.05, 15} \cdot se(\hat{y}) = 0.5119$$

$$\begin{aligned}\Rightarrow se(\hat{y}) &= \frac{\hat{y} - 0.5119}{t_{0.05, 15}} \\ &= \frac{2.663 - 0.5119}{2.131} \\ &= \mathbf{1.009}\end{aligned}$$



Example 3.3.8.2

For a multiple linear regression using 6 predictors on 20 observations, the 90% confidence interval for one of the regression coefficients, γ , is (-4.367, -1.064).

Determine which statements are true for testing $H_0 : \gamma = 0$ against $H_1 : \gamma \neq 0$.

- I. Reject H_0 at the 0.100 level.
- II. Reject H_0 at the 0.050 level.
- III. Reject H_0 at the 0.005 level.

Solution

With 90% confidence and $20 - 6 - 1 = 13$ degrees of freedom, the confidence interval uses

$$t_{1-0.9, 13} = t_{0.1, 13} = 1.771$$

as given by the exam table.

Let $\hat{\gamma}$ denote the estimate of γ . The midpoint of the confidence interval is $\hat{\gamma}$, which is

$$\frac{-4.367 + (-1.064)}{2} = -2.7155$$

The estimated standard error for γ can be solved using the upper or lower bound of the given 90% confidence interval. We demonstrate using the lower bound.

$$\begin{aligned} se &= \frac{\hat{\gamma} - (-4.367)}{t_{0.1, 13}} \\ &= \frac{-2.7155 - (-4.367)}{1.771} \\ &= 0.9325 \end{aligned}$$

In turn, the test statistic for this two-tailed t test is

$$t.s. = \frac{-2.7155 - 0}{0.9325} = -2.912$$

I is true because $|-2.912|$ is greater than the critical value at the 0.100 level, which is $t_{0.1, 13} = 1.771$.

II is true because $|-2.912|$ is greater than the critical value at the 0.050 level, which is $t_{0.05, 13} = 2.160$.

III is false. The critical value at the 0.005 level is $t_{0.005, 13}$, but it is not listed in the t table. However, note that $t_{0.01, 13} = 3.012$. The percentile 3.012 corresponds to a significance level of $\alpha = 0.01$. Therefore, $|-2.912| < 3.012$ means that we fail to reject H_0 at the 0.01 level. As a result, we also fail to reject H_0 at the 0.005 level because $t_{0.005, 13}$ must be even greater than 3.012.

Therefore, **only I and II are true.**

Alternative Solution

Recall that if the hypothesized value of the hypothesis test is within the $100(1 - \alpha)\%$ confidence interval, H_0 will fail to be rejected at the α level. Otherwise, H_0 will be rejected at the α level.

Thus, we are interested in the intervals with confidence levels of

$$\begin{aligned}\alpha = 0.1 &\Rightarrow k = 1 - 0.1 = 0.9 \\ \alpha = 0.05 &\Rightarrow k = 1 - 0.05 = 0.95 \\ \alpha = 0.005 &\Rightarrow k = 1 - 0.005 = 0.995\end{aligned}$$

I is true because 0 is not within the 90% confidence interval.

II is true. As increasing the confidence level will widen the interval, we only need to check whether the upper bound becomes positive; remaining negative means that 0 is still outside the interval, while changing to positive means that 0 is within the interval.

Recall that the midpoint of the given interval is -2.7155. Thus, the half-width of the 90% confidence interval can be computed as

$$\begin{aligned}t_{0.1, 13} \cdot se &= -1.064 - \hat{\gamma} \\ &= -1.064 - (-2.7155) \\ &= 1.6515\end{aligned}$$

We could solve for se , but a faster way to find the upper bound of the 95% confidence interval is to realize that we need the interval's half-width, which is

$$\begin{aligned}
 t_{0.05, 13} \cdot se &= t_{0.05, 13} \cdot \frac{t_{0.1, 13}}{t_{0.1, 13}} \cdot se \\
 &= \frac{t_{0.05, 13}}{t_{0.1, 13}} \cdot (t_{0.1, 13} \cdot se) \\
 &= \frac{2.160}{1.771} \cdot 1.6515 \\
 &= 2.01425
 \end{aligned}$$

This means the upper bound of the 95% confidence interval is

$$-2.7155 + 2.01425 \quad (\text{negative})$$

In conclusion, we reject H_0 because 0 is not within the 95% confidence interval.

III is false. We follow the same procedure, keeping in mind that we can only construct the 99% confidence interval with $t_{0.01, 13} = 3.012$ from the exam table. The half-width of the 99% confidence interval is

$$\begin{aligned}
 t_{0.01, 13} \cdot se &= \frac{t_{0.01, 13}}{t_{0.1, 13}} \cdot (t_{0.1, 13} \cdot se) \\
 &= \frac{3.012}{1.771} \cdot 1.6515 \\
 &= 2.8088
 \end{aligned}$$

This means the upper bound of the 99% confidence interval is

$$-2.7155 + 2.8088 \quad (\text{positive})$$

Since the 99% confidence interval contains 0, then the 99.5% confidence interval must also contain 0. Hence, we fail to reject H_0 .

Therefore, **only I and II are true.**



Coach's Remarks

We do not expect the computation of prediction intervals under multiple linear regression to be tested on the exam. On the other hand, the conceptual aspects of prediction intervals, as covered in Section 3.2.7, should still be fair game.

Nevertheless, it is not hard to infer that the prediction interval expression is

$$\hat{y}_{n+1} \pm t_{1-k, n-p-1} \cdot se(\hat{y}_{n+1})$$

where

$$se(\hat{y}_{n+1}) = \sqrt{\text{MSE} \left[1 + \mathbf{x}_{n+1}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_{n+1} \right]}$$

3.3 Summary

Multiple Linear Regression Notation

Symbol	Concept
β_j	The j^{th} regression coefficient
σ^2	Variance of response / Irreducible error
$\hat{\beta}_j$	Estimate/Estimator of β_j
MSE	Estimate of σ^2
\mathbf{X}	Design matrix
\mathbf{H}	Hat matrix
e	Residual
SST	Total sum of squares
SSR	Regression sum of squares
SSE	Error sum of squares
\hat{Y}	Estimator for $E[Y]$
se	Estimated standard error
df	Degrees of freedom
$t_{2(1-q), \text{df}}$	100 q^{th} percentile of a t -distribution
ndf	Numerator degrees of freedom
ddf	Denominator degrees of freedom
$F_{1-q, \text{ndf}, \text{ddf}}$	100 q^{th} percentile of an F -distribution

Assumptions

1. $Y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} + \varepsilon_i$
2. $x_{i,j}$'s are non-random
3. $E[\varepsilon_i] = 0$
4. $\text{Var}[\varepsilon_i] = \sigma^2$ (i.e. homoscedasticity)
5. ε_i 's are independent
6. ε_i 's are normally distributed

7. The predictor x_j is not a linear combination of the other p predictors, for $j = 0, 1, \dots, p$

Similar Concepts from Simple Linear Regression

ESTIMATION OF β_j

Ordinary least squares solves for $\hat{\beta}_j$ by minimizing SSE.

$$\begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix} = \hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \dots + \hat{\beta}_p x_p$$

$$\mathbf{H} = \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T, \quad \hat{\mathbf{y}} = \mathbf{H} \mathbf{y}$$

ESTIMATION OF σ^2

$$\text{MSE} = \frac{\text{SSE}}{n - p - 1}$$

$$\text{residual standard error} = \sqrt{\text{MSE}}$$

RESIDUALS

$$e = y - \hat{y}$$

SUM OF SQUARES

- $\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$ = total variability
- $\text{SSR} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ = explained variability
- $\text{SSE} = \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n e_i^2$ = unexplained variability
- $\text{SST} = \text{SSR} + \text{SSE}$

COEFFICIENT OF DETERMINATION

$$R^2 = \frac{\text{SSR}}{\text{SST}}$$

Adjusted R²

A measure of model quality that is suitable for comparing models, particularly those having different values of p .

$$\begin{aligned} R_{\text{adj.}}^2 &= 1 - \frac{\text{MSE}}{s_y^2} \\ &= 1 - (1 - R^2) \left(\frac{n - 1}{n - p - 1} \right) \end{aligned}$$

Special Variable Types

Types	Description
Higher-order variables	Variables raised to a power greater than 1
Dummy variables	Variables that take on 0 or 1; used for categorical predictors

Types	Description
Interactions	Product of predictors to represent dependence between them

Other Key Ideas

- R^2 is a poor measure for model comparison because it will increase simply by adding more predictors to a model.
- Polynomials do not change consistently by unit increases of its variable, i.e. no constant slope.
- Only $w - 1$ dummy variables are needed to represent w classes of a categorical predictor; one of the classes acts as a baseline.
- In effect, dummy variables define a distinct intercept for each class. Without the interaction between a dummy variable and a predictor, the dummy variable cannot additionally affect that predictor's regression coefficient.

Standard Errors

$$se(\hat{\beta}_j) = \sqrt{\widehat{\text{Var}}[\hat{\beta}_j]}$$

$$\widehat{\text{Var}}[\hat{\beta}] = \begin{bmatrix} \widehat{\text{Var}}[\hat{\beta}_0] & \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] & \dots & \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_p] \\ \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] & \widehat{\text{Var}}[\hat{\beta}_1] & \dots & \widehat{\text{Cov}}[\hat{\beta}_1, \hat{\beta}_p] \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_p] & \widehat{\text{Cov}}[\hat{\beta}_1, \hat{\beta}_p] & \dots & \widehat{\text{Var}}[\hat{\beta}_p] \end{bmatrix} = \text{MSE}(\mathbf{X}^T \mathbf{X})^{-1}$$

Estimating standard errors could also be done using bootstrapping.

Hypothesis Tests

T TESTS

$\text{df} = \text{denominator of MSE} = n - p - 1$

$$t. s. = \frac{\text{estimate} - \text{hypothesized value}}{\text{standard error}}$$

For an α level test, we reject H_0 if $p\text{-value} \leq \alpha$, or equivalently

Test Type	Critical Region
Left-tailed	$t. s. \leq -t_{2\alpha, n-p-1}$
Two-tailed	$ t. s. \geq t_{\alpha, n-p-1}$
Right-tailed	$t. s. \geq t_{2\alpha, n-p-1}$

F TESTS

Source	SS	df	MS
Regression	SSR	p	MSR
Error	SSE	$n - p - 1$	MSE
Total	SST	$n - 1$	s_y^2

- $\text{SSR} + \text{SSE} = \text{SST}$
- Regression df + Error df = Total df
- $\text{MS} = \frac{\text{SS}}{\text{df}}$

$$t. s. = \frac{\text{MSR}}{\text{MSE}} = \frac{\text{SSR} \div p}{\text{SSE} \div (n - p - 1)}$$

For an α level test, we reject H_0 if $p\text{-value} \leq \alpha$, or equivalently

$$t. s. \geq F_{\alpha, p, n-p-1}$$

PARTIAL F TESTS

The full model has more regression coefficients; the reduced model has fewer regression coefficients.

Source	SS	df
Reduced Regression	SSR_r	p_r
Difference	$\text{SSE}_r - \text{SSE}_f$ or $\text{SSR}_f - \text{SSR}_r$	$p_f - p_r$
Full Error	SSE_f	$n - p_f - 1$
Total	SST	$n - 1$

$$t. s. = \frac{\frac{\text{reduction in variability}}{(\text{SSE}_r - \text{SSE}_f)} \div \frac{\text{additional df spent}}{(p_f - p_r)}}{\text{SSE}_f \div (n - p_f - 1)}$$

For an α level test, we reject H_0 if $p\text{-value} \leq \alpha$, or equivalently

$$t. s. \geq F_{\alpha, p_f - p_r, n - p_f - 1}$$

Confidence Intervals

estimate \pm (t percentile) (standard error)

Quantity	Interval Expression
β_j	$\hat{\beta}_j \pm t_{1-k, n-p-1} \cdot se(\hat{\beta}_j)$
$E[Y]$	$\hat{y} \pm t_{1-k, n-p-1} \cdot se(\hat{y})$

3.4.0 Overview

 5m

The analysis of variance (ANOVA) concept was introduced in Section 3.3.7. This subsection covers additional detail on the topic. However, almost no new modeling technique is presented here; the models discussed are just specific setups of a multiple linear regression. Instead, the focus is on viewing the core concepts from a different perspective.

A different perspective is useful when the data is to be collected in a structured way, such as for a scientific experiment with specific design elements. Hence, an analyst might prefer to interpret and/or communicate the findings so that it directly relates to the experiment's design, and set aside the more generic multiple linear regression viewpoint.

3.4.1 One-Way ANOVA

A **one-way ANOVA model** is a multiple linear regression model with only **one** categorical variable predicting the response. It is also known as **one-factor ANOVA**, since "factor" is another term for a categorical predictor.

It is often that the levels of a factor are called **treatments** in this setting, which stems from subjects in an experiment made to undergo a treatment from w possible treatments. Then, we investigate whether the different treatments have any significant impact on the response.

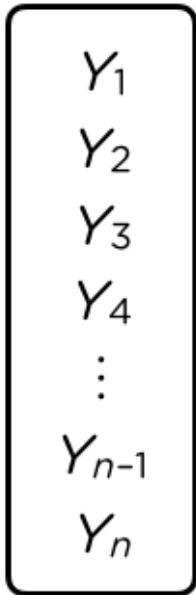
Recall that according to multiple linear regression, the model equation is

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_{w-1} x_{i,w-1} + \varepsilon_i$$

where $x_{i,1}, \dots, x_{i,w-1}$ are dummy variables for the i^{th} observation. For the sake of clarity, let the baseline category be treatment w (i.e. the last treatment), but remember this is an arbitrary choice. Specifically,

$$x_{i,j} = \begin{cases} 1, & \text{observation } i \text{ receives treatment } j \\ 0, & \text{otherwise} \end{cases}, \quad j = 1, \dots, w-1$$

To motivate the different form of the model equation used for one-way ANOVA, let's first consider how the data is now structured. The n total observations are divided into w groups for each treatment. We may view each treatment as having its own sample. The single sample (of n observations) becomes w samples (where Sample j has n_j observations, such that $\sum_{j=1}^w n_j = n$). Thus, we denote an observation using two indices rather than just one; Y_i now becomes $Y_{i,j}$. Let $Y_{i,j}$ be the response of the i^{th} observation for treatment j , where $i = 1, \dots, n_j$ and $j = 1, \dots, w$. The following diagram illustrates this configuration.



Coach's Remarks

Realize that the change in perspective alters the meaning of the index i . The original perspective uses i for the observations of the entire sample, whereas the new perspective uses i for the observations under a specific treatment.

The idea is that each treatment has its own mean response, so one version of the alternate model equation is

$$Y_{i,j} = \mu_j + \varepsilon_{i,j}$$

where $E[Y_{i,j}] = \mu_j$. Consequently, this equation uses the parameters μ_1, \dots, μ_w instead of the regression coefficients $\beta_0, \dots, \beta_{w-1}$. Yet, both model equations are inherently the same. To illustrate, imagine a treatment 1 observation, i.e. $x_{i,1} = 1$ and the other dummy variables are all 0. Depending on the equation, the mean response of this observation is either μ_1 or $\beta_0 + \beta_1$, meaning that $\mu_1 = \beta_0 + \beta_1$. So generally,

$$\mu_j = \begin{cases} \beta_0 + \beta_j, & j = 1, \dots, w-1 \\ \beta_0, & j = w \end{cases}$$

While writing the model equation this way seems intuitive, it can be tedious for other purposes, such as in hypothesis testing. Therefore, another version of the alternate model equation is

$$Y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j} \quad (3.4.1.1)$$

Effectively, μ_j is replaced by $\mu + \alpha_j$. This setup makes it problematic to estimate the $w+1$ parameters $(\mu, \alpha_1, \dots, \alpha_w)$ as is; there will not be a unique solution. In essence, it violates the 7th multiple linear regression model assumption listed in Section 3.3.1. To resolve this, we must "remove" one of these parameters.

CORNER POINT PARAMETERIZATION

One option is to let one of the α_j 's be 0, where the chosen parameter is arbitrary. In fact, the dummy variable coding of Section 3.3.4 performs corner point parameterization. This means selecting treatment w as the baseline category is equivalent to setting $\alpha_w = 0$. Consequently, $\mu + \alpha_j = \beta_0 + \beta_j$ leads to

$$\mu = \beta_0, \quad \alpha_j = \beta_j$$

for $j = 1, \dots, w-1$. This way, the interpretation of the parameters is straightforward using multiple linear regression knowledge:

- μ is the mean response for treatment w
- α_j is the change in the mean response for treatment j relative to treatment w

SUM-TO-ZERO CONSTRAINT

Another option is to let $\sum_{j=1}^w \alpha_j = 0$. This lets us express one of the α_j 's entirely in terms of the others, e.g. $\alpha_w = -\sum_{j=1}^{w-1} \alpha_j$. Then, the interpretation of the parameters is:

- μ is the average of the w mean responses

- α_j is the difference between the mean response for treatment j and the average of the w mean responses

Although the two approaches give different interpretations to the $w + 1$ parameters, the underlying model is the same, so the choice of approach ultimately does not matter. We proceed with corner point parameterization in this manual, keeping treatment w as the baseline category.

Parameter Estimation

Despite the exam not likely to emphasize it, calculating the parameter estimates is rather simple for this model. The key is that the mean response for treatment j will be estimated as the sample mean of the responses for treatment j observations. In math notation, this means \hat{y} for a treatment j observation equals \bar{y}_j , where

$$\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{i,j}}{n_j} \quad (3.4.1.2)$$

With our corner point parameterization, we get the equations

$$\begin{aligned}\hat{\mu} + \hat{\alpha}_1 &= \bar{y}_1 \\ &\vdots \\ \hat{\mu} + \hat{\alpha}_{w-1} &= \bar{y}_{w-1} \\ \hat{\mu} &= \bar{y}_w\end{aligned}$$

which produce the estimates of

$$\hat{\mu} = \bar{y}_w$$

$$\hat{\alpha}_j = \bar{y}_j - \bar{y}_w, \quad j = 1, \dots, w-1$$

ANOVA Table

We wish to test the hypotheses

- H_0 : The mean responses for the w treatments are all the same.
- H_1 : The mean responses are not all the same.

or equivalently,

- $H_0 : \alpha_1 = \dots = \alpha_w = 0$
- $H_1 : \text{At least one } \alpha_j \neq 0 \text{ for } j = 1, \dots, w$

As one may expect, this is done with an ordinary F test; we want to discover whether the factor as a whole contributes significantly to the model.

First, let's rewrite the familiar sums of squares in the one-way ANOVA style.

$$\begin{aligned} \text{SSR} &= \sum_{j=1}^w \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2 \\ &= \sum_{j=1}^w n_j (\bar{y}_j - \bar{y})^2 \end{aligned} \tag{3.4.1.3}$$

$$\text{SSE} = \sum_{j=1}^w \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2 \tag{3.4.1.4}$$

$$\text{SST} = \sum_{j=1}^w \sum_{i=1}^{n_j} (y_{i,j} - \bar{y})^2 \tag{3.4.1.5}$$

To be clear, \bar{y} is the regular sample mean of the response, i.e.

$$\bar{y} = \frac{\sum_{j=1}^w \sum_{i=1}^{n_j} y_{i,j}}{n}$$

Hence, the ANOVA table is

Source	SS	df	MS
Factor	SSR	$w - 1$	MSR

Source	SS	df	MS
Error	SSE	$n - w$	MSE
Total	SST	$n - 1$	s_y^2

where the test statistic

$$\begin{aligned}
 t.s. &= \frac{\text{MSR}}{\text{MSE}} \\
 &= \frac{\text{SSR} \div (w - 1)}{\text{SSE} \div (n - w)}
 \end{aligned} \tag{3.4.1.6}$$

comes from an F -distribution with $\text{ndf} = w - 1$ and $\text{ddf} = n - w$.

Coach's Remarks

Some resources use alternate ANOVA terminology such as:

- SSTR (treatment sum of squares) → SSR
- Residual source → Error source

Example 3.4.1.1

To study the impact of policyholder risk class (low, medium, high) on claim sizes, a one-way ANOVA model is used to analyze the following data:

Risk Class	Claims
Low	5 7 7 8
Medium	6 7 8 9
High	6 8 9 9

Calculate the test statistic that is used to evaluate whether risk class is a significant predictor of claim sizes.

Solution

The goal is to calculate

$$\frac{\text{SSR} \div (w - 1)}{\text{SSE} \div (n - w)}$$

Compute SSR with Equation 3.4.1.3 and SSE with Equation 3.4.1.4. First, determine the overall sample mean and the sample means for the three risk classes.

$$\bar{y} = \frac{5 + 7 + \dots + 9}{12} = 7.417$$

$$\bar{y}_1 = \frac{5 + 7 + 7 + 8}{4} = 6.75$$

$$\bar{y}_2 = \frac{6 + 7 + 8 + 9}{4} = 7.5$$

$$\bar{y}_3 = \frac{6 + 8 + 9 + 9}{4} = 8$$

$$\begin{aligned} \text{SSR} &= \sum_{j=1}^3 \sum_{i=1}^4 (\bar{y}_j - \bar{y})^2 \\ &\quad (6.75 - 7.417)^2 + \dots + (6.75 - 7.417)^2 \\ &=+ (7.5 - 7.417)^2 + \dots + (7.5 - 7.417)^2 \\ &\quad + (8 - 7.417)^2 + \dots + (8 - 7.417)^2 \\ &= 4(6.75 - 7.417)^2 + 4(7.5 - 7.417)^2 + 4(8 - 7.417)^2 \\ &= 3.167 \end{aligned}$$

$$\begin{aligned}
 SSE &= \sum_{j=1}^3 \sum_{i=1}^4 (y_{i,j} - \bar{y}_j)^2 \\
 &\quad (5 - 6.75)^2 + \dots + (8 - 6.75)^2 \\
 &=+ (6 - 7.5)^2 + \dots + (9 - 7.5)^2 \\
 &\quad + (6 - 8)^2 + \dots + (9 - 8)^2 \\
 &= 4.75 + 5 + 6 \\
 &= 15.75
 \end{aligned}$$

Therefore, the answer is

$$\frac{3.167 \div (3 - 1)}{15.75 \div (12 - 3)} = 0.905$$



Example 3.4.1.2

A one-way ANOVA model has the equation

$$Y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}$$

where $i = 1, \dots, n_j$ and $j = 1, 2, 3$. You are given:

- | Treatment, j | Sample size, n_j | Average response by treatment, \bar{y}_j |
|----------------|--------------------|--|
| 1 | 4 | 19.23 |
| 2 | 6 | 16.82 |
| 3 | 4 | 12.37 |

- The unbiased sample variance of the response is 26.45.

Determine whether the single factor is significant to the model at the 0.05 level.

Solution

To determine the test result, we must calculate SSR and SSE. Finding SSR requires knowing the sample mean of the response, \bar{y} .

$$\begin{aligned}\bar{y} &= \frac{\sum_{i=1}^{n_1} y_{i,1} + \sum_{i=1}^{n_2} y_{i,2} + \sum_{i=1}^{n_3} y_{i,3}}{n} \\ &= \frac{n_1 \bar{y}_1 + n_2 \bar{y}_2 + n_3 \bar{y}_3}{n_1 + n_2 + n_3} \\ &= \frac{4(19.23) + 6(16.82) + 4(12.37)}{4 + 6 + 4} \\ &= 16.237\end{aligned}$$

$$\begin{aligned}\text{SSR} &= \sum_{j=1}^3 \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2 \\ &= \sum_{j=1}^3 n_j (\bar{y}_j - \bar{y})^2 \\ &= 4(19.23 - 16.237)^2 + 6(16.82 - 16.237)^2 + 4(12.37 - 16.237)^2 \\ &= 97.686\end{aligned}$$

Next, compute SSE by first solving for SST. With $n = 4 + 6 + 4 = 14$ and $s_y^2 = 26.45$, obtain $\text{SST} = (14 - 1)26.45 = 343.85$. As a result,

$$\text{SSE} = \text{SST} - \text{SSR} = 246.164$$

Therefore, the test statistic is

$$\frac{97.686 \div (3 - 1)}{246.164 \div (14 - 3)} = 2.183$$

Since the test statistic is less than the critical value of $F_{0.05, 2, 11} = 3.982$, we conclude that the **single factor is not significant to the model at the 0.05 level**.

3.4.2 Two-Way ANOVA

As the name implies, a **two-way ANOVA model** is a multiple linear regression model with only **two** categorical variables predicting the response. We label the categorical variables as Factor A (having w levels) and Factor B (having v levels). This means the total number of treatments is $w \cdot v$, since each level from Factor A can be paired with each level from Factor B.

There are two types of models to discuss:

- Additive model
- Model with interactions

Additive Model

From the multiple linear regression viewpoint, the **additive model** has $p = w + v - 2$ predictors because there are

- $w - 1$ dummy variables from Factor A, plus
- $v - 1$ dummy variables from Factor B.

For this exam, we only need to work with a **balanced** dataset, meaning the observation counts for each treatment are all equal. We denote n_* as the number of observations for every treatment.

A version of the model equation in the two-way ANOVA style is

$$Y_{i,j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{i,j,k}, \quad (3.4.2.1)$$

$$i = 1, \dots, n_*, \quad j = 1, \dots, w, \quad k = 1, \dots, v$$

where

- $Y_{i,j,k}$ is the response of the i^{th} observation for treatment $\{j, k\}$
- α_j is the j^{th} **main effect** for Factor A
- β_k is the k^{th} **main effect** for Factor B

Analogous to the one-way ANOVA model, we will not obtain a unique solution by estimating these $w + v + 1$ parameters as is. We may use corner point parameterization or sum-to-zero constraints to "remove" two parameters, which means only $w + v - 1$ parameters will be estimated.

ANOVA TABLE

For the additive model, the ANOVA table associated with an ordinary F test is

Source	SS	df
Factor A and Factor B	SSR_{add}	$w + v - 2$
Error	SSE_{add}	$n - w - v + 1$
Total	SST	$n - 1$

In order to test hypotheses that involve the α_j 's only, or the β_k 's only, we partition the "Factor A and Factor B" source into two – one source per factor. This is done through one of two approaches:

1. Obtain results from the one-way ANOVA model with Factor A only
2. Obtain results from the one-way ANOVA model with Factor B only

Due to having balanced data, both approaches lead to the same outcome.

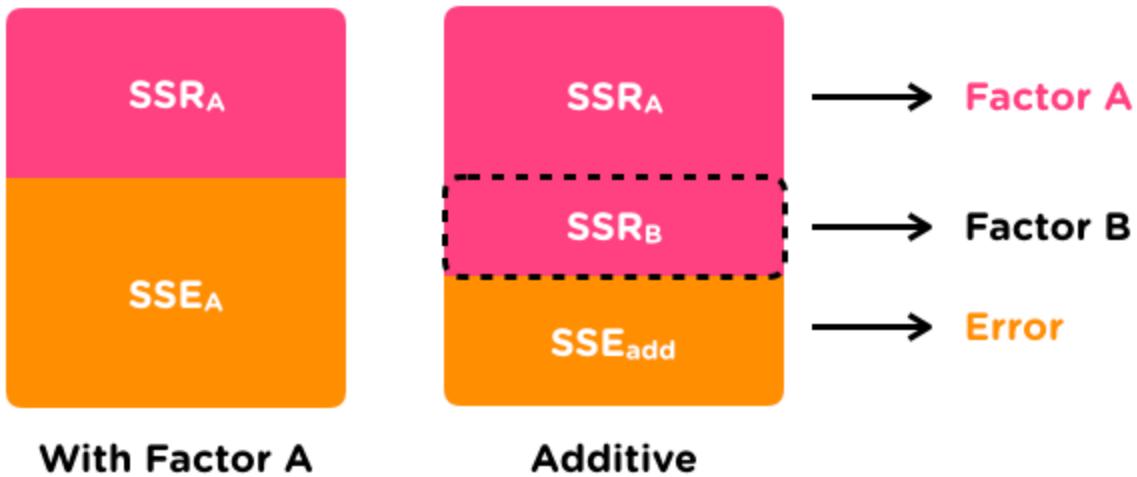
Let's assume the first approach. From the perspective of partial F tests in Section 3.3.7, we view the additive model as the full model, and the one-way ANOVA model with Factor A only as the reduced model.

We ultimately rewrite the additive model ANOVA table as

Source	SS	df
Factor A	SSR_A	$w - 1$
Factor B	SSR_B	$v - 1$
Error	SSE_{add}	$n - w - v + 1$
Total	SST	$n - 1$

where

$$\begin{aligned}\text{SSR}_B &= \text{SSE}_A - \text{SSE}_{\text{add}} \\ &= \text{SSR}_{\text{add}} - \text{SSR}_A\end{aligned}$$



Then, the hypotheses

- $H_0 : \alpha_1 = \dots = \alpha_w = 0$
- $H_1 : \text{At least one } \alpha_j \neq 0 \text{ for } j = 1, \dots, w$

can be evaluated by the test statistic

$$t. s. = \frac{SSR_A/(w-1)}{SSE_{add}/(n-w-v+1)} \quad (3.4.2.2)$$

which comes from an F -distribution with $\text{ndf} = w-1$ and $\text{ddf} = n-w-v+1$. This examines whether Factor A contributes significantly to the model.

On the other hand, the hypotheses

- $H_0 : \beta_1 = \dots = \beta_v = 0$
- $H_1 : \text{At least one } \beta_k \neq 0 \text{ for } k = 1, \dots, v$

can be evaluated by the test statistic

$$t. s. = \frac{SSR_B/(v-1)}{SSE_{add}/(n-w-v+1)} \quad (3.4.2.3)$$

which comes from an F -distribution with $\text{ndf} = v-1$ and $\text{ddf} = n-w-v+1$. This examines whether Factor B contributes significantly to the model.

Example 3.4.2.1

In considering ANOVA models to predict weight loss by participants in a fitness program, you are given:

- The available predictors are:
 - Initial weight class ("Overweight" or "Not overweight")
 - Age ("Under 30", "30 to 45", or "Over 45")
- The data is balanced with 18 participants total.

•

Predictors used	MSE
Age only	6.2333
Initial weight class and Age	1.8175

- The total sum of squares is 183.611.

Determine whether the initial weight class is a significant predictor at the 0.02 level.

Solution

Let the initial weight class be Factor A, and age be Factor B. Start by calculating the SSEs of the one-way model with age only and the additive model. With $n = 18$, $w = 2$, and $v = 3$,

$$6.2333 = \frac{\text{SSE}_B}{n - v} \quad \Rightarrow \quad \text{SSE}_B = 6.2333 (18 - 3) = 93.4995$$

$$1.8175 = \frac{\text{SSE}_{\text{add}}}{n - w - v + 1} \quad \Rightarrow \quad \text{SSE}_{\text{add}} = 1.8175 (18 - 2 - 3 + 1) = 25.445$$

The MSE formulas are essentially Equation 3.3.2.1, but they may be easier to remember using ANOVA tables. Rather than using formulas, the error degrees of freedom can be solved using multiple linear regression knowledge instead. For example, the additive model

has $p + 1 = (w - 1) + (v - 1) + 1 = 4$ regression coefficients: one for each dummy variable and the intercept. Therefore, $n - w - v + 1$ must equal $18 - 4 = 14$.

Next, determine the ANOVA table for the additive model with each factor given its own row. With $SST = 183.611$,

$$\begin{aligned} SSR_B &= SST - SSE_B \\ &= 183.611 - 93.4995 \\ &= 90.1115 \end{aligned}$$

$$\begin{aligned} SSR_A &= SSE_B - SSE_{\text{add}} \\ &= 93.4995 - 25.445 \\ &= 68.0545 \end{aligned}$$

Source	SS	df
Initial weight class	68.0545	1
Age	90.1115	2
Error	25.445	14
Total	183.611	17

We may sum the SS and df columns of the first three rows to check our work against the last row.

To test the significance of the initial weight class, we need the mean squares of the first and third rows of the ANOVA table. Therefore, the test statistic is

$$\frac{68.0545/1}{25.445/14} = 37.444$$

Since the test statistic is greater than the critical value of $F_{0.02, 1, 14} = 6.888$, we conclude that the **initial weight class is significant to the model at the 0.02 level.**

Model with Interactions

From the multiple linear regression viewpoint, the ***model with interactions*** has $p = wv - 1$ predictors because there are

- $w - 1$ dummy variables from Factor A, plus
- $v - 1$ dummy variables from Factor B, plus
- $(w - 1)(v - 1) = wv - w - v + 1$ interaction terms.

A version of the model equation in the two-way ANOVA style is

$$Y_{i,j,k} = \mu + \alpha_j + \beta_k + \gamma_{j,k} + \varepsilon_{i,j,k}, \quad (3.4.2.4)$$

$$i = 1, \dots, n_*, \quad j = 1, \dots, w, \quad k = 1, \dots, v$$

where

- $Y_{i,j,k}$, α_j , and β_k are defined the same as in the additive model
- $\gamma_{j,k}$ is an ***interaction effect***

Once again, we assume that the data is balanced, and may use corner point parameterization or sum-to-zero constraints to obtain unique parameter estimates.

ANOVA TABLE

For the model with interactions, the ANOVA table associated with an ordinary F test is

Source	SS	df
Factor A, Factor B, and Interaction	SSR_{int}	$wv - 1$
Error	SSE_{int}	$n - wv$
Total	SST	$n - 1$

We want to partition the "Factor A, Factor B, and Interaction" source into three – one source per factor and one for the interactions. This is achieved by following these steps:

1. Derive the Interaction source by using the additive model.

2. As described previously, derive the Factor A and Factor B sources by using the one-way model with Factor A only, or the one-way model with Factor B only.

So, by viewing the model with interactions as the full model, and the additive model as the reduced model, we get the ANOVA table

Source	SS	df
Factor A and Factor B	SSR_{add}	$w + v - 2$
Interaction	$\text{SSE}_{\text{add}} - \text{SSE}_{\text{int}}$ or $\text{SSR}_{\text{int}} - \text{SSR}_{\text{add}}$	$(w - 1)(v - 1)$
Error	SSE_{int}	$n - wv$
Total	SST	$n - 1$

Then, the first row is partitioned in exactly the same manner as shown in our additive model discussion. This produces the final table of

Source	SS	df
Factor A	SSR_A	$w - 1$
Factor B	SSR_B	$v - 1$
Interaction	$\text{SSE}_{\text{add}} - \text{SSE}_{\text{int}}$ or $\text{SSR}_{\text{int}} - \text{SSR}_{\text{add}}$	$(w - 1)(v - 1)$
Error	SSE_{int}	$n - wv$
Total	SST	$n - 1$

Coach's Remarks

We encourage you to not memorize the specific components of any ANOVA table in Section 3.4. Understanding the concepts in Section 3.3.7 is sufficient to derive them.

To test the hypotheses

- $H_0 : \text{All } \gamma_{j,k} = 0$
- $H_1 : \text{At least one } \gamma_{j,k} \neq 0$

we use the test statistic

$$t.s. = \frac{(\text{SSE}_{\text{add}} - \text{SSE}_{\text{int}})/[(w-1)(v-1)]}{\text{SSE}_{\text{int}}/(n-wv)} \quad (3.4.2.5)$$

which comes from an F -distribution with $\text{ndf} = (w-1)(v-1)$ and $\text{ddf} = n - wv$. This examines whether the interactions contribute significantly to the model.

To test the hypotheses

- $H_0 : \alpha_1 = \dots = \alpha_w = 0$
- $H_1 : \text{At least one } \alpha_j \neq 0 \text{ for } j = 1, \dots, w$

we use the test statistic

$$t.s. = \frac{\text{SSR}_A/(w-1)}{\text{SSE}_{\text{int}}/(n-wv)} \quad (3.4.2.6)$$

which comes from an F -distribution with $\text{ndf} = w-1$ and $\text{ddf} = n - wv$. This examines whether Factor A contributes significantly to the model.

To test the hypotheses

- $H_0 : \beta_1 = \dots = \beta_v = 0$
- $H_1 : \text{At least one } \beta_k \neq 0 \text{ for } k = 1, \dots, v$

we use the test statistic

$$t.s. = \frac{\text{SSR}_B/(v-1)}{\text{SSE}_{\text{int}}/(n-wv)} \quad (3.4.2.7)$$

which comes from an F -distribution with $\text{ndf} = v-1$ and $\text{ddf} = n - wv$. This examines whether Factor B contributes significantly to the model.

Example 3.4.2.2

In considering ANOVA models to predict weight loss by participants in a fitness program, you are given:

- The available predictors are:
 - Initial weight class ("Overweight" or "Not overweight")
 - Age ("Under 30", "30 to 45", or "Over 45")
- The data is balanced with 18 participants total.

-

Predictors used	MSE
Initial weight class only	7.2222
Age only	6.2333
Initial weight class, Age, and their interactions	1.7222

- The total sum of squares is 183.611.

Determine whether the interactions between initial weight class and age are significant to the model at the 0.05 level.

Solution

Let the initial weight class be Factor A, and age be Factor B. With $n = 18$, $w = 2$, and $v = 3$, we calculate the SSEs of the three specified models as

$$7.2222 = \frac{\text{SSE}_A}{n - w} \quad \Rightarrow \quad \text{SSE}_A = 7.2222 (18 - 2) = 115.5552$$

$$6.2333 = \frac{\text{SSE}_B}{n - v} \quad \Rightarrow \quad \text{SSE}_B = 6.2333 (18 - 3) = 93.4995$$

$$1.7222 = \frac{\text{SSE}_{\text{int}}}{n - wv} \quad \Rightarrow \quad \text{SSE}_{\text{int}} = 1.7222 [18 - 2(3)] = 20.6664$$

Again, the MSE formulas come from Equation 3.3.2.1, but remembering them through ANOVA tables might be easier.

With balanced data, we have

$$\text{SSR}_{\text{add}} = \text{SSR}_A + \text{SSR}_B$$

Thus, we calculate the SS for the Interaction source and populate the ANOVA table as follows:

$$\begin{aligned}\text{SSR}_{\text{int}} - \text{SSR}_{\text{add}} &= (\text{SST} - \text{SSE}_{\text{int}}) - (\text{SSR}_A + \text{SSR}_B) \\ &= (\text{SST} - \text{SSE}_{\text{int}}) - ([\text{SST} - \text{SSE}_A] + [\text{SST} - \text{SSE}_B]) \\ &= \text{SSE}_A + \text{SSE}_B - \text{SST} - \text{SSE}_{\text{int}} \\ &= 115.5552 + 93.4995 - 183.611 - 20.6664 \\ &= 4.7773\end{aligned}$$

Source	SS	df
Initial weight class	SSR_A	1
Age	SSR_B	2
Interaction	4.7773	2
Error	20.6664	12
Total	183.611	17

We do not need the values of SSR_A and SSR_B to solve this problem.

Calculate the test statistic by taking the ratio of the mean squares of the third and fourth rows of the ANOVA table.

$$\frac{4.7773/2}{20.6664/12} = 1.387$$

Since the test statistic is less than the critical value of $F_{0.05, 2, 12} = 3.885$, we conclude that the interactions are not significant to the model at the 0.05 level.

Example 3.4.2.3

You are given the following ANOVA tables that consider two factors for predicting a response variable:

- For the one-way ANOVA model with Factor X,

Source	Sum of squares	Degrees of freedom
Factor X	388.34	2
Error	1,254.50	20

- For the additive model with Factor X and Factor Y,

Source	Sum of squares	Degrees of freedom
Factor X and Factor Y	1,197.27	6
Error	445.57	16

- For the model with Factor X, Factor Y, and their interactions,

Source	Sum of squares	Degrees of freedom
Factor X, Factor Y, and Interaction	1,556.39	14
Error	86.45	8

Calculate the test statistic that examines whether Factor Y is a significant predictor.

Solution

Rewrite the last ANOVA table such that the first row is divided into three sources: Factor X, Factor Y, and Interaction. In this case, we skip including the Total source since it does not impact the solution. Start by viewing the model with interactions as the full model, and the additive model as the reduced model. Then, obtain

Source	SS	df
Factor X and Factor Y	1,197.27	6
Interaction	359.12	8
Error	86.45	8

where

$$\begin{aligned} 359.12 &= 445.57 - 86.45 \\ &= 1,556.39 - 1,197.27 \end{aligned}$$

and the Interaction df can be solved as $14 - 6 = 8$.

Repeat the process, where we now view the additive model as the full model, and the one-way model with Factor X only as the reduced model. Then, obtain

Source	SS	df
Factor X	388.34	2
Factor Y	808.93	4
Interaction	359.12	8
Error	86.45	8

where

$$\begin{aligned} 808.93 &= 1,254.50 - 445.57 \\ &= 1,197.27 - 388.34 \end{aligned}$$

and the Factor Y df can be solved as $6 - 2 = 4$.

Finally, calculate the test statistic using Equation 3.4.2.7; take the ratio of the mean squares of the second and fourth rows of the ANOVA table.

$$\frac{808.93/4}{86.45/8} = 18.71$$

Coach's Remarks

Notice that we used Equation 3.4.2.7 to calculate the test statistic rather than Equation 3.4.2.3, which would have resulted in

$$\frac{808.93/4}{445.57/16} = 7.26$$

This can be confusing, as both Equations 3.4.2.3 and 3.4.2.7 are test statistics for what appears to be the same hypothesis test. Clearly, the formulas have different denominators: the former has the MSE of the additive model, while the latter has the MSE of the model with interactions. The same can be said of Equations 3.4.2.2 and 3.4.2.6.

This difference technically leads to different hypothesis tests. But for this exam, this detail is not crucial. When the problem is not explicit on which models is the hypothesis test comparing, the default is to use the MSE of the model that has the most predictors in that context.

To illustrate, Example 3.4.2.1 uses the test statistic with the additive model's MSE in the denominator because it has the most predictors in that context. However, this example uses the test statistic with the model with interactions' MSE in the denominator because it has the most predictors in this context.

Example 3.4.2.4

Two categorical variables were considered for predicting a response variable. Some ANOVA results are provided in the table below:

Source	Sum of squares	Degrees of freedom
Categorical variable 1	k	2
Categorical variable 2	20.847	2
Interaction	12.221	4
Error	?	8

Additionally, the sum of squared residuals for the additive model is 40.646.

Determine the smallest value of k such that categorical variable 1 is significant to the model at the 0.05 level.

Solution

We are given $\text{SSE}_{\text{add}} = 40.646$, which means

$$\text{SSE}_{\text{add}} - \text{SSE}_{\text{int}} = 12.221$$

$$\begin{aligned}\Rightarrow \text{SSE}_{\text{int}} &= \text{SSE}_{\text{add}} - 12.221 \\ &= 40.646 - 12.221 \\ &= 28.425\end{aligned}$$

Categorical variable 1 is significant to the model at the 0.05 level when

$$\begin{aligned}t. s. &\geq F_{0.05, 2, 8} \\ \frac{k/2}{28.425/8} &\geq 4.459 \\ k &\geq 4.459 (28.425/8) (2) \\ &= 31.687\end{aligned}$$

Therefore, the smallest value of k that results in a significant categorical variable 1 at the 0.05 level is **31.687**.



3.4.3 Additive Model Without Replication

Replication refers to the plurality of observations for a certain treatment. Thus, a treatment without replication means that treatment has only one observation.

In keeping the structure of balanced data, an additive model without replication has $n_* = 1$ for all treatments, which implies $n = wv$. Moreover, there is no purpose for the index i ; the model equation simplifies to

$$Y_{j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{j,k}, \quad j = 1, \dots, w, \quad k = 1, \dots, v$$

The formulas of the relevant sums of squares are now relatively simple. Before discussing them, let's define the following sample means:

- $\bar{y}_{j\bullet}$ is the average of responses from level j of Factor A, across all v Factor B levels, i.e.

$$\bar{y}_{j\bullet} = \frac{\sum_{k=1}^v y_{j,k}}{v} \tag{3.4.3.1}$$

- $\bar{y}_{\bullet k}$ is the average of responses from level k of Factor B, across all w Factor A levels, i.e.

$$\bar{y}_{\bullet k} = \frac{\sum_{j=1}^w y_{j,k}}{w} \tag{3.4.3.2}$$

Then, the sums of squares have the following formulas:

$$\begin{aligned} \text{SSR}_A &= \sum_{k=1}^v \sum_{j=1}^w (\bar{y}_{j\bullet} - \bar{y})^2 \\ &= \sum_{j=1}^w v(\bar{y}_{j\bullet} - \bar{y})^2 \end{aligned} \tag{3.4.3.3}$$

$$\begin{aligned} \text{SSR}_B &= \sum_{k=1}^v \sum_{j=1}^w (\bar{y}_{\bullet k} - \bar{y})^2 \\ &= \sum_{k=1}^v w(\bar{y}_{\bullet k} - \bar{y})^2 \end{aligned} \tag{3.4.3.4}$$

$$\text{SSE}_{\text{add}} = \sum_{k=1}^v \sum_{j=1}^w (y_{j,k} - \bar{y}_{j\bullet} - \bar{y}_{\bullet k} + \bar{y})^2 \quad (3.4.3.5)$$

$$\text{SST} = \sum_{k=1}^v \sum_{j=1}^w (y_{j,k} - \bar{y})^2 \quad (3.4.3.6)$$

The following data records claim sizes in light of two policyholder attributes: gender and income level.

		Income Level		
		Low	Medium	High
Gender	Male	53	114	373
	Female	67	130	253

Calculate all the sums of squares associated with an ANOVA table for an additive model.

By letting gender be Factor A, and income level be Factor B, we have $w = 2$ and $v = 3$. To help organize the calculations, include another row and column to the table of data for the sample means.

		Income Level			Average
		Low	Medium	High	
Gender	Male	53	114	373	$\bar{y}_{1\bullet}$
	Female	67	130	253	$\bar{y}_{2\bullet}$
Average		$\bar{y}_{\bullet 1}$	$\bar{y}_{\bullet 2}$	$\bar{y}_{\bullet 3}$	\bar{y}

$$\bar{y}_{1\bullet} = \frac{53 + 114 + 373}{3} = 180, \quad \bar{y}_{2\bullet} = \frac{67 + 130 + 253}{3} = 150$$

$$\bar{y}_{\bullet 1} = \frac{53 + 67}{2} = 60, \quad \bar{y}_{\bullet 2} = \frac{114 + 130}{2} = 122, \quad \bar{y}_{\bullet 3} = \frac{373 + 253}{2} = 313$$

$$\bar{y} = \frac{53 + 114 + \dots + 253}{6} = 165$$

Therefore, the four sums of squares are

$$\begin{aligned} \text{SSR}_A &= \sum_{k=1}^3 \sum_{j=1}^2 (\bar{y}_{j\bullet} - \bar{y})^2 \\ &= (180 - 165)^2 + (180 - 165)^2 + (180 - 165)^2 \\ &+ (150 - 165)^2 + (150 - 165)^2 + (150 - 165)^2 \\ &= 3(180 - 165)^2 + 3(150 - 165)^2 \\ &= \mathbf{1,350} \end{aligned}$$

$$\begin{aligned} \text{SSR}_B &= \sum_{k=1}^3 \sum_{j=1}^2 (\bar{y}_{\bullet k} - \bar{y})^2 \\ &= (60 - 165)^2 + (122 - 165)^2 + (313 - 165)^2 \\ &+ (60 - 165)^2 + (122 - 165)^2 + (313 - 165)^2 \\ &= 2(60 - 165)^2 + 2(122 - 165)^2 + 2(313 - 165)^2 \\ &= \mathbf{69,556} \end{aligned}$$

$$\begin{aligned} \text{SSE}_{\text{add}} &= \sum_{k=1}^3 \sum_{j=1}^2 (y_{j,k} - \bar{y}_{j\bullet} - \bar{y}_{\bullet k} + \bar{y})^2 \\ &= (53 - 180 - 60 + 165)^2 + (114 - 180 - 122 + 165)^2 + (373 - 180 - 313 \\ &+ (67 - 150 - 60 + 165)^2 + (130 - 150 - 122 + 165)^2 + (253 - 150 - 313 \\ &= 3,038 + 3,038 \\ &= \mathbf{6,076} \end{aligned}$$

$$\begin{aligned} \text{SST} &= \sum_{k=1}^3 \sum_{j=1}^2 (y_{j,k} - \bar{y})^2 \\ &= (53 - 165)^2 + (114 - 165)^2 + \dots + (253 - 165)^2 \\ &= \mathbf{76,982} \end{aligned}$$

As a specific case of an additive model, all additive model concepts from the previous subsection still hold, e.g.

$$\text{SSR}_A + \text{SSR}_B + \text{SSE}_{\text{add}} = \text{SST}$$

Coach's Remarks

Using a model with interactions on balanced data without replication is not appropriate. This results in overfitting because the number of estimated parameters equals the number of observations. We will elaborate on this phenomenon in Section 3.5.1.

3.4.4 Miscellaneous Topics

There are two more topics that warrant discussion:

- ANCOVA models
- Uncorrected total

ANCOVA Models

An ***analysis of covariance (ANCOVA) model*** is a multiple linear regression model with quantitative predictors (i.e. covariates) in addition to categorical ones. We have already encountered ANCOVA models, such as the Commuting Chris's dummy variable model that was first introduced in Section 3.3.4; both Precip Chance and Season are used as predictors. Framing it as an ANCOVA model implies that the main interest is in testing the significance of the factor(s) in the model. By now, it should be clear how these partial F tests are performed. If not, review Sections 3.3.7 and 3.4.2.

Example 3.4.4.1

The costs of dental visits are analyzed by multiple linear regression using data on 20 patients. You are given:

- The available predictors are:
 - Age
 - Type of care ("preventative", "prosthetic", "braces", or "others")

-

Predictors used	R^2
Age only	0.478
Age and Type of care	0.654

- You want to perform the following hypothesis test:
 - H_0 : The simple linear regression model with Age is adequate.
 - H_1 : The ANCOVA model with both predictors is preferred.

Determine the test result at the 5% significance level.

Solution

Note that the simple linear regression with Age is the reduced model, while the ANCOVA model is the full model. The objective is to test whether the type of care is a significant predictor at the 0.05 level. Hence, the ANCOVA table we desire is

Source	SS	df
Age	SSR_r	1
Type of care	$\text{SSE}_r - \text{SSE}_f$ or $\text{SSR}_f - \text{SSR}_r$	3
Error	SSE_f	15
Total	SST	19

Solve for the degrees of freedom by noting:

- The Total df is 19, since $n = 20$
- The Age df is 1, since there is only one predictor for a simple linear regression, i.e. $p_r = 1$
- The Type of care df is 3, since the factor has 4 levels, thus requiring $4 - 1 = 3$ dummy variables
- The Error df is $19 - 1 - 3 = 15$, or alternatively, $n - p_f - 1 = 20 - 4 - 1 = 15$

Next, the test statistic is the ratio of the mean squares of the second and third rows of the ANCOVA table.

$$\frac{(\text{SSR}_f - \text{SSR}_r) \div 3}{\text{SSE}_f \div 15}$$

Since we are only given the R^2 's of the reduced and full models, we manipulate the test statistic to compute it as follows:

$$\begin{aligned}
 \frac{(\text{SSR}_f - \text{SSR}_r) \div 3}{\text{SSE}_f \div 15} \cdot \frac{1/\text{SST}}{1/\text{SST}} &= \frac{\left(\frac{\text{SSR}_f - \text{SSR}_r}{\text{SST}} \right) \div 3}{\left(\frac{\text{SSE}_f}{\text{SST}} \right) \div 15} \\
 &= \frac{\left(\frac{\text{SSR}_f}{\text{SST}} - \frac{\text{SSR}_r}{\text{SST}} \right) \div 3}{\left(1 - \frac{\text{SSR}_f}{\text{SST}} \right) \div 15} \\
 &= \frac{(0.654 - 0.478) \div 3}{(1 - 0.654) \div 15} \\
 &= 2.543
 \end{aligned}$$

Since the test statistic is less than the critical value of $F_{0.05, 3, 15} = 3.287$, we conclude that the **simple linear regression is adequate at the 0.05 level**; the type of care is not a significant enough predictor in the ANCOVA model.

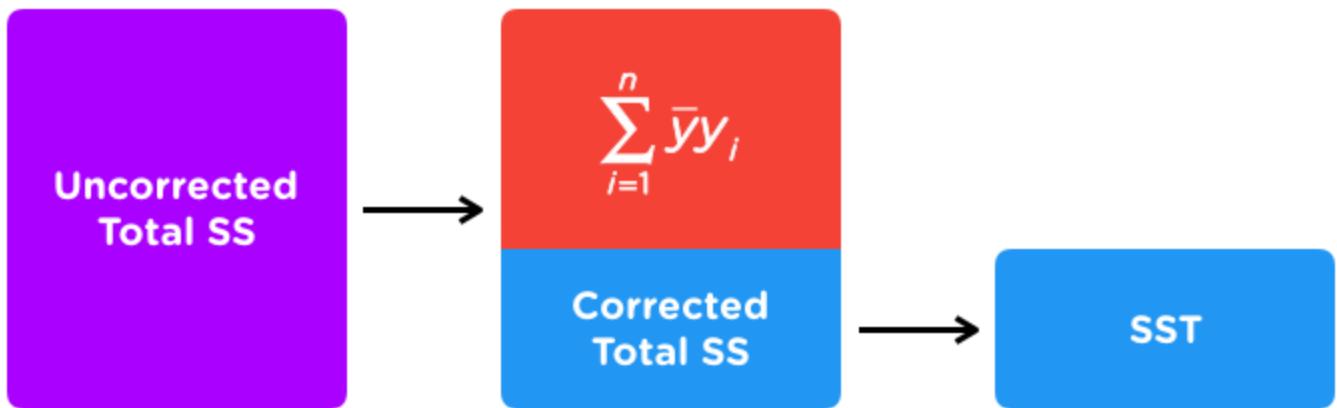


Uncorrected Total

SST is referred to as the "**corrected** total sum of squares" when it is necessary to be more specific. To motivate this idea, let's take a closer look at the SST formula. In regular, non-ANOVA notation, we can express SST as

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \sum_{i=1}^n \bar{y}y_i$$

This is demonstrated in the appendix at the end of the section. The first term of this alternate form, $\sum_{i=1}^n y_i^2$, is called the **uncorrected total sum of squares**. This makes SST a so-called corrected value, where the correction amount is $\sum_{i=1}^n \bar{y}y_i$.



Next, note that

$$\begin{aligned}
 \sum_{i=1}^n y_i^2 &= \sum_{i=1}^n \bar{y}y_i + SST \\
 &= \sum_{i=1}^n \bar{y}y_i + SSR + SSE
 \end{aligned}$$

This decomposition of the uncorrected total sum of squares makes $\sum_{i=1}^n \bar{y}y_i$ look like another type of sum of squares. To keep things simple, we say $\sum_{i=1}^n \bar{y}y_i$ is a sum of squares for a Mean source because it involves the sample mean, \bar{y} . Consequently, SST is also known as the "total sum of squares corrected for the mean".

It follows that an ANOVA/ANCOVA table may have another row for the Mean source, in which case the Total source refers to the uncorrected total instead, i.e.

Source	SS	df
Mean	$\sum_{i=1}^n \bar{y}y_i$	1
Regression	SSR	p
Error	SSE	$n - p - 1$
Total	$\sum_{i=1}^n y_i^2$	n

Moreover, some authors prefer a column header of "Parameters" rather than "Source", which would likely denote the "Mean" row as the "Intercept" row instead.

Realize that the hypothesis tests remain unchanged. The key is to identify whether the sources sum to the corrected total or the uncorrected total.

3.4 Summary

ANOVA and ANCOVA models are different ways to express multiple linear regression models that use categorical predictors (i.e. factors).

One-Way ANOVA

For a factor with w levels, the model equation can be written as

$$Y_{i,j} = \mu + \alpha_j + \varepsilon_{i,j}$$

$$i = 1, \dots, n_j, \quad j = 1, \dots, w$$

Let

$$\bar{y}_j = \frac{\sum_{i=1}^{n_j} y_{i,j}}{n_j}$$

thus

- $\text{SSR} = \sum_{j=1}^w \sum_{i=1}^{n_j} (\bar{y}_j - \bar{y})^2 = \sum_{j=1}^w n_j (\bar{y}_j - \bar{y})^2$
- $\text{SSE} = \sum_{j=1}^w \sum_{i=1}^{n_j} (y_{i,j} - \bar{y}_j)^2$
- $\text{SST} = \sum_{j=1}^w \sum_{i=1}^{n_j} (y_{i,j} - \bar{y})^2$

Source	SS	df
Factor	SSR	$w - 1$
Error	SSE	$n - w$
Total	SST	$n - 1$

Two-Way ANOVA

Let Factor A have w levels and Factor B have v levels. In addition, assume the data is balanced.

ADDITIVE MODEL

The model equation can be written as

$$Y_{i,j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{i,j,k}$$

$$i = 1, \dots, n_*, \quad j = 1, \dots, w, \quad k = 1, \dots, v$$

Source	SS	df
Factor A	SSR _A	$w - 1$
Factor B	SSR _B	$v - 1$
Error	SSE _{add}	$n - w - v + 1$
Total	SST	$n - 1$

$$\begin{aligned} \text{SSR}_B &= \text{SSE}_A - \text{SSE}_{\text{add}} \\ &= \text{SSR}_{\text{add}} - \text{SSR}_A \end{aligned}$$

ADDITIVE MODEL WITHOUT REPLICATION

With $n_* = 1$, the model equation can be written as

$$Y_{j,k} = \mu + \alpha_j + \beta_k + \varepsilon_{j,k}$$

$$j = 1, \dots, w, \quad k = 1, \dots, v$$

Let

$$\bar{y}_{j\bullet} = \frac{\sum_{k=1}^v y_{j,k}}{v}, \quad \bar{y}_{\bullet k} = \frac{\sum_{j=1}^w y_{j,k}}{w}$$

thus

- $\text{SSR}_A = \sum_{k=1}^v \sum_{j=1}^w (\bar{y}_{j\bullet} - \bar{y})^2 = \sum_{j=1}^w v(\bar{y}_{j\bullet} - \bar{y})^2$
- $\text{SSR}_B = \sum_{k=1}^v \sum_{j=1}^w (\bar{y}_{\bullet k} - \bar{y})^2 = \sum_{k=1}^v w(\bar{y}_{\bullet k} - \bar{y})^2$
- $\text{SSE}_{\text{add}} = \sum_{k=1}^v \sum_{j=1}^w (y_{j,k} - \bar{y}_{j\bullet} - \bar{y}_{\bullet k} + \bar{y})^2$
- $\text{SST} = \sum_{k=1}^v \sum_{j=1}^w (y_{j,k} - \bar{y})^2$

MODEL WITH INTERACTIONS

The model equation can be written as

$$Y_{i,j,k} = \mu + \alpha_j + \beta_k + \gamma_{j,k} + \varepsilon_{i,j,k}$$

$$i = 1, \dots, n_*, \quad j = 1, \dots, w, \quad k = 1, \dots, v$$

Source	SS	df
Factor A	SSR_A	$w - 1$
Factor B	SSR_B	$v - 1$
Interaction	$\text{SSE}_{\text{add}} - \text{SSE}_{\text{int}}$ or $\text{SSR}_{\text{int}} - \text{SSR}_{\text{add}}$	$(w - 1)(v - 1)$
Error	SSE_{int}	$n - wv$
Total	SST	$n - 1$

Other Key Ideas

- In testing whether a source is significant, the test statistic is the mean square of that source divided by the MSE of the model that has the most predictors in that context, which comes from an F -distribution.
- ANCOVA models have both quantitative and categorical predictors.
- The uncorrected total sum of squares is $\sum_{i=1}^n y_i^2$. The sources of an ANOVA/ANCOVA table may sum to the uncorrected total rather than the corrected total.

Appendix

🕒 5m

Alternate Form of Total Sum of Squares

Recall that

$$\text{SST} = \sum_{i=1}^n (y_i - \bar{y})^2$$

While we may expand this as is, let's use the result that was mentioned at the end of Example 3.2.2.1, which states

$$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2$$

Therefore,

$$\begin{aligned} \sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n y_i^2 - n\bar{y}^2 \\ &= \sum_{i=1}^n y_i^2 - \bar{y}(n\bar{y}) \\ &= \sum_{i=1}^n y_i^2 - \bar{y} \left(\sum_{i=1}^n y_i \right) \\ &= \sum_{i=1}^n y_i^2 - \sum_{i=1}^n \bar{y}y_i \end{aligned}$$

3.5.0 Overview

 5m

We can rely on the linear regression results when there are no severe violations of the model assumptions. On the flip side, depending on the results at face value is unwise. This subsection will detail the impact of violated assumptions, ways to detect these issues, and suggestions on how to correct them.

3.5.1 Violations and Issues

Here is a list of concerns when handling a multiple linear regression model:

1. Misspecified model equation
2. Residuals with non-zero averages
3. Heteroscedasticity
4. Dependent errors
5. Non-normal errors
6. Multicollinearity
7. Outliers
8. High leverage points
9. High dimensions

Note that the first six issues parallel the regression assumptions (excluding the 2nd assumption) given in Section 3.3.1.

Misspecified Model Equation

This refers to incorrectly assuming that the true form of f resembles Equation 3.3.1.1, or failing to include appropriate predictors. One example of the latter is to fit a linear relationship between response and predictor when there is evidence of a polynomial relationship.

As mentioned before, it is near impossible to know f 's true form in practice. However, we must be aware of the potential signs that show the shortcomings of the predictors, or that a linear regression is just inappropriate. Clearly, a severely wrong model equation can invalidate all regression results.

Residuals with Non-Zero Averages

Note that residuals (e) are the realizations of the random error terms (ε) assuming the regression model is correct. Therefore, attributes of the error terms should be evident when studying the residuals. For example, if the error terms have a true mean of 0, then the average of residuals should be close to 0. As a result, an average of residuals that is far from 0 suggests that some aspect of the linear regression is incorrect; this is typically a symptom of a model violation rather than a violation itself.

A common check is to average the residuals for observations that have similar predictions, \hat{y} , producing several averages that should all be close to 0. However, the check will not work by averaging all the residuals together because it is theoretically different from error terms with mean 0. In addition, ordinary least squares guarantees that the average of all the residuals is 0.

Heteroscedasticity

The opposite of homoscedasticity is **heteroscedasticity**. This means the variance of the error term is non-constant across the observations.

Since homoscedasticity means there is one variance parameter σ^2 , this violation implies there are several variance parameters. This leads to an unreliable MSE because it views the several parameters as one parameter. In turn, all outputs that depend on the MSE (e.g. estimation of standard errors and performing statistical inference) are questionable.

Dependent Errors

Dependent or correlated error terms will behave predictably from observation to observation. As an example of positively correlated error terms, if the realization of ε_i is positive, then the realization of ε_{i+1} would also tend to be positive, and vice versa for negative realizations.

When the error terms are dependent, the Y_i 's have non-zero covariances. Incorrectly assuming that the errors are independent typically leads to underestimated *se*'s. Consequently,

- confidence and prediction intervals will be narrower, and
- *p*-values will be smaller

than they should be. Then, it becomes possible to wrongly reject H_0 based on an incorrect, smaller *p*-value.

Non-Normal Errors

When the error terms do not follow a normal distribution, we likely cannot conclude that certain estimators follow a *t*-distribution or an *F*-distribution. Performing hypothesis tests with the wrong distribution will not make good inferences.

Multicollinearity

Multicollinearity, also known as **collinearity** or **aliasing**, is present when a predictor is close to being a linear combination of the other predictors. The term "perfect multicollinearity" refers to the violation of the 7th assumption in Section 3.3.1. In essence, we may fail to distinguish which predictors are truly meaningful to a model when a predictor is roughly similar to other predictors, whether it is in pairs or larger combinations.

This can lead to unstable estimates of the regression coefficients. This means a slight change in the dataset could result in vastly different $\hat{\beta}_j$'s. In other words, several dissimilar sets of $\hat{\beta}_j$ values will appear interchangeable with respect to minimizing the SSE. This is not desirable as the interpretations of the $\hat{\beta}_j$'s become less reliable.

The instability coincides with larger *se*'s. This affects inference on individual regression coefficients, such as making it harder to reject H_0 for *t* tests, even for coefficients whose explanatory variable is helpful in predicting the response.

Nevertheless, multicollinearity does not jeopardize the predictive power of \hat{y} , the reliability of MSE, and *F* test results.

Outliers and High Leverage Points

Outliers and high leverage points are observations that are unusual relative to the rest of the dataset. These observations are not bad by default, as they can provide useful insight. Note that an observation can be both an outlier and a high leverage point.

In linear regression, an **outlier** is an observation with an extreme residual. The standard for "extreme" is somewhat arbitrary; we will elaborate shortly. In any case, an extreme residual inflates the SSE and thus should be further investigated.

A **high leverage point** is an observation with an unusual set of predictor values. Keep in mind that an individual predictor value of a high leverage point could be typical for the predictor itself; the issue is whether the combination of values is strange. The $\hat{\beta}_j$'s are usually sensitive to these observations, thus making an investigation worthwhile to ensure reliable $\hat{\beta}_j$'s are obtained.

A related concept is an **influential point**: an observation that has a strong influence on model inferences. While there is no definitive way to measure such influence, an observation is likely to be influential if it is an outlier **and** a high leverage point.

High Dimensions

Linear regression is intended for datasets where n is much greater than p . Overfitting likely occurs with high-dimensional data, i.e. when p is too large. In particular, when $n \leq p + 1$, the fitted

equation will predict the training data responses perfectly. Said differently, there are no degrees of freedom associated with error, as all are "spent" to reduce SSE to 0. The model is certainly overfitting at that point – too flexible for the amount of data available.

With high dimensions leading to an unreasonably low SSE, many regression outputs should not be used, including $R^2_{\text{adj.}}$ and hypothesis test results. Moreover, multicollinearity is more likely to be present when p is large.

The issues in high dimensions can be summarized by the **curse of dimensionality**. A large n may contain a wealth of information in low dimensions, but the wealth tends to be sparse as p increases. In other words, the quantity of variables can dilute the quality of data; having more explanatory variables does not guarantee a better model.

As these ideas relate to flexibility and the bias-variance trade-off, other supervised learning approaches besides linear regression should also be wary of high-dimensional data.

Example 3.5.1.1

Determine which of the following issues can pose a challenge to interpreting the regression coefficient estimates in a multiple linear regression model.

- I. Misspecified model equation
- II. Multicollinearity
- III. High leverage points

Solution

A misspecified model equation does make interpreting the $\hat{\beta}_j$'s problematic. It becomes unlikely for the underlying parameters (β_j 's) to capture how the response and predictors are truly related, which renders the $\hat{\beta}_j$'s questionable.

Multicollinearity masks which predictors are actually meaningful to the model, potentially leading to unstable values of $\hat{\beta}_j$'s.

High leverage points tend to have a strong influence over the $\hat{\beta}_j$'s. Thus, we are less certain in the robustness of the $\hat{\beta}_j$'s.

Therefore, the answer is **I, II, and III.**



3.5.2 Leverage and Residuals

To better understand what makes an observation unusual or influential, let's discuss more concrete ideas in relation to leverage, residuals, and Cook's distance.

Leverage

The **leverage** of an observation measures its influence in predicting the response. We denote the i^{th} leverage as h_i , which is the i^{th} diagonal entry of the hat matrix, \mathbf{H} . Recall that the hat matrix was introduced in Section 3.2.3.

$$\mathbf{H} = \begin{bmatrix} h_1 & & & \\ & h_2 & & \\ & & \ddots & \\ & & & h_n \end{bmatrix}$$

In the case of simple linear regression,

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{u=1}^n (x_u - \bar{x})^2} \quad (3.5.2.1)$$

Observe from Equation 3.2.5.3 how h_i is used to calculate $se(\hat{y}_i)$.

Coach's Remarks

In Equation 3.5.2.1, u is also an index for the observations, similar to i in other formulas. Here, we let u iterate over the observations since i is fixed to denote a specific observation. If it helps to avoid confusion, you may choose to memorize that denominator as $(n - 1)s_x^2$ instead.

Here are other facts about leverage:

- It is a function of the x_j 's but not y .

- The larger h_i is, the more unusual the set of $x_{i,1}, \dots, x_{i,p}$ values is relative to other observations.
- $\sum_{i=1}^n h_i = p + 1$

There is nothing inherently bad about an observation with a high leverage. The concern is whether \hat{y} is mainly driven by a few observations that dominate a vast amount of the "total leverage" (i.e. $p + 1$), to the point that \hat{y} would alter drastically without these observations. One rule of thumb is that the i^{th} observation is a high leverage point if

$$h_i > 3 \left(\frac{p + 1}{n} \right)$$

Residuals

Recall that a residual is the difference between an actual and predicted response. This means a residual has the same unit as y . Consequently, the value of an extreme residual depends on the unit. For this reason, it is common to analyze a version of the residuals that is **unitless**. There are two types to discuss:

- Standardized residuals
- Studentized residuals

Standardized residuals are the residuals divided by an estimated standard error. The i^{th} standardized residual is

$$e_{\text{sta},i} = \frac{e_i}{\sqrt{\text{MSE}(1 - h_i)}} \quad (3.5.2.2)$$

Since the standard error is estimated, the standardized residuals are approximate realizations of the standard normal distribution, provided the model is correct.

Studentized residuals are similar to standardized residuals, except they are divided by a different estimated standard error. The i^{th} (externally) studentized residual is

$$e_{\text{stu},i} = \frac{e_i}{\sqrt{\text{MSE}_{(i)}(1 - h_i)}}$$

where $MSE_{(i)}$ is the MSE of the regression that excludes the i^{th} observation from the training data.

The studentized residuals are realizations of a t -distribution, provided the model is correct. Recall that a t -distribution converges to a standard normal as $\text{df} \rightarrow \infty$.

Therefore,

- roughly 95% of the population standardized/studentized residuals are between -2 and 2.
- roughly 99.7% of the population standardized/studentized residuals are between -3 and 3.

This allows us to choose a sensible cutoff in defining an outlier, regardless of the unit of the response. It is rather common to find the rule of thumb of exceeding 3 in absolute value used to identify outliers. Moreover, both types of residuals are often equally effective in determining outliers.

Cook's Distance

If we wish to combine leverage and residuals into a single measure, we can compute either **DFITS** or **Cook's distance** for each observation.

$$\text{DFITS}_i = e_{\text{sta}, i} \sqrt{\frac{h_i}{1 - h_i}} \quad (3.5.2.3)$$

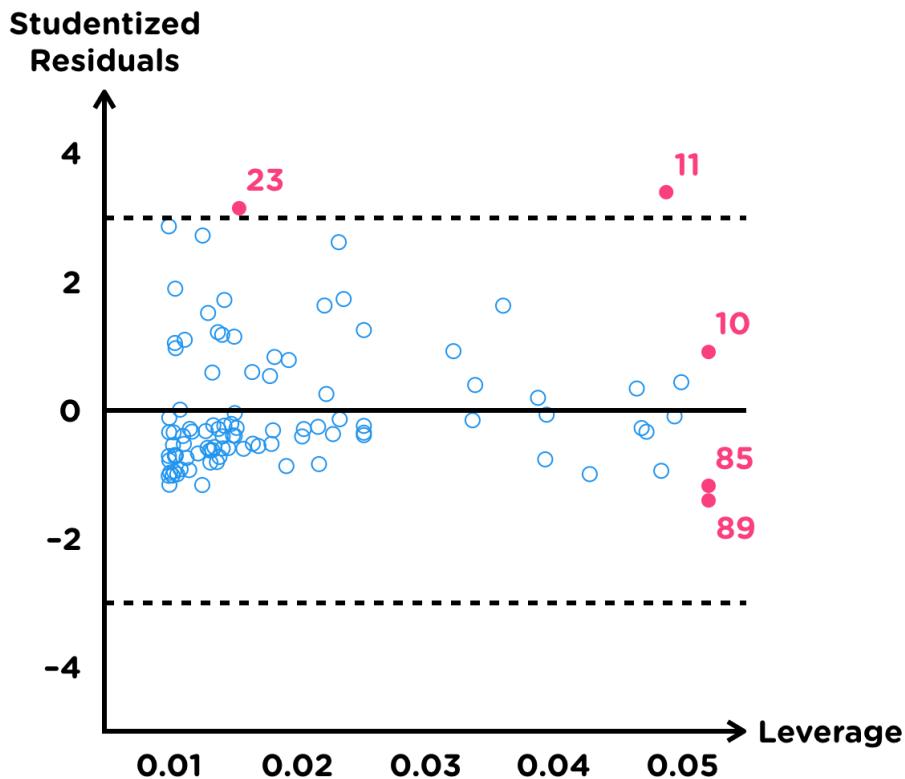
The i^{th} Cook's distance is

$$\begin{aligned} d_i &= \frac{\text{DFITS}_i^2}{p + 1} \\ &= \frac{e_{\text{sta}, i}^2 h_i}{(p + 1)(1 - h_i)} \end{aligned} \quad (3.5.2.4)$$

$$= \frac{e_i^2 h_i}{\text{MSE} (p + 1)(1 - h_i)^2} \quad (3.5.2.5)$$

One rule of thumb is that the i^{th} observation is an influential point if d_i exceeds **unity**, i.e. $d_i > 1$.

Let's plot the studentized residuals against the leverage values from the simple linear regression of Commute on Precip Chance from Section 3.2.2.



Assuming that an observation is

- an outlier if its studentized residual exceeds 3 in absolute value, and
- a high leverage point if its leverage exceeds $3 \left(\frac{2}{100} \right) = 0.06$,

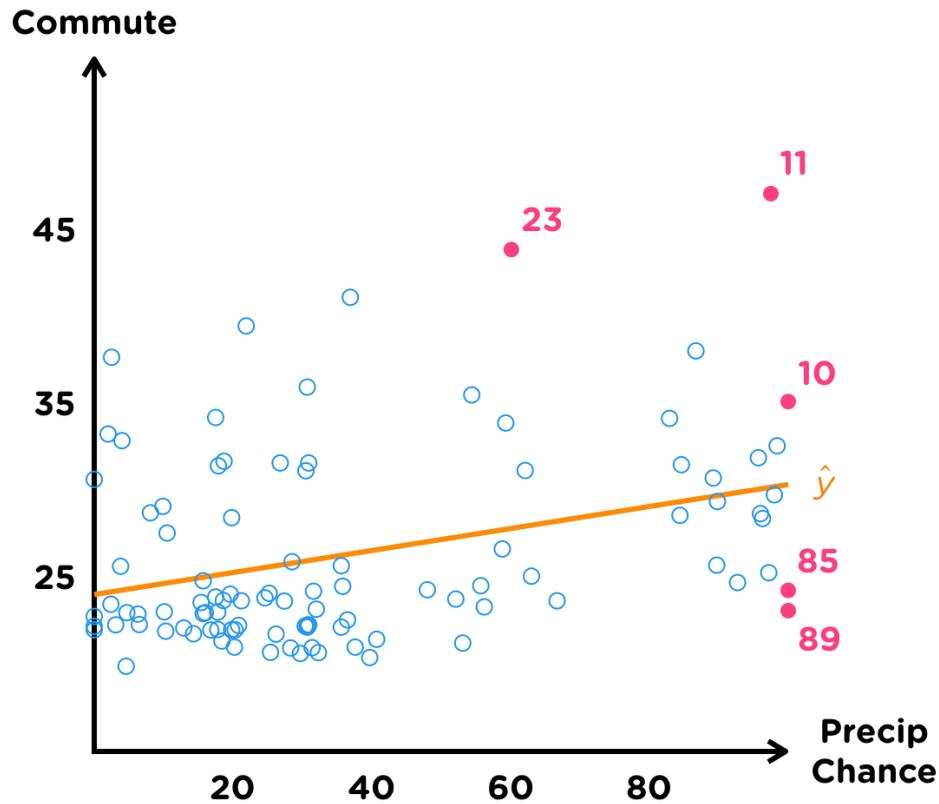
then observations 11 and 23 are outliers, whereas there are no high leverage points. The highest leverage belongs to observations 10, 85, and 89.

The table below summarizes the five observations of interest:

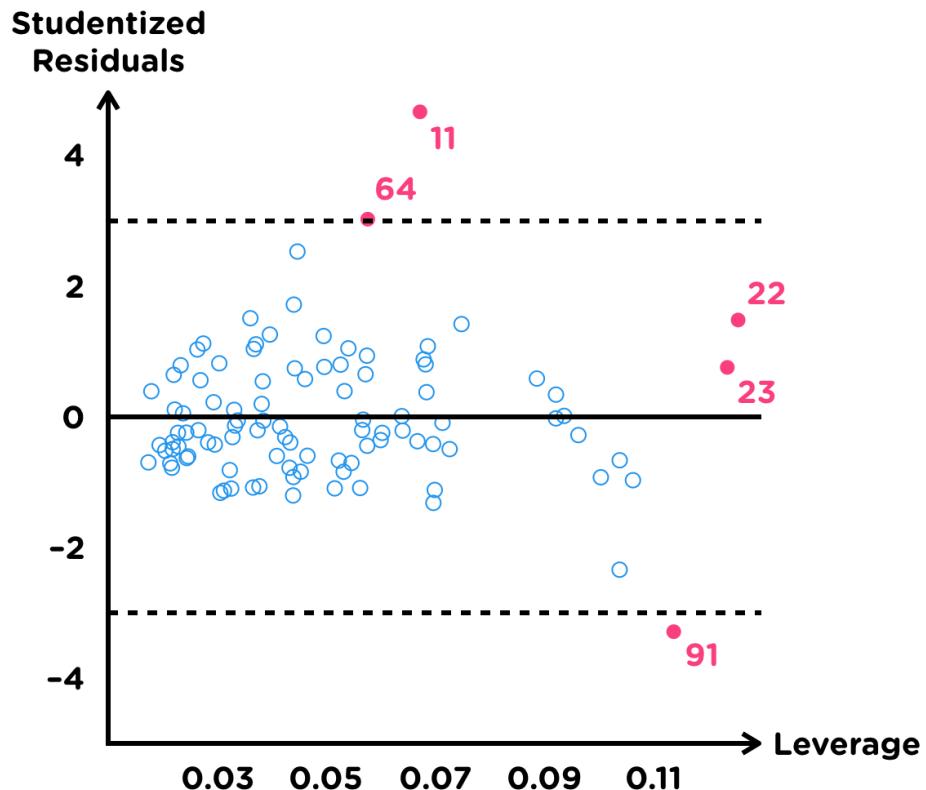
Observation	Precip Chance	Leverage	Studentized Residual
10	100.0	0.0521	0.9748
11	97.5	0.0488	3.4058
23	60.1	0.0155	3.1556
85	100.0	0.0521	-1.1707
89	100.0	0.0521	-1.3991

It should not be surprising that the leverage for observations 10, 85, and 89 are the same. This is because they all have the same predictor value of 100. It also makes intuitive sense that the largest possible Precip Chance value is considered unusual, hence producing the largest leverage.

Notice where these five observations are in the scatterplot between Commute and Precip Chance. It agrees with the plot of studentized residuals against leverage.



Let's also revisit the multiple linear regression with predictors Departure, Temp, Precip Chance, and Police from Section 3.3.2 by plotting its studentized residuals against leverage.



Assuming that an observation is

- an outlier if its studentized residual exceeds 3 in absolute value, and
- a high leverage point if its leverage exceeds $3 \left(\frac{5}{100} \right) = 0.15$,

then observations 11, 64, and 91 are outliers, but there are no high leverage points. The top three leverages belong to observations 22, 23, and 91.

The table below summarizes the five observations of interest:

Observation	Departure	Temp	Precip Chance	Police	Leverage	Studentized Residual
11	9.317	26.1	97.5	2	0.0673	4.6748
22	13.367	26.6	36.9	4	0.1257	1.4872
23	7.800	34.5	60.1	4	0.1237	0.7617
64	7.433	71.4	21.9	1	0.0577	3.0310
91	12.917	26.3	98.0	4	0.1139	-3.2855

It is more challenging to find patterns when working with several predictors. However, notice that the commute days with the top three leverages all record encountering four police vehicles on the commute route. This might be a hint as to why observations 22, 23, and 91 have somewhat unusual predictor values.

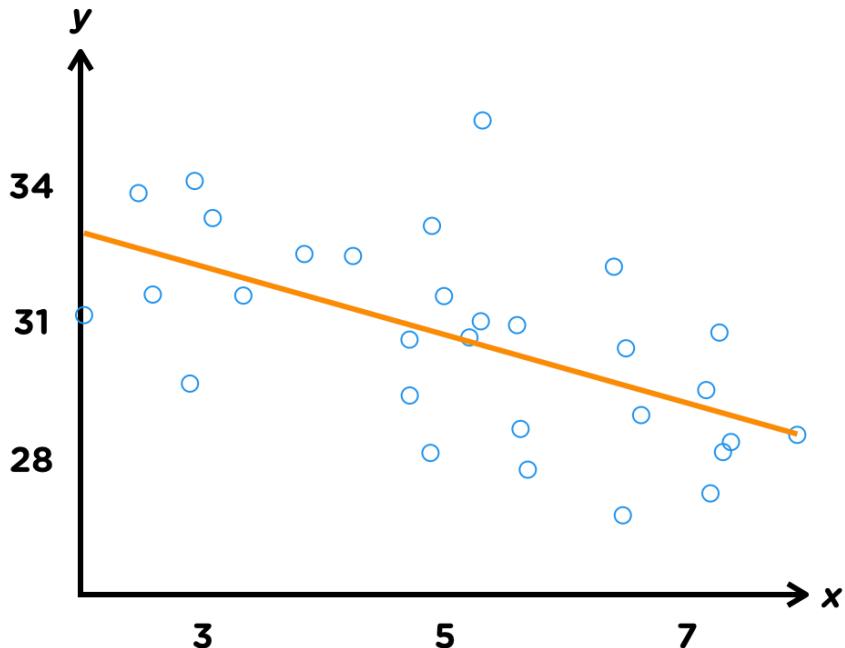
Other interesting points include:

- For the second model, observation 91 might be problematic, as it is an outlier and has a relatively high leverage.
- Observations 11 and 23 exhibit some concern in both models.

Discovering outliers and high leverage points is only the start of an investigation. Without further research, it is unclear whether any or all such observations should be excluded from the training data.

Example 3.5.2.1

You are given the scatterplot of a response variable y and an explanatory variable x with their line of best fit.



Determine which of the following best describes the data.

- There is nothing unusual about the data.
- The observation with the largest y is not a high leverage point but is an outlier.
- The observation with the largest y is not a high leverage point but is likely an outlier.
- The observation with the largest y is a high leverage point and an outlier.
- The observation with the largest y is a high leverage point and likely an outlier.

Solution

The observation with the largest y is unusual because it is apart from most of the data.

However, its x value is slightly greater than 5 which is not unusual for x . Observations that have typical predictor values have low leverage, meaning the unusual observation is not a high leverage point.

On the other hand, the unusual observation has a large residual. However, it is not possible to know the standardized or studentized residual of the observation based on the scatterplot. We are thus unable to determine if the large residual is extreme enough to be treated as an outlier.

Therefore, the answer is (C).



Example 3.5.2.2

You fit a linear regression with a model equation of

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \varepsilon_i \sim \text{Normal}(0, \sigma^2)$$

You are given the following values:

$$\mathbf{X} = \begin{bmatrix} 1 & 4 \\ 1 & 1 \\ 1 & 2 \\ 1 & 1 \end{bmatrix}; \quad \mathbf{y} = \begin{bmatrix} 8 \\ 6 \\ 8 \\ 5 \end{bmatrix}$$

$$\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T = \begin{bmatrix} \frac{11}{12} & -\frac{1}{12} & \frac{1}{4} & -\frac{1}{12} \\ -\frac{1}{12} & \frac{5}{12} & \frac{1}{4} & \frac{5}{12} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ -\frac{1}{12} & \frac{5}{12} & \frac{1}{4} & \frac{5}{12} \end{bmatrix}; \quad \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} = \begin{bmatrix} \frac{101}{12} \\ \frac{71}{12} \\ \frac{27}{4} \\ \frac{71}{12} \end{bmatrix}$$

Determine the number of observations that are influential based on a unity threshold for Cook's distance.

Solution

Use Equation 3.5.2.5 to calculate Cook's distance.

$$d_i = \frac{e_i^2 h_i}{\text{MSE} (p+1)(1-h_i)^2}$$

$p+1$ is the number of regression coefficients, which is the number of columns of \mathbf{X} . Thus, $p+1=2$.

Next, note that $\mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T$ is the hat matrix (Equation 3.2.3.2). As a result, its diagonal entries are the leverages.

$$h_1 = \frac{11}{12}, \quad h_2 = \frac{5}{12}, \quad h_3 = \frac{1}{4}, \quad h_4 = \frac{5}{12}$$

To solve for the residuals, recall that $\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$, producing

$$\begin{aligned} e_1 &= y_1 - \hat{y}_1 = 8 - \frac{101}{12} = -0.4167 \\ e_2 &= y_2 - \hat{y}_2 = 6 - \frac{71}{12} = 0.0833 \\ e_3 &= y_3 - \hat{y}_3 = 8 - \frac{27}{4} = 1.25 \\ e_4 &= y_4 - \hat{y}_4 = 5 - \frac{71}{12} = -0.9167 \end{aligned}$$

Now we can compute the MSE.

$$\begin{aligned}\text{MSE} &= \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n - p - 1} \\ &= \frac{(-0.4167)^2 + (0.0833)^2 + (1.25)^2 + (-0.9167)^2}{4 - 2} \\ &= 1.29167\end{aligned}$$

Finally, calculate Cook's distance for each observation.

$$\begin{aligned}d_1 &= \frac{(-0.4167)^2 \cdot \frac{11}{12}}{1.29167(2)\left(1 - \frac{11}{12}\right)^2} = 8.871 > 1 \\ d_2 &= \frac{(0.0833)^2 \cdot \frac{5}{12}}{1.29167(2)\left(1 - \frac{5}{12}\right)^2} = 0.003 < 1 \\ d_3 &= \frac{(1.25)^2 \cdot \frac{1}{4}}{1.29167(2)\left(1 - \frac{1}{4}\right)^2} = 0.269 < 1 \\ d_4 &= \frac{(-0.9167)^2 \cdot \frac{5}{12}}{1.29167(2)\left(1 - \frac{5}{12}\right)^2} = 0.398 < 1\end{aligned}$$

Therefore, only **one** observation is influential based on a unity threshold for Cook's distance.



3.5.3 Plots of Residuals

Running diagnostic tests on the residuals can assist in identifying potential issues and model violations. As mentioned, residuals are supposed to be the realizations of the error terms when the regression model is correct. Therefore, when the residuals display properties that differ from our expectations of the error terms, we should question the validity of those assumptions.

Here are three different plots of residuals to investigate:

1. e versus \hat{y}
2. e versus i
3. QQ plot of e

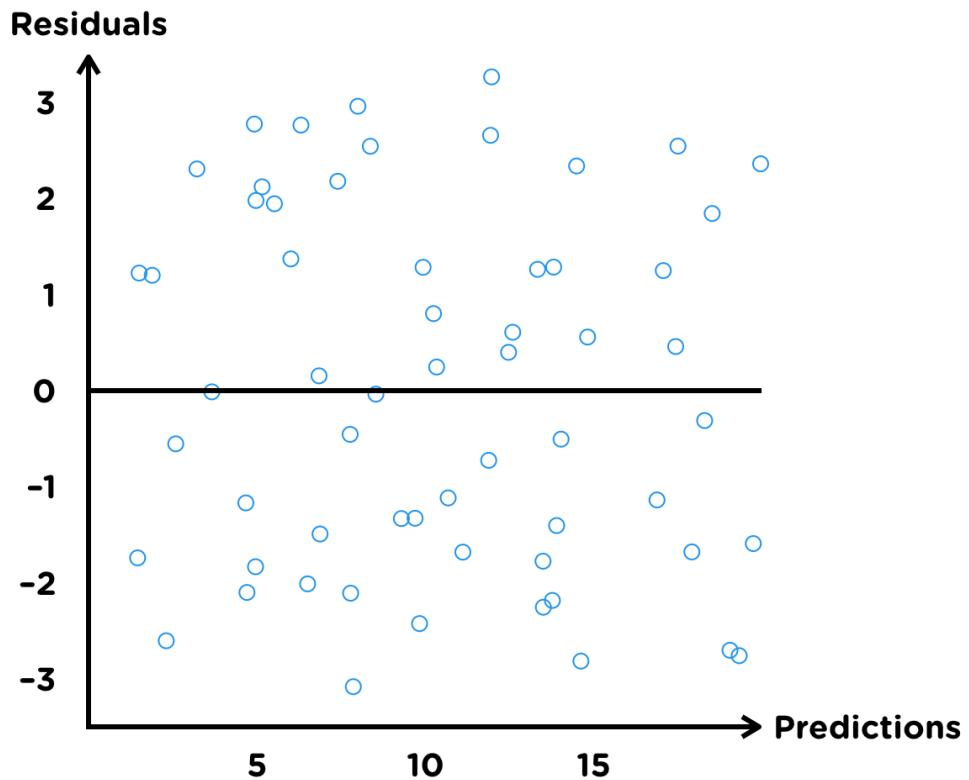
Coach's Remarks

All three plots focus on the overall pattern of the residuals rather than individual points. Therefore, it does not matter whether residuals, standardized residuals, or studentized residuals are used in these plots; an exam problem could use any.

In this subsection, we will only use residuals. Plots using a different type of residual will have the same overarching interpretations.

Plot Against Predictions

Here is a sample plot of well-behaved residuals that indicates no identifiable issues:

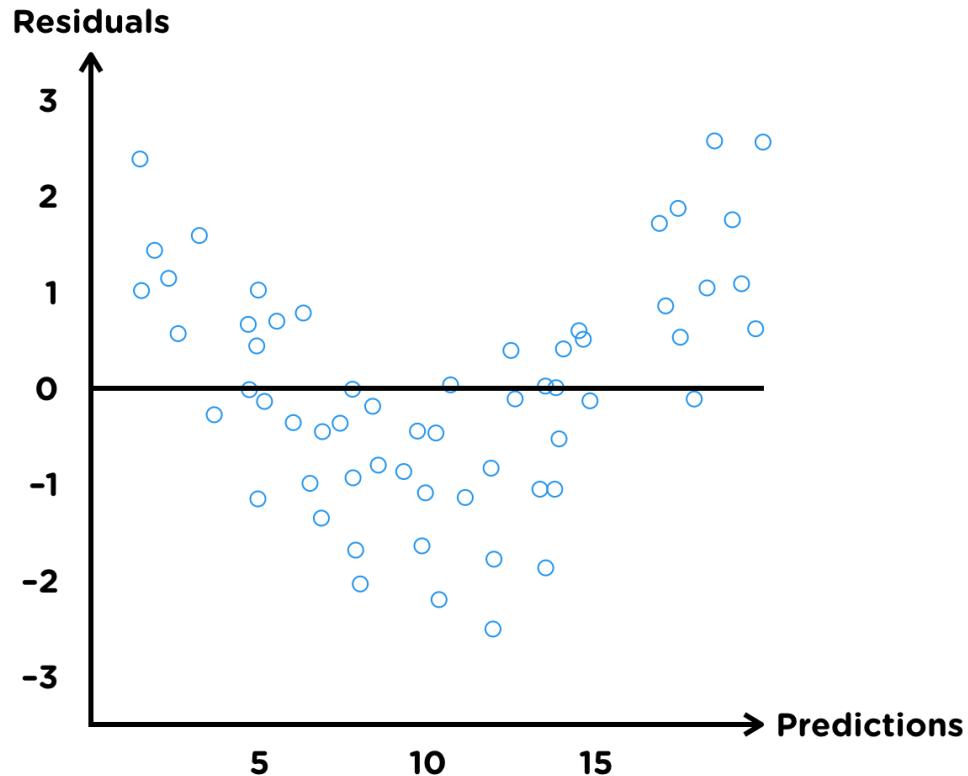


The residuals are well-behaved because

- the points appear to be randomly scattered and lacking any trend,
- at reasonable intervals of the predictions, the residuals seem to average to 0, and
- the spread of the residuals does not change throughout the plot.

DISCERNIBLE TRENDS

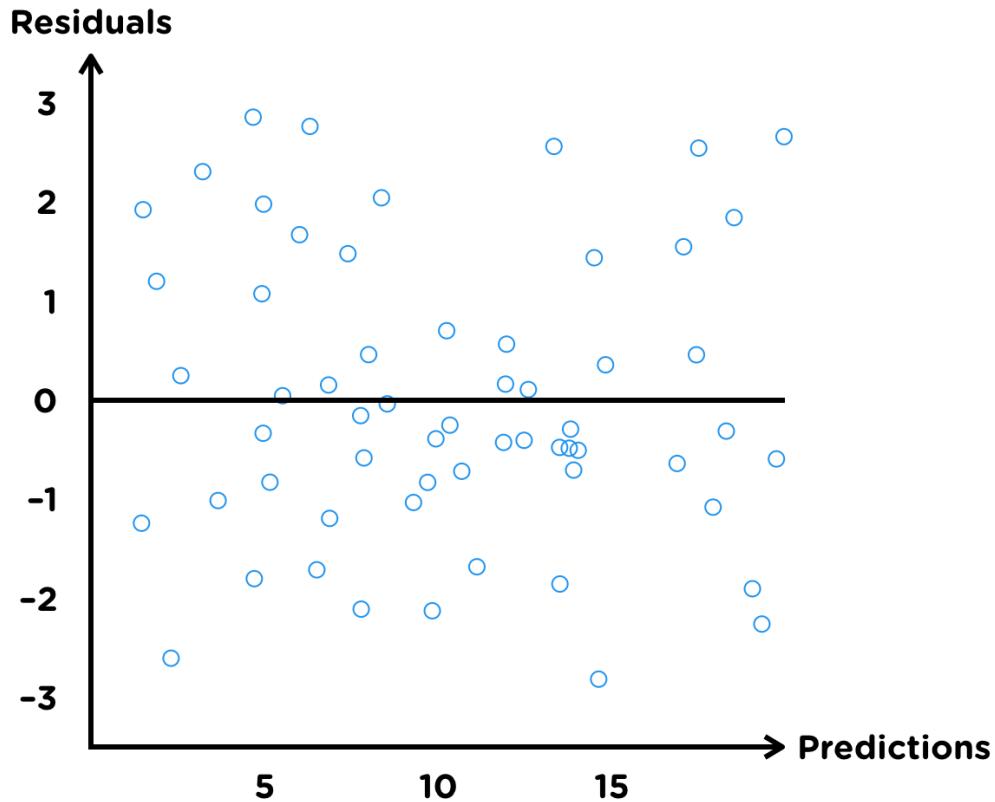
If the residuals exhibit a trend as a function of \hat{y} , then the model is likely missing a predictor that can explain the trend. A popular example of this is a u- or n-shaped pattern when a simple linear regression is used to fit data that have a quadratic relationship.



For predictions less than 5 or greater than 15, the residuals are almost always positive, i.e. the actual responses are almost always larger than the predictions. However, the predictions between 5 and 15 almost always have negative residuals. When the predictions are off in a systematic way similar to this, it is likely that the model equation is misspecified.

NON-ZERO AVERAGE

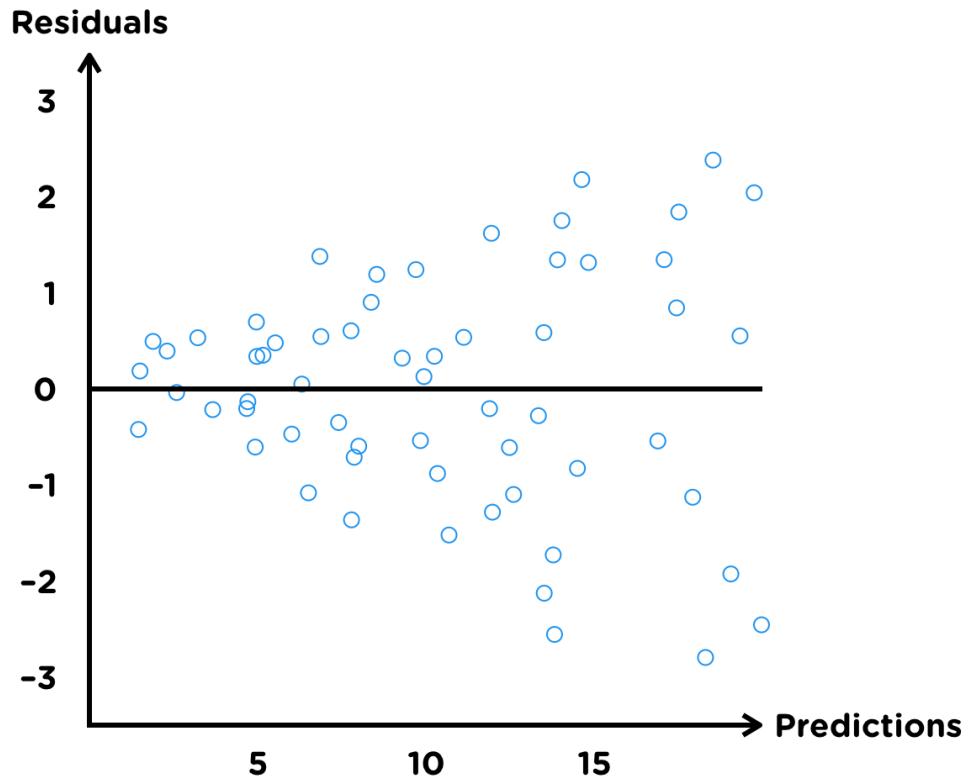
The assumption that the error term mean equals 0 is questionable if there is evidence of a non-zero average of residuals in some region of the plot. The following plot demonstrates this phenomenon:



For predictions between 8 and 15, there are more negative residuals than positive ones. Thus, the predictions in that interval tend to be larger than the actual response. This observation indicates that the average of residuals in this interval is non-zero.

HETROSCESTASTICITY

A multiple linear regression assumes that the error terms have the same variance for all observations. Therefore, residuals with an inconsistent spread throughout the plot is undesirable. A popular example of this is a funnel shape pattern:

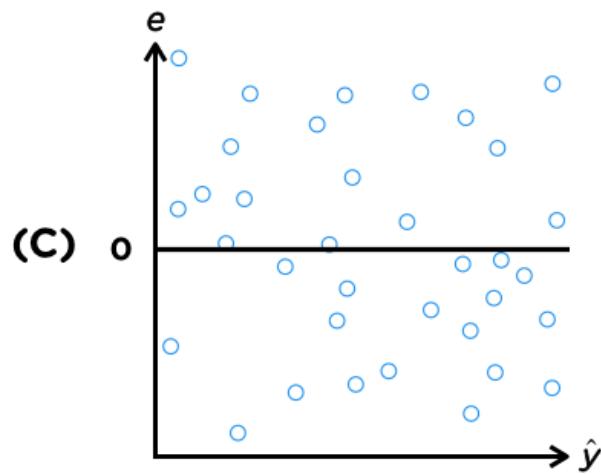
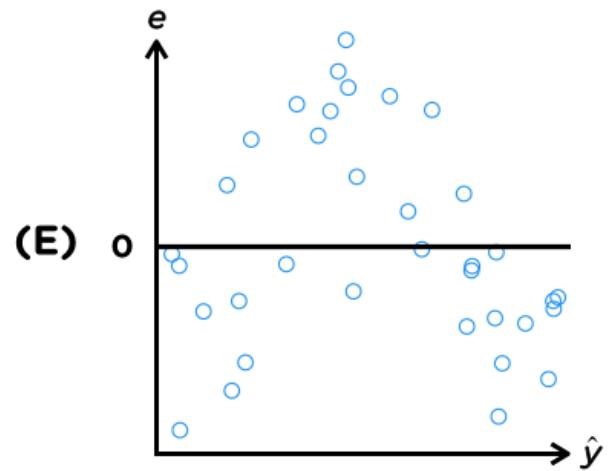
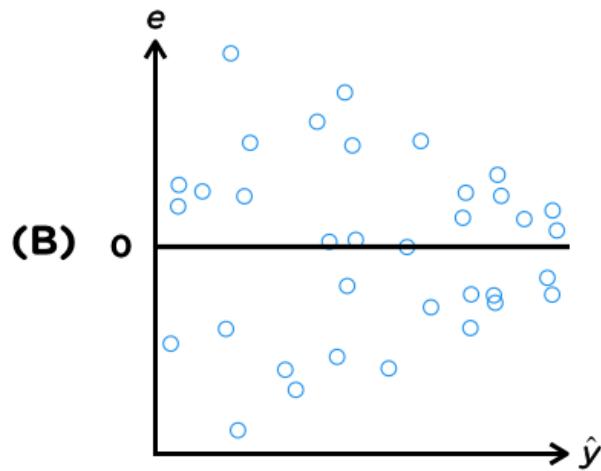
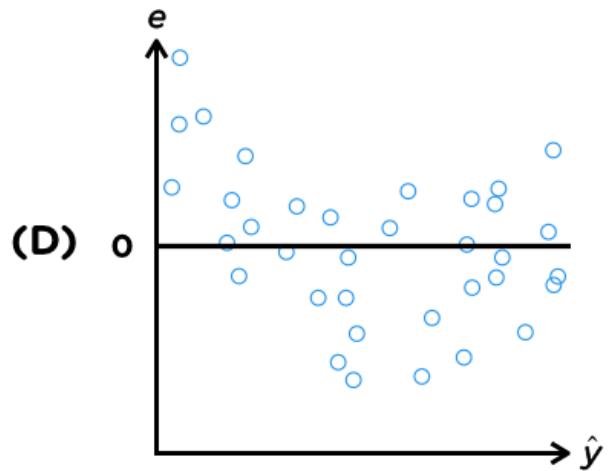
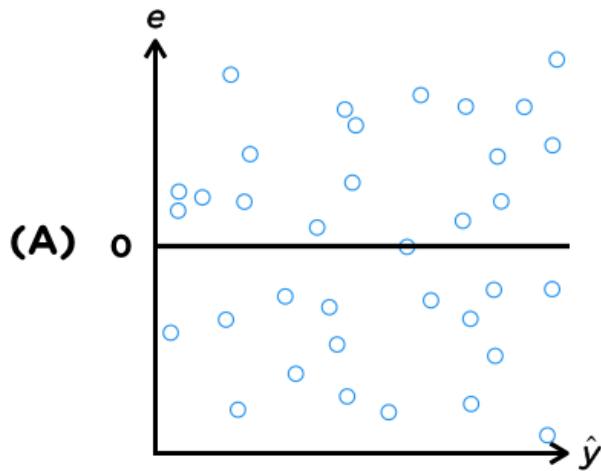


In this case, the spread of the residuals grows as the predictions are larger.

Example 3.5.3.1

In analyzing the residuals of a multiple linear regression, Jasmine noticed an issue: for small predictions, the residuals appear to average to a positive number, whereas for large predictions, they appear to average to a negative number. She also believes this is the only issue exhibited by the residuals.

Determine the residual plot that is most likely described by Jasmine.



Solution

Since the only issue with the residuals is the presence of non-zero averages, we can

eliminate option (B) which shows heteroscedasticity, as well as options (D) and (E) which have a distinct trend of a curve.

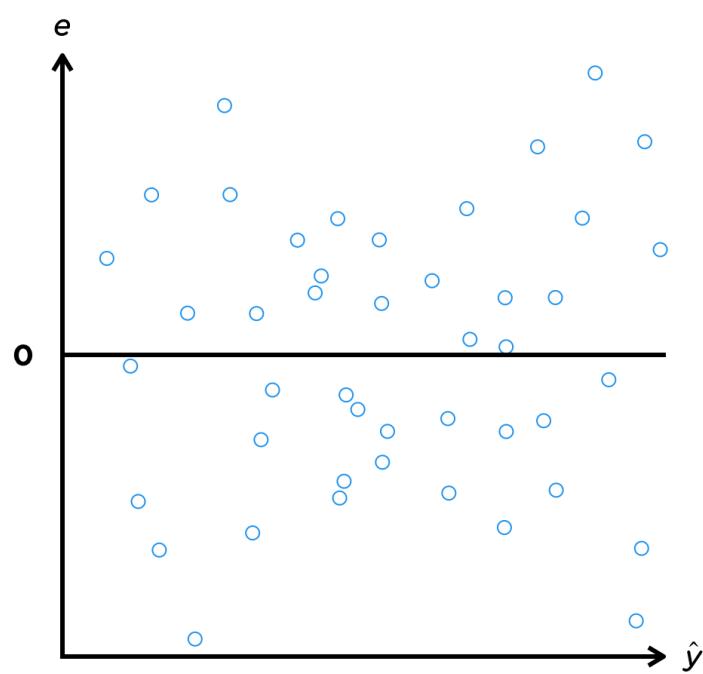
Option (A) has no issues in its plot.

For Option (C), the left region has more positive residuals than negative ones, and the right region has more negative residuals than positive ones. In addition, the residuals do not appear to form a trend, and they seem homoscedastic.

Therefore, the answer is (C). ■

Example 3.5.3.2

The following graph plots the residuals against the predicted values belonging to a multiple linear regression:



Based on the graph, determine which of the following model assumptions is likely to have been violated.

- I. Functional form of the mean response
- II. Error terms with mean 0
- III. Homoscedasticity

Solution

The graph does not show the residuals following a distinct trend, so it is not likely that the functional form of the mean response is inadequate.

At reasonable intervals of the predictions, the residuals appear to have an average of 0. It is unlikely that the regression has violated the assumption of error terms with mean 0.

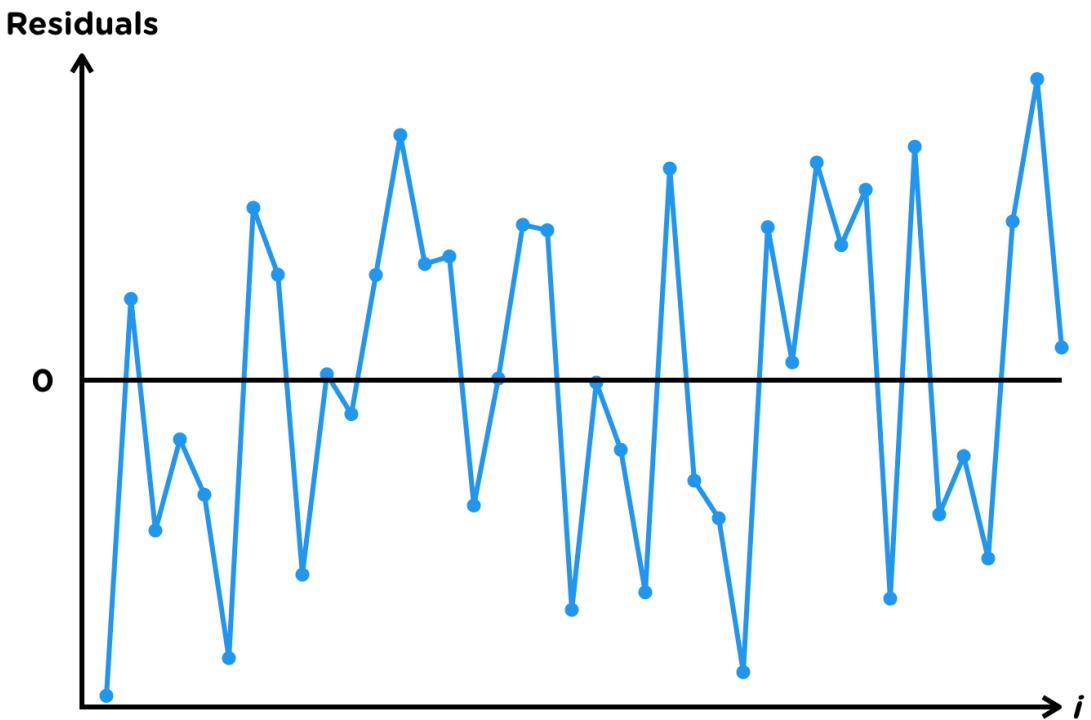
The residuals seem heteroscedastic, as the residual spread for both smaller and larger predictions appears wider compared to the spread for mid-sized predictions. This is a violation of homoscedasticity.

Therefore, the answer is **III only**.

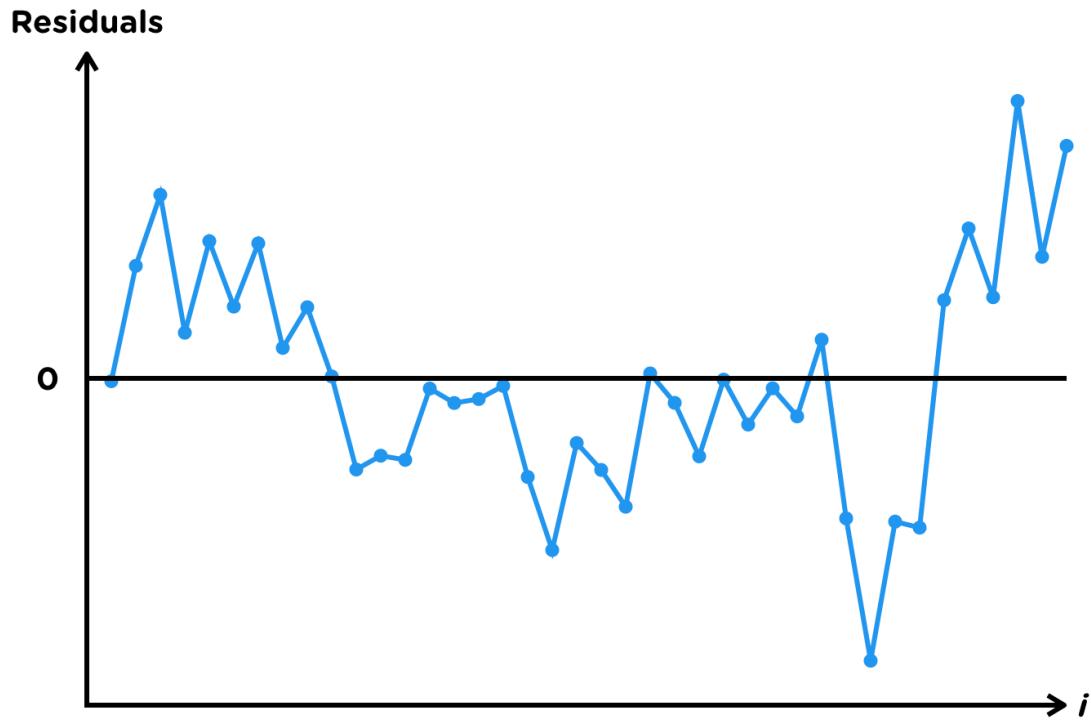


Plot Against Observation Index

If the error terms are independent, then we expect the residuals to be unpredictable from observation to observation. A sample plot of well-behaved residuals is:



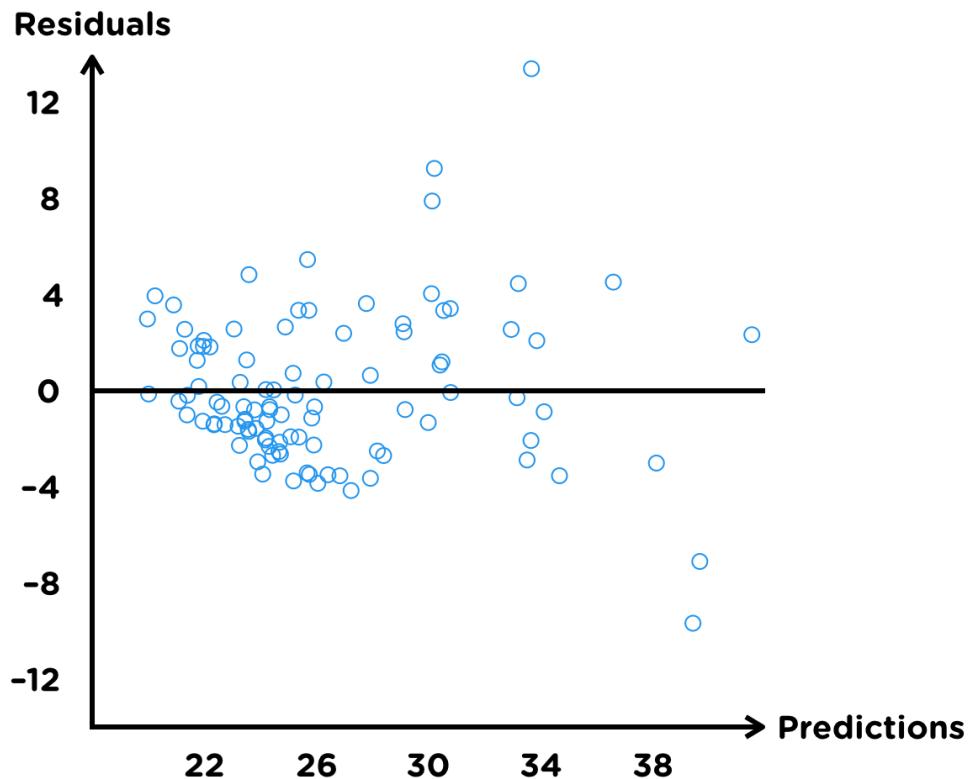
However, there are contexts where the error terms are likely dependent, especially for adjacent observations. In those situations, the residuals of adjacent observations will tend to be similar rather than unpredictable, such as in the following plot:



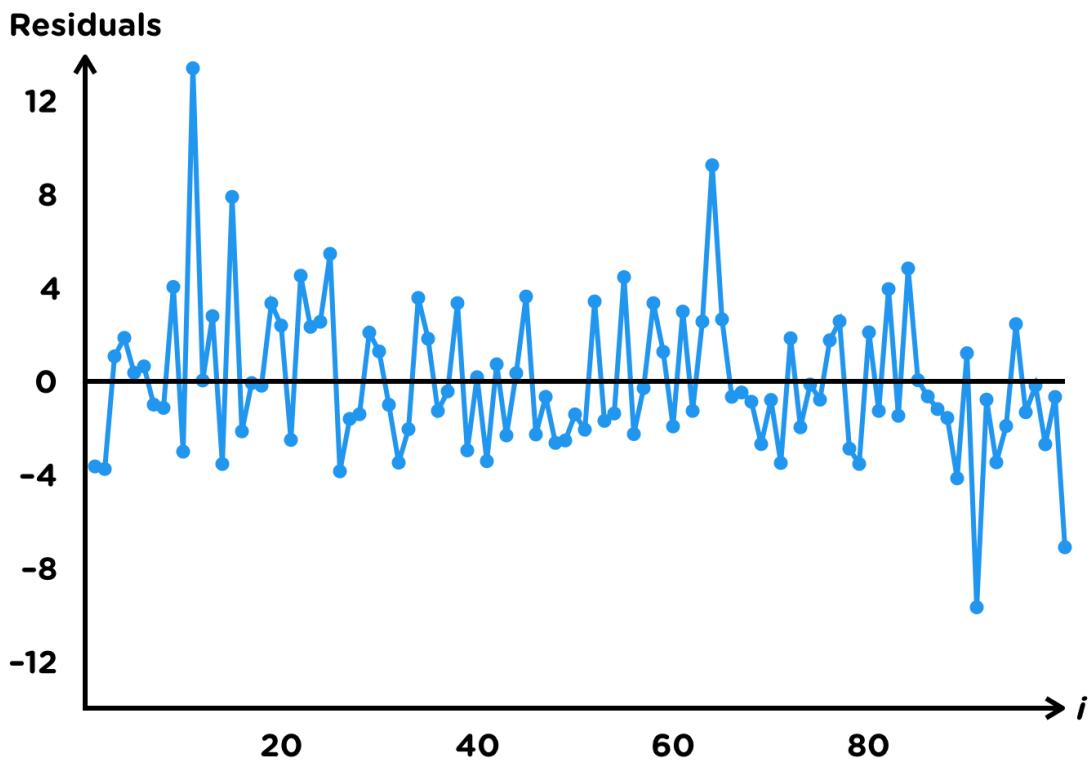
QQ Plot

Given that the model assumes normally distributed error terms, a QQ plot can check whether the distribution of the residuals resembles a normal distribution. The convention is to use the standard normal percentiles for the plot. Recall that a majority of points following the superimposed line signifies similar distribution shapes.

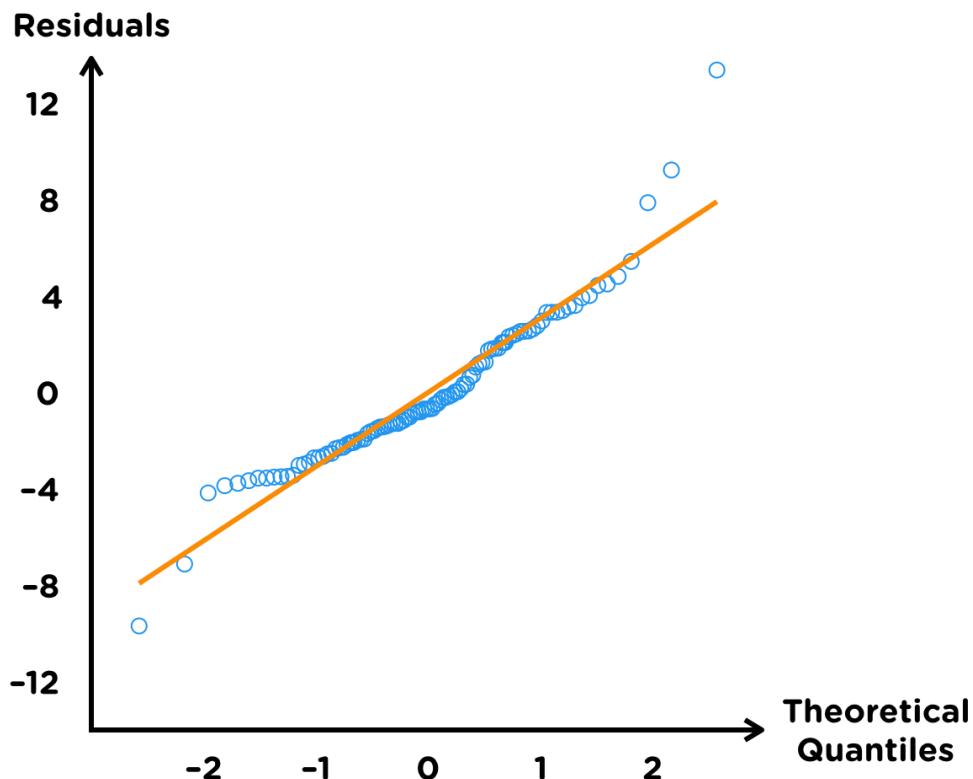
With the Commuting Chris scenario, let's consider these three plots for the multiple linear regression with predictors Departure, Temp, Precip Chance, and Police from Section 3.3.2. Remember to focus on the overall pattern of the plots, being careful to not be easily swayed by a few individual points.



There is no distinct trend formed by the residuals in this plot. It can be argued that the residuals are homoscedastic, especially when the three highly positive and the two highly negative residuals are ignored. Therefore, the biggest concern might be how the predictions under 23 minutes tend to produce large positive residuals, leading to a non-zero average in that region. While predictions between 28 and 32 minutes seem to have a similar issue, it is less convincing due to there being fewer observations available in that interval.



When the residuals are plotted against i , we do not see a predictable pattern or behavior. The assumption of independent error terms does not appear to be violated.



The QQ plot has many points aligned with the superimposed line. Only a few extreme residuals deviate significantly, which is expected since we found outliers in the studentized residuals analysis. As the distribution of the residuals is shaped similar to a standard normal, there is no sign of the normal errors assumption being violated.

3.5.4 Variance Inflation Factors

One way to measure multicollinearity is the **variance inflation factor (VIF)**. To calculate the VIF for the j^{th} predictor, first run a multiple linear regression with x_j acting as the response variable, predicted by the rest of the $p - 1$ predictors. Denote the coefficient of determination of that regression as R_j^2 . Then,

$$\text{VIF}_j = \frac{1}{1 - R_j^2} \quad (3.5.4.1)$$

Coach's Remarks

Intuitively, R_j^2 is itself a suitable measure for whether x_j is almost a linear combination of the other predictors. A reason to use VIF instead comes from how $se(\hat{\beta}_j)$ can be written as

$$\sqrt{\text{VIF}_j} \cdot \sqrt{\frac{\text{MSE}}{(n - 1)s_{x_j}^2}}$$

As a result, a large VIF leads to a large se , which could be problematic as discussed in Section 3.5.1.

Since $R_j^2 = 80\%$ is equivalent to $\text{VIF}_j = 5$, it is suggested that we should be concerned with multicollinearity if any of the p VIFs is 5 or greater.

The following are the VIFs for each predictor in the multiple linear regression from Section 3.3.2:

Predictors	VIF
Departure	1.082
Temp	2.805
Precip Chance	2.628
Police	1.269

Since all of the VIFs are less than 5, there is no obvious multicollinearity among the predictors. However, there can still be issues even when multicollinearity is not severe. The predictors Temp and Precip Chance – having the larger VIF values – might require further attention.

The table below lists the regression coefficient estimates for each predictor, as well as the sample correlation between each predictor and Commute.

Predictor	Coefficient Estimate	Correlation with Commute
Departure	-0.6329	-0.145
Temp	0.0558	-0.290
Precip Chance	0.0421	0.342
Police	4.0029	0.764

Notice that the signs of the coefficient estimate and the correlation are the same for each predictor except Temp. Even though the correlation detects a **negative** linear relationship between Temp and Commute, the regression **increases** the predicted commute time when temperature increases. This contradiction is perhaps a result of multicollinearity. One logical explanation is that Temp and Precip Chance are similar predictors, thus leading to unreliable estimates.

In practice, VIF is only one aspect to analyzing multicollinearity and its effect on the regression results.

Example 3.5.4.1

A linear regression model analyzes the variable y through predictors x_1 , x_2 , and x_3 . Based on ordinary least squares, you are given:

Regression	Response	Predictors	R^2
A	y	x_1, x_2, x_3	0.853
B	x_1	x_2, x_3	0.893
C	x_2	x_1, x_3	0.846
D	x_3	x_1, x_2	0.328

If any of the three predictors have a variance inflation factor that exceeds 8, then the model of interest is said to suffer from multicollinearity.

Determine which statement is true.

- I. The model suffers from multicollinearity.

- II. The model does not suffer from multicollinearity.
- III. It cannot be determined whether the model suffers from multicollinearity.

Solution

Notice we have the coefficient of determination for each predictor regressed on the other two predictors. Thus, we can calculate the VIF for each predictor, which makes statement III false.

Using Equation 3.5.4.1, the three VIFs are

$$\text{VIF}_1 = \frac{1}{1 - 0.893} = 9.346$$

$$\text{VIF}_2 = \frac{1}{1 - 0.846} = 6.494$$

$$\text{VIF}_3 = \frac{1}{1 - 0.328} = 1.488$$

Since $\text{VIF}_1 > 8$, the model of interest is said to suffer from multicollinearity. Therefore, **statement I is true**. Since neither x_1 nor x_2 explains x_3 well, it is likely that some similarity exists between x_1 and x_2 .



3.5.5 Potential Solutions

When there are signs of issues or violations, the best solution is usually not obvious. Additional research (e.g. trial-and-error) is commonly needed, especially if there appears to be

- a misspecified model equation,
- non-zero average of residuals,
- multicollinearity, or
- outliers, high leverage points, and/or influential points.

Among a variety of circumstances, there are a handful that merit some discussion.

Heteroscedasticity

The funnel shape of residuals with a larger spread for larger predictions is relatively common. One example is when studying a cost-related response variable that has a tendency to be more stable when small, but more volatile when large. To address this, one option is to transform the response variable using a **concave function**, such as by taking the natural logarithm or the square root of y . The idea is to shrink y 's larger values in hopes to obtain a more even volatility. So, the transformed y replaces the original y as the response in the linear model equation, e.g.

$$\ln Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

Dependent Errors

A typical context with dependent error terms involves time series data. A variable is called a time series when observations are recorded at fixed intervals of time (or space). A natural consequence is that the observations show patterns of dependency.

Non-Normal Errors

A violation of normality is when the response is discrete in nature. For example, the response may only take on the values 0 or 1, i.e. a binary variable. It is then more appropriate to use a model that incorporates a Bernoulli distribution rather than a normal distribution. Section 3.8 introduces these types of models.

Multicollinearity

After identifying predictors that contribute to multicollinearity, we may

- exclude all but one of those predictors from the model, or
- combine all of those predictors into one

to mitigate the issue. However, there are no best practices in executing either method. It may also be acceptable to make no changes and simply report its presence, especially if we only care about the several regression outputs that are not affected by multicollinearity.

ORTHOGONAL VARIABLES

Alternatively, we can use *orthogonal* predictors, which are also described as *uncorrelated* or *perpendicular*. A vector is orthogonal to another vector when their dot product equals 0. Thus, variables x_1 and x_2 are orthogonal when

$$\sum_{i=1}^n x_{i,1} \cdot x_{i,2} = 0$$

Collectively, orthogonal predictors have no multicollinearity whatsoever. To be more specific, first consider a predictor x_j that is orthogonal to a set of other predictors. Then, VIF_j will equal the minimum value of 1 (i.e. $R_j^2 = 0$). In addition, the estimated regression coefficients of the other predictors will be the same whether or not x_j is included in the model. Consequently, for a set of orthogonal predictors, all the VIFs equal 1, and their estimated coefficients remain the same even when other predictors are dropped.

While orthogonal predictors guarantee the absence of multicollinearity, it may not be feasible to record them. Hence, we may derive orthogonal predictors, such as principal components or partial least squares directions, from a dataset; these are detailed in Section 3.7. On the other hand, recording orthogonal predictors could be part of an experiment's design, which is commonly the case for ANOVA models.

Example 3.5.5.1

Determine which solution is appropriate when residuals exhibit a predictable pattern from

observation to observation.

- I. Transform the response with a concave function.
- II. Use a time series model.
- III. Remove observations that are high leverage points.

Solution

If residuals exhibit a predictable pattern from observation to observation, there is evidence of dependent error terms.

Transforming the response could resolve a specific heteroscedasticity pattern, but not dependent error terms.

A time series model attempts to account for dependent observations through dependent error terms.

High leverage points are observations with unusual sets of predictor values. They should be removed if they poorly represent the data. This has no clear link with dependent error terms.

Therefore, **statement II** is an appropriate solution when there is evidence of dependent error terms.



Example 3.5.5.2

Determine which statements are true regarding multicollinearity.

- I. It could be detected through variance inflation factors.
- II. It could be addressed by removing high leverage points.
- III. It is straightforward to resolve when detected.
- IV. It is absent in models that only use orthogonal variables to predict the response.

Solution

I is true because predictors with high VIFs indicate multicollinearity.

II is false because high leverage points are observations that have an unusual set of predictor values. This is not the same as predictors having a linear relationship among themselves. In other words, high leverage points pertain to the rows of the design matrix, whereas multicollinearity pertains to the columns.

III is false because there are no clear best practices in handling multicollinearity. Upon detection, more research is typically needed to determine an appropriate course of action.

IV is true because orthogonal variables are uncorrelated. Thus, they are unable to be written as or resemble a linear combination.

Therefore, **only I and IV are true.**



3.5 Summary

 5m

Violations and Issues

1. Misspecified model equation
2. Residuals with non-zero averages
3. Heteroscedasticity
4. Dependent errors
5. Non-normal errors
6. Multicollinearity
7. Outliers
8. High leverage points
9. High-dimensional issues

Negative Consequence	Potential Causes
Unreliable/unstable $\hat{\beta}$'s	<ul style="list-style-type: none"> • Misspecified model equation • Multicollinearity • High leverage points • High dimensions
Unreliable MSE	<ul style="list-style-type: none"> • Misspecified model equation • Heteroscedasticity • Outliers • High dimensions
Shrunken se 's (ignoring MSE)	<ul style="list-style-type: none"> • Dependent errors
Inflated se 's (ignoring MSE)	<ul style="list-style-type: none"> • Multicollinearity
Poor inferences (ignoring MSE and se)	<ul style="list-style-type: none"> • Misspecified model equation • Non-normal errors

Leverage

- h_i is the i^{th} diagonal entry of the hat matrix
- $\sum_{i=1}^n h_i = p + 1$

Standardized Residuals

$$e_{\text{sta}, i} = \frac{e_i}{\sqrt{\text{MSE}(1 - h_i)}}$$

DFITS and Cook's Distance

$$\text{DFITS}_i = e_{\text{sta}, i} \sqrt{\frac{h_i}{1 - h_i}}$$

$$d_i = \frac{\text{DFITS}_i^2}{p + 1} = \frac{e_i^2 h_i}{\text{MSE}(p + 1)(1 - h_i)^2}$$

Important Plots of Residuals

1. e versus \hat{y}
2. e versus i
3. QQ plot of e

Variance Inflation Factor

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

Other Key Ideas

- As realizations of an approximate standard normal distribution, standardized residuals can help identify outliers.
- As realizations of a t -distribution, studentized residuals can help identify outliers.
- When residuals have a larger spread for larger predictions, one solution is to transform the response variable with a concave function.
- Multicollinearity can be eliminated by using a set of orthogonal predictors.

3.6.0 Overview

5m

So far, we have introduced a few ways to compare models, specifically through R^2_{adj} , t tests, and F tests. This subsection covers model comparison more comprehensively in search of the "best" possible model. We typically desire a parsimonious model. **Parsimony** is the idea that a simpler model is preferred over a more complex model that does not substantially improve the simpler model.

In linear regression, a predictor is either included or excluded from the model equation, i.e. it has only two possibilities. So for g potential predictors, there are a total of 2^g models that can be constructed. Given that the number of possible models increases exponentially with g , it may be impractical to fit every model to find the best one. This could also lead to high-dimensional issues as discussed in Section 3.5.1. This subsection presents a few methods for finding an optimal model, while taking computational efficiency into account.

3.6.1 Subset Selection

An optimal linear model can be found by performing a subset selection procedure. Let's discuss:

1. Best subset

2. Stepwise

- Forward
- Backward
- Hybrids

To help explain these procedures, we let

- g be the total number of predictors in consideration
- p be the number of predictors for a specific model
- M_p be the "best" model among the $\binom{g}{p}$ models with p predictors

Coach's Remarks

It should be evident that there is only one possibility for M_0 and M_g :

- $M_0 \rightarrow$ null model
- $M_g \rightarrow$ model with all g predictors

Best Subset

The **best subset selection** has the following algorithm:

1. For $p = 0, 1, \dots, g$, fit all $\binom{g}{p}$ models with p predictors. The model with the largest R^2 is M_p .
2. Choose the best model among M_0, M_1, \dots, M_g using a selection criterion of choice, such as R^2_{adj} .

Recall that the largest R^2 is equivalent to the smallest SSE. In addition, note that R^2 can be used to determine the M_p 's because the comparisons are among models with the same value of p . By extension, it is **not** appropriate to use R^2 as the selection criterion in Step 2.

While the best subset selection considers all 2^g models, this method is computationally intensive for large values of g .

Stepwise

To avoid fitting and checking each model, we can employ **stepwise selection** in several different ways. The key feature of these procedures is that the optimal model is determined through an iteration (hence the name stepwise) of "best" models.

FORWARD SELECTION

Forward selection has the following algorithm:

1. Fit all g simple linear regression models (i.e. $p = 1$). The model with the largest R^2 is M_1 .
2. For $p = 2, \dots, g$, fit the models that add one of the remaining predictors to M_{p-1} . The model with the largest R^2 is M_p .
3. Choose the best model among M_0, M_1, \dots, M_g using a selection criterion of choice, such as $R^2_{\text{adj.}}$.

Forward selection is a greedy approach, in that it only adds the next best predictor as p increases rather than finding the best subset of predictors.

BACKWARD SELECTION

Backward selection has the following algorithm:

1. Fit the model with all g predictors, M_g .
2. For $p = g - 1, \dots, 1$, fit the models that drop one of the predictors from M_{p+1} . The model with the largest R^2 is M_p .
3. Choose the best model among M_0, M_1, \dots, M_g using a selection criterion of choice, such as $R^2_{\text{adj.}}$.

Instead of checking all 2^g models, both the forward and backward selection procedures only examine $1 + \frac{g(g+1)}{2}$ models. Moreover, these two algorithms force M_1, \dots, M_g to be nested. While streamlining the search helps with efficiency, there is no certainty that the absolute best model will be found.

When g is large, some of the models may exhibit high-dimensional issues. Specifically, models with p predictors where $n \leq p + 1$ do not have valid fits, as explained in Section 3.5.1. One important result is that only M_0, M_1, \dots, M_{n-2} would be unique and legitimate models. Even so, valid fits with $p \approx n - 2$ likely suffer from overfitting. Since backward selection starts with M_g , it should not (and sometimes, cannot) be performed in this situation. However, forward selection will still work; the algorithm would likely identify a legitimate, optimal model with $p \ll n - 2$.

HYBRIDS

The algorithms for forward and backward selection can be modified and/or combined to address some of their shortcomings. Consider the following stepwise regression algorithm as an example:

1. Start with the null model.
2. Add one of the remaining predictors to the current model. Select the predictor that, when added to the current model, produces the smallest p -value using two-tailed t tests. However, do not add the predictor if the p -value is above a specified threshold.
3. Using two-tailed t tests, remove the predictor with the largest p -value from the current model. However, do not remove the predictor if the p -value is below a specified threshold.
4. Repeat the previous two steps until the current model becomes stable, i.e. both steps result in no predictor being added or removed.

The main benefit of this algorithm over forward or backward selection is that a wider scope of models is considered. However, this introduces other sorts of issues, such as deciding the thresholds for adding and removing predictors.

In general, there is no perfect approach to determine the optimal model from g predictors. A critique against all subset selection procedures discussed here is the failure to account for any special knowledge that a researcher has, e.g.

- whether there are key predictors that should be in the model regardless
- how interaction terms should be handled to avoid violating the hierarchical principle

Let's work through the best subset selection for the Commuting Chris scenario. We consider all available predictors for Commute except Season. Since both Precip and Accident are categorical variables with two classes, they are each represented by one dummy variable. Therefore, $g = 6$. Let

- x_1 represent Departure
- x_2 represent Temp
- x_3 represent Precip Chance
- x_4 represent the dummy variable for Precip
- x_5 represent the dummy variable for Accident
- x_6 represent Police

When $p = 0$, there is only one model that uses none of the predictors: the null model. Thus, M_0 is the null model.

When $p = 1$, there are $\binom{6}{1} = 6$ models that use only one predictor. The table below sorts the models in descending order of R^2 .

Predictors	R^2
x_5	0.6748
x_6	0.5844
\vdots	\vdots
x_1	0.0210

Therefore, M_1 is the model with only the dummy variable for Accident as the predictor.

When $p = 2$, there are $\binom{6}{2} = 15$ models that use only two predictors. The table below sorts the models in descending order of R^2 .

Predictors	R^2
x_4, x_5	0.7866
x_3, x_5	0.7348
\vdots	\vdots
x_2, x_3	0.1181

Therefore, M_2 is the model with only the dummy variable for Precip and the dummy variable for Accident as the predictors.

This continues for $p = 3, \dots, 6$. The following table summarizes the seven M_p 's with their predictors and $R_{\text{adj.}}^2$:

M_p	Predictors	$R^2_{\text{adj.}}$
M_0	None	0.0000
M_1	x_5	0.6715
M_2	x_4, x_5	0.7822
M_3	x_1, x_4, x_5	0.8366
M_4	x_1, x_3, x_4, x_5	0.8487
M_5	x_1, x_2, x_3, x_4, x_5	0.8477
M_6	All	0.8461

If the selection criterion for the optimal model is $R^2_{\text{adj.}}$, then M_4 is the best. Moreover, it is interesting that the best subset selection produced nested M_p 's. When this occurs, we are guaranteed that the forward and backward selection procedures will produce the same M_p 's as the best subset selection.

To see why this is the case, let's briefly consider the forward selection algorithm. Realize that M_1 must be the same as it was for the best subset selection.

Then for $p = 2$, there are only $\binom{5}{1} = 5$ models to consider rather than 15. This is because forward selection only considers adding one of the remaining five predictors to M_1 , i.e. the model that already has x_5 . The table below sorts the models in descending order of R^2 .

Predictors	R^2
x_4, x_5	0.7866
x_3, x_5	0.7348
x_1, x_5	0.7062
x_2, x_5	0.7052
x_5, x_6	0.6906

Effectively, we are ignoring the 10 models from the best subset selection that exclude x_5 as a predictor. However, the $p = 2$ model with the largest R^2 is the same either way. Thus, we know in this case that M_1, \dots, M_6 under forward selection will match those under the best subset selection; the same can be said of backward selection for similar reasons.

Example 3.6.1.1

Five features are considered in a linear regression setting using ordinary least squares. For the following models, you are given:

Model	Features	RSS
A	x_1, x_2, x_3	378
B	x_1, x_2, x_4	371
C	x_1, x_3, x_4	367
D	x_2, x_3, x_4	382
E	x_2, x_3, x_5	375

Determine which statement is true.

- I. With forward selection, if the best model with two features has x_2 and x_3 , then we know Model D is the best with three features.
- II. With forward selection, if the best model with two features has x_1 and x_2 , then we know Model B is the best with three features.
- III. With backward selection, if the best model with four features has x_1, x_2, x_3 and x_4 , then we know Model C is the best with three features.
- IV. With backward selection, if the best model with four features has x_2, x_3, x_4 and x_5 , then we know Model E is the best with three features.

Solution

RSS refers to residual sum of squares, which we denote as SSE (**not** SSR) in this manual. To determine a "best" model with three features, we seek the smallest RSS since that is equivalent to the largest R^2 .

I is false. With forward selection, the best model with three features must add to the existing features of x_2 and x_3 . In this case, there are three possibilities: Models A, D, and E. Among them, Model E is the best because it has the lowest RSS.

II is false. With forward selection, the best model with three features must add to the existing features of x_1 and x_2 . In this case, there are three possibilities: Model A, Model B, and the model with features $\{x_1, x_2, x_5\}$. Since we do not have the RSS for all three models, we do not know which is the best.

III is true. With backward selection, the best model with three features must drop an existing feature among x_1, x_2, x_3 , and x_4 . In this case, there are four possibilities: Models A, B, C, and D. Among them, Model C is the best because it has the lowest RSS.

IV is false. With backward selection, the best model with three features must drop an existing feature among x_2, x_3, x_4 , and x_5 . In this case, there are four possibilities: Model D,

Model E, the model with features $\{x_2, x_4, x_5\}$, and the model with features $\{x_3, x_4, x_5\}$. Since we do not have the RSS for all four models, we do not know which is the best.

Therefore, **only III is true.**



Example 3.6.1.2

With k predictors that can be used in a linear regression, determine which statements are true in search of the best set of predictors.

- I. One advantage that backward selection has over the best subset selection is a higher computational efficiency.
- II. Forward selection can be implemented even when k is very large.
- III. When there are vastly more observations than predictors, the best model with k predictors is the same model for both forward and backward selection.
- IV. For the best subset selection, the best models of every subset size will be nested.

Solution

I is true because backward selection does not require checking as many models as the best subset selection in finding an optimal model.

II is true because forward selection is not prevented from finding the best models that have p predictors where $p \ll k$, of which the optimal model is likely among them. An optimal model is not likely to have many predictors since such models tend to overfit the data, some even to the extent of an invalid fit. In contrast, backward selection may not be feasible since it begins with finding M_k .

III is true because there is only one possible model with k predictors, i.e. the model that includes all available predictors. Hence, it is the best model with k predictors by default; the subset selection procedure of choice is irrelevant.

IV is false because the best subset selection considers all models at each subset size to find the best one. Its comprehensive search does not guarantee that the collection of best

models will be nested. In contrast, both forward and backward selection result in best models that are nested.

Therefore, **only I, II, and III are true.**



3.6.2 Selection Criteria

To find the optimal model among M_0, M_1, \dots, M_g (or in general, any collection of models), we should use an appropriate criterion that captures model quality. Besides $R_{\text{adj.}}^2$, we consider four other possibilities:

- Mallows' C_p
- Akaike information criterion (AIC)
- Bayesian information criterion (BIC)
- Cross-validation error

The first three options are appropriate because they are similar to the training MSE with a correction for flexibility, hence mimicking the test MSE. On the other hand, cross-validation error is a more direct estimate of the test MSE; we will address this option in the next subsection. Just as a lower test MSE indicates a better model, smaller values of these four quantities are preferred.

Coach's Remarks

The formulas for C_p , AIC, and BIC often have forms that differ across various resources. This is a minor detail because other forms usually come from dropping terms that are irrelevant when comparing the quantities.

The forms presented in the textbook by James et al. are covered here. In addition, note that formulas for AIC and BIC are given in the exam table, which is addressed in Section 3.8. Derivations for the formulas for AIC and BIC in this section using the ones in the exam table are included in the appendix at the end of the section.

Mallows' C_p

A formula for C_p is

$$C_p = \frac{1}{n}(\text{SSE} + 2p \cdot \text{MSE}_g) \quad (3.6.2.1)$$

where MSE_g is the MSE of the model that uses all g predictors. This formula is an unbiased estimate of the test MSE (Equation 3.1.3.1), which is attributed to MSE_g being an unbiased estimate of σ^2 .

To understand C_p intuitively, note that it can be written as the sum of two fractions, and that $\frac{\text{SSE}}{n}$ is the training MSE. Therefore, we can interpret C_p as the training MSE plus a penalty for each additional predictor. Recall that the training MSE decreases as p increases. So if the decrease in training MSE is not big enough to outweigh the penalty, then C_p will increase, suggesting that the model is worse than the one with a smaller p .

Mallows' C_p is also defined as

$$C'_p = \frac{\text{SSE}}{\text{MSE}_g} + 2p - n$$

This is equivalent to the version above where $C_p = \frac{1}{n} \text{MSE}_g (C'_p + n)$.

Coach's Remarks

It is unclear what the subscript p in C_p actually represents. It is probably either the number of predictors or regression coefficients. With this ambiguity, the subscript p in C_p is never evaluated; this quantity is always referred to as " C_p ".

AIC

In the context of multiple linear regression, a formula for AIC is

$$\text{AIC} = \frac{1}{n} (\text{SSE} + 2p \cdot \text{MSE}_g) \quad (3.6.2.2)$$

The best model by AIC will also be the best model by C_p .

BIC

In the context of multiple linear regression, a formula for BIC is

$$\text{BIC} = \frac{1}{n} (\text{SSE} + \ln n \cdot p \cdot \text{MSE}_g) \quad (3.6.2.3)$$

If the $\ln n$ is replaced with 2, then we will get the AIC formula. In other words, the penalty for each additional predictor is relative to the number of observations; the more observations that are available, the larger the penalty. Consequently, for $n \geq 8$, BIC favors models with a smaller p compared to AIC (as well as C_p).

Other miscellaneous ideas include:

- C_p , AIC, and BIC have theoretical justifications for being good measures of model quality, whereas $R^2_{\text{adj.}}$ does not have similar theoretical support.
- For overfitted models due to high dimensions, C_p , AIC, BIC, and $R^2_{\text{adj.}}$ are not reliable because they are functions of SSE.

Let's continue with the Commuting Chris scenario from the previous subsection. Using R, we can easily produce C_p and BIC to compare the models M_1, \dots, M_6 .

M_p	Predictors	C_p	BIC
M_1	x_5	113.15	-103.12
M_2	x_4, x_5	43.27	-140.63
M_3	x_1, x_4, x_5	9.88	-165.84
M_4	x_1, x_3, x_4, x_5	3.37	-169.95
M_5	x_1, x_2, x_3, x_4, x_5	5.03	-165.72
M_6	All	7.00	-161.14

The smallest C_p is 3.37, while the smallest BIC is -169.95. Thus, both C_p and BIC select M_4 as the best model; this is the same model selected using $R^2_{\text{adj.}}$. Even though BIC favors models with smaller p compared to C_p when there are at least 8 observations, both measures end up selecting the same model in this scenario. This is because adding x_3 (Precip Chance) to the model explained a substantial amount of variability, so much so the harsher penalty made no difference.

As a side note, these C_p and BIC values from R are computed using different formulas from the ones presented here.

Example 3.6.2.1

Three features are considered in a linear regression setting using ordinary least squares. You are given:

Model	Features	R^2	C_p
A	None	0.000	60.313
B	x_1	0.576	29.105
C	x_2	0.503	33.426
D	x_3	0.488	34.315
E	x_1, x_2	0.788	18.557
F	x_1, x_3	0.734	21.753
G	x_2, x_3	0.795	18.142
H	x_1, x_2, x_3	0.814	19.019

Let M_p denote the best model among those that have p features.

Determine which best describes the summary output.

- A. Using forward selection, Model G is M_2 .
- B. Using backward selection, there is not enough information to determine which model is M_1 .
- C. Using forward selection and AIC, Model E is the best.
- D. Using backward selection and C_p , Model C is the best.
- E. Using the best subset selection and C_p , there is not enough information to determine the best model.

Solution

By default, M_3 is Model H. Subject to the selection procedure, M_1 and M_2 are models with the highest R^2 .

Using the best subset selection, note that M_1 is Model B and M_2 is Model G. Among the four M_p 's, Model G has the smallest C_p and hence is the best model via best subset. Therefore, option (E) is incorrect.

Using backward selection, note that M_2 is Model G. This means there are only two candidates for M_1 : Models C and D. As a result, M_1 is Model C, so option (B) is incorrect. Among the four M_p 's, Model G has the smallest C_p and hence is the best model via backward selection. Therefore, option (D) is also incorrect.

Using forward selection, note that M_1 is Model B. This means there are only two candidates for M_2 : Models E and F. As a result, M_2 is Model E. Therefore, option (A) is incorrect. In addition, recall that C_p and AIC are minimized by the same model. Among the four M_p 's, Model E has the smallest C_p and hence is the best model via forward selection.

Therefore, the answer is (C). ■

Example 3.6.2.2

Given the same dataset, two actuaries are tasked with building a linear regression model using an intercept $\{I\}$ and four potential predictors $\{1, 2, 3, 4\}$.

- Actuary Ava chooses the best model based on forward stepwise selection
- Actuary Bello chooses the best model based on backward stepwise selection

Below are summaries for all candidate models:

Parameters	SSE	AIC
I	102.2	372.1
I, 1	100.8	372.0
I, 2	33.9	208.2
I, 3	24.5	160.1
I, 4	49.6	265.6
I, 1, 2	29.9	191.9
I, 1, 3	16.3	101.1
I, 1, 4	33.6	209.4
I, 2, 3	23.9	158.1
I, 2, 4	33.8	210.1
I, 3, 4	19.3	126.3
I, 1, 2, 3	14.4	84.7
I, 1, 2, 4	14.3	82.7
I, 1, 3, 4	15.8	98.3
I, 2, 3, 4	18.6	122.9
I, 1, 2, 3, 4	14.1	83.1

Calculate the absolute difference in AIC between Actuary Ava's chosen model and Actuary Bello's chosen model.

Solution

When selecting the "best" model among those with the same number of predictors, we choose the highest R^2 , or equivalently, the lowest SSE.

For forward selection, the following lists the "best" model at every number of predictors:

- $\{\}$
- $\{l, 3\}$
- $\{l, 1, 3\}$
- $\{l, 1, 2, 3\}$
- $\{l, 1, 2, 3, 4\}$

Among these five models, $\{l, 1, 2, 3, 4\}$ has the smallest AIC of 83.1, making it Actuary Ava's chosen model.

For backward selection, the following lists the "best" model at every number of predictors:

- $\{l, 1, 2, 3, 4\}$
- $\{l, 1, 2, 4\}$
- $\{l, 1, 2\}$
- $\{l, 2\}$
- $\{\}$

Among these five models, $\{l, 1, 2, 4\}$ has the smallest AIC of 82.7, making it Actuary Bello's chosen model.

As a result, the absolute difference in AIC between the chosen models is $83.1 - 82.7 = \mathbf{0.4}$.

3.6.3 Cross-Validation

In Section 3.1.3, we introduced the test MSE as a measure of model quality and how to estimate it using test data rather than training data. Without access to test data, estimating the test MSE can be done through cross-validation.

When cross-validation is used for model selection, the goal may be to

- directly select from a handful of models, or
- determine the proper amount of flexibility.

In the context of subset selection, the latter means cross-validation seeks the ideal p first, which then points to the optimal model.

There are three approaches of interest:

1. Validation set
2. k -fold cross-validation
3. Leave-one-out cross-validation (LOOCV)

Validation Set

The **validation set approach** randomly splits all available observations into two groups: the training set and the validation set. Let

- n_1 be the number of observations in the training set, and
- n_2 be the number of observations in the validation set.

After attaining the fitted equation \hat{y} using **only** the n_1 observations in the training set, we estimate the test MSE by using Equation 3.1.3.2 with **only** the n_2 observations in the validation set. This estimate is called the **validation set error**.

In the context of subset selection, we use the following algorithm:

1. Perform a procedure of choice (e.g. best subset selection) using only the n_1 training observations to fit the models.
2. Determine the p that corresponds to the best model among M_0, M_1, \dots, M_g , i.e. the one with the lowest validation set error.

3. Rerun the procedure of choice using **all** observations. The optimal model is the **new** M_p using the p noted in the previous step.

In the algorithm, note that the validation set errors are only used to discover the optimal p . Upon knowing the desirable flexibility level, we then determine the optimal model using all available observations in order to better estimate the regression coefficients. In addition, realize that the chosen predictors in the first step may be different from the chosen predictors in the third step.

The validation set approach has two issues worth discussing:

1. **The results are fickle** – because the observations are randomly split into two sets, the results can vary greatly when the approach is repeated.
2. **The validation set error tends to overestimate the test MSE** – using only n_1 observations (instead of all of them) to train a model will likely produce poorer fits, hence leading to higher estimates of the test MSE.

Let's revisit the Commuting Chris scenario in seeking the best subset among the $g = 6$ predictors with validation set error as the selection criterion. We begin by randomly dividing the 100 observations into the training and validation sets. For this demonstration, we chose $n_1 = 70$ and $n_2 = 30$.

Next, we fit the models with the same process detailed in Section 3.6.1, using only the 70 training observations. Recall that there are 6 models that use only one predictor. The table below sorts the models in descending order of R^2 .

Predictors	R^2
x_6	0.6887
x_5	0.6432
:	:
x_1	0.0198

In this case, M_1 is the model with only Police as the predictor.

The procedure continues as usual until all the M_p 's are identified. Then, using each \hat{y} from the M_p 's, we compute the validation set errors using the 30 observations in the validation set. The following table summarizes the seven M_p 's with their predictors and validation test error:

	Predictors	Validation Set Error
M_0	None	43.462
M_1	x_6	24.557
M_2	x_4, x_5	8.844
M_3	x_1, x_4, x_5	7.105

M_p	Predictors	Validation Set Error
M_4	x_1, x_4, x_5, x_6	10.374
M_5	x_1, x_3, x_4, x_5, x_6	9.685
M_6	All	9.756

Since the lowest validation set error belongs to M_3 , the proper amount of flexibility for the optimal model is to have 3 predictors. However, the optimal model is not M_3 from the table above, which was found using only the 70 training observations. The optimal model is actually the M_3 from the best subset selection performed on all 100 observations, i.e. the M_3 from Section 3.6.1. We expect the $\hat{\beta}_j$'s to be more accurate when using 100 observations compared to using only 70 observations.

In this case, whether training on the 70 or 100 observations, M_3 consists of the same three predictors. As a reminder, this is a coincidence; the chosen predictors could have been different.

k-Fold Cross-Validation

The ***k-fold cross-validation approach*** starts by randomly dividing all available observations into k groups called **folds** of roughly equal size. A model is fitted k times in total, following this procedure:

1. For $v = 1, \dots, k$, obtain the v^{th} fit by training with all observations **except** those in the v^{th} fold.
2. For $v = 1, \dots, k$, use \hat{y} from the v^{th} fit to calculate a test MSE estimate using Equation 3.1.3.2 with the observations in the v^{th} fold.
3. Average the k test MSE estimates in the previous step.

The average calculated in the third step is the test MSE estimate via k -fold cross-validation, which we denote as **CV error**. To visualize this procedure with 30 observations divided into 10 folds (i.e. 3 observations per fold), consider the animation below:

Observation, i
1
2
3
4
5
6
:
28
29
30

Each of the 30 observations is randomly assigned to one fold denoted by a specific color. In training the first model, we use all observations except those in the first fold (blue solid border). Then, we calculate a test MSE estimate using the three observations in the first fold (blue dashed border) based on the first model fit. We do this a total of 10 times, once for each fold. Averaging the 10 test MSE estimates produces the CV error.

Since a model is fitted k times, realize that each observation will be used for training $k - 1$ times; the one time it is not occurs when the observation is in the fold used for estimating the test MSE.

In the context of subset selection, the procedure of choice intertwines with cross-validation. Here is one way to describe the full algorithm:

1. Following the sequence of p based on the procedure of choice (i.e. $p = 0, \dots, g$ for best subset and forward selection; $p = g, \dots, 0$ for backward selection):
 - (a) For $v = 1, \dots, k$, determine M_p based on the procedure of choice by fitting models on all observations except those in the v^{th} fold.
 - (b) For $v = 1, \dots, k$, use M_p 's fitted equation from the v^{th} fit to calculate a test MSE estimate using the v^{th} fold.
 - (c) Compute the CV error associated with p by averaging the k test MSE estimates in the previous step.
2. Determine the p that has the lowest CV error.

3. Rerun the procedure of choice using all observations. The optimal model is the new M_p using the p noted in the previous step.

Let's see the algorithm in action. In using 5-fold cross-validation to find the best subset among the 6 predictors in the Commuting Chris scenario, the 100 observations are randomly split into folds of 20 at the beginning. Next, we focus on calculating the CV error for $p = 1$.

The first fold is excluded, and then M_1 is determined through the best subset selection. Then, the fitted equation of this M_1 is used to compute a test MSE estimate using the observations in the first fold, which equals 6.408.

Then, the second fold is excluded as another M_1 is determined through the best subset selection. The test MSE estimate from the second fold based on this other M_1 is 12.922.

The process resumes with the third, fourth, and fifth folds, resulting in the values of 11.723, 15.948, and 8.493 for each respective fold's test MSE estimate. Therefore, the CV error for $p = 1$ is the average of the five estimates from the folds, which equals 11.099.

At this point, the best subset procedure continues for the next value of p , and so on. The following table organizes the seven CV errors by their flexibility level:

p	CV Error
0	32.354
1	11.099
2	7.384
3	5.491
4	5.946
5	6.519
6	6.484

Since the lowest CV error belongs to $p = 3$, the optimal model according to best subset selection is M_3 using all 100 observations, i.e. the M_3 from Section 3.6.1. In our examples, both the k -fold cross-validation and validation set approaches agree on the appropriate level of flexibility.

Consider how k -fold cross-validation addresses the two issues faced by the validation set approach:

1. The results are typically less fickle because the CV error is an average that involves **all** observations.
2. As every fit uses a majority of the observations, they exhibit less bias and thus should not overestimate the test MSE as much.

LOOCV

The **leave-one-out cross-validation (LOOCV) approach** is the same as k -fold cross-validation when $k = n$, i.e. the number of folds equals the number of available observations. This means each fold has only one observation, and thus there is no random assignment of observations to folds. Consequently, LOOCV is computationally intensive for a large n , as computing the CV error for a single model would require performing n fits. It follows that each observation will be used for training $n - 1$ times. Realize that each fold (i.e. observation) produces a squared residual as its "test MSE estimate"; the LOOCV error is the average of these n squared residuals.

On the other hand, ordinary least squares estimation simplifies the LOOCV error calculation, such that only the fit using all observations is needed. The formula is similar to the training MSE, except the i^{th} residual is first divided by 1 minus the i^{th} leverage, i.e.

$$\text{LOOCV error} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2 \quad (3.6.3.1)$$

Relative to the issues with the validation set approach, LOOCV results are unchanging because there is no randomization of the observations. In addition, every fit only leaves out one observation. Hence, it is more difficult to overestimate the test MSE.

BIAS-VARIANCE TRADE-OFF

The validation set error is likely to have substantial bias as an estimator of the test MSE due to typically having fewer observations for training. As alluded to, this makes overestimating the test MSE more likely. The k -fold CV error is based on fits with typically more training observations, resulting in a lower bias. By extension, the LOOCV error has the least bias among the three for having fits that exclude only one observation from training.

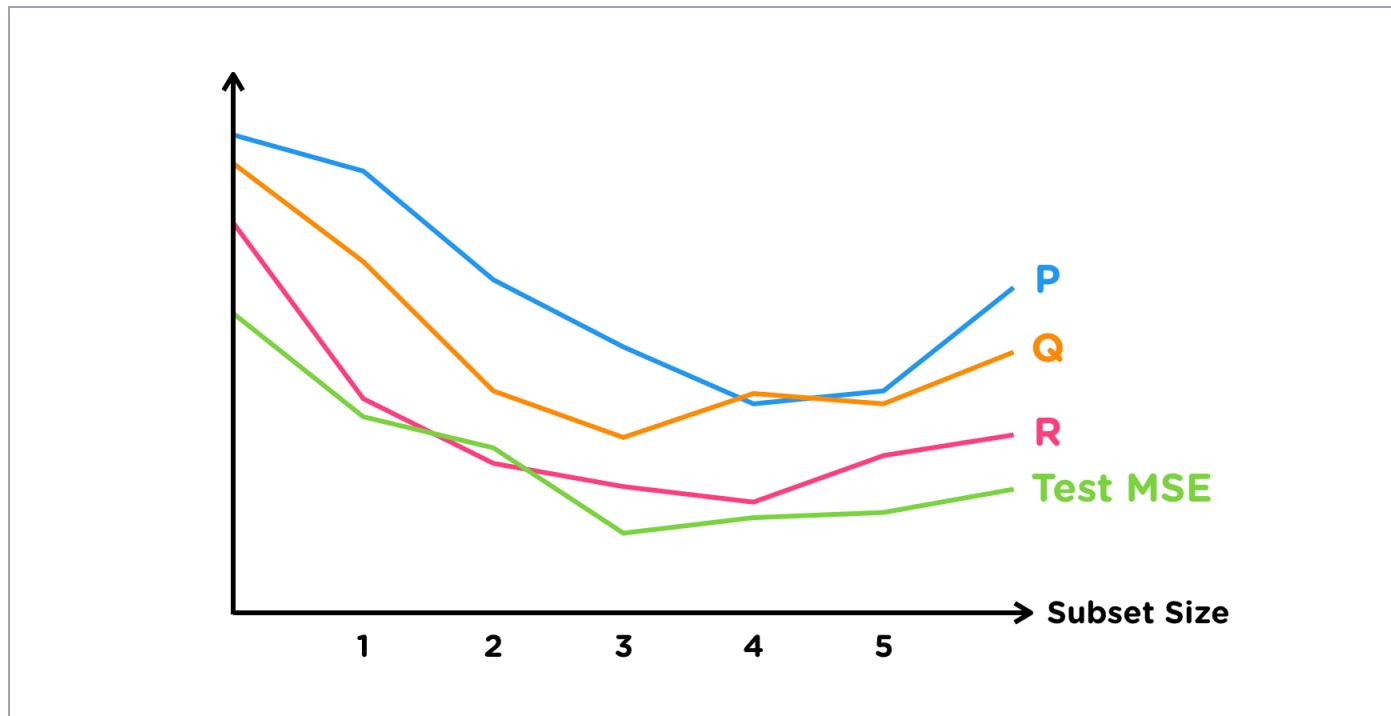
On the other hand, the LOOCV error is made of many quantities that have highly positive correlations due to big overlaps among the training sets, which leads to a high variance. By comparison, the k -fold CV error would have smaller training set overlaps, and thus lower variance. At the extreme is the validation set error with an even lower variance associated with only one training set.

Thus, to balance between the bias and variance, a rule of thumb is to use $k = 5$ or 10 folds in cross-validation.

Example 3.6.3.1

To determine the optimal model with a backward selection procedure, you consider estimating the test MSE using the approaches of validation set, k -fold cross-validation, and LOOCV.

Below is a plot of the test MSE estimates alongside the true test MSE as functions of subset size.



Determine which statements are true.

- I. It is most likely that R is the LOOCV error.
- II. Only Q agrees with the true test MSE on which subset size is optimal.
- III. It is most likely that P will not change when its calculation procedure is repeated.

Solution

I is true because R is the closest to the true test MSE, which is consistent with the LOOCV error having low bias. By that logic, Q is most likely the k -fold CV error, and P is most likely the validation set error.

II is true because Q is lowest at the subset size of 3 – just like the true test MSE – while P and R are lowest at the subset size of 4.

III is false because P is most likely the validation set error. It will likely result in different values when the training and validation sets are different due to randomly dividing the observations.

Therefore, **only I and II are true.**



Example 3.6.3.2

Two multiple linear regressions are considered for a certain dataset.

For Model A, you are given:

- $\hat{y} = 1.856 + 0.651x_1 + 0.709x_2 - 0.556x_3$
- $R^2_{\text{adj.}} = 0.856$

•

Cross-Validation Folds	Test MSE Estimates
1	0.1151
2	0.0870
3	0.1111

For Model B, you are given:

- $\hat{y} = 2.087 + 0.564x_1 + 0.763x_2 - 0.528x_3 - 0.305x_4$
- $R^2_{\text{adj.}} = 0.858$

•

Cross-Validation Folds	Test MSE Estimates
1	0.1116
2	0.0876
3	0.1237

You choose the best model based on 3-fold cross-validation, but your colleague chooses based on adjusted R^2 .

For an observation where $x_1 = 3$, $x_2 = 3.8$, $x_3 = 1.2$, and $x_4 = 1$, let f be the predicted response according to your model, and g be the predicted response according to your colleague's model.

Calculate $f - g$.

Solution

To determine your model, calculate the CV error for both models. For Model A,

$$\text{CV error} = \frac{0.1151 + 0.0870 + 0.1111}{3} = 0.1044$$

whereas for Model B,

$$\text{CV error} = \frac{0.1116 + 0.0876 + 0.1237}{3} = 0.1076$$

Thus, your model is Model A since it has the smaller CV error. On the other hand, your colleague's model is Model B since it has the larger $R^2_{\text{adj.}}$.

Solve for the answer as follows:

$$\begin{aligned} f &= 1.856 + 0.651(3) + 0.709(3.8) - 0.556(1.2) \\ &= 5.836 \end{aligned}$$

$$\begin{aligned} g &= 2.087 + 0.564(3) + 0.763(3.8) - 0.528(1.2) - 0.305(1) \\ &= 5.7398 \end{aligned}$$

$$f - g = 5.836 - 5.7398 = \mathbf{0.0962}$$



3.6 Summary

🕒 5m

Subset Selection

- Types:
 - Best subset selection
 - Forward selection
 - Backward selection
 - Hybrids
- Use R^2 to compare models with the same p ; use an appropriate selection criterion to compare models with different p .

	Scope of Models	Computationally Intensive	Suitable in High Dimensions
Best subset selection	All	Yes	No
Forward selection	Limited	No	Yes
Backward selection	Limited	No	No

Selection Criteria

- Types:
 - $R_{\text{adj.}}^2$
 - Mallows' C_p
 - AIC
 - BIC
 - Cross-validation error
- C_p and AIC will choose the same model as optimal.
- BIC will favor models with smaller p compared to C_p /AIC when $n \geq 8$.

Cross-Validation

- Estimates the test MSE with available data.
- Types:
 - Validation set approach
 - k -fold cross-validation approach
 - LOOCV approach
- The validation set approach has unstable results and will tend to overestimate the test MSE. The two other approaches mitigate these issues.
- With respect to bias, the validation set error has the most, followed by the k -fold CV error, then the LOOCV error.
- With respect to variance, the LOOCV error has the most, followed by the k -fold CV error, then the validation set error.

Appendix

🕒 5m

Deriving Equations 3.6.2.2 and 3.6.2.3 Using Equations 3.8.4.3 and 3.8.4.4

A multiple linear regression is essentially equivalent to a generalized linear model featuring a normal response, an identity link function, and the assumption of homoscedasticity.

The PDF of a normal distribution is given by:

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y-\mu)^2}{2\sigma^2}\right]$$

Hence, the likelihood function is the product of all n probability density functions:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(y_i-\mu_i)^2}{2\sigma^2}\right]$$

where

- σ does not vary with i because of homoscedasticity.
- $\mu_i = \mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}$ using the identity link function.

Then, the log-likelihood function is:

$$\begin{aligned} l(\boldsymbol{\beta}) &= \sum_{i=1}^n \left[-\frac{(y_i - \mu_i)^2}{2\sigma^2} - \ln(\sqrt{2\pi\sigma^2}) \right] \\ &= -\frac{\sum_{i=1}^n (y_i - \mu_i)^2}{2\sigma^2} - \sum_{i=1}^n \ln(\sqrt{2\pi\sigma^2}) \\ &= -\frac{\sum_{i=1}^n (y_i - \mu_i)^2}{2\sigma^2} - \frac{n}{2} \ln(2\pi\sigma^2) \end{aligned}$$

Note that there are $p + 2$ parameters, which includes $p + 1$ coefficients and σ .

Akaike Information Criterion (AIC) – Equation 3.6.2.2

Hence, the AIC is:

$$\begin{aligned}
 \text{AIC} &= -2 \cdot l(\beta) + 2(p + 2) \\
 &= -2 \left[-\frac{\sum_{i=1}^n (y_i - \mu_i)^2}{2\sigma^2} - \frac{n}{2} \ln(2\pi\sigma^2) \right] + 2p + 4 \\
 &= \frac{\sum_{i=1}^n (y_i - \mu_i)^2}{\sigma^2} + n \ln(2\pi\sigma^2) + 2p + 4 \\
 &= \frac{\sum_{i=1}^n (y_i - \mu_i)^2}{\sigma^2} + 2p + n \ln(2\pi\sigma^2) + 4 \\
 &= c_1 \left[\frac{1}{n} \left(\sum_{i=1}^n (y_i - \mu_i)^2 + 2p \cdot \sigma^2 \right) \right] + c_2
 \end{aligned} \tag{3.8.4.3}$$

where c_1 and c_2 are irrelevant constants, such that

- $c_1 = \frac{n}{\sigma^2}$
- $c_2 = n \ln(2\pi\sigma^2) + 4$

The term in the square bracket resembles Equation 3.6.2.2:

$$\text{AIC} = \frac{1}{n} (\text{SSE} + 2p \cdot \text{MSE}_g) \tag{3.6.2.2}$$

Bayesian Information Criterion (BIC) – Equation 3.6.2.3

Hence, the BIC is:

$$\begin{aligned}
 \text{BIC} &= -2 \cdot l(\beta) + \ln n \cdot (p + 2) & (3.8.4.4) \\
 &= -2 \left[-\frac{\sum_{i=1}^n (y_i - \mu_i)^2}{2\sigma^2} - \frac{n}{2} \ln (2\pi\sigma^2) \right] + \ln n \cdot (p + 2) \\
 &= \frac{\sum_{i=1}^n (y_i - \mu_i)^2}{\sigma^2} + n \ln (2\pi\sigma^2) + \ln n \cdot p + 2 \ln n \\
 &= \frac{\sum_{i=1}^n (y_i - \mu_i)^2}{\sigma^2} + \ln n \cdot p + n \ln (2\pi\sigma^2) + 2 \ln n \\
 &= c_1 \left[\frac{1}{n} \left(\sum_{i=1}^n (y_i - \mu_i)^2 + \ln n \cdot p \cdot \sigma^2 \right) \right] + c_2
 \end{aligned}$$

where c_1 and c_2 are irrelevant constants, such that

- $c_1 = \frac{n}{\sigma^2}$
- $c_2 = n \ln (2\pi\sigma^2) + 2 \ln n$

The term in the square bracket resembles Equation 3.6.2.3:

$$\text{BIC} = \frac{1}{n} (\text{SSE} + \ln n \cdot p \cdot \text{MSE}_g) \quad (3.6.2.3)$$

3.7.0 Overview

5m

Recall that multiple linear regression relies on ordinary least squares to determine the coefficient estimates.

In this subsection, we present alternatives that vary slightly from an MLR fit, namely:

- Shrinkage (regularization) methods – ridge and lasso
- Dimension reduction methods – principal components regression and partial least squares

These techniques may assume the variables used will be centered and/or scaled. We begin by explaining what this means.

3.7.1 Standardizing Variables

A **centered variable** is the result of subtracting the sample mean from a variable.

A **scaled variable** is the result of dividing a variable by its sample standard deviation.

A **standardized variable** is the result of first centering a variable, then scaling it.

Across various resources, it is common to find some that scale using the **biased** sample standard deviation (i.e. denominator of n) and others that use the unbiased version. In many cases, this choice does not have a large impact on the end results.

If an exam problem is not explicit, we anticipate scaling with the biased sample standard deviation. In this subsection, we let s represent the biased sample standard deviation.

Let's demonstrate these concepts using the following data.

y	x_1	x_2
35	9	17
44	5	5
31	11	8
47	8	9
43	7	11

Their sample means and biased sample standard deviations are:

- $\bar{y} = 40, \bar{x}_1 = 8, \bar{x}_2 = 10$
- $s_y = 6, s_{x_1} = 2, s_{x_2} = 4$

Considering $y_1 = 35$,

- the first centered value of y is $35 - 40 = -5$,
- the first scaled value of y is $\frac{35}{6} = 5.83$, and
- the first standardized value of y is $\frac{35-40}{6} = -0.83$.

To see how these transformations impact multiple linear regression, note that the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

leads to the coefficient estimates of

- $\hat{\beta}_0 = 59.1345$
- $\hat{\beta}_1 = -2.2873$
- $\hat{\beta}_2 = -0.0836$

CENTERED PREDICTORS

For the model

$$Y = \beta_0 + \underbrace{\beta_1(x_1 - 8)}_{\text{centered } x_1} + \underbrace{\beta_2(x_2 - 10)}_{\text{centered } x_2} + \varepsilon$$

the coefficient estimates are

- $\hat{\beta}_0 = 40$
- $\hat{\beta}_1 = -2.2873$
- $\hat{\beta}_2 = -0.0836$

Only $\hat{\beta}_0$ is different as compared to using the uncentered predictors. However, it is not a coincidence that this $\hat{\beta}_0$ is the same as \bar{y} . In fact, both models have identical fitted equations.

$$\begin{aligned}\hat{y} &= 40 - 2.2873(x_1 - 8) - 0.0836(x_2 - 10) \\ &= 40 + 8(2.2873) + 10(0.0836) - 2.2873x_1 - 0.0836x_2 \\ &= 59.1345 - 2.2873x_1 - 0.0836x_2\end{aligned}$$

SCALED PREDICTORS

For the model

$$Y = \beta_0 + \beta_1 \underbrace{\left(\frac{x_1}{2}\right)}_{\text{scaled } x_1} + \beta_2 \underbrace{\left(\frac{x_2}{4}\right)}_{\text{scaled } x_2} + \varepsilon$$

the coefficient estimates are

- $\hat{\beta}_0 = 59.1345$
- $\hat{\beta}_1 = -4.5745$
- $\hat{\beta}_2 = -0.3345$

Now $\hat{\beta}_1$ and $\hat{\beta}_2$ are different as compared to using the unscaled predictors. Even so, both models have identical fitted equations. This attribute is known as being **scale equivariant**.

$$\begin{aligned}\hat{y} &= 59.1345 - 4.5745 \left(\frac{x_1}{2} \right) - 0.3345 \left(\frac{x_2}{4} \right) \\ &= 59.1345 - 2.2873x_1 - 0.0836x_2\end{aligned}$$

When using ordinary least squares, note that:

- Centering and/or scaling the predictors does not intrinsically change the fitted equation.
- When centered predictors are used, the intercept estimate will equal the sample mean of the response, while the rest of the estimated regression coefficients remain unchanged.
- When scaled predictors are used, the intercept estimate does not change. However, the other coefficient estimates will change by a factor equal to the sample standard deviation of the associated explanatory variable.

Based on these facts, it is not difficult to infer that the coefficient estimates when using standardized x_1 and x_2 will be

- $\hat{\beta}_0 = 40$
- $\hat{\beta}_1 = -4.5745$
- $\hat{\beta}_2 = -0.3345$

This is demonstrated as follows:

$$\begin{aligned}\hat{y} &= 59.1345 - 2.2873x_1 - 0.0836x_2 \\&= 40 + 8(2.2873) + 10(0.0836) - 2.2873x_1 - 0.0836x_2 \\&= 40 - 2.2873(x_1 - 8) - 0.0836(x_2 - 10) \\&= 40 - 4.5745 \left(\frac{x_1 - 8}{2} \right) - 0.3345 \left(\frac{x_2 - 10}{4} \right)\end{aligned}$$

3.7.2 Ridge Regression

Recall that ordinary least squares (OLS) aims to find the regression coefficient estimates that minimize the SSE

$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_p x_{i,p} \right)^2$$

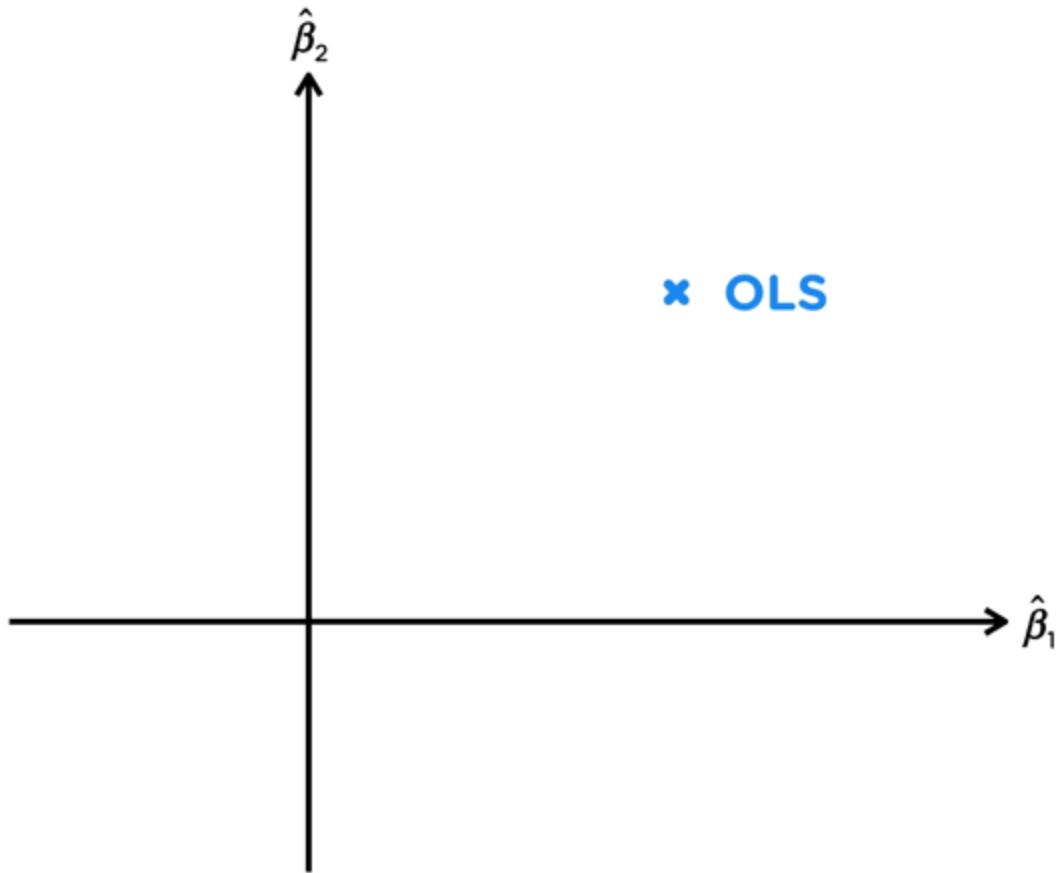
As p (i.e. flexibility) grows, OLS gets closer to having no bias. However, by the bias-variance trade-off, this is at the expense of increasing the method's variance. One way to reduce variance (while allowing a bit more bias) is to restrict the possible values of the coefficient estimates. Specifically, we restrict by forcing the coefficient estimates to be closer to or shrunken towards 0.

The first shrinkage method for consideration is *ridge regression*. It minimizes the same SSE expression with an added restriction of

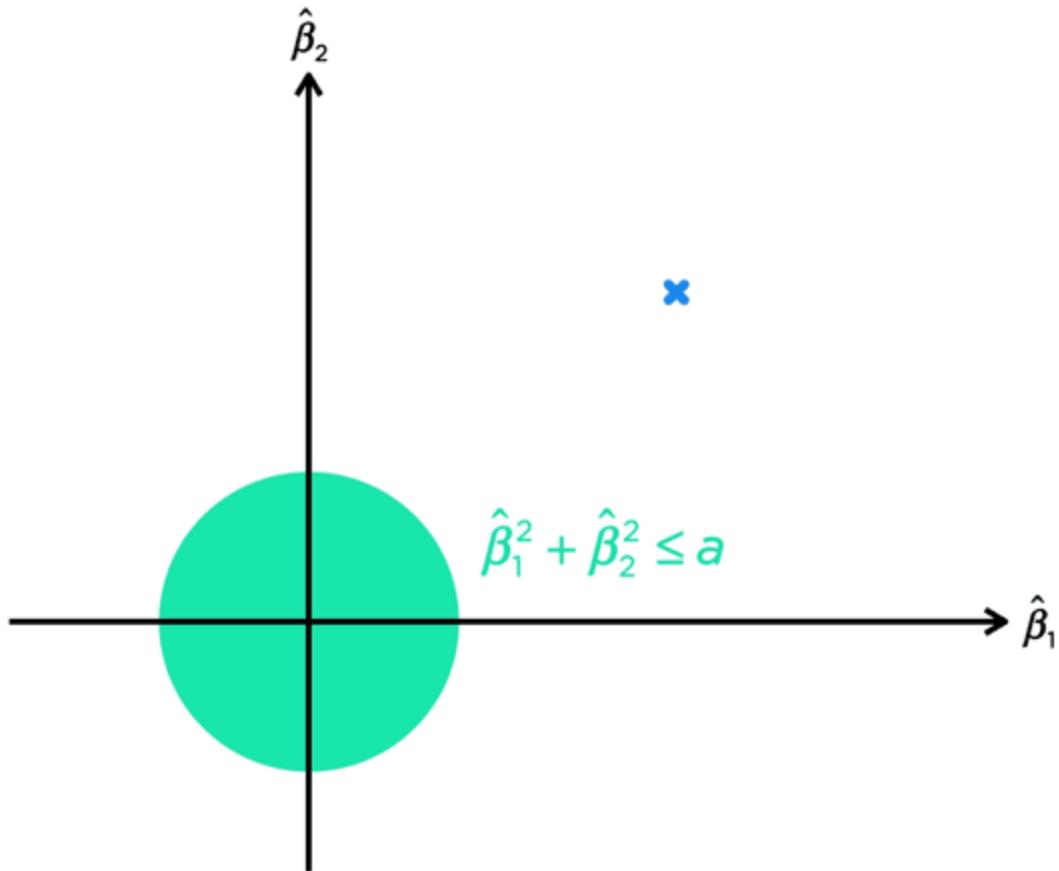
$$\sum_{j=1}^p \hat{\beta}_j^2 \leq a \tag{3.7.1.1}$$

for some constant a called the *budget parameter*. Notice that this restricts all the coefficient estimates except the intercept's.

We can depict this restriction for the case where $p = 2$. First, consider a plot of possible $\hat{\beta}_1$ and $\hat{\beta}_2$ values.



In the animation above, the cross marks the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ under OLS. Since OLS minimizes the SSE, any other point on this plot will result in a larger SSE relative to the cross. We illustrate SSE as a third dimension using contours. For each contour, every point along the contour has the **same** SSE value at those $\hat{\beta}_1$ and $\hat{\beta}_2$ coordinates. The further the contour is from the cross, the larger the SSE.



The animation above includes the region $\hat{\beta}_1^2 + \hat{\beta}_2^2 \leq a$ shaded in green. Ridge regression produces the coefficient estimates that minimize the SSE while restricted to the circle. Since this circle does not include the cross, the contour that first touches the edge of the circle satisfies the ridge regression criterion. Hence, their intersection is the closest to the cross (in terms of SSE) among all points within the circle, and denotes the ridge estimates.

The graphs demonstrate how the values of $\hat{\beta}_1$ and $\hat{\beta}_2$ would tend to shrink towards 0 for ridge regression. But it is possible that no shrinking occurs, as when the green circle is large enough to include the OLS estimates.

In addition, the restriction can be incorporated directly into an optimization problem. Ridge regression seeks to minimize the expression

$$\underbrace{\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_p x_{i,p} \right)^2}_{\text{SSE}} + \lambda \underbrace{\sum_{j=1}^p \hat{\beta}_j^2}_{\text{penalty}} \quad (3.7.1.2)$$

where λ is the **tuning parameter** that controls the strength of the shrinkage. Note that:

- When $\lambda = 0$, there is no shrinkage penalty, so $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$ will equal the OLS estimates.
- As λ approaches ∞ , the entire expression is minimized when $\hat{\beta}_1, \dots, \hat{\beta}_p$ equal 0.

This means λ is inversely related to flexibility. λ allows us to select a flexibility level between the inflexible null model ($\lambda \rightarrow \infty$) and the flexible multiple linear regression with all p predictors ($\lambda = 0$). The process of finding the best value of λ is called **tuning**, which can be achieved through cross-validation.

Furthermore, some might find the notation $\sum_{j=1}^p \hat{\beta}_j^2$ rather cumbersome. It can be simplified using the ℓ norm notation. First, let

$$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$$

Then, the ℓ_2 norm of $\hat{\boldsymbol{\beta}}$ is

$$\left\| \hat{\boldsymbol{\beta}} \right\|_2 = \sqrt{\sum_{j=1}^p \hat{\beta}_j^2} \quad (3.7.1.3)$$

Consequently, the restriction inequality and the expression to be minimized for ridge regression can be simplified as follows:

$$\left\| \hat{\boldsymbol{\beta}} \right\|_2^2 \leq a$$

$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_p x_{i,p} \right)^2 + \lambda \left\| \hat{\boldsymbol{\beta}} \right\|_2^2$$

Other important details regarding ridge regression include:

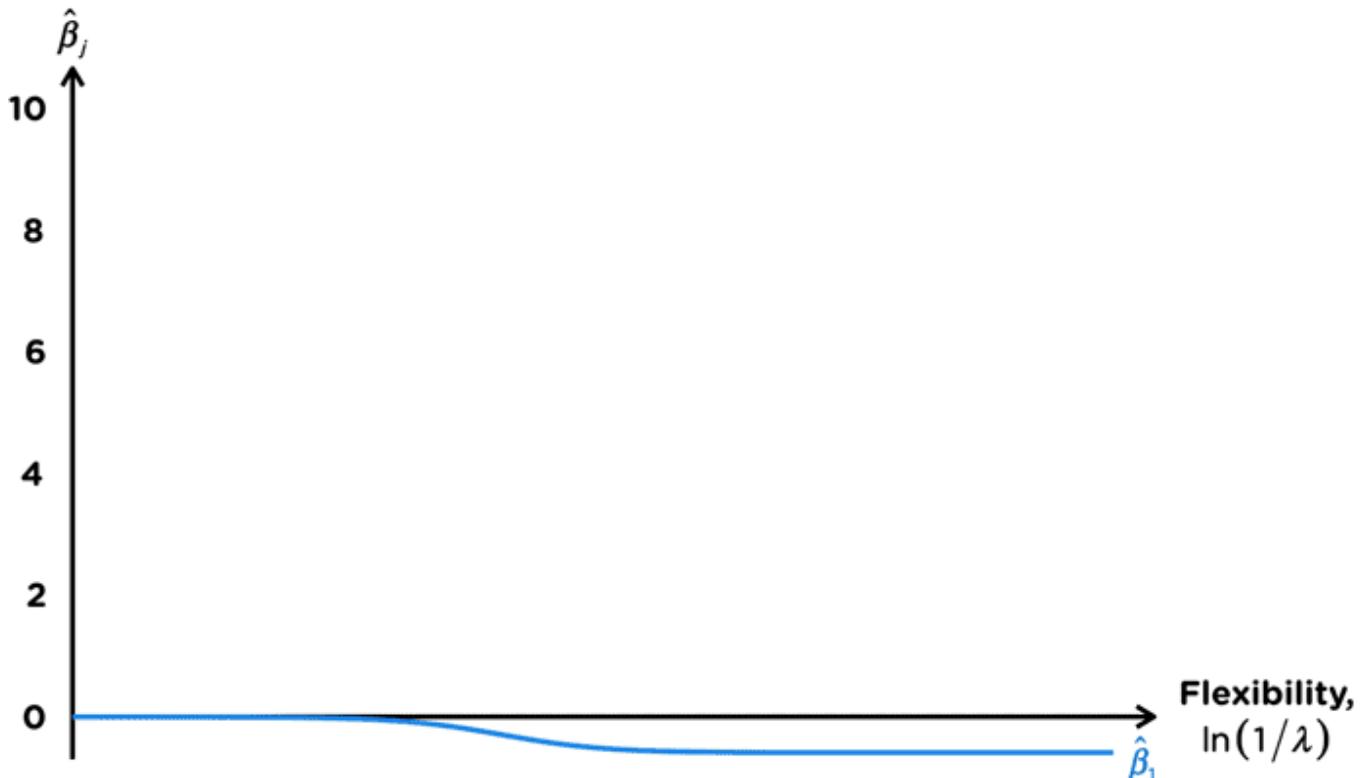
- In Equation 3.7.1.2, the x_j 's should be **scaled** variables of the original. This puts $\hat{\beta}_1, \dots, \hat{\beta}_p$ on the same scale, which is crucial because the shrinkage penalty puts an equal weight on each one. This means the ridge estimates are **not** scale equivariant.

- As λ increases, $\sum_{j=1}^p \hat{\beta}_j^2$ must decrease (i.e. be shrunk towards 0) to minimize SSE plus penalty. Yet, it is possible for an **individual** $\hat{\beta}_j$ to increase in absolute value (i.e. deviate away from 0) as λ increases.
 - Practically, none of the p ridge estimates will equal 0 as long as λ is finite. This means ridge regression does not drop variables from a model.
 - It is useful when dealing with high dimensions as only some of the predictors will have a meaningful estimated coefficient.
-

With the Commuting Chris scenario, we run ridge regression for 100 different values of λ (between 0.01 and 1,000,000) using all predictors except Season. Let

- x_1 represent Departure
- x_2 represent Temp
- x_3 represent Precip Chance
- x_4 represent the dummy variable for Precip
- x_5 represent the dummy variable for Accident
- x_6 represent Police

After scaling the predictors, running the ridge procedure, and reversing the scale back to the original units, we see how the ridge coefficients change as a function of flexibility in the graph below.



Since λ is inversely related to flexibility, we let the natural log of the reciprocal of λ denote flexibility. While the reciprocal itself can be a measure of flexibility, it has an extremely large interval of $[10^{-6}, 100]$; further taking the natural log narrows the interval to $[-13.816, 4.605]$, which is more suitable for plotting.

Where flexibility is low, the estimates are very close to 0; where flexibility is high, the estimates are practically the same as the OLS estimates. Therefore, shrinkage occurs from right to left.

Coach's Remarks

Note that a plot may use a penalty measure like λ in the horizontal axis instead of flexibility. Since flexibility and penalty are inversely related, both plots would look different despite having the same information; this can be confusing. Keep in mind that a plot with an increasing flexibility from left to right has an increasing penalty from right to left, and vice versa.

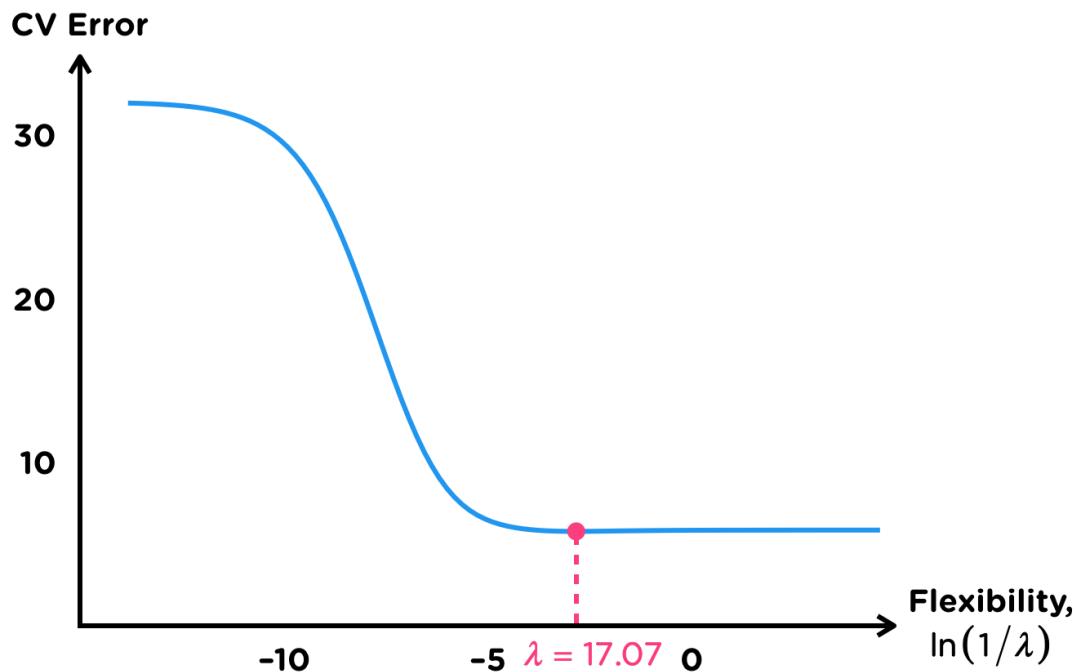
Here are the OLS and ridge estimates at the extreme values of λ :

Coefficients	OLS	Ridge,	Ridge,
$\hat{\beta}_1$	-0.58667	-0.58667	-0.00018

Coefficients	OLS	Ridge, $\lambda = 0.01$	Ridge, $\lambda = 1,000,000$
$\hat{\beta}_2$	-0.01342	-0.01342	-0.00005
$\hat{\beta}_3$	0.02095	0.02095	0.00004
$\hat{\beta}_4$	3.55843	3.55836	0.00339
$\hat{\beta}_5$	10.21613	10.21446	0.00630
$\hat{\beta}_6$	-0.08179	-0.08116	0.00216

In addition, notice the behavior of $\hat{\beta}_6$, the ridge estimate for Police. As flexibility decreases, the negative estimate first shrinks to 0 but keeps increasing and becomes positive. It eventually shrinks back to 0. This illustrates that an individual $\hat{\beta}_j$ may not always shrink towards 0 as the penalty increases.

Using 10-fold cross-validation, the lowest CV error coincides with $\lambda = 17.07$ as seen below.



However, it is clear that the CV error is practically the same as the minimum for many λ values. Since a lower flexibility with little to no loss in model accuracy is preferred (i.e. parsimony), a λ larger than 17.07 may be more suitable. In any case, the OLS and ridge estimates at $\lambda = 17.07$ are

Coefficients	OLS	Ridge,
$\hat{\beta}_0$	29.08082	28.69592
$\hat{\beta}_1$	-0.58667	-0.57412
$\hat{\beta}_2$	-0.01342	-0.00760

Coefficients	OLS	Ridge, $\lambda = 17.07$
$\hat{\beta}_3$	0.02095	0.01912
$\hat{\beta}_4$	3.55843	3.44500
$\hat{\beta}_5$	10.21613	8.98096
$\hat{\beta}_6$	-0.08179	0.36961

Example 3.7.1.1

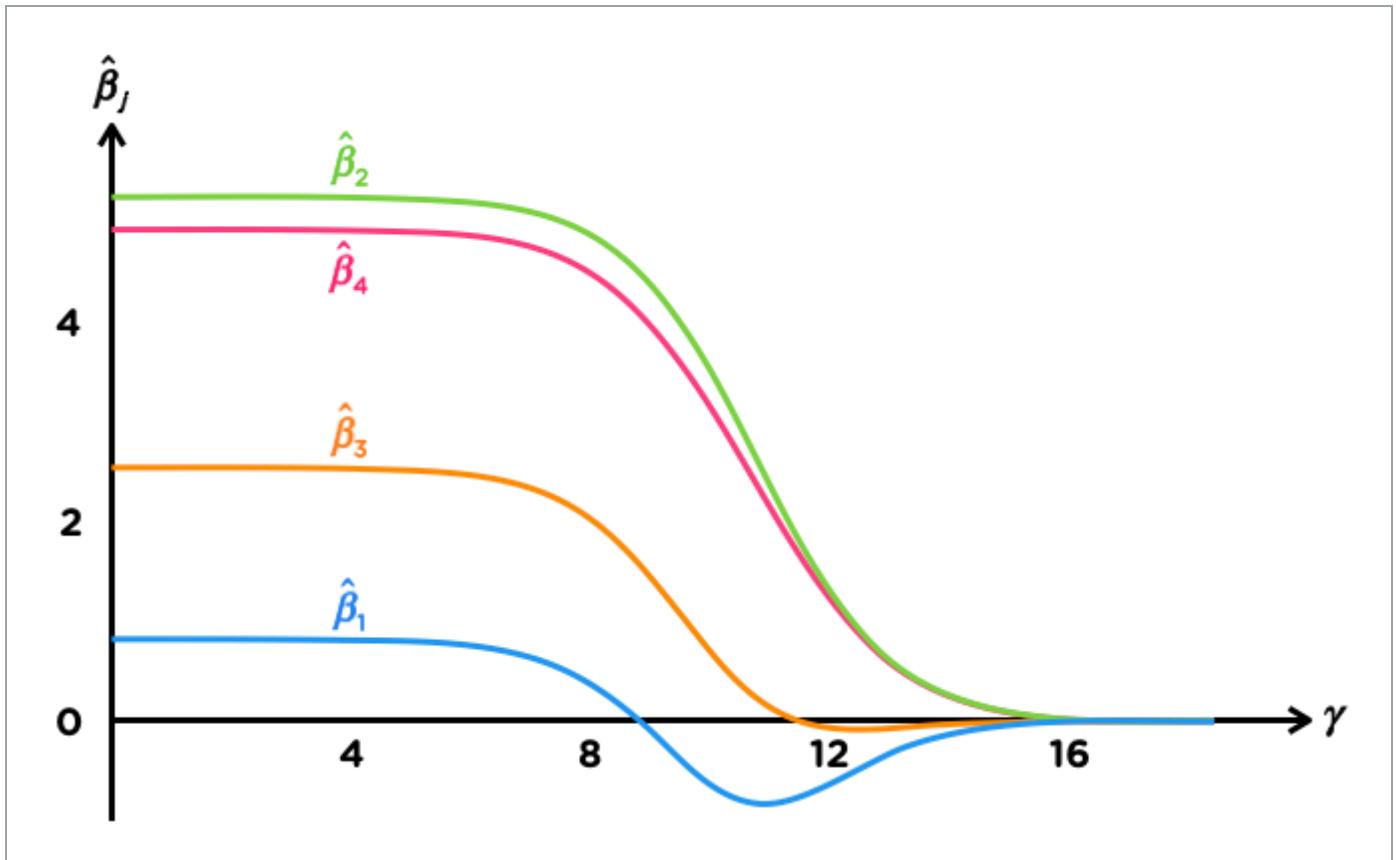
Charlie studies claim amounts using two categorical variables, each having three categories, that classify a group of insureds. He considers the model

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

where

- the baseline category is an insured with low risk and a policy in force for less than a year,
- x_1 and x_2 are dummy variables for medium and high risk, respectively, and
- x_3 and x_4 are dummy variables for a policy in force between 1 and 3 years and more than 3 years, respectively.

A ridge regression has its estimated coefficients of the four predictors plotted against a quantity γ in the graph below:



Determine which statement is true.

- I. Larger values of γ indicate a more flexible model, and vice versa.
- II. For $\gamma > 9$, the expected claims for a medium risk insured is lower than for a low risk insured, all else being equal.
- III. For $12 < \gamma < 14$, there is no practical difference in expected claims for a policy in force between 1 and 3 years versus more than 3 years, all else being equal.

Solution

I is false because the graph depicts shrinkage occurring from left to right. A stronger shrinkage or penalty leads to a less flexible fit.

II is true because $\hat{\beta}_1$ is negative for $\gamma > 9$, which leads to lower expected claims for a medium risk insured relative to a low risk insured.

III is false because the difference in expected claims for a policy in force between 1 and 3 years versus more than 3 years is captured by the absolute difference in $\hat{\beta}_3$ and $\hat{\beta}_4$, which is not close to zero for $12 < \gamma < 14$.

Therefore, **only II is true.**



Example 3.7.1.2

Determine which of the following statements about ridge regression are true.

- I. Each of the coefficient estimates will always shrink towards 0 as the tuning parameter increases.
- II. When the tuning parameter is 0, it simplifies to the null model.
- III. It is a method with higher variance compared to ordinary least squares.
- IV. Centering the predictors before performing the estimation is important to ensure the coefficients are shrunk based on the same scale.

Solution

I is false because each of the estimates could individually deviate away from 0 as the tuning parameter increases. Instead, it is $\sum_{j=1}^p \hat{\beta}_j^2$ that will always shrink towards 0.

II is false because a tuning parameter of 0 means there is no shrinkage penalty. Thus, the result is a multiple linear regression with all p predictors. By contrast, ridge regression simplifies to the null model when the tuning parameter approaches ∞ .

III is false because ridge regression has a lower variance relative to ordinary least squares by restricting the coefficient estimates through shrinkage.

IV is false because it is scaling the predictors rather than centering that achieves this goal.

Therefore, **none of the statements are true.**



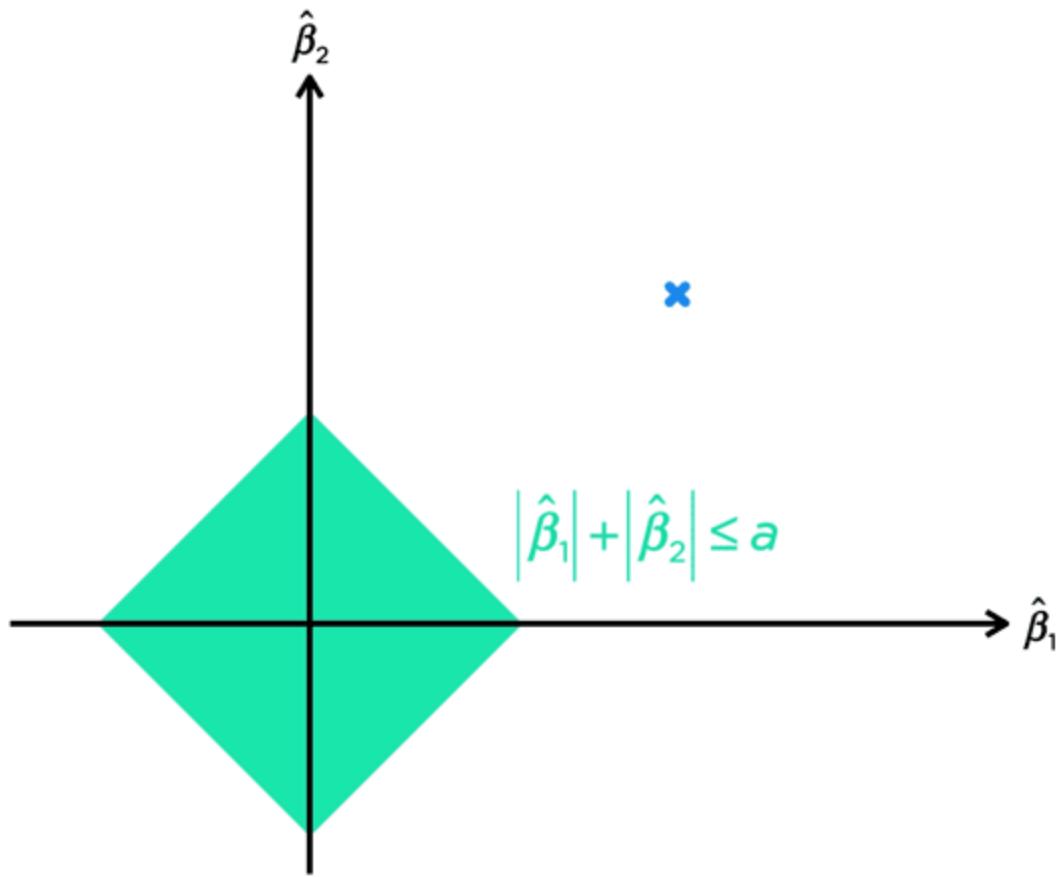
3.7.3 Lasso Regression

Lasso regression is another shrinkage method; it bears a close resemblance to ridge regression. In minimizing the SSE, lasso imposes the restriction of

$$\sum_{j=1}^p |\hat{\beta}_j| \leq a \quad (3.7.3.1)$$

with the budget parameter a . In short, where ridge takes the square of the $\hat{\beta}_j$'s, lasso takes their absolute value instead.

Let's visualize the restriction for the case where $p = 2$ with a plot of $\hat{\beta}_2$ against $\hat{\beta}_1$.



Again, the cross represents the OLS estimates. The region $|\hat{\beta}_1| + |\hat{\beta}_2| \leq a$ is shaded in green. Since this diamond region does not include the cross, the SSE contour that first touches the edge of the diamond reveals the lasso estimates.

The lasso estimates can be found by minimizing the expression

$$\underbrace{\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_p x_{i,p} \right)^2}_{\text{SSE}} + \lambda \sum_{j=1}^p \left| \hat{\beta}_j \right| \quad (3.7.3.2)$$

penalty

where λ is the tuning parameter that controls the shrinkage strength and functions in the same way as in ridge regression.

Moreover, the ℓ_1 norm of $\hat{\beta}$ is

$$\left\| \hat{\beta} \right\|_1 = \sum_{j=1}^p \left| \hat{\beta}_j \right| \quad (3.7.3.3)$$

Then, the restriction inequality and the expression to be minimized for lasso regression can be simplified as follows:

$$\left\| \hat{\beta} \right\|_1 \leq a$$

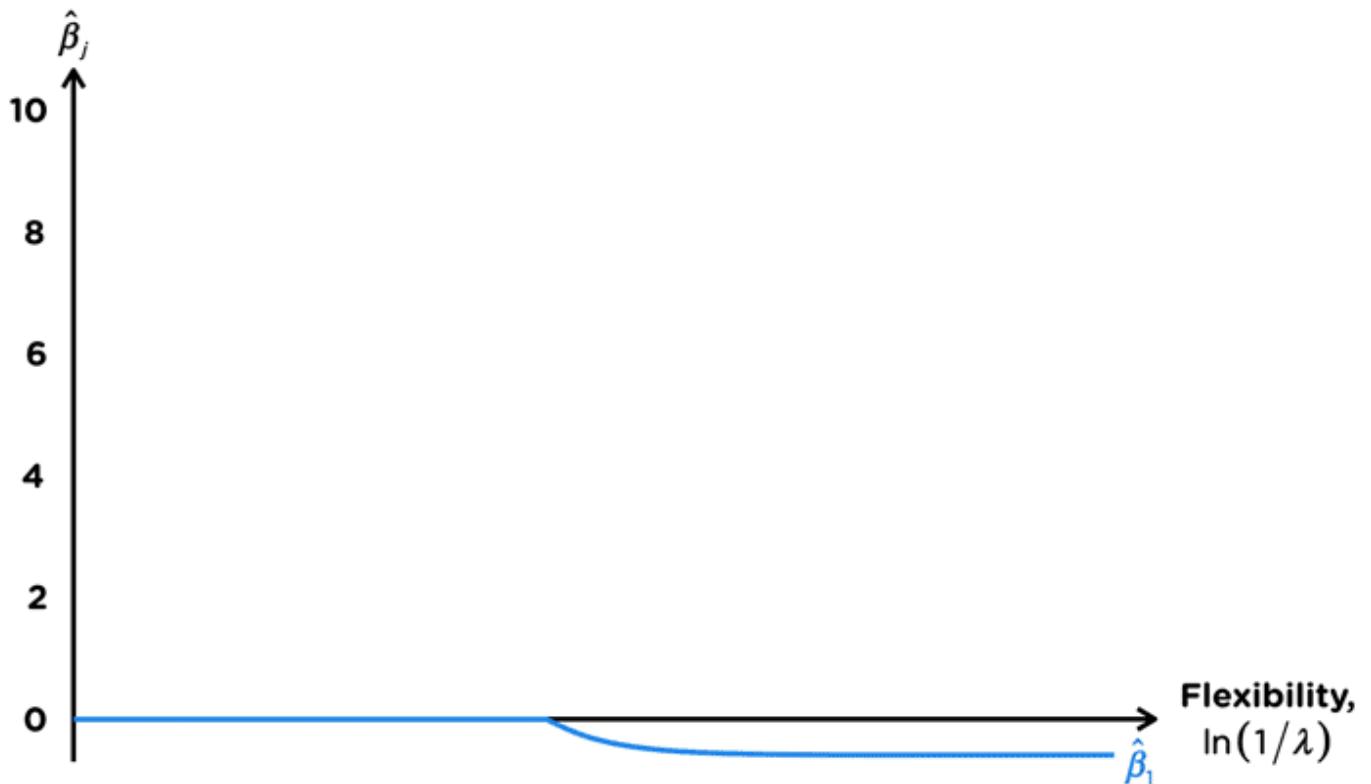
$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_p x_{i,p} \right)^2 + \lambda \left\| \hat{\beta} \right\|_1$$

Other important details regarding lasso regression include:

- In Equation 3.7.3.2, the x_j 's should be **scaled** variables of the original for the same reason mentioned in the previous subsection. Lasso estimates are **not** scale equivariant.
- As λ increases, $\sum_{j=1}^p \left| \hat{\beta}_j \right|$ must decrease to minimize SSE plus penalty. Like ridge regression, it is possible for an individual $\hat{\beta}_j$ to increase in absolute value as λ increases.
- Unlike ridge regression, the lasso estimates can equal 0 for a large enough, finite λ . This means lasso regression does drop variables from a model.
- It is useful when dealing with high dimensions.

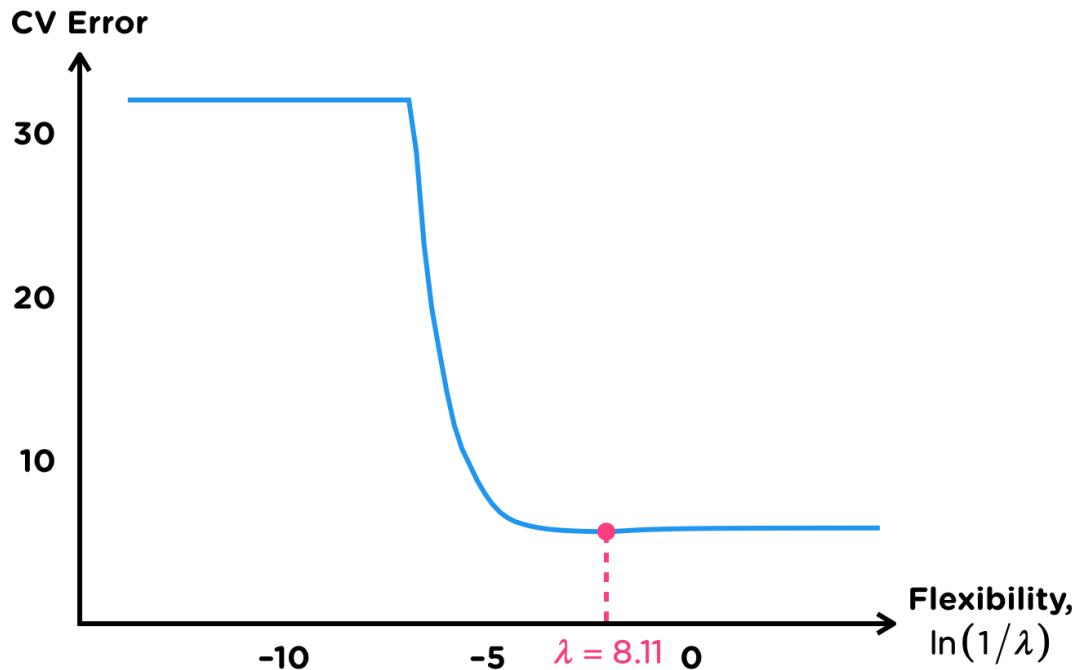
Let's perform lasso regression with the same 100 values of λ from the previous subsection for the Commuting Chris scenario. The six predictors (i.e. no Season) are first scaled before running the

lasso procedure. After reverting back to the original units, we see the lasso coefficients as a function of flexibility.



As the shrinkage becomes stronger from right to left, the lasso estimates converge and become 0 one after another. We also see a similar behavior with $\hat{\beta}_6$ as compared to the ridge regression – it first shrinks to and stays 0, then becomes positive, then shrinks back to 0. At roughly the flexibility level of -7 (specifically when $\lambda = 1,023.53$), all six lasso coefficients are 0 for the first time and remain so for greater penalties.

Using 10-fold cross-validation, the lowest CV error coincides with $\lambda = 8.11$ as seen below.



Much like the case with ridge regression, many λ values produce CV errors that are close to the minimum. Therefore, there is an argument for using a larger λ than 8.11. The OLS and lasso estimates at $\lambda = 8.11$ are

Coefficients	OLS	Lasso, $\lambda = 8.11$
$\hat{\beta}_0$	29.08082	28.72512
$\hat{\beta}_1$	-0.58667	-0.56542
$\hat{\beta}_2$	-0.01342	-0.01016
$\hat{\beta}_3$	0.02095	0.02038
$\hat{\beta}_4$	3.55843	3.50518
$\hat{\beta}_5$	10.21613	9.96472
$\hat{\beta}_6$	-0.08179	0

The output shows that the cross-validated lasso regression drops Police from the model. Lasso resembles a mix between subset selection and ridge.

Example 3.7.3.1

For a linear regression involving two scaled predictors, the following table gives the residual

sum of squares for pairs of $\hat{\beta}_1$ and $\hat{\beta}_2$ estimates, assuming these estimates are the only values in consideration.

		$\hat{\beta}_2$		
		1	2	3
$\hat{\beta}_1$	-1	99.4	96.0	79.3
	0	98.4	97.8	82.1
	1	105.5	114.7	90.8
	2	164.2	130.3	186.4

With a budget parameter of 2, calculate the absolute difference between the lasso estimate of $\hat{\beta}_1$ and the lasso estimate of $\hat{\beta}_2$.

Solution

A lasso regression with a budget parameter of 2 means the residual sum of squares (SSE) is minimized with the restriction of

$$|\hat{\beta}_1| + |\hat{\beta}_2| \leq 2$$

Only four pairs satisfy this inequality:

- $\hat{\beta}_1 = -1, \hat{\beta}_2 = 1$
- $\hat{\beta}_1 = 0, \hat{\beta}_2 = 1$
- $\hat{\beta}_1 = 0, \hat{\beta}_2 = 2$
- $\hat{\beta}_1 = 1, \hat{\beta}_2 = 1$

Among these four, the lowest SSE is 97.8, which is produced by the estimates $\hat{\beta}_1 = 0, \hat{\beta}_2 = 2$. Therefore, these are the lasso estimates; their absolute difference is **2**.

Example 3.7.3.2

Determine which of the following statements are true with a finite tuning parameter.

- I. Lasso regression is objectively better than ridge regression.
- II. Lasso regression is more easily interpretable than ridge regression, especially when considering many features.
- III. Lasso regression is useful for avoiding overfits, whereas ridge regression is not.
- IV. Lasso regression performs variable selection on the predictors, whereas ridge regression does not.

Solution

I is false. There is no clear advantage between either shrinkage method. For example, if a predictor's coefficient is truly close to 0, lasso may estimate it as 0 instead. On the other hand, if a predictor does not belong in a model, ridge will fail to estimate its coefficient as 0.

II is true. Since lasso may drop predictors that it deems as less important, only a handful of key predictors would remain, leading to results that are easier to interpret. This is not the case with ridge as it will retain all the predictors, making it more difficult to motivate or justify the inclusion of each predictor.

III is false. Both methods avoid overfits by finding the appropriate amount of flexibility through cross-validation.

IV is true. Lasso causes some of the coefficient estimates to equal 0, implying that the predictors with non-zero coefficient estimates are chosen to be in the model. Since all the predictors are kept in the model with ridge, no variable selection occurs.

Therefore, **only II and IV are true.**

3.7.4 Principal Components Analysis and Regression

Recall from Section 2.2.1 that a statistic is described as summarizing n random variables by mapping them to one value. Now consider summarizing a dataset with variables x_1, \dots, x_p into new and fewer variables. We can achieve this by performing principal components analysis. **Principal components analysis (PCA)** is an unsupervised way to obtain new variables that summarize x_1, \dots, x_p in a dataset. A more technical definition of PCA is: a statistical tool that finds a low-dimensional representation of a dataset without significant loss of information.

After explaining how PCA operates, we consider using these new variables as predictors in a multiple linear regression setting – this is called **principal components regression (PCR)**.

For this subsection, we define x_1, \dots, x_p as the **centered** variables of the original.

Principal Components Analysis

Just as a statistic summarizes random variables via a function, PCA summarizes x_1, \dots, x_p via a function as well. Specifically, a new variable is created by taking a linear combination of the original variables.

Denote such a new variable as z_m , which we call the m^{th} principal component. Then, z_m is defined as

$$z_m = \sum_{j=1}^p \phi_{j,m} x_j \quad (3.7.4.1)$$

The coefficients of this linear combination, $\phi_{1,m}, \dots, \phi_{p,m}$, are called the **loadings** of the m^{th} principal components.

Then, for each observation in the dataset, we can obtain an m^{th} principal component score. To find the **m^{th} principal component score** for the i^{th} observation, evaluate the variables in Equation 3.7.4.1 at the corresponding values for the i^{th} observation, i.e.

$$z_{i,m} = \sum_{j=1}^p \phi_{j,m} x_{i,j} \quad (3.7.4.2)$$

FIRST PRINCIPAL COMPONENT

How do we determine the principal component loadings? In PCA, the principal components are created in a sequence, so we start with the first principal component. The goal is for z_1 to explain the largest portion of variability in a dataset, compared to subsequent z_m 's. To achieve this, we maximize the (biased) sample variance of z_1 . In other words, the values of $\phi_{1,1}, \dots, \phi_{p,1}$ are determined by maximizing

$$\begin{aligned} \frac{\sum_{i=1}^n (z_{i,1} - \bar{z}_1)^2}{n} &= \frac{\sum_{i=1}^n z_{i,1}^2}{n} \\ &= \frac{\sum_{i=1}^n (\phi_{1,1}x_{i,1} + \dots + \phi_{p,1}x_{i,p})^2}{n} \end{aligned}$$

Moreover, we maximize while constrained by $\sum_{j=1}^p \phi_{j,1}^2 = 1$ in order to obtain non-trivial results.

Since centered variables are used, it can be shown that $\bar{z}_1 = 0$, which is why the expression above simplifies.

To solve for these loadings, a technique known as eigen decomposition is often used; learning this technique is not required for this exam. However, note that the loadings are determined using the x_1, \dots, x_p variables only. In the absence of a response variable, PCA is an unsupervised learning statistical method.

SECOND AND SUBSEQUENT PRINCIPAL COMPONENTS

Next, consider the second principal component. The goal is for z_2 to explain the next largest portion of remaining variability in a dataset that is not explained by z_1 . Aside from the analogous first principal component requirements, this means:

- z_2 is further constrained to be uncorrelated with z_1 . This is equivalent to saying that the vector of loadings for the first principal component is **orthogonal** or **perpendicular** to the vector of loadings for the second principal component, i.e.

$$\sum_{j=1}^p \phi_{j,1} \cdot \phi_{j,2} = 0$$

In turn, z_1 and z_2 are orthogonal variables. Recall the discussion on orthogonal variables in Section 3.5.5.

- The variability explained by the second principal component is less than the variability explained by the first principal component.

We can generalize this for all the subsequent principal components.

- All principal components are uncorrelated with one another. In other words,

$$\sum_{j=1}^p \phi_{j,m} \cdot \phi_{j,u} = 0 \quad (3.7.4.3)$$

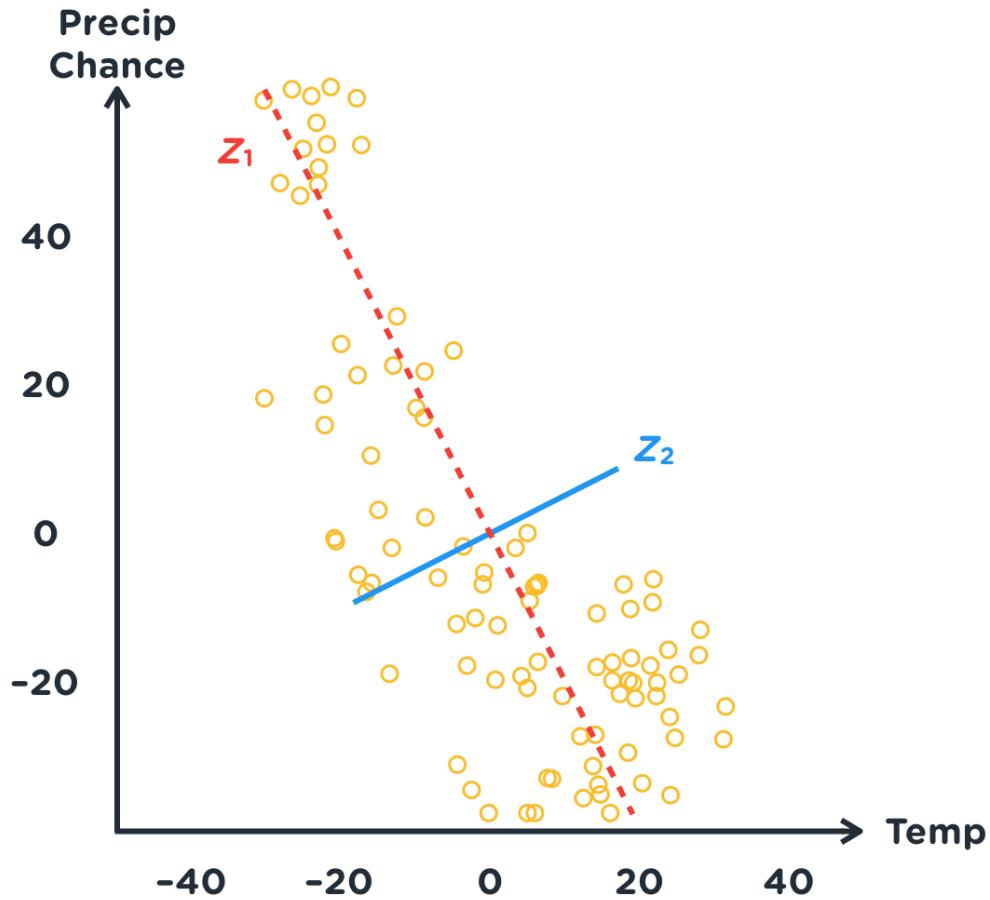
for all pairs of m and u where $m \neq u$.

- The variability explained by each subsequent principal component is always less than the variability explained by its previous principal component.

We usually are able to create up to p distinct principal components – the same number of original variables. All p principal components would collectively capture all of the variability from the dataset. The ideal situation is for the first **few** principal components to explain **much** of the dataset's variability. Thus, PCA is a **dimension reduction** technique, as all p original variables are being summarized/compressed into fewer variables.

COMMUTING CHRIS EXAMPLE

Let's apply PCA on two Commuting Chris variables: Temp and Precip Chance. The scatterplot below shows all observations of Precip Chance against Temp after centering. It is easy to visualize the first principal component on the same plot, as represented by the red dashed line, as well as the second principal component, being represented by the blue solid line.



The first principal component captures the largest variability present in the dataset of two variables. Visually, it is the direction where the variability of the points is the most. Since Precip Chance appears to vary more than Temp, it makes sense why z_1 is slanted more vertically than horizontally.

Note that the second principal component is uncorrelated with the first principal component. This is why the directions of both principal components are perpendicular to each other. Furthermore, $p = 2$ means that the second principal component captures the rest of the dataset's variability that the first principal component did not; there is no third principal component.

The first principal component loadings are $\phi_{1,1} = -0.4528$ and $\phi_{2,1} = 0.8916$. Therefore,

$$z_1 = -0.4528x_1 + 0.8916x_2$$

On the other hand, the second principal component loadings are $\phi_{1,2} = -0.8916$ and $\phi_{2,2} = -0.4528$. Hence,

$$z_2 = -0.8916x_1 - 0.4528x_2$$

In addition, note that:

- The principal components intersect at the origin. This is a consequence of centering both variables before performing PCA.
- The sum of squares of the (unrounded) first principal component loadings is 1, and likewise for the second.

$$\phi_{1,1}^2 + \phi_{2,1}^2 = (-0.4528)^2 + (0.8916)^2 = 1$$

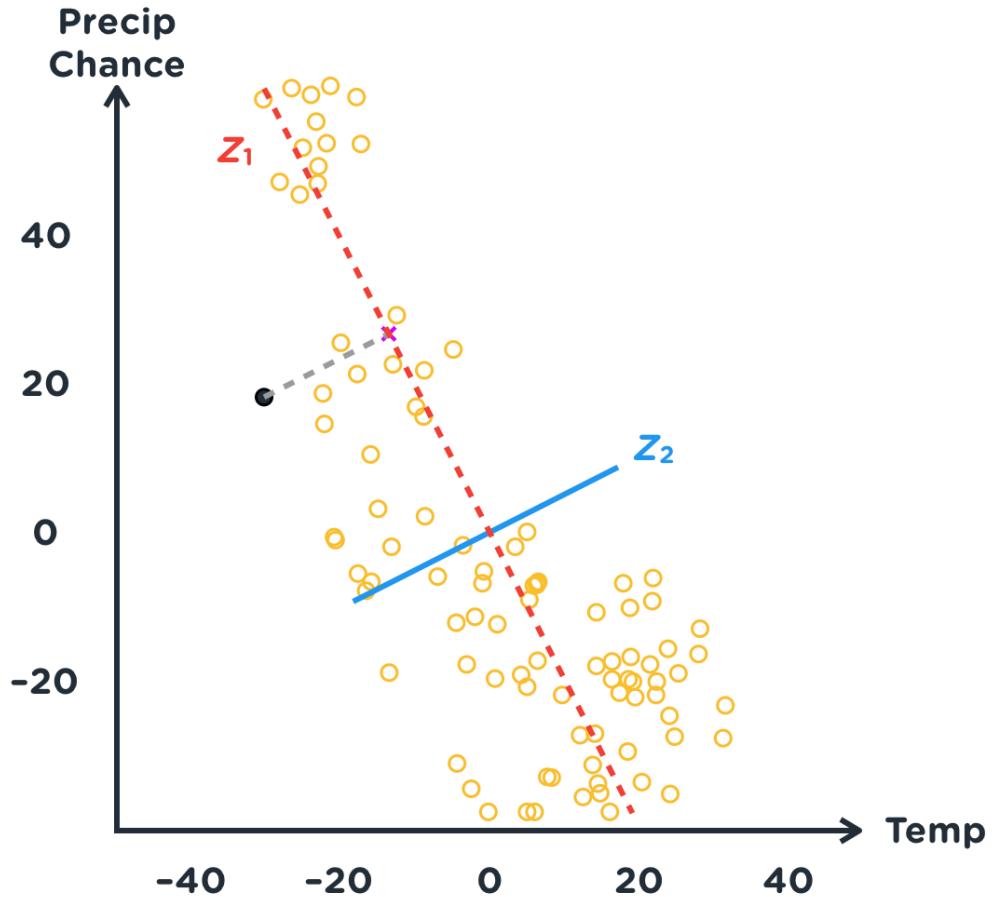
$$\phi_{1,2}^2 + \phi_{2,2}^2 = (-0.8916)^2 + (-0.4528)^2 = 1$$

- The principal components are uncorrelated. This means that the dot product of the loading vectors is zero, i.e.

$$\begin{aligned} \sum_{j=1}^2 \phi_{j,1} \cdot \phi_{j,2} &= \phi_{1,1} \cdot \phi_{1,2} + \phi_{2,1} \cdot \phi_{2,2} \\ &= -0.4528(-0.8916) + 0.8916(-0.4528) \\ &= 0 \end{aligned}$$

To better understand the principal component scores, let's study one observation in the dataset: the 12th observation. It is located towards the left of the scatterplot.

$$(x_{12,1}, x_{12,2}) = (-30.274, 18.154)$$



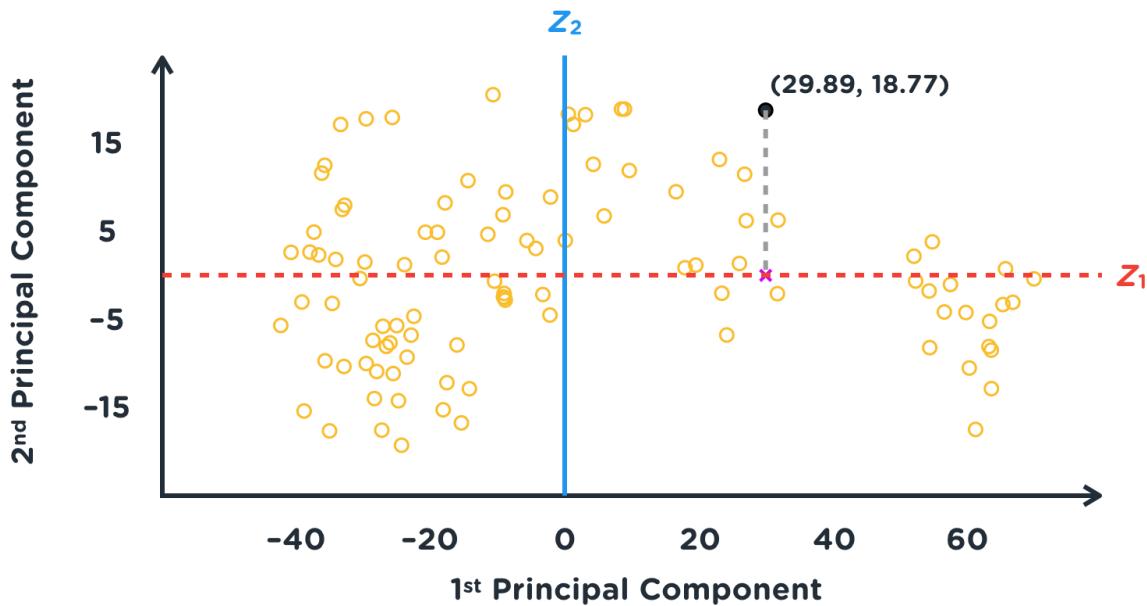
This observation has a first principal component score of

$$\begin{aligned} z_{12,1} &= -0.4528(-30.274) + 0.8916(18.154) \\ &= 29.89 \end{aligned}$$

and a second principal component score of

$$\begin{aligned} z_{12,2} &= -0.8916(-30.274) - 0.4528(18.154) \\ &= 18.77 \end{aligned}$$

To grasp the meaning of these scores, we rotate the scatterplot clockwise until the first and second principal components coincide with the horizontal and the vertical axes, respectively. This results in the following scatterplot where the second principal component scores are plotted against the first principal component scores.

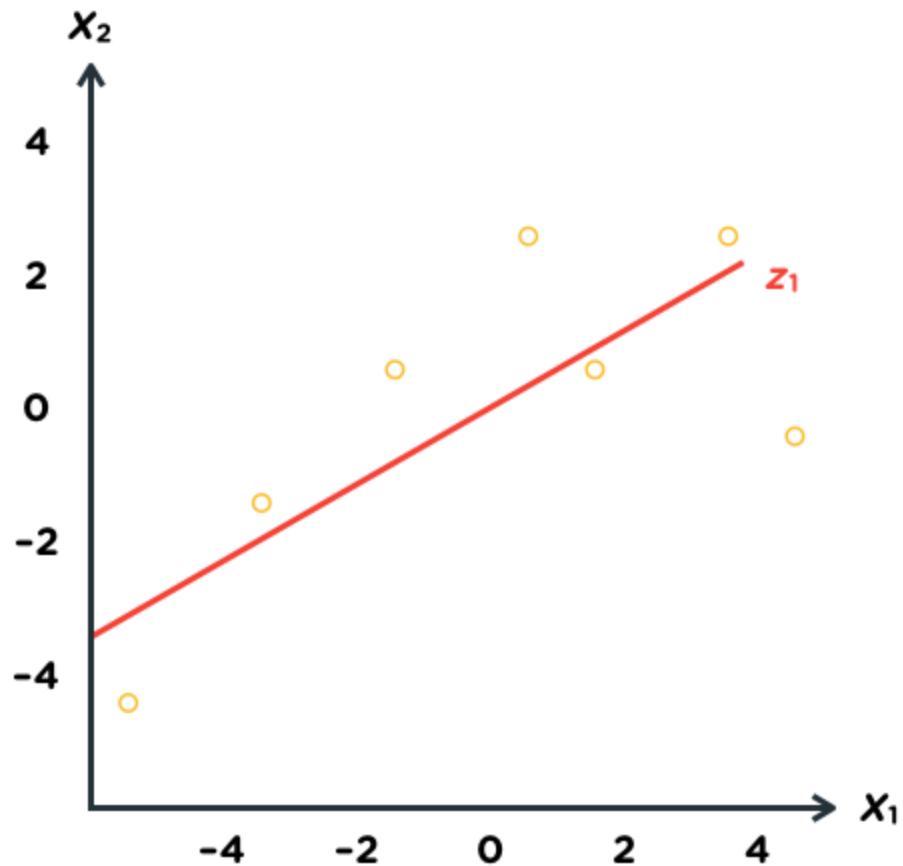


We can see that a first principal component score is the z_1 -direction distance from an observation to the origin. On the other hand, a second principal component score is the z_2 -direction distance from an observation to the origin. Therefore, 29.89 is the z_1 -direction distance from the 12th observation to the origin, and 18.77 is the z_2 -direction distance from the 12th observation to the origin.

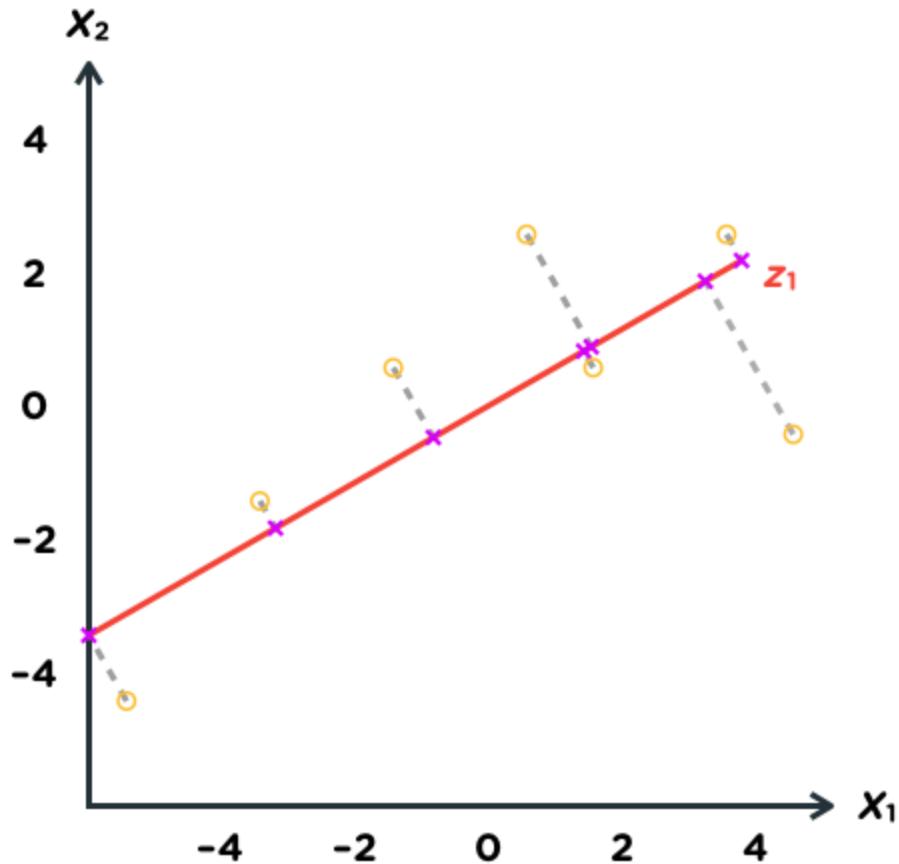
ALTERNATIVE INTERPRETATION

We have viewed principal components as the directions where the **data vary the most**. However, as the observations live in p -dimensional space, we can also interpret principal components as the lower-dimension surface that is **closest** to the observations.

To demonstrate, imagine a dataset with seven observations and two centered variables. So, the observations live in two-dimensional space. In this space, a "lower-dimension surface" can only be a line. The line that is the closest to the seven observations is the one given by the first principal component. The following scatterplot shows the seven observations and a red line representing the first principal component:



For any observation, the shortest distance to the red line is the perpendicular distance. The perpendicular distances are illustrated in gray in the following plot:

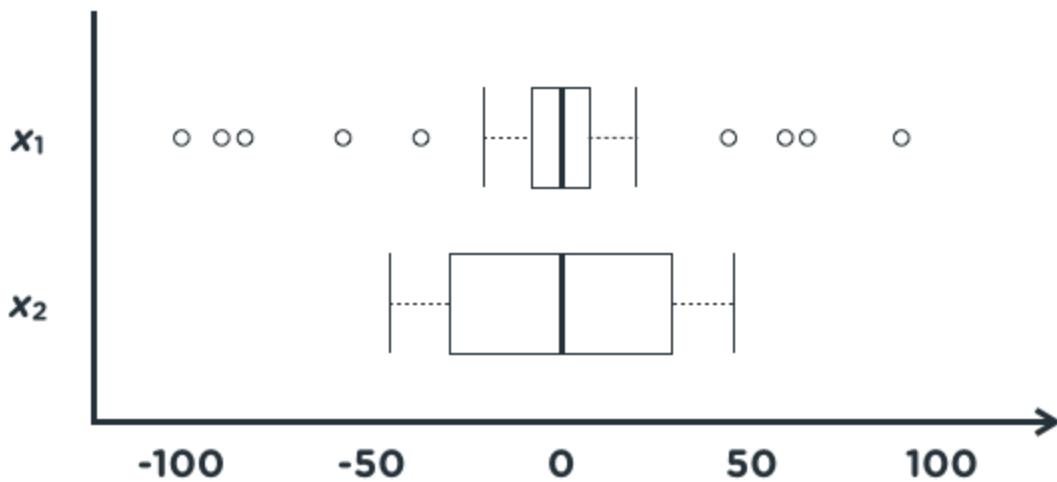


To say the first principal component is the closest to the observations is to say that the gray distances in aggregate is **minimized**. With no other line capable of getting closer to the observations, it is reasonable to consider z_1 as the best one-dimensional summary of the dataset.

STANDARDIZED VARIABLES

PCA is usually performed on a dataset with centered variables. However, we may wish to further scale the variables, which leads to a dataset with standardized variables.

Since the loadings are found by maximizing the sample variance of a principal component, an x_j with a large sample variance would receive a large loading in absolute value. This may not be desirable, since a large sample variance could be due to the scale that x_j is recorded on, as opposed to the "true" variability in x_j . To illustrate, consider this box plot of two variables:



Even though x_1 has a smaller spread than x_2 for most of the data, the sample variance of x_1 is likely to be larger simply because it records values on a wider range. Standardizing (or more generally, scaling) the variables puts them on the same scale. Thus, performing PCA on standardized variables produces loadings that are not swayed by varying scales among the original variables.

Example 3.7.4.1

You are given:

- A principal components analysis is performed on a dataset with two variables and 50 observations.
- The loadings for the first principal component are positive.
- The second principal component loading for the second variable is -0.8.
- For the 7th observation, the recorded values of the first and second variables are 8 and 12, respectively.
- The sample means of the first and second variables are 6.5 and 14.6, respectively.

Determine which of the following statements are true.

- I. The first principal component loading for the first variable is 0.8.
- II. The second principal component loading for the first variable is equal to the first principal component loading for the second variable.
- III. The first principal component score for the 7th observation is -0.36.

Solution

I is true. For this PCA, we have the following equations:

$$\sum_{j=1}^2 \phi_{j,1}^2 = 1$$

$$\sum_{j=1}^2 \phi_{j,2}^2 = 1$$

$$\sum_{j=1}^2 \phi_{j,1} \cdot \phi_{j,2} = 0$$

We are given that $\phi_{2,2} = -0.8$. From the second equation, we can conclude that $\phi_{1,2}$ is either equal to -0.6 or 0.6 .

$$\begin{aligned}\phi_{1,2}^2 + (-0.8)^2 &= 1 \\ \phi_{1,2}^2 &= 0.36 \\ \phi_{1,2} &= \pm 0.6\end{aligned}$$

Since we are given that both $\phi_{1,1}$ and $\phi_{2,1}$ are positive, we need $\phi_{1,2} = 0.6$ for the third equation to be true. If $\phi_{1,2} = -0.6$, we would have two negative terms, which cannot sum to zero:

$$\underbrace{\phi_{1,1} (+0.6)}_{-} + \underbrace{\phi_{2,1} (-0.8)}_{-} \neq 0$$

Then, from the third equation, we have

$$\begin{aligned}\phi_{1,1}(0.6) + \phi_{2,1}(-0.8) &= 0 \\ \phi_{2,1} &= \frac{0.6}{0.8} \phi_{1,1}\end{aligned}$$

Substitute into the first equation to solve for $\phi_{1,1}$:

$$\begin{aligned}\phi_{1,1}^2 + \left(\frac{0.6}{0.8} \phi_{1,1}\right)^2 &= 1 \\ \left(1 + \frac{0.6^2}{0.8^2}\right) \phi_{1,1}^2 &= 1 \\ \phi_{1,1}^2 &= \left(1 + \frac{0.6^2}{0.8^2}\right)^{-1} \\ &= 0.64 \\ \phi_{1,1} &= 0.8\end{aligned}$$

Therefore, the first principal component loading for the first variable is indeed 0.8.

II is true. Using the first equation, solve for $\phi_{2,1}$ as

$$\begin{aligned}0.8^2 + \phi_{2,1}^2 &= 1 \\ \phi_{2,1}^2 &= 0.36 \\ \phi_{2,1} &= 0.6\end{aligned}$$

This means that the second principal component loading for the first variable, $\phi_{1,2}$, and the first principal component loading for the second variable, $\phi_{2,1}$, both equal 0.6.

III is true. The first principal component score for the 7th observation is

$$z_{7,1} = \phi_{1,1} x_{7,1} + \phi_{2,1} x_{7,2}$$

where $x_{7,1}$ and $x_{7,2}$ are the 7th centered values of the first and second variables in the dataset. Consequently,

$$\begin{aligned}z_{7,1} &= 0.8(8 - 6.5) + 0.6(12 - 14.6) \\&= -0.36\end{aligned}$$

Therefore, **all of the statements are true.**



Example 3.7.4.2

For p variables in a dataset, determine which statements about principal components analysis (PCA) are true.

- I. With k principal components such that $k < p$, PCA finds the k -dimensional surface that is closest to the observations in p -dimensional space.
- II. It is possible that the first principal component explains less variability in the dataset than the aggregate variability explained by subsequent principal components.
- III. PCA performs variable selection among the p variables.

Solution

I is true. Principal components form the lower-dimension surface that is closest to the observations in p -dimensional space.

II is true. The first principal component must explain more variability than the second principal component, which must explain more variability than the third principal component, and so on. However, the total variability explained by the second principal component and beyond can exceed the variability explained by the first principal component.

III is false. PCA does not select variables to keep from among the p in the dataset. Instead, it performs dimension reduction, where all p original variables are represented by a few new variables.

Therefore, **only I and II are true.**



Principal Components Regression

Using the principal components as predictors in a multiple linear regression is known as the principal components regression (PCR) approach. Imagine using the first k principal components, z_1, \dots, z_k , in PCR. Then, the model equation is

$$Y = \theta_0 + \theta_1 z_1 + \dots + \theta_k z_k + \varepsilon \quad (3.7.4.4)$$

where $\theta_0, \dots, \theta_k$ are the regression coefficients. These coefficients are estimated using ordinary least squares as previously discussed.

If $k < p$ and the principal components z_1, \dots, z_k do explain most of the dataset's variability, then there is a good chance for PCR to outperform the multiple linear regression with the p original variables as predictors – the loss of information should be minor, and fewer coefficients would need to be estimated. However, if all principal components are used in PCR, i.e. $k = p$, then dimension reduction does not occur. In other words, the result would be equivalent to performing the multiple linear regression with the p original variables as predictors.

The ideal value of k can be determined through cross-validation; the number of principal components used in PCR is a flexibility measure.

COMMUTING CHRIS EXAMPLE

In this scenario, we perform PCA on four Commuting Chris variables: Departure, Temp, Precip Chance, and Police. Let's begin by using the first two principal components to predict Commute. The table below shows the coefficient estimates for this PCR:

	Estimate
(Intercept)	26.3875
PC1	1.7107
PC2	2.7434

Therefore, the fitted equation is

$$\hat{y} = 26.3875 + 1.7107z_1 + 2.7434z_2$$

Next, consider using all four principal components to predict Commute. This PCR produces the following result:

	Estimate
(Intercept)	26.3875
PC1	1.7107
PC2	2.7434
PC3	3.4323
PC4	-1.8602

Clearly, the fitted equation is now

$$\hat{y} = 26.3875 + 1.7107z_1 + 2.7434z_2 + 3.4323z_3 - 1.8602z_4$$

Moreover, there are several important details worth mentioning:

- $\hat{\theta}_0 = \bar{y}$, which is a consequence of the principal components each having a sample mean of 0.
- Notice $\hat{\theta}_1$ and $\hat{\theta}_2$ are unchanged, whether two or four principal components are used in PCR. This holds true for $\hat{\theta}_1, \dots, \hat{\theta}_k$ in general, since principal components are orthogonal variables (revisit Section 3.5.5 for details).
- The second fitted equation must be equivalent to the fitted equation given in Section 3.3.2, where the original predictors were used. This is because using all four principal components in the second PCR does not result in dimension reduction.

Example 3.7.4.3

You are given:

- An ordinary least squares regression is performed on a dataset with three predictors; the results are

	Estimate
(Intercept)	2.939
x_1	0.046

	Estimate
x_2	0.189
x_3	-0.001

- One of the observations is

y	x_1	x_2	x_3
10.4	44.5	39.3	45.1

- A principal components analysis is performed on the three predictors. All three principal components are then used in a principal components regression.

Predict the response for the observation under the principal components regression model.

Solution

Since all three principal components are used as predictors, PCR is equivalent to performing the ordinary least squares regression on the original predictors. Therefore, the predicted response for the given observation is

$$\begin{aligned}\hat{y} &= 2.939 + 0.046(44.5) + 0.189(39.3) - 0.001(45.1) \\ &= \mathbf{12.3686}\end{aligned}$$



Example 3.7.4.4

A number of multiple linear regressions are performed using a dataset that has four continuous variables without counting the response. Some of the regressions use the dataset's principal components as predictors.

Let x_j be the j^{th} variable in the dataset and z_m be the m^{th} principal component. The table below shows the predicted value of the sixth observation under the following regression models:

Model	Predictors	\hat{y}_6
I	z_1	27.065
II	z_1, z_2	23.271
III	z_1, z_2, z_3	21.208
IV	x_1, x_2, x_3, x_4	21.014

The principal component scores for the sixth observation are given as follows:

m	$z_{6,m}$
1	0.396
2	-1.383
3	-0.601
4	0.104

Calculate the estimate for the regression coefficient of z_2 under Model III.

Solution

Let $\hat{\theta}_2$ be the estimate of z_2 's coefficient. Since principal components are orthogonal variables, $\hat{\theta}_2$ is the same under Model II and Model III.

Notice that under Model I,

$$27.065 = \hat{\theta}_0 + \hat{\theta}_1 z_{6,1}$$

while under Model II,

$$23.271 = \hat{\theta}_0 + \hat{\theta}_1 z_{6,1} + \hat{\theta}_2 z_{6,2}$$

The values of $\hat{\theta}_0$ and $\hat{\theta}_1$ are the same between Model I and Model II. Therefore, we solve for the answer as follows:

$$\begin{aligned} 23.271 &= \hat{\theta}_0 + \hat{\theta}_1 z_{6,1} + \hat{\theta}_2 z_{6,2} \\ 23.271 &= 27.065 + \hat{\theta}_2 (-1.383) \\ \hat{\theta}_2 &= \frac{23.271 - 27.065}{-1.383} \\ &= \mathbf{2.7433} \end{aligned}$$



3.7.5 Partial Least Squares

Similar to principal components, the method of **partial least squares (PLS)** creates new predictors by taking linear combinations of the recorded variables. These new predictors are called **directions**.

Denote the m^{th} partial least squares direction as z_m . Upon finding k of these directions, we perform a multiple linear regression with z_1, \dots, z_k as predictors of the response.

The partial least squares directions differ from principal components in two main ways:

1. PLS directions are created using the response variable; they summarize the original predictors in a supervised way.
2. PLS directions are solely used as predictors; they have no other purpose. However, principal components are predictors only in the context of PCR; they can also be used in non-regression contexts.

In this subsection, we define x_1, \dots, x_p as the **standardized** variables of the original.

First PLS Direction

The first partial least squares direction is calculated as

$$z_1 = \sum_{j=1}^p \phi_j x_j$$

where ϕ_j is set to equal the OLS slope estimate of y regressed on x_j . Intuitively, the original predictors that are more strongly correlated to the response will contribute more in calculating z_1 .

Second and Subsequent PLS Directions

Similar to the z_1 formula,

$$z_2 = \sum_{j=1}^p \phi_j^* x_j^*$$

where x_j^* is the **residual** of the linear regression

- without an intercept term,
- with x_j as the response variable,
- with z_1 as the explanatory variable, and
- using ordinary least squares,

and ϕ_j^* is set to equal the OLS slope estimate of y regressed on x_j^* .

The x_1^*, \dots, x_p^* can be interpreted as the portion of x_1, \dots, x_p that was not captured by z_1 .

To generalize, calculating a partial least squares direction involves using the previous direction to "update the previous predictors". All subsequent partial least squares directions are computed iteratively by

- fitting models with the "previous predictors" explained by the previous direction, then
- using the residuals of those fits as the "current predictors" to calculate the current direction.

Coach's Remarks

Our manual follows the description in the textbook by James et al. for computing the linear combination coefficients (ϕ_j 's and ϕ_j^* 's). However, the directions produced by this description do not appear to be correct, such as not being orthogonal.

We believe this exam will not emphasize the specifics of the partial least squares procedure, but instead, focus on the ideas presented in the concluding paragraphs that follow.

Understanding PLS

Besides the aforementioned differences between principal components and PLS directions, the two concepts share many similarities. The PLS directions form a set of orthogonal variables. In addition, partial least squares performs dimension reduction as well; each direction summarizes all p original predictors. When $k = p$, no dimension reduction occurs, which leads to the same result as the regression of y on x_1, \dots, x_p . Furthermore, k is a flexibility measure whose ideal value can be determined via cross-validation.

As seen in the linear combination coefficients, the PLS directions are constructed based on information from the response. This is the key distinction between PLS directions and principal

components, where the latter is unsupervised in its construction. Hence, a drawback to PCR is that the principal components have no concrete reason to be good predictors of the response, even if they summarize the dataset well.

Example 3.7.5.1

Determine which of the following statements are true.

- I. For partial least squares with g directions, the regression's variance decreases as g decreases.
- II. Partial least squares may improve over ordinary least squares in the sense that it is not as biased.
- III. The partial least squares directions are a supervised, low-dimensional representation of the original features.

Solution

I is true. The number of directions used in partial least squares corresponds to flexibility, and a lower flexibility leads to lower variance.

II is false. Partial least squares may improve over ordinary least squares by reducing dimensions and thus reducing variance. By the bias-variance trade-off, this means partial least squares is more biased than ordinary least squares.

III is true. The directions reduce the features to a lower dimension in a way that makes use of the response variable.

Therefore, **only I and III are true.**



3.7 Summary

Shrinkage Methods

RIDGE REGRESSION

- The regression coefficients are estimated by minimizing the SSE while constrained by

$$\sum_{j=1}^p \hat{\beta}_j^2 \leq a$$

or equivalently, by minimizing the expression

$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_p x_{i,p} \right)^2 + \lambda \sum_{j=1}^p \hat{\beta}_j^2$$

with scaled predictors x_1, \dots, x_p .

- The ℓ_2 norm of $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$ is

$$\|\hat{\beta}\|_2 = \sqrt{\sum_{j=1}^p \hat{\beta}_j^2}$$

- λ is inversely related to flexibility. With a finite λ , none of the ridge estimates will equal 0.

LASSO REGRESSION

- The regression coefficients are estimated by minimizing the SSE while constrained by

$$\sum_{j=1}^p |\hat{\beta}_j| \leq a$$

or equivalently, by minimizing the expression

$$\sum_{i=1}^n \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i,1} - \dots - \hat{\beta}_p x_{i,p} \right)^2 + \lambda \sum_{j=1}^p |\hat{\beta}_j|$$

with scaled predictors x_1, \dots, x_p .

- The ℓ_1 norm of $\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \vdots \\ \hat{\beta}_p \end{bmatrix}$ is

$$\|\hat{\beta}\|_1 = \sum_{j=1}^p |\hat{\beta}_j|$$

- λ is inversely related to flexibility. With a finite λ , the lasso estimates could equal 0.

Principal Components

- PCA is an unsupervised learning technique that performs dimension reduction on p variables.
- The m^{th} principal component is a linear combination of the p centered variables:

$$z_m = \sum_{j=1}^p \phi_{j,m} x_j$$

subject to $\sum_{j=1}^p \phi_{j,m}^2 = 1$.

- All principal components are uncorrelated/orthogonal to one another, i.e.

$$\sum_{j=1}^p \phi_{j,m} \cdot \phi_{j,u} = 0$$

for all pairs of m and u where $m \neq u$.

- The variability explained by each subsequent principal component is always less than the variability explained by its previous principal component.
- Principal components form the lower dimension surface that is closest to the observations in p -dimensional space.
- Standardized variables affect the loadings by becoming resistant to varying scales among the original variables.

PRINCIPAL COMPONENTS REGRESSION

- PCR uses the first k principal components as predictors in a multiple linear regression.

$$Y = \theta_0 + \theta_1 z_1 + \dots + \theta_k z_k + \varepsilon$$

where $\theta_0, \dots, \theta_k$ are the regression coefficients.

- The number of principal components used, k , is a measure of flexibility. When $k = p$, PCR is equivalent to performing the multiple linear regression with the p original variables as predictors.

Partial Least Squares

- Partial least squares performs dimension reduction on p predictors in a supervised way.
- The PLS directions z_1, \dots, z_k are orthogonal and used as predictors in a multiple linear regression.
- The number of directions, k , is a measure of flexibility. When $k = p$, PLS is equivalent to performing the multiple linear regression with the p original predictors.
- The first PLS direction is a linear combination of the p standardized predictors, with coefficients that are based on the response y .
- Every subsequent PLS direction is calculated iteratively as a linear combination of "updated predictors" which are the residuals of fits with the "previous predictors" explained by the

previous direction.

3.8.0 Overview

 5m

Some linear regressions hold several strong assumptions, such as the response variable being normally distributed. **Generalized linear models (GLM)** have more relaxed assumptions, thus allowing for various possibilities that are all under a common umbrella. The commonality rests heavily on the exponential class of distributions.

We will discuss the main attributes of GLM and see how parameter estimation and statistical inference are performed. In addition, we will learn of important numerical outputs to better understand model fit. We then finish by looking at a special sub-family of distributions known as the Tweedie distributions.

3.8.1 Exponential Family

The exponential class or family of distributions was introduced in Section 2.2.5. Recall that a distribution belongs to this family if its probability function can be written in the form of Equation 2.2.5.1:

$$f(y) = \exp [a(y) \cdot b(\theta) + c(\theta) + d(y)] \quad (2.2.5.1)$$

In modeling the response variable Y , we are expected to understand specifics regarding its distribution. The probability function implies that the distribution of Y has only one parameter of interest, denoted generically as θ . All other parameters are assumed to be known/fixed; these are called ***nuisance parameters***.

In addition, the exponential family has several useful results:

$$\mathbb{E}[a(Y)] = -\frac{c'(\theta)}{b'(\theta)} \quad (3.8.1.1)$$

$$\text{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \quad (3.8.1.2)$$

The derivations can be found in the appendix at the end of the section.

In the GLM context, we focus on the situation where $a(y) = y$. The distribution is then said to be in ***canonical form***, and $b(\theta)$ is called the ***natural parameter*** of the distribution.

In linear regression, recall that the mean response is central in formulating a model. Likewise, GLM places a large emphasis on the mean as well. Therefore, two crucial takeaways are:

1. **The mean of Y can be written as a function of the parameter of interest** – there is a structured relation between the mean and θ .
2. **The variance of Y can be written as a function of the mean of Y** – there is a structured relation between the mean and variance.

The first result can be seen from Equation 3.8.1.1; the expression $-c'(\theta)/b'(\theta)$ is ultimately a function of θ . To prove the second result, notice that the variance is another function of θ from Equation 3.8.1.2. Then, given the connection between the mean and θ from the first result, it follows that any function of θ must also be a function of the mean.

The following table summarizes the relevant information regarding the distributions that are expected on the exam. Since these are straightforward to derive using Equation 2.2.5.1, we believe it is not necessary to memorize this table. However, you may do so to save yourself some time.

Distribution	θ	Natural Parameter, $b(\theta)$	$c(\theta)$
Binomial (fixed m)	q	$\ln\left(\frac{q}{1-q}\right)$	$m \ln(1-q)$
Normal (fixed σ^2)	μ	$\frac{\mu}{\sigma^2}$	$-\frac{\mu^2}{2\sigma^2}$
Poisson	λ	$\ln \lambda$	$-\lambda$
Gamma (fixed α)	θ	$-\frac{1}{\theta}$	$-\alpha \ln \theta$
Inverse Gaussian (fixed θ)	μ	$-\frac{\theta}{2\mu^2}$	$\frac{\theta}{\mu}$
Negative binomial (fixed r)	β	$\ln\left(\frac{\beta}{1+\beta}\right)$	$-r \ln(1+\beta)$

Show that the exponential distribution belongs to the exponential family, and that Equations 3.8.1.1 and 3.8.1.2 hold.

$$\begin{aligned}
 f(y) &= \frac{e^{-y/\theta}}{\theta} \\
 &= \exp\left[-\frac{y}{\theta}\right] \cdot \exp\left[\ln\left(\frac{1}{\theta}\right)\right] \\
 &= \exp\left[-\frac{y}{\theta} - \ln \theta\right] \\
 &= \exp\left[y\left(-\frac{1}{\theta}\right) - \ln \theta\right]
 \end{aligned}$$

In this form, note that

- $a(y) = y$
- $b(\theta) = -\frac{1}{\theta}$

- $c(\theta) = -\ln \theta$
- $d(y) = 0$

which proves that **the exponential distribution belongs to the exponential family**. Since a gamma distribution with $\alpha = 1$ is an exponential distribution, notice that we match the expressions of $b(\theta)$ and $c(\theta)$ provided in the table.

Next, solve for the necessary components in Equations 3.8.1.1 and 3.8.1.2.

$$\begin{aligned} b'(\theta) &= \frac{d}{d\theta} \left(-\frac{1}{\theta} \right) \\ &= -1 \cdot -\theta^{-2} \\ &= \theta^{-2} \end{aligned}$$

$$\begin{aligned} b''(\theta) &= \frac{d}{d\theta} \theta^{-2} \\ &= -2 \cdot \theta^{-3} \\ &= -2\theta^{-3} \end{aligned}$$

$$\begin{aligned} c'(\theta) &= \frac{d}{d\theta} (-\ln \theta) \\ &= -\frac{1}{\theta} \end{aligned}$$

$$\begin{aligned} c''(\theta) &= \frac{d}{d\theta} \left(-\frac{1}{\theta} \right) \\ &= \theta^{-2} \end{aligned}$$

Since the exponential distribution has mean θ and variance θ^2 , the two equations hold.

$$\begin{aligned}\mathbb{E}[Y] &= -\frac{c'(\theta)}{b'(\theta)} \\ &= -\frac{1}{-\frac{\theta}{\theta^{-2}}} \\ &= \theta\end{aligned}$$

$$\begin{aligned}\text{Var}[Y] &= \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3} \\ &= \frac{(-2\theta^{-3})\left(-\frac{1}{\theta}\right) - (\theta^{-2})(\theta^{-2})}{[\theta^{-2}]^3} \\ &= \frac{2\theta^{-4} - \theta^{-4}}{\theta^{-6}} \\ &= \theta^2\end{aligned}$$

3.8.2 Model Framework

As review, consider the following points regarding multiple linear regression:

- Y_i is normally distributed
- $E[Y_i]$ is allowed to depend on observation i
- $\mathbf{x}_i^T \boldsymbol{\beta} = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p}$ (also called **linear component**) is set equal to the mean
- $\text{Var}[Y_i]$ is constant

The essence of GLM is captured by these parallel concepts:

- Y_i has a distribution that belongs to the exponential family
- The mean μ_i (as a function of θ_i) is allowed to depend on observation i
- $\mathbf{x}_i^T \boldsymbol{\beta}$ is set equal to some function of the mean, $g(\mu_i)$
- The variance (as a function of μ_i) is allowed to depend on observation i

Note that while GLM allows for heteroscedasticity, it is not a requirement. For example, when the response is modeled as normally distributed, we have the option to let the variance be constant across the observations, resulting in homoscedasticity. On the other hand, this is usually not possible if a different distribution is used.

In summary, modeling a GLM requires making suitable choices for:

1. The random component – choice of distribution
2. The systematic component – choice of $g(\cdot)$

Choice of Distribution

The main criterion is to select a distribution with a domain that is compatible with the possible values of the response. For example, if the response is a count variable, then modeling with a continuous distribution such as normal is not appropriate.

Choice of Link Function

By default, a **link function**, $g(\cdot)$, is a function of the mean response that we model to equal $\mathbf{x}^T \boldsymbol{\beta}$. Therefore, it establishes how the mean response is linked with the predictors we wish to include. In

addition, $g(\cdot)$ must be a monotone and differentiable function.

Several well-known link functions are:

Function Name	$g(\mu)$
Identity link	μ
Logit link	$\ln\left(\frac{\mu}{1-\mu}\right)$
Logarithmic link	$\ln \mu$
Inverse link	$\frac{1}{\mu}$
Power link	μ^d

A key purpose of a link function is to account for the range of possible values for the mean. For a normal distribution, μ takes on any real number. The identity link (i.e. $\mu = \mathbf{x}^T \boldsymbol{\beta}$) works well in this case because $\mathbf{x}^T \boldsymbol{\beta}$ may also take on any real number; there are no restrictions on the values of the coefficient estimates.

On the other hand, the mean of a Bernoulli distribution is the probability of a "success"; μ must now be valid between 0 and 1. Using the identity link implies that $\mathbf{x}^T \boldsymbol{\beta}$ should also be between 0 and 1. In turn, the estimation procedure ought to comply with this restriction. Instead of using a restrictive procedure with the identity link, we may choose to use the logit link instead (i.e. $\ln\left(\frac{\mu}{1-\mu}\right) = \mathbf{x}^T \boldsymbol{\beta}$). With an input of μ between 0 and 1, the logit link outputs a real number. Then, no special requirements are needed in estimating the coefficients.

Link functions also serve a purpose with model interpretation. Note that

$$\mu = g^{-1}(\mathbf{x}^T \boldsymbol{\beta})$$

Therefore, it is desirable to use a $g(\cdot)$ such that $g^{-1}(\mathbf{x}^T \boldsymbol{\beta})$ is not overly complicated. A reasonably simple expression makes it easier to explain how the predictors relate to the mean response. We will see some concrete examples in Sections 3.9 and 3.10.

Example 3.8.2.1

Determine which of the following statements is true.

- I. It is appropriate to use the logit link function with an exponentially distributed response variable.

- II. To model insurance claim amounts using GLM, the gamma distribution is an unsuitable choice of distribution.
- III. From the modeling perspective, multiple linear regression is a special case of a GLM.

Solution

I is false. The mean of an exponential distribution can be any positive number. However, the logit link function only works for inputs valid between 0 and 1.

II is false. Claim amounts are continuous, positive quantities; they are compatible with gamma's domain. In addition, gamma belongs to the exponential family. There is no clear reason why the gamma distribution would be unsuitable for modeling claim amounts without additional context.

III is true. A GLM with a normally distributed response, the identity link function, and homoscedasticity is the same model as proposed by multiple linear regression.

Therefore, **only III is true.**



CANONICAL LINKS

Since any function of θ must also be a function of μ , realize that the natural parameter $b(\theta)$ is a function of μ .

Keep in mind that link functions are functions of μ . Thus, the specific function of μ that equals the natural parameter is called a **canonical link**. Consequently, a canonical link sets the natural parameter equal to $\mathbf{x}^T \boldsymbol{\beta}$.

As an example, consider the fact the Poisson distribution has a natural parameter of $b(\lambda) = \ln \lambda$. Since $\lambda = \mu$ in this scenario, the canonical link for a Poisson distribution is $\ln \mu$, which happens to be the logarithmic link.

The study note by Larsen provides the following table of distributions and their canonical links:

Distribution	Canonical Link
Normal	Identity link
Binomial	Logit link

Distribution	Canonical Link
Poisson	Logarithmic link
Gamma	Inverse link

Coach's Remarks

On a technical note, the table hints at the fact that link functions do not have very strict mathematical forms. To illustrate, the gamma distribution's natural parameter is $-\theta^{-1}$; when expressed as a function of the mean response, it equals $-\alpha\mu^{-1}$. Keep in mind that α is fixed in this scenario.

In other words, $-\alpha\mu^{-1}$ and μ^{-1} are both considered to be inverse link functions. This is acceptable because link functions are the same up to a multiplicative constant (which is $-\alpha$ in this case). Keeping or dropping the constant has no meaningful effect once the link function is set equal to the linear component.

Having said that, we disagree that the canonical link of a binomial distribution is the logit link; it is the canonical link of a Bernoulli distribution. Currently, we believe that exam questions will take the Larsen study note as correct.

For this exam, what you should know about canonical links is the table from the Larsen study note. Memorizing it is sufficient, but you may find it helpful to connect canonical links with natural parameters.

3.8.3 Parameter Estimation

GLM estimates β using maximum likelihood estimation. With independent Y_i 's, we wish to find the values of β that maximize the log-likelihood function:

$$\begin{aligned}
 l(\beta) &= \ln [L(\beta)] \\
 &= \ln \left\{ \prod_{i=1}^n \exp [y_i \cdot b(\theta_i) + c(\theta_i) + d(y_i)] \right\} \\
 &= \sum_{i=1}^n \ln \{\exp [y_i \cdot b(\theta_i) + c(\theta_i) + d(y_i)]\} \\
 &= \sum_{i=1}^n [y_i \cdot b(\theta_i) + c(\theta_i) + d(y_i)]
 \end{aligned} \tag{3.8.3.1}$$

It is important to remember that

- β is related to μ_i (via the link function), and
- μ_i is related to θ_i ,

meaning the model has β embedded in the θ_i 's.

Recall that for MLE we take partial derivatives of $l(\beta)$ with respect to each β_j , then set them equal to 0 to obtain the score equations. The solution to the score equations is the MLE estimates, $\hat{\beta}$. In many cases, $\hat{\beta}$ is not obtained analytically (i.e. using formulas), but rather, by a numerical algorithm that searches for the solution iteratively. In contrast, an analytical solution to the score equations exists under a multiple linear regression setup; it is the same as the ordinary least squares estimates.

Upon obtaining the MLE estimates, we let

$$\hat{\mu} = g^{-1}(\mathbf{x}^T \hat{\beta}) \tag{3.8.3.2}$$

to mirror the \hat{y} notation from linear regression. $\hat{\mu}$ is unbiased when the canonical link is used, but is biased otherwise.

Method of Scoring

The **method of scoring**, or **Fisher scoring**, is a numerical algorithm that is typically used in GLM to obtain the MLE estimates. It is based on the Newton-Raphson approximation — for a given function, it iteratively finds a point where the function equals 0; this is achieved with the function's derivative and an initial guess of the answer.

Let's begin by defining notation:

- u_j is the score function for β_j , i.e.

$$u_j = \frac{\partial}{\partial \beta_j} l(\boldsymbol{\beta}), \quad j = 0, 1, \dots, p$$

- \mathbf{u} is the vector of the u_j 's, i.e.

$$\mathbf{u} = \begin{bmatrix} u_0 \\ u_1 \\ \vdots \\ u_p \end{bmatrix}$$

- \mathbf{I} is the information matrix of $\boldsymbol{\beta}$.
- $\hat{\boldsymbol{\beta}}^{(m)}$ is the estimate of $\boldsymbol{\beta}$ at the end of the m^{th} iteration of the algorithm.
- $u_j^{(m)}$ is u_j after evaluating $\boldsymbol{\beta}$ at $\hat{\boldsymbol{\beta}}^{(m)}$.
- $\mathbf{u}^{(m)}$ is the vector of the $u_j^{(m)}$'s.
- $\mathbf{I}^{(m)}$ is \mathbf{I} after evaluating $\boldsymbol{\beta}$ at $\hat{\boldsymbol{\beta}}^{(m)}$.

With $l(\boldsymbol{\beta})$ given by Equation 3.8.3.1, the score functions become

$$u_j = \sum_{i=1}^n \frac{(y_i - \mu_i)x_{i,j}}{\text{Var}[Y_i] \cdot g'(\mu_i)} \tag{3.8.3.3}$$

and the $(j+1)^{\text{st}}$ row, $(j^*+1)^{\text{st}}$ column entry of \mathbf{I} is

$$\sum_{i=1}^n \frac{x_{i,j}x_{i,j^*}}{\text{Var}[Y_i] \cdot g'(\mu_i)^2}$$

for j and j^* taking on integers from 0 to p . Hence,

$$\mathbf{I} = \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^T}{\text{Var}[Y_i] \cdot g'(\mu_i)^2}$$

$$= \begin{bmatrix} \sum_{i=1}^n \frac{1}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} & \sum_{i=1}^n \frac{x_{i,1}}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} & \dots & \sum_{i=1}^n \frac{x_{i,p}}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} \\ \sum_{i=1}^n \frac{x_{i,1}}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} & \sum_{i=1}^n \frac{x_{i,1}^2}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} & \dots & \sum_{i=1}^n \frac{x_{i,1} x_{i,p}}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n \frac{x_{i,p}}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} & \sum_{i=1}^n \frac{x_{i,1} x_{i,p}}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} & \dots & \sum_{i=1}^n \frac{x_{i,p}^2}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} \end{bmatrix} \quad (3.8.3.4)$$

In summary,

- u_j consists of three components – \mathbf{y} , \mathbf{X} , and β .
- \mathbf{I} consists of two components – \mathbf{X} and β .

Then to compute $\hat{\beta}^{(m)}$, we use the formula:

$$\hat{\beta}^{(m)} = \hat{\beta}^{(m-1)} + [\mathbf{I}^{(m-1)}]^{-1} \mathbf{u}^{(m-1)} \quad (3.8.3.5)$$

where $[\mathbf{I}^{(m-1)}]^{-1}$ is the inverse of the matrix $\mathbf{I}^{(m-1)}$.

Thus, the method of scoring begins with an initial guess for the MLE estimates, which we denote as $\hat{\beta}^{(0)}$. This guess is used to compute $[\mathbf{I}^{(0)}]^{-1}$ and $\mathbf{u}^{(0)}$. Next, we use Equation 3.8.3.5 to improve on our initial guess by calculating $\hat{\beta}^{(1)}$. Because \mathbf{I} comes from taking derivatives of \mathbf{u} , the formula detects how close $\hat{\beta}^{(0)}$ is to satisfying the score equations $\mathbf{u} = \mathbf{0}$, and gets closer with $\hat{\beta}^{(1)}$. This concludes the first iteration. The algorithm continues until $\hat{\beta}^{(m)}$ converges, i.e. barely changes from the previous iteration. The converged values are taken to be the MLE estimates, $\hat{\beta}$.

Coach's Remarks

Other resources may denote the start of the numerical algorithm with $m = 1$ rather than $m = 0$. Doing so does not affect the algorithm itself.

Equation 3.8.3.5 can be restated in another useful form. To motivate it, first recall the formula for $\hat{\beta}$ under multiple linear regression (i.e. ordinary least squares):

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

Now imagine keeping all the assumptions of multiple linear regression except for homoscedasticity. We instead assume

$$\text{Var}[\varepsilon_i] = \frac{\sigma^2}{w_i}$$

where w_i is a predetermined weight for the i^{th} observation. This model is called **weighted least squares**. The formula for $\hat{\beta}$ is now

$$\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}$$

where \mathbf{W} is an $n \times n$ diagonal matrix (i.e. a matrix of 0's except in the diagonal) with the weights w_1, \dots, w_n as its entries.

Returning to GLM, let's further define:

$$w_i = \frac{1}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} \quad (3.8.3.6)$$

$$\mathbf{W} = \begin{bmatrix} w_1 & 0 & \dots & 0 \\ 0 & w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_n \end{bmatrix}$$

$$z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i) \quad (3.8.3.7)$$

$$\mathbf{z} = \begin{bmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{bmatrix}$$

Then, the alternate form of Equation 3.8.3.5 is

$$\hat{\beta}^{(m)} = \left(\mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)} \quad (3.8.3.8)$$

Hence, the method of scoring for GLM is an **iterative weighted least squares** procedure.

The next example demonstrates one iteration of the algorithm while solving for every necessary component. In contrast, exam problems are likely to provide summary expressions and/or values to make solving less tedious.

Example 3.8.3.1

You model a GLM with a Poisson response and a link function of

$$g(\lambda) = \ln \lambda = \beta_0 + \beta_1 x$$

for the following dataset:

Observation, i	Response, y_i	Feature, x_i
1	0	10
2	1	9
3	3	5

You use Fisher scoring to find the maximum likelihood estimate of $\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}$.

With starting values of $\begin{bmatrix} 3.0 \\ -0.2 \end{bmatrix}$, calculate the estimates after one iteration of the Fisher scoring algorithm.

Solution

Given that $\hat{\beta}^{(0)} = \begin{bmatrix} 3.0 \\ -0.2 \end{bmatrix}$, the goal is to solve for

$$\hat{\beta}^{(1)} = \hat{\beta}^{(0)} + \left[\mathbf{I}^{(0)} \right]^{-1} \mathbf{u}^{(0)}$$

Because $\lambda = \mu$ for Poisson, $g(\lambda) = g(\mu) = \ln \mu$. Start by finding $g'(\mu_i)$ and $\text{Var}[Y_i]$ in terms of the linear component. This requires inverting the link function to solve for μ_i .

$$\beta_0 + \beta_1 x_i = \mathbf{x}_i^T \boldsymbol{\beta}$$

$$\begin{aligned}\mu_i &= g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}) \\ &= \exp(\mathbf{x}_i^T \boldsymbol{\beta})\end{aligned}$$

$$\begin{aligned}g'(\mu_i) &= \frac{d}{d\mu_i} \ln \mu_i \\ &= \frac{1}{\mu_i} \\ &= \frac{1}{\exp(\mathbf{x}_i^T \boldsymbol{\beta})} \\ &= \exp(-\mathbf{x}_i^T \boldsymbol{\beta})\end{aligned}$$

$$\begin{aligned}\text{Var}[Y_i] &= \mu_i \\ &= \exp(\mathbf{x}_i^T \boldsymbol{\beta})\end{aligned}$$

Next, use Equation 3.8.3.3 to solve for the score equations.

$$\begin{aligned}u_0 &= \sum_{i=1}^3 \frac{(y_i - \mu_i) \overbrace{1}^{x_{i,0}}}{\text{Var}[Y_i] \cdot g'(\mu_i)} \\ &= \sum_{i=1}^3 \frac{y_i - \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\exp(\mathbf{x}_i^T \boldsymbol{\beta}) \cdot \exp(-\mathbf{x}_i^T \boldsymbol{\beta})} \\ &= \sum_{i=1}^3 y_i - \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \\ &= \sum_{i=1}^3 y_i - \sum_{i=1}^3 \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \\ &= \sum_{i=1}^3 y_i - \sum_{i=1}^3 \exp(\beta_0) \cdot \exp(\beta_1 x_i) \\ &= \left(\sum_{i=1}^3 y_i \right) - \exp(\beta_0) \left[\sum_{i=1}^3 \exp(\beta_1 x_i) \right]\end{aligned}$$

$$\begin{aligned}
u_1 &= \sum_{i=1}^3 \frac{(y_i - \mu_i)x_i}{\text{Var}[Y_i] \cdot g'(\mu_i)} \\
&= \sum_{i=1}^3 \frac{x_i y_i - x_i \exp(\mathbf{x}_i^T \boldsymbol{\beta})}{\exp(\mathbf{x}_i^T \boldsymbol{\beta}) \cdot \exp(-\mathbf{x}_i^T \boldsymbol{\beta})} \\
&= \sum_{i=1}^3 x_i y_i - \sum_{i=1}^3 x_i \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \\
&= \left(\sum_{i=1}^3 x_i y_i \right) - \exp(\beta_0) \left[\sum_{i=1}^3 x_i \exp(\beta_1 x_i) \right]
\end{aligned}$$

Thus, obtain $\mathbf{u}^{(0)}$ as follows:

$$\begin{aligned}
u_0^{(0)} &= (0 + 1 + 3) - e^3 \left[e^{-0.2(10)} + e^{-0.2(9)} + e^{-0.2(5)} \right] \\
&= -9.4275
\end{aligned}$$

$$\begin{aligned}
u_1^{(0)} &= [10(0) + 9(1) + 5(3)] - e^3 \left[10e^{-0.2(10)} + 9e^{-0.2(9)} + 5e^{-0.2(5)} \right] \\
&= -70.0092
\end{aligned}$$

$$\mathbf{u}^{(0)} = \begin{bmatrix} -9.4275 \\ -70.0092 \end{bmatrix}$$

Proceed to find the information matrix using Equation 3.8.3.4. Since we are estimating two parameters, the matrix is 2×2 .

$$\frac{1}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} = \frac{1}{\exp(\mathbf{x}_i^T \boldsymbol{\beta}) \cdot \exp(-\mathbf{x}_i^T \boldsymbol{\beta})^2} = \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

$$\begin{aligned}
 \mathbf{I} &= \begin{bmatrix} \sum_{i=1}^3 \frac{1}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} & \sum_{i=1}^3 \frac{x_i}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} \\ \sum_{i=1}^3 \frac{x_i}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} & \sum_{i=1}^3 \frac{x_i^2}{\text{Var}[Y_i] \cdot g'(\mu_i)^2} \end{bmatrix} \\
 &= \begin{bmatrix} \sum_{i=1}^3 \exp(\mathbf{x}_i^T \boldsymbol{\beta}) & \sum_{i=1}^3 x_i \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \\ \sum_{i=1}^3 x_i \exp(\mathbf{x}_i^T \boldsymbol{\beta}) & \sum_{i=1}^3 x_i^2 \exp(\mathbf{x}_i^T \boldsymbol{\beta}) \end{bmatrix} \\
 &= \exp(\beta_0) \begin{bmatrix} \sum_{i=1}^3 \exp(\beta_1 x_i) & \sum_{i=1}^3 x_i \exp(\beta_1 x_i) \\ \sum_{i=1}^3 x_i \exp(\beta_1 x_i) & \sum_{i=1}^3 x_i^2 \exp(\beta_1 x_i) \end{bmatrix}
 \end{aligned}$$

Thus, obtain $[\mathbf{I}^{(0)}]^{-1}$ as follows:

$$\begin{aligned}
 \mathbf{I}^{(0)} &= e^3 \begin{bmatrix} e^{-0.2(10)} + e^{-0.2(9)} + e^{-0.2(5)} & 10e^{-0.2(10)} + 9e^{-0.2(9)} + 5e^{-0.2(5)} \\ 10e^{-0.2(10)} + 9e^{-0.2(9)} + 5e^{-0.2(5)} & 10^2 e^{-0.2(10)} + 9^2 e^{-0.2(9)} + 5^2 e^{-0.2(5)} \end{bmatrix} \\
 &= \begin{bmatrix} 13.4275 & 94.0092 \\ 94.0092 & 725.4841 \end{bmatrix}
 \end{aligned}$$

$$\begin{aligned}
 [\mathbf{I}^{(0)}]^{-1} &= \frac{1}{(13.4275)(725.4841) - 94.0092^2} \begin{bmatrix} 725.4841 & -94.0092 \\ -94.0092 & 13.4275 \end{bmatrix} \\
 &= \begin{bmatrix} 0.8028 & -0.1040 \\ -0.1040 & 0.0149 \end{bmatrix}
 \end{aligned}$$

Finally,

$$\begin{aligned}
\hat{\beta}^{(1)} &= \begin{bmatrix} 3.0 \\ -0.2 \end{bmatrix} + \begin{bmatrix} 0.8028 & -0.1040 \\ -0.1040 & 0.0149 \end{bmatrix} \begin{bmatrix} -9.4275 \\ -70.0092 \end{bmatrix} \\
&= \begin{bmatrix} 3.0 \\ -0.2 \end{bmatrix} + \begin{bmatrix} 0.8028(-9.4275) + (-0.1040)(-70.0092) \\ -0.1040(-9.4275) + 0.0149(-70.0092) \end{bmatrix} \\
&= \begin{bmatrix} 3.0 - 0.2855 \\ -0.2 - 0.0595 \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{2.7145} \\ \mathbf{-0.2595} \end{bmatrix}
\end{aligned}$$

■

Coach's Remarks

The algorithm's next iteration leads to

$$\mathbf{u}^{(1)} = \begin{bmatrix} -2.7123 \\ -21.0390 \end{bmatrix}$$

Recall that the score equations are $\mathbf{u} = \mathbf{0}$, and notice that $\mathbf{u}^{(1)}$ is closer to $\mathbf{0}$ compared to $\mathbf{u}^{(0)}$. Therefore, $\hat{\beta}^{(1)}$ is closer to the MLE estimates than $\hat{\beta}^{(0)}$ is.

Coach's Remarks

An exam problem may also test the method of scoring outside of a GLM setup. For example, if the goal is to estimate a parameter of interest θ with MLE, then the score function and information would be for θ ; there is no β in the picture. Effectively, only Equation 3.8.3.5 remains relevant after making the proper changes, e.g.

$$\hat{\theta}^{(m)} = \hat{\theta}^{(m-1)} + \left[I(\hat{\theta}^{(m-1)}) \right]^{-1} l'(\hat{\theta}^{(m-1)})$$

3.8.4 Numerical Results

In multiple linear regression, R^2 is a useful quantity because SST is the sum of SSR and SSE. However, this is not true in general; while unnecessary to know for this exam, the actual formula for SST is

$$\underbrace{\sum_{i=1}^n (y_i - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}_{\text{SSR}} + \underbrace{\sum_{i=1}^n (y_i - \hat{y}_i)^2}_{\text{SSE}} + 2 \sum_{i=1}^n (y_i - \hat{y}_i) (\hat{y}_i - \bar{y})$$

The last term equals 0 for multiple linear regression. However, this is rarely the case with GLM. When the last term is non-zero, it becomes unclear how R^2 should even be defined; $\frac{\text{SSR}}{\text{SST}}$ and $1 - \frac{\text{SSE}}{\text{SST}}$ are no longer equal. Hence, we rely on other quantities in a GLM context.

Maximized Log-Likelihoods

Many measures of model quality are defined using maximized log-likelihoods. We introduced this concept in Section 2.5.4; we now clarify and expand on the details within the GLM context.

Given the MLE estimates $\hat{\beta}$, the maximized log-likelihood is simply $l(\hat{\beta})$, i.e. evaluate the log-likelihood at the estimates. Realize that we can also view the log-likelihood as a function of the mean responses, μ_i , rather than a function of β . Hence, a maximized log-likelihood is also the result of substituting each μ_i with $\hat{\mu}_i$ in the log-likelihood.

Consider the null model which has only one β parameter, so $\mathbf{x}^T \beta = \beta_0$. Then, $\hat{\beta}_0$ is the MLE estimate of β_0 . Let l_{null} denote the maximized log-likelihood for this model, i.e. $l(\hat{\beta}_0)$. Furthermore, for the distributions that are relevant to this exam, $\hat{\mu}$ would equal the sample mean \bar{y} . Since the null model estimates only one parameter, the results mentioned in Section 2.1.5 are applicable.

Now imagine a **saturated model** where β has length n ; there are as many parameters to estimate as there are observations. This will lead to

$$\hat{\mu}_i = y_i$$

Let l_{sat} denote the maximized log-likelihood for a saturated model, i.e. substitute each μ_i with y_i in the log-likelihood.

For nested models, more parameters in β lead to a larger maximized log-likelihood, and vice versa. This parallels the idea from linear regression that increasing flexibility decreases the SSE. As

a result,

$$l_{\text{null}} \leq l(\hat{\beta}) \leq l_{\text{sat}}$$

A good model fit has a reasonably high $l(\hat{\beta})$, as it measures the likelihood that the model and its estimated parameters match the data. Said differently, $l(\hat{\beta})$ indicates how close $\hat{\mu}$ is to the response. But when $l(\hat{\beta}) \approx l_{\text{sat}}$, there is danger of overfitting since $\hat{\mu}$ would then follow the response very closely.

Coach's Remarks

Some resources use alternate terminology such as:

- null model → minimal model
- saturated model → maximal model

Deviance

The **deviance** of a model with MLE estimates $\hat{\beta}$ is

$$D = 2 \left[l_{\text{sat}} - l(\hat{\beta}) \right] \quad (3.8.4.1)$$

Since the deviance is a difference of two log-likelihoods, realize that the $d(y_i)$ terms cancel.

A good fit has small deviance, but overfitting occurs when the deviance is too small, such as when $l(\hat{\beta}) \approx l_{\text{sat}}$. Intuitively, deviance is similar to the SSE from multiple linear regression.

For multiple linear regression,

$$D = \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{\sigma^2}$$

If interested in the proof, see the appendix at the end of the section. Since σ^2 is unknown in this context, the deviance is not useful. However, notice that

$$\sigma^2 D = \sum_{i=1}^n (y_i - \hat{\mu}_i)^2 = \text{SSE}$$

Here, $\sigma^2 D$ is called the **scaled deviance**.

Pseudo R²

An alternative for R^2 is the **pseudo R²**:

$$\begin{aligned} R_{\text{pse.}}^2 &= \frac{l_{\text{null}} - l(\hat{\beta})}{l_{\text{null}}} \\ &= 1 - \frac{l(\hat{\beta})}{l_{\text{null}}} \end{aligned} \tag{3.8.4.2}$$

It captures the proportional improvement of the fitted model relative to the null model.

Information Criteria

Both AIC and BIC are commonly used for comparing models in GLM. Their properties were discussed in Section 3.6.2; keep in mind that formulas given there are specific to multiple linear regression. The more generic formulas are:

$$\text{AIC} = -2 \cdot l(\hat{\beta}) + 2 \cdot (\# \text{ of estimated parameters}) \tag{3.8.4.3}$$

$$\text{BIC} = -2 \cdot l(\hat{\beta}) + \ln n \cdot (\# \text{ of estimated parameters}) \tag{3.8.4.4}$$

The number of estimated parameters is typically $p + 1$ for the regression coefficients. However, it could be more than $p + 1$, such as for multiple linear regression where σ^2 also requires estimation.

These formulas are provided in the exam table; note that its definition of p differs from our manual's definition.

Residuals

In GLM, **raw residuals** are equivalent to the residuals of linear regression, i.e. $e_i = y_i - \hat{\mu}_i$. However, raw residuals are not generally useful across the scope of GLM. Let's consider two other types of residuals; both can be assessed in ways similarly described in Section 3.5.

PEARSON RESIDUALS

$$e_i^P = \frac{e_i}{\sqrt{\widehat{\text{Var}}[Y_i]}} \quad (3.8.4.5)$$

$\widehat{\text{Var}}[Y_i]$ refers to $\text{Var}[Y_i]$ after evaluating β at $\hat{\beta}$. Then, the **Pearson chi-square statistic** is $\sum_{i=1}^n (e_i^P)^2$. Notice its similarity with SSE (the sum of squared residuals); the only difference is the type of residual.

Much like standardized residuals in Section 3.5.2, we can obtain standardized Pearson residuals:

$$e_{\text{sta}, i}^P = \frac{e_i^P}{\sqrt{1 - h_i}} \quad (3.8.4.6)$$

DEVIANCE RESIDUALS

Since l_{sat} and $l(\hat{\beta})$ are sums that are indexed by i , the same is true of the deviance. Thus,

$$D = \sum_{i=1}^n D_i$$

where D_i is the version of Equation 3.8.4.1 for the i^{th} observation. Then, the i^{th} deviance residual is

$$e_i^D = \pm \sqrt{D_i} \quad (3.8.4.7)$$

whose sign follows the i^{th} raw residual.

Likewise, we can obtain standardized deviance residuals:

$$e_{\text{sta}, i}^D = \frac{e_i^D}{\sqrt{1 - h_i}} \quad (3.8.4.8)$$

Example 3.8.4.1

Determine which of the following is an indication of a good model fit.

- I. A very large BIC
- II. A very large Pearson chi-square statistic
- III. A maximized log-likelihood at its largest possible value
- IV. A very large deviance
- V. A very small pseudo R^2

Solution

BIC mimics the behavior of the test MSE as a function of flexibility. Thus, statement I is not an indication of a good model fit.

The Pearson chi-squared statistic is the sum of squared Pearson residuals. Since residuals capture the disparity between the actual and predicted responses, statement II is not an indication of a good model fit.

A maximized log-likelihood at its largest possible value is l_{sat} . This is evidence of an overfitted model that predicts the actual responses perfectly. Thus, statement III is not an indication of a good model fit.

Deviance captures the disparity between l_{sat} and $l(\hat{\beta})$. A large disparity suggests that $\hat{\mu}$ does not predict the actual responses well, meaning statement IV is not an indication of a good model fit.

Pseudo R^2 is very small when $l(\hat{\beta}) \approx l_{\text{null}}$. This means the fitted model is almost as helpful as the model with no predictors, so statement V is not an indication of a good model fit.

Therefore, **none of the statements are an indication of a good model.**

3.8.5 Inference

Score Statistics

Let the **score statistic** U_j be the random variable counterpart of u_j , i.e.

$$U_j = \sum_{i=1}^n \frac{(Y_i - \mu_i)x_{i,j}}{\text{Var}[Y_i] \cdot g'(\mu_i)}$$

and let

$$\mathbf{U} = \begin{bmatrix} U_0 \\ U_1 \\ \vdots \\ U_p \end{bmatrix}$$

It can be proven that \mathbf{U} asymptotically follows a multivariate normal distribution with mean $\mathbf{0}$ and variance-covariance matrix \mathbf{I} , the information matrix of β . As a result, this translates to $\mathbf{U}^T \mathbf{I}^{-1} \mathbf{U}$ having an approximate chi-square distribution with $p + 1$ degrees of freedom, given that \mathbf{U} is a vector of length $p + 1$. It also means that

$$\frac{U_j}{\sqrt{(j+1)^{\text{st}} \text{ diagonal entry of } \mathbf{I}}}$$

approximately follows a standard normal distribution, and in turn, its square approximately follows a chi-square distribution with 1 degree of freedom.

Maximum Likelihood Estimators

Consistent with the introduction in Section 2.2.6, the maximum likelihood estimators, $\hat{\beta}$, asymptotically follow a multivariate normal distribution with mean β and variance-covariance matrix \mathbf{I}^{-1} . To be precise,

$$\mathbf{I}^{-1} = \begin{bmatrix} \widehat{\text{Var}}[\hat{\beta}_0] & \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] & \cdots & \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_p] \\ \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_1] & \widehat{\text{Var}}[\hat{\beta}_1] & \cdots & \widehat{\text{Cov}}[\hat{\beta}_1, \hat{\beta}_p] \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{\text{Cov}}[\hat{\beta}_0, \hat{\beta}_p] & \widehat{\text{Cov}}[\hat{\beta}_1, \hat{\beta}_p] & \cdots & \widehat{\text{Var}}[\hat{\beta}_p] \end{bmatrix}$$

As in multiple linear regression, we simplify notation by letting

$$\widehat{\text{Var}}[\hat{\beta}_j] = se(\hat{\beta}_j)^2, \quad j = 0, 1, \dots, p$$

after evaluating $\boldsymbol{\beta}$ at the estimate $\hat{\boldsymbol{\beta}}$.

WALD THEORY

The same conditions that lead to asymptotically normal maximum likelihood estimators permit statistical inference under what is called Wald theory. Given the maximum likelihood estimator $\hat{\boldsymbol{\beta}}$, the main result of Wald theory is that the **Wald statistic** $(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})^T \mathbf{I} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})$ (after \mathbf{I} has evaluated $\boldsymbol{\beta}$ at the estimator $\hat{\boldsymbol{\beta}}$) has an approximate chi-square distribution with $p + 1$ degrees of freedom.

While any hypothesis test that depends on Wald theory is a **Wald test**, we are mainly concerned with testing one β at a time for this exam. Consequently, the Wald tests we encounter can be described as two-tailed standard normal tests for a mean parameter. Specifically,

$$t. s. = \left[\frac{\hat{\beta}_j - h}{se(\hat{\beta}_j)} \right]^2 \tag{3.8.5.1}$$

comes from an approximate chi-square distribution with 1 degree of freedom, and is checked against a right-tailed critical region. Equivalently, this *t. s.* without the square comes from an approximate standard normal distribution, and can be checked against a two-tailed critical region. Even though both approaches are identical, we expect this exam to treat the chi-square approach as the default.

In the same vein, Wald confidence intervals for each of the β_j 's can be described as (two-sided) confidence intervals for a mean parameter based on the standard normal distribution.

OVERDISPERSION AND QUASI-LIKELIHOOD

Distributions in the exponential family have a specific connection between their mean and variance. For example, the Poisson distribution assumes that the mean and variance are equal. However, the data could exhibit a connection between the mean and variance that disagrees with the distribution.

Overdispersion occurs when the observed variability is larger than the variance estimated by the model. It could be caused by assuming independence among the Y_i 's when it is untrue. We may measure overdispersion by computing the ratio of the deviance to its expected value of $n - p - 1$; its severity is suggested by how much the ratio exceeds 1.

This issue causes the $se(\hat{\beta}_j)$'s to be unreliable for making inferences. One way to account for the larger variability observed in the data is to override the variance formula. With the variance of Y_i originally being

$$\frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^3}$$

the **quasi-likelihood method** redefines it as the original variance multiplied by a parameter, i.e.

$$\text{Var}[Y_i] = \phi \cdot \frac{b''(\theta_i)c'(\theta_i) - c''(\theta_i)b'(\theta_i)}{[b'(\theta_i)]^3}$$

where $\phi > 1$. Thus, we introduce a new parameter to gain more control in modeling the variance, which allows for more reliable inferences. On the other hand, this approach implies that the response no longer follows a member of the exponential family. In short, the quasi-likelihood method establishes a relationship between the mean and variance that is less restrictive, but it obscures the distribution of the response.

Likelihood Ratio Tests

Recall how F tests and partial F tests investigated hypotheses involving several regression coefficients simultaneously. In GLM, we use likelihood ratio tests instead; revisit Section 2.5.4 if you need a refresher on the basics.

The goal is to compare two nested models. Consistent with Section 3.3.7, the full model has **more** parameters in β ; the reduced model has **fewer** parameters in β . Let:

- p_f and $\hat{\beta}_f$ represent the number of predictors and the maximum likelihood estimates, respectively, for the full model.
- p_r and $\hat{\beta}_r$ represent the number of predictors and the maximum likelihood estimates, respectively, for the reduced model.

We wish to test the hypotheses

- H_0 : The reduced model is adequate; the additional $p_f - p_r$ regression coefficients are all 0.
- H_1 : The full model is better; at least one of the additional $p_f - p_r$ regression coefficients is not 0.

Identical to Equation 2.5.4.2, the test statistic is

$$t.s. = 2 \left[l(\hat{\beta}_f) - l(\hat{\beta}_r) \right] \quad (3.8.5.2)$$

which comes from an approximate chi-square distribution with $p_f - p_r$ degrees of freedom. Recall that likelihood ratio tests are right-tailed tests.

It is important to note that the test statistic can be written in terms of deviances. If D_f refers to the deviance of the full model and D_r refers to the deviance of the reduced model, then

$$\begin{aligned} t.s. &= 2 \left[l(\hat{\beta}_f) - l(\hat{\beta}_r) \right] \\ &= 2 \left[l(\hat{\beta}_f) - l(\hat{\beta}_r) + l_{\text{sat}} - l_{\text{sat}} \right] \\ &= 2 \left[l_{\text{sat}} - l(\hat{\beta}_r) - \{l_{\text{sat}} - l(\hat{\beta}_f)\} \right] \\ &= D_r - D_f \end{aligned} \quad (3.8.5.3)$$

Moreover, consider two ways of assessing a fitted model using this test. The first views the fitted model as the full model, and the null model as the reduced model. Such a likelihood ratio test examines whether at least one of the p predictors in the fitted model is significant; the degrees of freedom is p . For this exam, this assessment of the fitted model may be assumed when a likelihood ratio test **does not specify** which two models are being compared.

The second assessment views the saturated model as the full model, and the fitted model as the reduced model. Such a likelihood ratio test examines whether the fitted model is adequate or

perhaps missing a significant predictor. Note that in this scenario, the test statistic equals D (i.e. Equation 3.8.4.1) and the degrees of freedom is $n - p - 1$.

Example 3.8.5.1

For a generalized linear model, you consider the link function

$$g(\mu) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

You are given:

Predictors Used	Maximized Log-Likelihood
x_1, x_2	-539.85
x_1, x_3	-540.17
x_1, x_2, x_3, x_4	-536.41

Determine which conclusion can be made using a likelihood ratio test at the 5% significance level.

- A. $H_0 : \beta_1 = 0$ fails to be rejected.
- B. $H_0 : \beta_2 = \beta_3 = 0$ fails to be rejected.
- C. $H_0 : \beta_2 = \beta_4 = 0$ fails to be rejected.
- D. $H_0 : \beta_3 = \beta_4 = 0$ should be rejected.
- E. $H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$ should be rejected.

Solution

The full model uses all four predictors. There are only two likelihood ratio tests that can be performed with the given information:

- The reduced model uses x_1 and x_2 only, i.e. $H_0 : \beta_3 = \beta_4 = 0$.
- The reduced model uses x_1 and x_3 only, i.e. $H_0 : \beta_2 = \beta_4 = 0$.

This means options (A), (B), and (E) cannot be proven; they are thus eliminated.

Both tests examine the shared significance of two regression coefficients. At $\alpha = 0.05$, the critical value for both tests is $\chi^2_{0.95, 2} = 5.99$ according to the exam table.

For $H_0 : \beta_2 = \beta_4 = 0$, the test statistic is

$$2[-536.41 - (-540.17)] = 7.52$$

Since $7.52 > 5.99$, we reject this H_0 , which means option (C) is incorrect.

For $H_0 : \beta_3 = \beta_4 = 0$, the test statistic is

$$2[-536.41 - (-539.85)] = 6.88$$

Since $6.88 > 5.99$, we reject this H_0 . Therefore, the answer is (D).



Example 3.8.5.2

Determine which statement is true for a generalized linear model.

- I. Overdispersion occurs when the chosen distribution for the response variable assumes a larger variance than the variability observed in the data.
- II. Adjusting for overdispersion helps to correctly specify the mean of the response variable.
- III. The quasi-likelihood method assumes a relationship between the mean and variance of the response variable.
- IV. The quasi-likelihood method helps to improve the maximum likelihood estimates of the regression coefficients.

Solution

I is false. Overdispersion occurs when the variability in the data is larger than the variance

assumed by the distribution.

II is false. Adjusting for overdispersion helps to correctly specify the variance of the response variable.

III is true. For exponential family distributions, the variance is a function of the mean. The quasi-likelihood method retains the variance as a function of the mean by multiplying the original function with a parameter ϕ . Thus, there is still a relationship between the mean and variance, albeit a less restrictive one.

IV is false. The quasi-likelihood method seeks to produce reliable statistical inference, not obtain better coefficient estimates.

Therefore, **only III is true.**



3.8.6 Tweedie Distributions

Consider distributions from the exponential family whose mean and variance are specifically related by

$$\text{Var}[Y] = a \cdot \mathbb{E}[Y]^d \quad (3.8.6.1)$$

where a and d are constants. Such distributions are known as *Tweedie distributions*. These include:

- The normal distribution with $d = 0$
- The Poisson distribution with $d = 1$
- The compound Poisson-gamma distribution with $1 < d < 2$
- The gamma distribution with $d = 2$
- The inverse Gaussian distribution with $d = 3$

While a is also distribution-specific, it is not crucial to know for each distribution, and may be derived using Equation 3.8.6.1 if necessary.

Checking the data for such a connection between the mean and variance may help in selecting a suitable distribution. In particular, the compound Poisson-gamma distribution is appealing since it models a discrete mass for when the response is 0 and a continuous component for when the response is positive. It may help to review the basic structure of a compound Poisson process in Section 1.4.6.

Example 3.8.6.1

Tweedie distributions have the following relationship between their mean and variance:

$$\text{Var}[Y] = a(\mathbb{E}[Y])^p$$

Determine which statements are true.

- I. The exponential distribution is not a Tweedie distribution.
- II. The binomial distribution is not a Tweedie distribution.

III. For a Poisson distribution and an inverse Gaussian distribution with $\theta = 4$, the absolute difference in their a 's is 0.75.

Solution

I is false. We know that gamma is a Tweedie distribution, and that exponential is a gamma with $\alpha = 1$. Alternatively, exponential is a member of the exponential family, such that

$$\begin{aligned}\text{Var}[Y] &= \theta^2 \\ &= 1 \cdot \theta^2 \\ &= 1(\mathbb{E}[Y])^2\end{aligned}$$

II is true. For binomial,

$$\begin{aligned}\text{Var}[Y] &= mq(1-q) \\ &= (1-q)(\mathbb{E}[Y])^1\end{aligned}$$

where $1 - q$ is not a constant because it is a function of the parameter of interest, q .

III is true. For Poisson, we know that $p = 1$ and

$$\begin{aligned}\text{Var}[Y] &= \lambda \\ &= 1(\mathbb{E}[Y])^1\end{aligned}$$

i.e. $a = 1$.

For an inverse Gaussian with $\theta = 4$, we know that $p = 3$ and

$$\begin{aligned}\text{Var}[Y] &= \frac{\mu^3}{\theta} \\ &= \frac{1}{\theta} \cdot \mu^3 \\ &= \theta^{-1}(\mathbb{E}[Y])^3\end{aligned}$$

i.e. $a = \theta^{-1} = 4^{-1}$. As a result, the absolute difference between 1 and $\frac{1}{4}$ is 0.75.

Therefore, **only II and III are true.**



3.8 Summary

Notation

Symbol	Concept
$E[Y], \mu$	Mean response
$\hat{\beta}$	Maximum likelihood estimate/estimator for β
u_j, U_j	Score function/statistic for β_j
I	Information matrix of β
$l(\hat{\beta})$	Maximized log-likelihood
l_{null}	Maximized log-likelihood for null model
l_{sat}	Maximized log-likelihood for saturated model

Exponential Family

Distributions in the exponential family have probability functions in the form of

$$f(y) = \exp [a(y) \cdot b(\theta) + c(\theta) + d(y)]$$

Parameters other than θ are nuisance parameters.

$$E[a(Y)] = -\frac{c'(\theta)}{b'(\theta)}$$

$$\text{Var}[a(Y)] = \frac{b''(\theta)c'(\theta) - c''(\theta)b'(\theta)}{[b'(\theta)]^3}$$

For GLM,

- $a(y) = y$
- $b(\theta)$ is the natural parameter
- μ is a function of θ
- $\text{Var}[Y]$ is a function of μ

Model Framework

- The response has a distribution that belongs to the exponential family.
- A link function, $g(\cdot)$, is a function of μ that is set equal to $\mathbf{x}^T \boldsymbol{\beta}$.
 - Identity link: $g(\mu) = \mu$
 - Logit link: $g(\mu) = \ln \left(\frac{\mu}{1 - \mu} \right)$
 - Logarithmic link: $g(\mu) = \ln \mu$
 - Inverse link: $g(\mu) = \frac{1}{\mu}$
 - Power link: $g(\mu) = \mu^d$
- The canonical link of a distribution is the function of μ that equals the natural parameter.

Distribution	Canonical Link
Normal	Identity link
Binomial	Logit link
Poisson	Logarithmic link
Gamma	Inverse link

Parameter Estimation

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i \cdot b(\theta_i) + c(\theta_i) + d(y_i)]$$

$$\hat{\mu} = g^{-1}(\mathbf{x}^T \hat{\boldsymbol{\beta}})$$

$$u_j = \sum_{i=1}^n \frac{(y_i - \mu_i)x_{i,j}}{\text{Var}[Y_i] \cdot g'(\mu_i)}$$

$$\mathbf{I} = \sum_{i=1}^n \frac{\mathbf{x}_i \mathbf{x}_i^T}{\text{Var}[Y_i] \cdot g'(\mu_i)^2}$$

METHOD OF SCORING

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{(m)} &= \hat{\boldsymbol{\beta}}^{(m-1)} + \left[\mathbf{I}^{(m-1)} \right]^{-1} \mathbf{u}^{(m-1)} \\ &= \left(\mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{X} \right)^{-1} \mathbf{X}^T \mathbf{W}^{(m-1)} \mathbf{z}^{(m-1)}\end{aligned}$$

where \mathbf{W} is a diagonal matrix with entries w_1, \dots, w_n , and

$$w_i = \frac{1}{\text{Var}[Y_i] \cdot g'(\mu_i)^2}, \quad z_i = g(\mu_i) + (y_i - \mu_i)g'(\mu_i)$$

Each iteration attempts to bring $\hat{\boldsymbol{\beta}}^{(m)}$ closer to the estimate $\hat{\boldsymbol{\beta}}$.

Numerical Results

DEVIANCE

$$D = 2 \left[l_{\text{sat}} - l(\hat{\boldsymbol{\beta}}) \right]$$

PSEUDO R²

$$R_{\text{pse.}}^2 = \frac{l_{\text{null}} - l(\hat{\boldsymbol{\beta}})}{l_{\text{null}}} = 1 - \frac{l(\hat{\boldsymbol{\beta}})}{l_{\text{null}}}$$

INFORMATION CRITERIA

$$\text{AIC} = -2 \cdot l(\hat{\beta}) + 2 \cdot (\# \text{ of estimated parameters})$$

$$\text{BIC} = -2 \cdot l(\hat{\beta}) + \ln n \cdot (\# \text{ of estimated parameters})$$

RAW RESIDUAL

$$e_i = y_i - \hat{\mu}_i$$

PEARSON RESIDUAL

$$e_i^P = \frac{e_i}{\sqrt{\widehat{\text{Var}}[Y_i]}}, \quad e_{\text{sta}, i}^P = \frac{e_i^P}{\sqrt{1 - h_i}}$$

The Pearson chi-square statistic is $\sum_{i=1}^n (e_i^P)^2$.

DEVIANCE RESIDUAL

$$e_i^D = \pm \sqrt{D_i}, \quad e_{\text{sta}, i}^D = \frac{e_i^D}{\sqrt{1 - h_i}}$$

whose sign follows the i^{th} raw residual.

Inference

- Score statistics \mathbf{U} asymptotically follow a multivariate normal distribution with mean $\mathbf{0}$ and asymptotic variance-covariance matrix \mathbf{I} . Thus, $\mathbf{U}^T \mathbf{I}^{-1} \mathbf{U}$ follows an approximate chi-square distribution with $p + 1$ degrees of freedom.
- Maximum likelihood estimators $\hat{\beta}$ asymptotically follow a multivariate normal distribution with mean β and asymptotic variance-covariance matrix \mathbf{I}^{-1} .

- The Wald statistic $(\hat{\beta} - \beta)^T \mathbf{I} (\hat{\beta} - \beta)$ follows an approximate chi-square distribution with $p + 1$ degrees of freedom. To test the significance of predictor x_j ,

$$t.s. = \left[\frac{\hat{\beta}_j - h}{se(\hat{\beta}_j)} \right]^2$$

with a critical region of $t.s. \geq \chi^2_{1-\alpha, 1}$.

- For more reliable inferences, overdispersion can be addressed by the quasi-likelihood method, which changes the variance to

$$\text{Var}[Y_i] = \phi \cdot \text{original variance}$$

The mean and variance have a less restrictive relationship, but the distribution is obscured.

- For likelihood ratio tests, the full model has more regression coefficients, and the reduced model has fewer regression coefficients.

$$\begin{aligned} t.s. &= 2 \left[l(\hat{\beta}_f) - l(\hat{\beta}_r) \right] \\ &= D_r - D_f \end{aligned}$$

The critical region is $t.s. \geq \chi^2_{1-\alpha, p_f - p_r}$.

Tweedie Distributions

$$\text{Var}[Y] = a \cdot \mathbb{E}[Y]^d$$

- The normal distribution with $d = 0$
- The Poisson distribution with $d = 1$
- The compound Poisson-gamma distribution with $1 < d < 2$
- The gamma distribution with $d = 2$
- The inverse Gaussian distribution with $d = 3$

Connection with MLR

With a normally distributed response, identity link, and homoscedasticity, GLM is the same as MLR.
As a result,

- Maximum likelihood estimates are the same as the ordinary least squares estimates
- $\sigma^2 D = \text{SSE}$

Appendix

🕒 15m

Mean and Variance of Exponential Family

In order to derive the mean and variance of the exponential family, let's start with some key results.

$$\begin{aligned} \int f(y) dy &= 1 \\ \frac{d}{d\theta} \int f(y) dy &= \frac{d}{d\theta}(1) \\ \int \frac{df(y)}{d\theta} dy &= 0 \end{aligned} \tag{1}$$

$$\begin{aligned} \int f(y) dy &= 1 \\ \frac{d^2}{d\theta^2} \int f(y) dy &= \frac{d^2}{d\theta^2}(1) \\ \int \frac{d^2 f(y)}{d\theta^2} dy &= 0 \end{aligned} \tag{2}$$

Now, differentiate the probability function of the exponential family.

$$f(y) = \exp[a(y) \cdot b(\theta) + c(\theta) + d(y)]$$

$$\begin{aligned} \frac{df(y)}{d\theta} &= \exp[a(y) \cdot b(\theta) + c(\theta) + d(y)] \cdot [a(y) \cdot b'(\theta) + c'(\theta)] \\ &= f(y) \cdot [a(y) \cdot b'(\theta) + c'(\theta)] \end{aligned}$$

Substitute this into (1) and rearrange to obtain the mean.

$$\begin{aligned} \int f(y) \cdot [a(y) \cdot b'(\theta) + c'(\theta)] dy &= 0 \\ b'(\theta) \int a(y) \cdot f(y) dy + c'(\theta) \int f(y) dy &= 0 \\ b'(\theta) \cdot E[a(Y)] + c'(\theta) &= 0 \\ E[a(Y)] &= -\frac{c'(\theta)}{b'(\theta)} \end{aligned}$$

Now, differentiate the probability function twice.

$$\begin{aligned}
\frac{d^2 f(y)}{d\theta^2} &= \exp[a(y) \cdot b(\theta) + c(\theta) + d(y)] \cdot [a(y) \cdot b'(\theta) + c'(\theta)]^2 + \exp[a(y) \cdot b(\theta) + c(\theta) + d(y)] \cdot [a(y) \cdot b'(\theta) + c'(\theta)] \\
&= f(y) \cdot [a(y) \cdot b'(\theta) + c'(\theta)]^2 + f(y) \cdot [a(y) \cdot b''(\theta) + c''(\theta)] \\
&= f(y) \cdot [a(y) \cdot b'(\theta) - E[a(Y)] \cdot b'(\theta)]^2 + f(y) \cdot [a(y) \cdot b''(\theta) + c''(\theta)] \\
&= f(y) \cdot b'(\theta)^2 \cdot [a(y) - E[a(Y)]]^2 + f(y) \cdot [a(y) \cdot b''(\theta) + c''(\theta)]
\end{aligned}$$

Note that we substitute $c'(\theta) = -E[a(Y)] \cdot b'(\theta)$ in the third line above.

Substitute this into (2) and rearrange to obtain the variance.

$$\begin{aligned}
\int f(y) \cdot b'(\theta)^2 \cdot [a(y) - E[a(Y)]]^2 + f(y) \cdot [a(y) \cdot b''(\theta) + c''(\theta)] dy &= 0 \\
b'(\theta)^2 \int [a(y) - E[a(Y)]]^2 \cdot f(y) dy + b''(\theta) \int a(y) \cdot f(y) dy + c''(\theta) \int f(y) dy &= 0 \\
b'(\theta)^2 \cdot \text{Var}[a(Y)] + b''(\theta) \cdot E[a(Y)] + c''(\theta) &= 0 \\
\text{Var}[a(Y)] &= \frac{-b''(\theta) \cdot E[a(Y)] - c''(\theta)}{b'(\theta)^2} \\
&= \frac{b''(\theta) \cdot \frac{c'(\theta)}{b'(\theta)} - c''(\theta)}{b'(\theta)^2} \\
&= \frac{b''(\theta) \cdot c'(\theta) - c''(\theta)}{b'(\theta)^3}
\end{aligned}$$

Deviance under Multiple Linear Regression

The normal log-likelihood with constant variance is

$$l(\beta) = \sum_{i=1}^n \left[\frac{y_i \mu_i}{\sigma^2} - \frac{\mu_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} \right]$$

Under the fitted model:

$$l(\hat{\beta}) = \sum_{i=1}^n \left[\frac{y_i \hat{\mu}_i}{\sigma^2} - \frac{\hat{\mu}_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} \right]$$

Under the saturated model:

$$\hat{\mu}_i = y_i$$

$$\begin{aligned}
l_{\text{sat}} &= \sum_{i=1}^n \left[\frac{y_i^2}{\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} \right] \\
&= \sum_{i=1}^n \left[\frac{2y_i^2 - y_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} \right] \\
&= \sum_{i=1}^n \left[\frac{y_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} \right]
\end{aligned}$$

Therefore, the deviance is

$$\begin{aligned}
D &= 2 \left[l_{\text{sat}} - l(\hat{\beta}) \right] \\
&= 2 \left\{ \sum_{i=1}^n \left[\frac{y_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} \right] - \sum_{i=1}^n \left[\frac{y_i \hat{\mu}_i}{\sigma^2} - \frac{\hat{\mu}_i^2}{2\sigma^2} - \frac{y_i^2}{2\sigma^2} - \frac{\ln(2\pi\sigma^2)}{2} \right] \right\} \\
&= 2 \sum_{i=1}^n \left[\frac{y_i^2}{2\sigma^2} - \frac{y_i \hat{\mu}_i}{\sigma^2} + \frac{\hat{\mu}_i^2}{2\sigma^2} \right] \\
&= \frac{2}{2\sigma^2} \sum_{i=1}^n (y_i^2 - 2y_i \hat{\mu}_i + \hat{\mu}_i^2) \\
&= \frac{\sum_{i=1}^n (y_i - \hat{\mu}_i)^2}{\sigma^2}
\end{aligned}$$

3.9.0 Overview

 5m

Continuing the discussion on generalized linear models, we first focus on modeling the response variable with the binomial distribution. Moreover, the approach to handling a binomial response helps with formulating models for a categorical response. From Section 3.1.1, recall the two types of categorical variables: nominal and ordinal. A model for each type is presented as well.

3.9.1 Binomial Response

The PMF of a binomial distribution can be written as

$$\begin{aligned}
 p(y) &= \binom{m}{y} q^y (1-q)^{m-y} \\
 &= \exp \left[\ln \binom{m}{y} + y \ln q + (m-y) \ln (1-q) \right] \\
 &= \exp \left[y \ln q + m \ln (1-q) - y \ln (1-q) + \ln \binom{m}{y} \right] \\
 &= \exp \left[y \ln \left(\frac{q}{1-q} \right) + m \ln (1-q) + \ln \binom{m}{y} \right]
 \end{aligned}$$

where

- $a(y) = y$
- $b(q) = \ln \left(\frac{q}{1-q} \right)$
- $c(q) = m \ln (1-q)$
- $d(y) = \ln \binom{m}{y}$

This proves that the binomial distribution belongs to the exponential family. Moreover, recall that

$$\mathbb{E}[Y] = mq, \quad \text{Var}[Y] = mq(1-q)$$

Since Y has an exponential family distribution, these formulas agree with Equations 3.8.1.1 and 3.8.1.2.

Odds and Log Odds

Before going further, we define the **odds** of an event as the ratio of the probability that the event will occur to the probability that the event will not occur, i.e.

$$\text{odds} = \frac{q}{1-q} \tag{3.9.1.1}$$

The natural log of the odds is called the **log odds**.

Example 3.9.1.1

You are presented with a standard deck of 52 cards consisting of four suits (spades, hearts, clubs, and diamonds) with thirteen cards each. Cards are drawn without replacement.

Determine which of the following statements is true.

- I. The odds of the first card drawn being a spade are 1 to 4.
- II. The odds of the first two cards drawn being two clubs are 0.0625.
- III. Given the first four draws are hearts, the odds of the fifth card drawn being a heart are 0.1875.

Solution

I is false. The probability of drawing a spade is 0.25. The odds of that are

$$\frac{0.25}{1 - 0.25} = \frac{1}{3}$$

II is true. The probability of the first two draws being two clubs is

$$\frac{13}{52} \cdot \frac{12}{51} = \frac{1}{17}$$

The odds of that are

$$\frac{\frac{1}{17}}{1 - \frac{1}{17}} = \frac{1}{16} = 0.0625$$

III is false. Four hearts have been drawn; this means only nine hearts remain in the deck. The probability of drawing a heart on the fifth card is $\frac{9}{48} = 0.1875$. The odds of that are

$$\frac{0.1875}{1 - 0.1875} = 0.2308$$

Therefore, **only II is true.**



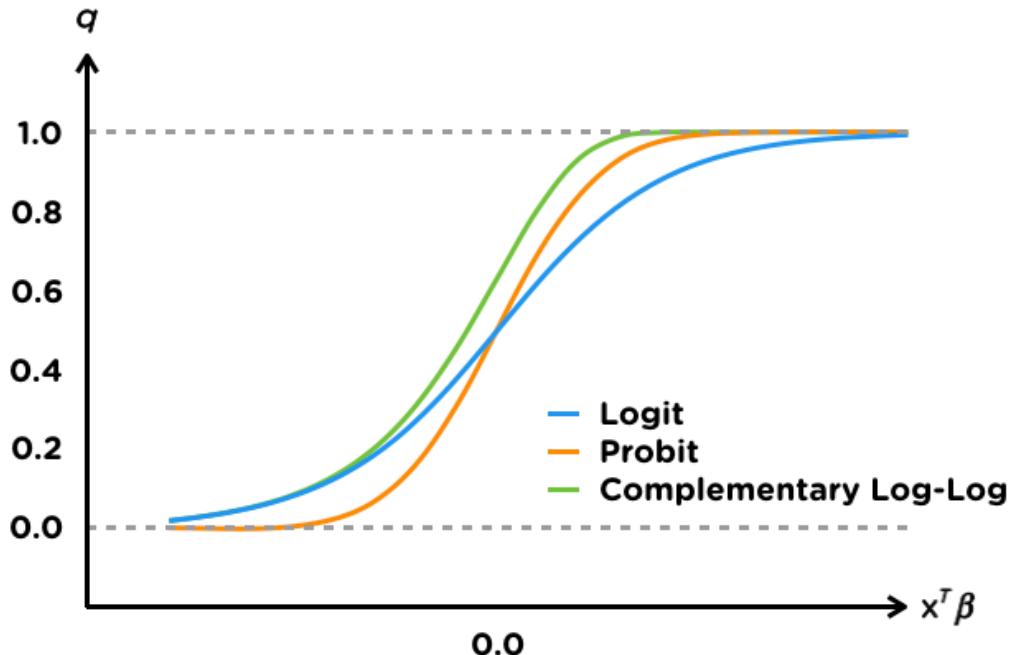
Link Functions

In Section 3.8.2, we use a link function to connect the mean response μ to the regression coefficients β . While that is a useful generic approach (i.e. not needing to specify a distribution), it is not strictly required. At minimum, we need a link function that connects the parameter of interest to β , i.e. $g(\theta) = \mathbf{x}^T \beta$. For a binomial response, recall that $\theta = q$. Therefore, the norm in this context is to consider link functions that are functions of q , rather than functions of $\mu = mq$.

The following link functions are suitable candidates for a binomial response, given that q must be between 0 and 1:

Function Name	$g(q)$
Logit link	$\ln\left(\frac{q}{1-q}\right)$
Probit link	$\Phi^{-1}(q)$
Complementary log-log link	$\ln[-\ln(1-q)]$

These link functions are suitable because they restrict the range of $g^{-1}(\mathbf{x}^T \beta)$ to be between 0 to 1, where $\mathbf{x}^T \beta$ is any real number. To illustrate, here is a plot of q against $\mathbf{x}^T \beta$ for each link function.



Model

Let Y_1, \dots, Y_n be independent binomial random variables with known m_i 's and $q_i = g^{-1}(\mathbf{x}_i^T \beta)$. Then, we obtain the following simplified expressions:

$$\begin{aligned}\mu_i &= m_i q_i \\ &= m_i \cdot g^{-1}(\mathbf{x}_i^T \beta)\end{aligned}$$

$$\begin{aligned}\hat{\mu}_i &= m_i \hat{q}_i \\ &= m_i \cdot g^{-1}(\mathbf{x}_i^T \hat{\beta})\end{aligned}$$

- Log-likelihood function:

$$l(\beta) = \sum_{i=1}^n \left[y_i \ln \left(\frac{q_i}{1 - q_i} \right) + m_i \ln (1 - q_i) + \ln \left(\frac{m_i}{y_i} \right) \right] \quad (3.9.1.2)$$

- Deviance:

$$D = 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) + (m_i - y_i) \ln \left(\frac{m_i - y_i}{m_i - \hat{\mu}_i} \right) \right] \quad (3.9.1.3)$$

If interested in the proof, see the appendix at the end of the section.

- Pearson residual:

$$e_i^P = \frac{y_i - m_i \hat{q}_i}{\sqrt{m_i \hat{q}_i (1 - \hat{q}_i)}} \quad (3.9.1.4)$$

- Pearson chi-square statistic:

$$\sum_{i=1}^n (e_i^P)^2 = \sum_{i=1}^n \frac{(y_i - m_i \hat{q}_i)^2}{m_i \hat{q}_i (1 - \hat{q}_i)} \quad (3.9.1.5)$$

Logistic Regression

A *logistic regression model* uses the logit link function, i.e.

$$\begin{aligned}g(q_i) &= \ln \left(\frac{q_i}{1 - q_i} \right) \\ &= \mathbf{x}_i^T \beta\end{aligned}$$

$$\begin{aligned}\Rightarrow q_i &= g^{-1}(\mathbf{x}_i^T \beta) \\ &= \frac{\exp(\mathbf{x}_i^T \beta)}{1 + \exp(\mathbf{x}_i^T \beta)}\end{aligned}$$

Coach's Remarks

Likewise, a probit regression model uses the probit link, and a complementary log-log regression model uses the complementary log-log link.

Since the link function is now $g(q)$ rather than $g(\mu)$, Equation 3.8.3.3 for the score function and Equation 3.8.3.4 for the information matrix no longer hold. For logistic regression, it is easier to memorize their formulas as follows:

$$u_j = \sum_{i=1}^n (y_i - \mu_i)x_{i,j} \quad (3.9.1.6)$$

$$\begin{aligned} \mathbf{I} &= \sum_{i=1}^n m_i q_i (1 - q_i) \mathbf{x}_i \mathbf{x}_i^T \\ &= \begin{bmatrix} \sum_{i=1}^n m_i q_i (1 - q_i) & \sum_{i=1}^n m_i q_i (1 - q_i) x_{i,1} & \cdots & \sum_{i=1}^n m_i q_i (1 - q_i) x_{i,p} \\ \sum_{i=1}^n m_i q_i (1 - q_i) x_{i,1} & \sum_{i=1}^n m_i q_i (1 - q_i) x_{i,1}^2 & \cdots & \sum_{i=1}^n m_i q_i (1 - q_i) x_{i,1} x_{i,p} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^n m_i q_i (1 - q_i) x_{i,p} & \sum_{i=1}^n m_i q_i (1 - q_i) x_{i,1} x_{i,p} & \cdots & \sum_{i=1}^n m_i q_i (1 - q_i) x_{i,p}^2 \end{bmatrix} \end{aligned}$$

Coach's Remarks

If you are curious, when the link function is a function of the parameter of interest instead of the mean response, the score function is

$$u_j = \sum_{i=1}^n (y_i - \mu_i) \frac{b'(\theta_i)}{g'(\theta_i)} x_{i,j}$$

and the $(j+1)^{\text{st}}$ row, $(j^* + 1)^{\text{st}}$ column entry of \mathbf{I} is

$$\sum_{i=1}^n \text{Var}[Y_i] \left[\frac{b'(\theta_i)}{g'(\theta_i)} \right]^2 x_{i,j} x_{i,j^*}$$

INTERPRETATION OF PARAMETER ESTIMATES

Specifically for logistic regression, interpreting $\hat{\beta}_j$ is not too difficult. To motivate this discussion, let's use the Commuting Chris scenario. We want to predict whether a commute is long or short using only Temp and Precip. A commute that takes more than 26 minutes is considered a long commute. Otherwise, it is a short commute.

$$Y = \begin{cases} 0, & \text{if commute is short} \\ 1, & \text{if commute is long} \end{cases}$$

Hence, the response follows a Bernoulli distribution in this case, i.e. we have binomial responses with $m_1 = \dots = m_{100} = 1$.

With a logit link, we have

$$\ln\left(\frac{q}{1-q}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where

- x_1 is the temperature at the time of departure, and
- x_2 is the dummy variable for the presence of precipitation during any point of the commute.

Here is the R output from the logistic regression:

	Estimate	Std. Error
(Intercept)	-1.2834	0.8629
Temp	-0.0108	0.0149
PrecipYes	2.3628	0.5363

With these coefficient estimates, we obtain the following equation to predict the probability that a commute is long given x_1 and x_2 .

$$\ln\left(\frac{\hat{q}}{1-\hat{q}}\right) = -1.2834 - 0.0108x_1 + 2.3628x_2$$

With a temperature of 0 °F and the presence of precipitation, the predicted odds that a commute is long are

$$\ln\left(\frac{\hat{q}}{1-\hat{q}}\right) = -1.2834 + 2.3628 \Rightarrow \widehat{\text{odds}} = e^{1.0794} = 2.9429$$

With a temperature of 0 °F and no precipitation, the predicted odds that a commute is long are

$$\ln \left(\frac{\hat{q}}{1 - \hat{q}} \right) = -1.2834 \quad \Rightarrow \quad \widehat{\text{odds}} = e^{-1.2834} = 0.2771$$

Notice that if we take the ratio of these odds, we get

$$\frac{2.9429}{0.2771} = 10.6206 = \exp(2.3628) = \exp(\hat{\beta}_2)$$

This ratio is called the **odds ratio**. It is the ratio of the odds of an event with the presence of a characteristic to the odds of the same event without the presence of that characteristic. Intuitively, an odds ratio is the change in odds expressed as a factor.

$$\left(\widehat{\text{odds}} \text{ with } x_2 = 1 \right) = \exp(\hat{\beta}_2) \cdot \left(\widehat{\text{odds}} \text{ with } x_2 = 0 \right)$$

Going back to the R output, we say that:

- The predicted odds of a long commute are $e^{2.3628}$ times larger with precipitation versus without precipitation, assuming the same temperature.
- For every 1 °F increase in temperature, the predicted odds of a long commute change by a factor of $e^{-0.0108}$, assuming the same precipitation status.

Probit Regression and Complementary Log-Log Regression

We repeat the above regression using the probit link and the complementary log-log link. Here are the R outputs and fitted equations for both models:

	Estimate	Std. Error
(Intercept)	-0.7515	0.5022
Temp	-0.0066	0.0086
PrecipYes	1.4189	0.3103

$$\Phi^{-1}(\hat{q}) = -0.7515 - 0.0066x_1 + 1.4189x_2$$

	Estimate	Std. Error
(Intercept)	-1.5373	0.6515
Temp	-0.0073	0.0105
PrecipYes	1.8911	0.4342

$$\ln [-\ln (1 - \hat{q})] = -1.5373 - 0.0073x_1 + 1.8911x_2$$

For a day with a temperature of 32 °F and the presence of precipitation, predict the probability of a long commute under each of the three regression models.

Under the logistic regression model,

$$\begin{aligned} \ln \left(\frac{\hat{q}}{1 - \hat{q}} \right) &= -1.2834 - 0.0108 (32) + 2.3628 \\ &= 0.7338 \\ \hat{q} &= \frac{e^{0.7338}}{1 + e^{0.7338}} \\ &= \mathbf{0.6756} \end{aligned}$$

Under the probit regression model,

$$\begin{aligned} \Phi^{-1}(\hat{q}) &= -0.7515 - 0.0066 (32) + 1.4189 \\ &= 0.4562 \\ \hat{q} &= \Phi(0.46) \\ &= \mathbf{0.6772} \end{aligned}$$

Under the complementary log-log regression model,

$$\begin{aligned} \ln [-\ln (1 - \hat{q})] &= -1.5373 - 0.0073 (32) + 1.8911 \\ &= 0.1202 \\ \hat{q} &= 1 - \exp [-\exp (0.1202)] \\ &= \mathbf{0.6762} \end{aligned}$$

Example 3.9.1.2

You are given several numerical outputs from a logistic regression model with a Bernoulli response:

$$\hat{\beta} = \begin{bmatrix} 0.6085 \\ -0.1129 \end{bmatrix}, \quad \mathbf{y} = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \hat{\mu} = \begin{bmatrix} 0.253 \\ 0.212 \\ 0.322 \\ 0.567 \\ 0.253 \\ 0.177 \\ 0.212 \\ 0.232 \\ 0.595 \\ 0.177 \end{bmatrix}, \quad \mathbf{I} = \begin{bmatrix} 1.887 & 23.582 \\ 23.582 & 367.152 \end{bmatrix}$$

Determine which of the following statements are true.

- I. The Pearson residual of the third observation is 1.451.
- II. The deviance of this model is 5.631.
- III. The estimated standard error for the intercept is 2.686.
- IV. The likelihood ratio test statistic is 0.956.

Solution

Since the response follows Bernoulli, $\mu = mq = q$. Therefore, $\hat{\mu}$ is the vector of predicted "success" probabilities for each observation, i.e. $\hat{q}_i = \hat{\mu}_i$.

I is true. The Pearson residual of the third observation is

$$\begin{aligned} e_3^P &= \frac{y_3 - m_3 \hat{q}_3}{\sqrt{m_3 \hat{q}_3 (1 - \hat{q}_3)}} \\ &= \frac{1 - 0.322}{\sqrt{0.322 (1 - 0.322)}} \\ &= 1.451 \end{aligned}$$

II is false. The deviance is

$$\begin{aligned}
D &= 2 \sum_{i=1}^{10} \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) + (m_i - y_i) \ln \left(\frac{m_i - y_i}{m_i - \hat{\mu}_i} \right) \right] \\
&= 2 \sum_{i=1}^{10} \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) + (1 - y_i) \ln \left(\frac{1 - y_i}{1 - \hat{\mu}_i} \right) \right] \\
&= 2 \left[\ln \left(\frac{1}{1 - 0.253} \right) + \ln \left(\frac{1}{0.212} \right) + \dots + \ln \left(\frac{1}{1 - 0.177} \right) \right] \\
&= 2 (0.292 + 1.551 + \dots + 0.195) \\
&= 11.262
\end{aligned}$$

III is false. The estimated standard error for the intercept is the square root of the first diagonal entry of \mathbf{I}^{-1} .

$$\begin{aligned}
\mathbf{I}^{-1} &= \frac{1}{(1.887)(367.152) - 23.582^2} \begin{bmatrix} 367.152 & -23.582 \\ -23.582 & 1.887 \end{bmatrix} \\
&= \begin{bmatrix} 2.686 & -0.173 \\ -0.173 & 0.014 \end{bmatrix}
\end{aligned}$$

$$se(\hat{\beta}_0) = \sqrt{2.686} = 1.639$$

IV is true. The default likelihood ratio test compares the fitted model and the null model. Thus, the test statistic is

$$2 [l(\hat{\beta}) - l_{\text{null}}]$$

First, solve for the log-likelihood. To calculate the maximized log-likelihoods with the given information, leave the log-likelihood in terms of the q_i 's.

$$\begin{aligned}
l(\beta) &= \sum_{i=1}^{10} \left[y_i \ln \left(\frac{q_i}{1 - q_i} \right) + 1 \cdot \ln(1 - q_i) + \ln \left(\frac{1}{y_i} \right) \right] \\
&= \sum_{i=1}^{10} \left[y_i \ln \left(\frac{q_i}{1 - q_i} \right) + \ln(1 - q_i) \right]
\end{aligned}$$

So for the fitted model,

$$\begin{aligned}
l(\hat{\beta}) &= \sum_{i=1}^{10} \left[y_i \ln \left(\frac{\hat{q}_i}{1 - \hat{q}_i} \right) + \ln(1 - \hat{q}_i) \right] \\
&= \ln(1 - 0.253) + \left[\ln \left(\frac{0.212}{1 - 0.212} \right) + \ln(1 - 0.212) \right] + \dots + \ln(1 - 0.177) \\
&= -0.292 + (-1.551) + \dots + (-0.195) \\
&= -5.631
\end{aligned}$$

As for the null model, recall that $\mathbf{x}_i^T \beta = \beta_0$. Note that \hat{q}_i will be the same for all i , i.e.

$$\hat{q}_i = \frac{\exp(\mathbf{x}_i^T \hat{\beta})}{1 + \exp(\mathbf{x}_i^T \hat{\beta})} = \frac{\exp(\hat{\beta}_0)}{1 + \exp(\hat{\beta}_0)}$$

where $\hat{\beta}_0$ satisfies the score equation $u_0 = 0$. Since we want to compute l_{null} , we may solve for \hat{q}_i directly. This means we seek

$$\begin{aligned}
u_0 &= \sum_{i=1}^{10} (y_i - \mu_i) \underbrace{1}_{x_{i,0}} \\
&= \sum_{i=1}^{10} y_i - \sum_{i=1}^{10} q_i \\
&= 3 - 10q_i = 0
\end{aligned}$$

$$\hat{q}_i = \frac{3}{10}$$

So,

$$\begin{aligned}
l_{\text{null}} &= \sum_{i=1}^{10} \left[y_i \ln \left(\frac{\hat{q}_i}{1 - \hat{q}_i} \right) + \ln(1 - \hat{q}_i) \right] \\
&= \ln(1 - 0.3) + \left[\ln \left(\frac{0.3}{1 - 0.3} \right) + \ln(1 - 0.3) \right] + \dots + \ln(1 - 0.3) \\
&= 10 \ln 0.7 + 3 \ln \left(\frac{3}{7} \right) \\
&= -6.109
\end{aligned}$$

$$2 \left[l(\hat{\beta}) - l_{\text{null}} \right] = 2 [-5.631 - (-6.109)] = 0.956$$



Therefore, **only I and IV are true.**

Example 3.9.1.3

You are given the following result on a probit regression model that predicts the number of high-salaried players on a major league baseball team:

	Coefficients	Standard Error
Intercept	-2.7802	0.3284
Home runs	0.0153	0.0047
Wins	-0.0024	0.0081
Years	0.1062	0.0191

The Padang Pandas had 105 home runs, 78 wins, and 20 years of major league experience. Their roster consists of 52 players.

Estimate:

1. the odds that a Padang Pandas player is not high-salaried, and
2. the variance of the number of high-salaried players for the Padang Pandas.

Solution to (1)

With the probit link, the fitted equation is

$$\Phi^{-1}(\hat{q}) = -2.7802 + 0.0153x_1 - 0.0024x_2 + 0.1062x_3$$

Substitute in the inputs, and solve for the predicted probability that a Padang Pandas player is high-salaried.

$$\begin{aligned}\Phi^{-1}(\hat{q}) &= -2.7802 + 0.0153(105) - 0.0024(78) + 0.1062(20) \\ &= 0.7631\end{aligned}$$

$$\Rightarrow \hat{q} = \Phi(0.76) = 0.7764$$

The predicted odds that a Padang Pandas player is not high-salaried are

$$\frac{1 - 0.7764}{0.7764} = \mathbf{0.2880}$$



Solution to (2)

For a binomial distribution, the variance equals $mq(1 - q)$. Hence, the estimated variance is

$$52(0.7764)(1 - 0.7764) = \mathbf{9.03}$$



3.9.2 Nominal Response

We now consider modeling a nominal response variable that has more than two categories. Recall that a nominal variable consists of categories without a meaningful order.

The multinomial distribution is suitable for modeling the response since it has multiple categories. Let Y_1, \dots, Y_g follow a multinomial distribution where

- there are m "trials", and
- π_c is the probability of being in category c , for $c = 1, \dots, g$.

The joint PMF is

$$p(y_1, \dots, y_g) = \frac{m!}{y_1! \cdot \dots \cdot y_g!} \cdot \pi_1^{y_1} \cdot \dots \cdot \pi_g^{y_g}$$

and the marginal mean and variance of Y_c are

$$\mathbb{E}[Y_c] = m\pi_c, \quad \text{Var}[Y_c] = m\pi_c(1 - \pi_c)$$

In addition, note that

$$\sum_{c=1}^g Y_c = m, \quad \sum_{c=1}^g \pi_c = 1$$

This means one of the random variables and its associated probability are redundant. We may simplify by letting one category be the **reference category**, which is similar to the baseline category for factors. If we arbitrarily set category g as the reference category, then we have Y_1, \dots, Y_{g-1} following a multinomial distribution where the joint PMF replaces y_g and π_g with

$$y_g = m - \sum_{c=1}^{g-1} y_c, \quad \pi_g = 1 - \sum_{c=1}^{g-1} \pi_c$$

In summary, a nominal response variable consisting of g categories is represented by $g - 1$ variables when using the multinomial distribution. While this is the first time discussing a

multivariate response, and thus must deviate from the usual GLM framework, we borrow key ideas from logistic regression to estimate the parameters of interest π_1, \dots, π_g .

Nominal Logistic Regression

Let $\pi_{i,c}$ be the probability that the i^{th} observation is classified as category c . In a *nominal logistic regression* model, we let

$$\ln \left(\frac{\pi_{i,t}}{\pi_{i,k}} \right) = \mathbf{x}_i^T \boldsymbol{\beta}_t \quad (3.9.2.1)$$

where k is the reference category and the coefficient vector $\boldsymbol{\beta}_t$ varies depending on category t , i.e.

$$\boldsymbol{\beta}_t = \begin{bmatrix} \beta_{0,t} \\ \beta_{1,t} \\ \vdots \\ \beta_{p,t} \end{bmatrix}$$

Category t is a category that is not the reference category, so t cannot equal k . In this model, we have Equation 3.9.2.1 for each category t , producing $g - 1$ equations and a total of $(p + 1)(g - 1)$ parameters to estimate by MLE. From those $g - 1$ equations and $\sum_{c=1}^g \pi_{i,c} = 1$, we can derive

$$\pi_{i,c} = \begin{cases} \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_c)}{1 + \sum_{\text{all } t} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_t)}, & c \neq k \\ \frac{1}{1 + \sum_{\text{all } t} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_t)}, & c = k \end{cases} \quad (3.9.2.2)$$

Coach's Remarks

Even though $\pi_{i,t}/\pi_{i,k}$ is **not** an odds by definition, calling it an "odds" is common because of the parallels between logistic regression and nominal logistic regression. So, it is not surprising for "log odds" and "odds ratio" to be also used in nominal logistic regression.

To motivate the nominal logistic regression, we revisit the Commuting Chris scenario. We want to predict whether a commute is short, medium, or long using only Temp and Precip.

A commute that takes less than 23 minutes is considered a short commute. A commute that takes more than 26 minutes is considered a long commute. A commute between those durations is considered a medium commute. We have a multinomial response with $g = 3$ and $m_1 = \dots = m_{100} = 1$.

With long commute as the reference category (i.e. $k = 3$), the two equations for this model are

$$\ln \left(\frac{\pi_1}{\pi_3} \right) = \beta_{0,1} + \beta_{1,1}x_1 + \beta_{2,1}x_2$$

$$\ln \left(\frac{\pi_2}{\pi_3} \right) = \beta_{0,2} + \beta_{1,2}x_1 + \beta_{2,2}x_2$$

Running the model produces the following R output:

	Estimate	Std. Error
(Intercept) : 1	0.5824	1.0954
(Intercept) : 2	0.2630	0.9219
Temp : 1	0.0169	0.0197
Temp : 2	0.0083	0.0158
PrecipYes : 1	-4.5622	1.1094
PrecipYes : 2	-1.3177	0.5825

From this output, we obtain the following two fitted equations:

$$\ln \left(\frac{\hat{\pi}_1}{\hat{\pi}_3} \right) = 0.5824 + 0.0169x_1 - 4.5622x_2$$

$$\ln \left(\frac{\hat{\pi}_2}{\hat{\pi}_3} \right) = 0.2630 + 0.0083x_1 - 1.3177x_2$$

Using these equations, we can estimate the probabilities for given inputs of the predictors. For example, if the temperature is 32 °F and there is precipitation during the commute, then the

equations reduce to

$$\ln \left(\frac{\hat{\pi}_1}{\hat{\pi}_3} \right) = -3.439, \quad \ln \left(\frac{\hat{\pi}_2}{\hat{\pi}_3} \right) = -0.7891$$

which we can use to solve for the predicted probabilities. With $\hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3 = 1$, solving this system of equations yields

$$\hat{\pi}_1 = 0.0216, \quad \hat{\pi}_2 = 0.3056, \quad \hat{\pi}_3 = 0.6728$$

Alternatively, Equation 3.9.2.2 can be used to estimate the probabilities directly, e.g.

$$\hat{\pi}_1 = \frac{\exp(-3.439)}{1 + \exp(-3.439) + \exp(-0.7891)} = 0.0216$$

Example 3.9.2.1

The following predictors are included in a nominal logistic regression model:

- Age – a continuous variable.
- Territory – a categorical variable with four categories.
- Gender – a categorical variable with two categories.
- Interaction terms between Territory and Gender.

These are used to predict the risk class of auto insurance policyholders: Preferred (1), Standard (2), Substandard (3), and Risky (4). Risky is the reference category.

Determine which of the following statements are true.

- I. A total of nine coefficients are estimated in this model.
- II. If $\hat{\beta}_{j,3} = 0.5$ where x_j represents Age, the estimated probability of being Substandard increases by a factor of $e^{0.5}$ for every unit increase in Age, all else being equal.
- III. If Preferred were to be the reference category, the model would produce different predicted probabilities.

Solution

I is false. We actually need to estimate 27 coefficients. This is because each of the $g - 1 = 4 - 1 = 3$ equations have nine coefficients:

- 1 for the intercept,
- 1 for Age,
- $4 - 1 = 3$ for Territory,
- $2 - 1 = 1$ for Gender, and
- $(4 - 1)(2 - 1) = 3$ interaction terms.

II is false. Interpreting the coefficient estimates is similar to the logistic regression interpretation. Risky as the reference category means $k = 4$. Thus, for c being any fixed value of Age,

$$\left(\frac{\hat{\pi}_3}{\hat{\pi}_4} \text{ with } x_j = c + 1 \right) = e^{0.5} \cdot \left(\frac{\hat{\pi}_3}{\hat{\pi}_4} \text{ with } x_j = c \right)$$

It is the "odds" $\hat{\pi}_3 / \hat{\pi}_4$ that increases by a factor of $e^{0.5}$, not $\hat{\pi}_3$. Another way to see this is by inspecting Equation 3.9.2.1: the change in $\hat{\pi}_3$ depends on $\hat{\pi}_4$, which requires knowing more than just $\hat{\beta}_{j,3}$.

III is false. Regardless of which category is chosen as the reference category, the predicted probabilities will always be the same. However, the coefficient estimates will be different because the coefficients themselves have different interpretations. This is analogous to choosing a baseline category for factors.

Therefore, **none of the statements are true.**

3.9.3 Ordinal Response

(L) 15m

To model an ordinal response variable, we may keep the multinomial distribution. Since an ordinal variable has a meaningful order to the categories, the model should attempt to capture this detail. One way is to focus on modeling the cumulative probabilities of the categories. Denote Π_c as

$$\Pi_c = \pi_1 + \dots + \pi_c$$

Since $\Pi_g = \sum_{c=1}^g \pi_c = 1$, we only need to model Π_1, \dots, Π_{g-1} .

Proportional Odds Model

In a *proportional odds model*, the coefficients for each predictor are assumed to not vary by category, but the intercept does. Let $\beta_{0,c}$ be the intercept for category c , and make slight adjustments to the \mathbf{x}_i and $\boldsymbol{\beta}$ vectors, i.e.

$$\mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

Then, we use

$$\ln \left(\frac{\Pi_{i,c}}{1 - \Pi_{i,c}} \right) = \beta_{0,c} + \mathbf{x}_i^T \boldsymbol{\beta} \quad (3.9.3.1)$$

to model the cumulative probabilities. Note that $g - 1$ equations are needed to model Π_1, \dots, Π_{g-1} . Along with the p coefficients in $\boldsymbol{\beta}$, there are $g - 1 + p$ parameters to be estimated.

Coach's Remarks

While category g seems to operate like the reference category, the idea of a reference category does not properly exist in this scenario. None of the components of the proportional odds model can be interpreted with category g as a point of reference to the other categories.

Consider the same Commuting Chris setup from the previous subsection, except now we treat the response as ordinal. The two equations for this proportional odds model are

$$\ln \left(\frac{\Pi_1}{1 - \Pi_1} \right) = \beta_{0,1} + \beta_1 x_1 + \beta_2 x_2$$

$$\ln \left(\frac{\Pi_2}{1 - \Pi_2} \right) = \beta_{0,2} + \beta_1 x_1 + \beta_2 x_2$$

Running the model produces the following summary outputs:

	Estimate	Std. Error
Temp	0.0084	0.0124
PrecipYes	-2.7625	0.5104

	Estimate	Std. Error
Short	-0.1331	0.7160
Medium	1.7084	0.7567

The second table provides details on the two intercepts. From this output, we get the following two fitted equations:

$$\ln \left(\frac{\hat{\Pi}_1}{1 - \hat{\Pi}_1} \right) = \ln \left(\frac{\hat{\pi}_1}{\hat{\pi}_2 + \hat{\pi}_3} \right) = -0.1331 + 0.0084x_1 - 2.7625x_2$$

$$\ln \left(\frac{\hat{\Pi}_2}{1 - \hat{\Pi}_2} \right) = \ln \left(\frac{\hat{\pi}_1 + \hat{\pi}_2}{\hat{\pi}_3} \right) = 1.7084 + 0.0084x_1 - 2.7625x_2$$

Using these equations, we can estimate the probabilities given inputs for the predictors. For example, if the temperature is 32 °F and there is precipitation during the commute, then the equations reduce to

$$\ln \left(\frac{\hat{\pi}_1}{\hat{\pi}_2 + \hat{\pi}_3} \right) = -2.6268, \quad \ln \left(\frac{\hat{\pi}_1 + \hat{\pi}_2}{\hat{\pi}_3} \right) = -0.7853$$

which we can use to solve for the predicted probabilities. With $\hat{\pi}_1 + \hat{\pi}_2 + \hat{\pi}_3 = 1$, solving this system of equations yields

$$\hat{\pi}_1 = 0.0674, \quad \hat{\pi}_2 = 0.2457, \quad \hat{\pi}_3 = 0.6868$$

In general, note that the fitted equations result in the following predicted odds:

$$\begin{aligned} \frac{\hat{\Pi}_1}{1 - \hat{\Pi}_1} &= \exp \left(\hat{\beta}_{0,1} + \mathbf{x}^T \hat{\boldsymbol{\beta}} \right) \\ \frac{\hat{\Pi}_2}{1 - \hat{\Pi}_2} &= \exp \left(\hat{\beta}_{0,2} + \mathbf{x}^T \hat{\boldsymbol{\beta}} \right) \\ &\vdots \\ \frac{\hat{\Pi}_{g-1}}{1 - \hat{\Pi}_{g-1}} &= \exp \left(\hat{\beta}_{0,g-1} + \mathbf{x}^T \hat{\boldsymbol{\beta}} \right) \end{aligned}$$

Because these are odds of cumulative probabilities, they are also known as **cumulative odds**.

Given a specific set of predictor inputs, notice that if we take the ratio of any two cumulative odds, the resulting value is not dependent on $\mathbf{x}^T \hat{\boldsymbol{\beta}}$, e.g.

$$\frac{\hat{\Pi}_1 \div (1 - \hat{\Pi}_1)}{\hat{\Pi}_2 \div (1 - \hat{\Pi}_2)} = \exp \left(\hat{\beta}_{0,1} + \mathbf{x}^T \hat{\boldsymbol{\beta}} - \hat{\beta}_{0,2} - \mathbf{x}^T \hat{\boldsymbol{\beta}} \right) = \exp \left(\hat{\beta}_{0,1} - \hat{\beta}_{0,2} \right)$$

Thus, in calculating the same ratio of cumulative odds for **another** set of predictor inputs, the result is unchanged. This relationship allows for potentially solving a $\hat{\Pi}_c$ without needing the coefficient estimates. We will demonstrate this in the following example.

Example 3.9.3.1

The proportional odds model is used to predict the risk class of a policyholder as Preferred, Standard, Substandard, or Risky. Only one categorical variable, Territory, is used in the model.

You are given the following information:

- The estimated probability of being a Preferred policyholder from Territory A is 0.36.
- The estimated probability of being a Preferred or Standard policyholder from Territory A is 0.68.
- The estimated probability of being a Risky policyholder from Territory A is 0.07.
- The estimated probability of being a Preferred policyholder from Territory B is 0.285.

Determine which of the following statements are true.

- I. The estimated odds of being a Substandard Territory A policyholder are 1 to 3.
- II. The estimated probability of being a Standard Territory B policyholder is 0.316.
- III. The estimated probability of being a Risky Territory B policyholder is 0.32.

Solution

Let π_c and Π_c be the probability and cumulative probability, respectively, of being in category c . Let

$$c = \begin{cases} 1, & \text{for Preferred} \\ 2, & \text{for Standard} \\ 3, & \text{for Substandard} \\ 4, & \text{for Risky} \end{cases}$$

We are given these estimated probabilities for a Territory A policyholder:

$$\hat{\pi}_1 = 0.36$$

$$\hat{\Pi}_2 = 0.68$$

$$\hat{\pi}_4 = 0.07$$

I is true. The estimated probability of being a Substandard policyholder from Territory A is

$$\begin{aligned}\hat{\pi}_3 &= 1 - \hat{\pi}_1 - \hat{\pi}_2 - \hat{\pi}_4 \\ &= 1 - \hat{\Pi}_2 - \hat{\pi}_4 \\ &= 1 - 0.68 - 0.07 \\ &= 0.25\end{aligned}$$

Then, the estimated odds are

$$\frac{0.25}{1 - 0.25} = \frac{1}{3}$$

II is true. Recall that the same ratio of cumulative odds is unchanged even for different predictor values. This means predicting for Territory A versus Territory B will lead to equivalent ratios. Then, the estimated probability of being a Preferred or Standard policyholder from Territory B is

$$\begin{aligned}\frac{0.68 \div (1 - 0.68)}{0.36 \div (1 - 0.36)} &= \frac{\hat{\Pi}_2 \div (1 - \hat{\Pi}_2)}{0.285 \div (1 - 0.285)} \\ \Rightarrow \hat{\Pi}_2 &= 0.6009\end{aligned}$$

Hence, the estimated probability of being a Standard policyholder from Territory B is

$$0.6009 - 0.285 = 0.3159$$

III is false. The estimated probability of being a Preferred, Standard, or Substandard policyholder from Territory B is

$$\begin{aligned}\frac{0.93 \div (1 - 0.93)}{0.36 \div (1 - 0.36)} &= \frac{\hat{\Pi}_3 \div (1 - \hat{\Pi}_3)}{0.285 \div (1 - 0.285)} \\ \Rightarrow \hat{\Pi}_3 &= 0.9040\end{aligned}$$

Hence, the estimated probability of being a Risky policyholder from Territory B is

$$1 - 0.9040 = 0.0960$$

Therefore, **only I and II are true.**



3.9 Summary

🕒 5m

Binomial Response Variable

- The odds of an event are the ratio of the probability that the event will occur to the probability that the event will not occur, i.e.

$$\text{odds} = \frac{q}{1-q}$$

- The odds ratio is the ratio of the odds of an event with the presence of a characteristic to the odds of the same event without the presence of that characteristic.
- Link functions:

Function Name	$g(q)$
Logit link	$\ln\left(\frac{q}{1-q}\right)$
Probit link	$\Phi^{-1}(q)$
Complementary log-log link	$\ln[-\ln(1-q)]$

- Log-likelihood:

$$l(\beta) = \sum_{i=1}^n \left[y_i \ln\left(\frac{q_i}{1-q_i}\right) + m_i \ln(1-q_i) + \ln\left(\frac{m_i}{y_i}\right) \right]$$

- Deviance:

$$D = 2 \sum_{i=1}^n \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (m_i - y_i) \ln\left(\frac{m_i - y_i}{m_i - \hat{\mu}_i}\right) \right]$$

- Pearson residual:

$$e_i^P = \frac{y_i - m_i \hat{q}_i}{\sqrt{m_i \hat{q}_i (1 - \hat{q}_i)}}$$

- Pearson chi-square statistic:

$$\sum_{i=1}^n \frac{(y_i - m_i \hat{q}_i)^2}{m_i \hat{q}_i (1 - \hat{q}_i)}$$

LOGISTIC REGRESSION

- Uses the logit link function.

$$q_i = \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta})}{1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta})}$$

- Score function:

$$u_j = \sum_{i=1}^n (y_i - \mu_i) x_{i,j}$$

- Information matrix:

$$\mathbf{I} = \sum_{i=1}^n m_i q_i (1 - q_i) \mathbf{x}_i \mathbf{x}_i^T$$

Nominal Response Variable

Under nominal logistic regression,

$$\ln \left(\frac{\pi_{i,t}}{\pi_{i,k}} \right) = \mathbf{x}_i^T \boldsymbol{\beta}_t$$

$$\pi_{i,c} = \begin{cases} \frac{\exp(\mathbf{x}_i^T \boldsymbol{\beta}_c)}{1 + \sum_{\text{all } t} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_t)}, & c \neq k \\ \frac{1}{1 + \sum_{\text{all } t} \exp(\mathbf{x}_i^T \boldsymbol{\beta}_t)}, & c = k \end{cases}$$

Ordinal Response Variable

Under the proportional odds model,

$$\ln \left(\frac{\Pi_{i,c}}{1 - \Pi_{i,c}} \right) = \beta_{0,c} + \mathbf{x}_i^T \boldsymbol{\beta}$$

where

$$\Pi_c = \pi_1 + \dots + \pi_c, \quad \mathbf{x}_i = \begin{bmatrix} x_{i,1} \\ \vdots \\ x_{i,p} \end{bmatrix}, \quad \boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}$$

A ratio of cumulative odds is not a function of the predictor values, e.g.

$$\frac{\hat{\Pi}_1 \div (1 - \hat{\Pi}_1)}{\hat{\Pi}_2 \div (1 - \hat{\Pi}_2)} = \exp(\hat{\beta}_{0,1} - \hat{\beta}_{0,2})$$

Appendix**Deviance for Binomial Response**

The binomial log-likelihood is

$$\begin{aligned}
 l(\beta) &= \sum_{i=1}^n \left[y_i \ln \left(\frac{q_i}{1-q_i} \right) + m_i \ln (1-q_i) + \ln \binom{m_i}{y_i} \right] \\
 &= \sum_{i=1}^n \left[y_i \ln \left(\frac{\mu_i}{1-\frac{\mu_i}{m_i}} \right) + m_i \ln \left(1 - \frac{\mu_i}{m_i} \right) + \ln \binom{m_i}{y_i} \right] \\
 &= \sum_{i=1}^n \left[y_i \ln \left(\frac{\mu_i}{m_i - \mu_i} \right) + m_i \ln \left(\frac{m_i - \mu_i}{m_i} \right) + \ln \binom{m_i}{y_i} \right] \\
 &= \sum_{i=1}^n \left[y_i \ln \mu_i + (m_i - y_i) \ln (m_i - \mu_i) - m_i \ln m_i + \ln \binom{m_i}{y_i} \right]
 \end{aligned}$$

Under the fitted model:

$$l(\hat{\beta}) = \sum_{i=1}^n \left[y_i \ln \hat{\mu}_i + (m_i - y_i) \ln (m_i - \hat{\mu}_i) - m_i \ln m_i + \ln \binom{m_i}{y_i} \right]$$

Under the saturated model:

$$\hat{\mu}_i = y_i$$

$$l_{\text{sat}} = \sum_{i=1}^n \left[y_i \ln y_i + (m_i - y_i) \ln (m_i - y_i) - m_i \ln m_i + \ln \binom{m_i}{y_i} \right]$$

Therefore, the deviance is

$$\begin{aligned}
 D &= 2 \left[l_{\text{sat}} - l(\hat{\beta}) \right] \\
 &= 2 \left\{ \sum_{i=1}^n \left[y_i \ln y_i + (m_i - y_i) \ln (m_i - y_i) - m_i \ln m_i + \ln \binom{m_i}{y_i} \right] - \sum_{i=1}^n \left[y_i \ln \hat{\mu}_i + (m_i - y_i) \ln (m_i - \hat{\mu}_i) \right] \right\} \\
 &= 2 \sum_{i=1}^n \left[y_i \ln y_i + (m_i - y_i) \ln (m_i - y_i) - y_i \ln \hat{\mu}_i - (m_i - y_i) \ln (m_i - \hat{\mu}_i) \right] \\
 &= 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) + (m_i - y_i) \ln \left(\frac{m_i - y_i}{m_i - \hat{\mu}_i} \right) \right]
 \end{aligned}$$

3.10.0 Overview

 5m

We now focus on modeling count response variables using the Poisson distribution. Exposures and contingency tables help us better understand the data; as such, they are key components of Poisson regression and log-linear models, respectively.

3.10.1 Poisson Regression

While it was proven in Section 2.2.5 that the Poisson distribution belongs to the exponential family, we repeat the steps here for convenience. The Poisson PMF can be written as

$$\begin{aligned} p(y) &= \frac{e^{-\lambda} \lambda^y}{y!} \\ &= \exp[-\lambda] \cdot \exp\left[\ln\left(\frac{\lambda^y}{y!}\right)\right] \\ &= \exp[-\lambda + \ln(\lambda^y) - \ln(y!)] \\ &= \exp[y \ln \lambda - \lambda - \ln(y!)] \end{aligned}$$

where

- $a(y) = y$
- $b(\lambda) = \ln \lambda$
- $c(\lambda) = -\lambda$
- $d(y) = -\ln(y!)$

Moreover, recall that

$$\text{E}[Y] = \lambda, \quad \text{Var}[Y] = \lambda$$

Since Y has an exponential family distribution, these formulas agree with Equations 3.8.1.1 and 3.8.1.2.

Exposures

A Poisson distribution models the frequency of an event within a specified scope. This scope is defined and measured in ***units of exposure***.

To motivate the concept of exposures, consider the Commuting Chris scenario. The variable Police is a count variable, which we view as a Poisson response in this situation. In counting the number of police vehicles along the route of commute, consider two possible cases:

- If Chris commutes to the same location by the same route each day, then the distance traveled is fixed. Using the commute distance as a measure of exposure, this is an example of equal exposures for all observations.
- If Chris commutes to different locations and/or uses different routes on certain days, then the distance traveled would vary. Using the commute distance as a measure of exposure, this is an example of varying exposures across the observations.

In the first case, we may ignore exposures altogether if our analysis assumes the same exposure amount throughout. However, the second case requires a model to account for exposures.

Poisson Regression

Let Y_1, \dots, Y_n be independent Poisson random variables with mean

$$\mu_i = a_i \lambda_i$$

where

- a_i is the exposure amount for the i^{th} observation, and
- λ_i is the mean per exposure for the i^{th} observation.

Using the logarithmic (or log) link function, we get

$$\begin{aligned} \ln \mu_i &= \ln (a_i \lambda_i) \\ &= \ln a_i + \ln \lambda_i \\ &= \ln a_i + \mathbf{x}_i^T \boldsymbol{\beta} \end{aligned}$$

$$\Rightarrow \mu_i = a_i \cdot \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

This means the predictors do not directly explain the Poisson mean at the observation level. Instead, the predictors explain the Poisson mean **per exposure**, which is then multiplied by an exposure amount to compute the Poisson mean for an observation.

The term $\ln a_i$ is called an **offset** since it adds to the usual linear component. One way to interpret the offset is as a predictor with a regression coefficient of 1.

Then, we obtain the following simplified expressions:

$$\hat{\mu}_i = a_i \cdot \exp\left(\mathbf{x}_i^T \hat{\beta}\right)$$

- Log-likelihood function:

$$l(\beta) = \sum_{i=1}^n [y_i \ln \mu_i - \mu_i - \ln(y_i!)] \quad (3.10.1.1)$$

- Score function:

$$u_j = \sum_{i=1}^n (y_i - \mu_i)x_{i,j} \quad (3.10.1.2)$$

- Information matrix:

$$\mathbf{I} = \sum_{i=1}^n \mu_i \mathbf{x}_i \mathbf{x}_i^T \quad (3.10.1.3)$$

- Deviance:

$$D = 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right] \quad (3.10.1.4)$$

If interested in the proof, see the appendix at the end of the section. Moreover, since the MLE estimates require

$$u_0 = \sum_{i=1}^n (y_i - \mu_i)$$

to equal 0, the deviance simplifies to

$$D = 2 \sum_{i=1}^n y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) \quad (3.10.1.5)$$

- Pearson residual:

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}} \quad (3.10.1.6)$$

- Pearson chi-square statistic:

$$\sum_{i=1}^n (e_i^P)^2 = \sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} \quad (3.10.1.7)$$

Coach's Remarks

The deviance formula of Equation 3.10.1.4 is relevant for two reasons:

- It helps define the deviance residual, i.e.

$$e_i^D = \pm \sqrt{2 \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right]} \\ \neq \pm \sqrt{2 \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) \right]}$$

- As explained, Equation 3.10.1.5 originates from estimating the intercept β_0 by MLE. Even if the linear component drops the intercept, Equation 3.10.1.4 remains the deviance formula.

INTERPRETATION OF PARAMETER ESTIMATES

Let's use the Commuting Chris scenario with Police as the response and commute distance in miles as the exposure unit. Here, we have

$$\ln \mu = \ln a + \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

where

- a represents commute distance,
- x_1 represents Temp, and
- x_2 represents the dummy variable for Precip.

Here is the R output from the Poisson regression:

	Estimate	Std. Error
(Intercept)	-1.8731	0.4174
Temp	-0.0292	0.0079
PrecipYes	0.5666	0.2706

This gives the following fitted equation:

$$\ln \hat{\mu} = \ln a - 1.8731 - 0.0292x_1 + 0.5666x_2$$

As an example, for a 12.6-mile commute with a temperature of 50 °F and no precipitation, the estimated mean of police vehicles is

$$\begin{aligned}\ln \hat{\mu} &= \ln 12.6 - 1.8731 - 0.0292(50) + 0 \\ &= \ln 12.6 - 3.3331 \\ \hat{\mu} &= 12.6 \cdot e^{-3.3331} \\ &= 0.45\end{aligned}$$

Next, take the ratio of $\hat{\mu}$ with precipitation versus without precipitation. This ratio is called **rate ratio**.

$$x_2 = 1 : \hat{\mu} = a \cdot \exp(-1.8731 - 0.0292x_1 + 0.5666)$$

$$x_2 = 0 : \hat{\mu} = a \cdot \exp(-1.8731 - 0.0292x_1)$$

$$\text{ratio} : \frac{a \cdot \exp(-1.8731 - 0.0292x_1 + 0.5666)}{a \cdot \exp(-1.8731 - 0.0292x_1)} = \exp(0.5666) = \exp(\hat{\beta}_2)$$

$$(\hat{\mu} \text{ with } x_2 = 1) = \exp(\hat{\beta}_2) \cdot (\hat{\mu} \text{ with } x_2 = 0)$$

In general, the estimated Poisson mean changes by a factor of $\exp(\hat{\beta}_j)$ per unit increase in x_j , assuming all other predictors and the exposure amount are held constant. Specifically for this Commuting Chris model, we say that:

- The estimated mean of police vehicles is $e^{0.5666}$ times larger with precipitation versus without precipitation, assuming the same temperature and commute distance.
- For every 1 °F increase in temperature, the estimated mean of police vehicles changes by a factor of $e^{-0.0292}$, assuming the same precipitation status and commute distance.

Example 3.10.1.1

You use the following data to run a Poisson regression:

Observation	Response	Feature	Exposure
1	0	66.7	11.9
2	1	84.4	13.1
3	4	98.0	13.4

The mean estimates for observations 1 and 3 are 0.0881 and 4.1147, respectively.

Calculate the Pearson residual of the second observation.

Solution

For this model, we have

$$\ln \mu = \ln a + \beta_0 + \beta_1 x \quad \Rightarrow \quad \hat{\mu} = a \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 x)$$

where a denotes the exposures and x denotes the feature. Thus, calculating a Pearson residual requires the values of $\hat{\beta}_0$ and $\hat{\beta}_1$.

We are given

$$11.9 \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 66.7) = 0.0881$$

$$13.4 \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 98.0) = 4.1147$$

Divide the two equations and solve for the coefficient estimates.

$$\frac{11.9 \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 66.7)}{13.4 \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 \cdot 98.0)} = \frac{0.0881}{4.1147}$$

$$\frac{\exp(\hat{\beta}_0) \exp(66.7\hat{\beta}_1)}{\exp(\hat{\beta}_0) \exp(98\hat{\beta}_1)} = \frac{0.0881}{4.1147} \cdot \frac{13.4}{11.9}$$

$$\exp[\hat{\beta}_1(66.7 - 98)] = 0.0241$$

$$\hat{\beta}_1 = -\frac{\ln 0.0241}{31.3}$$

$$= 0.119$$

$$11.9 \cdot \exp[\hat{\beta}_0 + 0.119(66.7)] = 0.0881$$

$$\exp(\hat{\beta}_0) = \frac{0.0881}{11.9 \cdot e^{0.119(66.7)}}$$

$$\hat{\beta}_0 = \ln \left[\frac{0.0881}{11.9 \cdot e^{0.119(66.7)}} \right]$$

$$= -12.844$$

As a result, the Pearson residual of the second observation is

$$\begin{aligned}
 e_2^P &= \frac{y_2 - \hat{\mu}_2}{\sqrt{\hat{\mu}_2}} \\
 &= \frac{y_2 - a_2 \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 x_2)}{\sqrt{a_2 \cdot \exp(\hat{\beta}_0 + \hat{\beta}_1 x_2)}} \\
 &= \frac{1 - 13.1 \cdot \exp[-12.844 + 0.119(84.4)]}{\sqrt{13.1 \cdot \exp[-12.844 + 0.119(84.4)]}} \\
 &= \mathbf{0.227}
 \end{aligned}$$

■

Example 3.10.1.2

In a basketball clinic, children are tasked to score as many three-point baskets as possible within 10 seconds. You are given the following information to predict the mean number of three-point baskets:

Response variable	3-point basket count
Response distribution	Poisson
Link	Log

	Coefficients	Standard Error
Intercept	-1.9903	0.62364
Proficiency Rating	0.0178	0.01035
Height		
Short	-0.2523	0.81324
Average	0.0000	—
Tall	2.0417	0.61202

On his attempt, Ricardo scored 1 three-point basket. He is classified as average height and had a proficiency rating of 71.

Determine which statements are true.

- I. At the same proficiency rating, the estimated mean number of three-point baskets for a tall participant is 2.3 times greater than for a short participant.
- II. The deviance residual corresponding to Ricardo is 0.65.
- III. A Wald test shows that the mean number of three-point baskets is plausibly the same between short and average height participants at the 5% significance level.

Solution

Since no exposure unit was specified, we assume equal exposures for all observations, i.e. $a = 1$.

Let

- x_1 represent Proficiency, and
- x_2 and x_3 represent the dummy variables for Short and Tall, respectively.

For this Poisson regression, we have

$$\ln \mu = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where μ is the mean number of three-point baskets. Therefore, the fitted equation for the mean is

$$\hat{\mu} = \exp(-1.9903 + 0.0178x_1 - 0.2523x_2 + 2.0417x_3)$$

I is false. The correct factor is calculated as $\hat{\mu}$ for a tall participant divided by $\hat{\mu}$ for a short participant, all while keeping x_1 constant.

$$\begin{aligned} \frac{\exp(-1.9903 + 0.0178x_1 + 2.0417)}{\exp(-1.9903 + 0.0178x_1 - 0.2523)} &= \exp[2.0417 - (-0.2523)] \\ &= e^{2.3} \end{aligned}$$

Therefore, the estimated mean for a tall participant is $e^{2.3}$ (not 2.3) times greater than for a short participant.

II is true. From Equation 3.8.4.7, a deviance residual is $\pm\sqrt{D_i}$, whose sign follows the i^{th} raw residual. Recall that D_i is the version of the deviance formula (Equation 3.10.1.4) for only one observation. This means we need

$$y_i = 1$$

$$\hat{\mu}_i = \exp [-1.9903 + 0.0178 (71) + 0] = 0.4836$$

Since the raw residual, i.e. $y_i - \hat{\mu}_i$, is positive, the deviance residual corresponding to Ricardo is

$$+\sqrt{2 \left[1 \cdot \ln \left(\frac{1}{0.4836} \right) - (1 - 0.4836) \right]} = 0.6482$$

III is true. The hypotheses for this test are $H_0 : \beta_2 = 0$ versus $H_1 : \beta_2 \neq 0$. The test statistic for this Wald test is

$$\left(\frac{-0.2523 - 0}{0.81324} \right)^2 = 0.096$$

With $\alpha = 0.05$, the critical value is $\chi^2_{0.95, 1} = 3.84$. Since $0.096 < 3.84$, we fail to reject H_0 . As β_2 is plausibly 0, short and average height participants may have the same mean.

Therefore, **only II and III are true.**



Example 3.10.1.3

50 observations are used to model two Poisson regressions and the null model. You are given with the following outputs:

-

	Null model	Model A	Model B
Number of predictors	0	3	5
$\sum_{i=1}^{50} y_i \ln \hat{\mu}_i$	1,259.203	1,280.007	1,283.107

- $\sum_{i=1}^{50} y_i = 469$
- $\sum_{i=1}^{50} \ln(y_i!) = 909.3959$

Determine which statement is true regarding model fit.

- The BIC for Model A is 214.0.
- The AIC for Model B is 200.6.
- Based on pseudo R^2 , Model A is preferred.
- Based on BIC, Model A is preferred.
- Based on AIC, Model A is preferred.

Solution

To calculate AIC, BIC, and pseudo R^2 , we need l_{null} and the maximized log-likelihood for both Models A and B. Using Equation 3.10.1.1, the formula for the maximized log-likelihood is

$$\begin{aligned} l(\hat{\beta}) &= \sum_{i=1}^{50} [y_i \ln \hat{\mu}_i - \hat{\mu}_i - \ln(y_i!)] \\ &= \sum_{i=1}^{50} y_i \ln \hat{\mu}_i - \sum_{i=1}^{50} \hat{\mu}_i - \sum_{i=1}^{50} \ln(y_i!) \end{aligned}$$

By default, GLMs are assumed to include the intercept in the linear component. As a consequence, satisfying $u_0 = 0$ means

$$\begin{aligned} \sum_{i=1}^{50} (y_i - \hat{\mu}_i) &= 0 \\ \sum_{i=1}^{50} y_i - \sum_{i=1}^{50} \hat{\mu}_i &= 0 \\ \sum_{i=1}^{50} \hat{\mu}_i &= \sum_{i=1}^{50} y_i \\ &= 469 \end{aligned}$$

Next, calculate the maximized log-likelihoods for all three models.

- Null model:

$$l_{\text{null}} = 1,259.203 - 469 - 909.3959 = -119.1929$$

- Model A:

$$l(\hat{\beta}) = 1,280.007 - 469 - 909.3959 = -98.3889$$

- Model B:

$$l(\hat{\beta}) = 1,283.107 - 469 - 909.3959 = -95.2889$$

Use Equations 3.8.4.2 to 3.8.4.4 to compute pseudo R^2 , AIC, and BIC for both Models A and B. In a Poisson regression, the parameters needing estimation are β_0, \dots, β_p only. Thus, the number of estimated parameters equals $p + 1$, where p is the number of predictors. These values are tabulated below:

	Model A	Model B
Pseudo R^2	0.1745	0.2005
AIC	204.8	202.6
BIC	212.4	214.0

Models with higher pseudo R^2 are preferred, whereas models with lower AIC/BIC are preferred. Therefore, the answer is (D).



3.10.2 Log-Linear Models

A contingency table is a useful way to organize and study two categorical variables (revisit Section 2.5.3 if needed). To model the counts in each cell of the contingency table, ***log-linear models*** are used. This means the response of a log-linear model is completely defined by the two factors. However, the main objective of these models is to assess the presence of association/dependence between the two factors, i.e. whether they have a significant interaction.

Model

Define Factor A as having w levels and Factor B as having v levels, such that they create a $w \times v$ contingency table. Let $Y_{j,k}$ be the count for cell $\{j, k\}$ of the contingency table. Furthermore, let

$$Y_{j\bullet} = \sum_{k=1}^v Y_{j,k}$$

$$Y_{\bullet k} = \sum_{j=1}^w Y_{j,k}$$

$$Y_{\bullet\bullet} = \sum_{k=1}^v \sum_{j=1}^w Y_{j,k}$$

These random variables can be organized in a contingency table.

		Factor B				Total
		1	2	...	v	
Factor A	1	$Y_{1,1}$	$Y_{1,2}$...	$Y_{1,v}$	$Y_{1\bullet}$
	2	$Y_{2,1}$	$Y_{2,2}$...	$Y_{2,v}$	$Y_{2\bullet}$
	:	:	:	..	:	:
	w	$Y_{w,1}$	$Y_{w,2}$...	$Y_{w,v}$	$Y_{w\bullet}$
Total		$Y_{\bullet 1}$	$Y_{\bullet 2}$...	$Y_{\bullet v}$	$Y_{\bullet \bullet}$

In addition, the $Y_{j,k}$'s are independent and each follows a Poisson distribution with mean $\mu_{j,k}$ (with equal exposures). This connects log-linear models to GLM. Given the role of log-likelihoods in GLM, let's discuss three scenarios that influence how the $Y_{j,k}$'s are jointly distributed.

SCENARIO 1

This scenario is the simplest: there are no restrictions to the counts. Therefore, the likelihood is

$$L(\boldsymbol{\beta}) = \prod_{k=1}^v \prod_{j=1}^w \frac{\exp(-\mu_{j,k}) \cdot \mu_{j,k}^{y_{j,k}}}{y_{j,k}!}$$

Using the log link,

$$\ln \mu_{j,k} = \mathbf{x}_{j,k}^T \boldsymbol{\beta}$$

where $\mathbf{x}_{j,k}$ is the vector of dummy variable inputs that correspond to level j of Factor A and level k of Factor B. It should be no surprise that $w + v - 2$ dummy variables are needed; a baseline category per factor is chosen arbitrarily.

This is described as a Poisson model.

SCENARIO 2

Now assume that the total count is known in advance. In other words, we know that $Y_{\bullet\bullet}$ equals a constant $n_{\bullet\bullet}$ due to the design of the study, e.g. a fixed number of 100 policyholders are sampled. This restriction means we must consider how the $Y_{j,k}$'s are jointly distributed, given $Y_{\bullet\bullet} = n_{\bullet\bullet}$. Without going through the proof, this leads to the $Y_{j,k}$'s following a multinomial distribution with $m = n_{\bullet\bullet}$ and $\pi_{j,k}$ denoting the probability of being in cell $\{j, k\}$. Therefore, the likelihood is

$$L(\boldsymbol{\beta}) = n_{\bullet\bullet}! \prod_{k=1}^v \prod_{j=1}^w \frac{\pi_{j,k}^{y_{j,k}}}{y_{j,k}!}$$

Since the mean count for cell $\{j, k\}$ is $n_{\bullet\bullet} \pi_{j,k}$, using the log link produces

$$\begin{aligned} \ln(n_{\bullet\bullet} \pi_{j,k}) &= \ln n_{\bullet\bullet} + \ln \pi_{j,k} \\ &= \ln n_{\bullet\bullet} + \mathbf{x}_{j,k}^T \boldsymbol{\beta} \end{aligned}$$

This is described as a multinomial model.

SCENARIO 3

Now assume that the subtotal counts for one of the factors is known in advance. To simplify, let the known subtotal counts belong to Factor A. This means we know that $Y_{j\bullet}$ equals a constant $n_{j\bullet}$, for every $j = 1, \dots, w$. For example, if Factor A represents gender, then a fixed number of 50 male and 50 female policyholders are sampled. This leads to a multinomial distribution for **each** Factor A level. Specifically, $Y_{j,1}, \dots, Y_{j,v}$ follows a multinomial distribution with parameters

- $m = n_{j\bullet}$
- $\pi_{j,1}$
- ...
- $\pi_{j,v}$

for every $j = 1, \dots, w$. Because there are w independent multinomial distributions, the likelihood is the product of the w joint PMFs, i.e.

$$L(\boldsymbol{\beta}) = \prod_{j=1}^w n_{j\bullet}! \prod_{k=1}^v \frac{\pi_{j,k}^{y_{j,k}}}{y_{j,k}!}$$

Since the mean count for cell $\{j, k\}$ is $n_{j\bullet} \pi_{j,k}$, using the log link produces

$$\begin{aligned}\ln(n_{j\bullet} \pi_{j,k}) &= \ln n_{j\bullet} + \ln \pi_{j,k} \\ &= \ln n_{j\bullet} + \mathbf{x}_{j,k}^T \boldsymbol{\beta}\end{aligned}$$

This is described as a *product multinomial* model.

Despite the differences between the three scenarios, estimating each model's $\boldsymbol{\beta}$ by MLE will result in the same predicted mean counts for the entire contingency table, all else being equal. To be precise, for every $\{j, k\}$ pair, the following expressions will be equal in value:

Scenario	Predicted Mean for Cell $\{j, k\}$
1	$\exp(\mathbf{x}_{j,k}^T \hat{\boldsymbol{\beta}})$
2	$n_{\bullet\bullet} \cdot \exp(\mathbf{x}_{j,k}^T \hat{\boldsymbol{\beta}})$
3	$n_{j\bullet} \cdot \exp(\mathbf{x}_{j,k}^T \hat{\boldsymbol{\beta}})$

More importantly, the likelihood ratio test that examines the interaction between Factor A and Factor B is also identical in each scenario. As a result, running a Poisson model is preferred regardless of the scenario due to its simplicity.

Testing the Interaction

Given the format of a $w \times v$ contingency table, log-linear models can look very similar to two-way ANOVA models.

In performing a likelihood ratio test, we have:

- Reduced model – predictors are the dummy variables without interaction terms, i.e. $p_r = w + v - 2$, analogous to the additive model
- Full model – predictors are the dummy variables with interaction terms, i.e. $p_f = wv - 1$, analogous to the model with interactions

Moreover, one count in each cell is analogous to having no replication. Therefore, the full model is also the saturated model; there are $p_f + 1 = wv$ parameters to estimate from wv data points. By viewing the model without the interaction terms as the fitted model, then the test statistic equals the fitted model's deviance. As there are $(w - 1)(v - 1)$ interaction terms, it is the degrees of freedom for the test. Similar to the chi-square test of independence, rejecting H_0 suggests that the two factors are dependent.

Example 3.10.2.1

A log-linear model is used to examine one factor with a levels and another factor with b levels.

Determine which statement is true.

- I. The main goal is to obtain estimates of the mean counts for each of the ab cells.
- II. It can be modeled as a generalized linear model with a Poisson response and identity link.
- III. The deviance of the model with interaction terms is 0.

Solution

I is false because the main goal is to study whether the two factors are independent or not.

II is false because a log-linear model can be modeled as a GLM with a Poisson response and log link.

III is true because the model with interaction terms is the saturated model, and its deviance is $2(l_{\text{sat}} - l_{\text{fit}}) = 0$.

Therefore, **only III is true.**



3.10 Summary

⌚ 5m

Poisson Regression

- $\mu_i = a_i \lambda_i$, where a_i is the exposure amount for the i^{th} observation.
- Uses the log link function.

$$\mu_i = a_i \cdot \exp(\mathbf{x}_i^T \boldsymbol{\beta})$$

- Log-likelihood function:

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i \ln \mu_i - \mu_i - \ln(y_i!)]$$

- Score function:

$$u_j = \sum_{i=1}^n (y_i - \mu_i) x_{i,j}$$

- Information matrix:

$$\mathbf{I} = \sum_{i=1}^n \mu_i \mathbf{x}_i \mathbf{x}_i^T$$

- Deviance:

$$\begin{aligned} D &= 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right] \\ &= 2 \sum_{i=1}^n y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) \end{aligned}$$

- Pearson residual:

$$e_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\hat{\mu}_i}}$$

- Pearson chi-square statistic:

$$\sum_{i=1}^n \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$$

Log-Linear Models

- The main objective is to assess whether there is an association or dependence between two factors. This is examined by checking whether they have a significant interaction.
- The response is the count in each cell of the contingency table created by the two factors; the predictors consist of dummy variables that represent the two factors.
- Key results of the multinomial model and the product multinomial model are shared with the Poisson model.
- In testing the interaction with a likelihood ratio test, the reduced model does not have the interaction terms as predictors, while the full model has the interaction terms. This is analogous to testing an additive model against a model with interactions for a two-way ANOVA setup without replication.

Appendix

🕒 5m

Deviance for Poisson Response

Recall that the Poisson log-likelihood is

$$l(\boldsymbol{\beta}) = \sum_{i=1}^n [y_i \ln \mu_i - \mu_i - \ln(y_i!)]$$

Under the fitted model:

$$l(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^n [y_i \ln \hat{\mu}_i - \hat{\mu}_i - \ln(y_i!)]$$

Under the saturated model:

$$\hat{\mu}_i = y_i$$

$$l_{\text{sat}} = \sum_{i=1}^n [y_i \ln y_i - y_i - \ln(y_i!)]$$

Therefore, the deviance is

$$\begin{aligned} D &= 2 \left[l_{\text{sat}} - l(\hat{\beta}) \right] \\ &= 2 \left\{ \sum_{i=1}^n [y_i \ln y_i - y_i - \ln(y_i!)] - \sum_{i=1}^n [y_i \ln \hat{\mu}_i - \hat{\mu}_i - \ln(y_i!)] \right\} \\ &= 2 \sum_{i=1}^n [y_i \ln y_i - y_i - y_i \ln \hat{\mu}_i + \hat{\mu}_i] \\ &= 2 \sum_{i=1}^n \left[y_i \ln \left(\frac{y_i}{\hat{\mu}_i} \right) - (y_i - \hat{\mu}_i) \right] \end{aligned}$$

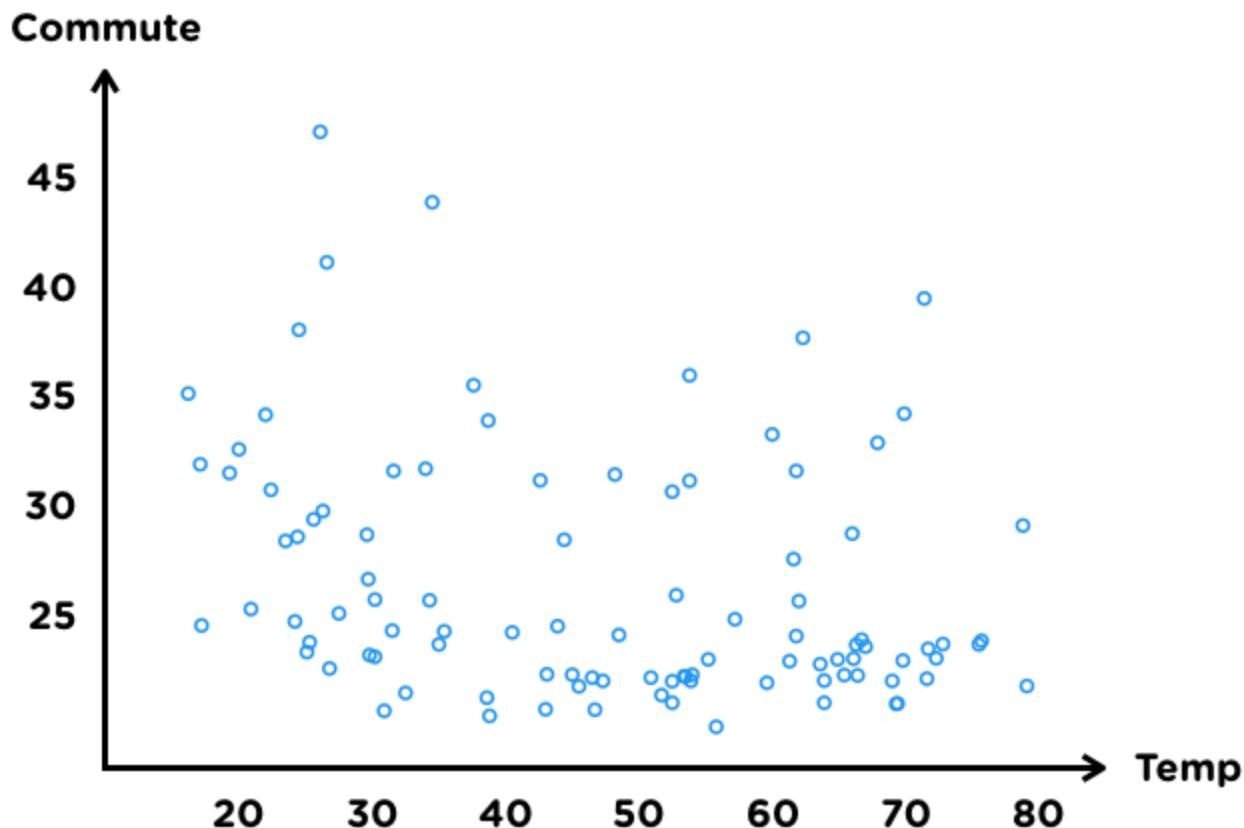
3.11.0 Overview

🕒 5m

To further investigate the capabilities of linear regression, we start by revisiting multiple linear regression. In particular, we begin with only one explanatory variable and consider predictors that are transformations of that variable. We examine a select scope of transformations under regression splines.

We also consider models that are relatively similar to regression splines, namely smoothing splines and local regression. The former has roots in shrinkage methods, while the latter resembles weighted least squares.

To motivate these models, the Commuting Chris scenario returns with Commute as the response and Temp as the single explanatory variable. Here is their scatterplot:



We end by expanding to more than one explanatory variable under generalized additive models (GAM).

3.11.1 Basis Functions

Recall the model equation for multiple linear regression:

$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p + \varepsilon$$

In limiting ourselves to **one** covariate x , consider the case where the predictor x_j is some function of x denoted by $b_j(x)$, for $j = 1, \dots, p$. This means we wish to transform the variable x using p different functions and use the resulting variables as predictors. The functions $b_1(\cdot), \dots, b_p(\cdot)$ are called **basis functions**. As a consequence, the model equation is

$$Y = \beta_0 + \beta_1 b_1(x) + \dots + \beta_p b_p(x) + \varepsilon$$

Effectively, we are resuming the discussion in Section 3.3.4. Let's consider three types of regressions and their particular choice of basis functions:

- Polynomial regression
- Step function approach
- Piecewise polynomial regression

Polynomial Regression

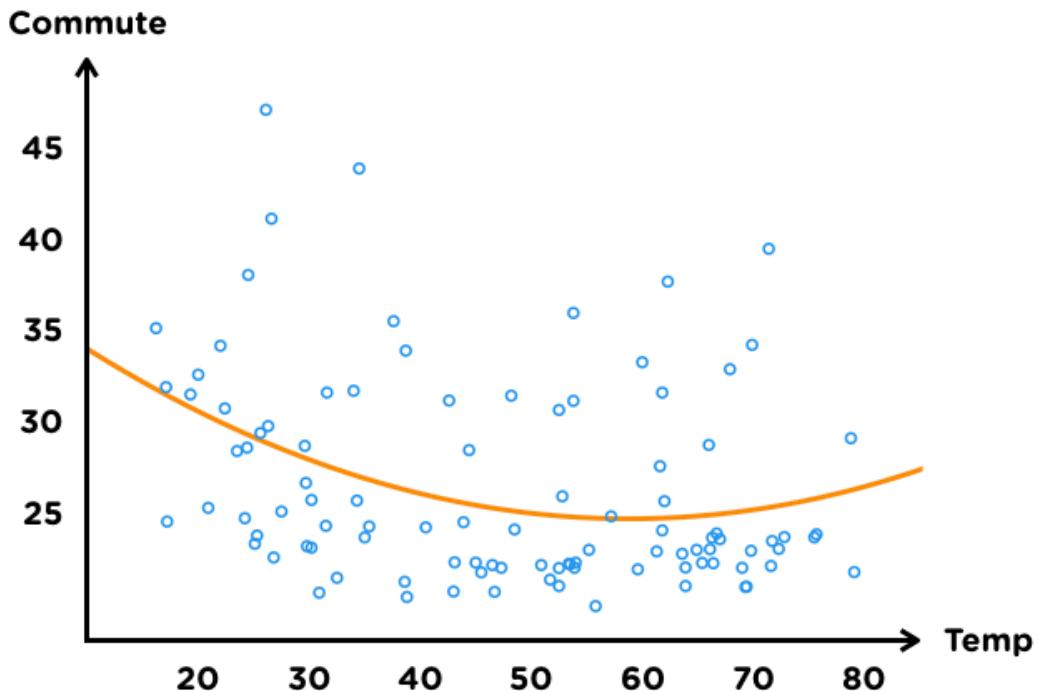
This regression type was already covered in Section 3.3.4. To model the response using x with a d^{th} order/degree polynomial, the model equation is

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_d x^d + \varepsilon$$

Here, the basis functions are

$$b_j(x) = x^j, \quad j = 1, \dots, d$$

By predicting Commute using Temp with a 2nd degree polynomial, we obtain the following quadratic fit:

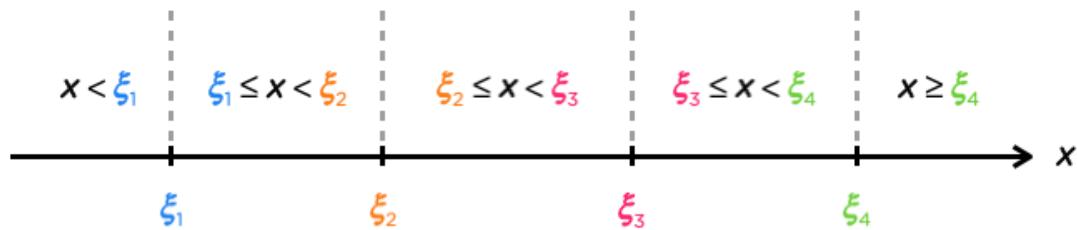


The fitted equation is

$$\hat{y} = 38.1898 - 0.4599x + 0.0039x^2$$

Step Function Approach

Start by dividing the range of x into $k + 1$ intervals or **bins**. Hence, there are k cutpoints or **knots** denoted as $\xi_1, \xi_2, \dots, \xi_k$ throughout the range of x . To illustrate, the following diagram shows how four knots create five intervals:



The basis functions here are dummy variables that indicate the interval where an x value is found. Specifically,

$$b_j(x) = \begin{cases} I(\xi_j \leq x < \xi_{j+1}), & j = 1, \dots, k-1 \\ I(x \geq \xi_k), & j = k \end{cases}$$

where $I(\cdot)$ is the indicator function. In effect, we have represented covariate x as a factor with $k + 1$ levels. These dummy variables are called **step functions**. Notice that the interval $x < \xi_1$ does not have a dummy variable, as it is the baseline category by default. As a result,

- β_0 is the mean response for $x \in (-\infty, \xi_1)$.
- $\beta_0 + \beta_j$ is the mean response for $x \in [\xi_j, \xi_{j+1})$ where $j = 1, \dots, k - 1$.
- $\beta_0 + \beta_k$ is the mean response for $x \in [\xi_k, \infty)$.

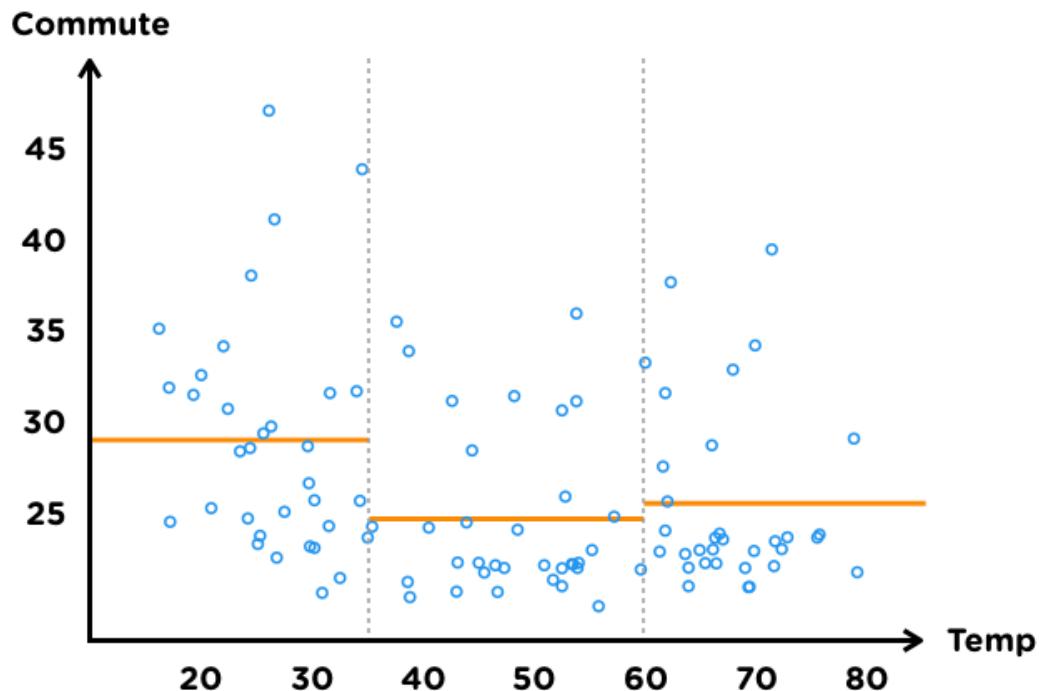
While we may arbitrarily decide the knot values, it is common to use sample percentiles of x spread uniformly, depending on the number of knots desired. To illustrate, let's divide the range of Temp into three intervals, thus requiring two knots. Then, it makes sense for

- $\xi_1 =$ the 33rd sample percentile of Temp, and
- $\xi_2 =$ the 67th sample percentile of Temp.

Using the smoothed empirical percentile approach, we get

$$\xi_1 = 35.132, \quad \xi_2 = 59.868$$

Predicting Commute with this setup produces the following fitted line:



The fitted equation is

$$\begin{aligned}\hat{y} &= 29.0121 - 4.3366 \cdot I(35.132 \leq x < 59.868) - 3.4853 \cdot I(x \geq 59.868) \\ &= \begin{cases} 29.0121, & x < 35.132 \\ 24.6755, & 35.132 \leq x < 59.868 \\ 25.5268, & x \geq 59.868 \end{cases}\end{aligned}$$

Piecewise Polynomial Regression

In Section 3.3.4, we studied the effect of a covariate interacting with a dummy variable – four linear functions of Precip Chance were fitted due to interacting with the three dummy variables of Season. We now tweak this by looking at polynomial functions of degree d , and set the dummy variables to be k step functions. This leads to the basis functions of

$$b_1(x) = x, \quad b_2(x) = x^2, \quad \dots, \quad b_d(x) = x^d,$$

$$b_{d+1}(x) = I(\xi_1 \leq x < \xi_2), \quad b_{d+2}(x) = I(\xi_2 \leq x < \xi_3), \quad \dots, \quad b_{d+k}(x) = I(x \geq \xi_k)$$

$$b_{d+k+1}(x) = x \cdot I(\xi_1 \leq x < \xi_2), \quad b_{d+k+2}(x) = x \cdot I(\xi_2 \leq x < \xi_3), \quad \dots, \quad b_{d+2k}(x) = x \cdot I(x \geq \xi_k)$$

$$\vdots$$

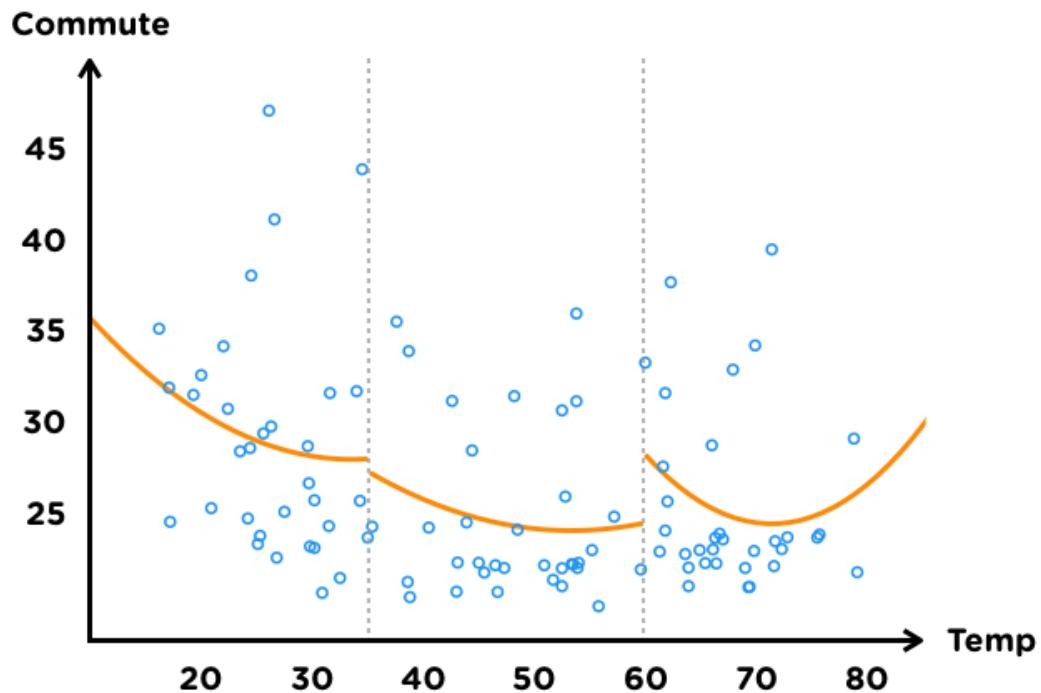
$$b_{d+dk+1}(x) = x^d \cdot I(\xi_1 \leq x < \xi_2), \quad b_{d+dk+2}(x) = x^d \cdot I(\xi_2 \leq x < \xi_3), \quad \dots, \quad b_{d+(d+1)k}(x) = x^d$$

While this may seem intimidating, we already have an intuitive idea of what is happening. Because interacting with dummy variables allows the baseline coefficients to be "updated" for the associated category, this setup creates a d^{th} degree polynomial **per category**. In this instance, the categories are the different intervals of x . Therefore, each of the $k + 1$ intervals has its own d^{th} degree polynomial. The result is a **piecewise polynomial regression**; it is more intuitive to write the model equation as follows:

$$Y = \begin{cases} \beta_{0,1} + \beta_{1,1}x + \dots + \beta_{d,1}x^d + \varepsilon, & x < \xi_1 \\ \beta_{0,2} + \beta_{1,2}x + \dots + \beta_{d,2}x^d + \varepsilon, & \xi_1 \leq x < \xi_2 \\ \vdots & \vdots \\ \beta_{0,k+1} + \beta_{1,k+1}x + \dots + \beta_{d,k+1}x^d + \varepsilon, & x \geq \xi_k \end{cases}$$

When $d = 0$, the regression simplifies to the step function approach; it is also called a piecewise **constant** regression for this reason.

Keeping the two knots $\xi_1 = 35.132$ and $\xi_2 = 59.868$, consider the piecewise quadratic regression (i.e. $d = 2$) of Commute on Temp.



The fitted equation is

$$\begin{aligned}
 \hat{y} &= 43.6522 - 0.9359x + 0.0139x^2 \\
 &\quad + 7.8452 \cdot I(35.132 \leq x < 59.868) + 129.5699 \cdot I(x \geq 59.868) \\
 &\quad - 0.0935x \cdot I(35.132 \leq x < 59.868) - 3.2328x \cdot I(x \geq 59.868) \\
 &\quad - 0.0043x^2 \cdot I(35.132 \leq x < 59.868) + 0.0153x^2 \cdot I(x \geq 59.868) \\
 &= \begin{cases} 43.6522 - 0.9359x + 0.0139x^2, & x < 35.132 \\ 51.4974 - 1.0294x + 0.0096x^2, & 35.132 \leq x < 59.868 \\ 173.2221 - 4.1687x + 0.0292x^2, & x \geq 59.868 \end{cases}
 \end{aligned}$$

Coach's Remarks

If you are still confused with how the model equation in piecewise form connects to the basis functions, try thinking it through using these few equalities:

$$\beta_{0,1} = \beta_0, \quad \beta_{1,1} = \beta_1, \quad \beta_{d,1} = \beta_d$$

$$\beta_{0,k+1} = \beta_0 + \beta_{d+k}, \quad \beta_{1,k+1} = \beta_1 + \beta_{d+2k}, \quad \beta_{d,k+1} = \beta_d + \beta_{d+(d+1)k}$$

The discontinuous fitted curve is unappealing; it would be better for it to be a continuous function and relatively smooth. We resume this discussion in the next subsection.

Degrees of Freedom

In the context of Section 3.11, the degrees of freedom refers to the number of regression coefficients, $p + 1$. Here, the degrees of freedom is used as an indicator of flexibility.

Example 3.11.1.1

You consider regression models with predictors based on a single variable.

Determine which of the following models has the most degrees of freedom.

- A. Simple linear regression
- B. Polynomial regression with 4 degrees of freedom
- C. Piecewise constant regression with 5 knots
- D. Piecewise linear regression with 4 knots
- E. Piecewise cubic regression with 2 knots

Solution

Simple linear regression has 1 predictor: x . Therefore, it has 2 degrees of freedom.

Piecewise constant regression with 5 knots has 5 predictors: the step functions $I(\xi_1 \leq x < \xi_2), \dots, I(x > \xi_5)$. Therefore, it has 6 degrees of freedom.

Piecewise linear regression with 4 knots has

- 1 from x , plus
- 4 step functions, plus
- $1 \cdot 4 = 4$ interaction terms,

for a total of 9 predictors. Alternatively, 4 knots means there are five intervals, and hence there are five linear functions of x . Since every linear function has two regression coefficients, we get $2 \cdot 5 = 10$ degrees of freedom.

Piecewise cubic regression with 2 knots has

- 3 from x, x^2 , and x^3 , plus
- 2 step functions, plus
- $3 \cdot 2 = 6$ interaction terms,

for a total of 11 predictors. Therefore, it has 12 degrees of freedom.



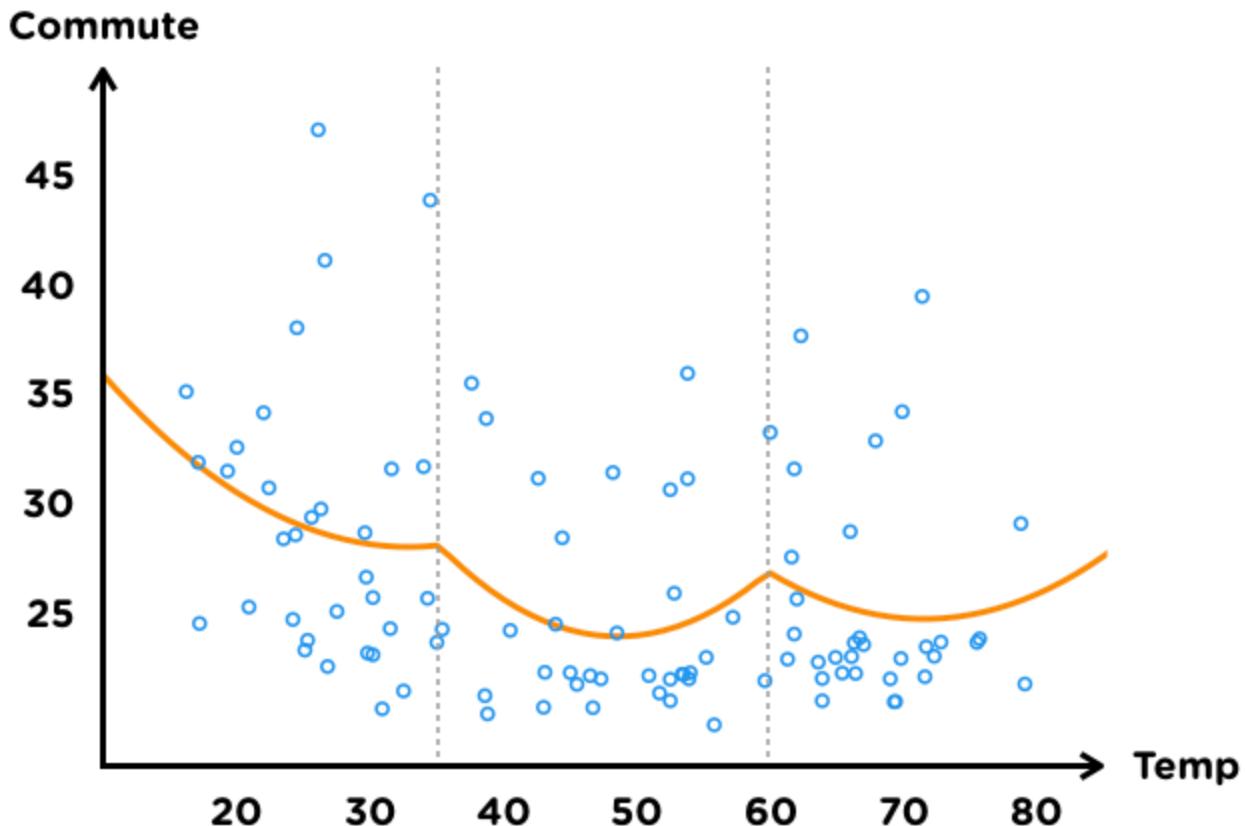
Therefore, the answer is (E).

Coach's Remarks

A polynomial regression with 4 degrees of freedom refers to a cubic regression; it has 3 predictors, so they must be the polynomial terms x , x^2 , and x^3 .

3.11.2 Regression Splines

The piecewise quadratic fit of Commute on Temp in the previous subsection was a discontinuous function of Temp. If we force the three quadratics to be continuous at the knots, we obtain the following fit:



This is called a **continuous** piecewise quadratic regression. To achieve this, we impose two constraints – one per knot – so that the quadratics are unbroken at the knots. Imposing two constraints implies that we drop from $2 + 2 + 2 \cdot 2 = 8$ predictors to $8 - 2 = 6$ predictors.

Even so, the fitted curve still looks awkward near the knots. To smooth the fitted curve, we can impose two more constraints: the first derivative of the piecewise quadratic must also be continuous at each knot. If we did this, then the regression would only consist of $6 - 2 = 4$ predictors, since another two constraints would be imposed.

Coach's Remarks

If it is unclear how constraints are reducing the number of predictors, we can illustrate with this model equation:

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

Now let's impose the constraint that $\beta_1 = \beta_2$. Then, the model equation simplifies to

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_1 x_2 + \varepsilon \\ &= \beta_0 + \beta_1 (x_1 + x_2) + \varepsilon \end{aligned}$$

The result is a simple linear regression whose sole predictor is $x_1 + x_2$; the model features drop from 2 to 1 due to the constraint.

In general, a constraint is an equation that allows one β to be completely expressed in terms of the other β 's.

Splines

A **regression spline** is a continuous piecewise polynomial regression that is smooth at the knots. In using a d^{th} degree polynomial, smoothness is attained by requiring the first $d - 1$ derivatives of the curve to all be continuous at the k knots. This means, for example, a cubic spline has

- continuity at the knots,
- first derivative continuity at the knots, and
- second derivative continuity at the knots.

Given that a cubic spline has three constraints per knot, its predictor count is

$$\begin{aligned} p &= 3 + k + \overbrace{3k}^{\text{interactions}} - \overbrace{3k}^{\text{constraints}} \\ &= 3 + k \end{aligned} \tag{3.11.2.1}$$

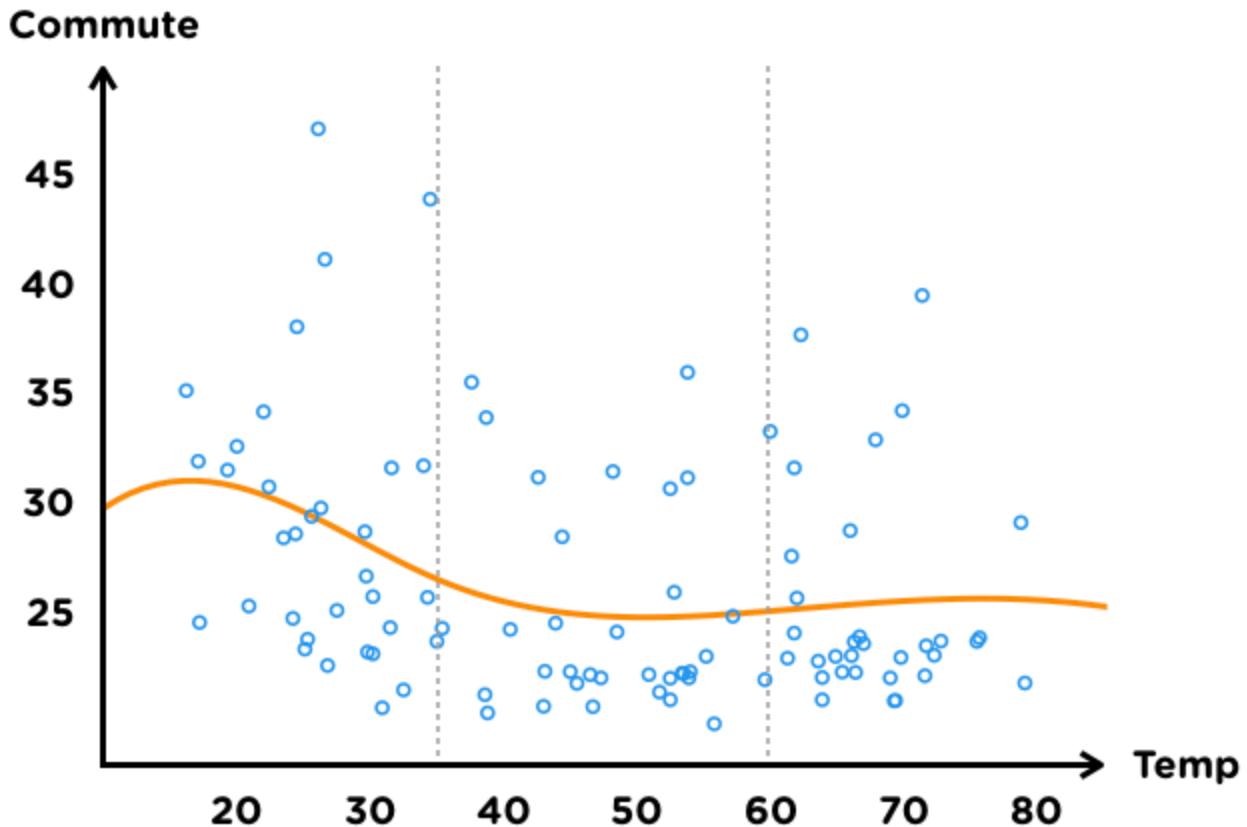
Since a cubic is often sufficiently flexible in a given interval, increasing the flexibility of a spline usually comes from increasing the number of knots rather than the polynomial order. In other words, the number of knots is a flexibility measure whose ideal value can be determined through cross-validation.

There are many ways to express the $3 + k$ basis functions for a cubic spline. For this exam, we use the following set of functions: x, x^2, x^3 , and the remaining k are

$$b_4(x) = (x - \xi_1)_+^3, \quad b_5(x) = (x - \xi_2)_+^3, \quad \dots, \quad b_{3+k}(x) = (x - \xi_k)_+^3$$

known as **truncated power** basis functions. The $+$ symbol has the same meaning as in Section 1.2.2; if $x > \xi$, then the basis function equals $(x - \xi)^3$, but equals 0 otherwise.

Keeping the two knots $\xi_1 = 35.132$ and $\xi_2 = 59.868$, the cubic spline of Commute on Temp has the following fitted curve:



Natural Splines

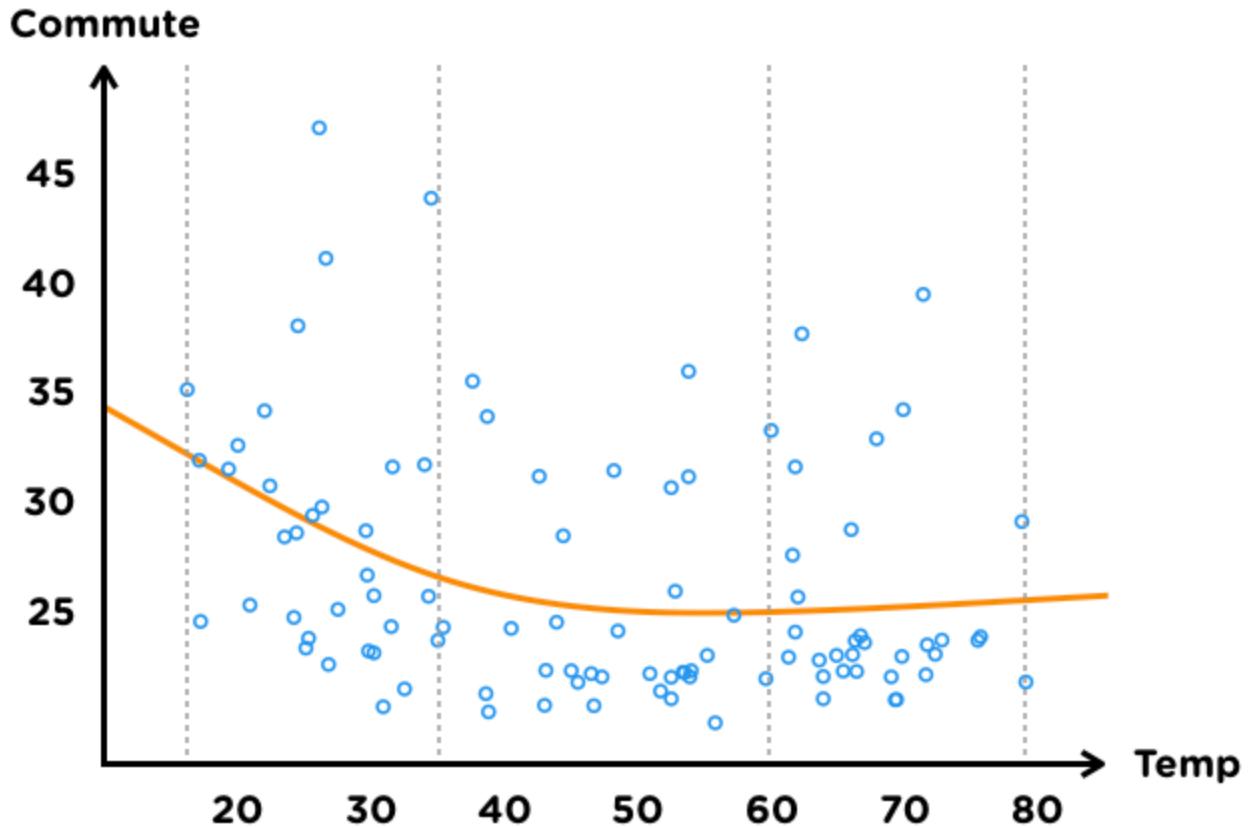
One disadvantage of fitting polynomials is that $se(\hat{y})$ is typically large near the lower and upper boundaries of the covariate. A method to minimize this for a cubic spline (or splines in general) is to assume that the curve becomes linear in the intervals $x < \xi_1$ and $x > \xi_k$. It is then implied that ξ_1 and ξ_k are now **boundary knots** that are respectively near the lower and upper boundaries of x ; the remaining knots are considered as **interior knots**.

This results in a **natural cubic spline**. Relative to a cubic spline, we make the first and last of the $k + 1$ cubics become lines. This further imposes four more constraints; we "drop the coefficients" of x^2 and x^3 in both boundary intervals. Therefore, a natural cubic spline operates with

$3 + k - 4 = k - 1$ predictors. Although there are fewer basis functions now, they become quite complex and are not shown here.

For a natural cubic spline of Commute on Temp, let the boundary knots be the smallest and largest values of Temp, and retain the 33rd and 67th sample percentiles as interior knots, i.e.

$$\xi_1 = 16.200, \quad \xi_2 = 35.132, \quad \xi_3 = 59.868, \quad \xi_4 = 79.100$$



With a d^{th} degree polynomial and k knots, the table below gives the number of predictors for the models we have discussed:

Model	Number of Predictors, p
Polynomial	d
Piecewise constant	k
Piecewise polynomial	$d + k + dk$
Continuous piecewise polynomial	$d + dk$
Cubic spline	$3 + k$
Natural cubic spline	$k - 1$

Here is a summary of the models we have considered for the Commuting Chris scenario so far:

Model	Number of Knots, k	Number of Predictors, p
Quadratic	—	2
Piecewise constant	2	2
Piecewise quadratic	2	8
Continuous piecewise quadratic	2	6
Cubic spline	2	5
Natural cubic spline	4	3

Example 3.11.2.1

A linear model has the following model equation:

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4 (x - 12)_+^3 + \varepsilon$$

Determine which of the following statements is true.

- I. The model is a natural cubic spline with one knot.
- II. The mean response is discontinuous at $x = 12$.
- III. The second derivative of the mean response is continuous at $x = 12$.

Solution

Based on the model's basis functions, the model is a cubic spline (not a **natural** cubic spline) with one knot at $x = 12$. Cubic splines have continuity at the knots, as well as second derivative continuity at the knots.

Therefore, **only III is true.**

Coach's Remarks

While you should know the continuity properties of a cubic spline by memory, it is possible to test statements II and III by hand.

Let $f(x)$ be the mean response, i.e.

$$f(x) = \begin{cases} \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3, & x \leq 12 \\ \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + \beta_4(x - 12)^3, & x > 12 \end{cases}$$

Then, show that

$$\begin{aligned} \beta_0 + \beta_1 \cdot 12 + \beta_2 \cdot 12^2 + \beta_3 \cdot 12^3 &= \\ \beta_0 + \beta_1 \cdot 12 + \beta_2 \cdot 12^2 + \beta_3 \cdot 12^3 + \beta_4(12 - 12)^3 & \end{aligned}$$

i.e. the mean response is continuous at $x = 12$. Do the same for $f''(x)$ to demonstrate second derivative continuity.

Example 3.11.2.2

An analyst runs a cubic spline on a dataset. The results indicate that the fitted spline is too smooth, failing to follow the patterns in the data points well.

Determine which of the following is an adjustment that the analyst should not make.

- A. Increase the number of knots
- B. Move the knots closer to the sample median of the covariate
- C. Consider a different covariate
- D. Run a natural cubic spline instead with all else equal
- E. Run a quintic (degree-5 polynomial) spline instead with all else equal

Solution

The analyst may make adjustment (A). Increasing the number of knots increases flexibility, which makes the fitted curve less smooth.

The analyst may make adjustment (B). Having more knots in any one region raises the flexibility in that region. So, positioning the knots near where the data points have special patterns and/or vary rapidly should improve the spline's quality, even if not guaranteed to increase model flexibility as a whole. The sample median is typically where most of the data points congregate, so intuitively speaking, it is a reasonable region to target.

The analyst may make adjustment (C). Perhaps the chosen covariate does not predict the response well, regardless of any manipulation. If so, it makes sense to forgo using it in the model.

The analyst should not make adjustment (D). With all else equal, the natural cubic spline is a constrained, less complex version of the cubic spline. This produces a less flexible, even smoother fit.

The analyst may make adjustment (E). Increasing the polynomial degree increases flexibility.

Therefore, the answer is **(D)**.



3.11.3 Smoothing Splines

For the types of regressions we have considered thus far, the SSE is

$$\sum_{i=1}^n (y_i - \hat{y}_i)^2 = \sum_{i=1}^n [y_i - \hat{\beta}_0 - \hat{\beta}_1 b_1(x_i) - \dots - \hat{\beta}_p b_p(x_i)]^2$$

As multiple linear regressions, realize the $\hat{\beta}_0, \dots, \hat{\beta}_p$ above are the OLS estimates.

Let's take a step back and imagine a more generic form for \hat{y} . It is ultimately a function of x , so let's denote it as $g(x)$. The function $g(x)$ contains the estimates $\hat{\beta}_0, \dots, \hat{\beta}_p$, but their relation to x is left unspecified. In other words, $\hat{\beta}_0, \dots, \hat{\beta}_p$ are not assumed to be OLS estimates.

Desiring to minimize the SSE

$$\sum_{i=1}^n [y_i - g(x_i)]^2$$

is problematic at this stage. So far, nothing about our setup prevents $g(x_i)$ from equaling y_i for all i , hence minimizing the expression is achieved by overfitting. Our goal is to establish a criterion that can prevent overfitting while having a smooth $g(x)$ that predicts the response well.

As overfitting means there is too much flexibility, we can incorporate a criterion that mimics a shrinkage method like ridge and lasso. Recall that these methods reduce flexibility by introducing a shrinkage penalty.

A **smoothing spline** seeks to minimize the expression

$$\underbrace{\sum_{i=1}^n [y_i - g(x_i)]^2}_{\text{SSE}} + \lambda \underbrace{\int_{-\infty}^{\infty} g''(t)^2 dt}_{\text{penalty}} \quad (3.11.3.1)$$

where λ is called a **smoothing parameter** or tuning parameter.

The shrinkage penalty involves $g''(\cdot)$ because the second derivative is an indicator of function roughness or wigginess. The more wiggly $g(\cdot)$ is, the larger the integral becomes. As a result,

minimizing Equation 3.11.3.1 promotes a smooth $g(x)$. The level of smoothness is controlled by λ ; as it increases, $g(x)$ must be smoother so that Equation 3.11.3.1 is minimized. The fitted curve $g(x)$ for a smoothing spline has the same form as the fitted natural cubic spline with knots at the n observed values of the covariate. To be precise, $g(x)$ is

- linear in the intervals $x < \min(x_1, \dots, x_n)$ and $x > \max(x_1, \dots, x_n)$,
- piecewise cubic between the ordered x_1, \dots, x_n , and
- has continuity, first derivative continuity, and second derivative continuity at x_1, \dots, x_n ,

when x_1, \dots, x_n are all unique.

However, a smoothing spline is distinct from a natural cubic spline with n knots at the x values because the latter has coefficients estimated by OLS. Moreover, such a natural cubic spline perfectly predicts the response because it estimates $p + 1 = (n - 1) + 1 = n$ regression coefficients. To understand how a smoothing spline is different, keep in mind that λ dictates the resulting $g(x)$ (and hence, the values of $\hat{\beta}_0, \dots, \hat{\beta}_p$ as well) by controlling its smoothness:

- When $\lambda = 0$, there is no shrinkage penalty, so SSE is minimized to 0 due to a perfect fit. This smoothing spline is the same as the OLS natural cubic spline (i.e. overfitted and very flexible).
- As λ approaches ∞ , the entire expression is minimized when the integral is 0. This means $g(x)$ must be linear in x ; this smoothing spline is the same as simple linear regression (i.e. smooth and inflexible).

A value of λ between the two extremes results in a shrunken version of the OLS natural cubic spline. Similar to ridge and lasso, λ is inversely related to flexibility. The appropriate amount of flexibility can be determined through cross-validation.

Effective Degrees of Freedom

In multiple linear regression, recall that \mathbf{H} is the hat matrix, such that

- $\hat{\mathbf{y}} = \mathbf{H}\mathbf{y}$
- The leverage h_i is the i^{th} diagonal entry of \mathbf{H}
- $\sum_{i=1}^n h_i = p + 1$

As already mentioned, $p + 1$ is the degrees of freedom in the sense that it indicates flexibility. However, like ridge regression, the flexibility of a smoothing spline cannot be measured by the number of regression coefficients.

The effective degrees of freedom measures flexibility by summing values similar to leverages. For clarity, we now use \hat{y} instead of $g(x)$ to represent the fitted response for a smoothing spline. Specifically, let $\hat{y}_{\lambda i}$ be the fitted response for the i^{th} observation given a tuning parameter value of λ . Then, there exists a matrix \mathbf{S}_λ such that

$$\hat{\mathbf{y}}_\lambda = \mathbf{S}_\lambda \mathbf{y} \quad (3.11.3.2)$$

Denote the i^{th} diagonal entry of \mathbf{S}_λ as $h_{\lambda i}$. As a result, the **effective degrees of freedom** is

$$\text{df}_\lambda = \sum_{i=1}^n h_{\lambda i} \quad (3.11.3.3)$$

As λ increases from 0 to ∞ , the effective degrees of freedom decreases from n to 2, alluding to

- the n degrees of freedom of the overfitted natural cubic spline, and
- the 2 degrees of freedom of simple linear regression.

Furthermore, recall that performing LOOCV does not require running n fits to calculate the LOOCV error when using OLS estimation, i.e. Equation 3.6.3.1 only needs the single fit with all the observations. This is exactly mirrored for smoothing splines when LOOCV is used to tune λ :

$$\text{LOOCV error} = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - \hat{y}_{\lambda i}}{1 - h_{\lambda i}} \right)^2 \quad (3.11.3.4)$$

Example 3.11.3.1

Determine which of the following statements are true regarding smoothing splines.

- I. Decreasing flexibility is not achieved by reducing the number of knots.
- II. The coefficients are estimated by ordinary least squares.
- III. The smoothing parameter is inversely related to the effective degrees of freedom.

Solution

I is true because flexibility decreases by increasing the smoothing parameter λ . The number of knots is always n for a smoothing spline.

II is false because the coefficients are estimated by shrinkage through minimizing the expression of Equation 3.11.3.1.

III is true because as λ increases from 0 to ∞ , the effective degrees of freedom decreases from n to 2; the former is inversely related to flexibility, while the latter is a direct measure of flexibility.

Therefore, **only I and III are true.**



3.11.4 Local Regression

In Section 3.8.3, we briefly presented weighted least squares. Recall that its coefficient estimates have the formula

$$\hat{\beta} = (\mathbf{X}^T \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^T \mathbf{W} \mathbf{y}$$

where \mathbf{W} is an $n \times n$ diagonal matrix with the weights w_1, \dots, w_n as its entries. Equivalently, the estimates minimize the expression

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

The point here is that the weight w_i indicates how important the i^{th} observation is in determining the fit. For example, if w_i is small, then the contribution of $(y_i - \hat{y}_i)^2$ to the whole expression is reduced, causing observation i to be deemphasized when estimating the coefficients.

Local regression implements this weighted least squares idea in an interesting way. This method is algorithmic in calculating the fitted values, and thus, we cannot express the fitted curve with an equation. To predict the response for the covariate having a value of x_* , the procedure is:

1. Select the **span** – a number between 0 and 1 that signifies a proportion of the observations. Let v denote the integer such that the span is $\frac{v}{n}$.
2. Identify the v observations whose x values are the closest to x_* .
3. Assign weights w_1, \dots, w_n based on each observation's closeness to x_* :
 - (a) For observations not among the v closest, their weights are all 0.
 - (b) For observations among the v closest, the weight is largest for the closest observation, and decreases until reaching 0 for the furthest observation. The weights are determined by a **weighting function**.
4. Solve for the estimates $\hat{\beta}_{0*}$ and $\hat{\beta}_{1*}$ by minimizing the expression

$$\sum_{i=1}^n w_i (y_i - \hat{\beta}_{0*} - \hat{\beta}_{1*} x_i)^2 \quad (3.11.4.1)$$

5. Calculate the fitted value as $\hat{\beta}_{0*} + \hat{\beta}_{1*}x_*$.

Therefore, local regression computes \hat{y} for an input x_* by fitting a line with only the observations that are close to the input, and also favors those that are closer. Realize that the coefficient estimates depend on the weights, which in turn depend on the choice of x_* , so a different input produces a different fitted line. However, the fitted lines are not the end goal; it is only the fitted values that we desire.

Coach's Remarks

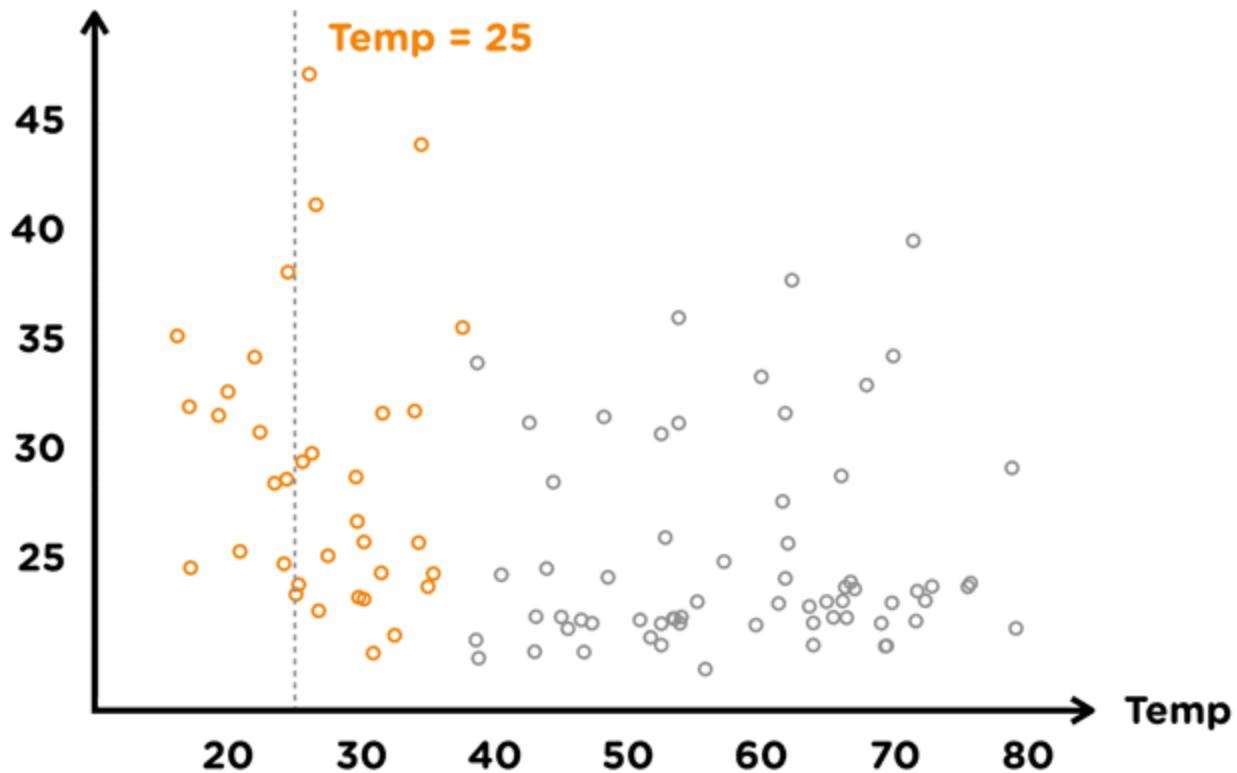
Although not found in the official reading, an example of an appropriate weighting function for the v closest observations is

$$\left(1 - \left| \frac{x_* - x}{x_* - x_{(v)}} \right|^3\right)^3$$

where $x_{(v)}$ is the x value of the furthest observation among the v closest to the input. This ensures a weight of 0 for that furthest observation, i.e. when $x = x_{(v)}$.

Let's run a local regression of Commute on Temp with a span of 0.35, i.e. $v = 35$ for this dataset. Consider the animation below that illustrates finding the fitted values at Temp values of 25 and 55:

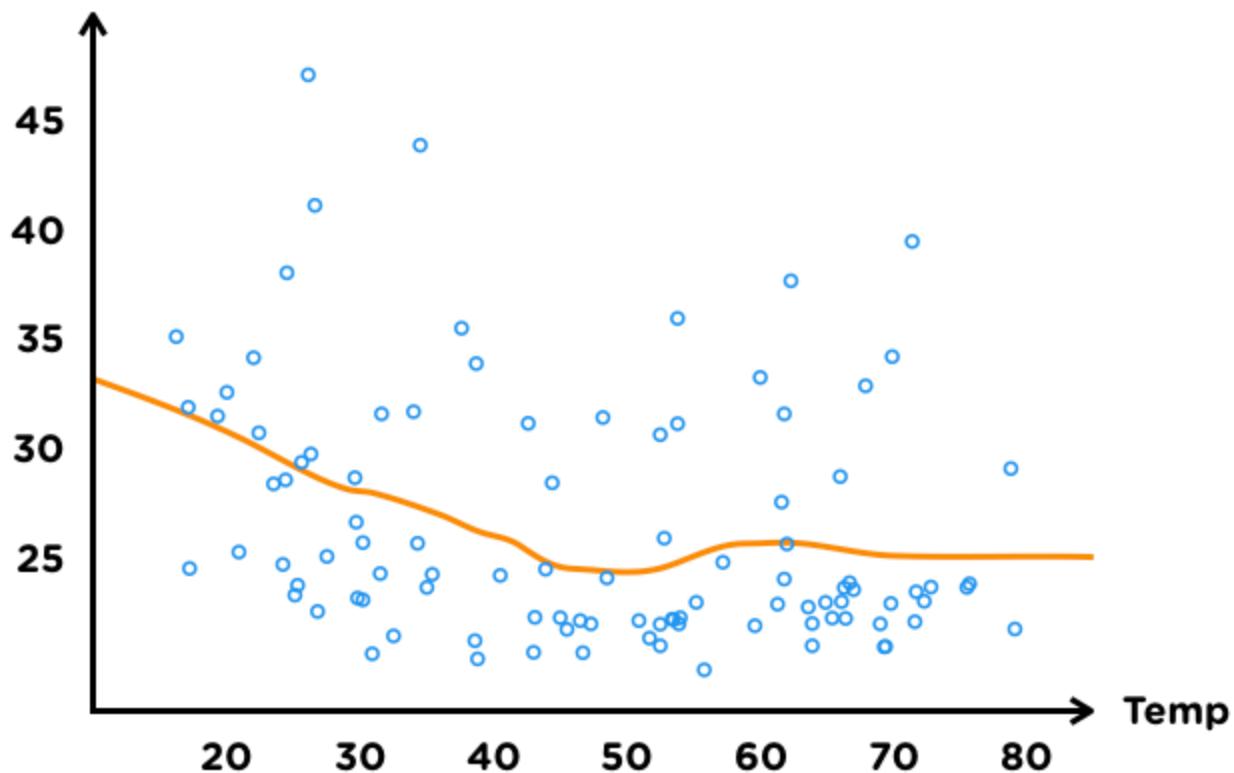
Commute



In each case, we only consider the 35 observations whose Temp values are closest to the input to obtain the fitted line. Yet, the fitted line itself is not the objective, being used only to determine the fitted value at the given input. The animation also shows the fitted values at several other inputs besides 25 and 55.

By repeating the procedure for many different Temp inputs, we obtain the following fitted curve by connecting the individual fitted values.

Commute



Here are other essential points about local regression:

- A small span means each fitted line depends on a few observations, resulting in fitted values that are highly sensitive to data patterns in narrow intervals. Conversely, a large span allows more observations to influence each fitted line, so the fitted values will be more stable. In other words, the span is inversely related to flexibility. Cross-validation can help determine the proper flexibility level.
- Instead of fitting lines to subsets of observations, we could choose to do constant or polynomial fits. This also impacts the regression's flexibility, but not as much as changing the span.

Higher Dimensions

Although we mainly focus on using one covariate, local regression can extend to multiple covariates. If we consider p covariates, then the procedure identifies the closest observations to the input in p -dimensional space. However, local regression is sensitive to the curse of dimensionality, meaning it does not perform well in high dimensions.

Example 3.11.4.1

Determine which of the following decisions will result in a rougher regression fit.

- I. Reducing the span for a local regression.
- II. Fitting a constant to the subsets of nearby observations in local regression.
- III. Running a cubic spline instead of a piecewise cubic regression with all else equal.
- IV. Increasing the number of knots for a natural cubic spline.
- V. Using cross-validation for a smoothing spline instead of setting $\lambda = 1$.

Solution

A rougher fit is synonymous with a more flexible, wiggly fit.

I will produce a rougher fit because reducing the span increases the flexibility of a local regression.

II will not produce a rougher fit because a constant is the lowest polynomial degree, making it harder for the fitted value to get close to the response data.

III will not produce a rougher fit because a cubic spline is less flexible than a piecewise cubic regression; the former has fewer predictors due to the added continuity constraints.

IV will produce a rougher fit because increasing the number of knots increases the number of predictors, which in turn increases flexibility.

V may not produce a rougher fit. Since cross-validation helps determine the optimal value of λ , the outcome could be lower or higher than 1; there is no indication whether $\lambda = 1$ is too flexible or inflexible.

Therefore, **only I and IV are true.**



3.11.5 Generalized Additive Models

We may extend all of the concepts discussed thus far for one covariate to g recorded variables. A **generalized additive model (GAM)** does this by assuming a model equation of

$$Y = \beta_0 + f_1(x_1) + \dots + f_g(x_g) + \varepsilon \quad (3.11.5.1)$$

For $m = 1, \dots, g$, the explanatory variable x_m contributes to the mean response through a function f_m , doing so independently of the other explanatory variables. f_m may have any functional form of regressions with one explanatory variable that we studied in previous subsections, and the functions are also free to vary across the g explanatory variables. For example, given two covariates (x_1, x_2) and one factor (x_3) , we can model

- f_1 as a cubic spline function,
- f_2 as a smoothing spline function, and
- f_3 as a function with dummy variables.

Incorporating these individual models in the larger generalized additive model gives a sense of **models within models**. From the perspective that the mean response $f(x_1, \dots, x_g)$ is a function of f_1, \dots, f_g , we can likewise say there is a **functions within functions** structure.

If f_1, \dots, f_g are all based on regressions that use OLS (i.e. not a smoothing spline nor local regression), then the GAM is a multiple linear regression. In other words, each $f_m(x_m)$ is a sum of basis functions with their coefficients (no intercept since β_0 handles it for the entire model), with all regression coefficients in Equation 3.11.5.1 estimated by OLS.

However, if at least one of f_1, \dots, f_g employs a smoothing spline or local regression, then fitting the GAM requires a different approach such as **backfitting**. In summary, backfitting runs separate fits for each x_m while fixing the other variables, and repeatedly uses a past fit to inform the next fit until the overall fit converges.

The additive nature of the model means the effect of each x_m on the response can be investigated individually, assuming the other variables are held constant. This includes conducting appropriate statistical inference. On the other hand, its additivity is also a restriction since no interactions are considered. One solution is to manually include interaction terms to the model, but it would no longer constitute a GAM.

Example 3.11.5.1

Determine which statements are true regarding a generalized additive model.

- I. It is in essence a multiple linear regression.
- II. It can model each predictor to have a non-linear relationship with the response.
- III. It does not account for possible interactions between predictors.

Solution

I is false because a GAM is only a multiple linear regression when every predictor is modeled by a regression that uses OLS.

II is true because each predictor may be modeled with any polynomial regression, spline, or local regression.

III is true because the model is additive, thus isolating the effects of each predictor on the response.

Therefore, **only II and III are true.**



3.11 Summary

With a d^{th} degree polynomial and k knots:

Model	Number of Predictors, p
Polynomial	d
Piecewise constant	k
Piecewise polynomial	$d + k + dk$
Continuous piecewise polynomial	$d + dk$
Cubic spline	$3 + k$
Natural cubic spline	$k - 1$

Step Functions

For knots ξ_1, \dots, ξ_k , the basis functions are:

$$b_j(x) = \begin{cases} I(\xi_j \leq x < \xi_{j+1}), & j = 1, \dots, k-1 \\ I(x \geq \xi_k), & j = k \end{cases}$$

Piecewise Polynomial Regression

The basis functions are:

- x, x^2, \dots, x^d
- k step functions
- dk interaction terms

A continuous piecewise polynomial regression imposes one constraint per knot.

Cubic Splines

A piecewise cubic with continuity, first derivative continuity, and second derivative continuity at the knots. The basis functions can be:

$$x, x^2, x^3, (x - \xi_1)_+^3, \dots, (x - \xi_k)_+^3$$

A natural cubic spline has two boundary knots, beyond which the curve is linear instead of cubic; this imposes four constraints.

Smoothing Splines

- The fitted curve $g(x)$ is determined by minimizing the expression

$$\sum_{i=1}^n [y_i - g(x_i)]^2 + \lambda \int_{-\infty}^{\infty} g''(t)^2 dt$$

where λ is inversely related to flexibility.

- $g(x)$ has the same form as the fitted natural cubic spline with knots at the n observed values of x .
- The effective degrees of freedom measures flexibility as the sum of the diagonal entries of \mathbf{S}_λ , where $\hat{\mathbf{y}}_\lambda = \mathbf{S}_\lambda \mathbf{y}$.

Local Regression

- Calculates the fitted value for a specific input x_* by mimicking weighted least squares.
- The weights are determined by the span and the weighting function, such that observations nearer to x_* are given larger weights.
- The span is inversely related to flexibility.
- Does not perform well in high dimensions.

Generalized Additive Models

- Each explanatory variable contributes to the mean response independently of the other explanatory variables; no interactions are considered.

- The effect of each explanatory variable on the response can be investigated individually, assuming the other variables are held constant.
- Backfitting can be used for fitting if ordinary least squares cannot.