Linear regression

Maria Jose Medina

Universidad de Santiago de Chile

Outline

Table of Contents

Multiple regression

Introduction

- The goal is to predict a quantitative response Y on the basis of p distinct predictors x_p.

Multiple regression

Introduction

- The goal is to predict a quantitative response Y on the basis of p distinct predictors x_p.
- It assumes that there is *approximately* a relationship between $x_1, x_2, \cdots, x_p$ and Y

Multiple regression

Introduction

- The goal is to predict a quantitative response Y on the basis of p distinct predictors x_p.
- It assumes that there is *approximately* a relationship between $x_1, x_2, \cdots, x_p$ and Y

$$Y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \cdots \beta_p x_p$$

- β_i are unknown constants called *model coefficients* or *parameters*.

Multiple regression

Introduction

- The goal is to predict a quantitative response Y on the basis of p distinct predictors x_p.
- It assumes that there is *approximately* a relationship between $x_1, x_2, \cdots, x_p$ and Y

$$Y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \cdots \beta_p x_p$$

- β_i are unknown constants called *model coefficients* or *parameters*.
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ be the prediction for Y based on the predictors of x_p.

Multiple regression

Introduction

- The goal is to predict a quantitative response Y on the basis of p distinct predictors x_p.
- It assumes that there is *approximately* a relationship between $x_1, x_2, \cdots, x_p$ and Y

$$Y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \cdots \beta_p x_p$$

- β_i are unknown constants called *model coefficients* or *parameters*.
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ be the prediction for Y based on the predictors of x_p.
  Then

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

Multiple regression

Introduction

- The goal is to predict a quantitative response Y on the basis of p distinct predictors x_p.
- It assumes that there is *approximately* a relationship between $x_1, x_2, \cdots, x_p$ and Y

$$Y \approx \beta_0 + \beta_1 x_1 + \beta_2 x_3 + \cdots \beta_p x_p$$

- β_i are unknown constants called *model coefficients* or *parameters*.
- Let $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_p x_p$ be the prediction for Y based on the predictors of x_p.
  Then

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

  represents the ith **residual**.

Multiple regression

Estimating the model coefficients

- Using *residuals*, we define the **residual sum of squares**(RSS) as

Multiple regression

Estimating the model coefficients

- Using *residuals*, we define the **residual sum of squares**(RSS) as

$$\begin{aligned} RSS &= e_1^2 + e_2^2 + \cdots e_n^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

Multiple regression

Estimating the model coefficients

- Using *residuals*, we define the **residual sum of squares**(RSS) as

$$\begin{aligned} RSS &= e_1^2 + e_2^2 + \cdots e_n^2 \\ &= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \hat{\beta}_p x_{ip})^2 \end{aligned}$$

- To estimate the coefficients we use the **least squares approach**, in which we seek to minimize RSS.

Multiple regression

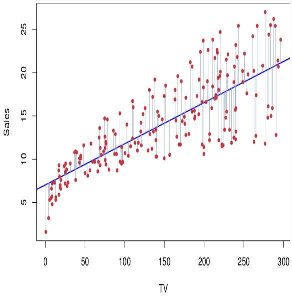
Estimating the model coefficients

- Using *residuals*, we define the **residual sum of squares**(RSS) as

$$
\begin{aligned}
RSS &= e_1^2 + e_2^2 + \cdots e_n^2 \\
&= \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \hat{\beta}_p x_{ip})^2
\end{aligned}
$$

- To estimate the coefficients we use the **least squares approach**, in which we seek to minimize RSS.

$$\hat{\beta} = (X'X)^{-1}X'y$$



Multiple regression

Some important questions

Now we have to evaluate the model:

1. Is at least one of the predictors $x_1, x_2, \cdots x_n$ useful in predicting the response?

Multiple regression

Some important questions

Now we have to evaluate the model:

1. Is at least one of the predictors $x_1, x_2, \cdots x_n$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?

Multiple regression

Some important questions

Now we have to evaluate the model:

1. Is at least one of the predictors $x_1, x_2, \cdots x_n$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?

Multiple regression

Some important questions

Now we have to evaluate the model:

1. Is at least one of the predictors $x_1, x_2, \cdots x_n$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Multiple regression

Some important questions

Now we have to evaluate the model:

1. Is at least one of the predictors $x_1, x_2, \cdots x_n$ useful in predicting the response?

Important questions

One: Is There a Relationship Between the Response and Predictors?

- In simple linear regression $(y = \beta_0 + \beta_1 x)$, we simply check whether $\beta_1 = 0$ through hypothesis testing.

Important questions

One: Is There a Relationship Between the Response and Predictors?

- In simple linear regression ($y = \beta_0 + \beta_1 x$), we simply check whether $\beta_1 = 0$ through hypothesis testing.
- Here, we extend that idea with p predictors.

Important questions

One: Is There a Relationship Between the Response and Predictors?

- In simple linear regression ($y = \beta_0 + \beta_1 x$), we simply check whether $\beta_1 = 0$ through hypothesis testing.
- Here, we extend that idea with p predictors.
- We need to ask:
  1. Are all regression coefficients zero?

Important questions

One: Is There a Relationship Between the Response and Predictors?

- In simple linear regression ($y = \beta_0 + \beta_1 x$), we simply check whether $\beta_1 = 0$ through hypothesis testing.
- Here, we extend that idea with p predictors.
- We need to ask:
  1. Are all regression coefficients zero?
  2. Are only a particular subset of regression coefficients zero?

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

To know whether all of the regression coefficients are zero, i.e. $\beta_1 = \beta_2 = \cdots = \beta_p = 0.$, we test the null hypothesis,

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

versus the alternative

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

To know whether all of the regression coefficients are zero, i.e. $\beta_1 = \beta_2 = \cdots = \beta_p = 0.$, we test the null hypothesis,

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

versus the alternative

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

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

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

Assuming homoscedasticity, the F-statistic is given by,

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

where

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

Assuming homoscedasticity, the F-statistic is given by,

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

where

- $TSS = \sum(y_i - \bar{y})^2$ is the **total sum of squares**. Measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

Assuming homoscedasticity, the F-statistic is given by,

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

where

- $TSS = \sum(y_i - \bar{y})^2$ is the **total sum of squares**. Measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.
- $RSS = \sum(y_i - \hat{y}_i)^2$ is the **residual sum of squares**. Measures the amount of variability that is left unexplained after performing the regression.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

Assuming homoscedasticity, the F-statistic is given by,

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

where

- $TSS = \sum(y_i - \bar{y})^2$ is the **total sum of squares**. Measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.
- $RSS = \sum(y_i - \hat{y}_i)^2$ is the **residual sum of squares**. Measures the amount of variability that is left unexplained after performing the regression.
- $TSS - RSS \rightarrow$ is the amount of variability in the response that is explained (or removed) by performing the regression.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

Assuming homoscedasticity, the F-statistic is given by,

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

where

- $TSS = \sum(y_i - \bar{y})^2$ is the **total sum of squares**. Measures the total variance in the response Y , and can be thought of as the amount of variability inherent in the response before the regression is performed.
- $RSS = \sum(y_i - \hat{y}_i)^2$ is the **residual sum of squares**. Measures the amount of variability that is left unexplained after performing the regression.
- $TSS - RSS \rightarrow$ is the amount of variability in the response that is explained (or removed) by performing the regression.

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

- If we assume homoscedasticity then, $\mathbb{E}\{RSS/(n-p-1)\} = \sigma^2$

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

- If we assume homoscedasticity then, $\mathbb{E}\{RSS/(n-p-1)\} = \sigma^2$
- Assuming that H_0 is true, then $\mathbb{E}\{(TSS - RSS)/p\} = \sigma^2$.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

- If we assume homoscedasticity then, $\mathbb{E}\{RSS/(n-p-1)\} = \sigma^2$
- Assuming that H_0 is true, then $\mathbb{E}\{(TSS - RSS)/p\} = \sigma^2$.
- Therefore, if H_0 is true (i.e. there is no relationship between $x_1, \cdots x_p$ and Y), F-statistic is closer to 1.

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

- If we assume homoscedasticity then, $\mathbb{E}\{RSS/(n-p-1)\} = \sigma^2$
- Assuming that H_0 is true, then $\mathbb{E}\{(TSS - RSS)/p\} = \sigma^2$.
- Therefore, if H_0 is true (i.e. there is no relationship between $x_1, \cdots x_p$ and Y), F-statistic is closer to 1.
- On other hand, if H_0 is not true, $\mathbb{E}\{(TSS - RSS)/p\} > \sigma^2 \Rightarrow F > 1$

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

- If we assume homoscedasticity then, $\mathbb{E}\{RSS/(n-p-1)\} = \sigma^2$
- Assuming that H_0 is true, then $\mathbb{E}\{(TSS - RSS)/p\} = \sigma^2$.
- Therefore, if H_0 is true (i.e. there is no relationship between $x_1, \cdots x_p$ and Y), F-statistic is closer to 1.
- On other hand, if H_0 is not true, $\mathbb{E}\{(TSS - RSS)/p\} > \sigma^2 \Rightarrow F > 1$
- How large does the F-statistic need to be before we can reject H_0 and conclude that there is a relationship?

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

- If we assume homoscedasticity then, $\mathbb{E}\{RSS/(n-p-1)\} = \sigma^2$
- Assuming that H_0 is true, then $\mathbb{E}\{(TSS - RSS)/p\} = \sigma^2$.
- Therefore, if H_0 is true (i.e. there is no relationship between $x_1, \cdots x_p$ and Y), F-statistic is closer to 1.
- On other hand, if H_0 is not true, $\mathbb{E}\{(TSS - RSS)/p\} > \sigma^2 \Rightarrow F > 1$
- How large does the F-statistic need to be before we can reject H_0 and conclude that there is a relationship?
  $\rightarrow$ It depends on the values of n and p.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

When n is large,

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.
- a F-statistic that is just a little larger than 1 might still provide evidence against H_0.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.
- a F-statistic that is just a little larger than 1 might still provide evidence against H_0.

When n is small,

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.
- a F-statistic that is just a little larger than 1 might still provide evidence against H_0.

When n is small,

- A larger F-statistic is is needed to reject H_0.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.
- a F-statistic that is just a little larger than 1 might still provide evidence against H_0.

When n is small,

- A larger F-statistic is is needed to reject H_0.

For any value n and p,

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.
- a F-statistic that is just a little larger than 1 might still provide evidence against H_0.

When n is small,

- A larger F-statistic is is needed to reject H_0.

For any value n and p,

- Compute the $p - value$: The p-value indicates how likely it is to observe the results due to chance, assuming H_0.

Important questions

One: Is There a Relationship Between the Response and Predictors? - Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.
- a F-statistic that is just a little larger than 1 might still provide evidence against H_0.

When n is small,

- A larger F-statistic is is needed to reject H_0.

For any value n and p,

- Compute the $p - value$: The p-value indicates how likely it is to observe the results due to chance, assuming H_0.
  - Small p-value: very unlikely that H_0 is true $\rightarrow$ Reject H_0.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.
- a F-statistic that is just a little larger than 1 might still provide evidence against H_0.

When n is small,

- A larger F-statistic is is needed to reject H_0.

For any value n and p,

- Compute the $p - value$: The p-value indicates how likely it is to observe the results due to chance, assuming H_0.
  - Small p-value: very unlikely that H_0 is true $\rightarrow$ Reject H_0.
  - Large p-value: very likely that H_0 is true $\rightarrow$ Fail to reject H_0.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are all regression coefficients zero?

When n is large,

- F-statistic approximately follows an F-distribution, even if the errors are not normally distributed.
- a F-statistic that is just a little larger than 1 might still provide evidence against H_0.

When n is small,

- A larger F-statistic is is needed to reject H_0.

For any value n and p,

- Compute the $p - value$: The p-value indicates how likely it is to observe the results due to chance, assuming H_0.
  - Small p-value: very unlikely that H_0 is true $\rightarrow$ Reject H_0.
  - Large p-value: very likely that H_0 is true $\rightarrow$ Fail to reject H_0.

Typical p-value cutoffs for rejecting the null hypothesis are 5% or 1%.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

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

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

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

versus the alternative,

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

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

versus the alternative,

$$H_a : \text{ One or more than } q \text{ restrictions assuming } H_0 \text{ does not stand.}$$

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

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

versus the alternative,

$$H_a : \text{ One or more than } q \text{ restrictions assuming } H_0 \text{ does not stand.}$$

To test the hypothesis:

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

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

versus the alternative,

$$H_a : \text{ One or more than } q \text{ restrictions assuming } H_0 \text{ does not stand.}$$

To test the hypothesis:

- Fit a regression model y_0 that uses all variables except q.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

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

versus the alternative,

$$H_a : \text{ One or more than } q \text{ restrictions assuming } H_0 \text{ does not stand.}$$

To test the hypothesis:

- Fit a regression model y_0 that uses all variables except q.
- Calculate the residual sum of squares for that model, RSS_0.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

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

versus the alternative,

$$H_a: \text{ One or more than } q \text{ restrictions assuming } H_0 \text{ does not stand.}$$

To test the hypothesis:

- Fit a regression model y_0 that uses all variables except q.
- Calculate the residual sum of squares for that model, RSS_0.
- Compute the appropiate F-statistic

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}.$$

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

Sometimes we want to test the hypothesis if a particular subset q of the coefficients are zero,

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

versus the alternative,

$$H_a : \text{ One or more than } q \text{ restrictions assuming } H_0 \text{ does not stand.}$$

To test the hypothesis:

- Fit a regression model y_0 that uses all variables except q.
- Calculate the residual sum of squares for that model, RSS_0.
- Compute the appropiate F-statistic

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(n-p-1)}.$$

- Compute p-values.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

- The approach of using an F -statistic to test for any association between the predictors and the response works when p is relatively small compared to n.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

- The approach of using an F -statistic to test for any association between the predictors and the response works when p is relatively small compared to n.
- If $p > n$ then there are more coefficients β_j to estimate than observations from which to estimate them.

Important questions

One: Is There a Relationship Between the Response and Predictors?
- Are only a particular subset of regression coefficients zero?

- The approach of using an F -statistic to test for any association between the predictors and the response works when p is relatively small compared to n.
- If $p > n$ then there are more coefficients β_j to estimate than observations from which to estimate them.
- We cannot even fit the multiple linear regression model using least squares, so the **F-statistic cannot be used**.

Multiple regression

Some important questions

1. Is at least one of the predictors $x_1, x_2, \cdots x_n$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Multiple regression

Some important questions

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

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors.

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,
2. A model containing x_1 only,

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,
2. A model containing x_1 only,
3. A model containing x_2 only, and

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,
2. A model containing x_1 only,
3. A model containing x_2 only, and
4. A model containing both x_1 and x_2.

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,
2. A model containing x_1 only,
3. A model containing x_2 only, and
4. A model containing both x_1 and x_2.

Then we can use Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2 to select the best model.

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,
2. A model containing x_1 only,
3. A model containing x_2 only, and
4. A model containing both x_1 and x_2.

Then we can use Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2 to select the best model.
Unfortunately, there are a total of 2^p models that contain subsets of p variables.

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,
2. A model containing x_1 only,
3. A model containing x_2 only, and
4. A model containing both x_1 and x_2.

Then we can use Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2 to select the best model.
Unfortunately, there are a total of 2^p models that contain subsets of p variables.

- If if $p = 2$, then there are $2^2 = 4$ models to consider.

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,
2. A model containing x_1 only,
3. A model containing x_2 only, and
4. A model containing both x_1 and x_2.

Then we can use Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2 to select the best model.
Unfortunately, there are a total of 2^p models that contain subsets of p variables.

- If if $p = 2$, then there are $2^2 = 4$ models to consider.
- if $p = 30$, then we must consider $2^{30} = 1.073.741.824$ models!

Important questions

Two: Deciding on Important Variables

Ideally, we would like to perform **variable selection** by trying out a lot of different models, each containing a different subset of the predictors. For example, if $p = 2$, then we can consider,

1. A model containing no variables,
2. A model containing x_1 only,
3. A model containing x_2 only, and
4. A model containing both x_1 and x_2.

Then we can use Akaike information criterion (AIC), Bayesian information criterion (BIC), and adjusted R^2 to select the best model.
Unfortunately, there are a total of 2^p models that contain subsets of p variables.

- If if $p = 2$, then there are $2^2 = 4$ models to consider.
- if $p = 30$, then we must consider $2^{30} = 1.073.741.824$ models!
  $\rightarrow$ **This is not practical!**

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.
  3. Add to the null model the variable that results in the lowest RSS, this create a new model y_1.

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.
  3. Add to the null model the variable that results in the lowest RSS, this create a new model y_1.
  4. Repeat the step 2 and 3 but now with $p-1$ regressors.

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.
  3. Add to the null model the variable that results in the lowest RSS, this create a new model y_1.
  4. Repeat the step 2 and 3 but now with $p-1$ regressors.
  5. Continue until some stopping rule is satisfied.

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.
  3. Add to the null model the variable that results in the lowest RSS, this create a new model y_1.
  4. Repeat the step 2 and 3 but now with $p-1$ regressors.
  5. Continue until some stopping rule is satisfied.
- **Backward selection**

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.
  3. Add to the null model the variable that results in the lowest RSS, this create a new model y_1.
  4. Repeat the step 2 and 3 but now with $p - 1$ regressors.
  5. Continue until some stopping rule is satisfied.
- **Backward selection**
  1. We start with all the variables in the model y_p.

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.
  3. Add to the null model the variable that results in the lowest RSS, this create a new model y_1.
  4. Repeat the step 2 and 3 but now with $p-1$ regressors.
  5. Continue until some stopping rule is satisfied.
- **Backward selection**
  1. We start with all the variables in the model y_p.
  2. Remove the variable with the largest p-value.

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.
  3. Add to the null model the variable that results in the lowest RSS, this create a new model y_1.
  4. Repeat the step 2 and 3 but now with $p-1$ regressors.
  5. Continue until some stopping rule is satisfied.
- **Backward selection**
  1. We start with all the variables in the model y_p.
  2. Remove the variable with the largest p-value.
  3. The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.

Two: Deciding on Important Variables

So when p is not small, we can consider these automated approaches:

- **Forward selection**
  1. We begin with by fitting the *null model* y_0.
  2. Fit p simple linear regressions and compute its RSS.
  3. Add to the null model the variable that results in the lowest RSS, this create a new model y_1.
  4. Repeat the step 2 and 3 but now with $p-1$ regressors.
  5. Continue until some stopping rule is satisfied.
- **Backward selection**
  1. We start with all the variables in the model y_p.
  2. Remove the variable with the largest p-value.
  3. The new $(p-1)$-variable model is fit, and the variable with the largest p-value is removed.
  4. Continue until some stopping rule is satisfied.

Two: Deciding on Important Variables

- **Mixed selection**: a combination of forward and backward selection.

Two: Deciding on Important Variables

- **Mixed selection**: a combination of forward and backward selection.
    1. We start with the null model and then we successively add the variables that provides the best fit (lowest RSS).

Two: Deciding on Important Variables

- **Mixed selection**: a combination of forward and backward selection.
  1. We start with the null model and then we successively add the variables that provides the best fit (lowest RSS).
  2. If at any point the p-value for one of the variables in the model rises above a certain threshold, remove that variable.

Two: Deciding on Important Variables

- **Mixed selection**: a combination of forward and backward selection.
  1. We start with the null model and then we successively add the variables that provides the best fit (lowest RSS).
  2. If at any point the p-value for one of the variables in the model rises above a certain threshold, remove that variable.
  3. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Two: Deciding on Important Variables

- **Mixed selection**: a combination of forward and backward selection.
    1. We start with the null model and then we successively add the variables that provides the best fit (lowest RSS).
    2. If at any point the p-value for one of the variables in the model rises above a certain threshold, remove that variable.
    3. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Important notes:

Two: Deciding on Important Variables

- **Mixed selection**: a combination of forward and backward selection.
  1. We start with the null model and then we successively add the variables that provides the best fit (lowest RSS).
  2. If at any point the p-value for one of the variables in the model rises above a certain threshold, remove that variable.
  3. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Important notes:

- Backward selection cannot be used if $p > n$.

Two: Deciding on Important Variables

- **Mixed selection**: a combination of forward and backward selection.
  1. We start with the null model and then we successively add the variables that provides the best fit (lowest RSS).
  2. If at any point the p-value for one of the variables in the model rises above a certain threshold, remove that variable.
  3. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Important notes:

- Backward selection cannot be used if $p > n$.
- Forward selection can always be used but might include variables early that later become redundant.

Two: Deciding on Important Variables

- **Mixed selection**: a combination of forward and backward selection.
    1. We start with the null model and then we successively add the variables that provides the best fit (lowest RSS).
    2. If at any point the p-value for one of the variables in the model rises above a certain threshold, remove that variable.
    3. We continue to perform these forward and backward steps until all variables in the model have a sufficiently low p-value, and all variables outside the model would have a large p-value if added to the model.

Important notes:

- Backward selection cannot be used if $p > n$.
- Forward selection can always be used but might include variables early that later become redundant.
- Mixed selection can remedy redundant variables.

Multiple regression

Some important questions

1. Is at least one of the predictors $x_1, x_2, \cdots x_n$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Multiple regression

Some important questions

3. How well does the model fit the data?

Important questions

Three: Model fit

The quality of a linear regression fit is typically assessed using two related quantities:

Important questions

Three: Model fit

The quality of a linear regression fit is typically assessed using two related quantities:

1. Residual standard error (RSE)

Important questions

Three: Model fit

The quality of a linear regression fit is typically assessed using two related quantities:

1. Residual standard error (RSE)
2. R^2 statistic.

Important questions

Three: Model fit - Residual standard error (RSE)

- Recall that from every model, there is some error term ϵ associated with each observation.

Important questions

Three: Model fit - Residual standard error (RSE)

- Recall that from every model, there is some error term ϵ associated with each observation.
- The RSE is an estimate of the standard deviation of ϵ.

Important questions

Three: Model fit - Residual standard error (RSE)

- Recall that from every model, there is some error term ϵ associated with each observation.
- The RSE is an estimate of the standard deviation of ϵ.
- Roughly speaking, it is the average amount that the response will deviate from the true regression line.

Important questions

Three: Model fit - Residual standard error (RSE)

- Recall that from every model, there is some error term ϵ associated with each observation.
- The RSE is an estimate of the standard deviation of ϵ.
- Roughly speaking, it is the average amount that the response will deviate from the true regression line.

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

Important questions

Three: Model fit - Residual standard error (RSE)

- Recall that from every model, there is some error term ϵ associated with each observation.
- The RSE is an estimate of the standard deviation of ϵ.
- Roughly speaking, it is the average amount that the response will deviate from the true regression line.

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

- If $\hat{y}_i \approx y_i \, \forall i \in n$, then RSE is small. $\rightarrow$ The model fits the data well.

Important questions

Three: Model fit - Residual standard error (RSE)

- Recall that from every model, there is some error term ϵ associated with each observation.
- The RSE is an estimate of the standard deviation of ϵ.
- Roughly speaking, it is the average amount that the response will deviate from the true regression line.

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

- If $\hat{y}_i \approx y_i \, \forall i \in n$, then RSE is small. $\rightarrow$ The model fits the data well.
- If $\hat{y}_i$ is very far from y_i for one or more observations, then RSE may be quite large. $\rightarrow$ The model doesn't fit the data well.

Important questions

Three: Model fit - R^2 statistic

- The R^2 it's the proportion of variance explained.

Important questions

Three: Model fit - R^2 statistic

- The R^2 it's the proportion of variance explained.
- It always takes on a value between 0 and 1, and is independent of the scale of Y.

Important questions

Three: Model fit - R^2 statistic

- The R^2 it's the proportion of variance explained.
- It always takes on a value between 0 and 1, and is independent of the scale of Y.

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

- TSS measures the total variance in Y

Important questions

Three: Model fit - R^2 statistic

- The R^2 it's the proportion of variance explained.
- It always takes on a value between 0 and 1, and is independent of the scale of Y.

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

- TSS measures the total variance in Y
- RSS measures the variability that is left unexplained after performing the regression.

Important questions

Three: Model fit - R^2 statistic

- The R^2 it's the proportion of variance explained.
- It always takes on a value between 0 and 1, and is independent of the scale of Y.

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

- TSS measures the total variance in Y
- RSS measures the variability that is left unexplained after performing the regression.
- TSS - RSS is the amount of variability in Y that is explained (or removed) by performing the regression.

Important questions

Three: Model fit - R^2 statistic

- The R^2 it's the proportion of variance explained.
- It always takes on a value between 0 and 1, and is independent of the scale of Y.

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

- TSS measures the total variance in Y
- RSS measures the variability that is left unexplained after performing the regression.
- TSS - RSS is the amount of variability in Y that is explained (or removed) by performing the regression.
- R^2 measures the proportion of variability in Y that can be explained using X.

Important questions

Three: Model fit - R^2 statistic

- The R^2 it's the proportion of variance explained.
- It always takes on a value between 0 and 1, and is independent of the scale of Y.

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

- TSS measures the total variance in Y
- RSS measures the variability that is left unexplained after performing the regression.
- TSS - RSS is the amount of variability in Y that is explained (or removed) by performing the regression.
- R^2 measures the proportion of variability in Y that can be explained using X.

Multiple regression

Some important questions

1. Is at least one of the predictors $x_1, x_2, \cdots x_n$ useful in predicting the response?
2. Do all the predictors help to explain Y, or is only a subset of the predictors useful?
3. How well does the model fit the data?
4. Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Multiple regression

Some important questions

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

Important questions

Four: Predictions

The accuracy of our predictions depends on three sorts of uncertainty:

Important questions

Four: Predictions

The accuracy of our predictions depends on three sorts of uncertainty:

1. **The reducible error**
   $\rightarrow$ Related to the inaccuracy in the coefficient estimates.

Important questions

Four: Predictions

The accuracy of our predictions depends on three sorts of uncertainty:

1. **The reducible error**
$\rightarrow$ Related to the inaccuracy in the coefficient estimates.
$\rightarrow$ We compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$.

Important questions

Four: Predictions

The accuracy of our predictions depends on three sorts of uncertainty:

1. **The reducible error**
   $\rightarrow$ Related to the inaccuracy in the coefficient estimates.
   $\rightarrow$ We compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$.
2. **Model bias**

Important questions

Four: Predictions

The accuracy of our predictions depends on three sorts of uncertainty:

1. **The reducible error**
   $\rightarrow$ Related to the inaccuracy in the coefficient estimates.
   $\rightarrow$ We compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$.
2. **Model bias**
   $\rightarrow$ Assuming a linear model for $f(X)$ is an approximation of reality.

Important questions

Four: Predictions

The accuracy of our predictions depends on three sorts of uncertainty:

1. **The reducible error**
   $\rightarrow$ Related to the inaccuracy in the coefficient estimates.
   $\rightarrow$ We compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$.
2. **Model bias**
   $\rightarrow$ Assuming a linear model for $f(X)$ is an approximation of reality.
3. **The irreducible error.**

Important questions

Four: Predictions

The accuracy of our predictions depends on three sorts of uncertainty:

1. **The reducible error**
   $\rightarrow$ Related to the inaccuracy in the coefficient estimates.
   $\rightarrow$ We compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$.
2. **Model bias**
   $\rightarrow$ Assuming a linear model for $f(X)$ is an approximation of reality.
3. **The irreducible error.**
   $\rightarrow$ We use prediction intervals to answer how much will $\hat{Y}$ vary from Y.

Important questions

Four: Predictions

The accuracy of our predictions depends on three sorts of uncertainty:

1. **The reducible error**
   $\rightarrow$ Related to the inaccuracy in the coefficient estimates.
   $\rightarrow$ We compute a confidence interval in order to determine how close $\hat{Y}$ will be to $f(X)$.
2. **Model bias**
   $\rightarrow$ Assuming a linear model for $f(X)$ is an approximation of reality.
3. **The irreducible error.**
   $\rightarrow$ We use prediction intervals to answer how much will $\hat{Y}$ vary from Y.
   $\rightarrow$ Prediction intervals are always wider than confidence intervals.

Table of Contents

Other considerations in the Regression model

Qualitative predictors: Predictors with only two levels

- Suppose that we wish to investigate differences in credit card balance between people who own a house vs those who don't.

Other considerations in the Regression model

Qualitative predictors: Predictors with only two levels

- Suppose that we wish to investigate differences in credit card balance between people who own a house vs those who don't.
- We can adding the qualitative predictor to the regression as a dummy variable.

Other considerations in the Regression model

Qualitative predictors: Predictors with only two levels

- Suppose that we wish to investigate differences in credit card balance between people who own a house vs those who don't.
- We can adding the qualitative predictor to the regression as a dummy variable.

$$x_i = \begin{cases} 1 & \text{if } i\text{th person owns a house,} \\ 0 & \text{if } i\text{th person does not own a house,} \end{cases}$$

- This results in the model,

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

Other considerations in the Regression model

Qualitative predictors

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

- β_0 is the average credit card balance among those who do not own a house.

Other considerations in the Regression model

Qualitative predictors

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

- β_0 is the average credit card balance among those who do not own a house.
- $\beta_0 + \beta_1$ is the average credit card balance among those who do own their house.

Other considerations in the Regression model

Qualitative predictors

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

- β_0 is the average credit card balance among those who do not own a house.
- $\beta_0 + \beta_1$ is the average credit card balance among those who do own their house.
- β_1 is the average difference in credit card balance between owners and non-owners.

Other considerations in the Regression model

Qualitative predictors

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

- β_0 is the average credit card balance among those who do not own a house.
- $\beta_0 + \beta_1$ is the average credit card balance among those who do own their house.
- β_1 is the average difference in credit card balance between owners and non-owners.
- To test significance, we test $H_0 : \beta_1 = 0$

Other considerations in the Regression model

Qualitative predictors

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

- β_0 is the average credit card balance among those who do not own a house.
- $\beta_0 + \beta_1$ is the average credit card balance among those who do own their house.
- β_1 is the average difference in credit card balance between owners and non-owners.
- To test significance, we test $H_0 : \beta_1 = 0$

Note

$\rightarrow$ The decision to code owners as 1 and non-owners as 0 is arbitrary, and has no effect on the regression fit.

Other considerations in the Regression model

Qualitative predictors

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person owns a house,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person does not.} \end{cases}$$

- β_0 is the average credit card balance among those who do not own a house.
- $\beta_0 + \beta_1$ is the average credit card balance among those who do own their house.
- β_1 is the average difference in credit card balance between owners and non-owners.
- To test significance, we test $H_0 : \beta_1 = 0$

Note

$\rightarrow$ The decision to code owners as 1 and non-owners as 0 is arbitrary, and has no effect on the regression fit.
$\rightarrow$ That decision only alter the interpretation of the coefficients.

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

- Suppose now that we want to investigate differences in credit card balance between people who lives in one of these **regions**: South, West, East.

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

- Suppose now that we want to investigate differences in credit card balance between people who lives in one of these **regions**: South, West, East.
- Single dummy variable cannot represent all possible values, so we create additional ones.

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

- Suppose now that we want to investigate differences in credit card balance between people who lives in one of these **regions**: South, West, East.
- Single dummy variable cannot represent all possible values, so we create additional ones.

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South,} \\ 0 & \text{if } i\text{th person is not from the South,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West,} \\ 0 & \text{if } i\text{th person is not from the West,} \end{cases}$$

- This results in the model,

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

- Suppose now that we want to investigate differences in credit card balance between people who lives in one of these **regions**: South, West, East.
- Single dummy variable cannot represent all possible values, so we create additional ones.

$$x_{i1} = \begin{cases} 1 & \text{if } i\text{th person is from the South,} \\ 0 & \text{if } i\text{th person is not from the South,} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if } i\text{th person is from the West,} \\ 0 & \text{if } i\text{th person is not from the West,} \end{cases}$$

- This results in the model,

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South,} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th is from the East.} \end{cases}$$

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South,} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th is from the East.} \end{cases}$$

- β_0 is the average credit card balance for individuals from the East.

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South,} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th is from the East.} \end{cases}$$

- β_0 is the average credit card balance for individuals from the East.
- β_1 is the difference in the average balance between people from the South versus the East.

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South,} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th is from the East.} \end{cases}$$

- β_0 is the average credit card balance for individuals from the East.
- β_1 is the difference in the average balance between people from the South versus the East.
- β_2 is the difference in the average balance between those from the West versus the East.

Notes

$\rightarrow$ The level selected as the baseline category is arbitrary, and the final predictions for each group will be the same regardless of this choice.

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South,} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th is from the East.} \end{cases}$$

- β_0 is the average credit card balance for individuals from the East.
- β_1 is the difference in the average balance between people from the South versus the East.
- β_2 is the difference in the average balance between those from the West versus the East.

Notes

→ The level selected as the baseline category is arbitrary, and the final predictions for each group will be the same regardless of this choice.
→ The coefficients and their p-values **do depend** on the choice of dummy variable coding.

Other considerations in the Regression model

Qualitative predictors: Predictors with More that Two Levels

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is from the South,} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if } i\text{th person is from the West,} \\ \beta_0 + \epsilon_i & \text{if } i\text{th is from the East.} \end{cases}$$

- β_0 is the average credit card balance for individuals from the East.
- β_1 is the difference in the average balance between people from the South versus the East.
- β_2 is the difference in the average balance between those from the West versus the East.

Notes

→ The level selected as the baseline category is arbitrary, and the final predictions for each group will be the same regardless of this choice.
→ The coefficients and their p-values **do depend** on the choice of dummy variable coding.
→ To test significance, we can use F-test on $H_0 : \beta_1 = \beta_0 = 0$. **This does not depend on the coding**.

Other considerations in the Regression model

Removing the Additive Assumption

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$$

According to this,

Other considerations in the Regression model

Removing the Additive Assumption

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$$

According to this,

- A one-unit increase in X_1 is associated with an average increase in Y of β_1 units.

Other considerations in the Regression model

Removing the Additive Assumption

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$$

According to this,

- A one-unit increase in X_1 is associated with an average increase in Y of β_1 units.
- Notice that the presence of X_2 does not alter this statement.

Other considerations in the Regression model

Removing the Additive Assumption

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$$

According to this,

- A one-unit increase in X_1 is associated with an average increase in Y of β_1 units.
- Notice that the presence of X_2 does not alter this statement.
- We can extend this model to include an **interaction term** with X_1 and X_2.

Other considerations in the Regression model

Removing the Additive Assumption

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$$

According to this,

- A one-unit increase in X_1 is associated with an average increase in Y of β_1 units.
- Notice that the presence of X_2 does not alter this statement.
- We can extend this model to include an **interaction term** with X_1 and X_2.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

Other considerations in the Regression model

Removing the Additive Assumption

Consider the standard linear regression model with two variables,

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 + \epsilon$$

According to this,

- A one-unit increase in X_1 is associated with an average increase in Y of β_1 units.
- Notice that the presence of X_2 does not alter this statement.
- We can extend this model to include an **interaction term** with X_1 and X_2.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

- Now Y can be rewritten as,

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2)X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

- Since $\tilde{\beta}_1$ is now a function of X_1, the association between X_1 and Y is no longer constant.

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2)X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

- Since $\tilde{\beta}_1$ is now a function of X_1, the association between X_1 and Y is no longer constant.
- A change in the value of X_2 will change the association between X_1 and Y.

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

- Since $\tilde{\beta}_1$ is now a function of X_1, the association between X_1 and Y is no longer constant.
- A change in the value of X_2 will change the association between X_1 and Y.
- Similarly, a change in the value of X_1 changes the association between X_2 and Y.

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

- Since $\tilde{\beta}_1$ is now a function of X_1, the association between X_1 and Y is no longer constant.
- A change in the value of X_2 will change the association between X_1 and Y.
- Similarly, a change in the value of X_1 changes the association between X_2 and Y.

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2)X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$

with $\tilde{\beta}_1 = \beta_1 + \beta_3 X_2$.

The hierarchical principle

If we include an **interaction** in a model, we should also include the **main effects**, even if the p-values associated with principle their coefficients are not significant.

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

The hierarchical principle

If we include an **interaction** in a model, we should also include the **main effects**, even if the p-values associated with principle their coefficients are not significant.

Why?

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

The hierarchical principle

If we include an **interaction** in a model, we should also include the **main effects**, even if the p-values associated with principle their coefficients are not significant.

Why?

- If $X_1 \times X_2$ is related to the response, then whether or not the coefficients of X_1 or X_2 are exactly zero is of little interest.

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

The hierarchical principle

If we include an **interaction** in a model, we should also include the **main effects**, even if the p-values associated with principle their coefficients are not significant.

Why?

- If $X_1 \times X_2$ is related to the response, then whether or not the coefficients of X_1 or X_2 are exactly zero is of little interest.
- $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

Other considerations in the Regression model

Removing the Additive Assumption

$$Y = \beta_0 + (\beta_1 X_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon$$
$$Y = \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon$$
$$\text{with } \tilde{\beta}_1 = \beta_1 + \beta_3 X_2.$$

The hierarchical principle

If we include an **interaction** in a model, we should also include the **main effects**, even if the p-values associated with principle their coefficients are not significant.

Why?

- If $X_1 \times X_2$ is related to the response, then whether or not the coefficients of X_1 or X_2 are exactly zero is of little interest.
- $X_1 \times X_2$ is typically correlated with X_1 and X_2, and so leaving them out tends to alter the meaning of the interaction.

Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

Potential Problems

When we fit a linear regression model to a particular data set, many problems may occur. Most common among these are the following:

1. Non-linearity of the response-predictor relationships.

Potential Problems

Non-linearity of the Data

- The linear regression model assumes that there is a straight-line relationship between the predictors and the response.

Potential Problems

Non-linearity of the Data

- The linear regression model assumes that there is a straight-line relationship between the predictors and the response.
- If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect.

Potential Problems

Non-linearity of the Data

- The linear regression model assumes that there is a straight-line relationship between the predictors and the response.
- If the true relationship is far from linear, then virtually all of the conclusions that we draw from the fit are suspect.
- *Residual plots* are a useful graphical tool for identifying non-linearity.
- In the case of a multiple regression model, we plot the residuals versus the predicted (or fitted) values $\hat{y}_i$.

Potential Problems

Non-linearity of the Data

- Ideally, the residual plot will show no fitted discernible pattern.

Potential Problems

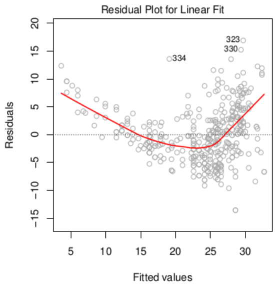
Non-linearity of the Data

- Ideally, the residual plot will show no fitted discernible pattern.
- The presence of a pattern may indicate a problem with some aspect of the linear model.

Potential Problems

Non-linearity of the Data

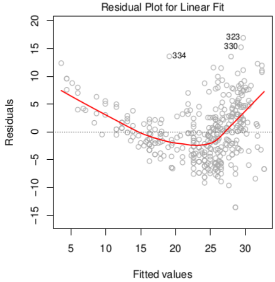
- Ideally, the residual plot will show no fitted discernible pattern.
- The presence of a pattern may indicate a problem with some aspect of the linear model.



Potential Problems

Non-linearity of the Data

- Ideally, the residual plot will show no fitted discernible pattern.
- The presence of a pattern may indicate a problem with some aspect of the linear model.



Note

If the residual plot indicates that there are non-linear associations in the data, then a simple approach is to use **non-linear transformations** of the predictors, such as $\log X$ and $\sqrt{X}$, in the regression model.

Potential Problems

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

Potential Problems

2. Correlation of error terms.

Potential Problems

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, are uncorrelated.

Potential Problems

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, are uncorrelated.
  $\rightarrow$ i.e. the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.

Potential Problems

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, are uncorrelated.
  $\rightarrow$ i.e. the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.
- Is there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors.

Potential Problems

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, are uncorrelated.
  $\rightarrow$ i.e. the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.
- Is there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors.
- As a result, confidence and prediction intervals will be narrower than they should be.

Potential Problems

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, are uncorrelated.
  $\rightarrow$ i.e. the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.
- Is there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors.
- As a result, confidence and prediction intervals will be narrower than they should be.
- In addition, p-values associated with the model will be lower than they should be; this could cause us to erroneously conclude that a parameter is statistically significant.

Potential Problems

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, are uncorrelated.
  $\rightarrow$ i.e. the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.
- Is there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors.
- As a result, confidence and prediction intervals will be narrower than they should be.
- In addition, p-values associated with the model will be lower than they should be; this could cause us to erroneously conclude that a parameter is statistically significant.
- Why might correlations among the error terms occur?

Potential Problems

Correlation of Error Terms

- An important assumption of the linear regression model is that the error terms, $\epsilon_1, \epsilon_2, \cdots, \epsilon_n$, are uncorrelated.
  $\rightarrow$ i.e. the fact that ϵ_i is positive provides little or no information about the sign of ϵ_{i+1}.
- Is there is correlation among the error terms, then the estimated standard errors will tend to underestimate the true standard errors.
- As a result, confidence and prediction intervals will be narrower than they should be.
- In addition, p-values associated with the model will be lower than they should be; this could cause us to erroneously conclude that a parameter is statistically significant.
- Why might correlations among the error terms occur?
  $\rightarrow$ Such correlations frequently occur in the context of **time series data**.

Potential Problems

Correlation of Error Terms

- In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time.

Potential Problems

Correlation of Error Terms

- In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time.
- If the errors are uncorrelated, then there should be no discernible pattern.

Potential Problems

Correlation of Error Terms

- In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time.
- If the errors are uncorrelated, then there should be no discernible pattern.
- If the error terms are positively correlated, then we may see tracking in the residuals.

Potential Problems

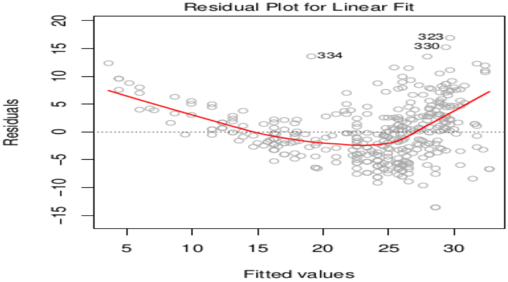
Correlation of Error Terms

- In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time.
- If the errors are uncorrelated, then there should be no discernible pattern.
- If the error terms are positively correlated, then we may see tracking in the residuals. $\rightarrow$ adjacent residuals may have similar values.

Potential Problems

Correlation of Error Terms

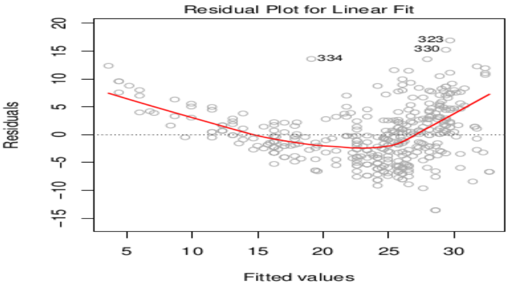
- In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time.
- If the errors are uncorrelated, then there should be no discernible pattern.
- If the error terms are positively correlated, then we may see tracking in the residuals. $\rightarrow$ adjacent residuals may have similar values.



Potential Problems

Correlation of Error Terms

- In order to determine if this is the case for a given data set, we can plot the residuals from our model as a function of time.
- If the errors are uncorrelated, then there should be no discernible pattern.
- If the error terms are positively correlated, then we may see tracking in the residuals. $\rightarrow$ adjacent residuals may have similar values.



- Good experimental design is crucial in order to mitigate the risk of such correlations.

Potential Problems

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

Potential Problems

3. Non-constant variance of error terms.

Potential Problems

Non-constant Variance of Error Terms

- Another important assumption of the linear regression model is that the error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$

Potential Problems

Non-constant Variance of Error Terms

- Another important assumption of the linear regression model is that the error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$
- One can identify non-constant variances in the errors, or heteroscedasticity, from the presence of a funnel shape in the residual plot.

Potential Problems

Non-constant Variance of Error Terms

- Another important assumption of the linear regression model is that the error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$
- One can identify non-constant variances in the errors, or heteroscedasticity, from the presence of a funnel shape in the residual plot.
- When faced with this problem, one possible solution is to transform the response Y using a **concave** function such as $log(Y)$ or $\sqrt{Y}$.

Potential Problems

Non-constant Variance of Error Terms

- Another important assumption of the linear regression model is that the error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$
- One can identify non-constant variances in the errors, or heteroscedasticity, from the presence of a funnel shape in the residual plot.
- When faced with this problem, one possible solution is to transform the response Y using a **concave** function such as $log(Y)$ or $\sqrt{Y}$.
- Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity.

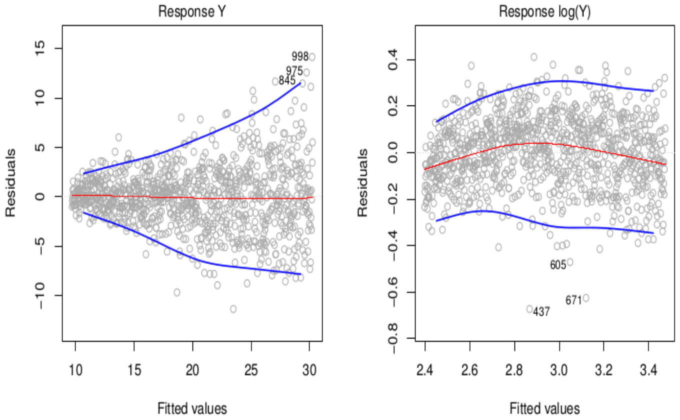
Potential Problems

Non-constant Variance of Error Terms

- Another important assumption of the linear regression model is that the error terms have a constant variance, $Var(\epsilon_i) = \sigma^2$
- One can identify non-constant variances in the errors, or heteroscedasticity, from the presence of a funnel shape in the residual plot.
- When faced with this problem, one possible solution is to transform the response Y using a **concave** function such as $log(Y)$ or $\sqrt{Y}$.
- Such a transformation results in a greater amount of shrinkage of the larger responses, leading to a reduction in heteroscedasticity.

Potential Problems

Non-constant Variance of Error Terms



Potential Problems

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

Potential Problems

4. Outliers.

Potential Problems

Outliers

- An outlier is a point for which yi is far from the value predicted by the outlier model.

Potential Problems

Outliers

- An outlier is a point for which yi is far from the value predicted by the outlier model.
- Include outliers in the regression fit can cause alterations on the RSE values, which are further used to compute confidence intervals and p-values.

Potential Problems

Outliers

- An outlier is a point for which yi is far from the value predicted by the outlier model.
- Include outliers in the regression fit can cause alterations on the RSE values, which are further used to compute confidence intervals and p-values.
- A single data point can have implications for the interpretation of the fit.

Potential Problems

Outliers

- An outlier is a point for which yi is far from the value predicted by the outlier model.
- Include outliers in the regression fit can cause alterations on the RSE values, which are further used to compute confidence intervals and p-values.
- A single data point can have implications for the interpretation of the fit.
- To identify outliers, we can compute the **studentized residuals** by dividing each residual e_i by its estimated standard error.

Potential Problems

Outliers

- An outlier is a point for which yi is far from the value predicted by the outlier model.
- Include outliers in the regression fit can cause alterations on the RSE values, which are further used to compute confidence intervals and p-values.
- A single data point can have implications for the interpretation of the fit.
- To identify outliers, we can compute the **studentized residuals** by dividing each residual e_i by its estimated standard error.
- Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

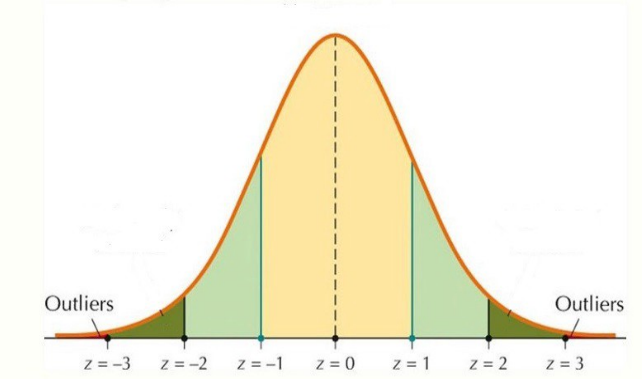
Potential Problems

Outliers

- An outlier is a point for which yi is far from the value predicted by the outlier model.
- Include outliers in the regression fit can cause alterations on the RSE values, which are further used to compute confidence intervals and p-values.
- A single data point can have implications for the interpretation of the fit.
- To identify outliers, we can compute the **studentized residuals** by dividing each residual e_i by its estimated standard error.
- Observations whose studentized residuals are greater than 3 in absolute value are possible outliers.

Potential Problems

Outliers

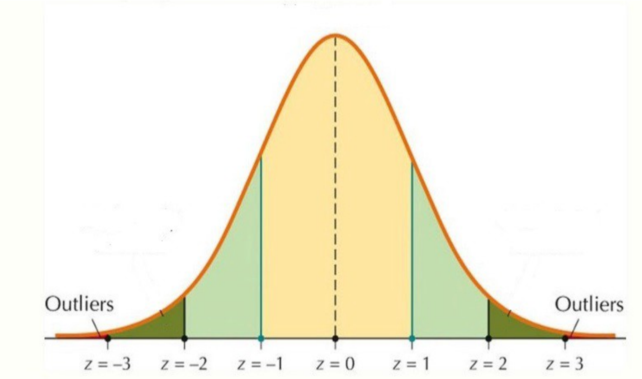


Notes

- If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation.

Potential Problems

Outliers



Notes

- If we believe that an outlier has occurred due to an error in data collection or recording, then one solution is to simply remove the observation.
- However, care should be taken, since an outlier may instead indicate a

Potential Problems

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

Potential Problems

5. High-leverage points.

Potential Problems

Outliers

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i.

Potential Problems

Outliers

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i.
- In contrast, observations with **high leverage** have an **unusual value for** x_i.

Potential Problems

Outliers

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i.
- In contrast, observations with **high leverage** have an **unusual value for** x_i.
- High leverage observations tend to have a sizable impact on the estimated regression line.

Potential Problems

Outliers

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i.
- In contrast, observations with **high leverage** have an **unusual value for** x_i.
- High leverage observations tend to have a sizable impact on the estimated regression line.
- In order to quantify an observation's leverage, we compute the leverage statistic:

Potential Problems

Outliers

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i.
- In contrast, observations with **high leverage** have an **unusual value for** x_i.
- High leverage observations tend to have a sizable impact on the estimated regression line.
- In order to quantify an observation's leverage, we compute the leverage statistic:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_i' - \bar{x})^2}$$

Potential Problems

Outliers

- We just saw that outliers are observations for which the response y_i is unusual given the predictor x_i.
- In contrast, observations with **high leverage** have an **unusual value for** x_i.
- High leverage observations tend to have a sizable impact on the estimated regression line.
- In order to quantify an observation's leverage, we compute the leverage statistic:

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_i' - \bar{x})^2}$$

Potential Problems

Outliers

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_i' - \bar{x})^2}$$

Potential Problems

Outliers

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_i' - \bar{x})^2}$$

- A large value of this statistic indicates an observation with high leverage.

Notes

$\rightarrow$ h_i increases with the distance of x_i from $\bar{x}$.

Potential Problems

Outliers

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x'_i - \bar{x})^2}$$

- A large value of this statistic indicates an observation with high leverage.

Notes

$\rightarrow$ h_i increases with the distance of x_i from $\bar{x}$.
$\rightarrow$ The average leverage for all the observations is always equal to $(p+1)/n$.

Potential Problems

Outliers

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n} (x_i' - \bar{x})^2}$$

- A large value of this statistic indicates an observation with high leverage.

Notes

$\rightarrow$ h_i increases with the distance of x_i from $\bar{x}$.
$\rightarrow$ The average leverage for all the observations is always equal to $(p+1)/n$.
$\rightarrow$ If a given observation has a leverage statistic that greatly exceeds $(p+1)/n$, then we may suspect that the corresponding point has high leverage.

Potential Problems

1. Non-linearity of the response-predictor relationships.
2. Correlation of error terms.
3. Non-constant variance of error terms.
4. Outliers.
5. High-leverage points.
6. Collinearity.

Potential Problems

6. Collinearity.

Potential Problems

Collinearity

- Collinearity refers to the situation in which two or more predictor variables collinearity are closely related to one another.

Potential Problems

Collinearity

- Collinearity refers to the situation in which two or more predictor variables collinearity are closely related to one another.
- The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.

Potential Problems

Collinearity

- Collinearity refers to the situation in which two or more predictor variables collinearity are closely related to one another.
- The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.
- Collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow.

Potential Problems

Collinearity

- Collinearity refers to the situation in which two or more predictor variables collinearity are closely related to one another.
- The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.
- Collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow.
- Recall that the *t-statistic* for each predictor is calculated by dividing $\hat{\beta}_j$ by its standard error.

Potential Problems

Collinearity

- Collinearity refers to the situation in which two or more predictor variables collinearity are closely related to one another.
- The presence of collinearity can pose problems in the regression context, since it can be difficult to separate out the individual effects of collinear variables on the response.
- Collinearity reduces the accuracy of the estimates of the regression coefficients, it causes the standard error for $\hat{\beta}_j$ to grow.
- Recall that the *t-statistic* for each predictor is calculated by dividing $\hat{\beta}_j$ by its standard error.
- Consequently, collinearity results in a decline in the t-statistic. As a result, in the presence of collinearity, we may fail to reject $H_0 : \beta_j = 0$.

Potential Problems

Collinearity

How to detect collinearity (or multicollinearity):

- Look at the correlation matrix of the predictors.

Potential Problems

Collinearity

How to detect collinearity (or multicollinearity):

- Look at the correlation matrix of the predictors.
  $\rightarrow$ An element of this matrix that is large in absolute value indicates a pair of highly correlated variables.

Potential Problems

Collinearity

How to detect collinearity (or multicollinearity):

- Look at the correlation matrix of the predictors.
  $\rightarrow$ An element of this matrix that is large in absolute value indicates a pair of highly correlated variables.
- Compute the variance inflation factor (VIF).

Potential Problems

Collinearity

How to detect collinearity (or multicollinearity):

- Look at the correlation matrix of the predictors.
  $\rightarrow$ An element of this matrix that is large in absolute value indicates a pair of highly correlated variables.
- Compute the variance inflation factor (VIF).
  $\rightarrow$ The VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the *full model* divided by the variance of $\hat{\beta}_j$ if fit on *its own*.

Potential Problems

Collinearity

How to detect collinearity (or multicollinearity):

- Look at the correlation matrix of the predictors.
  $\rightarrow$ An element of this matrix that is large in absolute value indicates a pair of highly correlated variables.
- Compute the variance inflation factor (VIF).
  $\rightarrow$ The VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the *full model* divided by the variance of $\hat{\beta}_j$ if fit on *its own*.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{Xj|X_{-j}}},$$

Potential Problems

Collinearity

How to detect collinearity (or multicollinearity):

- Look at the correlation matrix of the predictors.
  $\rightarrow$ An element of this matrix that is large in absolute value indicates a pair of highly correlated variables.
- Compute the variance inflation factor (VIF).
  $\rightarrow$ The VIF is the ratio of the variance of $\hat{\beta}_j$ when fitting the *full model* divided by the variance of $\hat{\beta}_j$ if fit on *its own*.

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{Xj|X_{-j}}},$$

where $R^2_{Xj|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

Potential Problems

Collinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{Xj|X_{-j}}},$$

where $R^2_{Xj|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

Potential Problems

Collinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{Xj|X_{-j}}},$$

where $R^2_{Xj|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

- If $VIF \simeq 1 \rightarrow$ complete absence of collinearity.

Potential Problems

Collinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{Xj|X_{-j}}},$$

where $R^2_{Xj|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

- If $VIF \simeq 1 \rightarrow$ complete absence of collinearity.
- If $1 < VIF < 5 \rightarrow$ moderate collinearity.

Potential Problems

Collinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{X_j|X_{-j}}},$$

where $R^2_{X_j|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

- If $VIF \simeq 1 \rightarrow$ complete absence of collinearity.
- If $1 < VIF < 5 \rightarrow$ moderate collinearity.
- If $VIF > 5 \rightarrow$ high collinearity.

Potential Problems

Collinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{Xj|X_{-j}}},$$

where $R^2_{Xj|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

- If $VIF \simeq 1 \rightarrow$ complete absence of collinearity.
- If $1 < VIF < 5 \rightarrow$ moderate collinearity.
- If $VIF > 5 \rightarrow$ high collinearity.

Solutions:

Potential Problems

Collinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{Xj|X_{-j}}},$$

where $R^2_{Xj|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

- If $VIF \simeq 1 \rightarrow$ complete absence of collinearity.
- If $1 < VIF < 5 \rightarrow$ moderate collinearity.
- If $VIF > 5 \rightarrow$ high collinearity.

Solutions:

1. Drop one of the problematic variables from the regression.

Potential Problems

Collinearity

$$VIF(\hat{\beta}_j) = \frac{1}{1 - R^2_{Xj|X_{-j}}},$$

where $R^2_{Xj|X_{-j}}$ is the R^2 from a regression of X_j onto all of the other predictors.

- If $VIF \simeq 1 \rightarrow$ complete absence of collinearity.
- If $1 < VIF < 5 \rightarrow$ moderate collinearity.
- If $VIF > 5 \rightarrow$ high collinearity.

Solutions:

1. Drop one of the problematic variables from the regression.
2. Combine the collinear variables together into a single predictor.

Thank you!

Any question?