LECTURE 15:
OUTLIER AND MISSING VALUES

BUS 211A-3

GROUP WORK

Any help?

Sometimes a dataset can contain extreme values that are outside the range of what is expected and unlike the other data

We will discover outliers and how to identify and remove them from your dataset

MORE CLOSELY

MORE CLOSELY

1. What are Outliers?

MORE CLOSELY

1. What are Outliers?
2. Standard Deviation Method

MORE CLOSELY

1. What are Outliers?
2. Standard Deviation Method
3. Interquartile Range Method

MORE CLOSELY

1. What are Outliers?
2. Standard Deviation Method
3. Interquartile Range Method
4. High Leverage Points

- Define outliers as samples that are exceptionally far from the mainstream of the data

- Define outliers as samples that are exceptionally far from the mainstream of the data
- Outliers can have many causes, such as:

- Define outliers as samples that are exceptionally far from the mainstream of the data
- Outliers can have many causes, such as:
  - Measurement or input error

- Define outliers as samples that are exceptionally far from the mainstream of the data
- Outliers can have many causes, such as:
  - Measurement or input error
  - Data corruption

- Define outliers as samples that are exceptionally far from the mainstream of the data
- Outliers can have many causes, such as:
    - Measurement or input error
    - Data corruption
    - True outlier observation

- Define outliers as samples that are exceptionally far from the mainstream of the data
- Outliers can have many causes, such as:
    - Measurement or input error
    - Data corruption
    - True outlier observation
- There is no precise way to define and identify outliers in general because of the specifics of each dataset

- Define outliers as samples that are exceptionally far from the mainstream of the data
- Outliers can have many causes, such as:
  - Measurement or input error
  - Data corruption
  - True outlier observation
- There is no precise way to define and identify outliers in general because of the specifics of each dataset
- Nevertheless, we can use statistical methods to identify observations that appear to be rare or unlikely given the available data

STANDARD DEVIATION METHOD

STANDARD DEVIATION METHOD

- If we know the distribution of values in the sample is Gaussian-like, we can use the standard deviation of the sample as a cut-off for identifying outliers

STANDARD DEVIATION METHOD

- If we know the distribution of values in the sample is Gaussian-like, we can use the standard deviation of the sample as a cut-off for identifying outliers
- What is the Gaussian distribution?

STANDARD DEVIATION METHOD

- If we know the distribution of values in the sample is Gaussian-like, we can use the standard deviation of the sample as a cut-off for identifying outliers
- What is the Gaussian distribution?
  - summarize the percentage of values in the sample based on mean and s.d. only

STANDARD DEVIATION METHOD

- If we know the distribution of values in the sample is Gaussian-like, we can use the standard deviation of the sample as a cut-off for identifying outliers
- What is the Gaussian distribution?
  - summarize the percentage of values in the sample based on mean and s.d. only
  - e.g. within one s.d. of the mean will cover 68 percent of the data

EXAMPLE

EXAMPLE

- If the mean is 50 and the s.d. is 5

EXAMPLE

- If the mean is 50 and the s.d. is 5
  - then all data in the sample between 45 and 55 will account for about 68 percent of the sample

EXAMPLE

- If the mean is 50 and the s.d. is 5
  - then all data in the sample between 45 and 55 will account for about 68 percent of the sample
- We can cover more of the data sample if we expand the range as follows:

EXAMPLE

- If the mean is 50 and the s.d. is 5
  - then all data in the sample between 45 and 55 will account for about 68 percent of the sample
- We can cover more of the data sample if we expand the range as follows:
  - 1 Standard Deviation from Mean: 68 percent

EXAMPLE

- If the mean is 50 and the s.d. is 5
  - then all data in the sample between 45 and 55 will account for about 68 percent of the sample
- We can cover more of the data sample if we expand the range as follows:
  - 1 Standard Deviation from Mean: 68 percent
  - 2 Standard Deviations from Mean: 95 percent

EXAMPLE

- If the mean is 50 and the s.d. is 5
  - then all data in the sample between 45 and 55 will account for about 68 percent of the sample
- We can cover more of the data sample if we expand the range as follows:
  - 1 Standard Deviation from Mean: 68 percent
  - 2 Standard Deviations from Mean: 95 percent
  - 3 Standard Deviations from Mean: 99.7 percent

EXAMPLE

- If the mean is 50 and the s.d. is 5
  - then all data in the sample between 45 and 55 will account for about 68 percent of the sample
- We can cover more of the data sample if we expand the range as follows:
  - 1 Standard Deviation from Mean: 68 percent
  - 2 Standard Deviations from Mean: 95 percent
  - 3 Standard Deviations from Mean: 99.7 percent
- A value that falls outside of 3 standard deviations is rare event at approximately 1 in 370 samples

- Three s.d. from the mean is a common cut-off in practice for identifying outliers in a Gaussian distribution

- Three s.d. from the mean is a common cut-off in practice for identifying outliers in a Gaussian distribution
  - for smaller samples of data, perhaps a value of 2 s.d. can be used

- Three s.d. from the mean is a common cut-off in practice for identifying outliers in a Gaussian distribution
  - for smaller samples of data, perhaps a value of 2 s.d. can be used
  - for larger samples, perhaps a value of 4 s.d. (99.9 percent) can be used

TEST DATA: PSEUDO CODE

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)
- Define outliers

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)
- Define outliers
  - cut_off := data_std * 3

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)
- Define outliers
  - cut_off := data_std * 3
  - lower, upper := data_mean - cut_off, data_mean + cut_off

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)
- Define outliers
  - cut_off := data_std * 3
  - lower, upper := data_mean - cut_off, data_mean + cut_off
- Remove outliers

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)
- Define outliers
  - cut_off := data_std * 3
  - lower, upper := data_mean - cut_off, data_mean + cut_off
- Remove outliers
  - data < upper and data > lower

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)
- Define outliers
  - cut_off := data_std * 3
  - lower, upper := data_mean - cut_off, data_mean + cut_off
- Remove outliers
  - data < upper and data > lower
- Check the outcomes

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)
- Define outliers
  - cut_off := data_std * 3
  - lower, upper := data_mean - cut_off, data_mean + cut_off
- Remove outliers
  - data < upper and data > lower
- Check the outcomes
  - Identified outliers: 29

TEST DATA: PSEUDO CODE

- Generate a population 10,000 random numbers from Gaussian dist with (50, 5)
  - data_mean, data_std := mean(data), std(data)
- Define outliers
  - cut_off := data_std * 3
  - lower, upper := data_mean - cut_off, data_mean + cut_off
- Remove outliers
  - data < upper and data > lower
- Check the outcomes
  - Identified outliers: 29
  - Non-outlier observations: 9971

INTERQUARTILE RANGE METHOD (IQR)

INTERQUARTILE RANGE METHOD (IQR)

- Not all data is normal

INTERQUARTILE RANGE METHOD (IQR)

- Not all data is normal
- A good statistic for summarizing a non-Gaussian distribution sample of data is the IQR

INTERQUARTILE RANGE METHOD (IQR)

- Not all data is normal
- A good statistic for summarizing a non-Gaussian distribution sample of data is the IQR
  - remember a box (whisker plot)?

INTERQUARTILE RANGE METHOD (IQR)

- Not all data is normal
- A good statistic for summarizing a non-Gaussian distribution sample of data is the IQR
  - remember a box (whisker plot)?
- Refer to the percentiles as quartiles (quart meaning 4) because the data is divided into four groups via the 25th, 50th and 75th values

INTERQUARTILE RANGE METHOD (IQR)

- Not all data is normal
- A good statistic for summarizing a non-Gaussian distribution sample of data is the IQR
  - remember a box (whisker plot)?
- Refer to the percentiles as quartiles (quart meaning 4) because the data is divided into four groups via the 25th, 50th and 75th values
- Identify outliers by defining limits on the sample values that are a factor k of the IQR below the 25th percentile or above the 75th percentile

INTERQUARTILE RANGE METHOD (IQR)

- Not all data is normal
- A good statistic for summarizing a non-Gaussian distribution sample of data is the IQR
  - remember a box (whisker plot)?
- Refer to the percentiles as quartiles (quart meaning 4) because the data is divided into four groups via the 25th, 50th and 75th values
- Identify outliers by defining limits on the sample values that are a factor k of the IQR below the 25th percentile or above the 75th percentile
  - common k: 1.5

TEST DATA: PSEUDO CODE

TEST DATA: PSEUDO CODE

- Use the same data

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers
    - cut_off := iqr * 1.5

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers
    - cut_off := iqr * 1.5
    - lower, upper := q25 - cut_off, q75 + cut_off

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers
    - cut_off := iqr * 1.5
    - lower, upper := q25 - cut_off, q75 + cut_off
- Remove outliers

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers
    - cut_off := iqr * 1.5
    - lower, upper := q25 - cut_off, q75 + cut_off
- Remove outliers
  - data < upper and data > lower

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers
    - cut_off := iqr * 1.5
    - lower, upper := q25 - cut_off, q75 + cut_off
- Remove outliers
  - data < upper and data > lower
- Check the outcomes

TEST DATA: PSEUDO CODE

- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers
    - cut_off := iqr * 1.5
    - lower, upper := q25 - cut_off, q75 + cut_off
- Remove outliers
  - data < upper and data > lower
- Check the outcomes
  - Percentiles: 25th=46.685, 75th=53.359, IQR=6.674

TEST DATA: PSEUDO CODE

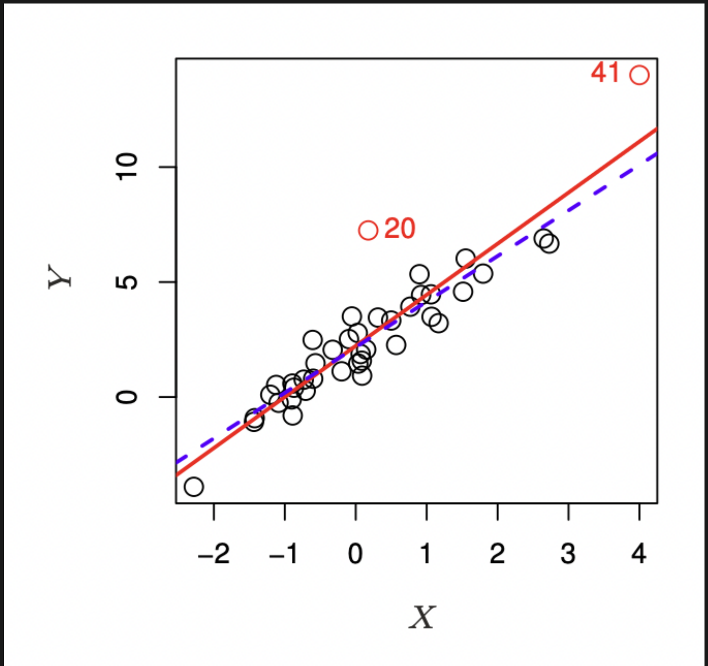
- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers
    - cut_off := iqr * 1.5
    - lower, upper := q25 - cut_off, q75 + cut_off
- Remove outliers
  - data < upper and data > lower
- Check the outcomes
  - Percentiles: 25th=46.685, 75th=53.359, IQR=6.674
  - Identified outliers: 81

TEST DATA: PSEUDO CODE

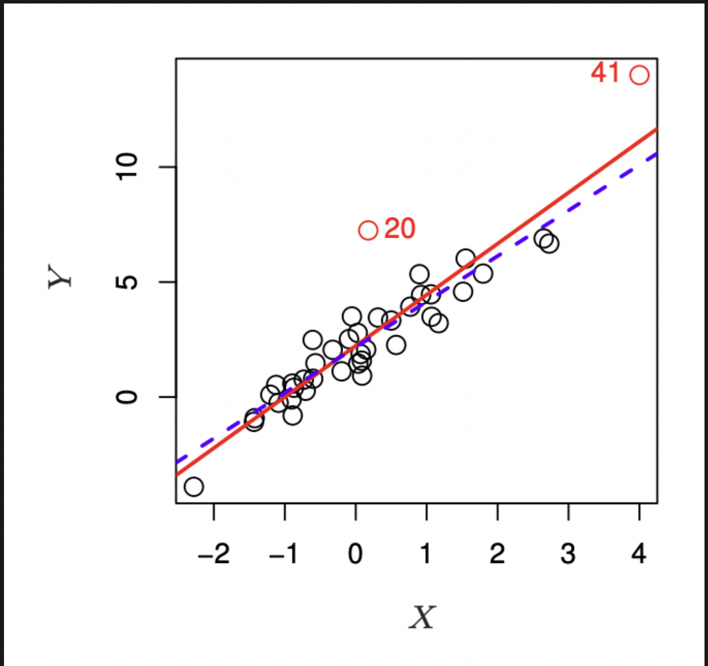
- Use the same data
  - calculate interquartile range
    - iqr = q75 - q25
  - Define outliers
    - cut_off := iqr * 1.5
    - lower, upper := q25 - cut_off, q75 + cut_off
- Remove outliers
  - data < upper and data > lower
- Check the outcomes
  - Percentiles: 25th=46.685, 75th=53.359, IQR=6.674
  - Identified outliers: 81
  - Non-outlier observations: 9919

HIGH LEVERAGE POINTS

HIGH LEVERAGE POINTS

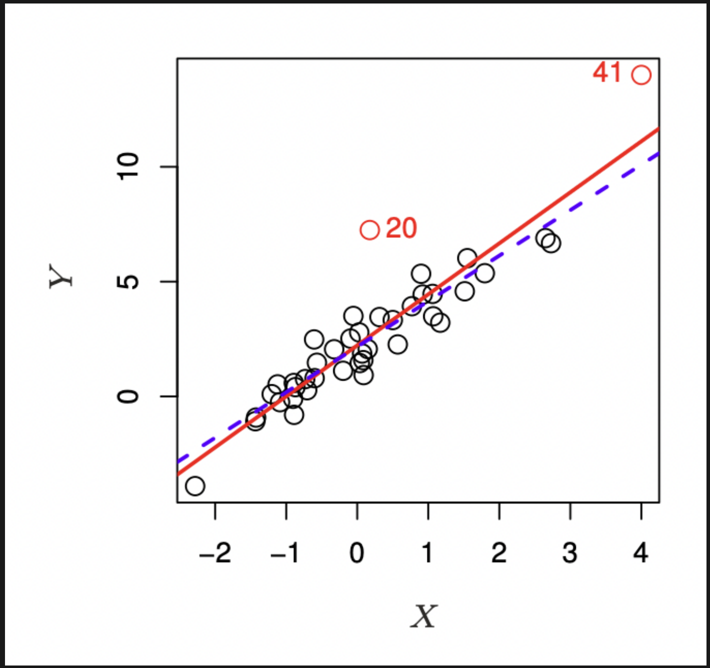

41
20
10
5
0
Y
−2
−1
0
1
2
3
4
X

HIGH LEVERAGE POINTS



- 41 has high leverage: the predictor value for this observation is large relative to the others

HIGH LEVERAGE POINTS

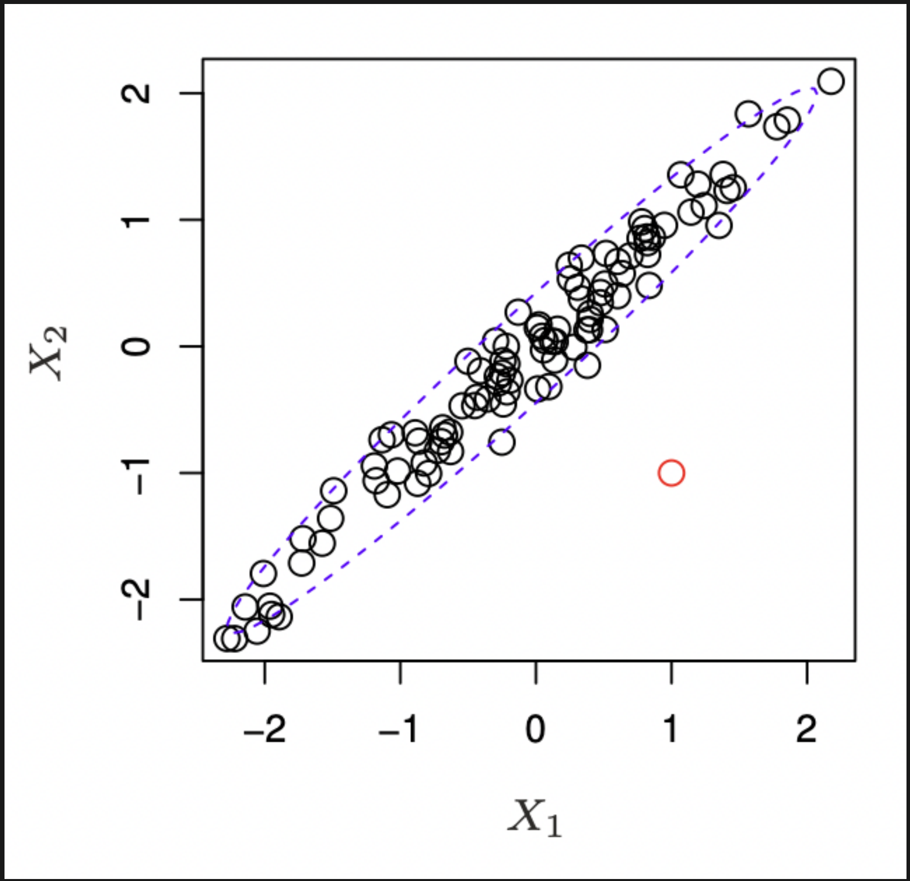


- 41 has high leverage: the predictor value for this observation is large relative to the others
- The red solid line is the least squares fit to the data, while the blue dashed line is the fit produced when observation 41 is removed

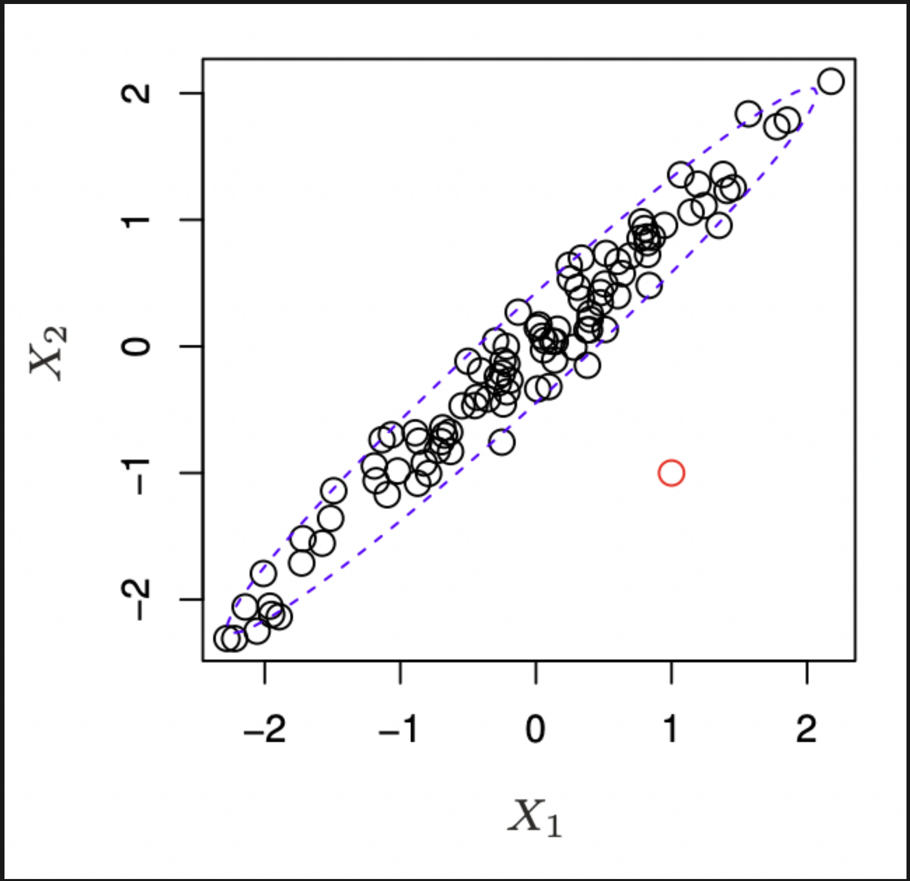
- High leverage observations tend to have a sizable impact on the estimated regression line

- High leverage observations tend to have a sizable impact on the estimated regression line
- It is cause for concern if the least squares line is heavily affected by just a few observations, because any problems with these points may invalidate the entire fit

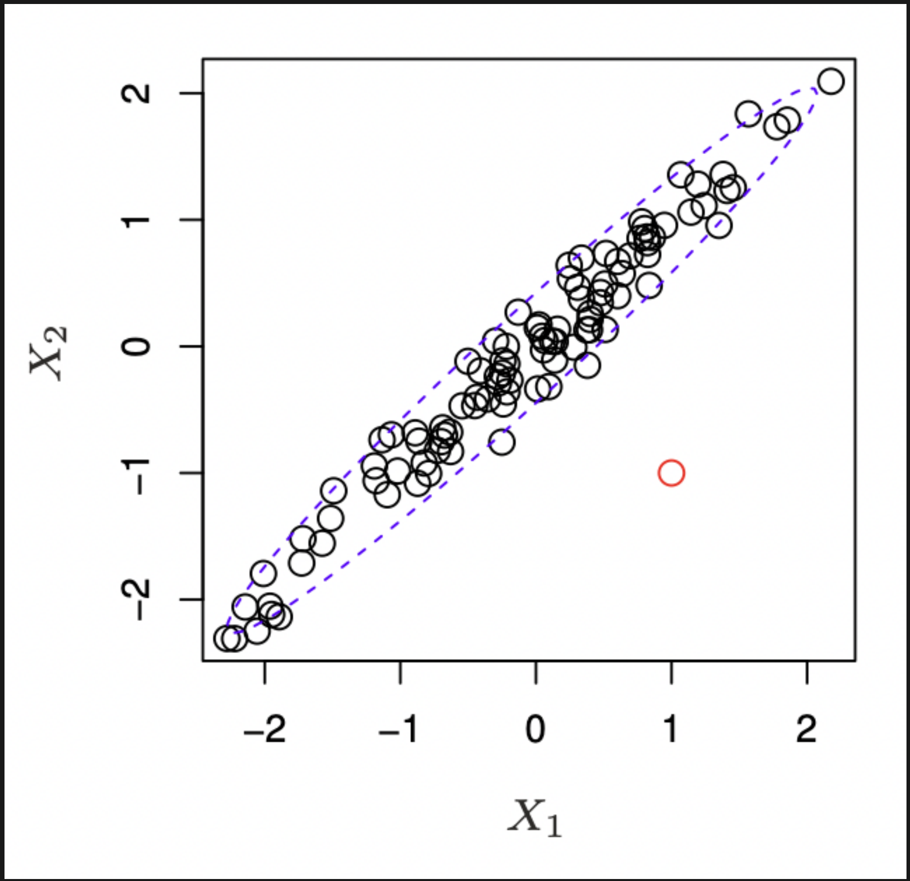
- High leverage observations tend to have a sizable impact on the estimated regression line
- It is cause for concern if the least squares line is heavily affected by just a few observations, because any problems with these points may invalidate the entire fit
- In a simple linear regression, high leverage observations are fairly easy to identify



2
1
0
−1
−2
X_2
−2
−1
0
1
2
X_1



- Neither its value for X_1 nor for X_2 is unusual



- Neither its value for X_1 nor for X_2 is unusual
- So if we examine just X_1 or just X_2, we will fail to notice this high leverage point

- In order to quantify an observation's leverage, we compute the **leverage statistic**

- In order to quantify an observation's leverage, we compute the **leverage statistic**
- A large value of this statistic indicates an observation with high leverage statistic leverage

- In order to quantify an observation's leverage, we compute the **leverage statistic**
- A large value of this statistic indicates an observation with high leverage statistic leverage
- For a simple linear regression

- In order to quantify an observation's leverage, we compute the **leverage statistic**
- A large value of this statistic indicates an observation with high leverage statistic leverage
- For a simple linear regression
- $$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i'=1}^{n}(x_i - \bar{x})^2}$$

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

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

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

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

- h_i increases with the distance of x_i from $\bar{x}$
- There is a simple extension of h_i to multiple predictors. Use the software.

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

- h_i increases with the distance of x_i from $\bar{x}$
- There is a simple extension of h_i to multiple predictors. Use the software.
- h_i is always between 1/n and 1, and the average leverage for all the observations is always equal to (p + 1)/n

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

- h_i increases with the distance of x_i from $\bar{x}$
- There is a simple extension of h_i to multiple predictors. Use the software.
- h_i is always between 1/n and 1, and the average leverage for all the observations is always equal to (p + 1)/n
- So 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

DUPLICATE DATA

DUPLICATE DATA

- Removing duplicate data will be an important step in ensuring your data can be accurately used

DUPLICATE DATA

- Removing duplicate data will be an important step in ensuring your data can be accurately used
- For example, if you are using a train/test split, then it is possible for a duplicate row or rows to appear in both train and test datasets and any evaluation of the model on these rows will be correct

DUPLICATE DATA

- Removing duplicate data will be an important step in ensuring your data can be accurately used
- For example, if you are using a train/test split, then it is possible for a duplicate row or rows to appear in both train and test datasets and any evaluation of the model on these rows will be correct
- How do we define it?

DUPLICATE DATA

- Removing duplicate data will be an important step in ensuring your data can be accurately used
- For example, if you are using a train/test split, then it is possible for a duplicate row or rows to appear in both train and test datasets and any evaluation of the model on these rows will be correct
- How do we define it?
  - identifier / time

DUPLICATE DATA

- Removing duplicate data will be an important step in ensuring your data can be accurately used
- For example, if you are using a train/test split, then it is possible for a duplicate row or rows to appear in both train and test datasets and any evaluation of the model on these rows will be correct
- How do we define it?
  - identifier / time
  - must understand data first

DUPLICATE DATA

- Removing duplicate data will be an important step in ensuring your data can be accurately used
- For example, if you are using a train/test split, then it is possible for a duplicate row or rows to appear in both train and test datasets and any evaluation of the model on these rows will be correct
- How do we define it?
  - identifier / time
  - must understand data first
- How do you identify it?

DUPLICATE DATA

- Removing duplicate data will be an important step in ensuring your data can be accurately used
- For example, if you are using a train/test split, then it is possible for a duplicate row or rows to appear in both train and test datasets and any evaluation of the model on these rows will be correct
- How do we define it?
  - identifier / time
  - must understand data first
- How do you identify it?
- How do you delete it?