Statistics and Machine Learning

Linear regression and beyond: bootstrapping and model complexity

Week 4 02/08 — 02/12 2001

Contents of Week 4

Random data set and statistical meaning

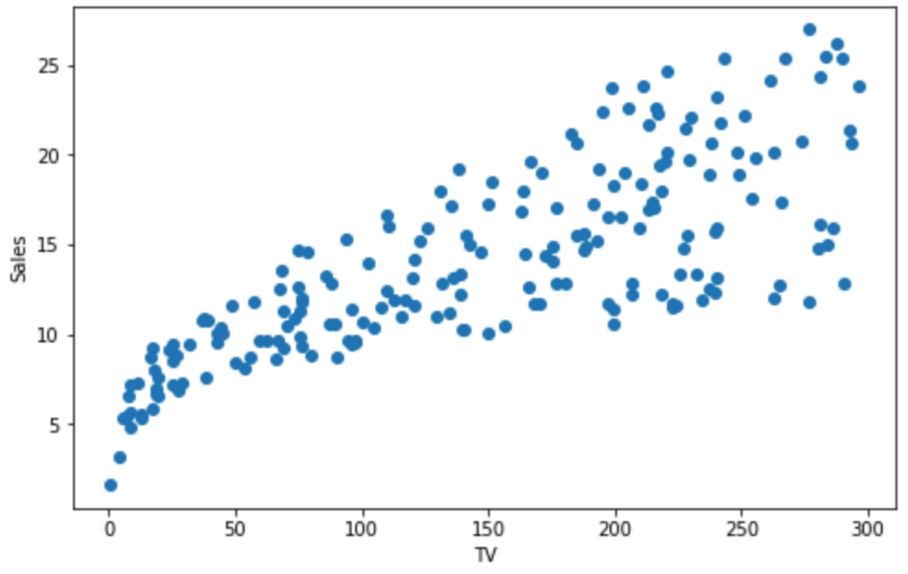
- Random data set
- Interpretation of linear model
- Is ‘newspaper’ irrelevant to ‘sales’?
- Bootstrapping data set
- Normal distribution: one sigma, two sigma, three sigma,…,

Review of linear regression

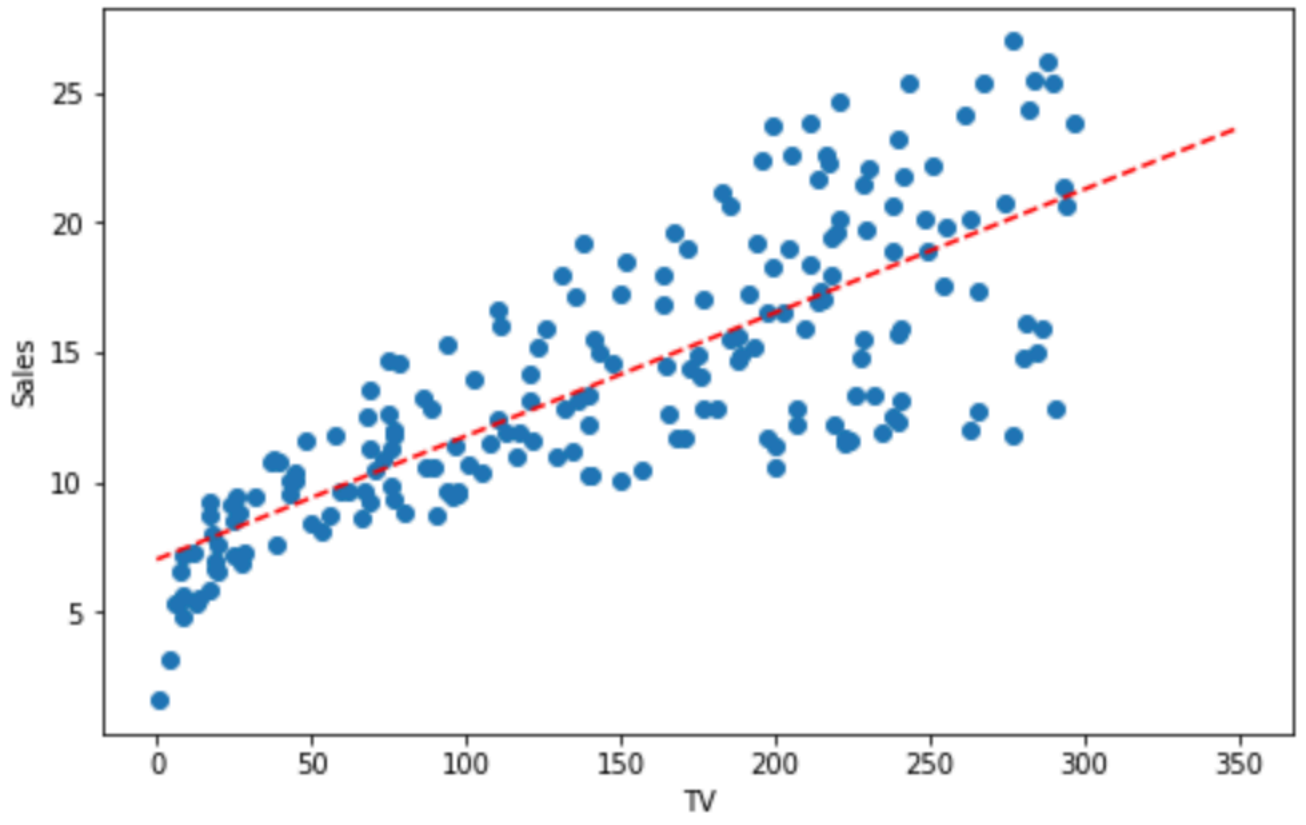
Finding the line which causes the minimum quadratic loss

Model

Data set



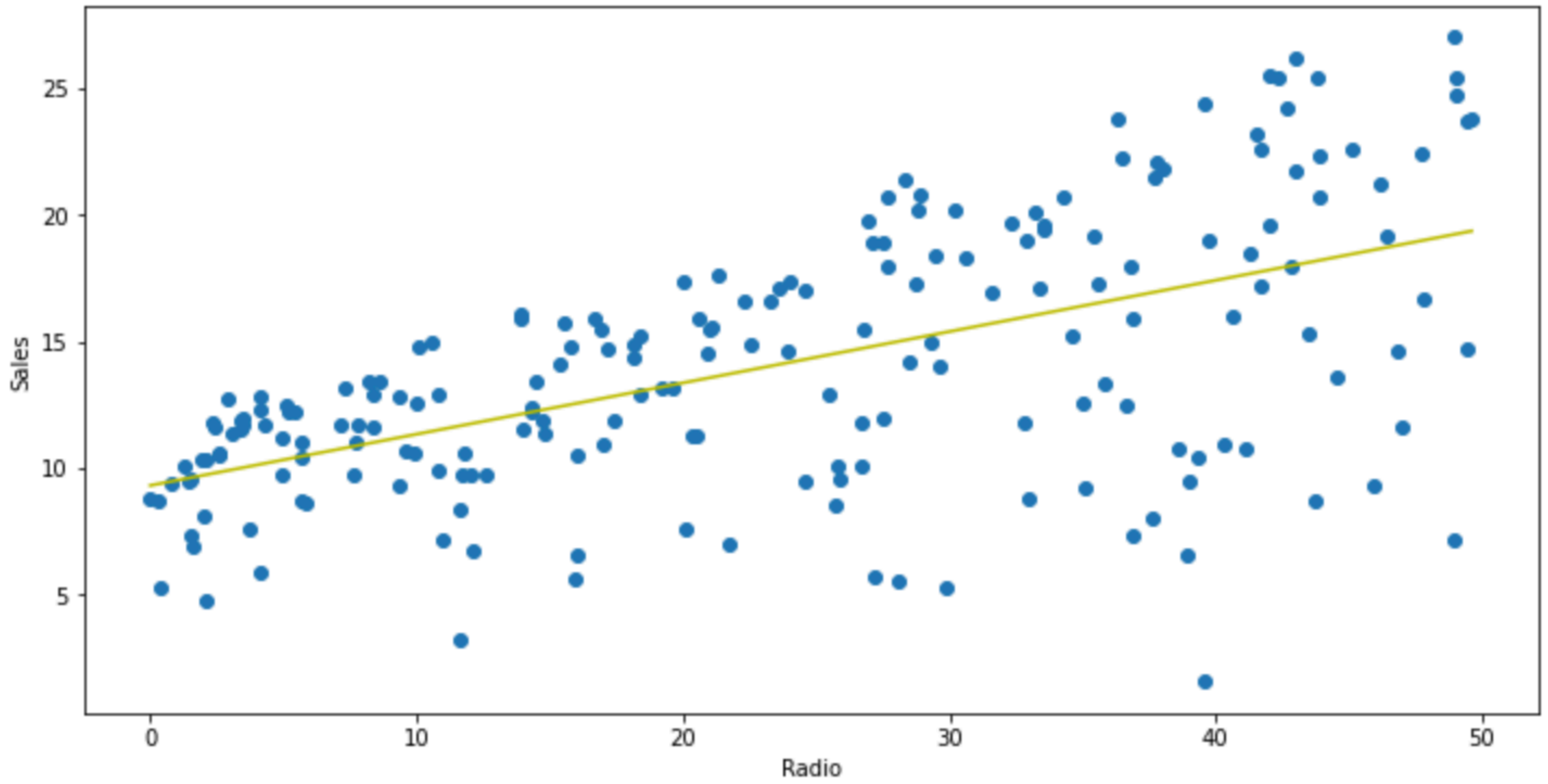
Linear regression

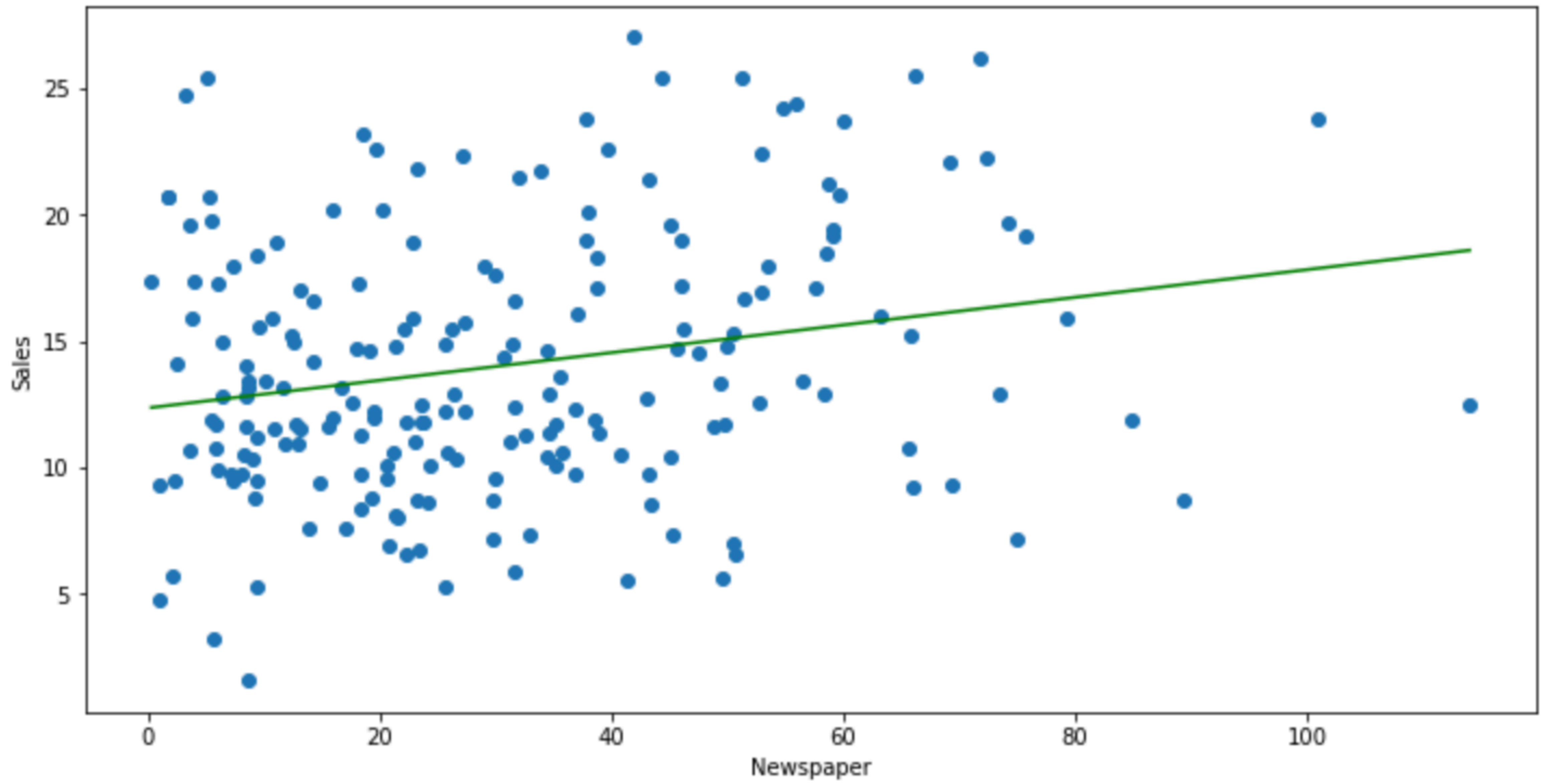


.fit()
a<—.coef_
b <— .intercept_

The linear model: $y = ax + b$

Application to (radio, sales) and (newspaper, sales)

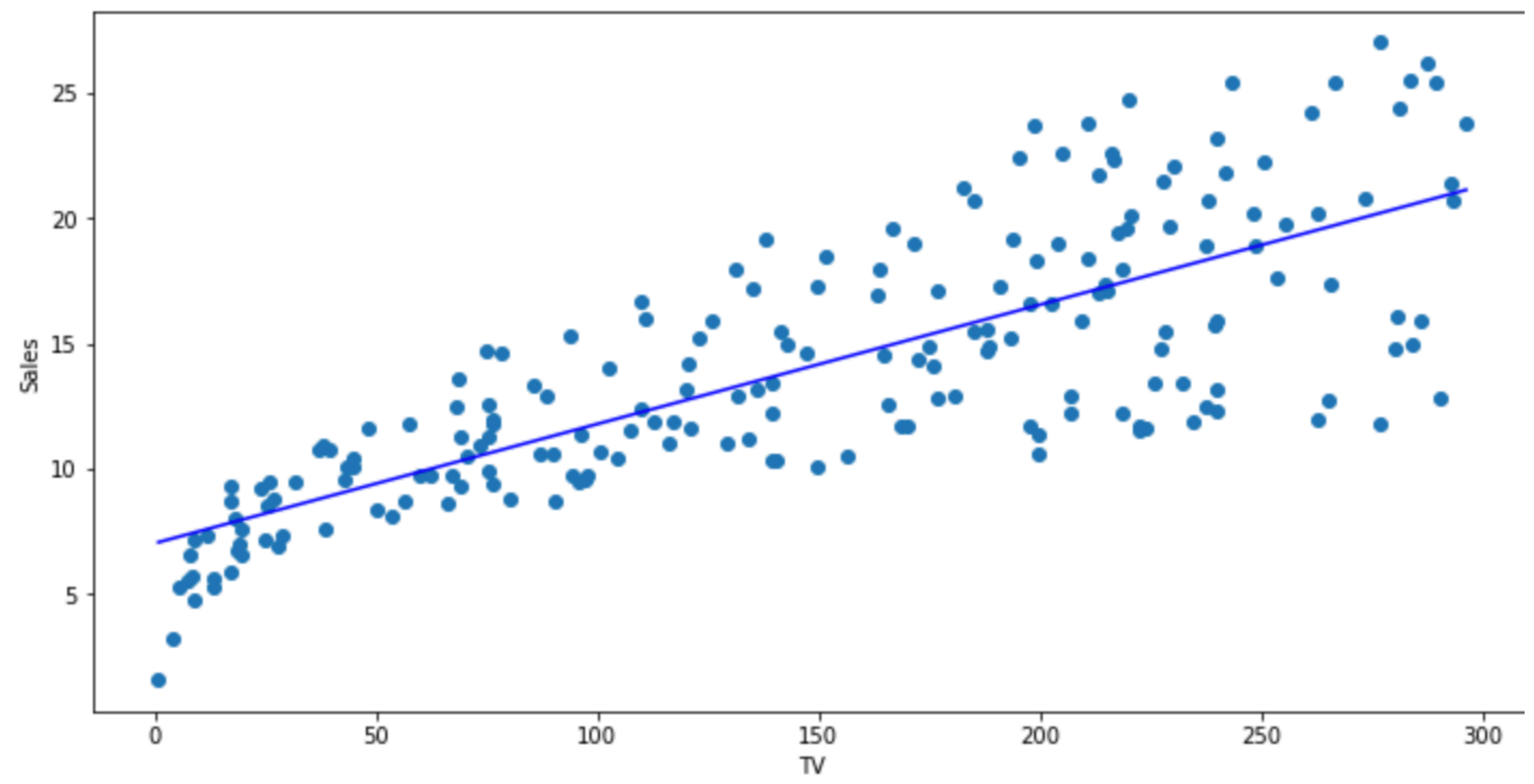




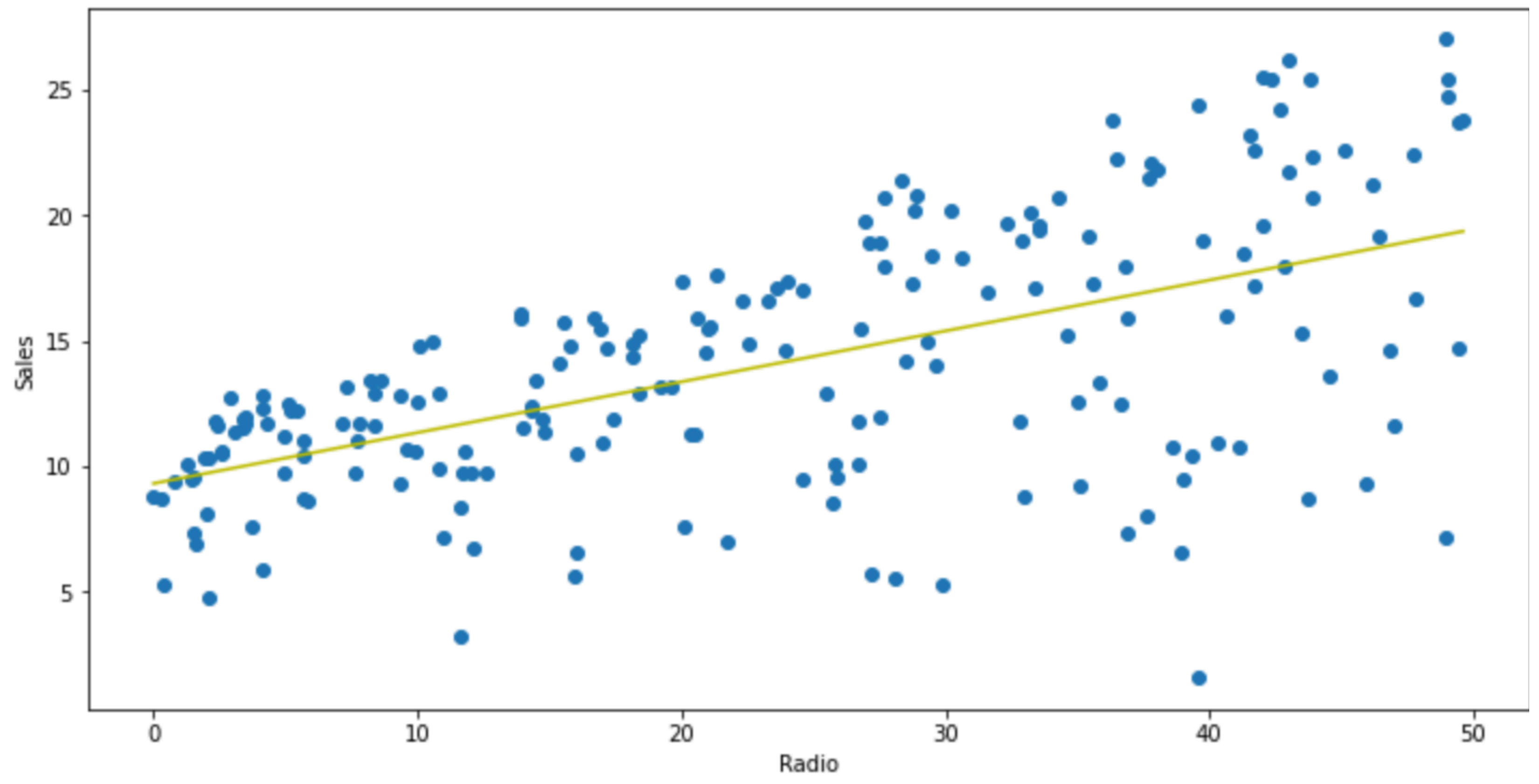
‘TV’, ‘Radio’, and ‘Newspaper’ versus ‘Sales’

Which one is the least relevant?

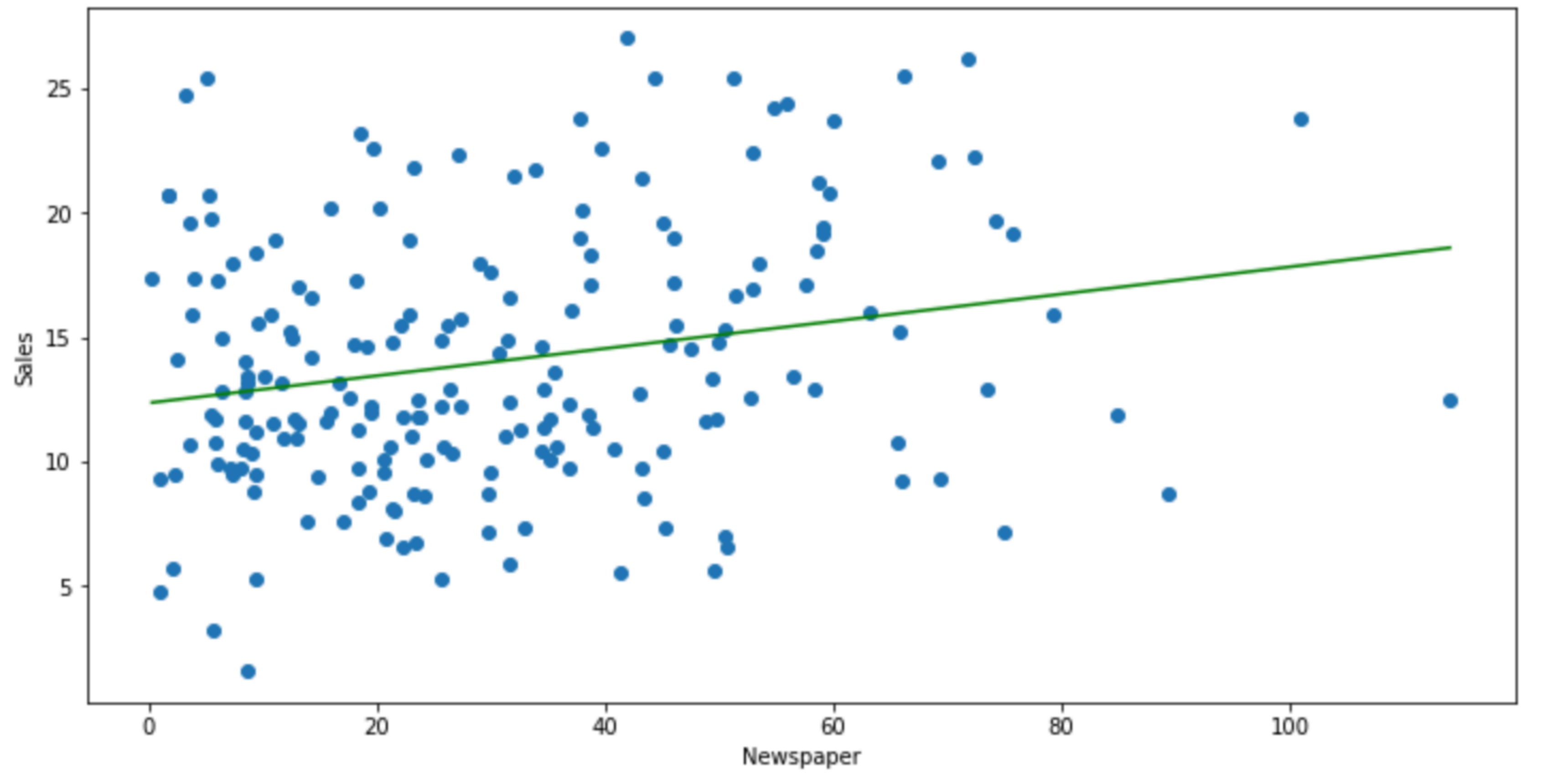
Observation: if the model parameter of slope is close to zero, it means the x input is not relevant to the output!




```
regr.coef_, regr.intercept_
(array([[0.04753664]]), array([7.03259355]))
```




```
regr_radio.coef_, regr_radio.intercept_
(array([[0.20249578]]), array([9.3116381]))
```




```
regr_news.coef_, regr_news.intercept_
(array([[0.0546931]]), array([12.35140707]))
```

Slope = 0.0475

Slope = 0.2

Slope = 0.05

Before we jump to conclusion,

The caveats are

‘Sales’ is the result of all three attributes ‘tv’, ‘radio’, and ‘newspaper’. Every time we consider one attribute and neglect the other two. It might be problematic!

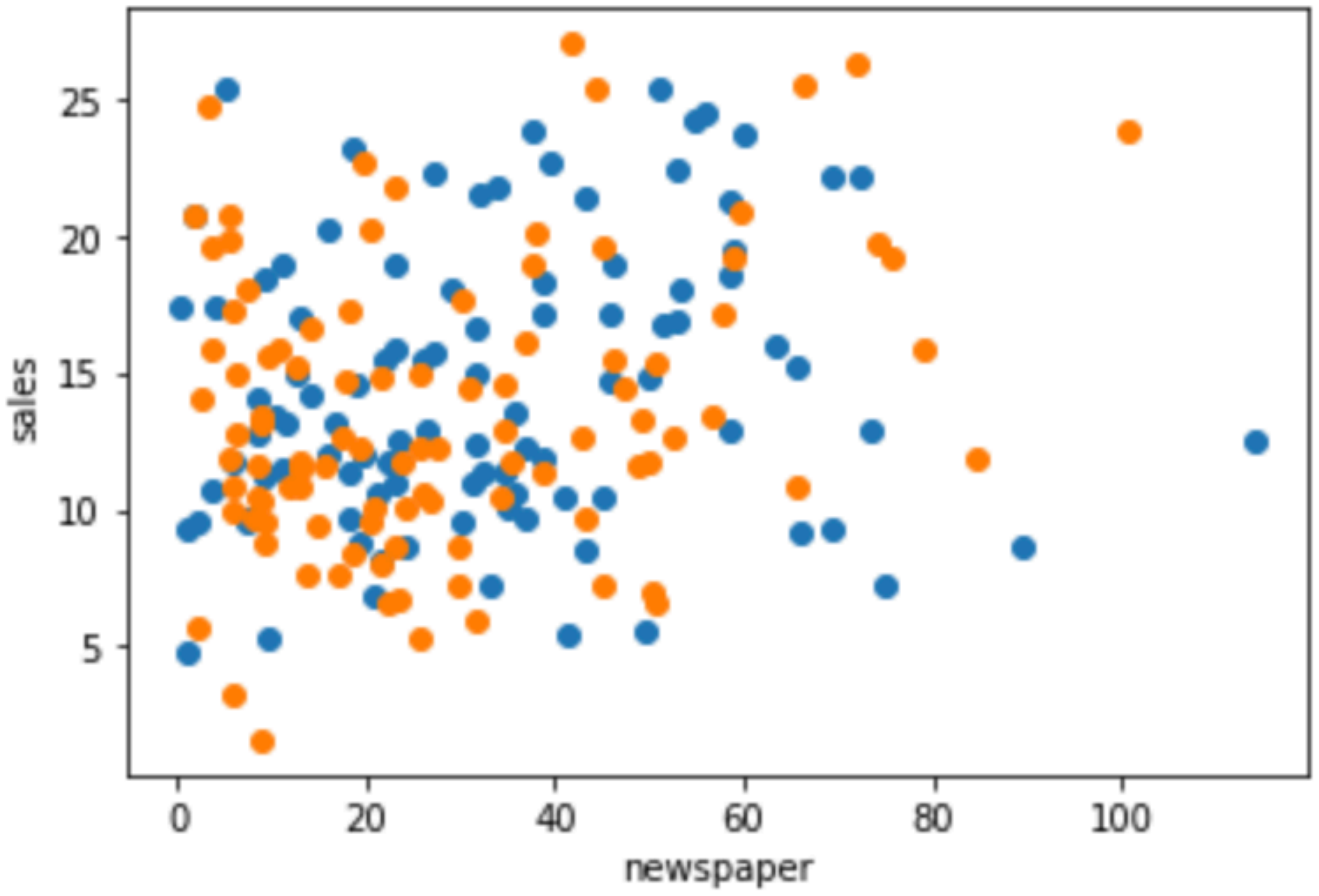
We do not know what does the unit for every attribute mean? So it might lead to comparing apple and orange.

Do we have a way out? yes.

Different data led to different conclusions

Example: sales versus newspaper

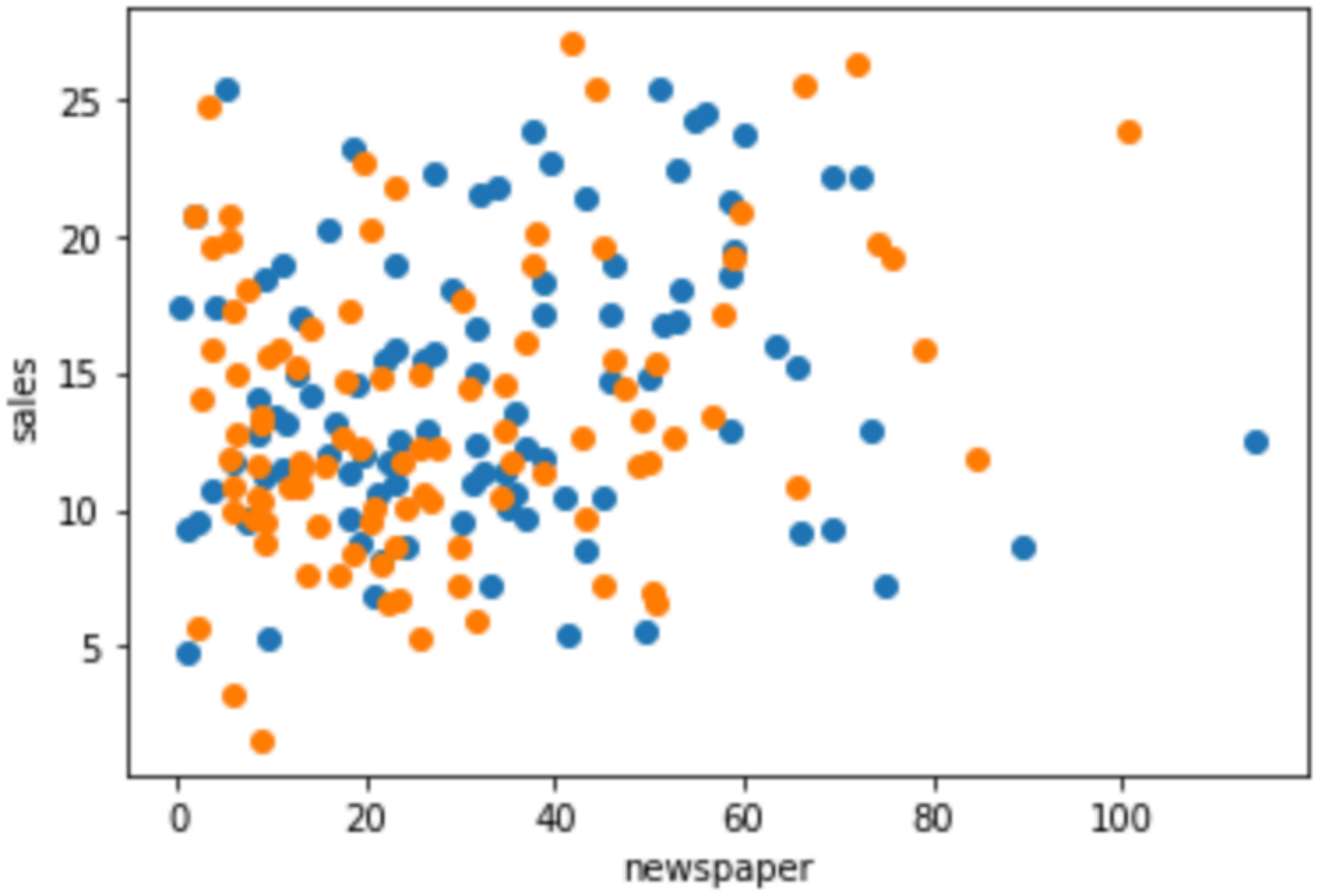
200 data split into 100 + 100 data

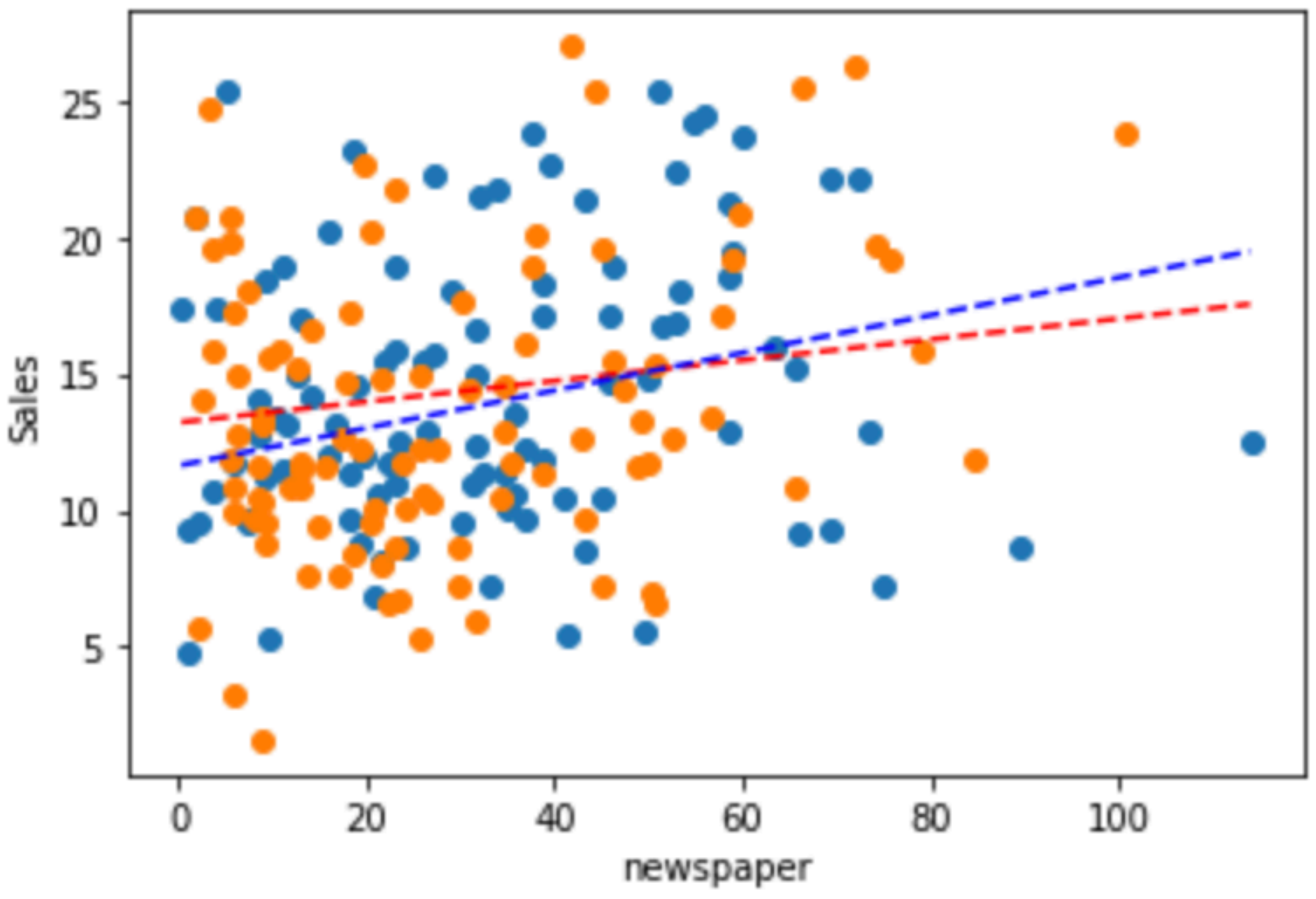


Different data led to different conclusions

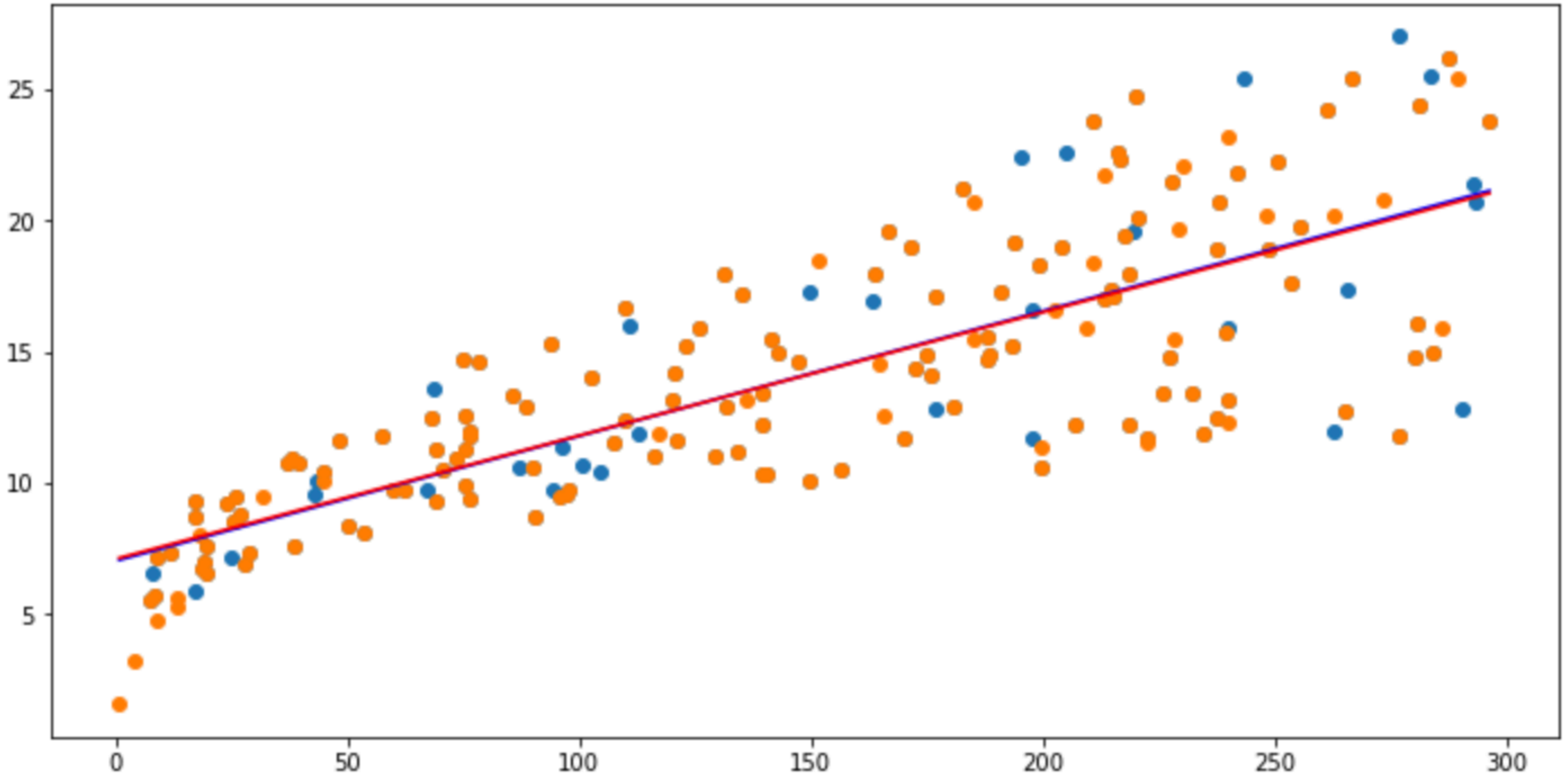
Example: sales versus newspaper

200 data split into 100 + 100 data



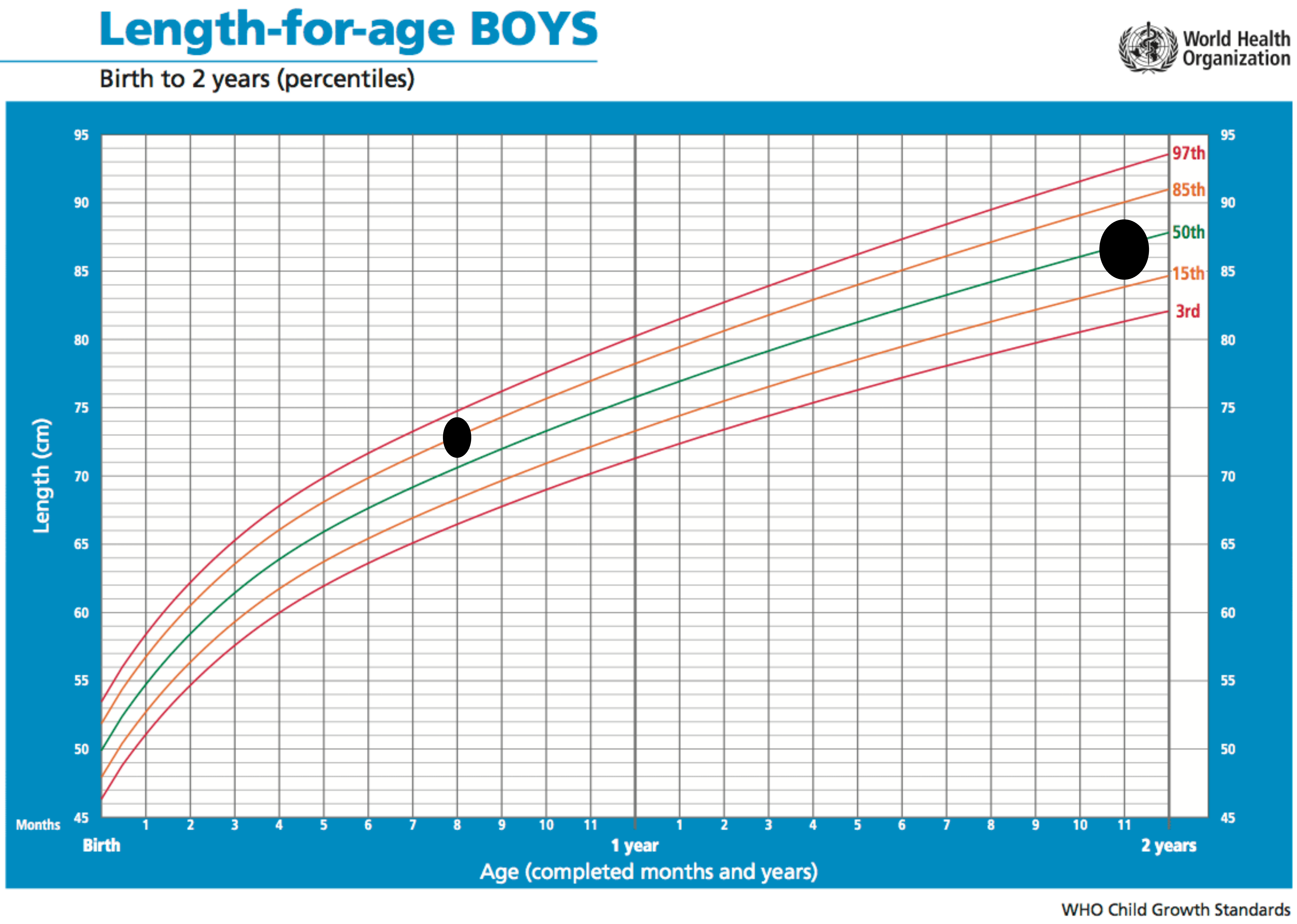


For (tv, sales), different data sets lead to almost identical linear model



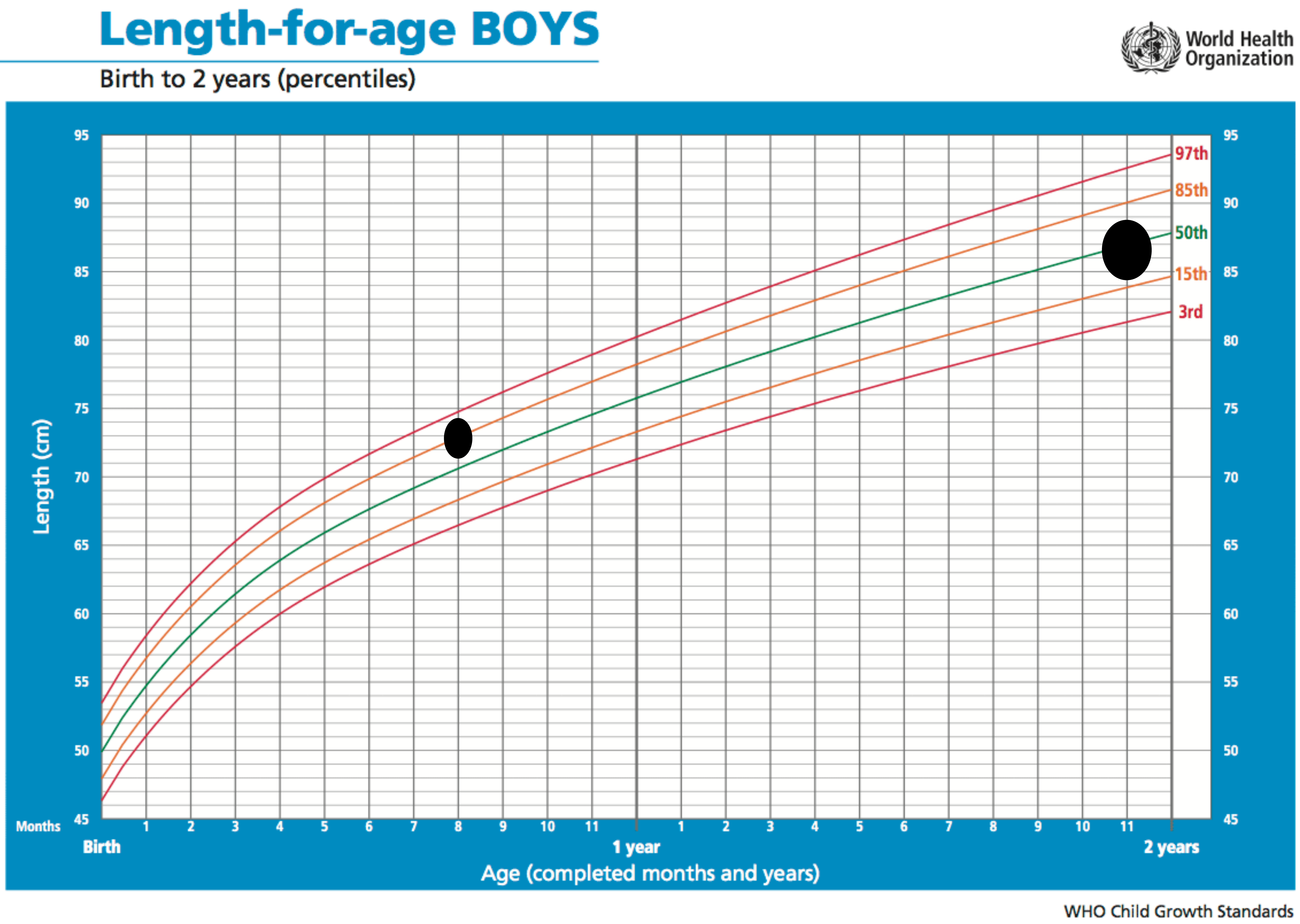
How to compare who is taller between elder and younger brothers?

Standardized data



How to compare who is taller between elder and younger brothers?

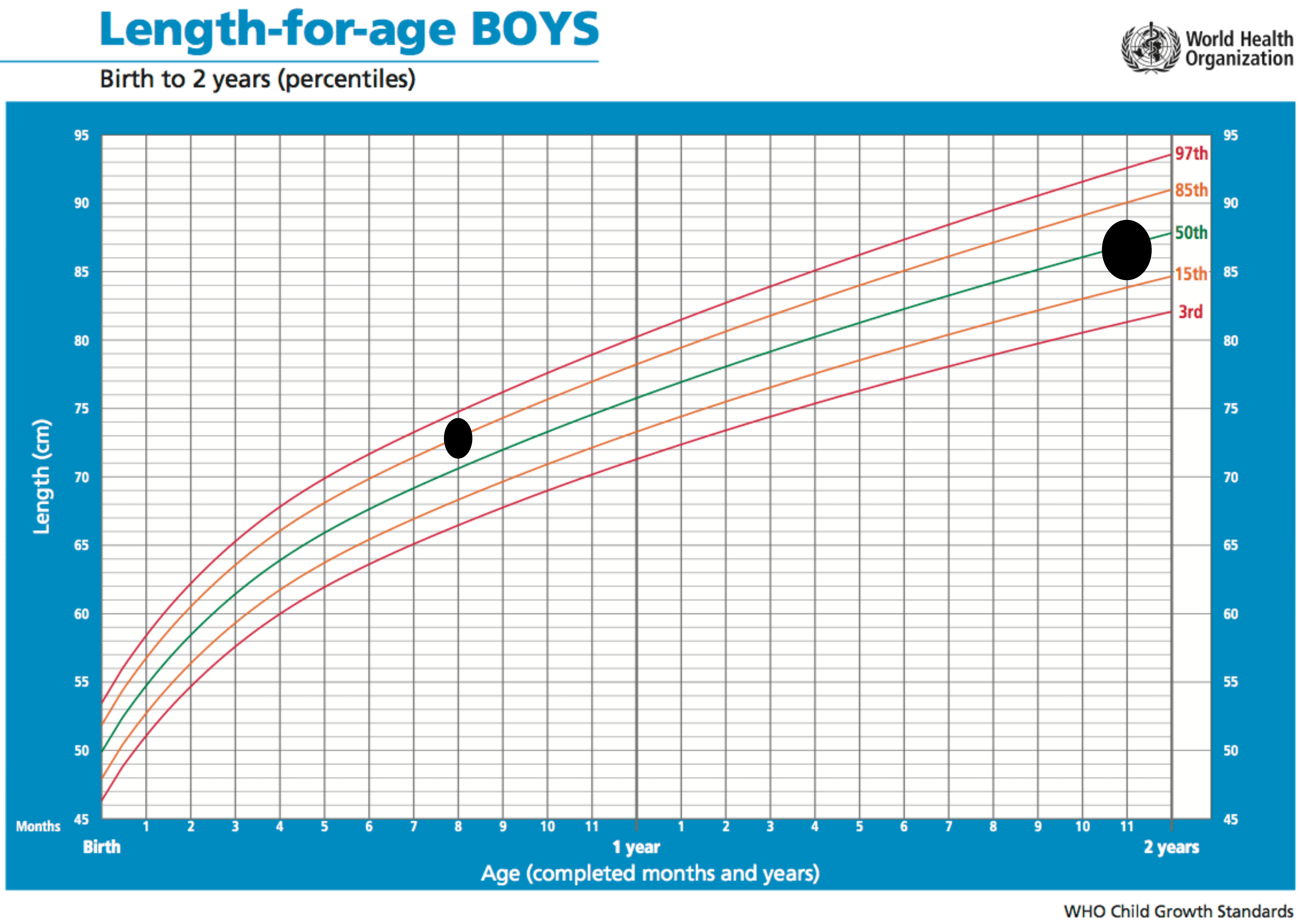
Standardized data



Elder brother is 23 months old and 87 cm tall
Younger brother is 8 month and 74 cm tall

How to compare who is taller between elder and younger brothers?

Standardized data

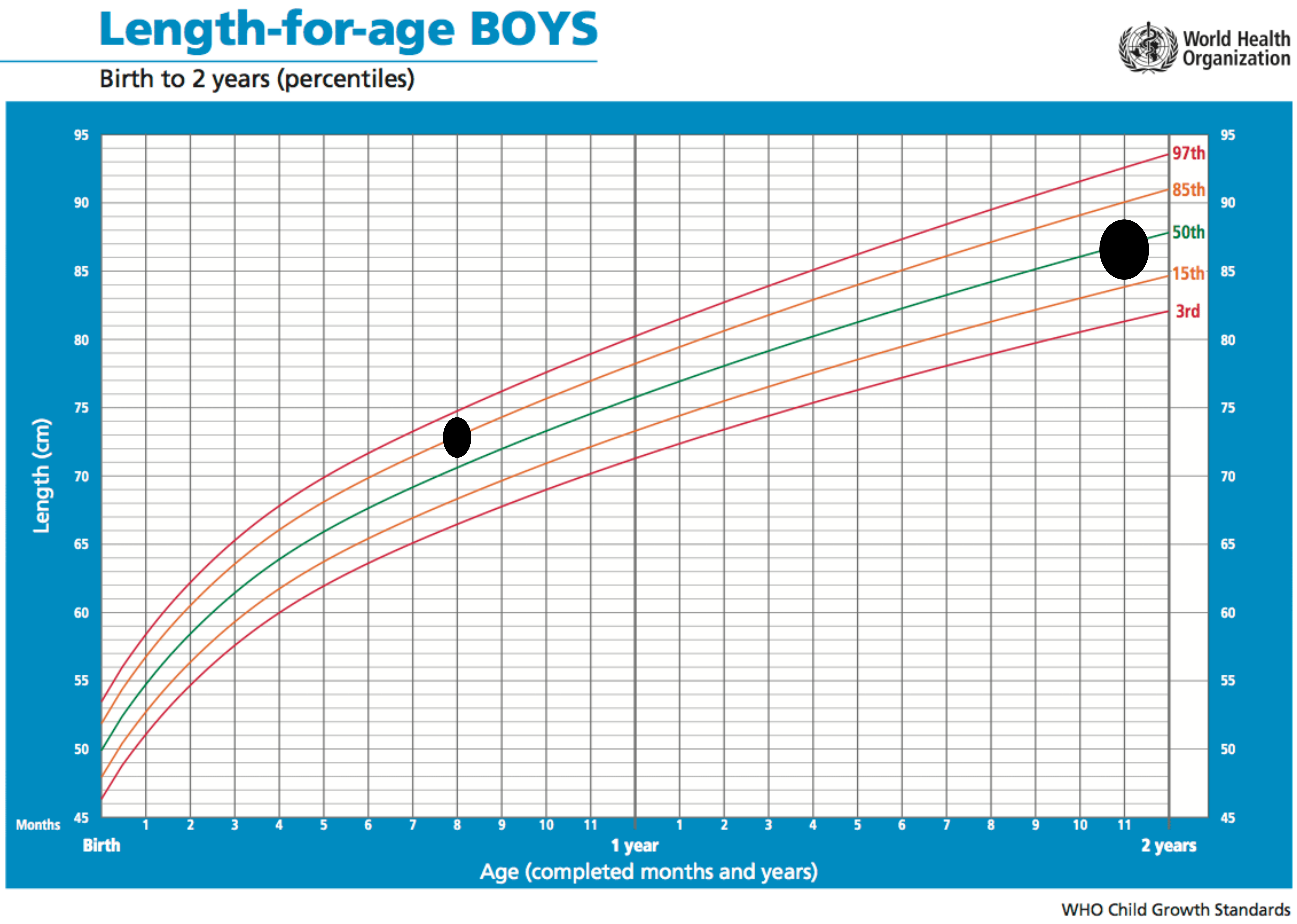


Elder brother is 23 months old and 87 cm tall
Younger brother is 8 month and 74 cm tall

Compare elder brother with his peer, he is at average.

How to compare who is taller between elder and younger brothers?

Standardized data



Elder brother is 23 months old and 87 cm tall
Younger brother is 8 month and 74 cm tall

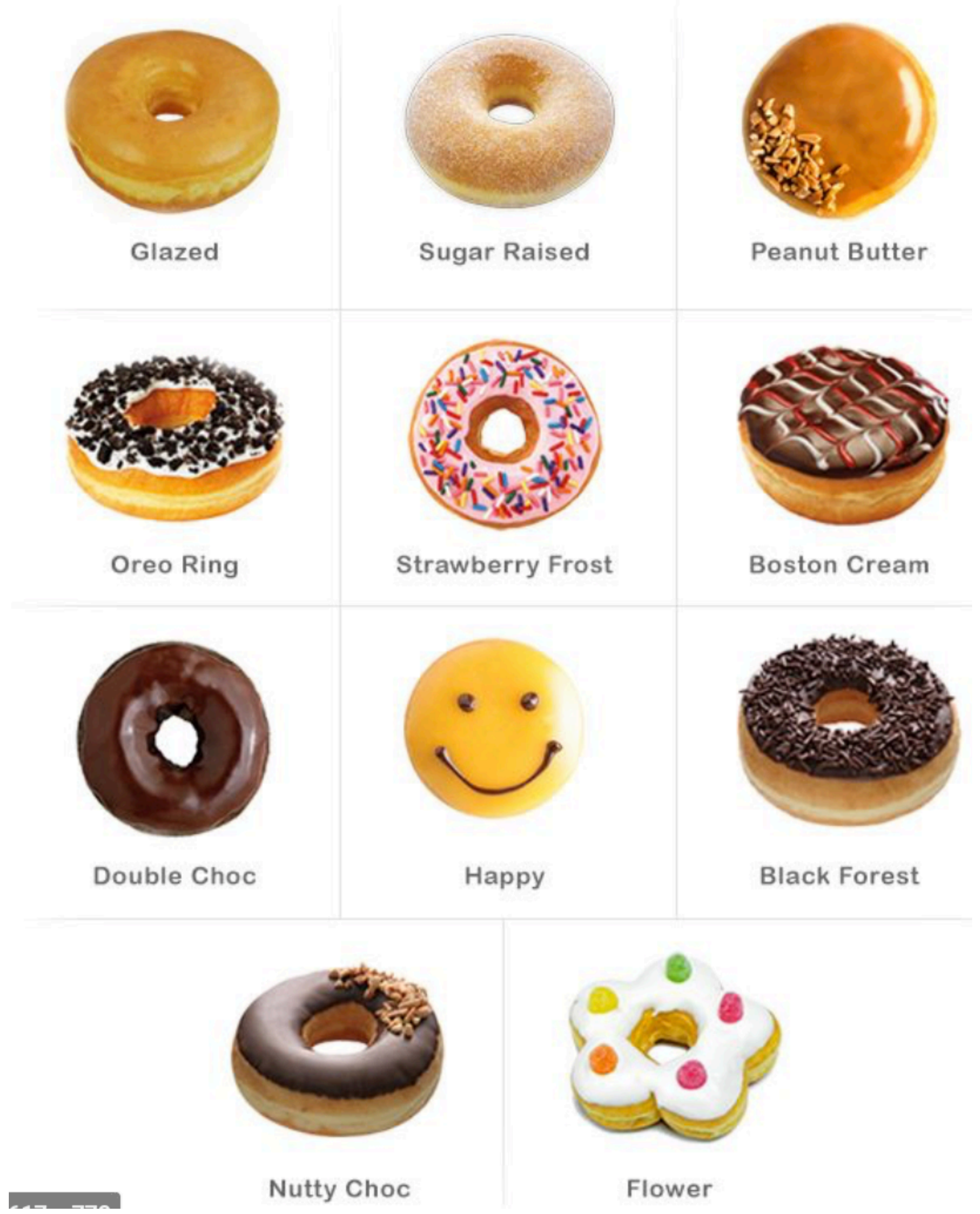
Compare elder brother with his peer, he is at average.

As for younger brother, he is 1 sigma taller than his average peer.

Bootstrapping data set

Please read the textbook from p.187 — p.190

Menu of dunking donut



Customer 1:

Customer 2:

Customer 3:

Customer 4:

Simulation on bootstrapping data sets

Collecting all model parameters

Original data set

Resampling

Data set 1

$f(x) = a^{(1)}x + b^{(1)}$

Data set 2

$f(x) = a^{(2)}x + b^{(2)}$

....

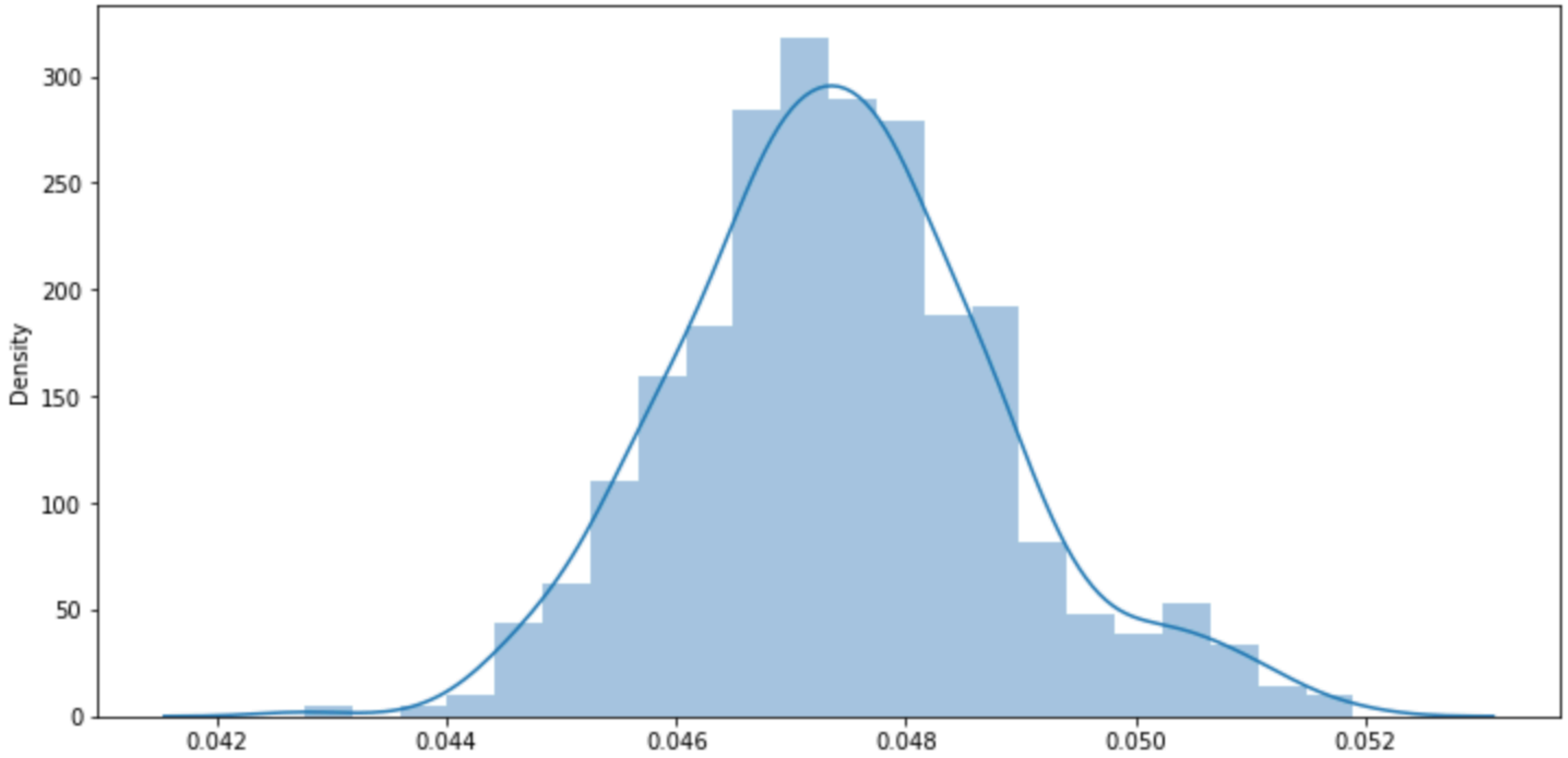
Data set (N-1)

$f(x) = a^{(N-1)}x + b^{(N-1)}$

Data set N

$f(x) = a^{(N)}x + b^{(N)}$

Question: where is the zero slope in tv versus sales?

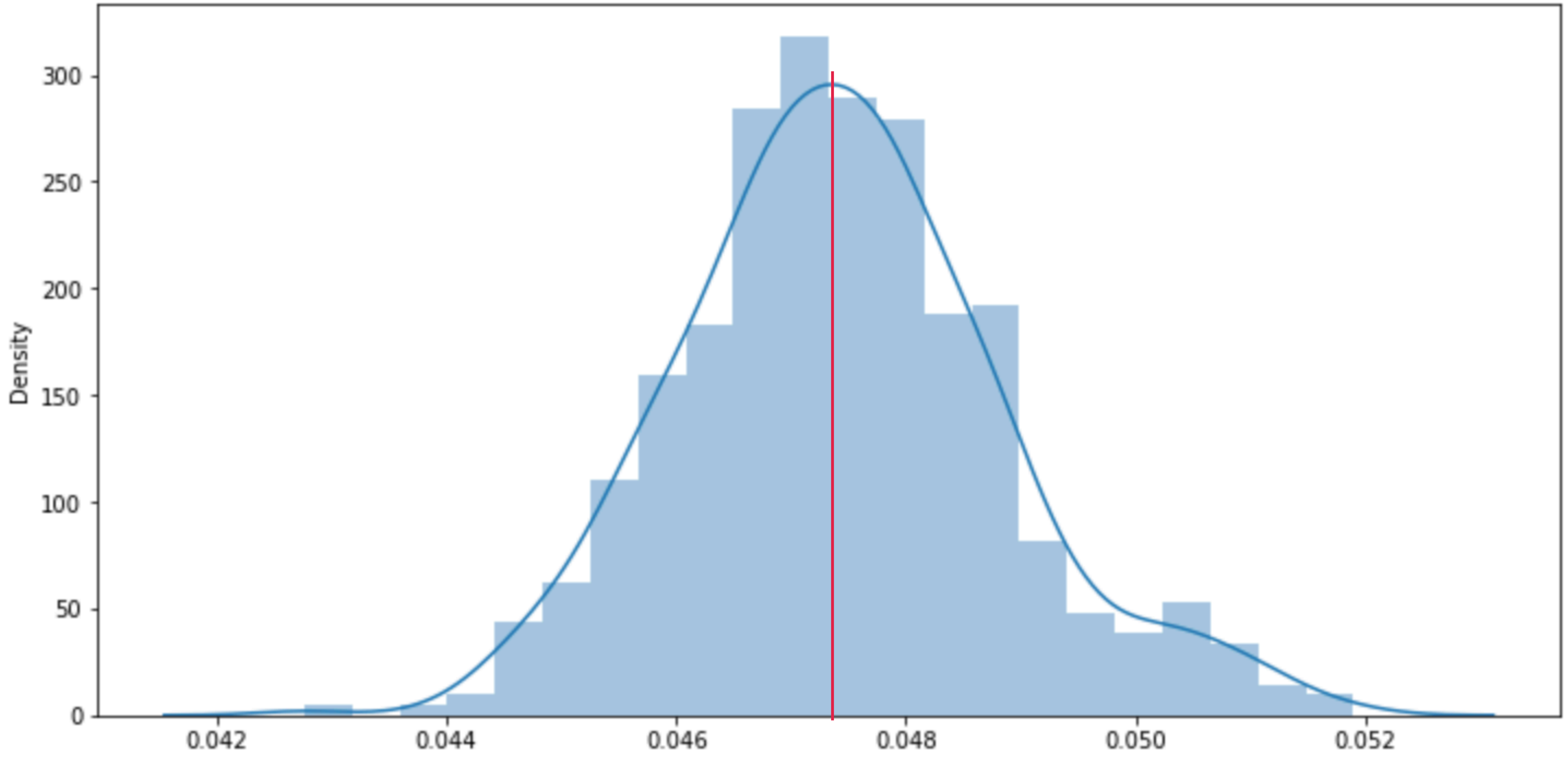


0.00

```
print(np.mean(m_list), np.std(m_list))
```

0.0474348697630429 0.0014079599492549707

Question: where is the zero slope in tv versus sales?

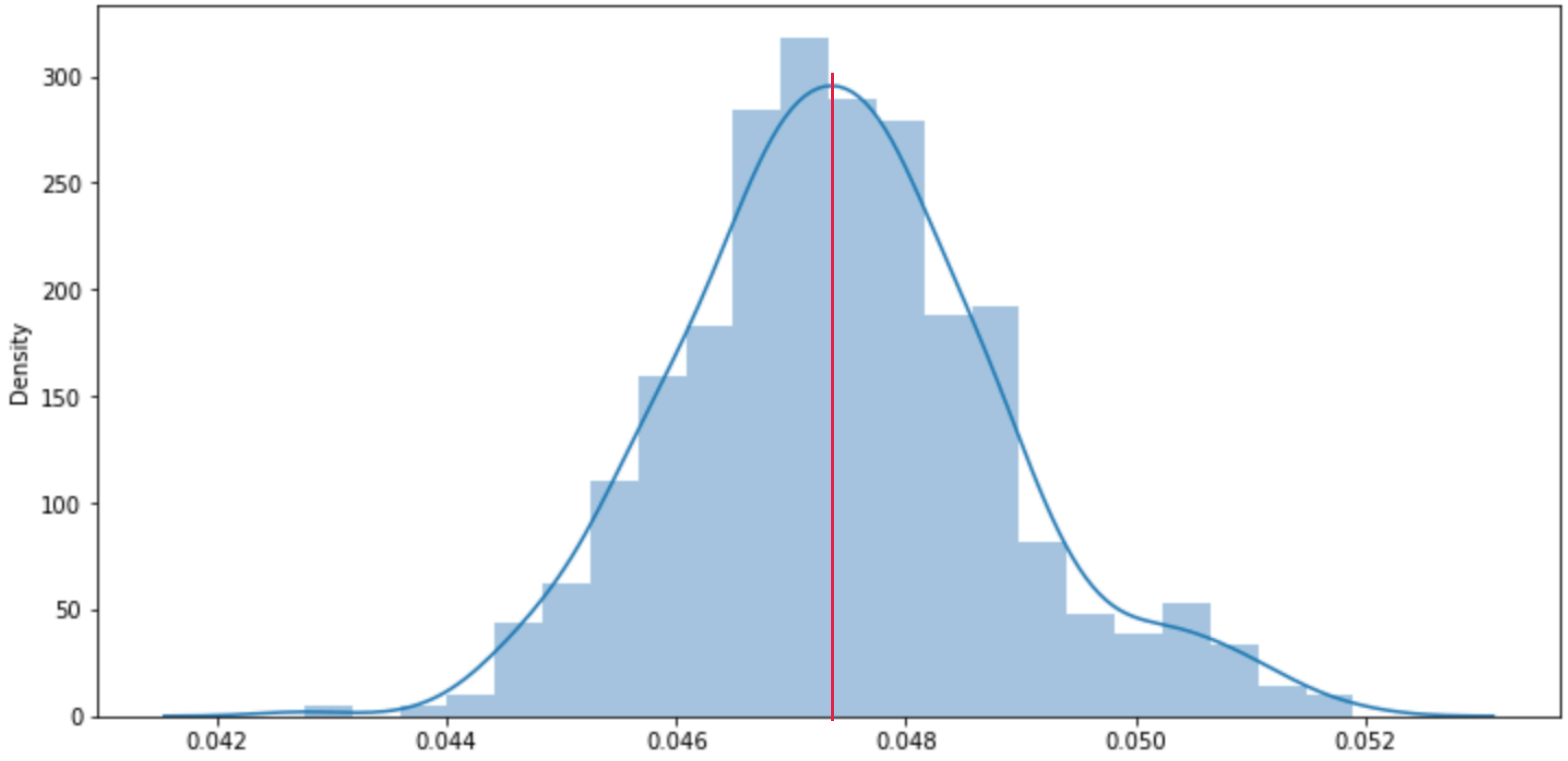


0.00

```
print(np.mean(m_list), np.std(m_list))
```

0.0474348697630429 0.0014079599492549707

Question: where is the zero slope in tv versus sales?

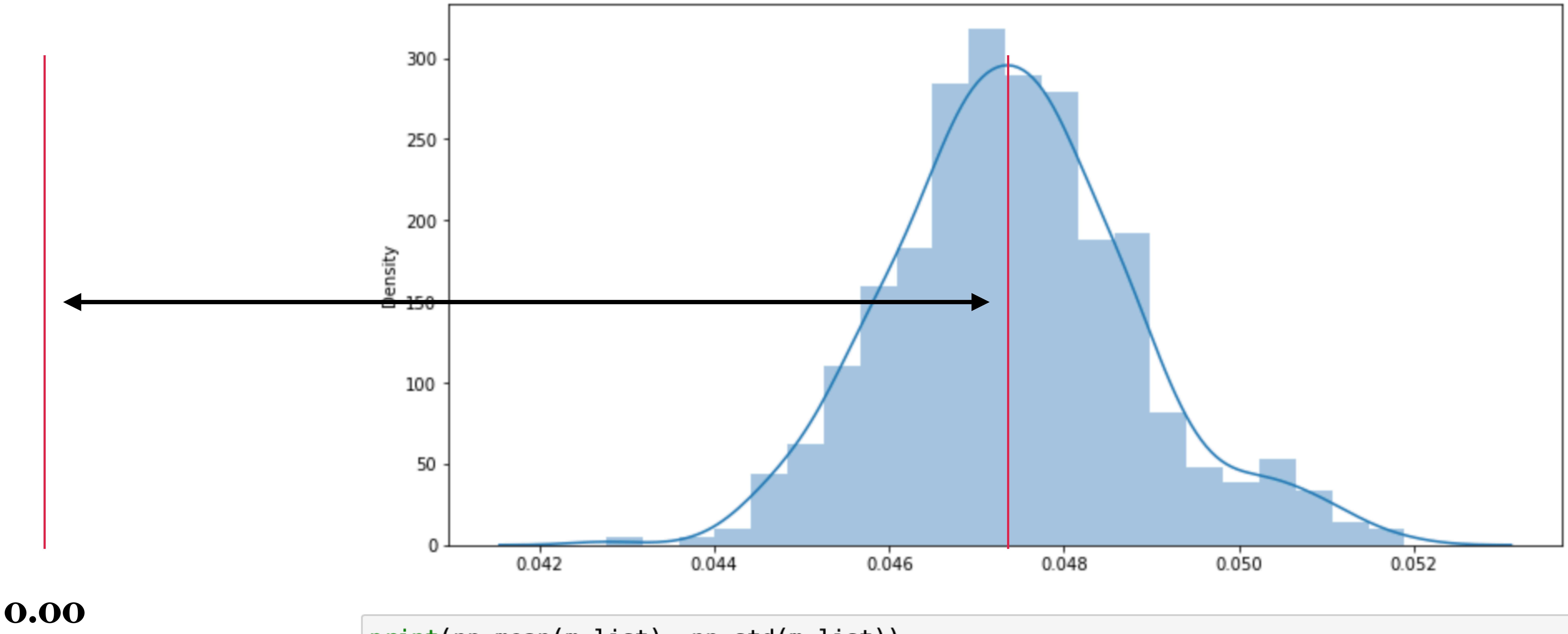


0.00

```
print(np.mean(m_list), np.std(m_list))
```

0.0474348697630429 0.0014079599492549707

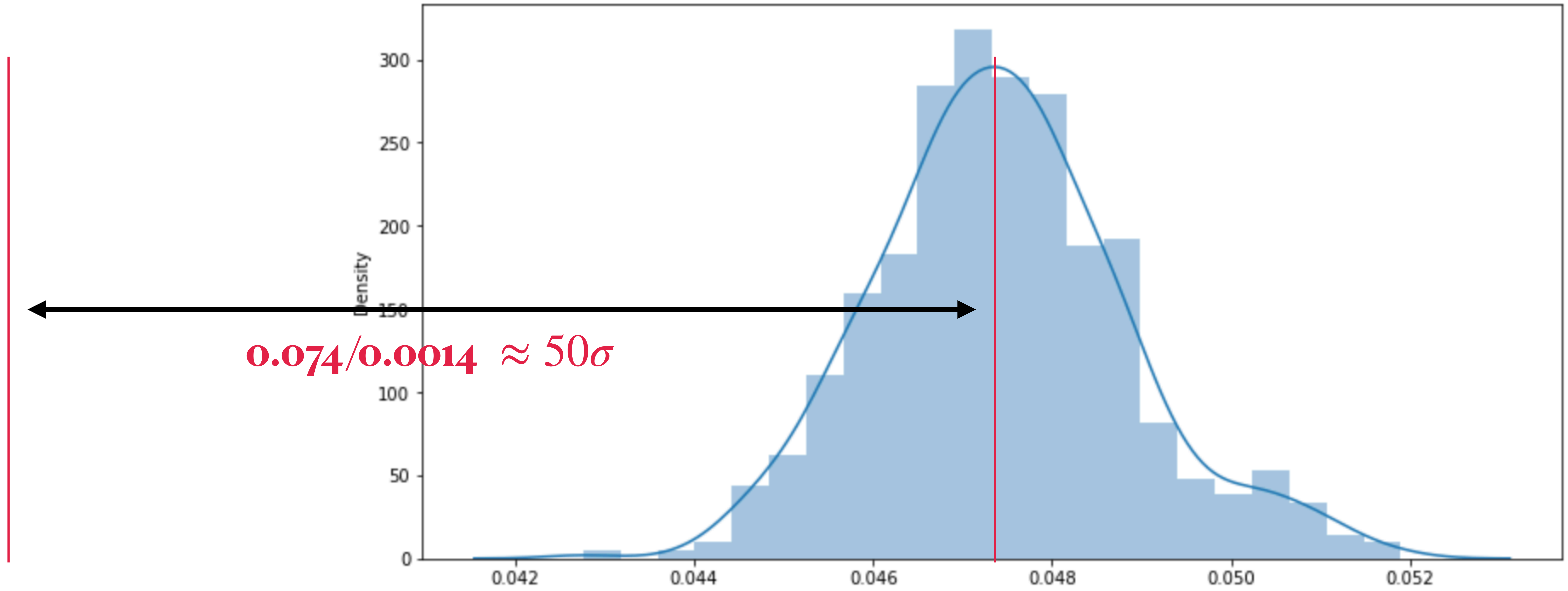
Question: where is the zero slope in tv versus sales?




```
print(np.mean(m_list), np.std(m_list))
```

0.0474348697630429 0.0014079599492549707

Question: where is the zero slope in tv versus sales?



$0.074/0.0014 \approx 50\sigma$

0.00

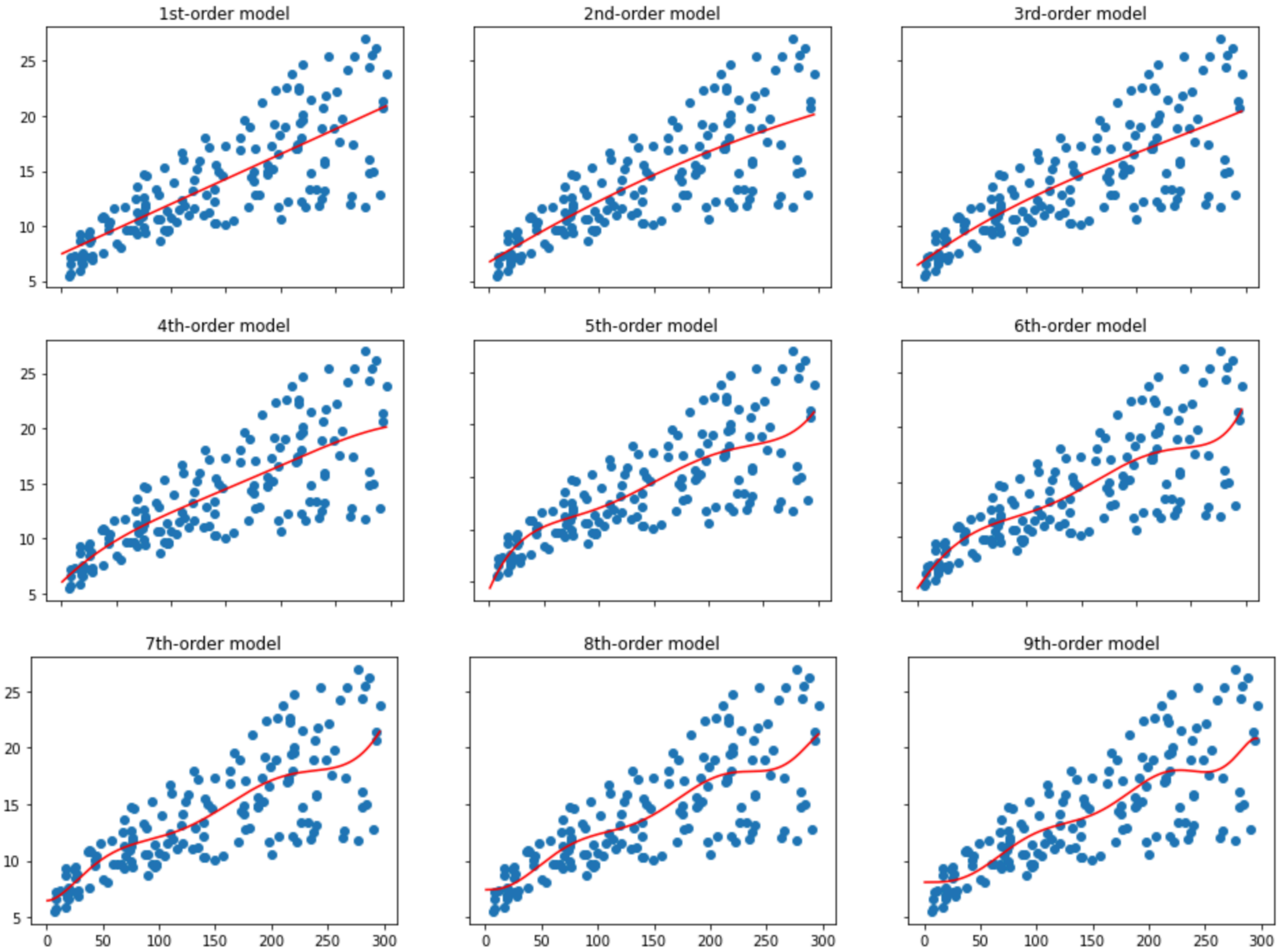
```
print(np.mean(m_list), np.std(m_list))
```

0.0474348697630429 0.0014079599492549707

You can try the same simulation to (radio,sales) and (newspaper, sales)

It would be the next homework

Beyond linear regression model



How to choose which one would work best?