The Basic Practice of Statistics
Ninth Edition

David S. Moore   William I. Notz

Chapter 5
Regression

Lecture Slides

In Chapter 5, we cover …

- Regression lines

- The least-squares regression line

- Example in Python

- Facts about least-squares regression

- Residuals

- Outliers and influential observations

- Cautions about correlation and regression

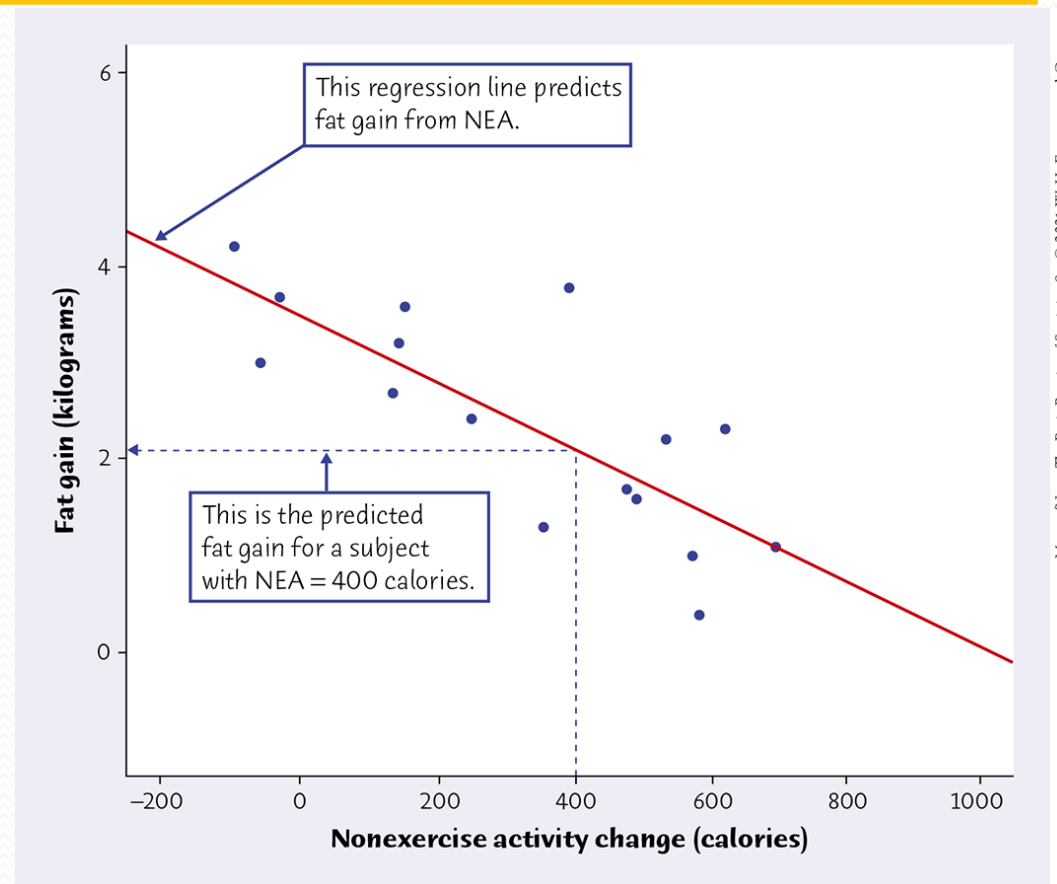- Association does not imply causation

A regression line is a straight line that describes how a response variable y changes as an explanatory variable x changes. We often use a regression line to predict the value of y for a given value of x, when we believe the relationship between y and x is linear.

Example: Predict the gain in fat (in kilograms) based on the change in energy use (in calories) from nonexercise activity.

❑ **If the NEA change is 400 calories, what is the expected fat gain?**

NEA: any other activity we perform than deliberate exercise (e.g., walking to work, going upstairs, typing, washing the dishes, etc.)

This regression line predicts fat gain from NEA.

This is the predicted fat gain for a subject with NEA = 400 calories.

Fat gain (kilograms)

Nonexercise activity change (calories)

REVIEW OF STRAIGHT LINES

- Suppose that y is a response variable (plotted on the vertical axis) and x is an explanatory variable (plotted on the horizontal axis). A straight line relating y to x has an equation of the form

$$y = a + bx$$

- In this equation, b is the **slope**—the amount by which y changes when x increases by one unit. The number a is the **intercept**—the value of y when $x = 0$.

REVIEW OF STRAIGHT LINES

- If you know two points on a line, (x_1, y_1) and (x_2, y_2), the line slope is given by:

$$b = \frac{y_2 - y_1}{x_2 - x_1}$$

- The intercept can be obtained using one of the points and the straight-line equation:

$$a = y_1 - bx_1$$

or

$$a = y_2 - bx_2$$

REVIEW OF STRAIGHT LINES

- To **plot the line** on the scatterplot, use the equation to find the predicted y for two values of x, one near each end of the range of x in the data. Plot each y above its x-value and draw the line through the two points.

The least-squares regression line (1 of 2)

The least-squares regression line of y on x is the line that makes the sum of the squares of the vertical distances of the data points from the line as small as possible.
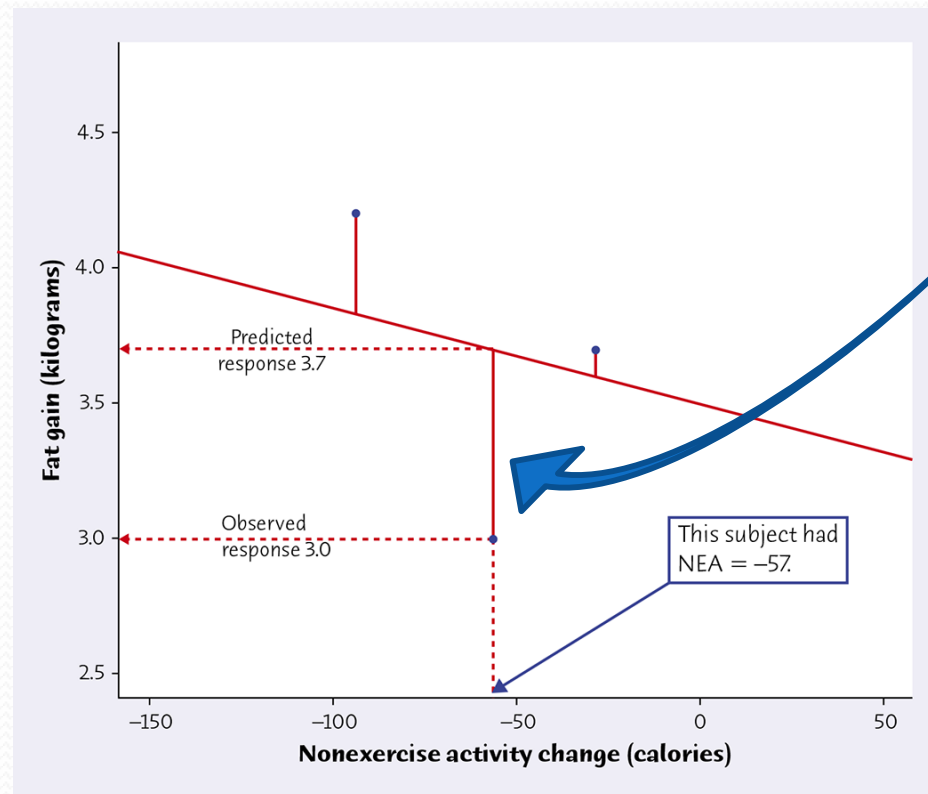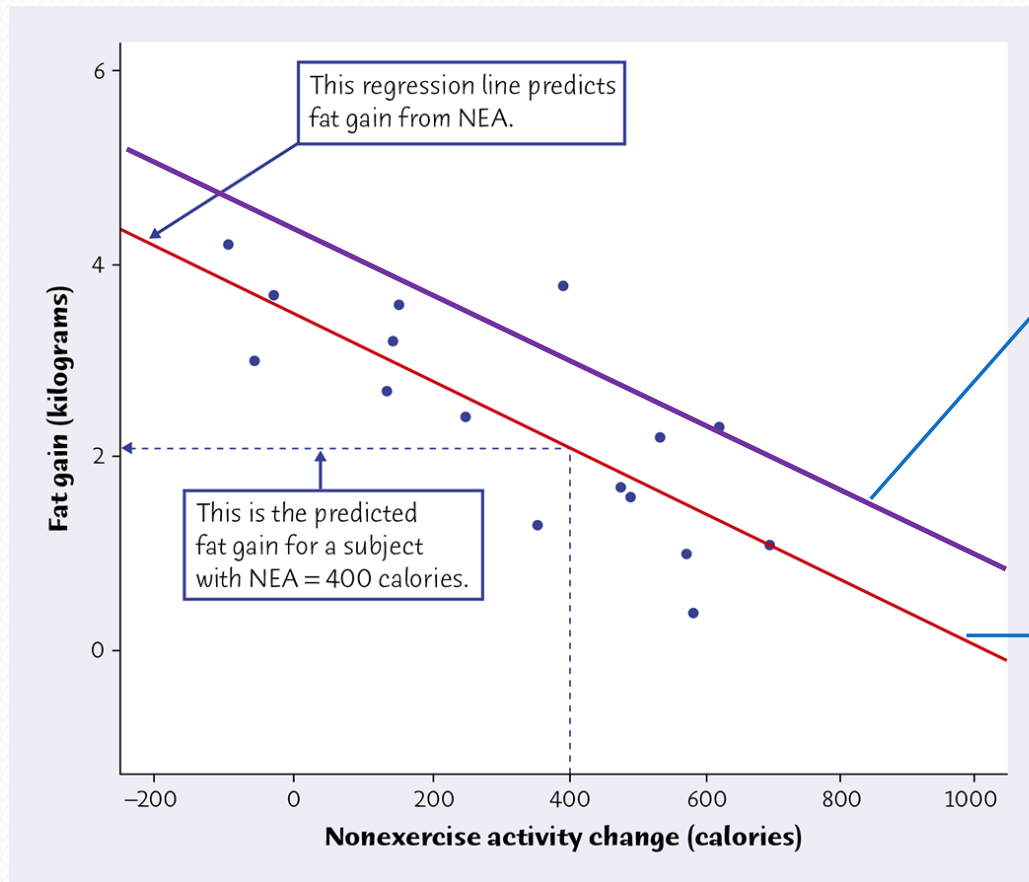
If we were to fit the purple line to the data, it would give larger squared distances to the data points than the red line. The purple line cannot be the regression line .

The red line is the regression line because it gives the smallest squared distances to the data points.

Within the figure:

6 — This regression line predicts fat gain from NEA.

This is the predicted fat gain for a subject with NEA = 400 calories.

Fat gain (kilograms)

Nonexercise activity change (calories)

−200   0   200   400   600   800   1000

EQUATION OF THE LEAST-SQUARES REGRESSION LINE

- We have data on an explanatory variable x and a response variable y for n individuals. From the data, calculate the means $\bar{x}$ and $\bar{y}$ and the standard deviations s_x and s_y of the two variables, as well as their correlation r. The least-squares regression line is the line

$$\hat{y} = a + bx$$

- with slope

$$b = r\frac{s_y}{s_x}$$

- and intercept

$$a = \bar{y} - b\bar{x}$$

Prediction via regression line

- For the nonexercise activity example, the least-squares regression line is
$$\hat{y} = 3.505 - 0.0034x$$

- $\hat{y}$ is the predicted fat gain (in kg) with x calories of nonexercise activity.

- Suppose we know someone has an increase of 400 calories of NEA. What would we predict for fat gain?

- For someone with 400 calories of NEA, we would predict fat gain of

$$3.505 - 0.00344(400) = 2.13 \text{ kg}$$
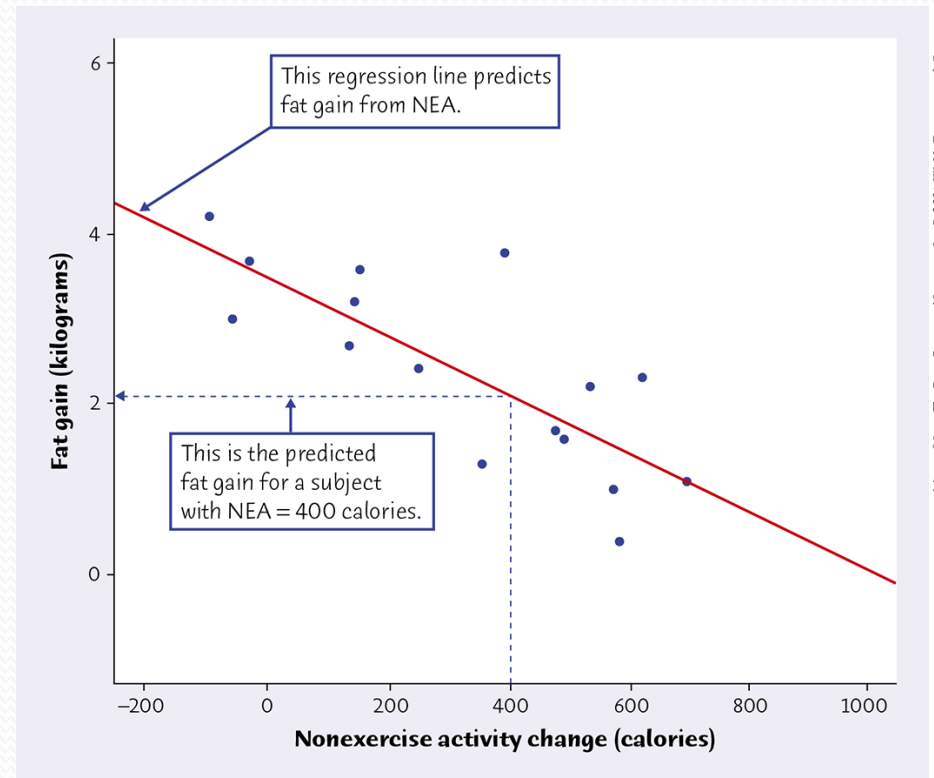
Example using Python

Example 5.1 of the textbook
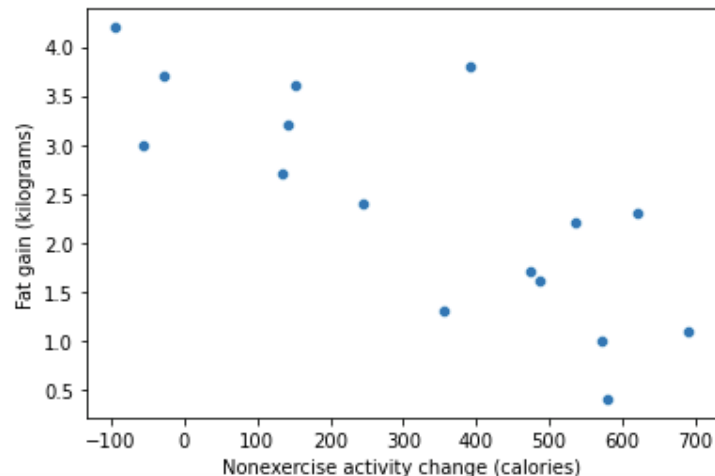
```
In [1]: import numpy as np
        import pandas as pd
        import seaborn as sns
        import matplotlib.pyplot as plt
        from scipy import stats
```

```
In [2]: fatgain = pd.read_csv("eg05-01fatgain.csv")
        fatgain.head()
```

Out[2]:

|	NEA	Fat
0	-94	4.2
1	-57	3.0
2	-29	3.7
3	135	2.7
4	143	3.2

```
In [3]: sns.scatterplot(x = "NEA", y = "Fat", data = fatgain)
        plt.xlabel("Nonexercise activity change (calories)")
        plt.ylabel("Fat gain (kilograms)")
        plt.show()
```


Example: Predict the gain in fat (in kilograms) based on the change in energy use (in calories) from nonexercise activity.

Example using Python

```
In [4]:  # sns.lmplot adds to a scatterplot the regression line that best fits the data points
         sns.lmplot(x = "NEA", y = "Fat", data = fatgain, ci=None)
         plt.xlabel("Nonexercise activity change (calories)")
         plt.ylabel("Fat gain (kilograms)")
         plt.show()
```
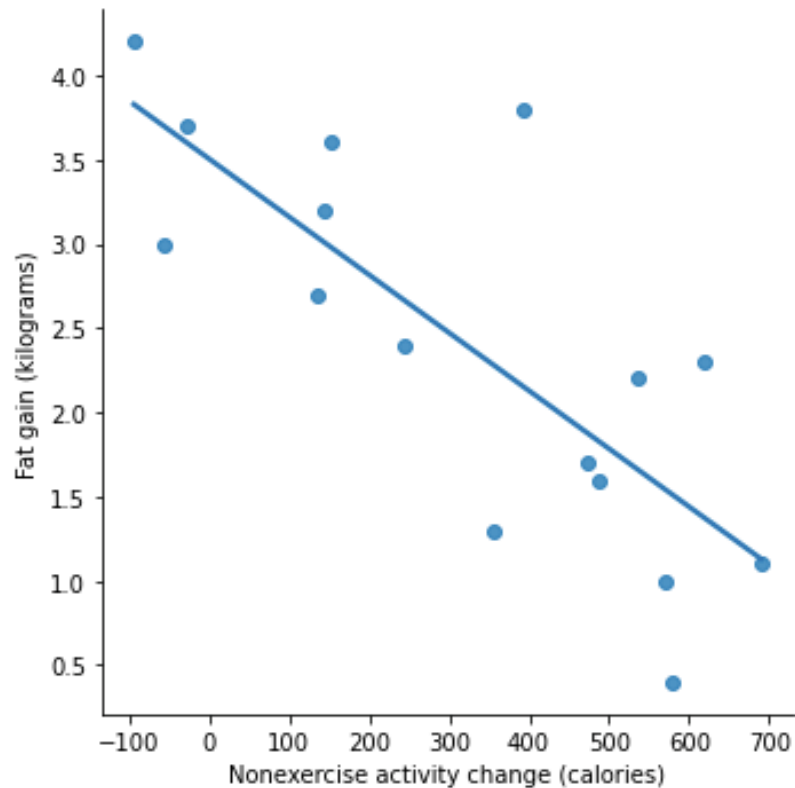

Example: Predict the gain in fat (in kilograms) based on the change in energy use (in calories) from nonexercise activity.

Example using Python

```
In [5]: lm = stats.linregress(x=fatgain['NEA'],y=fatgain['Fat'])
        lm.intercept # intercept
Out[5]: 3.5051229156310724

In [6]: lm.slope # slope
Out[6]: -0.0034414870381249342

In [7]: lm.rvalue # correlation
Out[7]: -0.7785558457058473

In [8]: lm.rvalue**2 # r^2
Out[8]: 0.6061492048827472
```

Example: Predict the gain in fat (in kilograms) based on the change in energy use (in calories) from nonexercise activity.

$$\hat{y} = 3.505 - 0.0034x$$

- A change of 1 calorie in NEA results in a decrease of 0.0034 kg (3.4 g) in fat gain
- Or a change of 100 calories in NEA results in a decrease of 0.340 kg (340 g) in fat gain

Example using Python

```
In [5]: lm = stats.linregress(x=fatgain['NEA'],y=fatgain['Fat'])
        lm.intercept # intercept

Out[5]: 3.5051229156310724

In [6]: lm.slope # slope

Out[6]: -0.0034414870381249342

In [7]: lm.rvalue # correlation

Out[7]: -0.7785558457058473

In [8]: lm.rvalue**2 # r^2

Out[8]: 0.6061492048827472
```
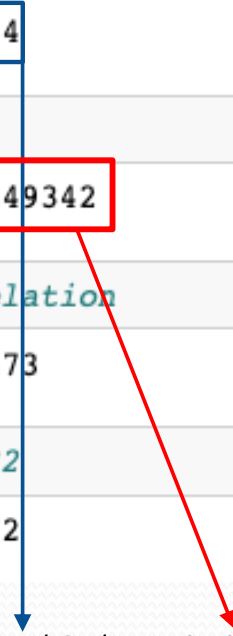
$$\hat{y} = 3.505 - 0.0034x$$

Example: Predict the gain in fat (in kilograms) based on the change in energy use (in calories) from nonexercise activity.

❑ **If the NEA change is 400 calories, what is the expected fat gain?**

For someone with 400 calories of NEA, what would be the predicted fat gain?

```
In [9]: # what would be the fat gain for someone with 400 calories in NEA change?
        yhat = lm.intercept + lm.slope*400
        yhat

Out[9]: 2.1285281003810987   → 2.13 kg
```

Facts about least-squares regression

- The distinction between explanatory variables and response variables is essential in regression.

- There is a close connection between correlation and the slope of the least-squares line. The slope is

$$b = r\frac{s_y}{s_x}$$

- The slope b and correlation r always have the same sign.

- Along the regression line, a change of 1 standard deviation in x corresponds to a change of r standard deviations in y.

- The least-squares regression line always passes through

- $(\bar{x}, \bar{y})$ on the graph of y against x.

- The correlation r describes the strength of a straight-line relationship.

The square of the correlation

- The square of the correlation, r^2, is the fraction of the variation in the values of y that is explained by the least-squares regression of y on x.

$$r^2 = \frac{\text{variation in } \hat{y} \text{ along the regression line as } x \text{ varies}}{\text{total variation in observed values of } y}$$

- *Caution:* You can find a regression line for any relationship between two quantitative variables, but the usefulness of the line for prediction depends on the strength of the linear relationship.

- r^2 near 0 means a *weak* linear relationship.

- r^2 near 1 means a *strong* linear relationship.

- Note that with r^2, we lose information about the *direction* of the association.

```
In [5]: lm = stats.linregress(x=fatgain['NEA'],y=fatgain['Fat'])
        lm.intercept # intercept

Out[5]: 3.5051229156310724


In [6]: lm.slope # slope

Out[6]: -0.0034414870381249342


In [7]: lm.rvalue # correlation

Out[7]: -0.7785558457058473


In [8]: lm.rvalue**2 # r^2

Out[8]: 0.6061492048827472
```

Example: Predict the gain in fat (in kilograms) based on the change in energy use (in calories) from nonexercise activity.

Residuals (1 of 2)

A residual is the difference between an observed value of the response variable and the value predicted by the regression line. That is, a residual is the prediction error that remains after we have chosen the regression line:

$$\text{residual} \quad = \quad \text{observed } y - \text{predicted } y$$
$$= \quad y - \hat{y}$$

Residuals represent "leftover" variation in the response after fitting the regression line.
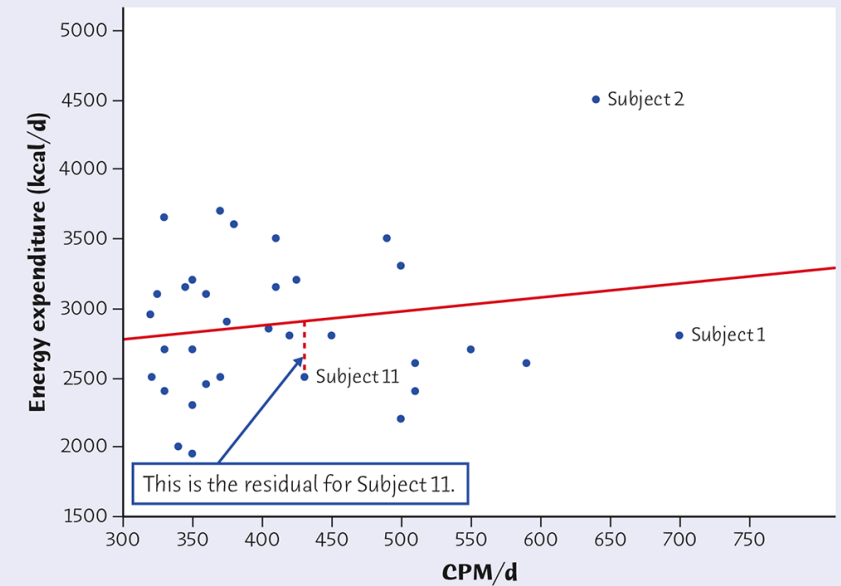
Example 5.5: researchers measured physical activity of 34 subjects, in mean counts per minute per day (CPM/d), using wearable accelerometers (fitbits). They also measured the total energy expenditure in kilocalories per day (kcal/d) for each subject.

Residuals (2 of 2)

The residuals from the least-squares line have a special property: **the mean of the least-squares residuals is always zero**.

RESIDUAL PLOTS

- A residual plot is a scatterplot of the regression residuals against the explanatory variable.

- Residual plots help us assess how well a regression line fits the data.

- Look for a "random" scatter around zero.
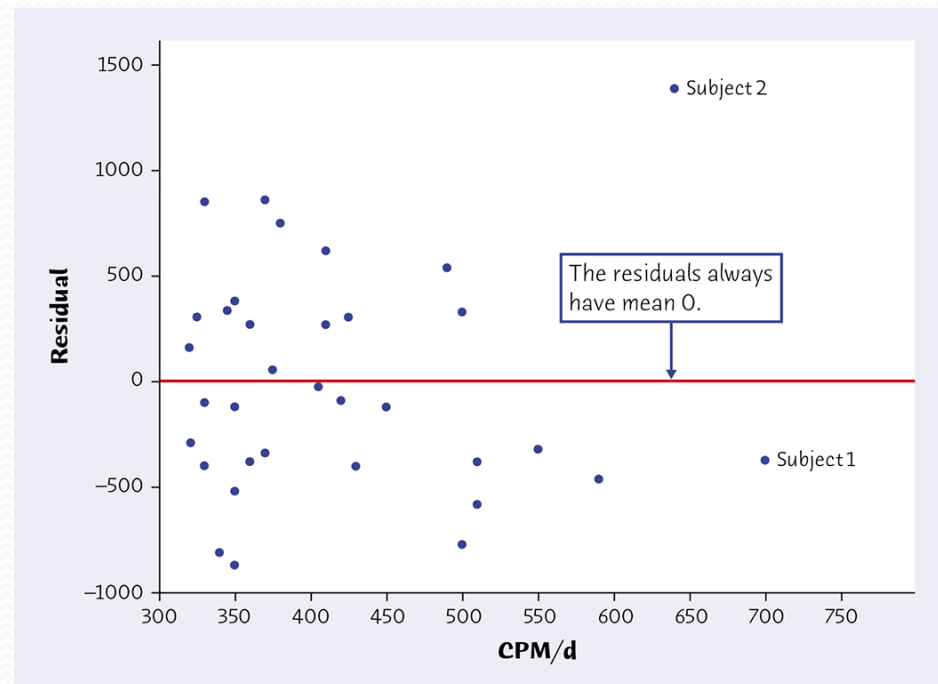
Example 5.5 in Python

```
In [10]: exercise = pd.read_csv("eg05-05exerc.csv")
         exercise.head()
```

Out[10]:

	Subject	CPM/d	EnergyExpenditure
0	1	700	2800
1	2	640	4500
2	3	590	2600
3	4	550	2700
4	5	510	2400

Example 5.5: researchers measured physical activity of 34 subjects, in mean counts per minute per day (CPM/d), using wearable accelerometers (fitbits). They also measured the total energy expenditure in kilocalories per day (kcal/d) for each subject.

```
In [11]: sns.lmplot(x = "CPM/d", y = "EnergyExpenditure", data = exercise,ci=None)
         plt.xlabel("CPM/d")
         plt.ylabel("Energy expenditure (kcal/d)")
         plt.show()
```
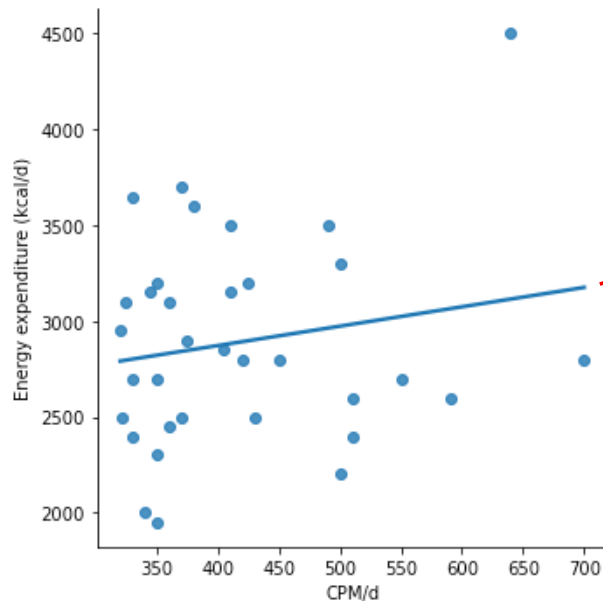
Regression line:

energy expenditure = 2467.64 + 1.01(CPM/d)


```
In [12]: lm2 = stats.linregress(x=exercise['CPM/d'],y=exercise['EnergyExpenditure'])
         print(lm2.intercept) # intercept
         print(lm2.slope) #slope
```

```
2467.549357864118
1.011082124022258
```

Example 5.5 in Python

```
In [13]: #calculating the residuals
         x = exercise['CPM/d']
         y_hat = lm2.intercept + lm2.slope*x ## predicted energy expenditures
         y = exercise['EnergyExpenditure'] ## observed energy expenditures
         residuals = y - y_hat
```

```
In [14]: sns.scatterplot(x = exercise['CPM/d'], y = residuals)
         plt.xlabel('CPM/d')
         plt.ylabel("Residuals")
         plt.axhline(y=0,color="black",linestyle="--")
         plt.show()
```
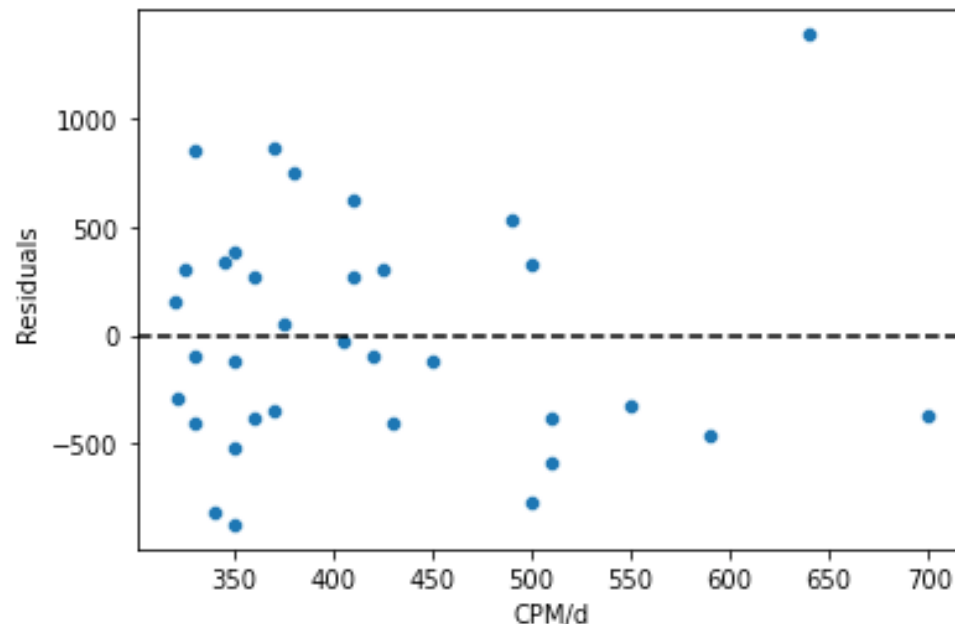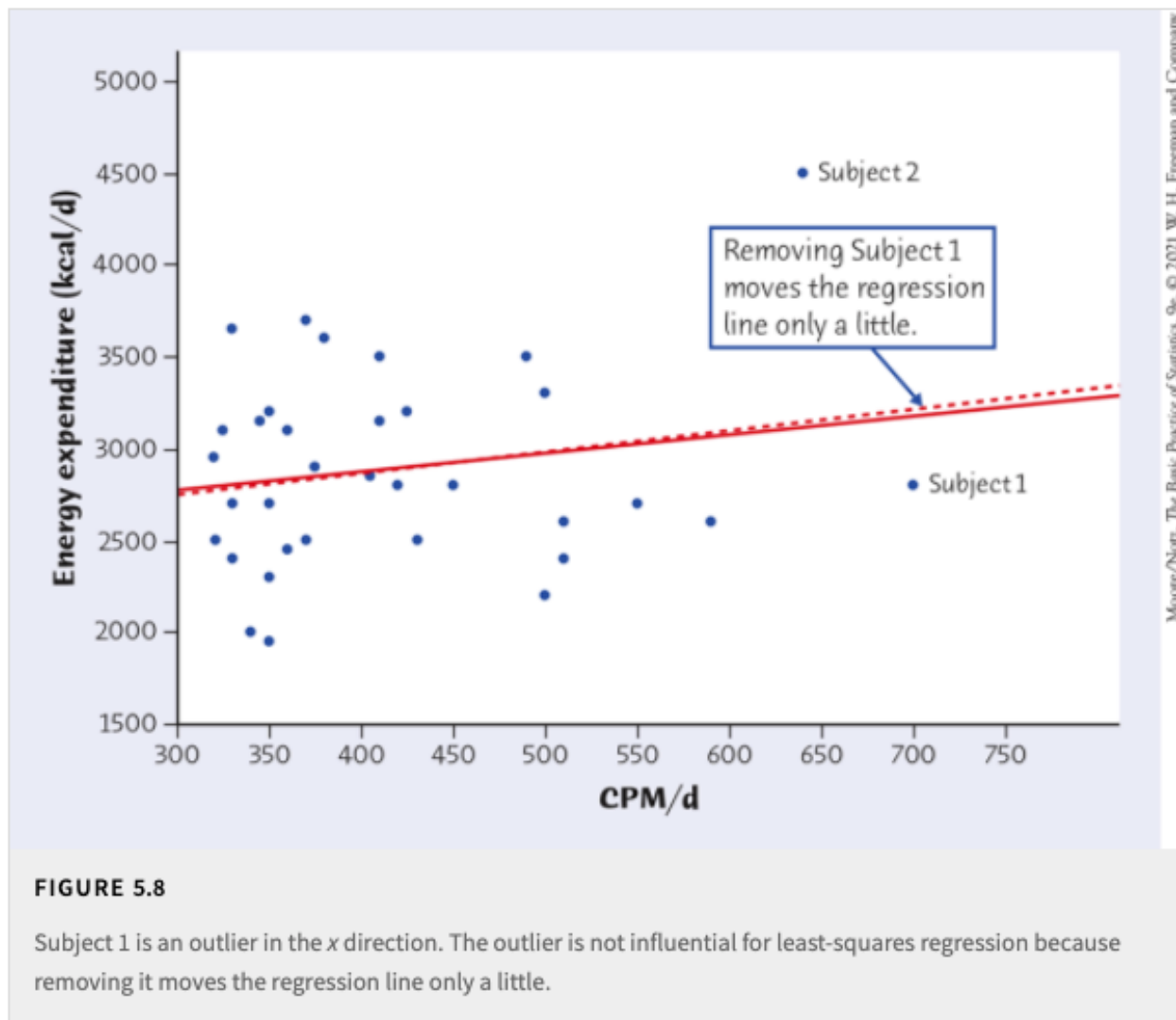

Residual Plot

Influential observations

- An observation is influential for a statistical calculation if removing it would markedly change the result of the calculation.

- The result of a statistical calculation may be of little practical use if it depends strongly on a few influential observations.

- Points that are outliers, in either the x or the y direction of a scatterplot, are often influential for the correlation.

- In the regression setting, however, not all outliers are influential.

- The outlier has influence on the regression line is that it lies well above the regression line calculated from the other observations.
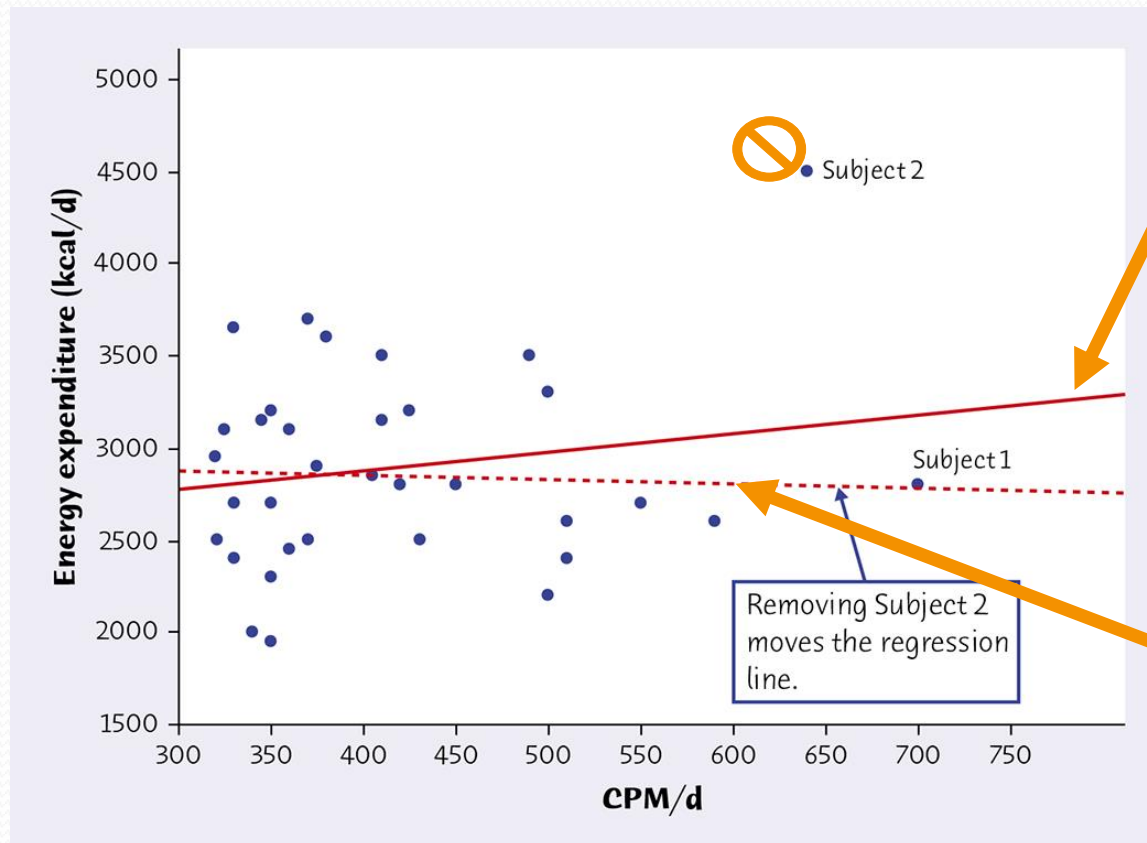
Outliers and influential points

PHYSICAL ACTIVITY AND WEIGHT LOSS

FIGURE 5.8

Subject 1 is an outlier in the *x* direction. The outlier is not influential for least-squares regression because removing it moves the regression line only a little.

Outliers and influential points

PHYSICAL ACTIVITY AND WEIGHT LOSS

From all of the data
$$r = 0.181$$

Without Subject 2
$$r = -0.043$$

Cautions about correlation and regression

- Correlation and regression lines describe only *linear* relationships.

- Correlation and least-squares regression lines are not *resistant*.

- Beware *ecological correlation*, or correlation based on *averages* rather than individuals.

- Beware of *extrapolation*—predicting outside of the range of x.

- Beware of *lurking variables*—these have an important effect on the relationship among the variables in a study but are not included in the study.

- Correlation does not imply causation!

Example of *ecological* correlation

- There is a large positive correlation between *average* income and number of years of education.
- The correlation is smaller if we compare the incomes of *individuals* with number of years of education.
- The correlation based on average income ignores the large variation in the incomes of individuals having the same amount of education.
- The variation from individual to individual increases the scatter in a scatterplot, reducing the correlation.
- The correlation between average income and education overstates the strength of the relationship between the incomes of individuals and number of years of education.

⚠️ *Correlations based on averages can be misleading if they are interpreted to be about individuals.*

# Caution: Beware of extrapolation (1 of 2)

Sarah's height was plotted against her age.

- Can you predict her height at age 42 months?

- Can you predict her height at age 30 years (360 months)?
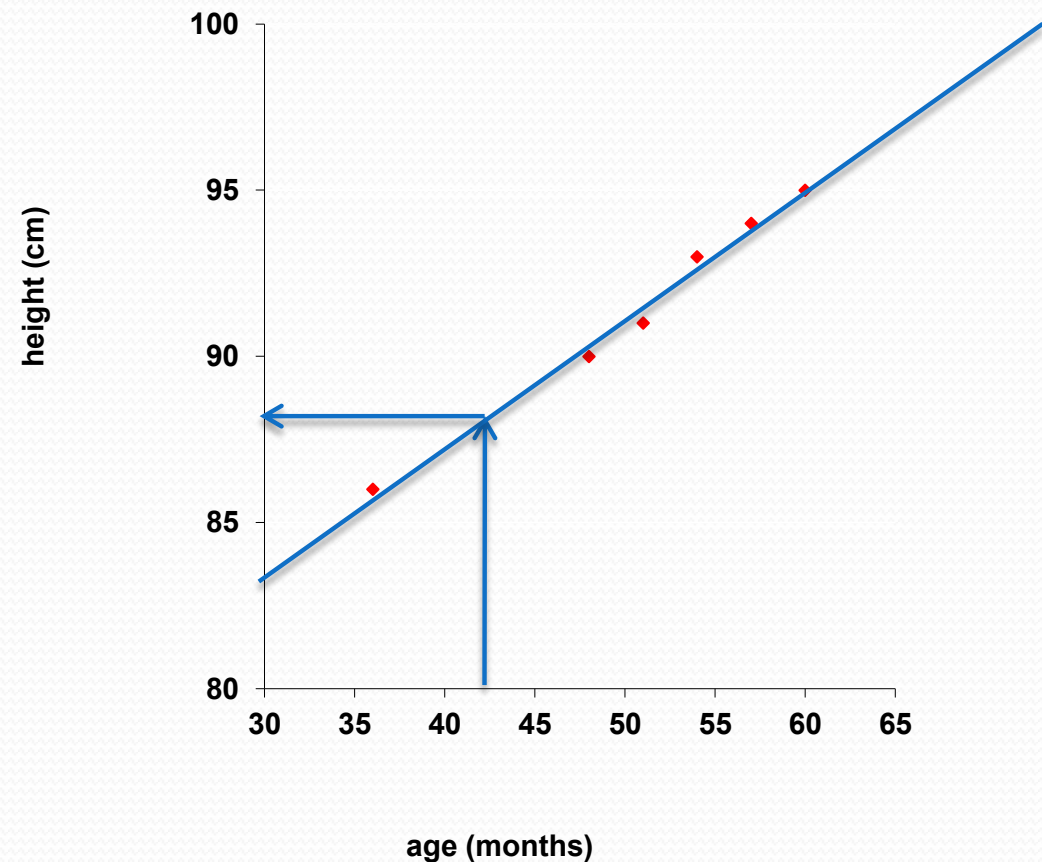
Caution: Beware of extrapolation (2 of 2)

- Regression line:
$$\hat{y} = 71.95 + 0.383x$$

- Predicted height at age 42 months?
$$\hat{y} = 98$$

- Predicted height at age 30 years?
$$\hat{y} = 209.8$$

- She is predicted to be 6 feet 10½ inches tall at age 30!
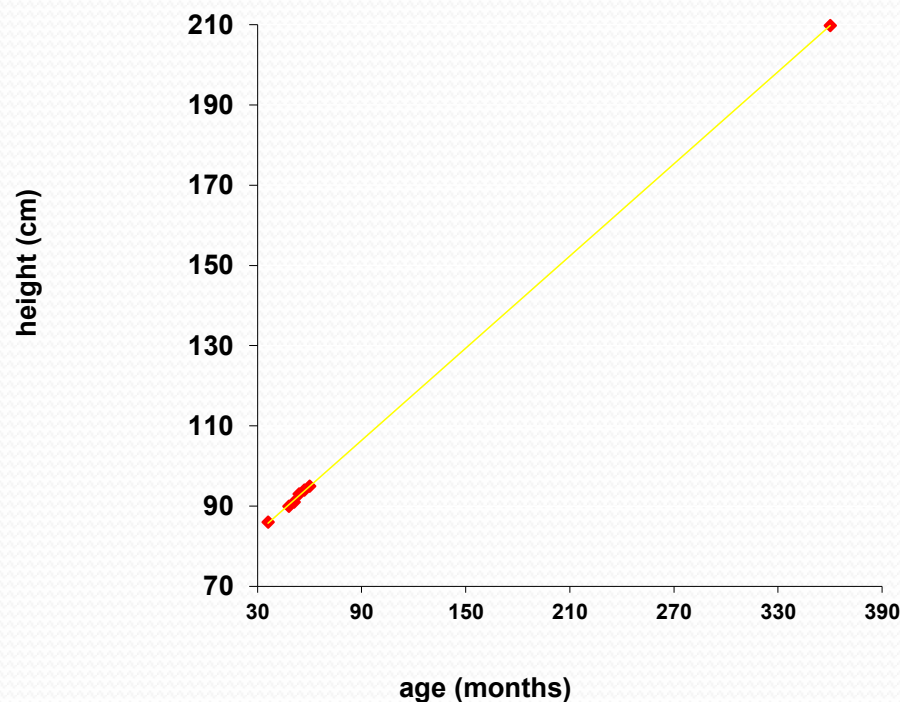
Caution: Beware of lurking variables

- Explanatory variable: Observed meditation practice (yes/no)
- Response variable: Level of age-related enzyme

Example: Meditation and Aging
 (*Noetic Sciences Review*, Summer 1993, p. 28)

General concern for one's well-being may also be affecting the response (and the decision to try meditation).

⚠️ *The relationship between two variables can often be understood only by taking other variables into account.*

Correlation does not imply causation

Weekly tissue sales and weekly hot chocolate sales for a city shows a high positive correlation.

Does that mean that hot chocolate consumption causes tissue consumption? What other factors can explain this relationship?

Correlation does not imply causation

- Even very strong correlations may not correspond to a real causal relationship (changes in x actually causing changes in y).

- Correlation may be explained by a lurking variable.

Social Relationships and Health

House, J., Landis, K., and Umberson, D. "Social Relationships and Health," *Science*, Vol. 241 (1988), pp. 540–545.

Does lack of social relationships cause people to become ill?

(*There was a strong correlation.*)

- ❏ **Or** are unhealthy people less likely than others to establish and maintain social relationships? (*reversed relationship*)

- ❏ **Or** is there some other factor that predisposes people both to have lower social activity and to become ill?

Evidence of causation

A properly conducted **randomized experiment** may establish causation.

Other considerations when we cannot do an experiment:

- The association is *strong*.

- The association is *consistent*.

- *Higher* doses are associated with *stronger* responses.

- Alleged cause *precedes* the effect *in time*.

- Alleged cause is *plausible* (reasonable explanation).

EXAMPLE 5.11 Does Smoking Cause Lung Cancer? (1 of 3)

- Doctors had long observed that most lung cancer patients were smokers.

- Comparison of smokers and "similar" nonsmokers showed a very strong association between smoking and death from lung cancer.

- Could the association be explained by lurking variables? Might there be, for example, a genetic factor that predisposes people both to nicotine addiction and to lung cancer? Smoking and lung cancer would then be positively associated, even if smoking had no direct effect on the lungs. **How were these objections overcome?**

EXAMPLE 5.11 Does Smoking Cause Lung Cancer? (2 of 3)

How is it possible to build a strong case for causation in the absence of experiments?

- *The association is strong*. The association between smoking and lung cancer is very strong.

- *The association is consistent*. Many studies of different kinds of people in many countries link smoking to lung cancer. That reduces the chance that a lurking variable specific to one group or one study explains the association.

- *Higher doses are associated with stronger responses*. People who smoke more cigarettes per day or who smoke over a longer period get lung cancer more often. People who stop smoking reduce their risk.

EXAMPLE 5.11 Does Smoking Cause Lung Cancer? (3 of 3)

- *The alleged cause precedes the effect in time*. Lung cancer develops after years of smoking. The number of men dying of lung cancer rose as smoking became more common, with a lag of about 30 years. Lung cancer kills more men than any other form of cancer. Lung cancer was rare among women until women began to smoke. Lung cancer in women rose along with smoking, again with a lag of about 30 years.

- *The alleged cause is plausible*. Experiments with animals show that tars from cigarette smoke do cause cancer.

Joy of Stats – Correlation between smoking and lung cancer

https://www.youtube.com/watch?v=6RzDMEW5omc