Probability and Mathematical Statistics in Data Science

Lecture 33: Section 12.1: The Simple Linear Regression Model

The Simple Linear Regression Model

- Model has two variables: response (Y) & (x) predictor or explanatory variable

- **Assumptions**: response is a linear function of the predictor (signal) + random error (noise), where the noise has a **normal** distribution, centered at 0.

- The signal is not random, but the response is, because the noise is random:

response = signal + noise

The Regression Line

- For each i, we want to get as close as we can to the *signal* $\beta_0 + \beta_1 x_i$
- There is some "true" regression line $\beta_0 + \beta_1 x$ that we cannot observe. We estimate this line by minimizing the squared observed error.
- Estimate of the line given the data is $Y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$, where $\hat{\beta}_0$ and $\hat{\beta}_1$ are the estimates of the intercept and slope, respectively, given the data.

A Linear Probabilistic Model

- For the deterministic model $Y = \beta_0 + \beta_1 x$, the actual observed value of *y* is a linear function of *x*.

- The appropriate generalization of this to a probabilistic model assumes that *the expected value of Y is a linear function of x*, but that for fixed *x* the variable *Y* differs from its expected value by a random amount.

A Linear Probabilistic Model

The Simple Linear Regression Model

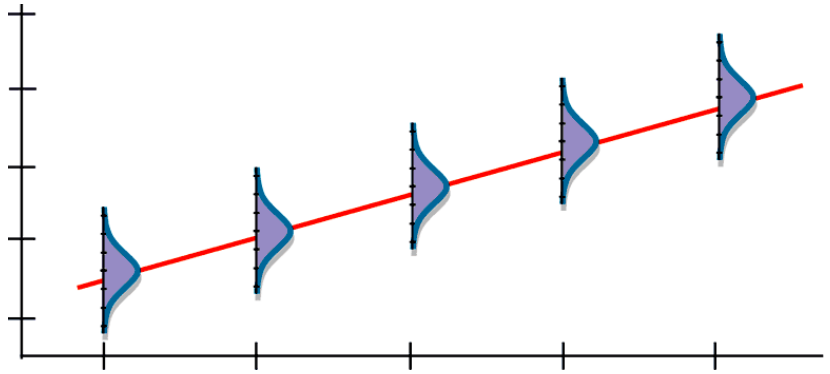
There are parameters β_0, β_1, and σ^2, such that for any fixed value of the independent variable x, the dependent variable is a random variable related to x through the **model equation**

$$Y = \beta_0 + \beta_1 x + \epsilon \qquad (12.1)$$

The quantity ϵ in the model equation is a random variable, assumed to be normally distributed with $E(\epsilon) = 0$ and $V(\epsilon) = \sigma^2$.

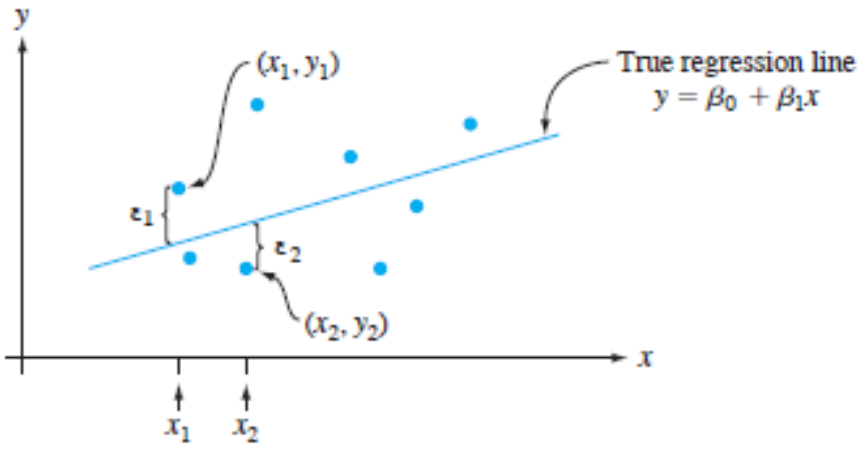
The Individual Response Y_i

- For any fixed i, Y_i is the sum of the signal and the noise.
- The signal is not random, but the noise is random with $\epsilon_i \sim N(0, \sigma^2)$
- Therefore what is the distribution of the Y_i ?



Therefore **the response Y_i of individual i has the normal distribution with mean $\beta_0 + \beta_1 x_i$ and variance σ^2.**

The Simple Linear Regression Model



$\mu_{Y \cdot x^*}$ = the expected (or mean) value of Y when x has value x^*

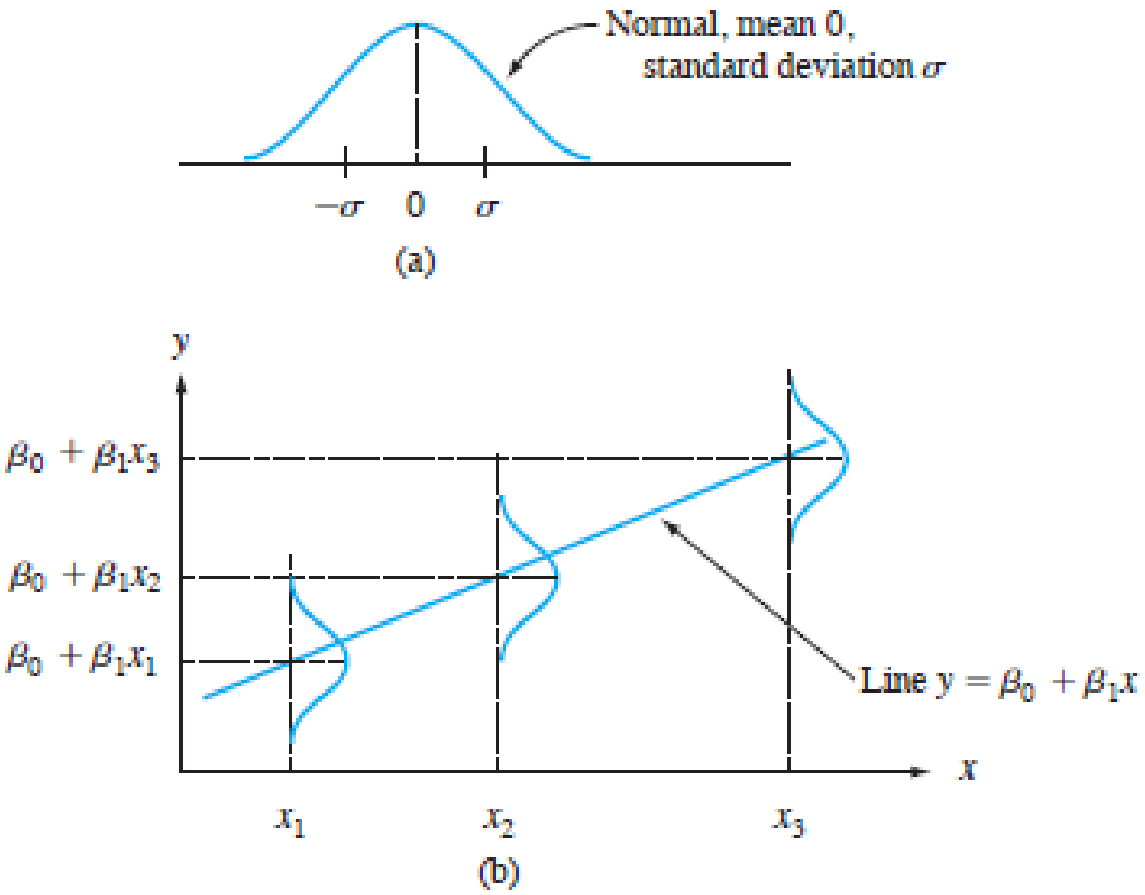
$\sigma^2_{Y \cdot x^*}$ = the variance of Y when x has value x^*

$$\mu_{Y \cdot x^*} = E(\beta_0 + \beta_1 x^* + \epsilon) = \beta_0 + \beta_1 x^* + E(\epsilon) = \beta_0 + \beta_1 x^*$$

$$\sigma^2_{Y \cdot x^*} = V(\beta_0 + \beta_1 x^* + \epsilon) = V(\beta_0 + \beta_1 x^*) + V(\epsilon) = 0 + \sigma^2 = \sigma^2$$

Devore 7th edition

The Simple Linear Regression Model



The Simple Linear Regression Model

The simple linear regression equation provides an estimate of the population regression line

Estimated (or predicted) Y value for observation i

Estimate of the regression intercept

Estimate of the regression slope

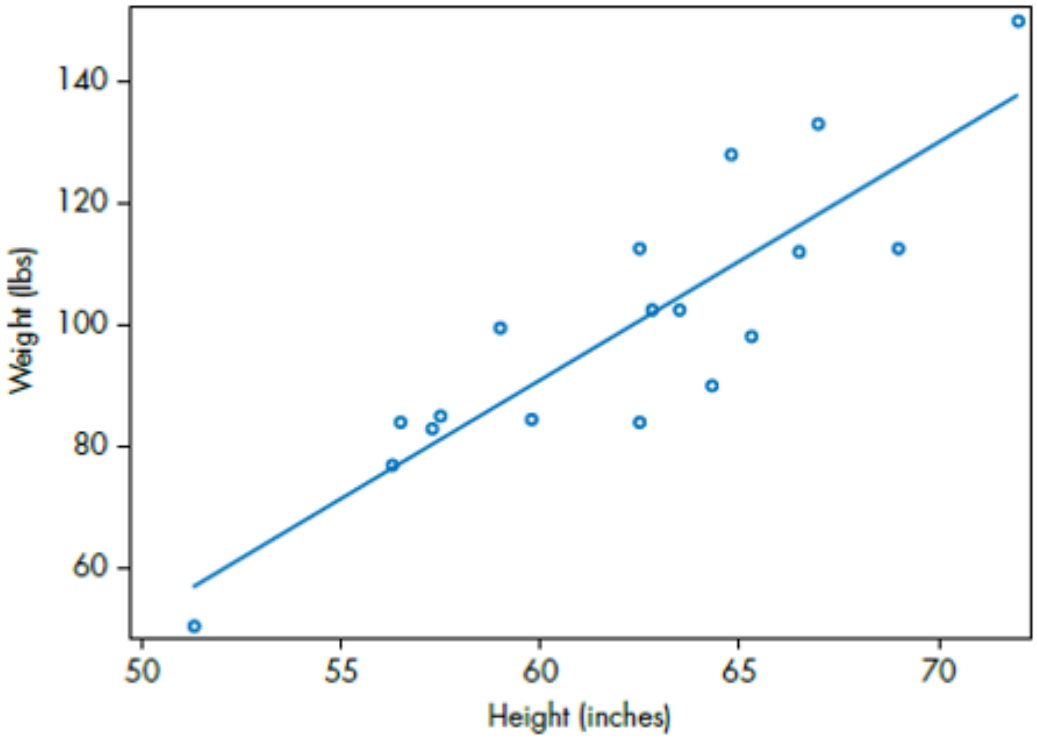
Value of X for observation i

$$\hat{Y}_i = b_0 + b_1 X_i$$

Interpretation of the Slope and the Intercept

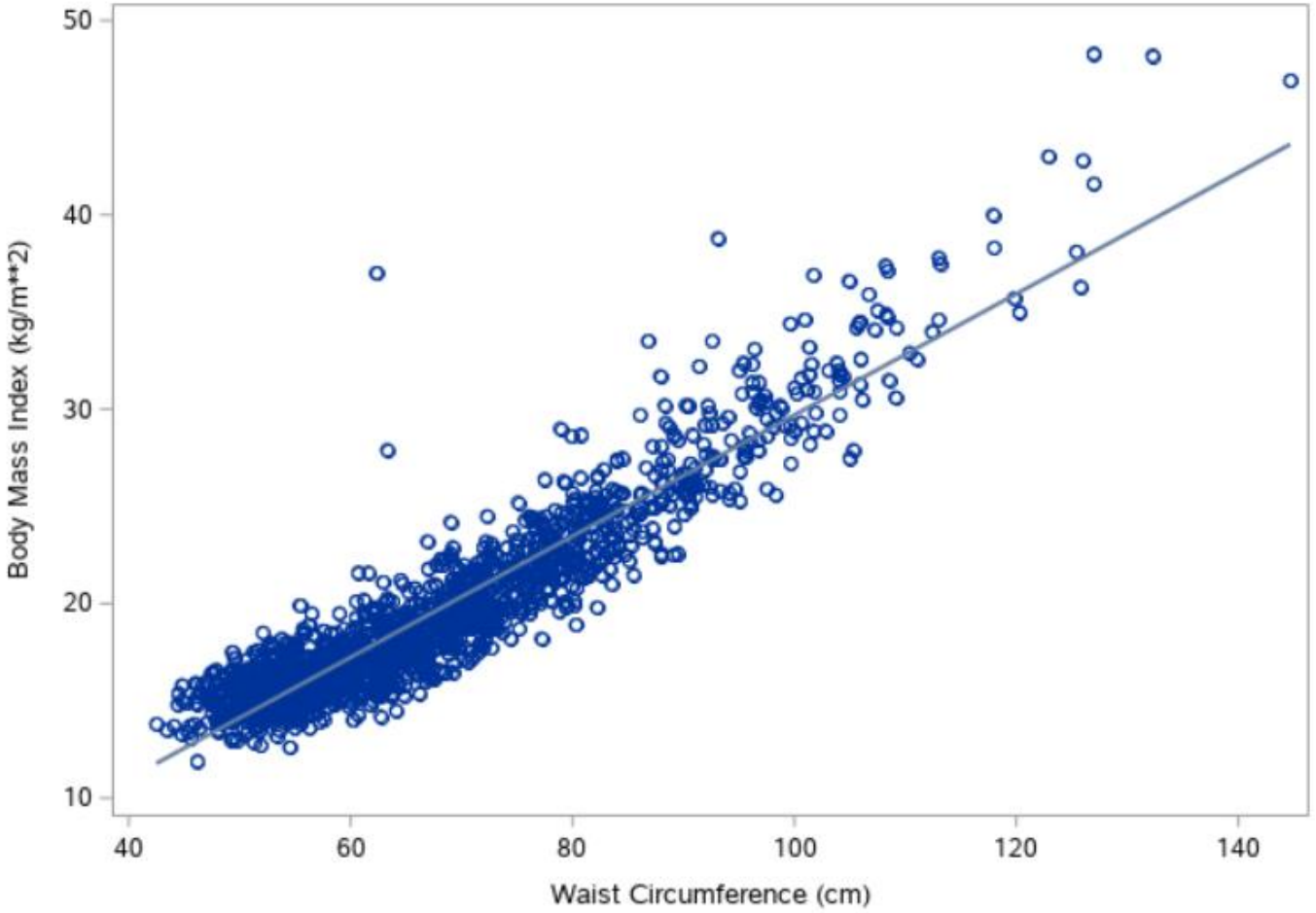
- b_0 is the estimated average value of Y when the value of X is zero
- b_1 is the estimated change in the average value of Y as a result of a one-unit increase in X

Specifying Linear Relationships with Linear Regression

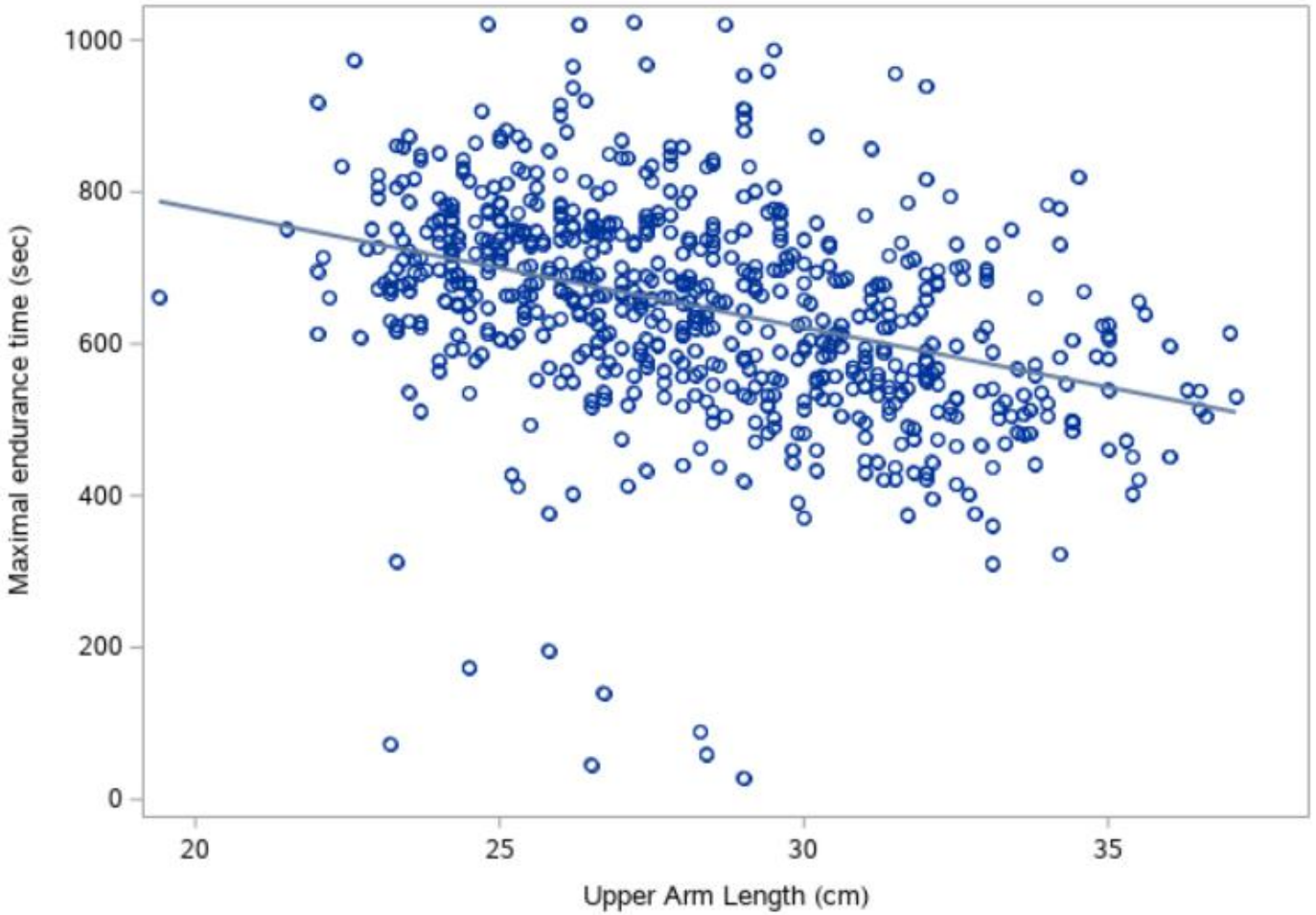


- Our aim is to fit a line to the data that gets as close to the data points as possible.
- For this reason, the line is often called the **line of best fit**.

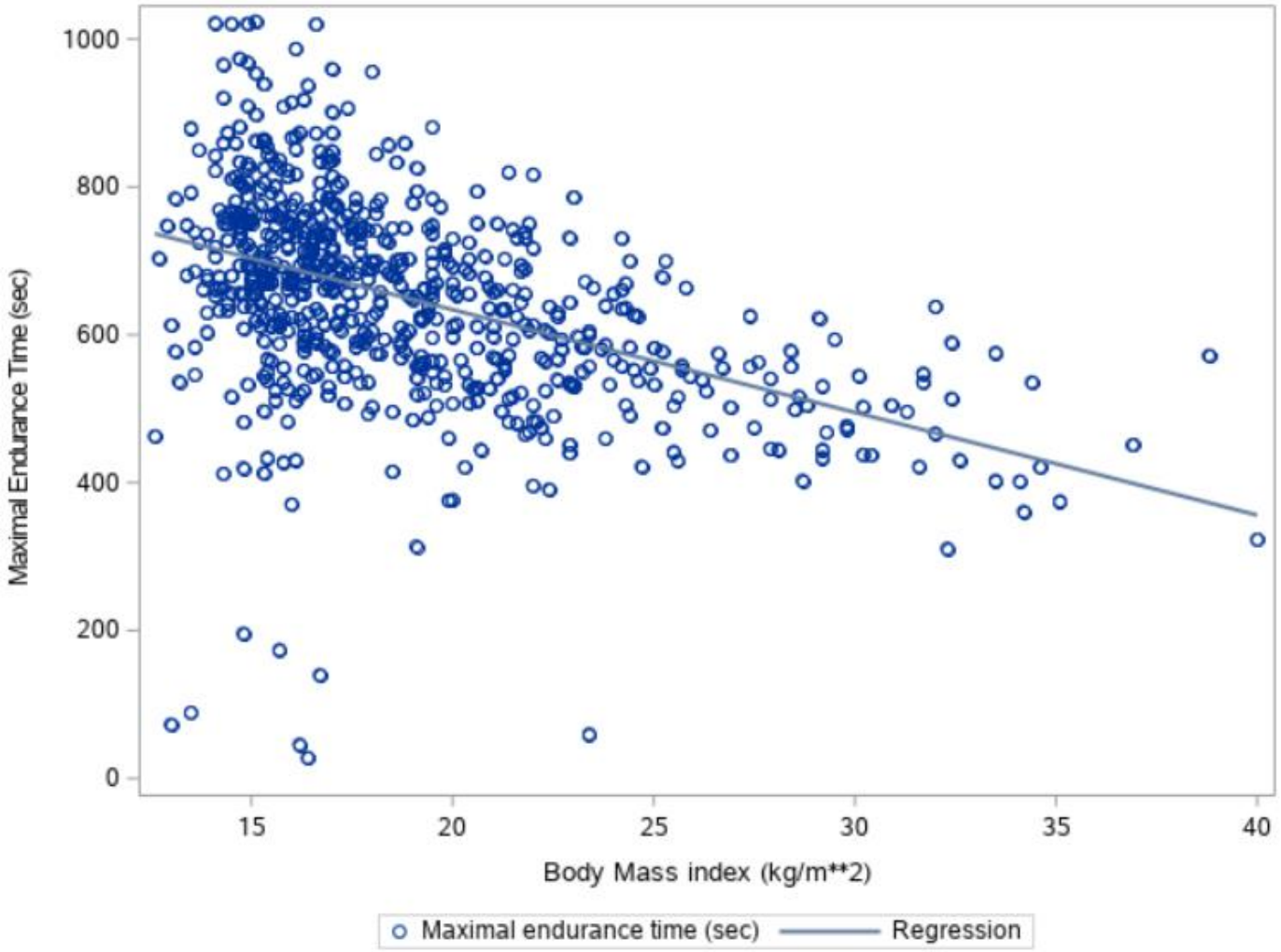
Specifying Linear Relationships with Linear Regression



Specifying Linear Relationships with Linear Regression



Specifying Linear Relationships with Linear Regression



The Equation of the Line of Best Fit

The equation for the regression line can be written as follows:

$$\hat{y}_i = b_0 + b_1 x_i$$

For our BMI versus Maximal Endurance Time (MET) example:

Predicted MET = Intercept + Slope x BMI

Predicted MET = 911.9 - 13.9 x BMI

We can use the slope to explain the relationship between BMI and MET

We can enter different values of BMI into the equation to make predictions for MET for those particular values of BMI

NHANES National Youth Fitness Survey

Predicted MET = 911.9 - 13.9 x BMI

BMI = 25: Predicted MET = 911.9 - 13.9 x (25)

= 564.4 secs

BMI = 26: Predicted MET = 911.9 - 13.9 x (26)

= 550.5 secs

Q. What do you think the the slope of -13.9 means? Explain in terms of increasing the value of BMI by one unit of measurement and the resulting change in the predicted value of MET