# Probability and Mathematical Statistics in Data Science

Lecture 35: Section 12.2: The Distribution of the Estimated Slope

# Distribution of the Estimated Slope $\hat{\beta}_1$

- $E(\hat{\beta}_1) = \beta_1$ indicating that $\hat{\beta}_1$ is an unbiased estimator of $\beta_1$
- Recall that the common variance of the errors $\epsilon_i$ is $\sigma^2$
- FACT: $Var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$
- What you want to note is that the numerator is constant, so as we have more data, the denominator gets larger, and our estimated slope gets closer to the true slope.
- We will need to estimate $\sigma^2$ since it is an unknown parameter.

# SD of the estimated slope $\hat{\beta}_1$

- Need to estimate $\sigma$, which we will do by using the SD of the residuals.

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

- We will denote this estimated $SD(\hat{\beta}_1)$ by $\boldsymbol{SE(\hat{\beta}_1)}$.

$$SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

# Making Statistical Decisions Regarding the Value of the Population Slope

- We will now apply the statistical methods of confidence intervals and hypothesis testing to linear regression. We are making statistical decisions with regard to what is the population slope

- We will attempt to capture the **population slope** within the limits of a **confidence interval**

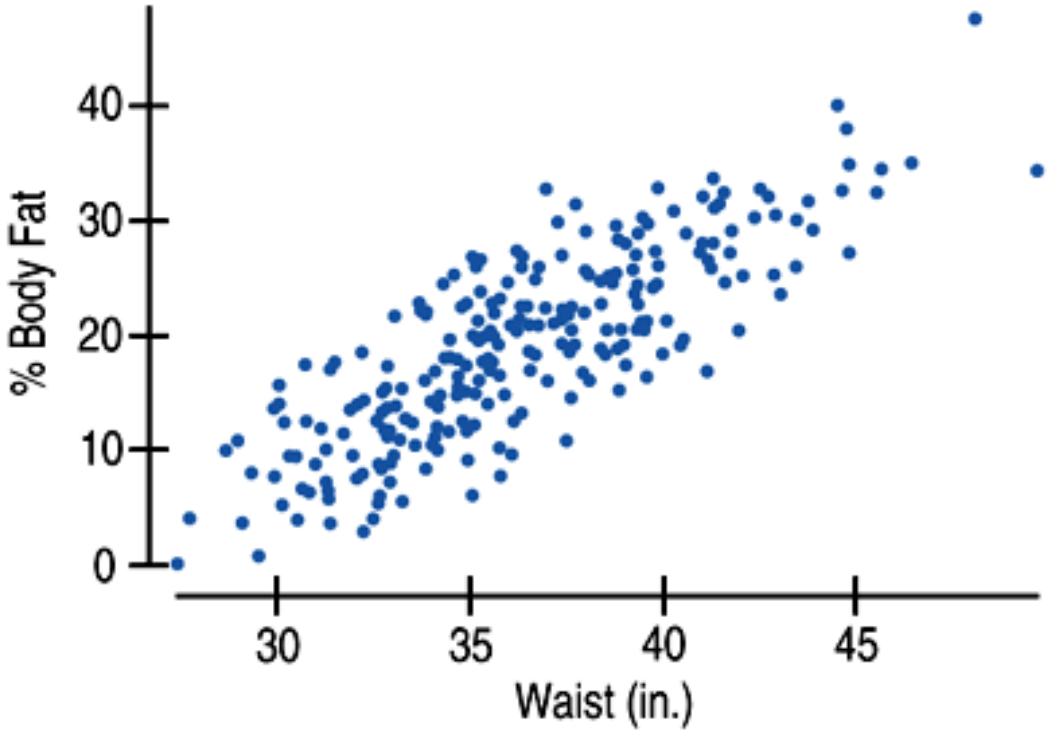
- We will conduct a **hypothesis test** to determine if the value of the sample slope is enough **statistical evidence** to state that there is a **linear relationship** in the population

- In other words, we will test to see if we have found statistical evidence that the population slope is not equal to zero

# Assumptions and Conditions for Regression Analysis

## 1. Linearity Assumption:

- **Straight Enough Condition:** Check the scatterplot—the shape must be linear or we can't use regression at all.



# Assumptions and Conditions for Regression Analysis

## 1. Linearity Assumption:

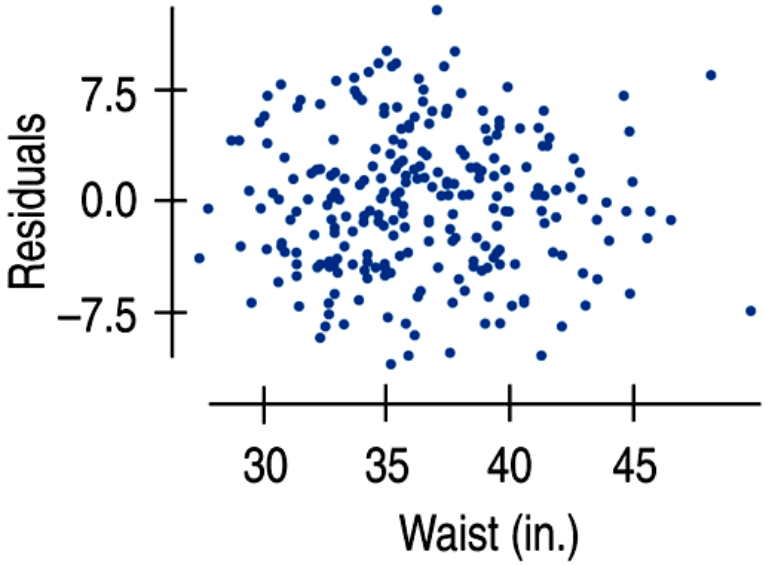
- If the scatterplot is straight enough, we can go on to some assumptions about the errors. If not, stop here, or consider re-expressing the data to make the scatterplot more nearly linear.

- Check the **Quantitative Data Condition**. The data must be quantitative for this to make sense.

# Assumptions and Conditions for Regression Analysis

## 2. Independence Assumption:

- **Randomization Condition**: the individuals are a representative sample from the population.
- Check the residual plot —the residuals should appear to be randomly scattered.




Residuals versus waist size

# Assumptions and Conditions for Regression Analysis

## 3. Equal Variance Assumption:

- **Does The Plot Thicken? Condition:** Residual plot to check the assumption about error terms have common variance
  - Homoscedasticity: same variance;
  - Heteroscedasticity: different variance.

# Assumptions and Conditions for Regression Analysis

## 4. Normal Population Assumption:

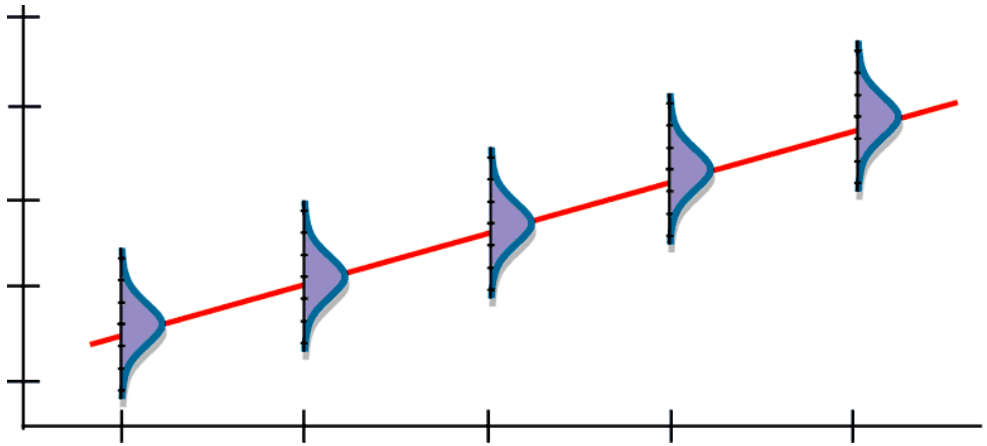
- **Nearly Normal Condition**

- Normal probability plot to check the assumption about error terms have normal distribution

- **Outlier Condition:** Check for outliers.

# Assumptions and Conditions for Regression Analysis

- If all four assumptions are true, the idealized regression model would look like this:



- At each value of *x* there is a distribution of *y*-values that follows a Normal model, and each of these Normal models is centered on the line and has the same standard deviation.

# Intuition About Regression Inference

- We expect any sample to produce a $b_1$ whose expected value is the true slope, $\beta_1$.
- What about its standard error of the sample slope?
- What aspects of the data affect how much the slope varies from sample to sample?

# Standard Error for the Estimated Slope

- Three aspects of the scatterplot affect the standard error of the regression slope:
  - spread around the line, $s_e$
  - spread of $x$ values, $s_x$
  - sample size, $n$.

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

- The formula for the standard error (which you will probably never have to calculate by hand) is:

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\, s_x}$$

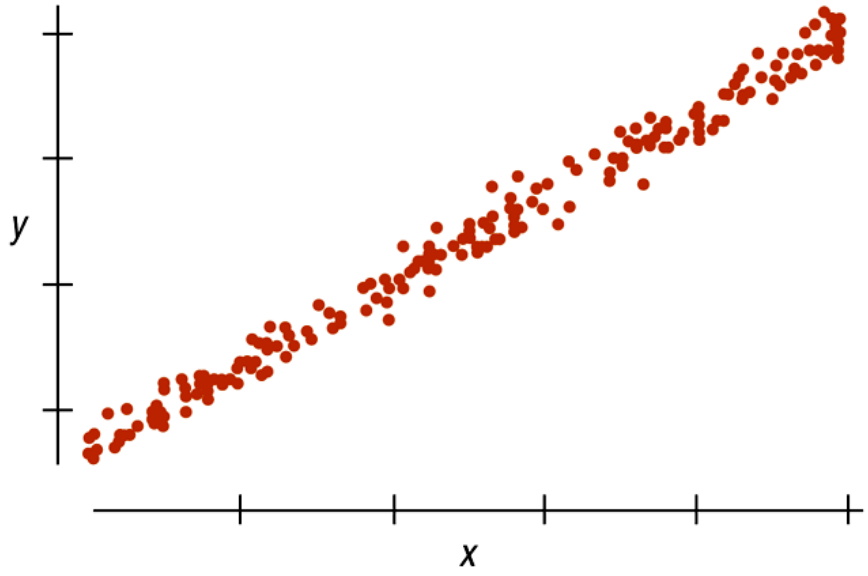
$$SD(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}}$$

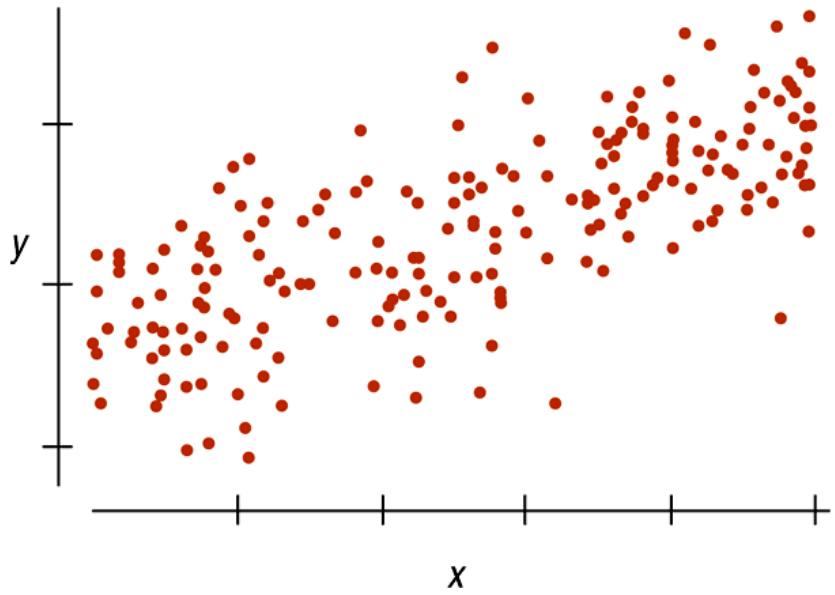
# Intuition About Regression Inference

- **Spread around the line:**
  - Less scatter around the line means the slope will be more consistent from sample to sample.
  - The spread around the line is measured with the residual standard deviation $s_e$.
  - You can always find $s_e$ in the regression output, often just labeled *s*.

# Intuition About Regression Inference

- **Spread around the line:**

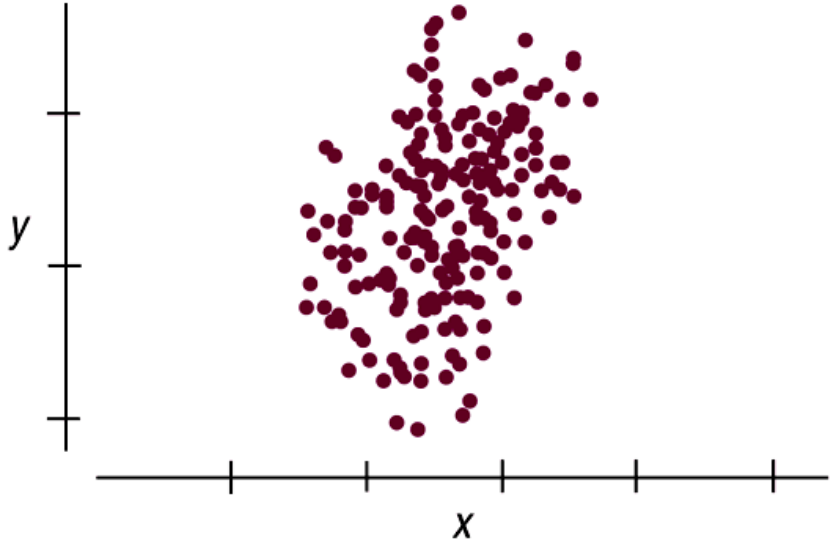


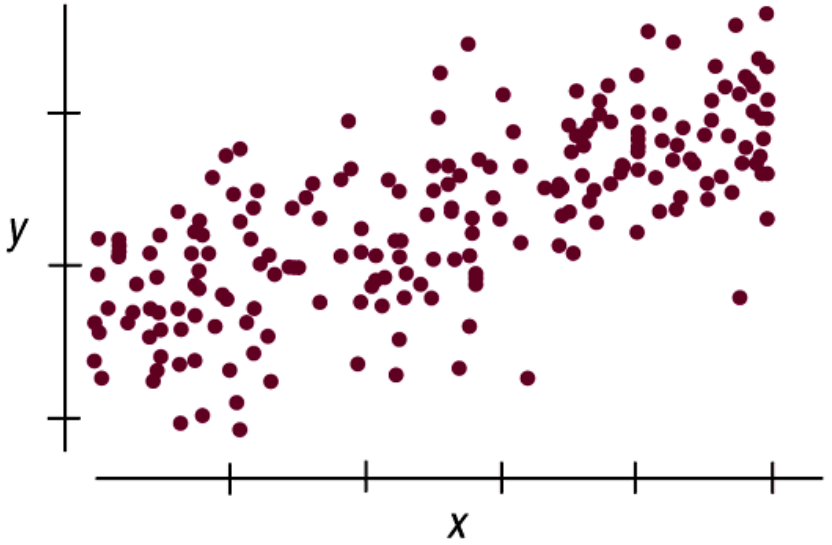




Less scatter around the line means the slope will be more consistent from sample to sample.

# Intuition About Regression Inference

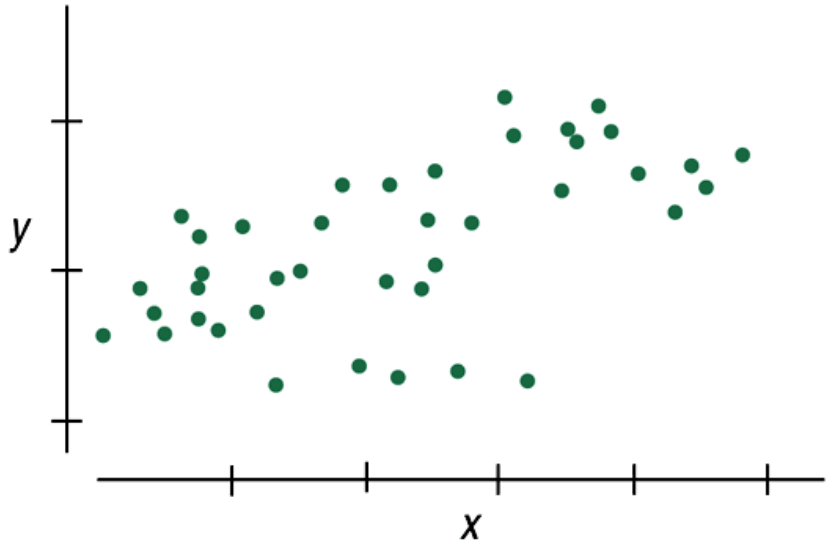
- **Spread of the *x*'s:** A large standard deviation of *x* provides a more stable regression.

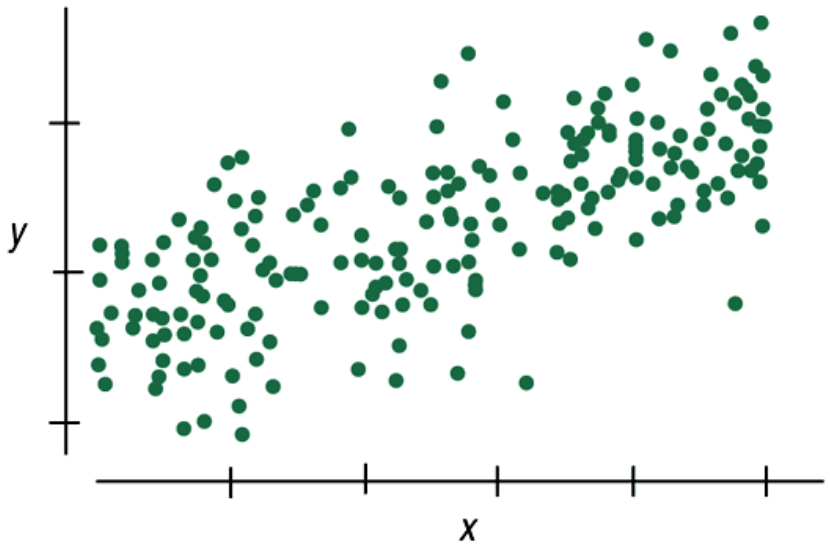



# Intuition About Regression Inference

- **Sample size:** Having a larger sample size, *n*, gives more consistent estimates.




# Sampling Distribution for Regression Slopes

- When the conditions are met, the standardized estimated regression slope

$$t = \frac{b_1 - \beta_1}{SE(b_1)}$$

follows a Student's *t*-model with $n - 2$ degrees of freedom.

# Sampling Distribution for Regression Slopes

- We estimate the standard error with

$$SE(b_1) = \frac{s_e}{\sqrt{n-1}\,s_x}$$

where:

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

- $n$ is the number of data values
- $s_x$ is the ordinary standard deviation of the $x$-values.

# Regression Analysis

- A null hypothesis of a zero slope questions the entire claim of a linear relationship between the two variables—often just what we want to know.
- To test $H_0: \beta_1 = 0$, we find

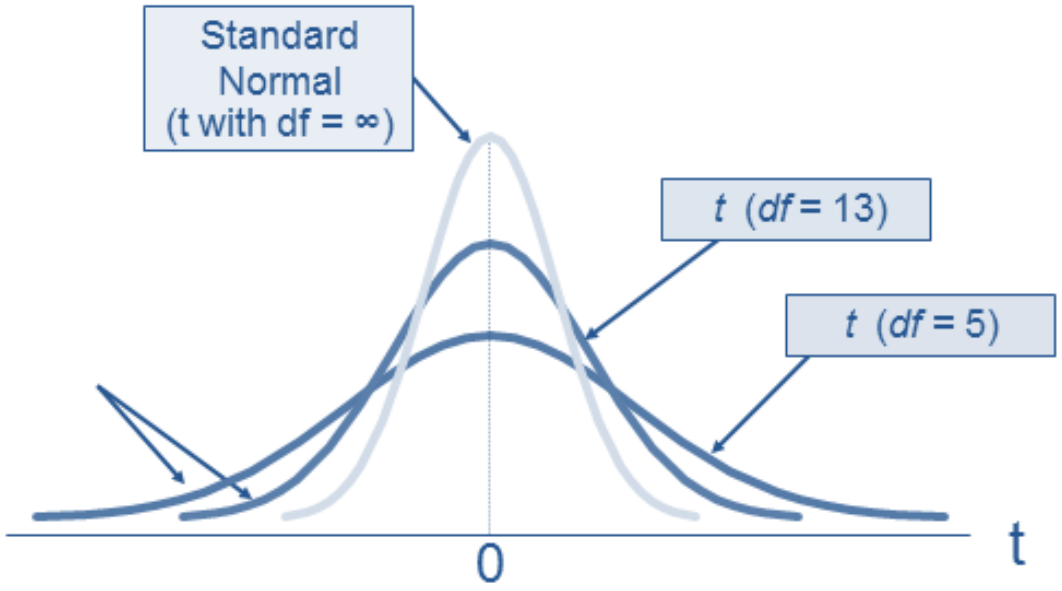
$$t_{n-2} = \frac{b_1 - 0}{SE(b_1)}$$

and continue as we would with any other *t*-test.

- The formula for a confidence interval for $\beta_1$ is

$$b_1 \pm t^*_{n-2} \times SE(b_1)$$

# The t-distribution




https://financetrain.com/students-t-distribution/

- There is a t-distribution for every sample size used. The larger the sample size the closer the shape of the t-distribution is to the standard normal or z-distribution. For sample sizes of 60 or greater they are practically the same distribution

# Degrees of Freedom

- If only we knew the true population mean, $\mu$, we would find the sample standard deviation as

$$s = \sqrt{\frac{\sum (y - \mu)^2}{n}}.$$

- But, we use $\bar{y}$ instead of $\mu$, though, and that causes a problem.
- When we use $\sum \left(y - \bar{y}\right)^2$ instead of $\sum \left(y - \mu\right)^2$ to calculate s, our standard deviation estimate would be too small.
- The amazing mathematical fact is that we can compensate for the smaller sum exactly by dividing by $n - 1$ which we call the degrees of freedom.

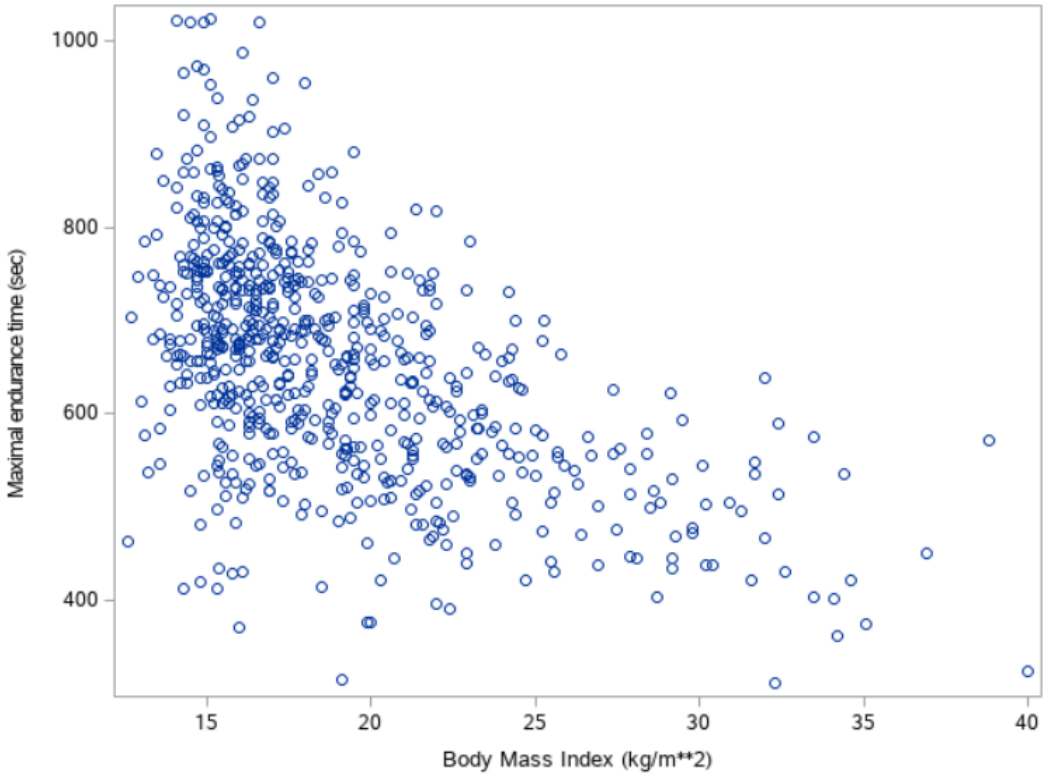
# NHANES National Youth Fitness Survey

- We will construct a confidence interval to try and capture the population slope (for the relationship between BMI and MET) within its limits

- We will conduct a hypothesis test known as a model utility test. It will test to see if there is statistical evidence in the data of a linear relationship between the two quantitative variables (BMI and MET) or not

# Checking the Assumptions and Conditions

## Linearity




- It is fair to assume that the relationship between MET and BMI in this sample is linear

# Checking the Assumptions and Conditions

**Independence**

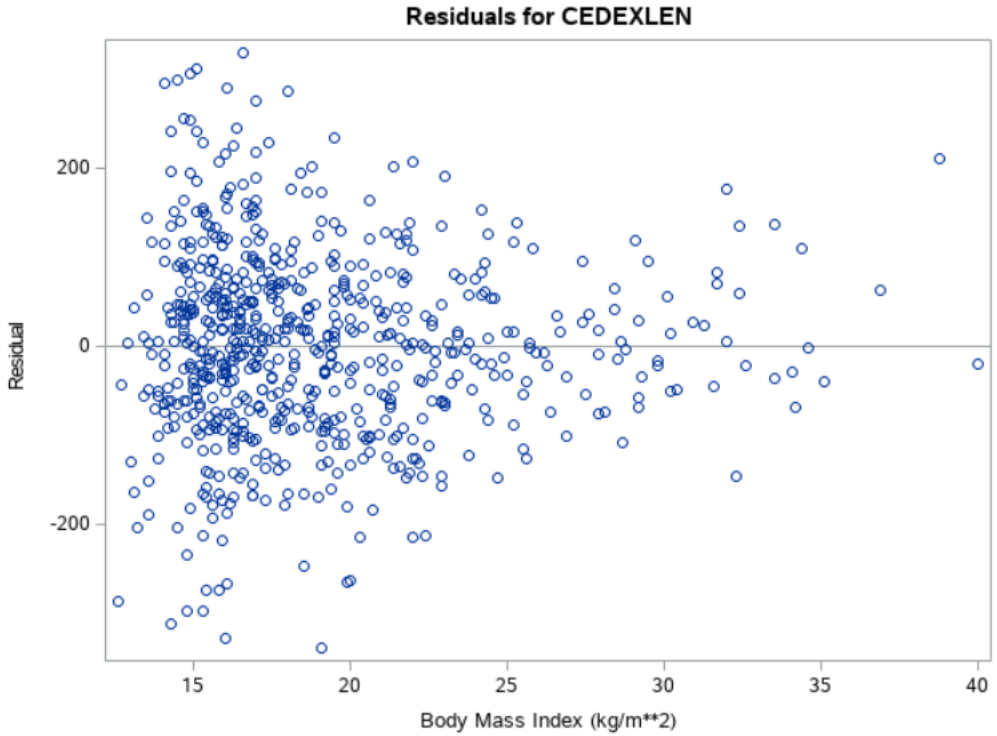
- The NHANES National Youth Fitness Survey (NNYFS) was a one year survey conducted in 2012. The NNYFS collected nationally representative (random sample of) data on physical activity and fitness levels of children and adolescents in the United States through interviews and fitness tests.

- **Q.** Do you think the MET and BMI measurements will be independent across children in the survey?

- **Q.** Regarding MET, do you think independence across values of this measurement also depends on how the study was conducted? Explain

# Checking the Assumptions and Conditions

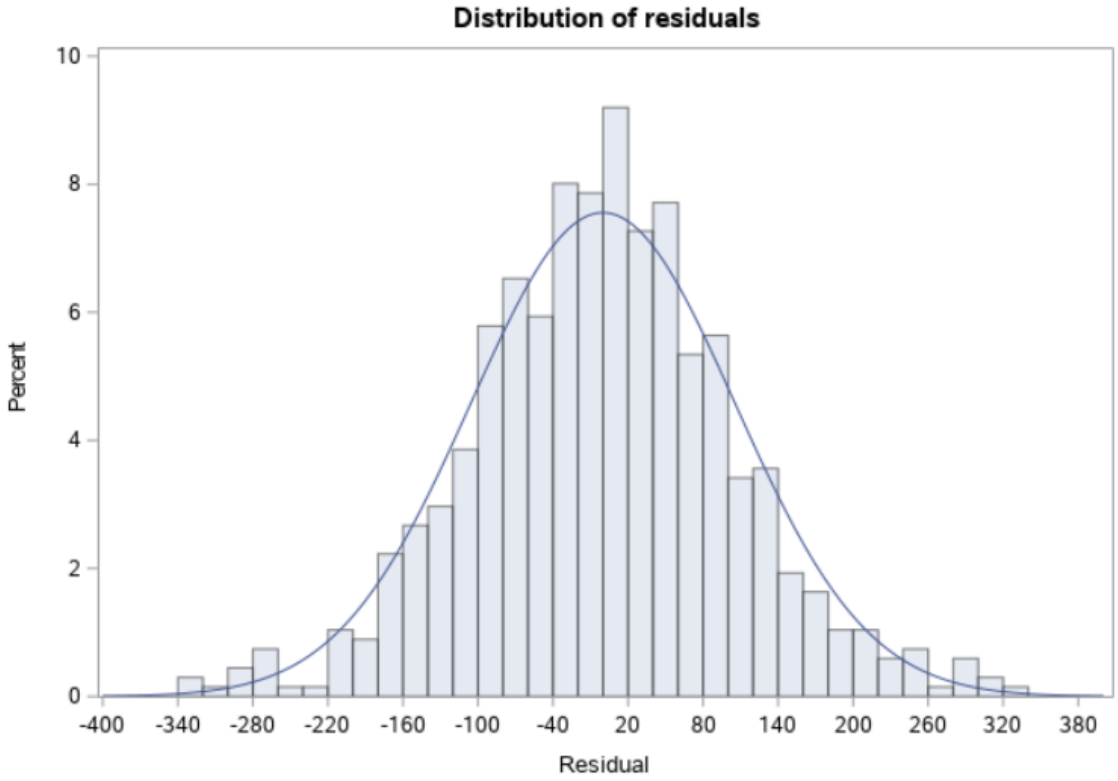
## Equal Variance




There is more variability in residuals for smaller values of BMI compared to larger values of BMI. However, we can say that the assumption of equal variance is approximately valid

# Checking the Assumptions and Conditions

## Normal Distribution (of residuals)




The residuals are approximately normally distributed

# NHANES National Youth Fitness Survey

- Since the assumptions and conditions are (approximately) valid, we will complete a statistical analysis of the data. Our 95% confidence interval can be calculated as follows:

$$\text{sample slope} \pm 2 \times (\text{standard error})$$

$$-14.85 \pm 2 \times (0.90)$$

$$[-16.65, -13.05]$$

**Q**. What does the confidence interval mean?

**Q.** From the CI can we state that we found evidence of a linear relationship between MET and BMI?

# NHANES National Youth Fitness Survey

- We want to run a hypothesis test to see if we have statistical evidence in the data that there is a linear relationship between BMI and MET

**Step 1**

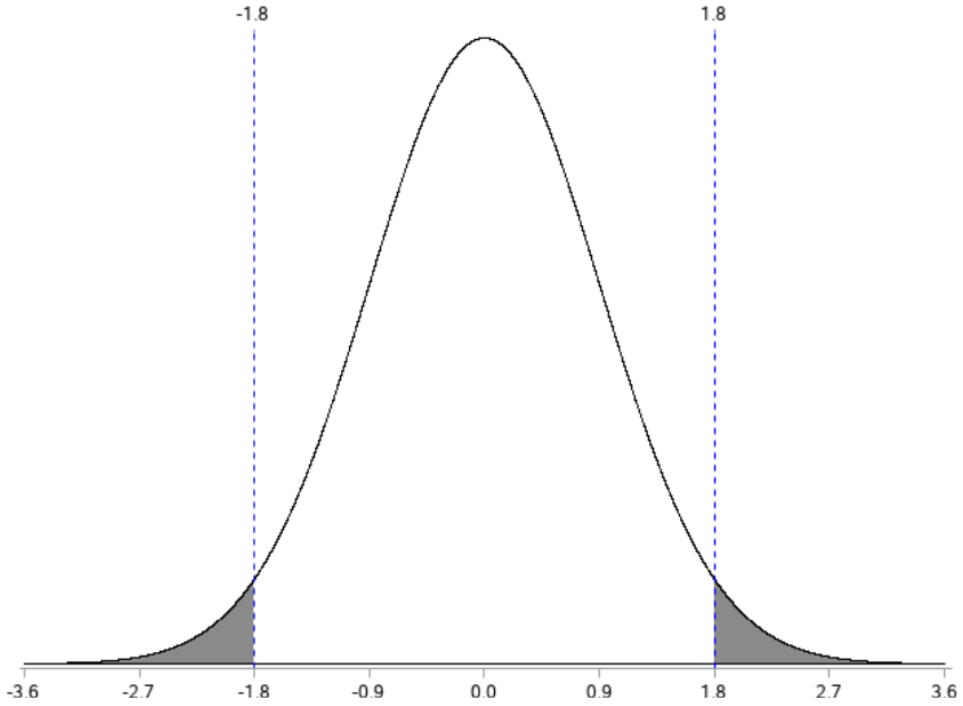
**Null Hypothesis:** Population Slope is equal to zero ($\mathbf{H_0: \beta_1 = 0}$)

**Alternative Hypothesis:** Population Slope is not equal to zero($\mathbf{H_a: \beta_1 \neq 0}$)

We are testing to see if we have evidence that the population slope does not equal zero

# NHANES National Youth Fitness Survey

**Step 2: The Model**




The standard error of the sample slope is equal to 0.90. If the null hypothesis is true, we expect 95% of sample slope to fall between -1.8 and 1.8

# NHANES National Youth Fitness Survey

**Step 3 : Calculations**

test statistic = (sample slope – null value)/standard error

= (-14.85 – 0)/0.90

= -16.5

With such a large (negative) test statistic, the p-value is less than 0.001

# NHANES National Youth Fitness Survey

**Step 4: Conclusion**

Since the p-value is far less than 0.05, we will reject the null hypothesis in favor of the alternative

There is strong evidence in the data of a linear relationship between BMI and MET

Furthermore, we are 95% confident that the population slope could be anywhere between -16.65 and -13.05

# Measuring the Amount of Variation Explained by the Linear Regression Model

- **R-Square** (the correlation squared) is a sample statistic that measures the proportion (or percentage) of variation in the response variable that can be explained by a linear relationship with the explanatory variable.

- In our example looking at the relationship between MET and BMI, R-square is equal to 0.28 (-0.53 x -0.53).

- This means that 0.28 (or 28%) of the variability in MET scores can be explained by a linear relationship with BMI

- It is also known as the **coefficient of determination**

# Body Fat Example: Excel Output

**The regression equation is:**

$$\widehat{\text{Body Fat}}(\%) = -27.376 + 0.2499\,(\text{weight})$$

SUMMARY OUTPUT

| *Regression Statistics* | |
|---|---|
| Multiple R | 0.69663276 |
| R Square | 0.485297203 |
| Adjusted R Square | 0.456702603 |
| Standard Error | 7.049132279 |
| Observations | 20 |

ANOVA

| | *df* | *SS* | *MS* | *F* | *Significance F* |
|---|---|---|---|---|---|
| Regression | 1 | 843.325214 | 843.3252 | 16.97164 | 0.000643448 |
| Residual | 18 | 894.424786 | 49.69027 | | |
| Total | 19 | 1737.75 | | | |

| | *Coefficients* | *Standard Error* | *t Stat* | *P-value* | *Lower 95%* | *Upper 95%* |
|---|---|---|---|---|---|---|
| Intercept | -27.37626233 | 11.54742832 | -2.37077 | 0.029119 | -51.63650899 | -3.116015659 |
| Weight | 0.249874137 | 0.060653997 | 4.119665 | 0.000643 | 0.122444818 | 0.377303457 |

# Inferences About the Slope

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

**From Excel output:**

| | Coefficients | Standard Error | t Stat | P-value |
|---|---|---|---|---|
| Intercept | -27.37626233 | 11.54742832 | -2.37077 | 0.029119 |
| X Variable 1 | 0.249874137 | 0.060653997 | 4.119665 | 0.000643 |

$b_1$

$S_{b_1}$

$$t_{STAT} = \frac{b_1 - \beta_1}{S_{b_1}} = \frac{0.249874 - 0}{0.060654} = 4.1197$$

# Inferences About the Slope:

$H_0: \beta_1 = 0$

$H_1: \beta_1 \neq 0$

| | Coefficients | Standard Error | t Stat | P-value |
|---|---|---|---|---|
| Intercept | -27.37626233 | 11.54742832 | -2.37077 | 0.029119 |
| X Variable 1 | 0.249874137 | 0.060653997 | 4.119665 | 0.000643 |

p-value

Decision: Reject $H_0$, since p-value < $\alpha$=0.05

There is sufficient evidence that weight is related to % Body Fat.

# F Test for Overall Significance

- F Test statistic:

$$F_{STAT} = \frac{MSR}{MSE}$$

$$MSR = \frac{SSR}{k}$$

$$MSE = \frac{SSE}{n-k-1}$$

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n-2}}$$

where $F_{STAT}$ follows an F distribution with k numerator and (n – k - 1) denominator degrees of freedom

(k = the number of independent variables in the regression model)

# F-Test for Overall Significance

$$s_e = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

| Regression Statistics | |
|---|---|
| Multiple R | 0.69663276 |
| R Square | 0.485297203 |
| Adjusted R Square | 0.456702603 |
| Standard Error | 7.049132279 |
| Observations | |

$$F_{STAT} = \frac{MSR}{MSE} = 843.3252 / 49.69027 = 16.97164$$

**With 1 and 18 degrees of freedom**

**p-value for the F-Test**

| ANOVA | df | SS | MS | F | Significance F |
|---|---|---|---|---|---|
| Regression | 1 | 843.325214 | 843.3252 | 16.97164 | 0.000643448 |
| Residual | 18 | 894.424786 | 49.69027 | | |
| Total | 19 | 1737.75 | | | |