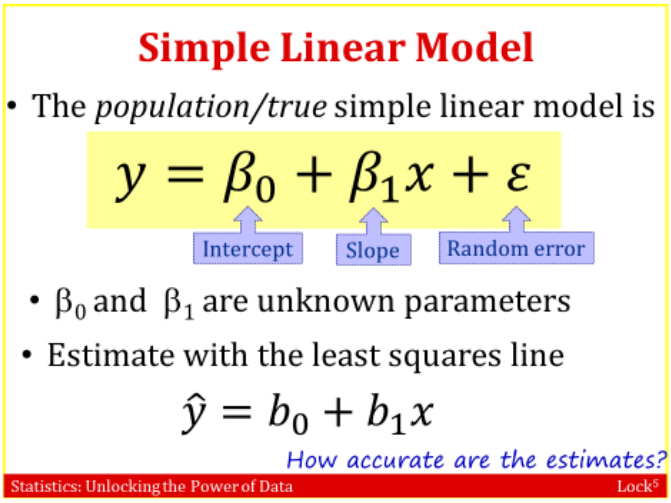
Section 9.1: Inference for Slope and Correlation

In chapter 2 we looked out how to summarize the information in a scatterplot using a regression line. In this section, we will learn some new notation and vocabulary used to formally describe the regression line.

If we had a scatterplot of population data we could use a line to describe the points. When we use a line to describe population data we call it a simple linear model and we can use the following notation to describe the line:



If we had a scatterplot of sample data we could use a line to describe the points. When we use a line to describe sample data we call it a regression line and we can use the above notation to describe the line.

Notice that we now have 2 new parameters with 2 corresponding statistics.

Parameters	Statistics
β_0	b_0
β_1	b_1

We can use b_0 and b_1 to learn about β_0 and β_1 respectively

Example: With your group answer the following: Recall the previous example with crickets. Previously we found that …

$$\widehat{Temp} = 37.7 + 0.23\ Chirps$$

- What is the symbol and value of the intercept statistic? b_0 =37.7

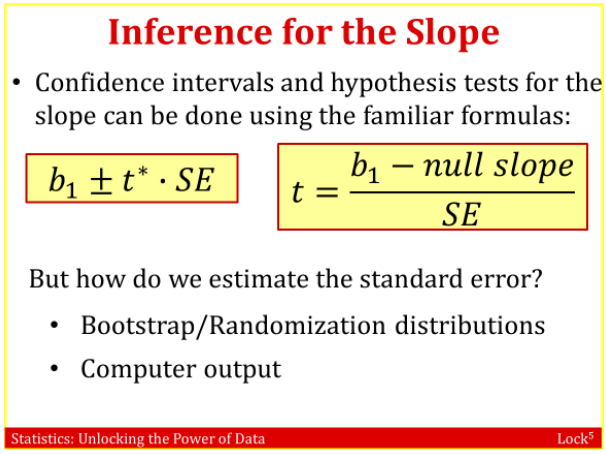
- What is the symbol for the intercept parameter? β_0
- What is the symbol and value of the slope statistic? b_1 =0.23
- What is the symbol for the slope parameter? β_1

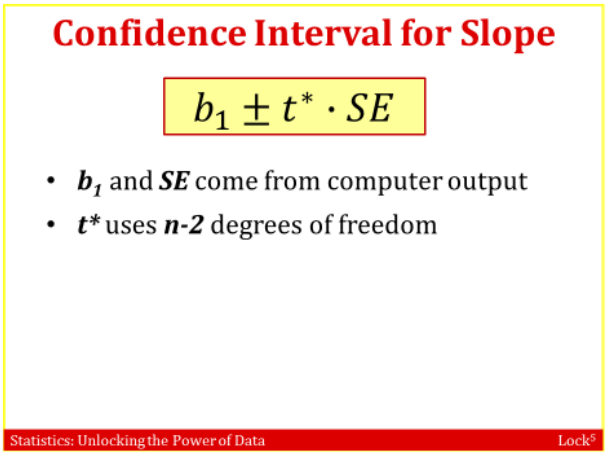
Usually we are interested in learning about the slope. The slope can help us determine if there is a linear relationship between the explanatory and response variable.

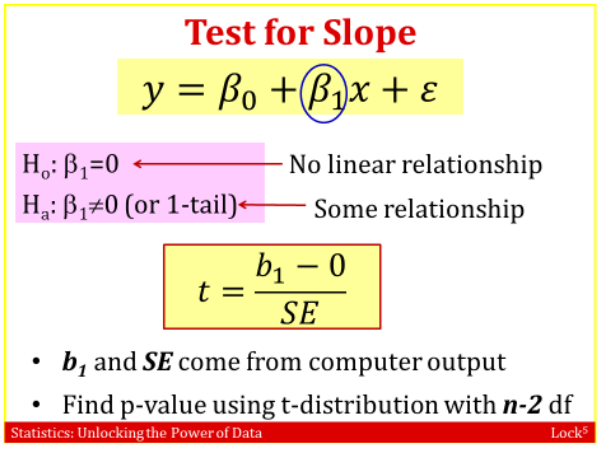
Confidence Interval for Slope:

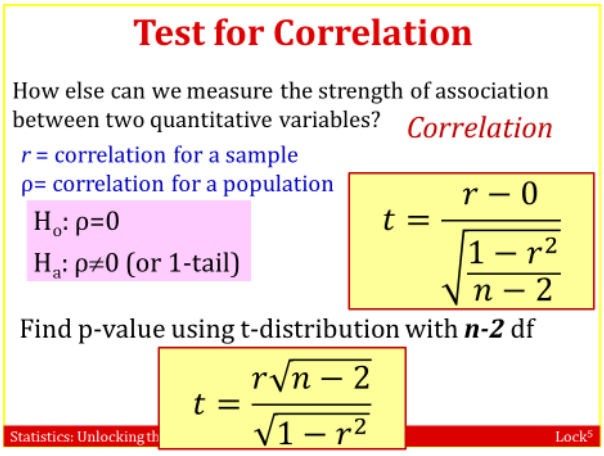
We can create a confidence interval to give us a range of reasonable values for the population slope (β_1).

The formula is:









Coefficient of Determination, R^2

Recall that for correlation: $-1 \le r \le 1$

If we *square* the correlation, r^2, we get a number between 0 and 1 that can be interpreted as a percentage

R^2 = proportion of ***variability*** in response variable *Y* that is "explained" by the model based on the predictor *X*.

(by convention we use a capital R^2, although the value is just r^2 for a single predictor)

Checking Conditions

$$y = \beta_0 + \beta_1 x + \varepsilon$$

For a simple linear model, we assume the errors (ε) are randomly distributed above and below the line.

Quick check : (more details in Section 10.2)

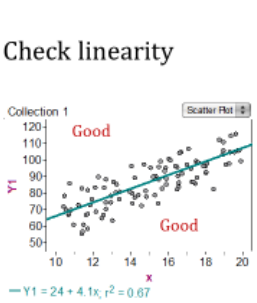
Look at a scatterplot with regression line on it

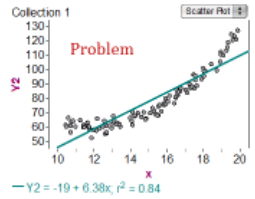
Watch out for:

- Curved (nonlinear) patterns in the data
- Consistently changing variability
- Outliers and influential points

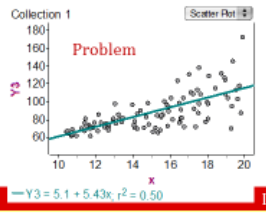
Scatterplot with Regression Line

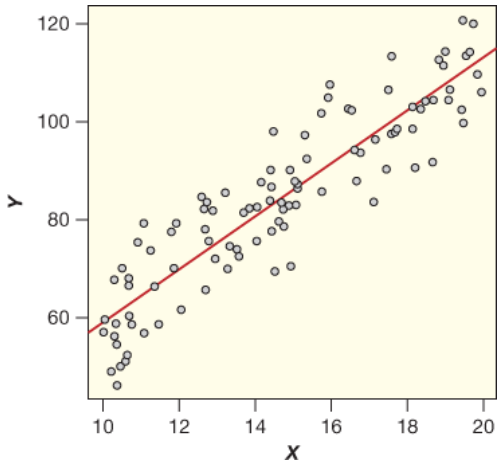
Check linearity

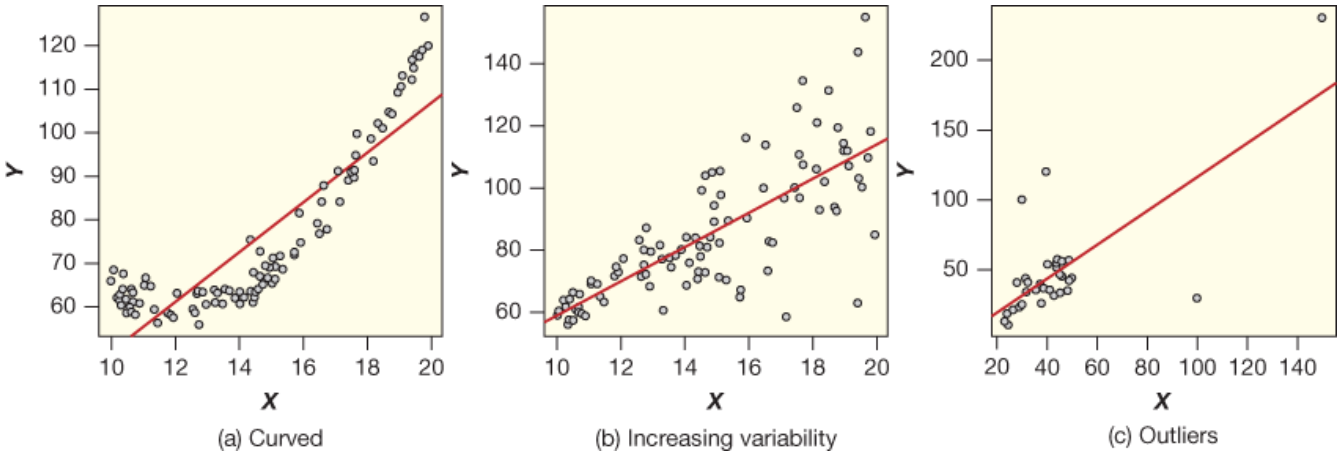




Check consistent variability







Example 1: Depression and Missed Classes

Is depression a possible factor in students missing classes? Two of the variables in the dataset **SleepStudy** are *DepressionScore*, which gives score on a standard depression scale with higher numbers indicating greater depression, and *Classes Missed*, the number of classes missed during the semester,

for a sample of 253 college students. Computer output is shown below for correlation between these two variables and the regression line to predict classes missed based on the depression score.

```
Pearson correlation of DepressionScore and classes Missed = 0.154

P-Value = 0.014

The regression equation is ClassesMissed = 1.78 + 0.0831 DepressionScore

Predictor            Coef       SE Coef        T       P

Constant            1.7771       0.2671      6.65    0.000

Depression Score   0.08312       0.03368     2.47    0.014

S = 3.20806    R-Sq = 2.4%    R-Sq(adj) = 2.0%
```

(a) What is the sample correlation? 0.154 What is the p-value for testing the correlation? 0.014
Give the conclusion of the test in context.
At a 5% level, we reject H_0 and conclude that there is evidence that $\rho \neq 0$, which means there is evidence some association between depression scores and classes missed.

(b) What is the regression line?

$$\widehat{ClassesMissed} = 1.78 + 0.0831\ DepressionScore$$

Find the predicted value and residual for an individual with a depression score of 7 who has missed 4 classes.
The predicted number of missed classes is: $\widehat{ClassesMissed} = 1.78 + 0.0831\ (7) = 2.36$
The residual is $4 - 2.36 = 1.64$

(c) Interpret the slope of the line in context.
If the depression score were __one___ point higher, the predicted number of missed classes would be _____0.0831 ___ higher.

(d) What is the p-value for a test of the slope? *__0.014___* Give the conclusion of the test in context.
At a 5% level, we reject H_0 and conclude that there is evidence that $\beta_1 \neq 0$, which means there is evidence that a student's depression score is an effective predictor of the number of classes missed.

(e) What is the standard error of the slope? __0.03368___ Find and interpret a 95% confidence interval for the slope of the regression line.

We use df = n – 2 = 251, and find that t = 1.970. We have*

$$b_1 \pm t^* \cdot SE$$
$$0.0831 \pm 1.970(0.03368)$$
$$0.0831 \pm 0.0664$$
$$0.0167 \text{ to } 0.1495$$

We are 95% sure that the slope of the linear model to predict classes missed from the depression score for all students at this college is between 0.0167 and 0.1495.

(f) Compare the two p-values, from the test for correlation and the test for slope.
They are the same, as expected.

(g) What is the value of R^2 for this model? Interpret it in context.
R^2=2.4%. The depression scores only explain 2.4% of the variability in number of classes missed.

Quick Self-Quiz: Inference for Regression: Alcohol and Missed Classes

Another possible explanatory variable for the *ClassesMissed* variable in the **SleepStudy** dataset, described in Example 1, is *Drinks*, the number of alcoholic drinks in a week that the college students say they have. Computer output is shown below for correlation between these two variables and the regression line to predict classes missed based on the number of alcoholic drinks.

```
Pearson correlation of Drinks and ClassesMissed = 0.078

P-Value = 0.215

The regression equation is ClassesMissed = 1.86 + 0.0620 Drinks

Predictor      Coef  SE Coef     T      P
Constant     1.8644   0.3439  5.42  0.000
Drinks      0.06196  0.04979  1.24  0.215

S = 3.23679   R-Sq = 0.6%   R-Sq(adj) = 0.2%
```

---

1. What is the correlation?

   $r = \pm\sqrt{0.006}$ r=0.07746 (because slope is positive, so r is positive)

2. Use the computer output to test the correlation to determine whether the number of alcoholic drinks is an effective predictor of classes missed. State the null and alternative hypotheses.

   H_0: $\rho = 0$ H_a: $\rho \neq 0$

3. Based on the available information, what is the test statistic for correlation?

$$t = \frac{r-0}{\sqrt{(1-r^2)/(n-2)}} = \frac{0.07746-0}{\sqrt{(1-0.006)/(253-2)}} = 1.231$$

   df = n-2 = 253-2 = 251

4. What is the p–value for testing the correlation?

   p-value = 2 tcdf(1.231,100,251)=0.222

5. What is the conclusion of testing the correlation?

   We do not reject H_0 and do not find evidence of a linear relationship between the number of classes missed and number of alcoholic drinks consumed.

6. What is the regression line?

$$\widehat{ClassesMissed} = 1.86 + 0.062\ Drinks$$

7. Find the predicted value for an individual who drinks 6 alcoholic drinks a week and has missed 2 classes.

   The predicted number of missed classes is: $\widehat{ClassesMissed} = 1.86 + 0.062\ (6) = 2.23.$

8. Find the residual for an individual who drinks 6 alcoholic drinks a week and has missed 2 classes.
   The residual is $2 - 2.23 = -0.23$

9. What is the slope?
   Slope= 0.06196

10. Interpret the slope of the line in context.
    If the the number of alcoholic drinks were 1 point higher, the predicted number of missed classes would be 0.06196 higher.

11. Interpret the intercept of the line in context.
    intercept=1.8644

    for students drink 0 alcoholic drinks will expect them miss 1.8644 classes

12. Based on the available information, what is the test statistic for slope?

$$t = \frac{b_1 - 0}{SE} = \frac{0.06196 - 0}{0.04979} = 1.244$$

13. What is the p-value for a test of the slope?
    p-value = 2 tcdf(1.244,100,251)=0.215

14. Give the conclusion of the test in context.
    We do not reject H_0 *and do not find evidence that the number of alcoholic drinks is effective at predicting the number of classes missed.*

15. What is the standard error of the slope?
    SE= 0.04979

16. What are the degrees of freedom for constructing a confidence interval for, or performing a test about, the population slope?
    df = n-2 = 253-2 = 251

17. To construct a 95% confidence interval for the population slope. find t^* = ?
    t^* = 1.969

18. Find a 95% confidence interval for the slope of the regression line.

$$b_1 \pm t^* \cdot SE$$
$$0.062 \pm 1.970(0.04979)$$
$$0.062 \pm 0.098$$
$$-0.036 \text{ to } 0.160$$

19. Interpret a 95% confidence interval for the slope of the regression line.
    We are 95% sure that the slope of the linear model to predict classes missed from the number of alcoholic drinks per week for all students at this college is between –0.036 and 0.160.

20. Compare the two p-values, from the test for correlation and the test for slope.
    They are about the same, as expected

21. What is the value of R^2 for this model?
    R^2 *=0.6%.*

22. Interpret it in context.

Only 0.6% of the variability in number of classes missed was explain by The number of alcoholic drinks per week