

# **MSIT 431 Probability and Statistical Methods**

Chapter 10 Inference for Regression

Dongning Guo

Fall 2017

## **Chapter 10 Inference for Regression**

**10.1 Simple Linear Regression**

**10.2 More Detail about Simple Linear Regression\***

## 10.1 Simple Linear Regression



- Statistical model for linear regression
- Simple linear regression model
- Estimating the regression parameters
- Confidence intervals and significance tests
- Confidence interval for mean response
- Prediction interval

### Introduction

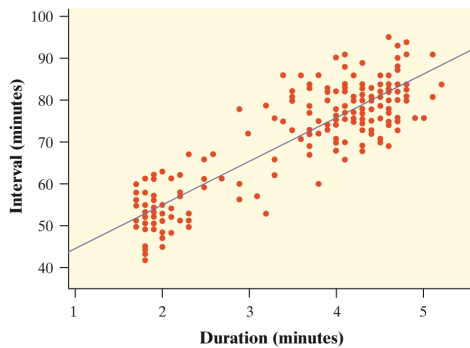
When a scatterplot shows a linear relationship between a quantitative explanatory variable  $x$  and a quantitative response variable  $y$ , we can use the least-squares line fitted to the data to predict  $y$  for a given value of  $x$ .

If the data are a random sample from a larger population, we need statistical inference to answer questions like these:

- ✓ Is there really a linear relationship between  $x$  and  $y$  in the population, or could the pattern we see in the scatterplot plausibly happen just by chance?
- ✓ What is the slope (rate of change) that relates  $y$  to  $x$  in the population, including a margin of error for our estimate of the slope?
- ✓ If least-squares regression line is used to predict  $y$  for a given value of  $x$ , how accurate is our prediction (again, with a margin of error)?

## Introduction

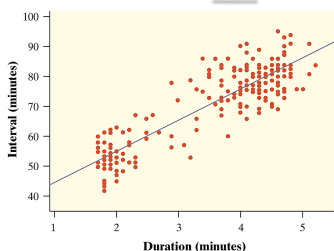
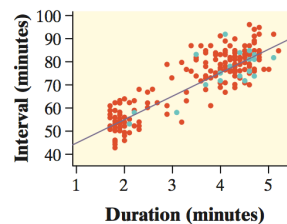
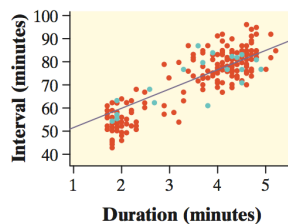
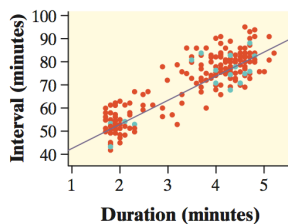
Researchers have collected data on eruptions of the Old Faithful geyser. Below is a scatterplot of the duration and interval of time until the next eruption for all 222 recorded eruptions in a single month. The least-squares regression line for this population of data has been added to the graph. It has slope 10.36 and y-intercept 33.97. Regarding all 222 eruptions as the population, this line is the **population regression line** (or true regression line) because it uses all the observations that month.



Suppose we take an SRS of 20 eruptions from the population and calculate the least-squares regression line  $\hat{y} = b_0 + b_1x$  for the sample data. How does the slope of the **sample regression line** (also called the estimated regression line, or LSRL) relate to the slope of the population regression line?

## Introduction

The figures below show the results of taking three different SRSs of 20 Old Faithful eruptions in this month. The green points in each graph are the selected points and the line is the LSRL for that sample of 20.



Notice that the slopes of the sample regression lines—10.2, 7.7, and 9.5—vary quite a bit from the slope of the population regression line, 10.36.

The pattern of variation in the slope  $b$  is described by its sampling distribution.

## Conditions for Regression Inference

- The slope and intercept of the least-squares line are *statistics* (calculated from the sample). They have sample variability. To do inference, think of  $b_0$  and  $b_1$  as estimates of unknown parameters  $\beta_0$  and  $\beta_1$  that describe the population of interest.

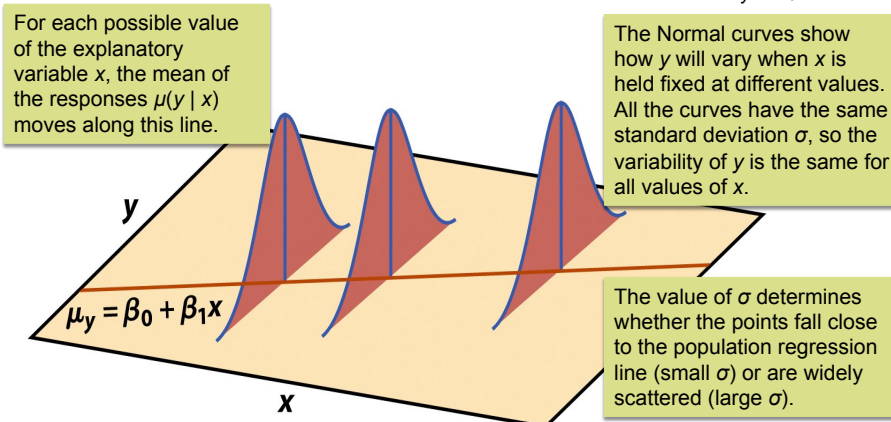
### Conditions for Regression Inference

We have  $n$  observations on an explanatory variable  $x$  and a response variable  $y$ . Our goal is to study or predict the behavior of  $y$  for given values of  $x$ .

- For any fixed value of  $x$ , the response  $y$  varies according to a **Normal distribution**. Repeated responses  $y$  are **independent** of each other.
- The mean response  $\mu_y$  has a **straight line relationship** with  $x$  given by a population regression line  $\mu_y = \beta_0 + \beta_1 x$ .
- The slope and intercept are unknown parameters.
- The standard deviation of  $y$  (call it  $\sigma$ ) is the same for all values of  $x$ . The value of  $\sigma$  is unknown.

## Conditions for Regression Inference

The figure below shows the regression model when the conditions are met. The line in the figure is the population regression line  $\mu_y = \beta_0 + \beta_1 x$ .



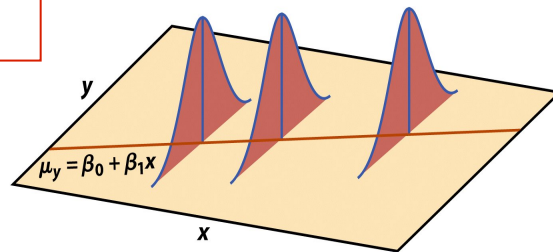
## Simple Linear Regression Model

In the population, the linear regression equation is  $\mu_y = \beta_0 + \beta_1 x$ .

Sample data fits **simple linear regression model**:

$$\begin{array}{lcl} \text{Data} = & \boxed{\text{Fit}} & + \boxed{\text{Error}} \\ y_i = & \boxed{(\beta_0 + \beta_1 x_i)} & + \boxed{(\varepsilon_i)} \end{array}$$

where the  $\varepsilon_i$  are **independent** and **Normally** distributed  $N(0, \sigma)$ .



Linear regression assumes **equal variance of y** ( $\sigma$  is the same for all values of  $x$ ).

## Estimating the Parameters

$$\mu_y = \beta_0 + \beta_1 x$$

The intercept  $\beta_0$ , the slope  $\beta_1$ , and the standard deviation  $\sigma$  of  $y$  are the unknown parameters of the regression model. We rely on the random sample data to provide unbiased estimates of these parameters.

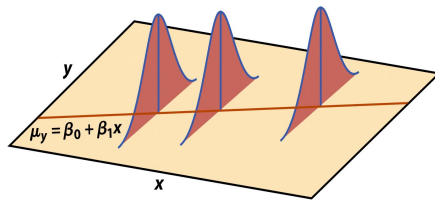
- The value of  $\hat{y}$  from the least-squares regression line is really a prediction of the mean value of  $y$  ( $\mu_y$ ) for a given value of  $x$ .
- The least-squares regression line ( $\hat{y} = b_0 + b_1 x$ ) obtained from sample data is the best estimate of the true population regression line ( $\mu_y = \beta_0 + \beta_1 x$ ).

$\hat{y}$  is an unbiased estimate for mean response  $\mu_y$ .

$b_0$  is an unbiased estimate for intercept  $\beta_0$ .

$b_1$  is an unbiased estimate for slope  $\beta_1$ .

## Estimating the Parameters



The **population standard deviation**  $\sigma$  for  $y$  at any given value of  $x$  represents the spread of the normal distribution of the  $\varepsilon_i$  around the mean  $\mu_y$ .

The **predicted values** are  $\hat{y}_i = b_0 + b_1 x_i$ ,  $i = 1, \dots, n$ , and the **residuals** are  $y_i - \hat{y}_i$ ,  $i = 1, \dots, n$ .

The **regression standard error**,  $s$ , for  $n$  sample data points is calculated from the **residuals** ( $y_i - \hat{y}_i$ ):

$$s = \sqrt{\frac{\sum \text{residual}^2}{n-2}} = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

$s$  is an essentially unbiased estimate of the regression standard deviation  $\sigma$ .

## Checking the Conditions

- You can fit a least-squares line to any set of explanatory-response data when both variables are quantitative. If the scatterplot doesn't show a roughly linear pattern, the fitted line may be almost useless.
- Before you can trust the results of inference, you must check the conditions for inference one by one.

- ✓ **The relationship is linear in the population.**
- ✓ **The response varies Normally about the population regression line.**
- ✓ **Observations are independent.**
- ✓ **The standard deviation of the responses is the same for all values of  $x$ .**

You can check all of the conditions for regression inference by looking at graphs of the residuals, or **residual plots**.

## Confidence interval for slope

The slope  $\beta_1$  of the population regression line  $\mu_y = \beta_0 + \beta_1 x$  is the rate of change of the mean response as the explanatory variable increases. The slope  $b_1$  of the sample regression line is our point estimate for  $\beta_1$ .

The confidence interval for  $\beta_1$  has the familiar form:

$$\text{Estimate} \pm t^* \cdot (\text{standard deviation of estimate})$$

Because we use the statistic  $b$  as our estimate, the confidence interval is:

$$b_1 \pm t^* SE_{b_1}$$

Confidence Interval for Regression Slope

A level  $C$  **confidence interval for the slope  $\beta_1$**  of the population regression line is:

$$b_1 \pm t^* SE_{b_1}$$

Here  $t^*$  is the critical value for the  $t$  distribution with  $df = n - 2$  having area  $C$  between  $-t^*$  and  $t^*$ .

## Significance test for regression slope

When the conditions for inference are met, we can perform a significance test to determine whether a specified value of  $\beta_1$  is plausible. The null hypothesis has the general form

$$H_0: \beta_1 = \text{hypothesized value.}$$

To do a test, standardize to get the test statistic:

$$\text{test statistic} = \frac{\text{statistic} - (\text{hypothesized value})}{\text{standard deviation of statistic}}$$

$$t = \frac{b_1 - \beta_1}{SE_{b_1}}$$

To find the  $P$ -value, use a  $t$  distribution with  $n - 2$  degrees of freedom. Here are the details for the  **$t$  test for the slope**.

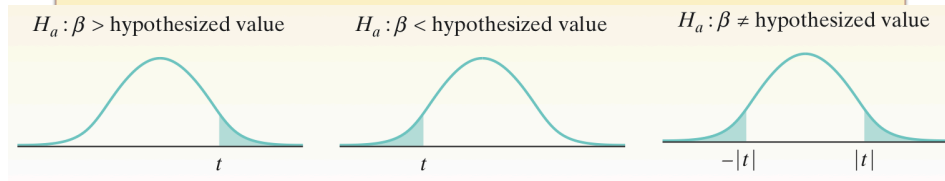
## Significance test for regression slope

### Significance Test for Regression Slope

To test the hypothesis  $H_0: \beta_1 = \text{hypothesized value}$ , compute the test statistic:

$$t = \frac{b_1 - \text{hypothesized value}}{SE_{b_1}}$$

Find the  $P$ -value by calculating the probability of getting a  $t$  statistic this large or larger in the direction specified by the alternative hypothesis  $H_a$ . Use the  $t$  distribution with  $df = n - 2$ .



## Testing the hypothesis of no relationship

We may look for evidence of a **significant relationship** between variables  $x$  and  $y$  in the population from which our data were drawn.

For that, we can test the hypothesis that the regression slope parameter  $\beta$  is equal to zero.

$$H_0: \beta_1 = 0 \text{ vs. } H_a: \beta_1 \neq 0$$

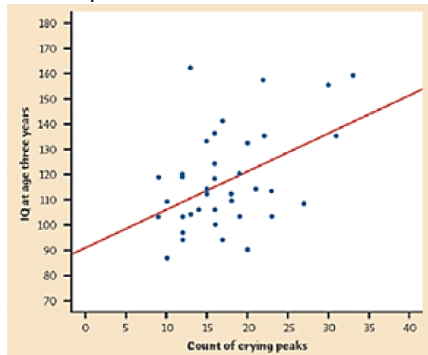
Testing  $H_0: \beta_1 = 0$  is equivalent to testing the **hypothesis of no correlation** between  $x$  and  $y$  in the population.

**Note:** A test of hypothesis for  $\beta_0$  is seldom of interest, mainly because  $\beta_0$  often has no practical interpretation.



## Example

Infants who cry easily may be more easily stimulated than others. This may be a sign of higher IQ. Child development researchers explored the relationship between the crying of infants 4 to 10 days old and their later IQ test scores. A snap of a rubber band on the sole of the foot caused the infants to cry. The researchers recorded the crying and measured its intensity by the number of peaks in the most active 20 seconds. They later measured the children's IQ at age three years using the Stanford-Binet IQ test. A scatterplot and Minitab output for the data from a random sample of 38 infants is below.



| Regression Analysis: IQ versus Crycount  |        |         |       |       |
|--|--------|---------|-------|-------|
| Predictor                                | Coef   | SE Coef | T     | P     |
| Constant                                 | 91.268 | 8.934   | 10.22 | 0.000 |
| Crycount                                 | 1.4929 | 0.4870  | 3.07  | 0.004 |
| S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5% |        |         |       |       |

**Do these data provide convincing evidence that there is a positive linear relationship between crying counts and IQ in the population of infants?**

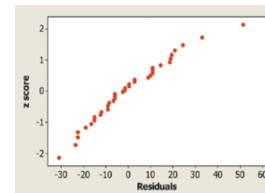
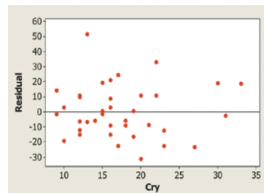
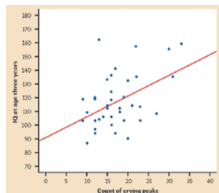
## Example: the test

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 > 0$$

$\beta_1$  is the true slope of the population regression line relating crying count to IQ.

- The scatterplot suggests a moderately positive linear relationship between crying peaks and IQ. The residual plot shows a random scatter of residuals about the line  $y = 0$ .



- IQ scores of individual infants should be independent.
- The Normal probability plot of the residuals shows a slight curvature, which suggests that the responses may not be Normally distributed about the line at each  $x$ -value. With such a large sample size ( $n = 38$ ), however, the  $t$  procedures are robust against departures from Normality.
- The residual plot shows a fairly equal amount of scatter around the horizontal line at 0 for all  $x$ -values.

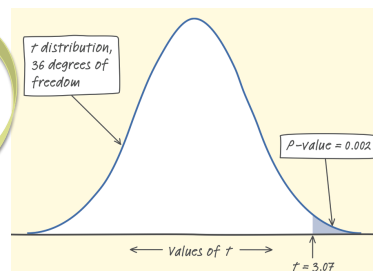
## Example

With no obvious violations of the conditions, we proceed to inference. The test statistic and  $P$ -value can be found in the Minitab output.

| Regression Analysis: IQ versus Crycount  |        |         |       |       |
|--|--------|---------|-------|-------|
| Predictor                                | Coef   | SE Coef | T     | P     |
| Constant                                 | 91.268 | 8.934   | 10.22 | 0.000 |
| Crycount                                 | 1.4929 | 0.4870  | 3.07  | 0.004 |
| S = 17.50 R-Sq = 20.7% R-Sq(adj) = 18.5% |        |         |       |       |

$$t = \frac{b_1}{SE_{b_1}} = \frac{1.4929}{0.4870} = 3.07$$

The Minitab output gives  $P = 0.004$  as the  $P$ -value for a two-sided test. The  $P$ -value for the one-sided test is half of this,  $P = 0.002$ .



The  $P$ -value, 0.002, is less than our  $\alpha = 0.05$  significance level, so we have enough evidence to reject  $H_0$  and conclude that there is a positive linear relationship between intensity of crying and IQ score in the population of infants.

## Confidence interval for mean response

We can calculate a confidence interval for the population mean  $\mu_y$  of all responses  $y$  when  $x$  takes the value  $x^*$  (within the range of data).

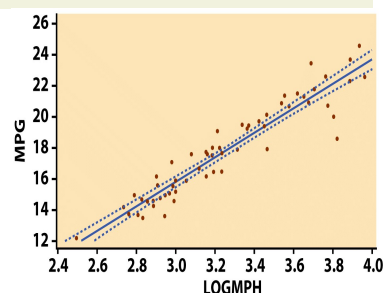
The **level  $C$  confidence interval for the mean response  $\mu_y$**  at a given value  $x^*$  of  $x$  is:

$$\hat{\mu}_y \pm t^* SE_{\hat{\mu}}$$

where  $t^*$  is the value such that the area under the  $t(n - 2)$  density curve between  $-t^*$  and  $t^*$  is  $C$ .

A separate confidence interval could be calculated for  $\mu_y$  along all the values that  $x$  takes.

Graphically, the series of confidence intervals is shown as a continuous interval on either side of  $\hat{y}$ .

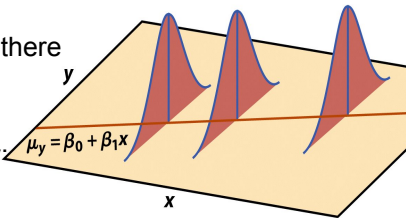


## Prediction Intervals

One use of regression is for **predicting** the value of  $y$  at some value of  $x$  within the range of data tested. Reliable predictions require statistical inference.

To estimate an *individual* response  $y$  for a given value of  $x$ , we use a **prediction interval**.

If we randomly sampled many times, there would be many different values of  $y$  obtained for a particular  $x$  following  $N(0, \sigma)$  around the mean response  $\mu_y$ .



## Prediction Intervals

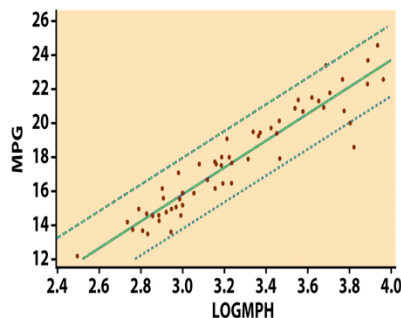
The **level C prediction interval for a single observation** on  $y$  when  $x$  takes the value  $x^*$  is:

$$\hat{y} \pm t^* SE_{\hat{y}}$$

$t^*$  is the critical value for the  $t(n-2)$  distribution with area  $C$  between  $-t^*$  and  $+t^*$ .

The prediction interval accounts for error in estimating  $\beta_0$  and  $\beta_1$  as well as uncertainty about the value of  $y$  being predicted.

Graphically, the series of prediction intervals is shown as a continuous interval on either side of  $\hat{y}$ . These intervals are wider than the corresponding confidence intervals for  $\mu_y$ .



## 10.2 More Detail About Simple Linear Regression\*



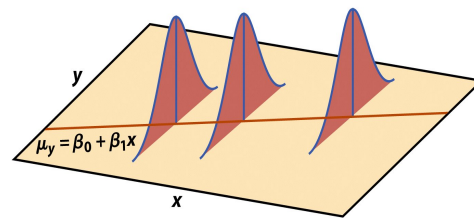
- Calculations for regression inference
- Inference for correlation

### Analysis of Variance for Regression

The regression model is:

$$\begin{aligned} \text{Data} &= \boxed{\text{fit}} + \boxed{\text{error}} \\ y_i &= \boxed{(\beta_0 + \beta_1 x_i)} + \boxed{(\varepsilon_i)} \end{aligned}$$

where the  $\varepsilon_i$  are **independent** and **Normally** distributed  $N(0, \sigma)$ , and  $\sigma$  is the same for all values of  $x$ .



It resembles an ANOVA, which also assumes equal variance, where

$$\begin{aligned} \text{SST} &= \boxed{\text{SS model}} + \boxed{\text{SS error}} \quad \text{and} \\ \text{DFT} &= \boxed{\text{DF model}} + \boxed{\text{DF error}} \end{aligned}$$

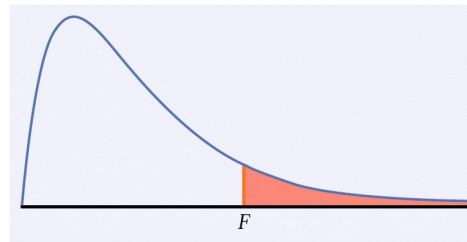
## The ANOVA $F$ Test

For a simple linear relationship, the ANOVA tests the hypotheses

$$H_0: \beta_1 = 0 \text{ versus } H_a: \beta_1 \neq 0$$

by comparing MSM (model) to MSE (error):  $F = \text{MSM}/\text{MSE}$

When  $H_0$  is true,  $F$  follows the  $F(1, n - 2)$  distribution. The  $P$ -value is  $P(F \geq f)$ .



The ANOVA test and the two-sided  $t$ -test for  $H_0: \beta_1 = 0$  yield the same  $P$ -value.

Software output for regression may provide  $t$ ,  $F$ , or both, along with the  $P$ -value.

## The ANOVA Table

| Source | Sum of squares SS                            | DF      | Mean square MS  | $F$       | $P$ -value          |
|--------|--|---------|-----------------|-----------|---------------------|
| Model  | $SSM = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$ | 1       | $MSM = SSM/DFM$ | $MSM/MSE$ | Tail area above $F$ |
| Error  | $SSE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$     | $n - 2$ | $MSE = SSE/DFE$ |           |                     |
| Total  | $SST = \sum_{i=1}^n (y_i - \bar{y})^2$       | $n - 1$ |                 |           |                     |

$$SST = SSM + SSE$$

$$DFT = DFM + DFE$$

$$F = MSM/MSE$$

The standard deviation,  $s$ , of the  $n$  residuals  $e_i = y_i - \hat{y}_i$ ,  $i = 1, \dots, n$ , is calculated from the following quantity:

$$s^2 = \frac{\sum e_i^2}{n - 2} = \frac{\sum (y_i - \hat{y}_i)^2}{n - 2} = \frac{SSE}{DFE} = MSE$$

$s$  is an approximately unbiased estimate of the regression standard deviation  $\sigma$ .

## Calculations for regression inference

To assess variation in the estimates of  $\beta_0$  and  $\beta_1$ , we calculate the standard errors for the estimated regression coefficients.

**The standard error of the slope estimate  $b_1$  is:**

$$SE_{b_1} = \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}}$$

**The standard error of the intercept estimate  $b_0$  is:**

$$SE_{b_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

## Calculations for regression inference

To estimate mean responses or predict future responses, we calculate the following standard errors:

**The standard error of the estimate of the mean response  $\mu_y$  is:**

$$SE_{\hat{\mu}} = s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

**The standard error for predicting an individual response  $y$  is:**

$$SE_{\hat{y}} = s \sqrt{1 + \frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_i - \bar{x})^2}}$$

## Inference for correlation

To test the null hypothesis of no linear association, we have the choice of also using the **correlation parameter  $\rho$** .

$$b_1 = r \frac{s_y}{s_x}$$

- When  $x$  is clearly the explanatory variable, this test is equivalent to testing the hypothesis  $H_0: \beta = 0$ .
- When there is no clear explanatory variable (e.g., arm length vs. leg length), a regression of  $x$  on  $y$  is no less legitimate than one of  $y$  on  $x$ . In that case, the correlation test of significance should be used.

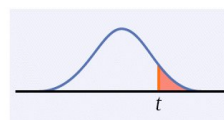
## Inference for correlation

The test of significance for  $\rho$  uses the one-sample  $t$ -test for:  $H_0: \rho = 0$ .

We compute the  $t$  statistic for sample size  $n$  and correlation coefficient  $r$ .

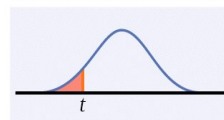
$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

$H_a: \rho > 0$  is  $P(T \geq t)$

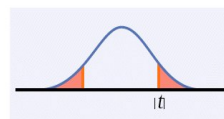


The  $P$ -value is the area under  $t(n-2)$  for values of  $T$  as or more extreme than  $t$  in the direction of  $H_a$ .

$H_a: \rho < 0$  is  $P(T \leq t)$



$H_a: \rho \neq 0$  is  $2P(T \geq |t|)$



## Inference for Correlation

When the hypothesis  $H_0: \rho = 0$  is rejected, it is safe to assume that there is some sort of relationship between the variables  $x$  and  $y$ .

Recall, though, from Chapter 2 that correlation measures *linear* relationships. As such it is not always a reliable indicator of *nonlinear* relationships.

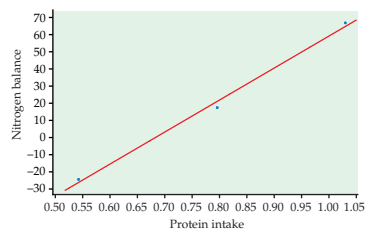
When the hypothesis  $H_0: \rho = 0$  is *not* rejected, do not assume that the variables are unrelated.

- ✓ First of all, it is possible that a Type II error may have occurred!
- ✓ Secondly, it is possible that  $x$  and  $y$  are related in a nonlinear way that the correlation coefficient  $r$  has no chance of detecting.
- ✓ A good way to investigate the second possibility is to examine a scatterplot.

### Example 10.16

- Nitrogen balance studies are used to determine protein requirements for people. Each subject is fed three different controlled diets during three separate experimental periods. The three diets are similar with regard to all nutrients except protein. Nitrogen balance is the difference between the amount of nitrogen consumed and the amount lost in feces and urine and by other means. Since virtually all of the nitrogen in a diet comes from protein, nitrogen balance is an indicator of the amount of protein retained by the body. The protein requirement for an individual is the intake corresponding to a balance of zero.
- Linear regression is used to model the relationship between nitrogen balance, measured in milligrams of nitrogen per kilogram of body weight per day (mg/kg/d), and protein intake, measured in grams of protein per kilogram of body weight per day (g/kg/d). Here are the data for one subject:

|                          |       |       |       |
|--------------------------|-------|-------|-------|
| Protein intake ( $x$ )   | 0.543 | 0.797 | 1.030 |
| Nitrogen balance ( $y$ ) | -23.4 | 17.8  | 67.3  |





## Linear model

| Variable  | Mean                 | Standard deviation | Correlation      | Sample size |
|-----------|----------------------|--------------------|------------------|-------------|
| Intake    | $\bar{x} = 0.79000$  | $s_x = 0.24357545$ | $r = 0.99698478$ | $n = 3$     |
| N balance | $\bar{y} = 20.56667$ | $s_y = 45.4132506$ |                  |             |

$$\begin{aligned}
 b_1 &= r \frac{s_y}{s_x} \\
 &= 0.99698478 \frac{45.4132506}{0.24357545} \\
 &= 185.882
 \end{aligned}$$

$$\begin{aligned}
 b_0 &= \bar{y} - b_1 \bar{x} \\
 &= 20.56667 - (185.882)(0.79000) \\
 &= -126.280
 \end{aligned}$$

$$\hat{y} = -126 + 186x$$

$$\begin{aligned}
 \hat{y}_1 &= b_0 + b_1 x_1 \\
 &= -126.280 + (185.882)(0.543) \\
 &= -25.346
 \end{aligned}$$

$$\begin{aligned}
 e_1 &= y_1 - \hat{y}_1 \\
 &= -23.4 - (-25.346) \\
 &= 1.946
 \end{aligned}$$

## Are x and y correlated?

$$\begin{aligned}
 s^2 &= \frac{\sum e_i^2}{n - 2} \\
 &= \frac{(1.946)^2 + (-4.068)^2 + (2.122)^2}{1} \\
 &= 24.84
 \end{aligned}$$

$$\begin{aligned}
 SE_{b_1} &= \frac{s}{\sqrt{\sum (x_i - \bar{x})^2}} \\
 &= \frac{4.984}{\sqrt{0.118658}} \\
 &= 14.47
 \end{aligned}$$

$$\begin{aligned}
 t &= \frac{b_1}{SE_{b_1}} \\
 &= \frac{185.882}{14.47} = 12.85
 \end{aligned}$$

$$P < 0.05.$$

$$\begin{aligned}
 b_1 \pm t^* SE_{b_1} &= 185.882 \pm (12.71)(14.47) \\
 &= 186 \pm 184
 \end{aligned}$$

## Prediction interval

$$\begin{aligned}\hat{\mu} &= b_0 + b_1 x_1 \\ &= -126.280 + (185.882)(0.7) \\ &= 3.837\end{aligned}$$

$$\begin{aligned}SE_{\hat{\mu}} &= s \sqrt{\frac{1}{n} + \frac{(x^* - \bar{x})^2}{\sum (x_j - \bar{x})^2}} \\ &= 4.984 \sqrt{\frac{1}{3} + \frac{(0.70 - 0.79)^2}{0.118658}} \\ &= 3.158\end{aligned}$$

$$\begin{aligned}\hat{\mu} \pm t^* SE_{\hat{\mu}} &= 3.837 \pm (12.71)(3.158) \\ &= 3.837 \pm 40.138 \\ &= 4 \pm 40\end{aligned}$$