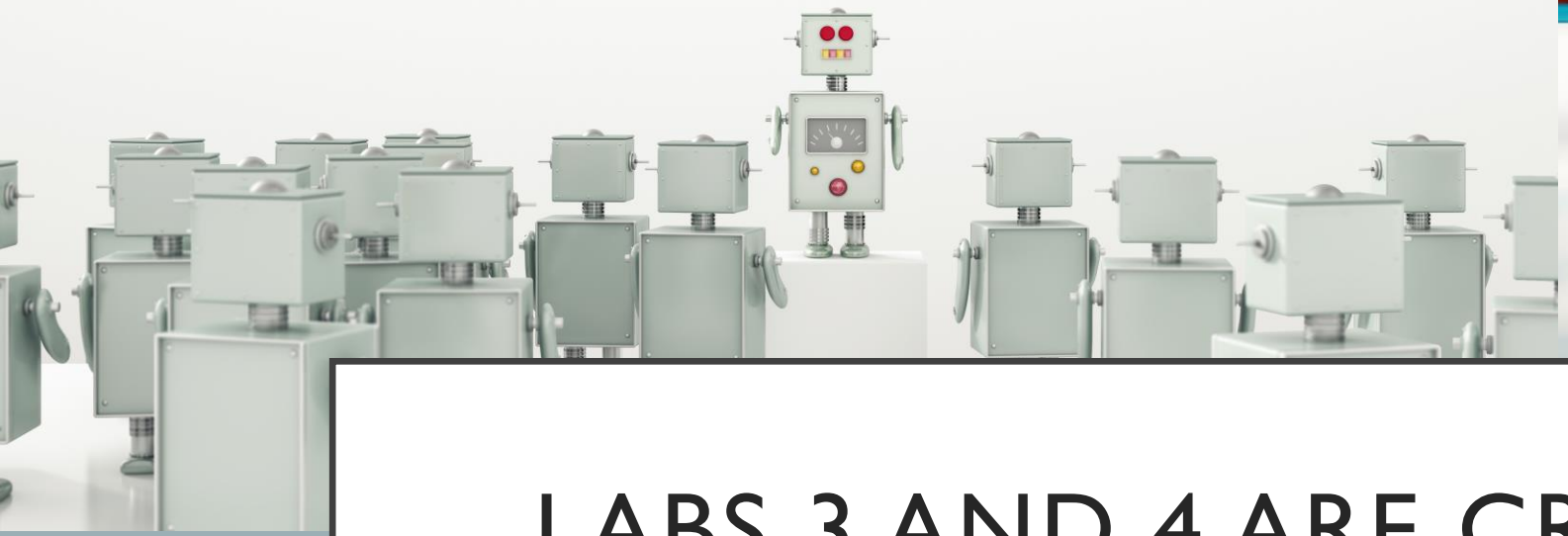
CHAPTER 10

Wayne Stewart



LABS 3 AND 4 ARE CRITICAL!!

Pvalue song


P-value
• Left side:
• P-value
Middle:
Getting so small
Right side:
Kick Ho
• Everyone:
• OUT the door

Ass –ump-TIONS

- I see a linear model
- I see assumptions in epsilon
- Independent Identically Distributed
- Normal Zero sigma squared.

X2

$$\varepsilon_i \overset{iid}{\sim} N(0, \sigma^2)$$

Supplementary Applied Exercises

10.72 *Quantum tunneling.* At temperatures approaching absolute zero (273 degrees below zero Celsius), helium exhibits traits that defy many laws of conventional physics. An experiment has been conducted with helium in solid form at various temperatures near absolute zero. The solid helium is placed in a dilution refrigerator along with a solid impure substance, and the proportion (by weight) of the impurity passing through the solid helium is recorded. (This phenomenon of solids passing directly through solids is known as *quantum tunneling*.) The data are given in the table.

HELIUM

Proportion of Impurity Passing Through Helium	Temperature
y	x, °C
.315	−262
.202	−265
.204	−256
.620	−267
.715	−270
.935	−272
.957	−272
.906	−272
.985	−273
.987	−273

a. Construct a scattergram of the data.
b. Find the least-squares line for the data and plot it on your scattergram.
c. Define β_1 in the context of this problem.
d. Test the hypothesis (at $\alpha = .05$) that temperature contributes no information for the prediction of the proportion of impurity passing through helium when a linear model is used. Draw the appropriate conclusions.
e. Find a 90% confidence interval for β_1. Interpret your results.
f. Find the coefficient of correlation for the given data.
g. Find the coefficient of determination for the linear model you constructed in part **b.** Interpret your result.
h. Find a 99% prediction interval for the proportion of impurity passing through helium when the temperature is set at −270°C.
i. Estimate the mean proportion of impurity passing through helium when the temperature is set at −270°C. Use a 99% confidence interval.

10.73 *Snow geese feeding trial.* Botanists at the University of Toronto conducted a series of experiments to investigate the feeding habits of baby snow geese. (*Journal of Applied Ecology*, Vol. 32, 1995.) Goslings were deprived of food until their guts were empty, then were allowed to feed for 6 hours on a diet of plants or Purina Duck Chow. For each feeding trial, the change in the weight of the gosling after 2.5 hours was recorded as a percentage of initial weight. Two other variables recorded were digestion efficiency (measured as a percentage) and amount of acid-detergent


002
He
HELIUM
4.00

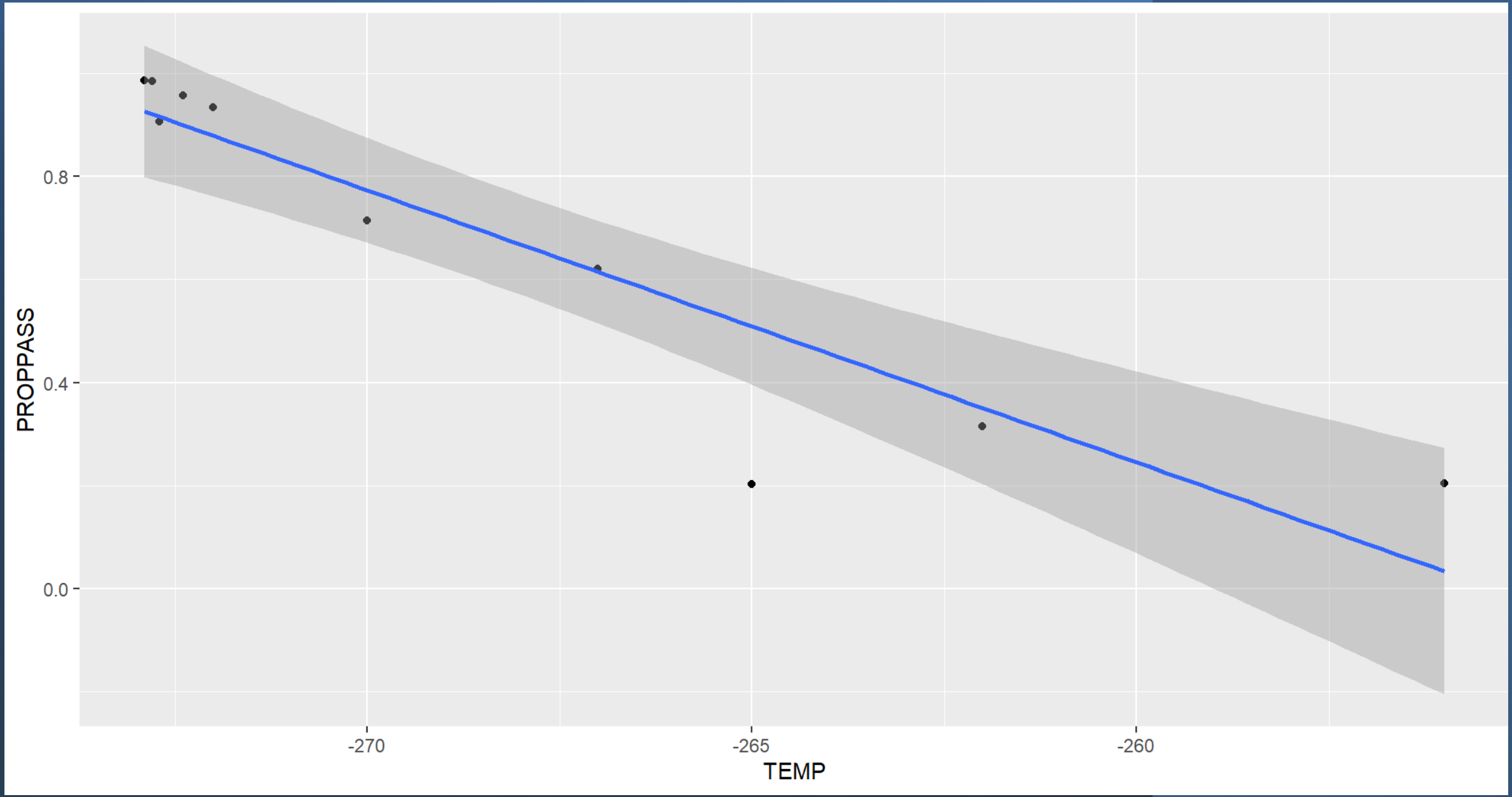
```
# 10.72
v <- Intro2R::myreadxl()
helium <- v$HELIUM
names(helium)
ylm <- lm(PROPPASS ~ TEMP, helium)

library(ggplot2)
g <- ggplot(helium, aes( x = TEMP, y = PROPPASS)) + geom_point()

g <- g + geom_smooth(method = "lm", formula = y ~ x)

g

summary(ylm)
```


PROPPASS
0.8
0.4
0.0
-270
-265
-260
TEMP

Summary information

```
> summary(ylm)

Call:
lm(formula = PROPPASS ~ TEMP, data = helium)

Residuals:
      Min        1Q    Median        3Q       Max
-0.30732 -0.02940  0.03045  0.05943  0.17014

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -13.490347   2.073772  -6.505 0.000187 ***
TEMP         -0.052829   0.007728  (-6.836)^2 0.000133 *
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1333 on 8 degrees of freedom
Multiple R-squared:  0.8538,    Adjusted R-squared:  0.8356
F-statistic: 46.73 on 1 and 8 DF,  p-value: 0.0001329
```

You must know how to interpret ALL of this output!!

?

$\hat{\sigma}$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{SS_{XX}}\right)$$

$$\hat{\beta}_0 \sim N\left(\beta_0, \frac{\frac{\sigma^2}{n}\sum x_i^2}{SS_{XX}}\right)$$

$$\widehat{\sigma^2} = S^2 = \frac{RSS}{n-2}$$

F STATISTIC
LAST LINE OF SUMMARY

Source	SS	df	MS	F	Pvalue
Regression	SSM	1	$\frac{SSM}{1}$	$\frac{MS_M}{MS_R}$	Use pf()
Residual error	SSR	n-2	$\frac{SSR}{n-2}$		
Total	SST	n-1			

WHAT WE NEED TO COVER

CONTENTS

SOME DEFINITIONS

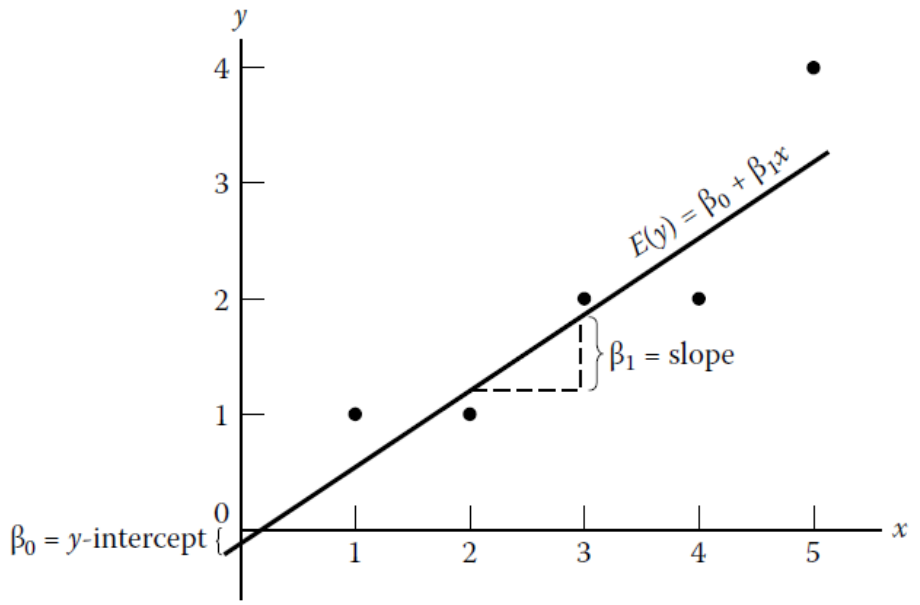
Definition 10.1

The variable to be predicted (or modeled), y, is called the **dependent** (or **response**) **variable**.

Definition 10.2

The variables used to predict (or model) y are called **independent variables** and are denoted by the symbols x_1, x_2, x_3, etc.

FIGURE 10.2
A graph of the data points of Table 10.1 and the hypothetical line of means, $E(y) = \beta_0 + \beta_1 x$



THE MODEL

A Simple Linear Regression (Probabilistic) Model

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

where

y = Dependent variable

x = Independent variable

$E(y) = \beta_0 + \beta_1 x$ is the deterministic component (the equation of a straight line)

ε (epsilon) = **Random error** component

β_0 (beta-zero) = ***y*-intercept** of the line, i.e., point at which the line intercepts or cuts through the y-axis (see Figure 10.2)

β_1 (beta-one) = **Slope** of the line, i.e., amount of increase (or decrease) in the deterministic component y for every 1 unit increase in x (see Figure 10.2)

ASSUMPTIONS

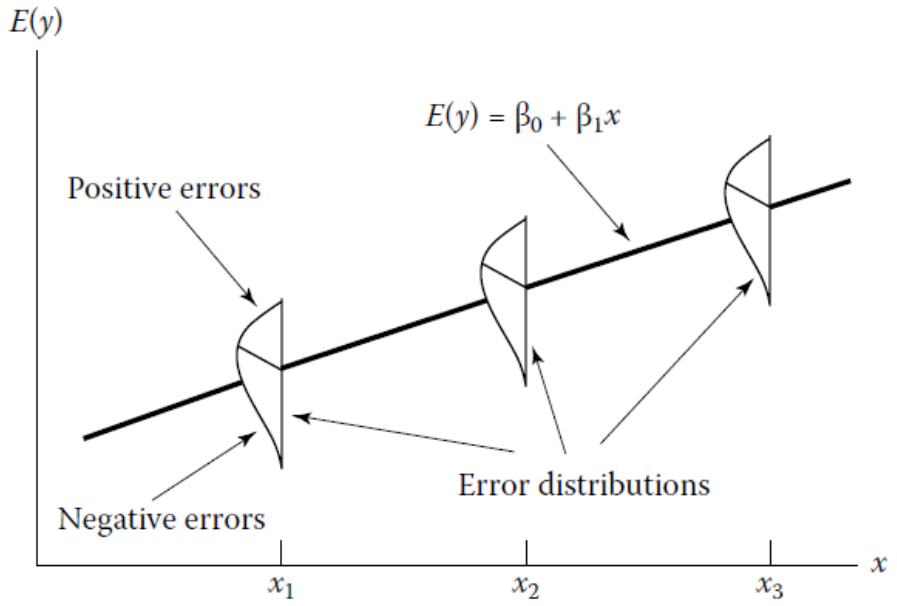
Assumption 1 The mean of the probability distribution of ε is 0. That is, the average of the errors over an infinitely long series of experiments is 0 for each setting of the independent variable x. This assumption implies that the mean value of y, $E(y)$, for a given value of x is $E(y) = \beta_0 + \beta_1 x$.

Assumption 2 The variance of the probability distribution of ε is constant for all settings of the independent variable x. For our straight-line model, this assumption means that the variance of ε is equal to a constant, say, σ^2, for all values of x.

Assumption 3 The probability distribution of ε is normal.

Assumption 4 The errors associated with any two different observations are independent. That is, the error associated with one value of y has no effect on the errors associated with other y values.

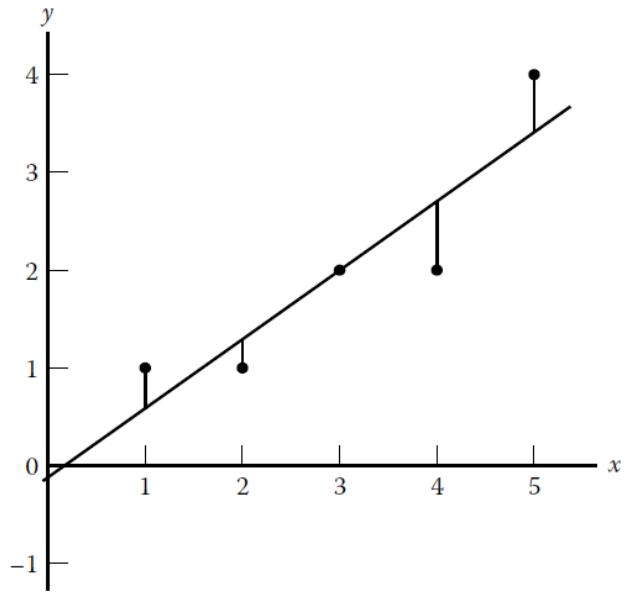
FIGURE 10.3
The probability distribution of ε



ERROR DISTRIBUTIONS

RESIDUALS

FIGURE 10.4
Graph showing the deviations of the points about a line



NOTICE THE BOOK USES SSE NOT SSR

Definition 10.3

A regression **residual** $\hat{\epsilon}$ is defined as the difference between an observed y value and its corresponding predicted value:

$$\hat{\epsilon} = y - \hat{y}$$

Definition 10.4

The **least-squares line** is one that has a smaller SSE than any other straight-line model.

The values of $\hat{\beta}_0$ and $\hat{\beta}_1$ that minimize

$$\text{SSE} = \sum_{i=1}^{n} [y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)]^2$$

Theoretical Exercises

10.19 Show that

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \sum\left[\frac{1}{n} - \frac{\bar{x}(x_i - \bar{x})}{SS_{xx}}\right]y_i$$

[*Hint*: Note that

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{SS_{xx}}$$

$$= \frac{\sum(x_i - \bar{x})y_i}{SS_{xx}} - \frac{\bar{y}\sum(x_i - \bar{x})}{SS_{xx}}$$

$$= \frac{\sum(x_i - \bar{x})y_i}{SS_{xx}}$$

since $\sum(x_i - \bar{x}) = 0$.]

10.20 We showed in Example 10.2 that $\hat{\beta}_1$, the least-squares estimator of the slope β_1, is an unbiased estimator of β_1, i.e., $E(\hat{\beta}_1) = \beta_1$. Use the result from Exercise 10.17 to show that $E(\hat{\beta}_0) = \beta_0$.

10.21 In Exercise 10.19, you showed that $\hat{\beta}_0$ could be written as a linear function of independent random variables. Use Theorem 6.8 to show that

$$V(\hat{\beta}_0) = \frac{\sigma^2}{n}\left(\frac{\sum x_i^2}{SS_{xx}}\right)$$

CAN YOU PROVE THESE?

THE FOLLOWING THEOREM IS PROVED IN MATH 4773 WHERE LA IS USED EXTENSIVELY

$$S^2 = \frac{RSS}{n-2}$$

THEOREM 10.1

THEOREM 10.1

Let $s^2 = \text{SSE}/(n-2)$. Then, when the assumptions of Section 10.2 are satisfied, the statistic

$$\chi^2 = \frac{\text{SSE}}{\sigma^2}$$

$$= \frac{(n-2)s^2}{\sigma^2}$$

possesses a chi-square distribution with $\nu = (n-2)$ degrees of freedom.

S^2

Estimation of σ^2

$$s^2 = \frac{\text{SSE}}{\text{Degrees of freedom for error}} = \frac{\text{SSE}}{n-2}$$

where

$$\text{SSE} = \sum (y_i - \hat{y}_i)^2 = \text{SS}_{yy} - \hat{\beta}_1 \text{SS}_{xy}$$

$$\text{SS}_{yy} = \sum (y_i - \bar{y})^2 = \sum y_i^2 - \frac{\left(\sum y_i\right)^2}{n}$$

Warning: When performing these calculations, you may be tempted to round the calculated values of SS_{yy}, β_1, and SS_{xy}. Be certain to carry at least six significant figures for each of these quantities to avoid substantial errors in the calculation of SSE.

INTERPRETATION OF S

Interpretation of s, the Estimated Standard Deviation of ε

We expect most of the observed y values to lie within $2s$ of their respective least-squares predicted values, $\hat{y}$.

Sampling Distribution of $\hat{\beta}_1$

If we make the four assumptions about ε (see Section 10.2), then the sampling distribution of $\hat{\beta}_1$, the least-squares estimator of slope, will be a normal distribution with mean β_1 (the true slope) and standard error

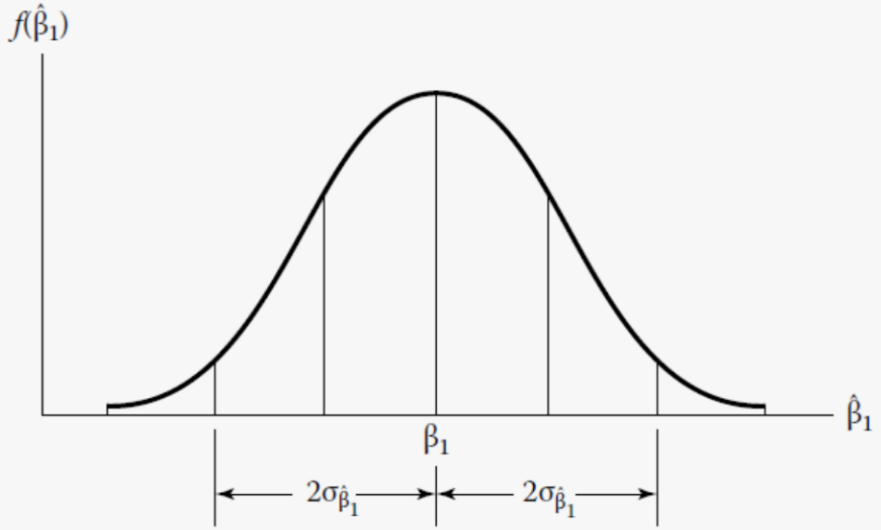
$$\sigma_{\hat{\beta}_1} = \frac{\sigma}{\sqrt{SS_{xx}}} \approx \frac{s}{\sqrt{SS_{xx}}} \qquad \text{(see Figure 10.8)}$$

Since σ will usually be unknown, the appropriate test statistic will generally be a Student's T statistic formed as follows:

$$T = \frac{\hat{\beta}_1 - \text{Hypothesized value of } \beta_1}{s_{\hat{\beta}_1}} \qquad \text{where } s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

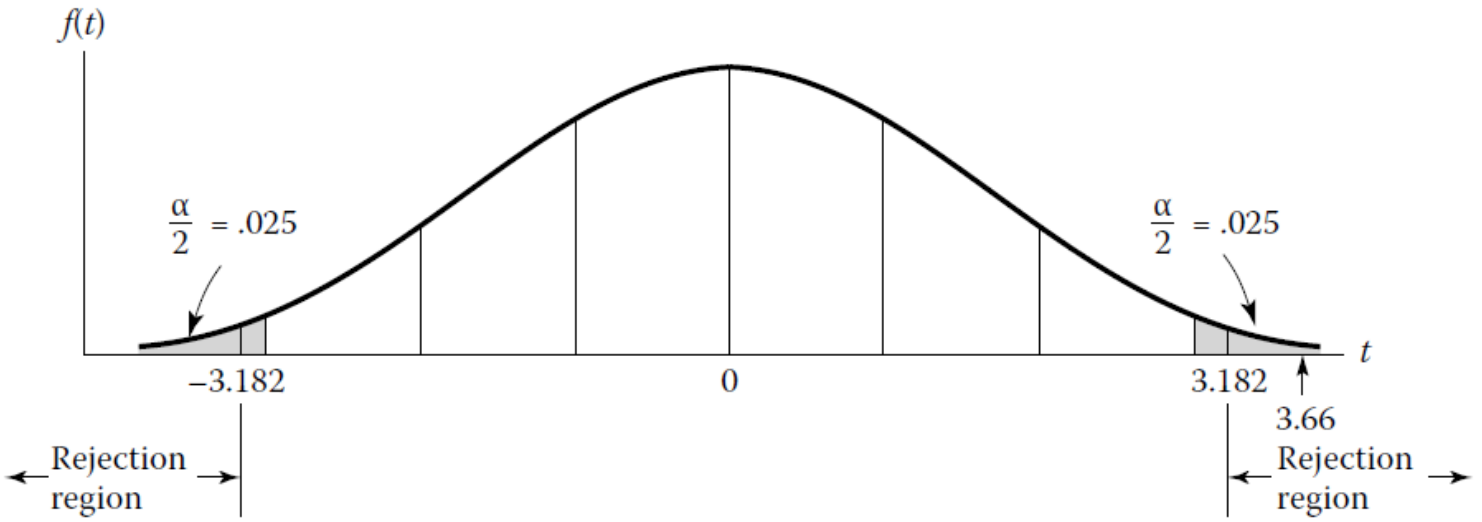
$$= \frac{\hat{\beta}_1 - 0}{s/\sqrt{SS_{xx}}}$$

FIGURE 10.8
Sampling distribution of $\hat{\beta}_1$



RAR MAN!!

FIGURE 10.9
Rejection region and calculated t value for testing whether the slope $\beta_1 = 0$



CI FOR β_1

A $(1-\alpha)100\%$ Confidence Interval for the Slope β_1

$$\hat{\beta}_1 \pm t_{\alpha/2} s_{\hat{\beta}_1} \qquad \text{where} \quad s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

and $t_{\alpha/2}$ is based on $(n-2)$ df

$$\hat{\beta}_1 \pm t_{.025} s_{\hat{\beta}_1} = .7 \pm 3.182\left(\frac{s}{\sqrt{SS_{xx}}}\right)$$

$$= .7 \pm 3.182\left(\frac{.61}{\sqrt{10}}\right) = .7 \pm .61 = (.09, 1.31)$$

TEDIOUS CALCULATIONS DONE BY R THROUGH lm()

CORRELATION

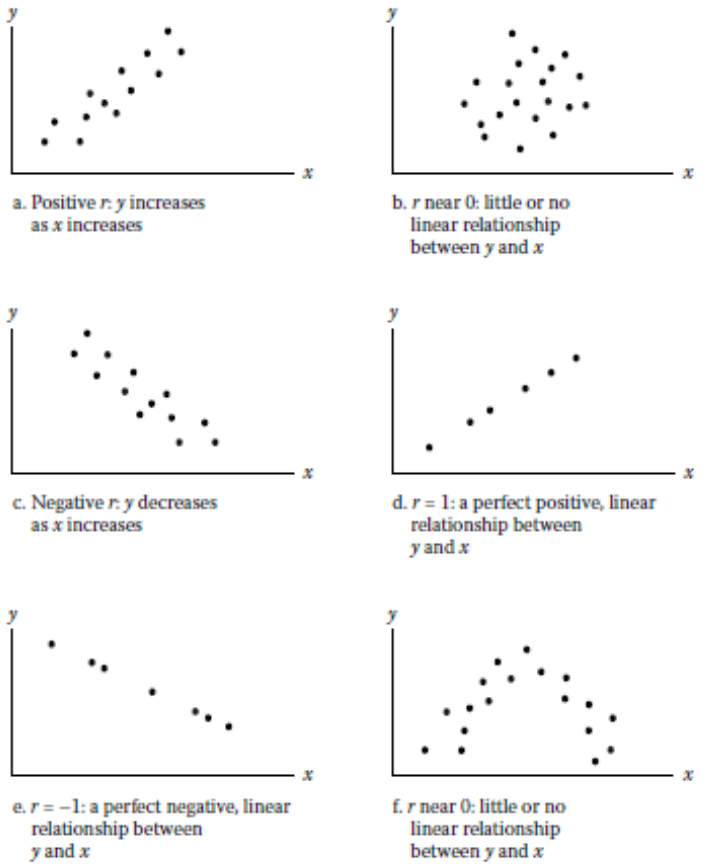
Definition 10.5

The **Pearson product moment coefficient of correlation *r*** is a measure of the strength of the linear relationship between two variables *x* and *y* in the sample. It is computed (for a sample of *n* measurements on *x* and *y*) as follows:

$$r = \frac{SS_{xy}}{\sqrt{SS_{xx}SS_{yy}}}$$

LINEAR ASSOCIATION

FIGURE 10.11
Values of r and their implications



a. Positive r: y increases as x increases

b. r near 0: little or no linear relationship between y and x

c. Negative r: y decreases as x increases

d. $r = 1$: a perfect positive, linear relationship between y and x

e. $r = -1$: a perfect negative, linear relationship between y and x

f. r near 0: little or no linear relationship between y and x

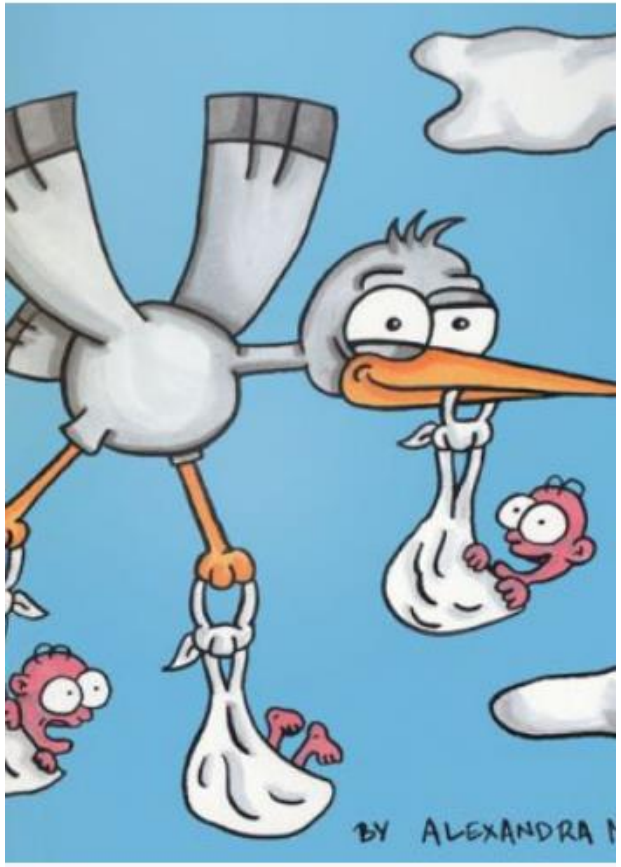
Warning

High correlation does not imply causality. If a large positive or negative value of the sample correlation coefficient r is observed, it is incorrect to conclude that a change in x *causes* a change in y. The only valid conclusion is that a linear trend *may* exist between x and y.

CORRELATION ≠ CAUSALITY

STORKS DELIVER BABIES?

- Pioneering statistician George Udny Yule, author of the seminal 1911 textbook *Introduction to the Theory of Statistics*, explained confounding factors with a pleasing reference to reproduction. He noted that in Alsatian villages numbers of human newborns are correlated with numbers of storks nesting locally. It is tempting to conclude that storks do actually deliver babies, but the real explanation is far more mundane. Larger villages have more houses with chimneys for storks to build nests, and more babies are of course delivered in larger villages. The confounding factor is village size.



cor.test() in R

Test of Hypothesis for Linear Correlation

One-Tailed Test

H_0: $\rho = 0$

H_a: $\rho > 0$ (or $\rho < 0$)

Two-Tailed Test

H_0: $\rho = 0$

H_a: $\rho \neq 0$

Test statistic: $T_c = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Rejection region: $T_c > t_\alpha$ (or $T_c < -t_\alpha$)

Rejection region: $|T_c| > t_{\alpha/2}$

p-value: $P(T > T_c)$ [or $P(T < T_c)$]

p-value: $2P(T > |T_c|)$

where t_α and $t_{\alpha/2}$ are the critical values based on $(n-2)$df obtained from Table 7 of Appendix B.

Assumptions: The sample of (x, y) values is randomly selected from a (bivariate) normal population.*

Definition 10.6

The **coefficient of determination** is

$$r^2 = \frac{SS_{yy} - SSE}{SS_{yy}} = 1 - \frac{SSE}{SS_{yy}}$$

It represents the proportion of the sum of squares of deviations of the y values about their predicted values ($\hat{y}$) that can be attributed to a linear relation between y and x. (In simple linear regression, it may also be computed as the square of the coefficient of correlation r.)

COEFFICIENT OF DETERMINATION OR MULTIPLE R^2

Note that r^2 is always between 0 and 1, because r is between -1 and $+1$. Thus, $r^2 = .60$ means that the sum of squares of deviations of the y values about their predicted values has been reduced 60% by the use of $\hat{y}$, instead of $\overline{y}$, to predict y. Or, more practically, $r^2 = .60$ implies that the straight-line model relating y to x can explain (or account for) 60% of the variation present in the sample of y values.

SIZE

Practical Interpretation of the Coefficient of Determination, r^2

About $100(r^2)\%$ of the total sum of squares of deviations of the sample y values about their mean $\bar{y}$ can be explained by (or attributed to) using x to predict y in the straight-line model.

PRACTICAL INTERPRETATION

ADJUSTED $R^2 = R_a^2$ HERE $k = 1$

The Adjusted Multiple Coefficient of Determination

The **adjusted multiple coefficient of determination** is given by

$$R_a^2 = 1 - \frac{(n-1)}{n-(k+1)}\left(\frac{\text{SSE}}{\text{SS}_{yy}}\right) = 1 - \frac{n-1}{n-(k+1)}(1-R^2)$$

Unlike R^2, R_a^2 takes into account ("adjusts for") both the sample size n and the number of β parameters in the model. R_a^2 will always be smaller than R^2, and more importantly, cannot be "forced" to 1 by simply adding more and more independent variables to the model. Consequently, some analysts prefer the more conservative R_a^2 when choosing a measure of model adequacy.

Introduction

There is a very important identity we use in regression called the “Sum of Squares” identity.

This identity is used in many different contexts and can be rewritten to explain ANOVA for example.

See https://en.wikipedia.org/wiki/Linear_least_squares#Derivation_of_the_normal_equations for more detail.

$$TSS = MSS + RSS$$

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

where TSS is the total sum of squares, MSS is the model sum of squares and RSS is the residual sum of squares.

Please note that many texts use slightly different notation – for example MS uses ESS for RSS.

Proof

To prove this we will need the normal equations. These are derived from taking two partial derivatives, $\frac{\partial RSS}{\partial \hat{\beta}_0}$ and $\frac{\partial RSS}{\partial \hat{\beta}_1}$ and forming two simultaneous equations by setting each to zero.

$$n\hat{\beta}_0 + n\bar{x}\hat{\beta}_1 - n\bar{y} = 0$$

$$n\hat{\beta}_0\bar{x} + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 - \sum_{i=1}^{n} x_i y_i = 0$$

We will approach this problem by turning the left hand side of the SS identity into the RHS.

$$\begin{aligned} TSS &= \sum_{i=1}^{n}(y_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\ &= RSS + MSS + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \end{aligned}$$

To prove the result we need only show that the cross product term is zero.

The following proof is specific to SLR but we can approach this differently and make a more general proof by using matrix methods.

To simplify the last term we will need a couple of results

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$$

also,

$$\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$$

$$\begin{aligned}
\sum_{i=1}^{n}(y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) &= \sum_{i=1}^{n}(y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i))(\hat{\beta}_0 + \hat{\beta}_1 x_i - (\hat{\beta}_0 + \hat{\beta}_1 \bar{x})) \\
&= \hat{\beta}_1 \sum_{i=1}^{n}((y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i)))(x_i - \bar{x}) \\
&= \hat{\beta}_1 \sum_{i=1}^{n}(y_i x_i - y_i \bar{x} - (\hat{\beta}_0 + \hat{\beta}_1 x_i)(x_i - \bar{x})) \\
&= \hat{\beta}_1 \sum_{i=1}^{n}(y_i x_i - y_i \bar{x} - \hat{\beta}_0 x_i + \hat{\beta}_0 \bar{x} - \hat{\beta}_1 x_i^2 + \hat{\beta}_1 x_i \bar{x}) \\
&= \hat{\beta}_1 \left\{ \sum_{i=1}^{n} x_i y_i - n\bar{x}\bar{y} - n\bar{x}\hat{\beta}_0 + n\hat{\beta}_0 \bar{x} - \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 + n\hat{\beta}_1 \bar{x}^2 \right\} \\
&= 0
\end{aligned}$$

The last line follows by the normal equations.

Theoretical Exercises

10.51 Verify that

$$\hat{\beta}_1 = r\sqrt{\frac{SS_{yy}}{SS_{xx}}} \quad \text{and} \quad SSE = SS_{yy}(1 - r^2)$$

10.52 Use the result of Exercise 10.51 to show that

$$\frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}} = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$$

cor.test() in R

Test of Hypothesis for Linear Correlation

One-Tailed Test

H_0: $\rho = 0$

H_a: $\rho > 0$ (or $\rho < 0$)

Two-Tailed Test

H_0: $\rho = 0$

H_a: $\rho \neq 0$

Test statistic: $T_c = \dfrac{r\sqrt{n-2}}{\sqrt{1-r^2}}$

Rejection region: $T_c > t_\alpha$ (or $T_c < -t_\alpha$)

Rejection region: $|T_c| > t_{\alpha/2}$

p-value: $P(T > T_c)$ [or $P(T < T_c)$]

p-value: $2P(T > |T_c|)$

where t_α and $t_{\alpha/2}$ are the critical values based on $(n-2)$df obtained from Table 7 of Appendix B.

Assumptions: The sample of (x, y) values is randomly selected from a (bivariate) normal population.*

Sampling Errors for the Estimator of the Mean of *y*, *E(y)*, and the Predictor for an Individual *y*

1. The standard deviation of the sampling distribution of the estimator $\hat{y}$ of $E(y)$ at a particular value of x, say, x_p, is

$$\sigma_{\hat{y}} = \sigma\sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

   where σ is the standard deviation of the random error ε.

2. The standard deviation of the prediction error for the predictor $\hat{y}$ of an individual y value for $x = x_p$ is

$$\sigma_{(y-\hat{y})} = \sigma\sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

   where σ is the standard deviation of the random error ε.

ESTIMATION AND PREDICTION

A $(1-\alpha)100\%$ Confidence Interval for the Mean Value of y, $E(y)$, for $x = x_p$

$$\hat{y} \pm t_{\alpha/2}(\text{Estimated standard deviation of } \hat{y})$$

or

$$\hat{y} \pm t_{\alpha/2} s \sqrt{\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on $(n-2)$ df

A $(1-\alpha)100\%$ Prediction Interval for an Individual y for $x = x_p$

$$\hat{y} \pm t_{\alpha/2}[\text{Estimated standard deviation of } (y - \hat{y})]$$

or

$$\hat{y} \pm t_{\alpha/2} s \sqrt{1 + \frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}}}$$

where $t_{\alpha/2}$ is based on $(n - 2)$ df

PREDICTION INTERVAL

Example 10.11

Deriving $v(\hat{y})$

Find the variance of $\hat{y}$ when $x = x_p$.

Solution

When $x = x_p$, we have $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_p$, where $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$. Substituting this value of $\hat{\beta}_0$ into the expression for $\hat{y}$, we obtain

$$\begin{aligned}\hat{y} &= (\bar{y} - \hat{\beta}_1 \bar{x}) + \hat{\beta}_1 (x_p) \\ &= \bar{y} + \hat{\beta}_1 (x_p - \bar{x})\end{aligned}$$

The next step is to express $\hat{y}$ as a linear function of the random y values, $y_1, y_2, \ldots, y_n$ so that we can obtain $V(\hat{y})$ as the variance of a linear function of independent random variables. We now write

$$\begin{aligned}\hat{y} &= \bar{y} + \hat{\beta}_1 (x_p - \bar{x}) \\ &= \sum \frac{y_i}{n} + \frac{(x_p - \bar{x})}{SS_{xx}} \sum (x_i - \bar{x}) y_i \\ &= \sum \frac{y_i}{n} + \sum \frac{(x_p - \bar{x})(x_i - \bar{x})}{SS_{xx}} y_i\end{aligned}$$

EXAMPLE 10.11

We can now express $\hat{y}$ as a single summation:

$$\hat{y} = \sum \left[\frac{1}{n} + \frac{(x_p - \bar{x})(x_i - \bar{x})}{SS_{xx}} \right] y_i$$

i.e., $\hat{y}$ is a linear function of the independent random variables, $y_1, y_2, \ldots, y_n$, where the coefficient of y_i is

$$\left[\frac{1}{n} + \frac{(x_p - \bar{x})(x_i - \bar{x})}{SS_{xx}} \right]$$

Then, by Theorem 6.8,

$$V(\hat{y}) = \sum \left[\frac{1}{n} + \frac{(x_p - \bar{x})(x_i - \bar{x})}{SS_{xx}} \right]^2 V(y_i)$$

where $V(y_i) = \sigma^2, i = 1, 2, \ldots, n$. Therefore,

$$V(\hat{y}) = \sum \left[\frac{1}{n^2} + \frac{2}{n} \frac{(x_p - \bar{x})(x_i - \bar{x})}{SS_{xx}} + \frac{(x_p - \bar{x})^2 (x_i - \bar{x})^2}{(SS_{xx})^2} \right] \sigma^2$$

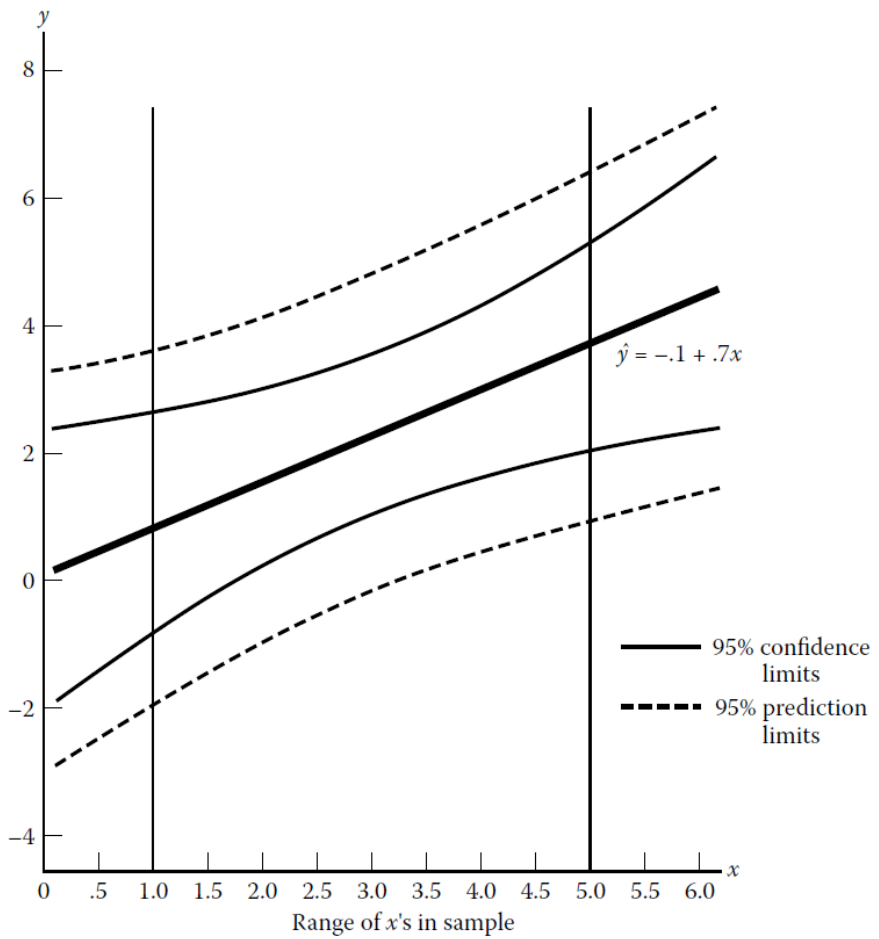
$$= \left[\frac{n}{n^2} + \frac{2}{n} \frac{(x_p - \bar{x})}{SS_{xx}} \sum (x_i - \bar{x}) + \frac{(x_p - \bar{x})^2}{(SS_{xx})^2} \sum (x_i - \bar{x})^2 \right] \sigma^2$$

$$= \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{(SS_{xx})^2} SS_{xx} \right] \sigma^2 \qquad \text{since } \sum (x_i - \bar{x}) = 0$$

$$= \sigma^2 \left[\frac{1}{n} + \frac{(x_p - \bar{x})^2}{SS_{xx}} \right]$$

You can see that this agrees with the formula for $V(\hat{y})$ given previously in this section.

FIGURE 10.19
Comparison of widths of 95% confidence and prediction intervals



OUTSIDE THE RANGE – WATCH OUT!!

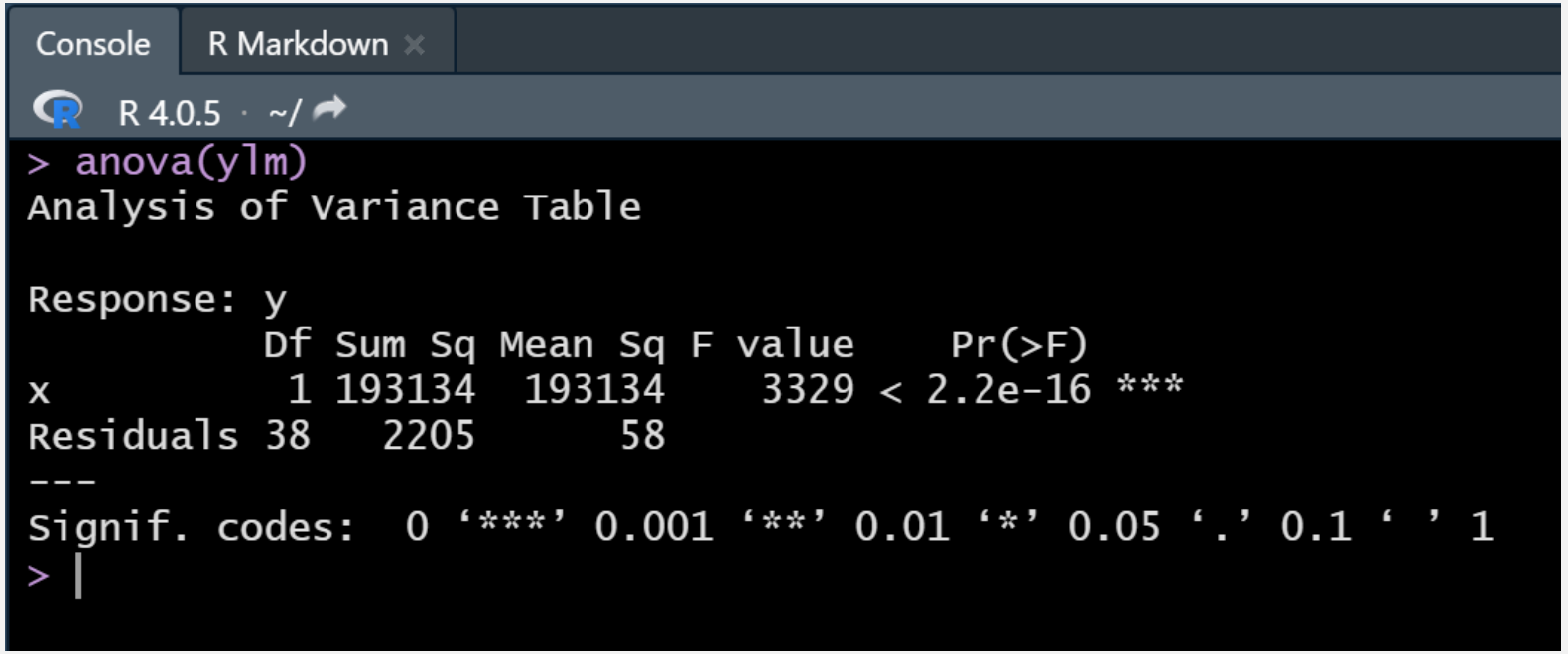
Warning

Using the least-squares prediction equation to estimate the mean value of y or to predict a particular value of y for values of x that fall *outside* the range of values of x contained in your sample data may lead to errors of estimation or prediction that are much larger than expected. Although the least-squares model may provide a very good fit to the data over the range of x values contained in the sample, **it could give a poor representation of the true model for values of x outside this region**.

F-STAT

Source	SS	df	MS	F	Pvalue
Regression	SSM	1	$\frac{SSM}{1}$	$\frac{MS_M}{MS_R}$	Use pf()
Residual error	SSR	n-2	$\frac{SSR}{n-2}$		
Total	SST	n-1			

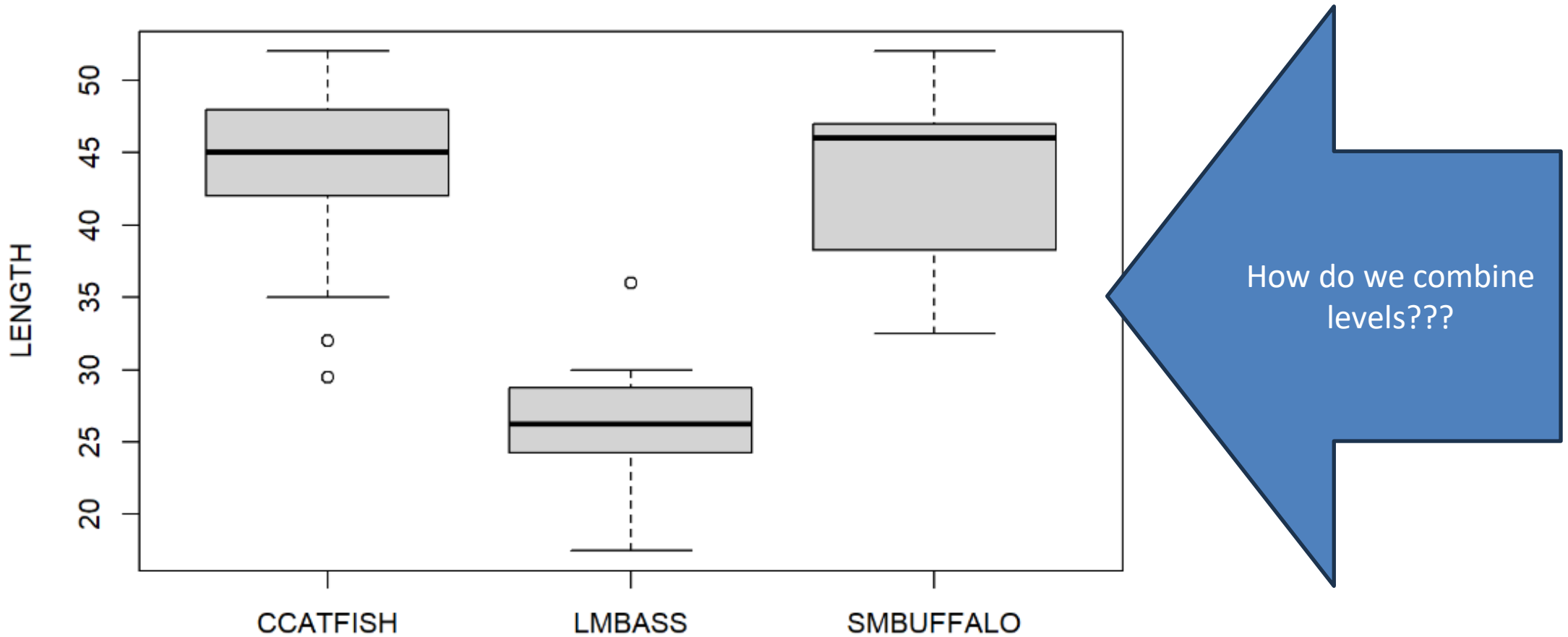
IN R, $H_0: \beta_1 = 0$



Steps to Follow in a Simple Linear Regression Analysis

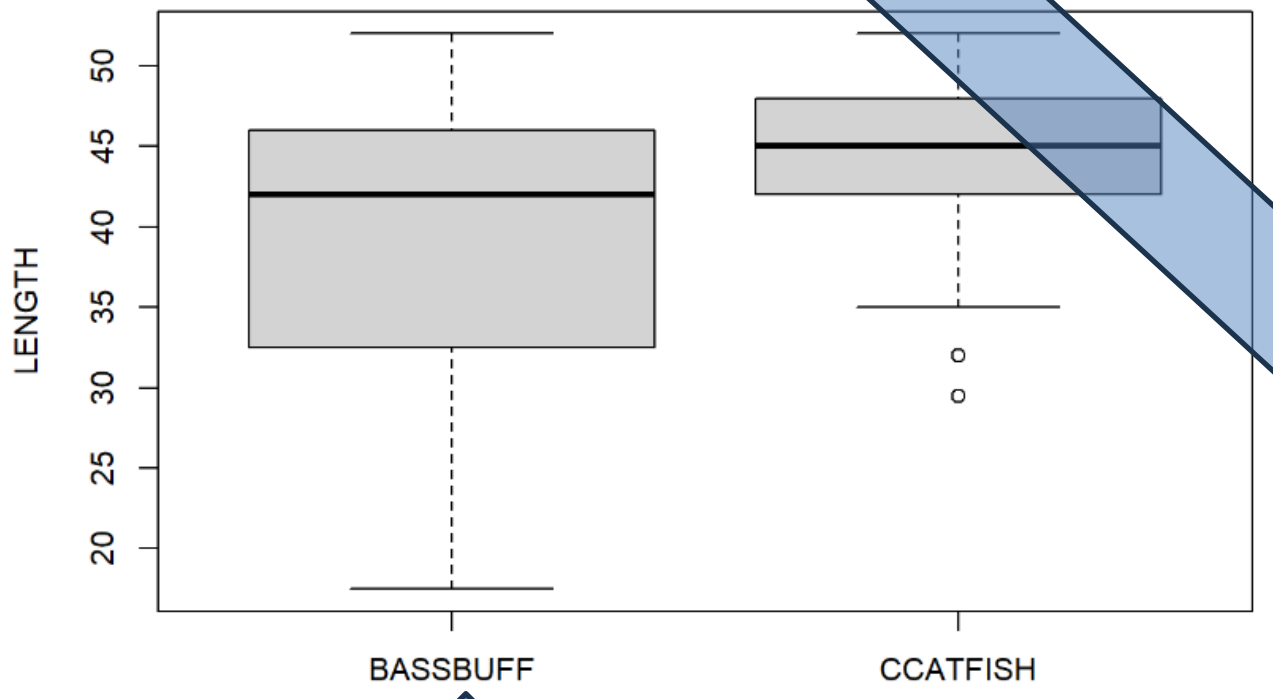
1. The first step is to hypothesize a **probabilistic model**. In this chapter, we confined our attention to the **straight-line model**, $y = \beta_0 + \beta_1 x + \varepsilon$.
2. The second step is to use the method of least squares to estimate the unknown parameters in the **deterministic component**, $\beta_0 + \beta_1 x$. The least-squares estimates yield a model $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$ with a **sum of squared errors (SSE)** that is smaller than the SSE for any other straight-line model.
3. The third step is to specify the probability distribution of the **random error component ε**. Conduct a **residual analysis** to check the validity of these assumptions.
4. The fourth step is to assess the utility of the hypothesized model. Included here are making inferences about the **slope β_1**, calculating the **coefficient of correlation r**, and calculating the **coefficient of determination r^2**.
5. Finally, if we are satisfied with the model, we are prepared to use it. We used the model to **estimate the mean y value, $E(y)$**, for a given x value and to **predict an individual y value** for a specific value of x.

```
with(ddt, boxplot(LENGTH ~ SPECIES))
```



```
ddt2 <- ddt %>% mutate(SPECIES2 = recode(SPECIES, LMBASS  = "BASSBUFF", SMBUFFALO = "BASSBUFF")) %>% mutate(SPECI
ES2 = factor(SPECIES2, levels = c("BASSBUFF", "CCATFISH")))

with(ddt2, boxplot(LENGTH ~ SPECIES2))
```

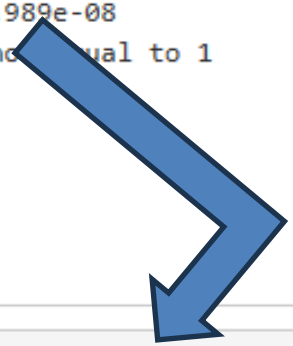


Here we recode levels

We can order the levels by using “factor()”

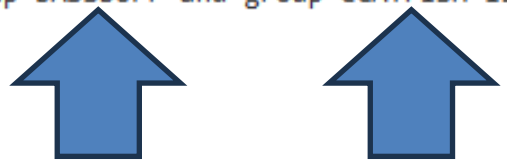
```
with(ddt2, var.test(LENGTH ~ SPECIES2))
```

```
##
##  F test to compare two variances
##
## data:  LENGTH by SPECIES2
## F = 3.7734, num df = 47, denom df = 95, p-value = 3.989e-08
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  2.339752 6.349847
## sample estimates:
## ratio of variances
##           3.773412
```



```
with(ddt2, t.test(LENGTH ~ SPECIES2, mu = 0, var.equal = FALSE))
```

```
##
##  Welch Two Sample t-test
##
## data:  LENGTH by SPECIES2
## t = -4.2069, df = 59.762, p-value = 8.814e-05
## alternative hypothesis: true difference in means between group BASSBUFF and group CCATFISH is not equal to 0
## 95 percent confidence interval:
##  -8.484254 -3.015746
## sample estimates:
## mean in group BASSBUFF mean in group CCATFISH
##               38.97917               44.72917
```



BASSBUFF first – why?

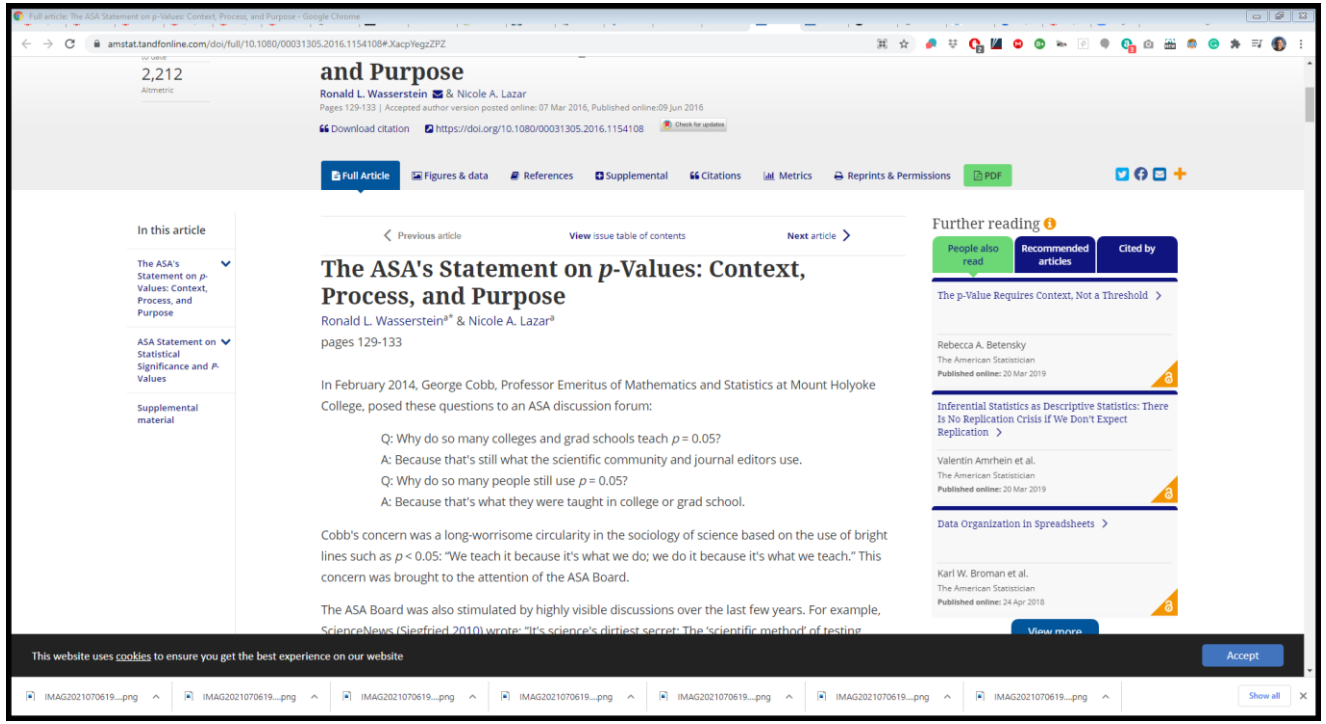
OTHER ISSUES

STATISTICAL ERRORS

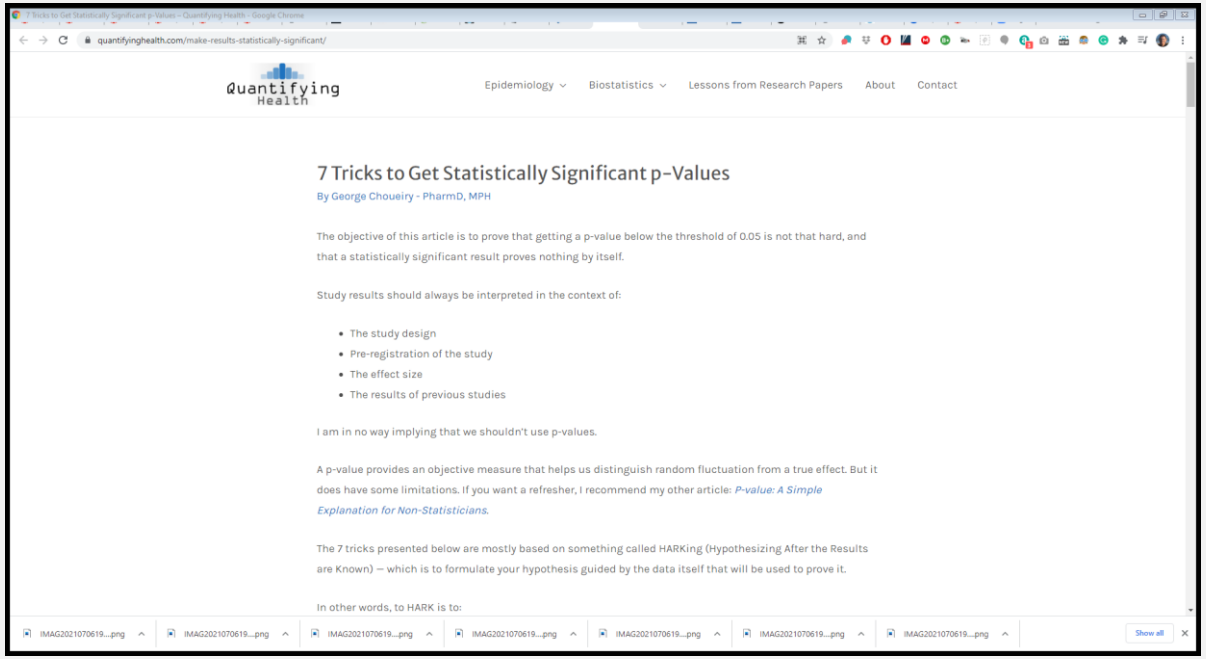
P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume.

BY REGINA NUZZO

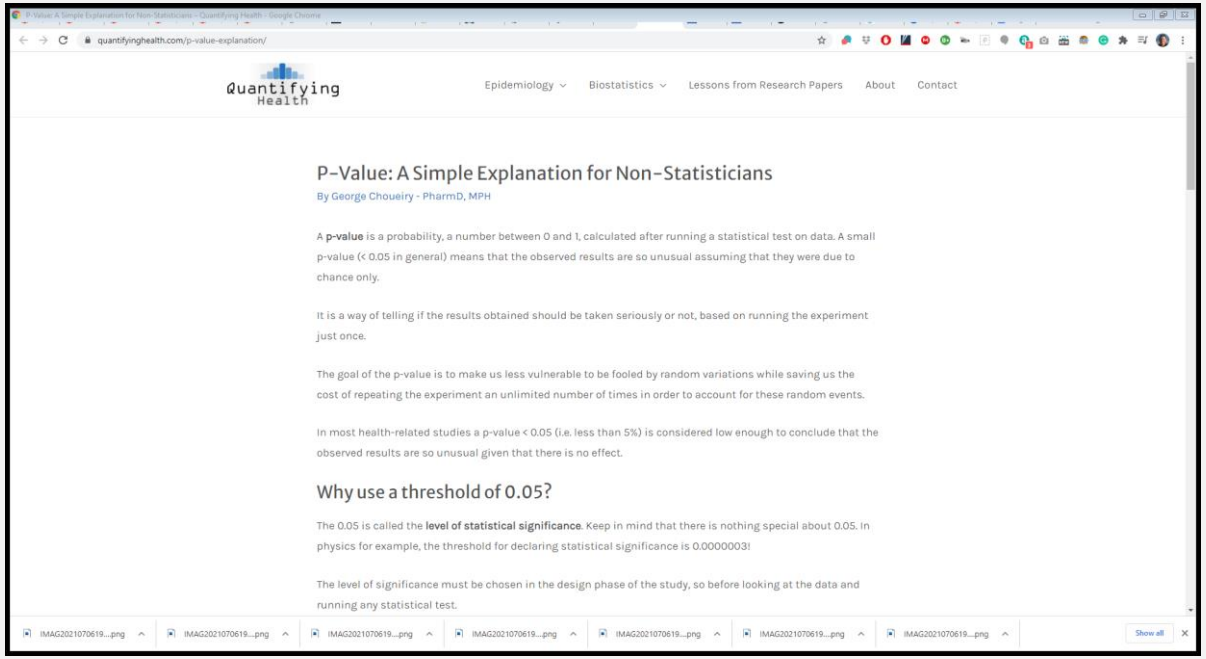
BE AWARE OF THE ISSUES RELATED TO THE CLASSICAL PARADIGM (P-VALUES)



EASY TO UNDERSTAND WARNINGS



P-VALUE


quantifyinghealth.com/p-value-explanation/
Quantifying Health
Epidemiology
Biostatistics
Lessons from Research Papers
About
Contact
P-Value: A Simple Explanation for Non-Statisticians
By George Choueiry - PharmD, MPH
A p-value is a probability, a number between 0 and 1, calculated after running a statistical test on data. A small p-value (< 0.05 in general) means that the observed results are so unusual assuming that they were due to chance only.
It is a way of telling if the results obtained should be taken seriously or not, based on running the experiment just once.
The goal of the p-value is to make us less vulnerable to be fooled by random variations while saving us the cost of repeating the experiment an unlimited number of times in order to account for these random events.
In most health-related studies a p-value < 0.05 (i.e. less than 5%) is considered low enough to conclude that the observed results are so unusual given that there is no effect.
Why use a threshold of 0.05?
The 0.05 is called the level of statistical significance. Keep in mind that there is nothing special about 0.05. In physics for example, the threshold for declaring statistical significance is 0.0000003!
The level of significance must be chosen in the design phase of the study, so before looking at the data and running any statistical test.

Mindless statistics

Gerd Gigerenzer*

Max Planck Institute for Human Development, Lentzeallee 94, 14195 Berlin, Germany

Rituals call for cognitive illusions. Their function is to make the final product, a significant result, appear highly informative, and thereby justify the ritual. Try to answer the following question (Oakes, 1986; Haller and Krauss, 2002):

> Suppose you have a treatment that you suspect may alter performance on a certain task. You compare the means of your control and experimental groups (say 20 subjects in each sample). Further, suppose you use a simple independent means *t*-test and your result is significant ($t = 2.7$, d.f. $= 18$, $p = 0.01$). Please mark each of the statements below as "true" or "false." "False" means that the statement does not follow logically from the above premises. Also note that several or none of the statements may be correct.

1.You have absolutely disproved the null hypothesis (that is, there is no difference between the population means).

[] true/false []

2. You have found the probability of the null hypothesis being true.

[] true/false []

3. You have absolutely proved your experimental hypothesis (that there is a difference between the population means).

[] true/false []

4. You can deduce the probability of the experimental hypothesis being true.

[] true/false []

5.You know, if you decide to reject the null hypothesis, the probability that you are making the wrong decision.

[] true/false []

6. You have a reliable experimental finding in the sense that if, hypothetically, the experiment were repeated a great number of times, you would obtain a significant result on 99% of occasions.

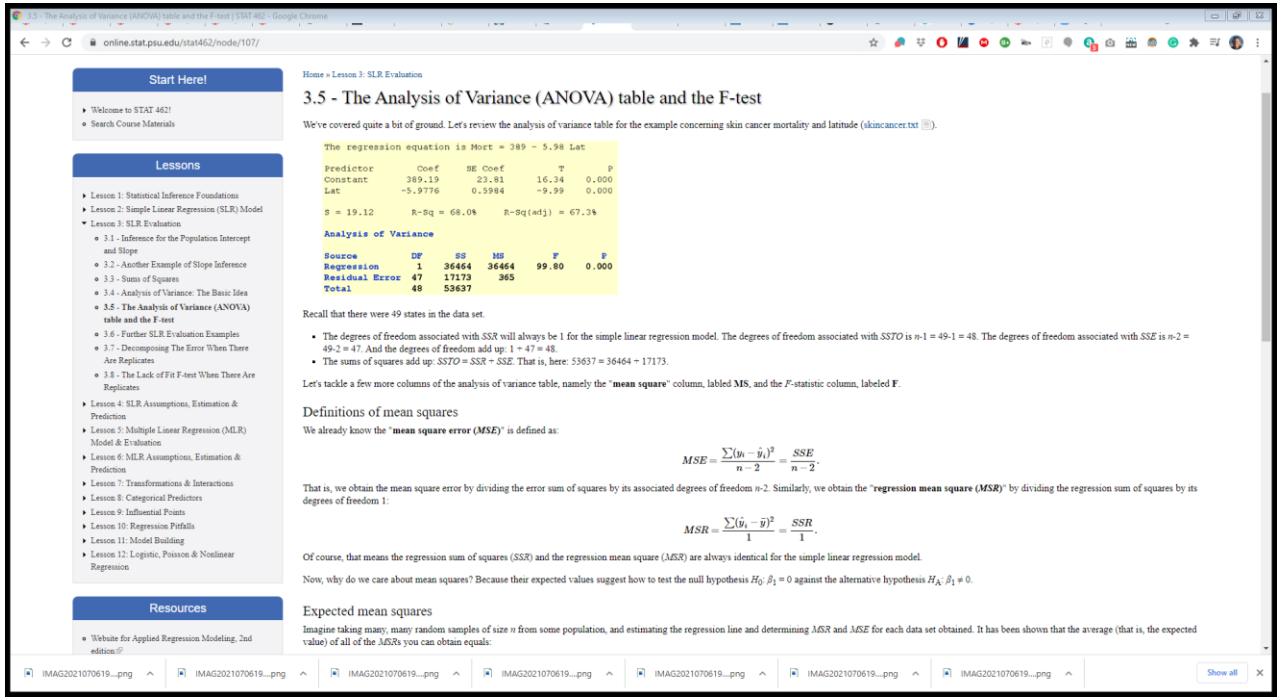
[] true/false []

Statements 1 and 3 are easily detected as being false, because a significance test can never disprove the null hypothesis or the (undefined) experimental hypothesis. They are instances of the *illusion of certainty* (Gigerenzer, 2002).

Statements 2 and 4 are also false. The probability $p(D|H_0)$ is not the same as $p(H_0|D)$, and more generally, a significance test does not provide a probability for a hypothesis. The statistical toolbox, of course, contains tools that would allow estimating probabilities of hypotheses, such as Bayesian statistics. Statement 5 also refers to a probability of a hypothesis. This is because if one rejects the null hypothesis, the only possibility of making a wrong decision is if the null hypothesis is true. Thus, it makes essentially the same claim as Statement 2 does, and both are incorrect.

Statement 6 amounts to the replication fallacy (Gigerenzer, 1993, 2000). Here, $p = 1\%$ is taken to imply that such significant data would reappear in 99% of the repetitions. Statement 6 could be made only if one knew that the null hypothesis was true. In formal terms, $p(D|H_0)$ is confused with $1 - p(D)$.

To sum up, all six statements are incorrect. Note that all six err in the same direction of wishful thinking: They make a p-value look more informative than it is.



MORE LESSONS (VERY USEFUL)

FIND A

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6949     4.6645  ##A##    ####
x             8.2396     ##B##   31.360   <2e-16 ***
---
```

FIND B

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6949     4.6645  ##A##    ####
x             8.2396     ##B##   31.360   <2e-16 ***
---
```

FIND C

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6949     4.6645  -0.149    ##C##
x             8.2396     0.2627  31.360   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.46 on 28 degrees of freedom
Multiple R-squared:  0.9723,     Adjusted R-squared:  0.9713
F-statistic: ##D## on 1 and ##E## DF,  p-value: ##F##
```

FIND D

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6949     4.6645  -0.149    ##C##
x             8.2396     0.2627  31.360   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.46 on 28 degrees of freedom
Multiple R-squared:  0.9723,     Adjusted R-squared:  0.9713
F-statistic: ##D## on 1 and ##E## DF,  p-value: ##F##
```

FIND E

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6949     4.6645  -0.149    ##C##
x             8.2396     0.2627  31.360   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.46 on 28 degrees of freedom
Multiple R-squared:  0.9723,     Adjusted R-squared:  0.9713
F-statistic: ##D## on 1 and ##E## DF,  p-value: ##F##
```

FIND F

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6949     4.6645  -0.149    ##C##
x             8.2396     0.2627  31.360   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.46 on 28 degrees of freedom
Multiple R-squared:  0.9723,     Adjusted R-squared:  0.9713
F-statistic: ##D## on 1 and ##E## DF,  p-value: ##F##
```

FIND THE SAMPLE PEARSON CORRELATION COEFFICIENT r

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.6949     4.6645  -0.149    ##C##
x             8.2396     0.2627  31.360   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.46 on 28 degrees of freedom
Multiple R-squared:  0.9723,     Adjusted R-squared:  0.9713
F-statistic: ##D## on 1 and ##E## DF,  p-value: ##F##
```