BUSINESS STATISTICS

introduction to multiple regression analysis

STUDY GUIDE

MICHAEL J McQUAID

THIS REVISION PRODUCED AUGUST 5, 2014

Contents

I. Introduction . . . . . 5.
II. Review summary statistics . . . . . 8.
III. Standard deviation review . . . . . 11.
IV. HW 1.27: Calculate some summary statistics . . . . . 12.
V. HW 1.28: Calculate summary statistics . . . . . 12.
VI. Normal distribution . . . . . 13.
VII. HW 1.32: Calculating z scores . . . . . 17.
VIII. HW 1.33: Sketch the graphs of probabilities . . . . . 18.
IX. Central limit theorem . . . . . 20.
X. Normal distribution example problem . . . . . 20.
XI. Exam questions for z-scores . . . . . 23.
XII. Estimate a population parameter . . . . . 25.
XIII. Elements of a hypothesis test . . . . . 26.
XIV. Mistakes typically made on quizzes . . . . . 31.
XV. Revisit population parameters . . . . . 33.
XVI. Identify small and large samples . . . . . 33.
XVII. Estimate a population mean . . . . . 33.
XVIII. HW 1.48: Practice estimating a population mean . . . . . 35.
XIX. Test a hypothesis about a population mean . . . . . 38.
XX. HW 1.53: Identify rejection region marker . . . . . 38.
XXI. HW 1.61: Test a hypothesis about a mean . . . . . 40.
XXII. Linear regression introduction . . . . . 43.
XXIII. HW 3.1: Graph lines . . . . . 43.
XXIV. HW 3.3: Equations of lines . . . . . 43.
XXV. HW 3.4: Graphing equations of lines . . . . . 43.
XXVI. HW 3.5: Finding beta 0 and beta 1 . . . . . 44.
XXVII. The simplest linear regression model . . . . . 44.
XXVIII. A linear regression example . . . . . 44.
XXIX. HW 3.6: Use the method of least squares . . . . . 49.
XXX. HW 3.7: Use the method of least squares . . . . . 50.
XXXI. Output of a computerized regression analysis . . . . . 51.
XXXII. HW 3.11: Compare scatterplots . . . . . 52.
XXXIII. Check the 4 model assumptions for any given model . . . . . 53.

XXXIV. HW 3.17: SUM OF SQUARED RESIDUALS . . . . . 54.
XXXV. HW 3.18: VARIANCE OF EPSILON . . . . . 54.
XXXVI. HW 3.21: READING REGRESSION ANALYSIS OUTPUT . . . . 56.
XXXVII. HW 3.23 DOES BETA SUB ONE EQUAL ZERO?. . . . . 56.
XXXVIII. HW 3.24 INTERPRET COMPUTERIZED CONF INTVL OUTPUT . . 57.
XXXIX. HW 3.31 CONDUCT A SIMPLE LINEAR REGRESSION . . . . . 58.
XL. FIND AND INTERPRET THE CORRELATION COEFFICIENT . . . 58.
XLI. FIND AND INTERPRET THE COEFFICIENT OF DETERMINATION . 60.
XLII. HW 3.34 CORRELATION: POSITIVE, NEGATIVE, WEAK . . . . 61.
XLIII. HW 3.35: COEFFICIENTS OF CORRELATION AND DETERMINATION 61.
XLIV. HW 3.48: PREDICTION AND CONFIDENCE INTERVALS . . . . 61.
XLV. HW 3.49: REGRESSION EXAMPLE . . . . . 62.
XLVI. HW 3.60: INTERPRET A REGRESSION ANALYSIS . . . . . 63.
XLVII. HW 3.61: CONDUCT A LINEAR REGRESSION ANALYSIS . . . . 63.
XLVIII. REVIEW OF PREVIOUS PART OF COURSE . . . . . 64.
XLIX. MULTIPLE REGRESSION . . . . . 66.
L. THE F DISTRIBUTION . . . . . 66.
LI. TEST OVERALL MODEL UTILITY . . . . . 68.
LII. HW 4.1: DEGREES OF FREEDOM . . . . . 69.
LIII. ADJUSTED MULTIPLE COEFFICIENT OF DETERMINATION . . . 70.
LIV. HW 4.2: ESTIMATE A MULTIPLE REGRESSION MODEL . . . . 71.
LV. HW 4.4: TEST A HYPOTHESIS ABOUT BETA I . . . . . 73.
LVI. HW 4.6: ESTIMATE ANOTHER MODEL . . . . . 75.
LVII. HW 4.7: OBTAIN F FROM R SQUARED. . . . . 76.
LVIII. RELATE EXPLAINED TO UNEXPLAINED VARIANCE . . . . . 78.
LIX. HW 4.9: CONFIDENCE INTERVALS FOR BETA 1 TO K . . . . 78.
LX. HW 4.10: READ COMPUTER OUTPUT. . . . . 79.
LXI. HW 4.12: ANOTHER COMPUTER OUTPUT PROBLEM . . . . 79.
LXII. HW 4.16: TEST MODEL ADEQUACY . . . . . 82.
LXIII. HW 4.18: PREDICT AND ESTIMATE . . . . . 85.
LXIV. HW 4.20: PREDICTION AND CONFIDENCE INTERVALS . . . . 85.
LXV. QUADRATIC MODELS . . . . . 87.
LXVI. HW 4.34: QUADRATIC MODEL EXAMPLE. . . . . 88.
LXVII. HW 4.36: ANOTHER QUADRATIC MODEL . . . . . 89.
LXVIII. NESTED MODELS . . . . . 91.

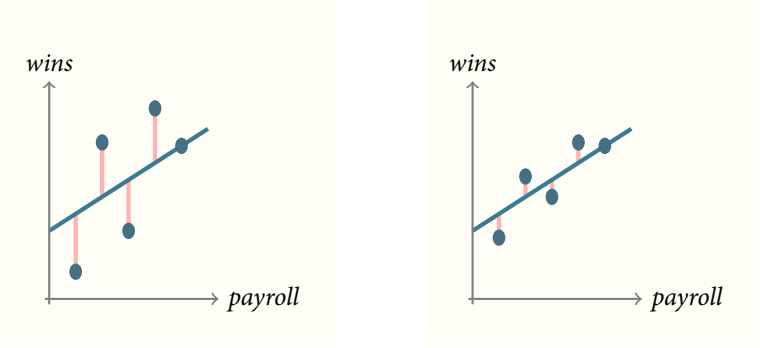
LXIX. HW 4.86: NESTED MODEL . . . . . 92.
LXX. INTERACTION . . . . . 93.
LXXI. SECOND-ORDER MODELS . . . . . 93.
LXXII. HW 5.15: SECOND-ORDER MODEL EXAMPLES . . . . . 93.
LXXIII. HW 5.16: MORE SECOND-ORDER MODEL EXAMPLES . . . . 94.
LXXIV. HW 5:31: QUALITATIVE AND QUANTITATIVE VARIABLES . . . 94.
LXXV. HW 5.33: COMBINATIONS OF INDICATOR VARIABLES . . . . 97.
LXXVI. HW 6.1: STEPWISE REGRESSION PROCESS . . . . . 97.
LXXVII. HW 6.2: STEPWISE REGRESSION EXAMPLE . . . . . 99.
LXXVIII. HW 6.6: STEPWISE REGRESSION USING SOFTWARE . . . . 101.
LXXIX. HW 7.2 MULTICOLLINEARITY . . . . . 101.
LXXX. HW 7.4: MULTICOLLINEARITY EXAMPLE . . . . . 103.
LXXXI. THE REST . . . . . 104.
Appendices . . . . . 104.
A. PRACTICE FIRST QUIZ . . . . . 104.
B. PRACTICE SECOND QUIZ . . . . . 105.
C. PRACTICE THIRD QUIZ . . . . . 107.
D. PRACTICE MIDTERM EXAM . . . . . 109.
E. PRACTICE FOURTH QUIZ . . . . . 117.
F. PRACTICE FIFTH QUIZ . . . . . 118.
G. PRACTICE SIXTH QUIZ . . . . . 121.
H. PRACTICE FINAL EXAM . . . . . 122.
I. SOLUTIONS TO ALL QUIZZES AND EXAMS . . . . . 133.
J. TERMINOLOGY . . . . . 167.
K. TABLES . . . . . 174.

I. INTRODUCTION

This course helps you to interpret the variability in data. The goal is to help you estimate and predict unknown quantities, a common task in business as elsewhere in life. For example, you might try to estimate future sales by taking an average of past sales. In a previous course, you should have learned how to do that. In this course, you will learn a method that may yield better results. Suppose you know that there is a relationship between advertising and sales. You may be able to form a better estimate of future sales by taking what you know about that relationship into account. The technique you'll learn here for predicting or estimating a quantity like

sales, taking into account what you know about its relationship to another quantity, like advertising, is called regression.

Quantify the difference between two prediction lines.



The prediction line on the right is better than that on the left by an amount proportional to the difference between the total length of red lines in the two pictures.

This course is about finding and assessing the best line.

- Consider a desired outcome (e.g., wins)
- Identify one or more factors contributing (e.g., payroll dollars)
- Find the slope and intercept to predict how much of the factor leads to how much of the outcome
- Figure out how good or bad the prediction is

… and there you have a simplified view of regression, the heart of this course.

Think about the prediction lines as models of reality.

- reduces reality to a manageable fiction
- requires deep knowledge of the subject you model
- easy to do badly
- hard to figure out what aspects of reality to leave in and what to take out
- example: death penalty in Florida

The process I've just described is a process called modeling. Some people call the equations of lines models. It is more correct though to say that the model includes both the equations and a set of statements related to the equations. When we work with models we need to pause and think about what models are and why we make them.

A model is a parsimonious description of some aspect of reality. The word parsimonious literally means frugal or stingy. In science it means we leave out the aspects of reality that are not absolutely necessary. Opinions differ about what is absolutely necessary.

The more realistic model is the more expensive it becomes to construct: In business we want to find the minimum amount of stuff we have to keep track of in order to construct a model that useful enough to manage.

People in decision science see management as making choices and following through on executing those choices. We need models that are good enough to make choices and good enough to help us monitor how our execution of choices is going. Consider a simple example of fishing. When I arrived in Eugene, a man told me that within a 50 mile radius there are 300 miles of fishing coastline. He claimed that most people here fish. Do you? Suppose two fishermen each tell you their favorite spot is the best for catching big walleye. It might be impolite to ask them to prove it. One weekend you accompany fisherman Q to his favorite spot and catch some walleye. The polite thing to say is that these are the finest and biggest walleye you have ever seen. When you get home and fisherman Q is not around you weigh the walleye. The next weekend you do the same with fisherman R. The next weekend both fishermen want you to go to their spot at the same time. Both fishermen are equally fun to be around. Both bring the finest cold beer with them. The only difference you can think of is that perhaps one stream has bigger walleye and thus better helps you to feed your family. Which do you choose?

Notice the ways in which this model is not realistic. All this model includes is our location and weight and nothing else. Is that a good description of reality? No two people are equally fun. No two beers are equally cold. No two spots next to streams are equally comfortable. Can you think of more variables we just left out? You should be able to. Much of the art of management is figuring out which variables are important. Some of the art of management is being able to reframe the problem. For example you could perhaps convince both of the fishermen to try a new third spot, even if it is just to prove their spots are better.

Death penalty in Florida example. Many analysts tried to determine whether the death penalty in Florida was being applied in a racist manner. Many failed to do so. Most tried to predict the likelihood that a murderer would receive the death penalty (y) as a function of the race of the murderer (x). Various analysts tried to control for different factors, such as the relative populations of each race from which murderers are convicted. Finally, one analyst considered a different model: predicting the likelihood that a convicted murderer would receive the death penalty (y) as a function of the

race of the victim of the murder (x). Using this model, the analyst showed that a murderer is overwhelmingly more likely to receive the death penalty for murdering a white victim than for a murdering a victim of color. Why had this seemingly simple model eluded analysts for so many years? The answer, in part, is that modeling is very hard. Choosing the right variables is very hard.

This is a technical course in which you will work with the mechanics of models. Many of the most important issues in modeling are outside the scope of this course. We believe it will take you 10 weeks to figure out the mechanics of building and analyzing regression models. The task of figuring out which models are worthwhile to build and analyze for managerial purposes may take your whole career.

II. REVIEW SUMMARY STATISTICS

A prerequisite should have covered these.

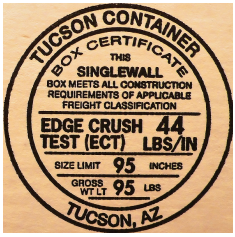
- Distinguish between samples and populations.
- Know how to calculate the arithmetic mean.
- Know how to calculate standard deviation.
- Know the definition of median.
- Review other summary statistics.

We begin following the textbook with section 1.5, *Describing quantitative data numerically*. This section introduces the mean, the distinction between sample mean and population mean, the range, the sample and population variance, and the sample and population standard deviation. The textbook also displays a stem-and-leaf plot and a set of descriptive statistics, including N, mean, median, Trimmed mean, standard deviation, Standard Error Mean, minimum, maximum, first quartile, and third quartile.

Guidelines for interpreting standard deviation are (a) for any data set, at least three-fourths of the measurements will lie within two standard deviations of the mean, and (b) for most data sets with enough measurements (25 or more) and a mound-shaped distribution, about 95 percent of the measurements will lie within two standard deviations of the mean.

We use samples to make inferences about populations.

Shipping boxes bear certificates of testing as a benefit of the Uniform Commercial Code, which allows businesses in each of the fifty United States to make legally-binding assumptions about the businesses they work with in other states. Unfortunately, you have to destroy the box to conduct the crush test.



Therefore a sample of boxes are tested to estimate the parameters of the population of boxes and we need terminology to talk about both sample and population.

The act of finding a sum is denoted like this.

The Greek letter Sigma, Σ, usually means to sum the values represented by the expression that follows:

$$\sum_{i=1}^{n} y_i$$

which is the same as

$$y_1 + y_2 + \cdots + y_n$$

Sigma notation may be inconsistent.

You may see Σ used in an inconsistent way in math and stats:

$$\sum_{i=1}^{n} y_i$$

may be replaced by a synonymous shortcut like

$$\sum_{i} y_i \quad \text{or} \quad \sum y$$

The arithmetic mean is the average of a set of values.

Usually when we use the word *mean*, we refer to

$$\bar{y} = \frac{\sum_{i=1}^{n} y_i}{n}$$

which is the same as

$$\bar{y} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

We use the sample mean to estimate the population mean.

The sample mean is denoted in the textbook as $\bar{y}$.

The population mean is called the expected value of y and denoted in the textbook as

$$E(y) = \mu$$

and in the case of the boxes, we would have to destroy all of them to be sure of its value, so we destroy a sample to estimate μ.

A sample's range is the difference between its max and min.

If the grades of a sample of six students are

$$(2, 2, 3, 3, 4, 4)$$

then the range is

$$4 - 2 = 2$$

The mean of the sample is

$$\bar{y} = (2 + 2 + 3 + 3 + 4 + 4)/6 = 3$$

Standard deviation is used to describe data variation.

The standard deviation of a population is σ and of a sample is s. It's painfully easy to confuse the spreadsheet functions for σ and s.

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n}}$$

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n - 1}}$$

Find the standard deviation of the grade sample.

- sum: $2 + 2 + 3 + 3 + 4 + 4 = 18$
- mean: $18/6 = 3$

- deviations: $(2-3)^2+(2-3)^2+(3-3)^2+(3-3)^2+(4-3)^2+(4-3)^2$
- deviations part two: $1+1+0+0+1+1$
- divide: $4/(6-1) = 4/5 = 0.8$
- square root: just write $\sqrt{0.8}$ unless you're allowed a calculator / computer

The *deviations part two* step is the numerical version of what I previously showed in the graph with green lines between data and some imaginary line. In this case, the imaginary line is $\bar{y}$.

III. STANDARD DEVIATION REVIEW

Calculating s emphasizes its interpretation.

Previously I asked you to add up the deviations, even though there is a math shortcut that prevents you from having to do so. I asked you to do it this way because the shortcut does not make an obvious connection between the pink lines in the graph at the beginning of the introduction and the result of the equation. In future, I will specify whether to use the previous formula or its shortcut,

$$s = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}{n-1}}$$

Keep in mind that they give equivalent results. The s in this equation is the same s as above.

two rough guidelines to interpret s

(i) For any data set, at least three-fourths of the measurements will lie within two standard deviations of the mean.

(ii) For most data sets with enough measurements (25 or more) and a mound-shaped distribution, about 95 percent of the measurements will lie within two standard deviations of the mean. (We'll study mound-shaped distributions in the next lecture.)

These are rough guidelines. Confidence intervals and prediction intervals offer much more precise guidelines we'll learn later but will require that certain assumptions are met. We'll always use confidence intervals or prediction intervals in preference to these rough guidelines *if* we can meet these assumptions.

Standard deviation and mean work as a pair.

When you want to describe a set of data, the two most frequently used numbers, used as a pair, are mean and standard deviation. Suppose two websites, tra.com and la.com, both sell used phones. The last five sales of

the ZZ11 on tra.com, in chronological order, were \$36, \$29, \$59, \$18, \$23, \$35, \$25, \$63, \$69, and \$43.

The last fives sales of the ZZ11 on la.com, in chronological order, were \$44, \$36, \$47, \$38, \$35, \$36, \$37, \$38, \$50, and \$39. Using only this info, what is the expected value of the next sale in each market? Our best estimate of the expected value is the sample mean, which is \$40 in both cases. What about the standard deviation? Calculating the standard deviation will give you about \$17.95 for tra.com and \$5.16 for la.com. So the most likely value in each market is \$40 but we have a lot more confidence in that estimate for la.com. It's not that we expect a different value. It's that the values are more widely dispersed in the first case than in the second case. It's more likely that we'll get \$40 for our used phone if the sale prices are all near \$40. If half the sales in bla.com were for \$5 and half were for \$75, the expected value would still be \$40, even though the comparison of the bla.com market to the other two markets provides a striking difference in risk. Which market would you prefer? Surely it would depend in part on how much risk you were willing to absorb. The *safe* bet would be on the market with a standard deviation of \$5.16 and the riskiest would be the bla.com market with a standard deviation of \$36.89.

So we conclude this section by saying that the standard deviation really enriches our description of a data set beyond the description given by the mean.

IV. HW 1.27: CALCULATE SOME SUMMARY STATISTICS

Textbook problem 1.27 asks you to use the SHIPSANIT dataset bundled with the textbook to calculate some summary statistics. You may use a calculator, spreadsheet, or statistics app to do so. Check the resources folder on Blackboard for videos demonstrating each of these methods.

(a) $\bar{y} = 94.91124, s = 4.825323$

(b) $\bar{y} \pm 2s = (85.2606, 104.5619)$

(c) The percentage of scores in the above interval is 97.63314%. The guideline on page 21 indicates that approximately 95% should fall within two standard deviations, so my opinion is that this result agrees with the guideline. You may disagree if your definition of *approximately* is more stringent.

V. HW 1.28: CALCULATE SUMMARY STATISTICS

Textbook problem 1.28 asks you to use the WPOWER50 dataset bundled with the textbook to calculate some summary statistics. As with the pre-

vious problem, you may use a calculator, spreadsheet, or statistics app to do so. Check the resources folder on Blackboard for videos demonstrating each of these methods.

(a) $\bar{y} = 50.02, s = 6.444393$

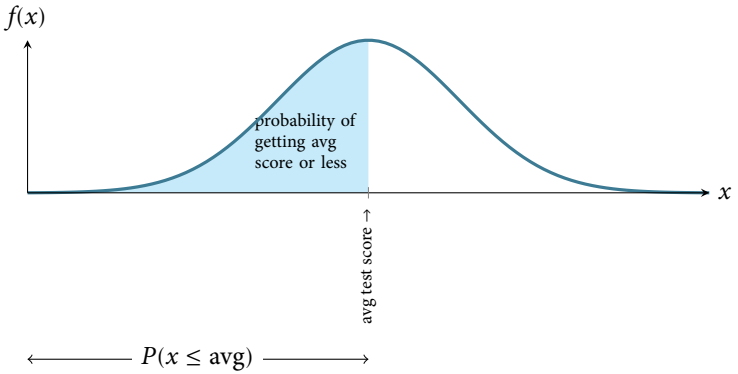
(b) If you interpret *highly likely* as 95%, then the interval in the previous problem, $\bar{y} \pm 2s$ should be sufficient: (37.13121, 62.90879). That interval encloses 94% of the sample, which I regard as a good approximation.

As this review continues, bear in mind that, overall, our goal is to make inferences from data, not an easy task.

> *It's tough to make predictions, especially about the future.*
>
> Yogi Berra

VI. NORMAL DISTRIBUTION

This picture illustrates the normal distribution. The mound-shaped curve represents the probability density function and the area between the curve and the horizontal line represents the value of the cumulative distribution function.



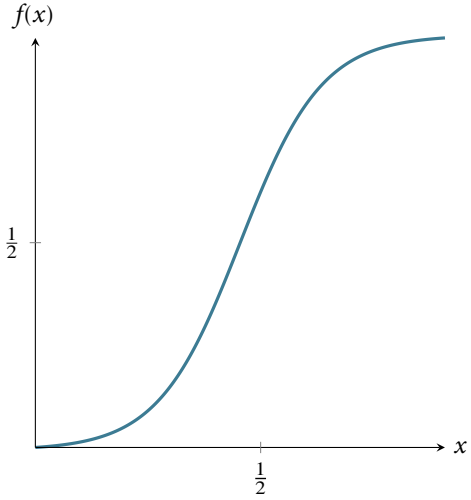
Consider a normally distributed nationwide test.

The total shaded area between the curve and the straight horizontal line can be thought of as one hundred percent of that area. In the world of probability, we measure that area as 1. The curve is symmetrical, so measure all the area to the left of the highest point on the curve as 0.5. That is half, or fifty percent, of the total area between the curve and the horizontal line at the bottom. Instead of saying *area between the curve and the horizontal line at the bottom*, people usually say *the area under the curve.*

For any value along the x-axis, the y-value on the curve represents the value of the probability density function.

The area bounded by the vertical line between the x-axis and the corresponding y-value on the curve, though, is what we are usually interested in because that area represents probability.

Here is a graph of the *size* of that area. It's called the cumulative distribution function.



The above graph can be read as having an input and output that correspond to the previous graph of the probability density function. As we move from right to left on the x-axis, the area that would be to the left of a given point on the probability density function is the y-value on this graph. For example, if we go half way across the x-axis of the probability density function, the area to its left is one half of the total area, so the y-value on the cumulative distribution function graph is one half.

The shape of the cumulative distribution function is called a sigmoid curve. You can see how it gets this shape by looking again at the probability density function graph above. As you move from left to right on that graph, the area under the curve increases very slowly, then more rapidly, then slowly again. The places where the area grows more rapidly and then more slowly on the probability density function curve correspond to the s-shaped bends on the cumulative distribution curve.

At the left side of the cumulative distribution curve, the y-value is zero meaning zero probability. When we reach the right side of the cumulative distribution curve, the y-value is 1 or 100 percent of the probability.

Let's get back to the example of a nationwide test. If we say that students nationwide took an test that had a mean score of 75 and that the score was normally distributed, we're saying that the value on the x-axis in the center of the curve is 75. Moreover, we're saying that the area to the left of 75 is one half of the total area. We're saying that the probability of a score less than 75 is 0.5 or fifty percent. We're saying that half the students got a score below 75 and half got a score above 75.

That is called the frequentist interpretation of probability. In general, that interpretation says that a probability of 0.5 is properly measured by saying that, if we could repeat the event enough times, we would find the event happening half of those times.

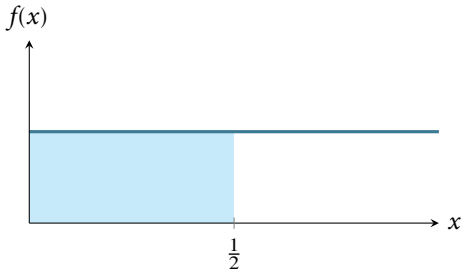
Furthermore, the frequentist interpretation of the normal distribution is that, if we could collect enough data, such as administering the above test to thousands of students, we would see that the graph of the frequency of their scores would look more and more like the bell curve in the picture, where x is a test score and y is the number of students receiving that score.

Suppose we have the same test and the same distribution but that the mean score is 60. Then 60 is in the middle and half the students are on each side. That is easy to measure. But what if, in either case, we would like to know the probability associated with scores that are not at that convenient midpoint?

It's hard to measure any other area under the normal curve except for x-values in the middle of the curve, corresponding to one half of the area. Why is this?

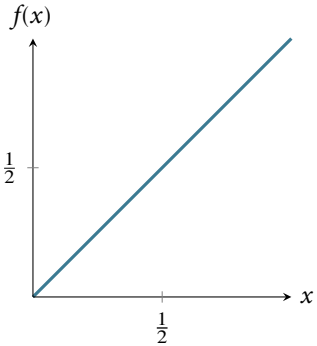
To see why it's hard to measure the area corresponding to any value except the middle value, let's first consider a different probability distribution, the uniform distribution. Suppose I have a machine that can generate any number between 0 and 1 at random. Further, suppose that any such number is just as likely as any other such number.

Here's a graph of the the uniform distribution of numbers generated by the machine. The horizontal line is the probability density function and the shaded area is the cumulative distribution function from 0 to 1/2. In other words, the probability of the machine generating numbers from 0 to 1/2 is 1/2. The probability of generating numbers from 0 to 1 is 1, the area of the entire rectangle.



It's very easy to calculate any probability for this distribution, in contrast to the normal distribution. The reason it is easy is that you can just use the formula for the area of a rectangle, where area is base times side. The probability of being in the entire rectangle is $1 \times 1 = 1$, and the probability of being in the part from $x = 0$ to $x = 1/4$ is just $1 \times (1/4) = 1/4$.

The cumulative distribution function of the uniform distribution is simpler than that of the normal distribution because area is being added at the same rate as we move from left to right on the above graph. Therefore it is just a straight diagonal line from (0,1) on the left to (1,1) on the right.



Reading it is the same as reading the cumulative distribution function for the normal distribution. For any value on the x-axis, say, 1/2, go up to the diagonal line and over to the value on the y-axis. In this case, that value is 1/2. That is the area under the horizontal line in the probability density function graph from 0 to 1/2 (the shaded area). For a rectangle, calculating area is trivial.

Calculating the area of a curved region like the normal distribution can be more difficult. If you've studied any calculus, you know that there are techniques for calculating the area under a curve. These techniques are called integration techniques. In the case of the normal distribution the formula for the height of the curve at any point on the x-axis is

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

and the area is the integral of that quantity from $-\infty$ to x, which can be rewritten as

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2} dt = (1/2)\left(1 + \text{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right)$$

The integral on the left is difficult to evaluate so people use numerical approximation techniques to find the expression on the right in the above equation. Those techniques are so time-consuming that, rather than recompute them every time they are needed, a very few people used to write the results into a table and publish it and most people working with probability would just consult the tables. Only in the past few decades have calculators become available that can do the tedious approximations. Hence, most statistics books, written by people who were educated decades ago, still teach you how to use such tables. There is some debate as to whether there is educational value in using the tables vs using calculators or smartphone apps or web-based tables or apps. We'll demonstrate them and assume that you use a calculator or smartphone app on exams.

Using the normal distribution

Since there is no convenient integration formula, people used tables until recently. Currently you can google tables or apps that do the work of tables. We're going to do two exercises with tables that help give you an idea of what's going on. You can use your calculator afterward. The main reason for what follows is so you understand the results produced by your calculator to avoid ridiculous mistakes.

VII. HW 1.32: CALCULATING Z SCORES

$$z = \frac{y-\mu}{\sigma}$$

Calculations use the fact that the *bell curve* is symmetric and adds up to 1, so you can calculate one side and add it, subtract it, or double it

Textbook problem 1.32 provides some examples.

(a) $P(-1 \le z \le 1)$: Since the bell curve is symmetric, find the area from $z = 0$ to $z = 1$ and double that area. The table entry for $z = 1$ gives the relevant area under the curve, .3413. Doubling this gives the area from -1 to 1: .6826. Using a calculator may give you .6827

since a more accurate value than that provided by the table would be .6826895. This is an example where no points would be deducted from your score for using either answer.

(b) $P(-1.96 \leq z \leq 1.96)$: This is one of the three most common areas of interest in this course, the other two being the one in part (c) below and the one I will add on after I show part (d) below. Here again, we can read the value from the table as .4750 and double it, giving .95. This is really common because because 95% is the most commonly used confidence interval.

(c) $P(-1.645 \leq z \leq 1.645)$: The table does not have an entry for this extremely commonly desired value. A statistical calculator or software package will show that the result is .45, which can be doubled to give .90, another of the three most frequently used confidence intervals. If you use interpolation, you will get the correct answer in this case. Interpolation means to take the average of the two closest values, in this case $(.4495 + .4505)/2$. You will rarely, if ever need to use interpolation in real life because software has made the tables obsolete and we only use them to try to drive home the concept of z-scores relating to area under the curve, rather than risking the possibility that you learn to punch numbers into an app without understanding them. Our hope is that, by first learning this method, you will be quick to recognize the results of mistakes, rather than naively reporting wacky results like *the probability is 1.5* just because you typed a wrong number.

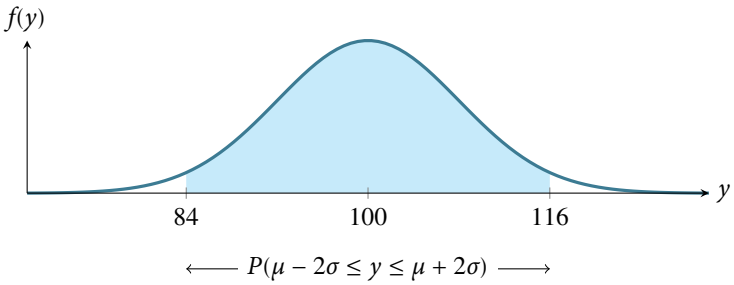
(d) $P(-3 \leq z \leq 3)$: The table gives .4987 and doubling that gives .9974. A calculator would give the more correct (but equally acceptable in this course) result of .9973.

The other common confidence interval I mentioned above is the 99% confidence interval, used in cases where the calculation relates to something life-threatening, such as a question involving a potentially life-saving drug or surgery. The textbok provides a small table on page 35 showing that the z-score that would lead to this result is 2.576. So if you were asked to compute $P(-2.576 \leq z \leq 2.576)$, the correct answer would be .99 or 99%. To use a calculator or statistical app to find the z-score given the desired probability, you would look in an app for something called a quantile function.

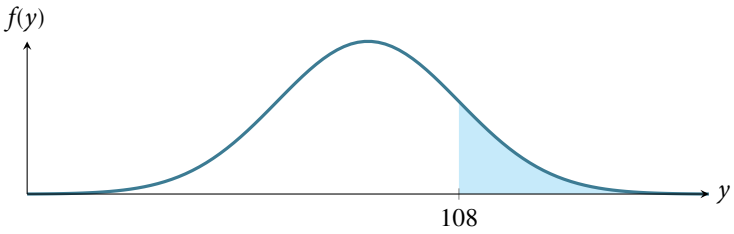
VIII. HW 1.33: SKETCH THE GRAPHS OF PROBABILITIES

Textbook problem 1.33 asks you to sketch the normal curve six times, identifying a different region on it each time. For these graphs, let $y \sim$

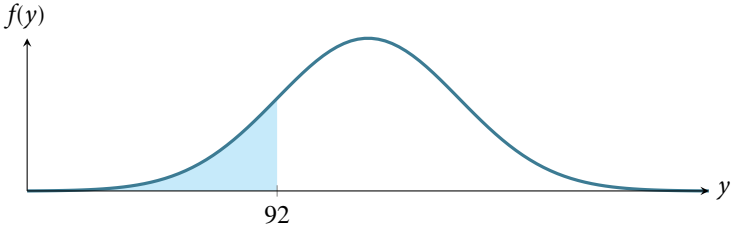
$N(100, 8)$. The textbook does not identify the standard deviation but rather the variance, which is the square of the standard deviation. The six graphs are as follows.



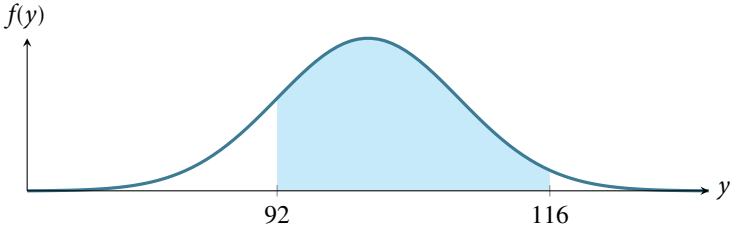
$\longleftarrow P(\mu - 2\sigma \leq y \leq \mu + 2\sigma) \longrightarrow$



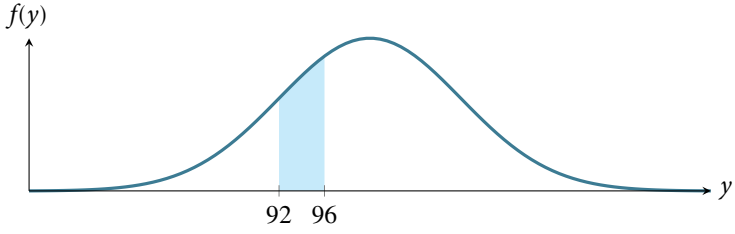
$\longleftarrow P(108 \leq y) \longrightarrow$



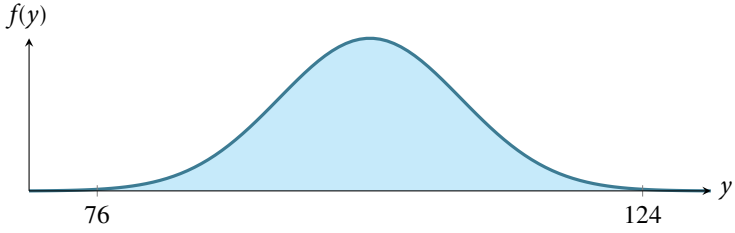
$\longleftarrow P(y \leq 92) \longrightarrow$



$\longleftarrow P(92 \leq y \leq 116) \longrightarrow$



$P(92 \leq y \leq 96)$



$P(76 \leq y \leq 124)$

IX. CENTRAL LIMIT THEOREM

Here's the definition of the central limit theorem.

For large sample sizes, the sample mean $\bar{y}$ from a population with mean μ and standard deviation σ has a sampling distribution that is approximately normal, regardless of the probability distribution of the sampled population.

Why care about the central limit theorem?

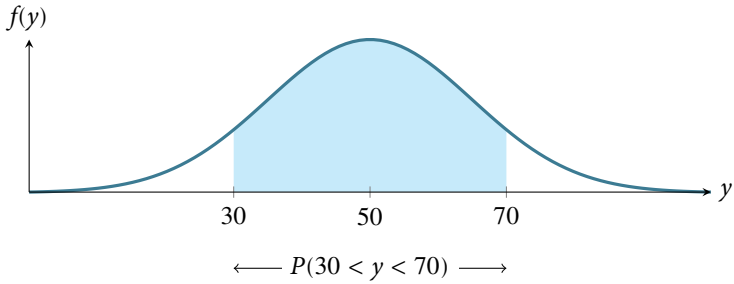
In business, distributions of phenomena like waiting times and customer choices from a catalog are typically not normally distributed, but instead long-tailed. The central limit theorem means that resampling the mean of any of these distributions can be done on a large scale using the normal distribution assumptions without regard to the underlying distribution. This simplifies many real-life calculations. For instance, waiting times at each bus stop are exponentially distributed but if we take the mean waiting time at each of 100 bus stops, the mean of those 100 times is normally distributed, even though the individual waiting times are drawn from an exponential distribution.

X. NORMAL DISTRIBUTION EXAMPLE PROBLEM

Review the normal distribution example.

The textbook gives an example for calculating a z-score where $x \sim N(50, 15)$, which is a statistical notation for saying x is a random normal variable with mean 50 and standard deviation 15. It is also read as if you said that x has the normal distribution with mean 50 and standard deviation 15.

Picture the example.



Here's what we input into the z calculation.

To identify the size of the shaded area, we can use the table of z-scores by standardizing the parameters we believe, as if they were the population parameters μ and σ. We only do this if we have such a large sample that we have reason to believe that the sample values approach the population parameters. For the more typical case of limited amount of data, we'll learn a more advanced technique that you will use more frequently in practice.

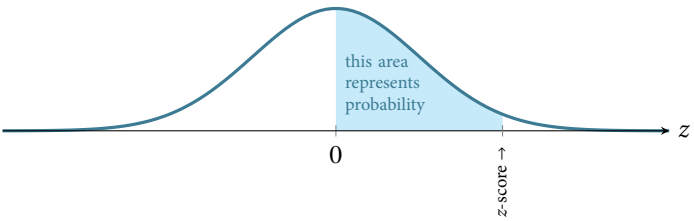
We input a z to get an output probability.

The table in our textbook contains an input in the left and top margins and an output in the body. The input is a z-score, the result of the calculation

$$z = \frac{y - \mu}{\sigma}$$

where $z \geq 0$. The output is a number in the body of the table, expressing the probability for the area between the normal curve and the axis, from the mean (0) to z. The values of z start at the mean and grow toward the right end of the graph. If z were ∞, the shaded area would be 0.5.

This is the z-score table concept.



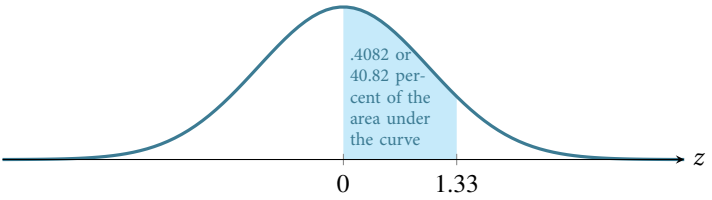
Calculate the z-score to input.

For now, let's calculate the z-score as

$$z = \frac{y - \mu}{\sigma} = \frac{70 - 50}{15} = 1.33$$

giving half of the answer we're seeking:

Apply the z-score to the table.



Obtain an intermediate result.

Now use this to read the table. The input is 1.33 and you'll use the left and top margins to find it. The output is the corresponding entry in the body of the table, .4082, also known as 40.82 percent of the area under the curve.

Finish the example.

Recall that our initial problem was to find $P(30 < y < 70)$ and what we've just found, .4082, is $P(50 < y < 70)$. We must multiply this result by 2 to obtain the correct answer, .8164 or 81.64 percent. That is to say that the probability that y is somewhere between 30 and 70 is .8164 or 81.64 percent. As a reality check, all probabilities for any single event must sum to 1, and the total area under the curve is 1, so it is a relief to note that the answer we've found is less than 1. It's also comforting to note that the shaded area in the original picture of the example looks like it could plausibly represent about 80 percent of the total area. It is easy to get lost in the mechanics of calculations and come up with a wildly incorrect answer because of a simple arithmetic error.

Some z-score tables differ from ours.

Bear in mind that anyone can publish a z-score table using their own customs. Students have found such tables that define z as starting at the extreme left of the curve. If we used such a table for the above example, the output would have been .9082 instead of .4082 and we would have had to subtract the left side, .5, from that result before multiplying by 2.

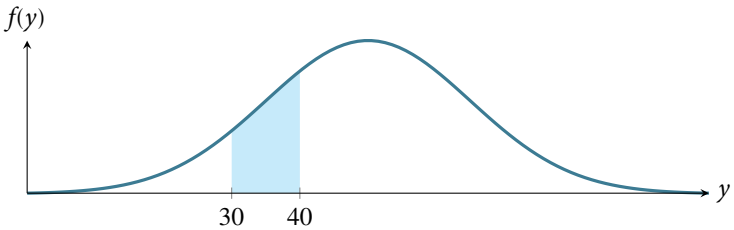
Many exam-style problems will ask questions such that you must do more or less arithmetic with the result from the table. Consider these questions, still using the above example where $y \sim N(50, 15)$: What is the probability that: y is greater than 50? y is greater than 70? y is less than 30? y is between 30 and 50?

Answer the preceding questions.

Each of these questions can be answered using $z = 1.33$ except the first. Since we know that y is normally distributed, we also know that the probability of y being greater than its mean is one half, so the answer to the first question is 0.5 or fifty percent. The second question simply requires us to subtract the result from the table, .4082, from .5 to find the area to the right of 1.33, which is .0918 or 9.18 percent. The third question is symmetrical with the second, so we can just use the method from the second question to find that it is also .0918. Similarly, the fourth question is symmetrical with the first step from the book example, so the answer is the answer to that first step, .4082.

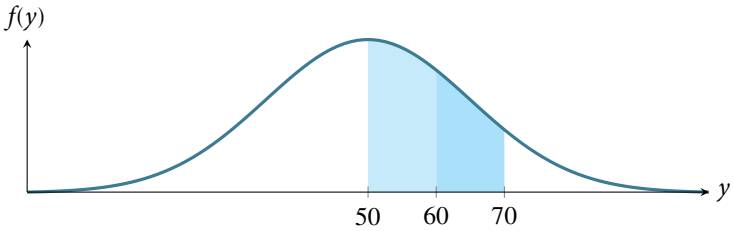
This one takes an extra step.

What is the probability that y is between 30 and 40?



Find the difference between areas.

Subtract the probability that y is between 50 and 60 from the probability that y is between 50 and 70.

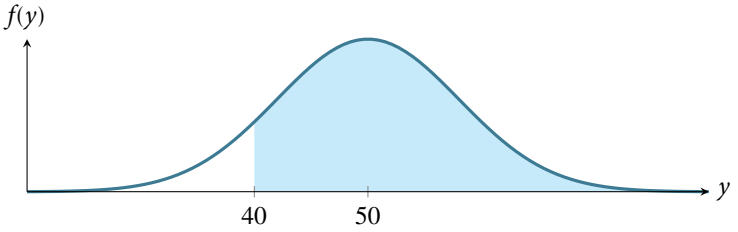


How do you find areas in z-score tables?

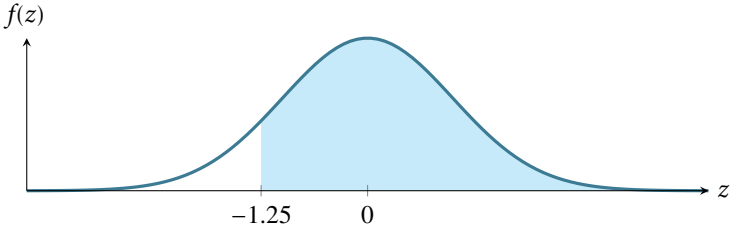
- draw picture to help you understand the question
- standardize the picture so you can use a table or a
- draw the standardized picture
- pick one of three kinds of tables / apps
- write the standardized result (may do this multiple times)
- fit the standardized result(s) into the original problem

Let's look at these steps with an example. Suppose that $y \sim N(50, 8)$. In words, this means that y has the normal distribution with true mean 50 and true standard deviation 8. Let's answer the question *What's the probability that $y > 40$?*

Step 1 is to draw a picture to make sense of the question. The picture shows the area under the curve where the scale marking is to the right of 40. This picture tells you right away that the number that will answer the question is less than 1 (the entire curve would be shaded if it were 1) and more than 1/2 (the portion to the right of 50 would be 1/2 and we have certainly shaded more than that).



Step 2 is to standardize the question so we can use a table or app to find the probability / area. We use the equation $z = (y - \mu)/\sigma$ with values from the original question: $(50 - 40)/8 = -10/8 = -1.25$. Now we know the labels that would go on a standardized picture similar to the picture above. Now we can ask the standardized question *What's the probability that $z > -1.25$?*

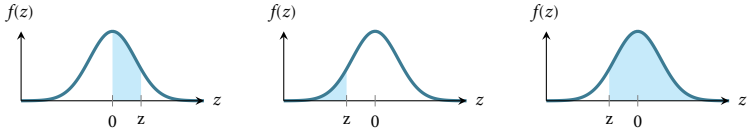


Step 3 is to draw that standardized picture. It's the same as the picture above except standardized so that it fits the tables / apps for calculating

probability / area. Now, instead of y we're looking for z and the probability associated with z on a standardized table will be the same as for y on a table for the parameters given in the original question.

Step 4 is to pick one of the three kinds of tables / apps to input the standardized z score to get a probability as output. In this example, we only have to do this step once because we only want to know the area greater than y. If we wanted to know a range between two y values, we'd need the z scores for each of them so we'd have to do it twice.

The three kinds of output from tables / apps are as follows.



The left figure shows how the table works in Mendenhall. It provides the value from 0 to $|z|$. If z is negative, using this table requires you to input $|z|$ instead. In this question, the value you get will be .3944. To get the ultimate answer to this question, Step 5 will be to add this value to .5, giving a final answer of .8944.

The middle figure shows how most tables and apps work. It gives the value from $-\infty$ to z. In this question, the value you get will be .1056. To get the ultimate answer to this question, Step 5 will be to subtract this value from 1, giving a final answer of .8944.

The right figure shows how some tables and apps work. It gives the value from z to $+\infty$. In this question, the value you get will be .8944. To get the ultimate answer to this question, Step 5 will be to simply report this value, giving a final answer of .8944.

Notice that all three types of tables / apps lead to the same result by different paths. In this case, the right figure is the most convenient but, for other questions, one of the others may be more convenient. For this class, you should probably pick one and stick with it so you can make the calculation rapidly and correctly.

Step 5, the final step is to use the value you got from a table or app in conjunction with the original picture you drew in Step 1. Since the procedure for step 5 depends on the table / app you use, I gave the procedure for Step 5 above in the paragraphs for left, middle, and right figure.

XII. ESTIMATE A POPULATION PARAMETER

Use a large-sample confidence interval.

A large-sample $100(1-\alpha)\%$ confidence interval for a population mean, μ, is given by

$$\bar{y} \pm z_{\alpha/2}\sigma_y \approx \bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

XIII. Elements of a hypothesis test

- Null hypothesis
- Alternative hypothesis
- Test statistic
- Level of significance
- Rejection region marker
- p-value
- Conclusion

The frequentist approach to statistics, which is taught in nearly all undergraduate statistics courses in the USA, is distinguished from the Bayesian approach in part by its reliance on tests of hypotheses. These tests have a binary outcome. From this point on in the course, whenever we calculate a statistic, it will be either to test a hypothesis or estimate a parameter. Any hypothesis test is taught as taking the following steps in order. In practice, some steps are omitted or reordered in ways that directly contradict the teaching of frequentist statistics.

Identify the null hypothesis.

The null hypothesis must be an assertion with a truth value that can be rejected if we see evidence refuting it. In other words, we need to be able to say that we have seen evidence that this assertion is not true. If the null hypothesis represents something unknowable or controversial, it can not meet this test.

For example, you can assert that I will never commit a crime. Unless you follow me around for my entire life and have a good grasp of the laws wherever I go, you can not resolve this. Opinions as to which is the best drink or the best candidate or the best show are problematic. If we agree on a proxy that can be tested, such as which drink will experience the most sales, which candidate will win, or which show will have the highest ratings, we can develop a testable assertion. But there are many assertions that can not be tested to everyone's satisfaction. Other examples might be which fabric feels best against the skin or which celebrity is most loved by fans. In each case, a society that worships money can easily identify a proxy for love or

other feelings but societies and certainly individuals exist who reject these proxies.

The null hypothesis must be able to be expressed in an ordinal way. By *ordinal* I mean that I have to be able to compare two values to determine which is greater. For example, I can test customer satisfaction on a scale of one to five and my null hypothesis is that the new model generates no greater satisfaction than the old model. In order to test, I ask customers to rate the models on a scale of one to five. If they all say three for both, then my null hypothesis has been supported. If they all say three for one model and two for the other model, I have seen evidence to refute my null hypothesis.

In most of this course, we will observe an additional restriction on the null hypothesis. We'll say that it must be quantifiable on a scale where the distance between 1 and 2 is the same length as the distance between 2 and 3. The customer satisfaction model given above may not fit this. A temperature scale certainly fits this notion, as does money. Even in these cases, though, we may see exceptions. Take temperature for example. Water has a boiling point and a freezing point. A famous novel takes its title from the temperature at which paper burns, *Fahrenheit 451*. Even though the markings on the scale are evenly spaced, some of the gaps portend a greater change than others. The same can be said of money. Progressive taxation arises from the belief that humans need some certain minimum income for food and shelter and that, beyond some basic level, income is more likely to be spent on luxuries than necessities. This is coupled with a belief that it is less objectionable to tax luxuries than to tax necessities. There are other ways in which the scale of money values, while evenly spaced, has some gaps that mean more than others. For instance, the basis for a major political philosophy is a division between capital and labor where the possession of a certain amount of money allows one to purchase all human necessities without the need for labor.

All these examples point to the limitations of tests of hypotheses and help to explain the rising popularity of Bayesian statistics. You should think about these things in the context of this class to give you a better idea of which problems can and can not be solved by tests of hypotheses.

The null hypothesis is the hypothesis we will postulate if the test of hypothesis is inconclusive. Therefore, the null hypothesis is chosen to minimize the possibility of error in the face of inconclusive results. Type I error is defined to be the incorrect rejection of the null hypothesis when the null hypothesis is true.

Finally, the null hypothesis is expressed as an equality. If the two quantities being measured are close enough that they may be equal, we will fail to

reject the null hypothesis. More will be said about that in step 7, stating the conclusion.

To summarize, a null hypothesis is the less controversial of the two assertions in question, it will be assumed if results are inconclusive, and it will be expressed as an equality between two scale numbers whose values are observable and agreed upon by all parties involved in the test.

Identify the alternative hypothesis.

The alternative hypothesis is a statement that motivates the test. Usually in business it will be a statement that one choice, such as of new advertising content, will have a measurably more favorable outcome than another choice, usually representing some status quo, such as current advertising content.

The alternative hypothesis has the same constraints as the null hypothesis except that it is the more controversial of the two assertions and is not measured as an equality. It may be that two quantities are not equal, that one quantity is greater than another quantity, or that one quantity is less than the other quantity.

Identify a test statistic.

The test statistic is a single number calculated from the sample. I will identify this as t_c or F_c where the subscript c means calculated. It is generally a ratio between the expected value of what is observed in the sample and the variability of the sample. Hence, smaller values of test statistics indicate higher variability in sample data than do larger values of the same statistic.

Choose a significance level.

The pioneers of frequentist statistics believed that a morally correct investigator should choose a significance level before conducting a test. They asserted that it would be cheating to choose a significance level *after* the result of the test was known.

The name *significance level* is a misnomer because the dictionary defines level as a boundary separating regions. The *significance level* is a region, not a boundary. This confusion may be heightened by the misnomer in the next step, rejection region, which *does* refer to a boundary between regions! Unfortunately, the term *significance level* is frequently used in statistics, so I will test whether you understand that it refers to a region.

Specifically, the significance level is always one of three probabilities, 0.1, 0.05, or 0.01. These three probabilities correspond to the stakes involved in the outcome of the test. If the stakes are life and death, the significance level of 0.01 should be chosen. If the stakes are money, 0.05 should be chosen.

If the test is in involved in an exploratory study without significant adverse consequences to a wrong outcome, a significance level of 0.1 is appropriate.

Find the rejection region.

The rejection region is actually *not* a region. In statistics, regions are probabilities. The rejection region marks the boundary of the region named by the significance level. I call it the rejection region marker instead of the rejection region to emphasize that it is not a probability. The textbook calls it the rejection region.

Find the p value.

The p value is the probability associated with the test statistic. It is the probability of seeing the observed level the test statistic or a larger value if the null hypothesis is true.

The statement above sounds clumsy. Therefore students try to reword it to make it easier to understand. In general, these attempts to reword it convert it into a false statement. In other words, you are not likely to better understand the above statement by converting it into a statement that is easier to parse. I will put such statements on the multiple choice exams to tempt you to choose a wrong answer. For example, it is FALSE to say that the p value represents the probability that the null hypothesis is true. Again, that's FALSE. If I put that statement on an exam as a choice, you should understand that it's false and that it differs from the correct interpretation of the p value:

It is the probability of seeing the observed level the test statistic or a larger value if the null hypothesis is true.

State the conclusion.

The conclusion can be only one of two statements:

Reject H_0 : the bla is equal to the blabla.

Fail to reject H_0: the bla is equal to the blabla.

These two conclusions contain the symbol of the null hypothesis, H_0, a colon, and a statement of the null hypothesis in words.

These two conclusions may be abbreviated as follows.

Reject H_0.

Fail to reject H_0.

Notice that the above abbreviation omits the colon. If we were to include the colon, it would imply that the next statement defines the null hypothesis. If we abbreviate the null hypothesis, we must do so unambiguously.

If we fail to reject the null hypothesis, we may follow the above statement by saying something like:

> Conclude that we have seen no evidence to refute the null hypothesis.

If we reject the null hypothesis, we may follow the above statement by saying something like:

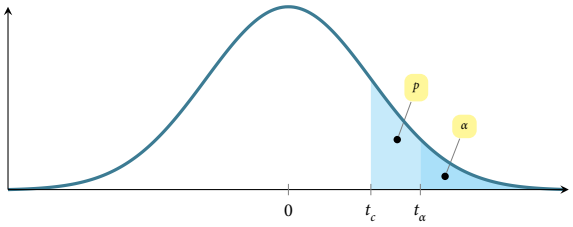
> Conclude instead that we have seen evidence to refute the null hypothesis.
>
> Conclude instead that we have seen evidence to support the alternative hypothesis.

We would never (except as a false choice in a multiple choice question) say any of these:

> We accept the null hypothesis.
>
> We accept the alternative hypothesis.
>
> We reject the alternative hypothesis.
>
> We fail to reject the alternative hypothesis.
>
> We reject the null hypothesis and conclude that we have seen no evidence to support the alternative hypothesis.

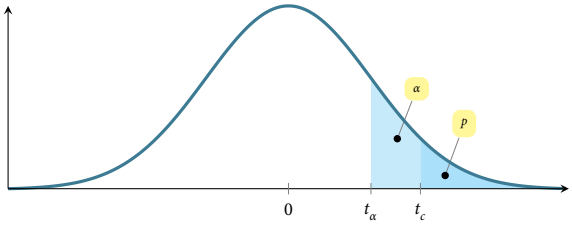
Picture the possible situations a test leads to.

The test of a hypothesis leads to two situations. The following two pictures show the possible outcomes of a t test such as we will see throughout this course.



The above possibility shows the situation where $t_\alpha > t_c$, which is equivalent to saying that $\alpha < p$. The x-axis scale is centered at the mean, 0 in this case, and values of $|t|$ to the increase from left to right. At the same time, the shaded regions decrease in size from left to right. Note that the entire shaded region above is p, while only the darker region at the right is α. In this picture α is a subset of p.

The situation pictured below is the reverse. In this situation, we reject the null hypothesis and conclude instead that we have seen evidence in this sample to support the alternative hypothesis. In this case $t_\alpha < t_c$, which is equivalent to saying that $\alpha > p$.



XIV. MISTAKES TYPICALLY MADE ON QUIZZES

Arithmetic with negative numbers causes mistakes.

- You don't like to work with negative numbers.
- A formula calls for the computation $(y_i - \bar{y})^2$ where $y_i = 15$ and $\bar{y} = 17$ so you should get $15 - 17 = (-2)^2 = 4$.
- The negative number makes you uncomfortable so you switch them around, saying $17 - 15 = 2$.
- You have to do this several times and you become confused because 15 is in the location where you expect $\bar{y}$ so you start subtracting 15 from the other values of y_i instead of 17 …

Bite the bullet and say $(15 - 17)^2 + (15 - 17)^2 + (17 - 17)^2 + (19 - 17)^2 + (19 - 17)^2$ with the mean in the same place every time!

Squares and square roots cause confusion.

- The variance is the square of the standard deviation.
- The standard deviation is the square root of the variance.
- s is the sample standard deviation.
- s^2 is the sample variance.
- σ is the population standard deviation.
- σ^2 is the population variance.

You should *not* need to take the square root of a negative number in this course! The result of a square or square root operation is a single number, not $\pm$.

Why do we use squares and square roots?

- We want what the absolute value function would give us, but we want to be able to prove that the result would always be true.
- The absolute value function is v-shaped, has a discontinuity at the point of the v, and is not differentiable at that point.
- The square and square root functions are u-shaped and everywhere differentiable, so we can use math tricks to prove that they will always work.
- Wherever you take a square root in this course, it will be to undo the effects of taking a square while leaving a positive result.
- If you get a negative result in a standard deviation formula, check your work.

Work out a standard deviation example.

Exam question: ten samples of gas mileage have been calculated for a new model of car by driving ten examples of the car each for a week. The average of the ten samples is 17. The sum of the squares of the sample is 3281. What is their standard deviation?

Recognize appropriate formulas.

The information in the problem statement hints that you should use

$$s = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}{n-1}}$$

so you write

$$s = \sqrt{\frac{3281 - 10(17)^2}{10-1}} = \sqrt{\frac{3281 - 2890}{9}} = \sqrt{43.4444} = 6.5912$$

Why was the subtraction result positive?

The sample of ten gas mileage estimates is 17, 18, 16, 20, 14, 17, 21, 13, 22, 12, 17. The sum of their squares is inevitably larger than or equal to the mean squared times the number of values. The easiest way to see this is to use a series of identical values. Hence, finding the sum of the squares is the same as calculating the mean squared times the number of values. There is no variance at all in such a sample, so it makes sense to arrive at a standard deviation of zero. Is there any way to alter such a sample so that the sum of the squared values falls below the mean? No.

What should you do when you don't understand?

The previous example was developed as an answer to the question, *what do I do if I need to do a negative square root?* You can figure out that you will never need to do so by the preceding process of finding a way to make $\sum y_i^2 = n(\bar{y}^2)$ and then trying to alter the values to decrease the left side or increase the right side.

XV. REVISIT POPULATION PARAMETERS

Make an inference about a population parameter.

At the beginning of the course, I said that statistics describes data and makes inferences about data. This course is about the latter, making inferences. You can make two kinds of inferences about a population parameter: estimate it or test a hypothesis about it.

XVI. IDENTIFY SMALL AND LARGE SAMPLES

First distinguish between small and large.

You may have a small sample or a large sample. The difference in the textbook is typically given as a cutoff of 30. Less is small, more is large. Other cutoffs are given, but this is the most prevalent.

Large samples with a normal distribution can be used to estimate a population mean using a z-score. Small samples can be used to estimate a population mean using a t-statistic.

Sampling leads to a new statistic.

If we take more and more samples from a given population, the variability of the samples will decrease. This relationship gives rise to the standard error of an estimate

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

which is not exactly the standard deviation. It is the standard deviation divided by a function of the sample size and it shrinks as the sample size grows.

XVII. ESTIMATE A POPULATION MEAN

Find a confidence interval.

If the sample is larger than or equal to 30, use the formula for finding a large-sample confidence interval to estimate the mean.

A large-sample $100(1-\alpha)\%$ confidence interval for a population mean, μ, is given by

$$\overline{y} \pm z_{\alpha/2}\sigma_{\overline{y}} \approx \overline{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

For a small sample, use t.

If the sample is smaller than 30, calculate a t-statistic and use the formula for finding a small-sample confidence interval to estimate the mean.

The t-statistic you calculate from the small sample is

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}$$

Does your calculated t fit within a $100(1-\alpha)\%$ confidence interval? Find out by calculating that interval. A small-sample $100(1-\alpha)\%$ confidence interval for a population mean, μ, is given by

$$\overline{y} \pm t_{\alpha/2}s_{\overline{y}} \approx \overline{y} \pm t_{\alpha/2,\nu}\frac{s}{\sqrt{n}}$$

Here, you are using a t statistic based on a chosen α to compare to your calculated t-statistic. The Greek letter ν, pronounced nyoo, represents the number of degrees of freedom.

Find a probability to estimate a mean.

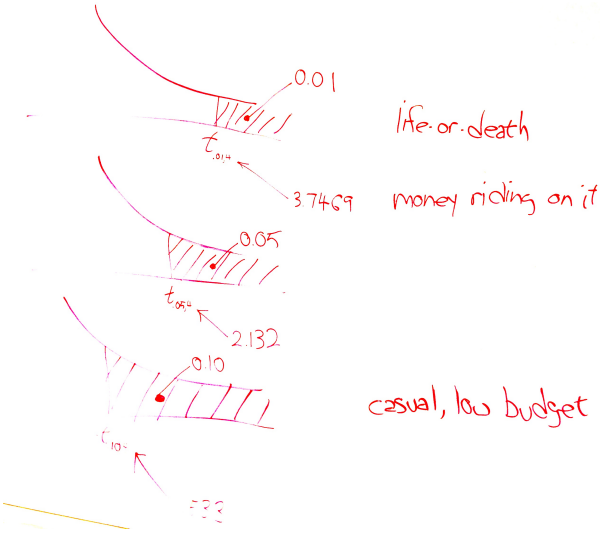
Estimating a mean involves finding a region where the mean is likely to be. The first question to ask is *how likely?* and to do so, we use a concept called alpha, denoted by the Greek letter α. Traditionally, people have used three values of α for the past century, .10, .05, and .01. These correspond to regions of $1-\alpha$ under the normal distribution curve, 90 percent of the area, 95 percent of the area, and 99 percent of the area. What we mean, for instance, if $\alpha = 0.01$ is that we are 99 percent confident that the true population mean lies within the region we've calculated, $\overline{y} \pm 2.576\sigma_{\overline{y}}$

α selection is driven by criticality.

Traditionally, $\alpha = 0.01$ is used in cases where life could be threatened by failure.

Traditionally, $\alpha = 0.05$ is used in cases where money is riding on the outcome.

Traditionally, $\alpha = 0.10$ is used in cases where the consequences of failing to capture the true mean are not severe.



The above picture shows these three cases. The top version, *life-or-death*, has the smallest rejection region. Suppose the test is whether a radical cancer treatment gives longer life than a traditional cancer treatment. Let's say that the traditional treatment gives an average of 15 months longer life. The null hypothesis is that the new treatment also gives 15 months longer life. The alternative hypothesis is that the radical treatment gives 22 months more life on average based on on only five patients who received the new treatment. A patient can only do one treatment or the other. The status quo would be to take the traditional treatment unless there is strong evidence that the radical treatment provides an average longer life. In the following picture, the shaded area is where the test statistic would have to fall for us to say that there is strong evidence that the radical treatment provides longer life. We want to make that shaded region as small as possible so we minimize the chance our test statistic lands in it by mistake.

We can afford to let that shaded area be bigger (increasing the chance of mistakenly landing in it) if only money, not life, is riding on the outcome. And we can afford to let it be bigger still if the consequences of the mistake are small. To choose an α level, ask yourself how severe are the consequences of a Type I error.

XVIII. HW 1.48: PRACTICE ESTIMATING A POPULATION MEAN

Textbook problem 1.48 describes an outlet store that purchased used display panels to refurbish and resell. The relevant statistics for failure time of the sample are $n = 50, \overline{y} = 1.9350, s = 0.92865$.

An SPSS printout accompanying the problem gives a lot more numbers, but these are the only necessary numbers to solve the problem. The standard error in the SPSS printout, .1313 can be calculated from

$$s/\sqrt{n} = 0.92865/\sqrt{50} = .1313$$

(Bear in mind that, on an exam, I could provide an SPSS printout with everything except these three quantities — $s, \bar{y}, n$ — blotted out and ask you to solve the same problem.)

(a) Part (a) of problem 1.48 asks for the 95% confidence interval, which is also given in the SPSS output, but can be calculated from the above information. Since the sample is greater than or equal to 30, we use the large sample formula:

$$\begin{aligned}\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} &= 1.9350 \pm 1.96(.1313) \\ &= 1.9350 \pm 0.2573 \\ &= (1.6777, 2.1932)\end{aligned}$$

The correct answer to part (a) is simply that pair of numbers, 1.6777 and 2.1932. (Bear in mind that the approximations given by SPSS may differ a little from those given by your calculator, largely because of different policies in rounding intermediate results. I am only likely to subtract points from your score if the difference is so great that I don't believe you used the correct inputs. This is because, in real life, your use of approximation is likely to be determined by your employer, as is the particular software or calculator you will use.)

(b) Part (b) of problem 1.48 asks for an interpretation. This result says that we are 95% confident that the true average time to failure for these panels is somewhere between 1.6777 years and 2.1932 years. It is tempting to rephrase the result. Be careful that you don't say something with a different meaning. Suppose the store wants to offer a warranty on these panels. Knowing that we are 95% confident that the true mean is in the given range helps the store evaluate the risk of different warranty lengths. The correct answer is to say that we are 95% confident that the true mean time to failure for these panels is somewhere between 1.6777 years and 2.1932 years.

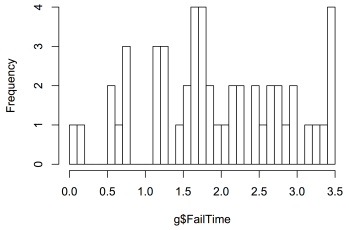
(c) Part (c) of problem 1.48 provides the standard rephrasing of the result and basically asks you to agree with it. The correct answer to part (c) is *95%*. This is because the meaning of the 95 percent confidence

interval is that, if we repeatedly resample the population, computing a 95% confidence interval for each sample, we expect 95% of the confidence intervals generated to capture the true mean.

Any statistics software can also offer a graphical interpretation, such as a stem-and-leaf plot or histogram. The stem-and-leaf plot uses the metaphor of a stem bearing some number of leaves. In the following stem-and-leaf plot, the stem represents the first digit of a two-digit number. The top row of the plot has the stem 0 and two leaves, 0 and 2. Each leaf represents a data point as the second digit of a two-digit number. If you count the leaves (the digits to the right of the vertical bar), you will see that there are fifty of them, one for each recorded failure time. You can think of each stem as holding all the leaves in a certain range. The top stem holds all the leaves in the range .00 to .04 and there are two of them. The next stem holds the six leaves in the range .05 to .09. The third stem holds all six leaves in the range .10 to .15. The stem-and-leaf plot resembles a sideways bar chart and helps us see that the distribution of the failure times is somewhat mound-shaped. The main advantages of the stem-and-leaf plot are that it is compact for the amount of information it conveys and that it does not require a graphics program or even a computer to quickly construct it from the raw data. The programming website `http://rosettacode.org` uses the stem-and-leaf plot as a programming task, demonstrating how to create one in 37 different programming languages.

```
0 | 02
0 | 567788
1 | 122222
1 | 55666667888999
2 | 0223344
2 | 666788
3 | 00233
3 | 5555
```

Most statistics programs offer many different histogram types. The simplest is equivalent to a barchart as follows.



XIX. TEST A HYPOTHESIS ABOUT A POPULATION MEAN

Choose a null and alternative hypothesis.

Testing a hypothesis must be done modestly. As an example of what I mean by modestly, consider criminal trials in the USA. The suspect on trial is considered innocent until proven guilty. The more modest hypothesis (the null hypothesis) is that the person is innocent. The more immodest hypothesis (the alternative hypothesis) carries the burden of proof.

Type I error is worse than Type II error.

The traditional view of the legal system in the USA is that if you imprison an innocent person, it constitutes a more serious error than letting a guilty person go free.

Imprisoning an innocent person is like a Type I error: we rejected the null hypothesis when it was true. Letting a guilty person go free is like a Type II error: we failed to reject the null hypothesis when it was false.

XX. HW 1.53: IDENTIFY REJECTION REGION MARKER

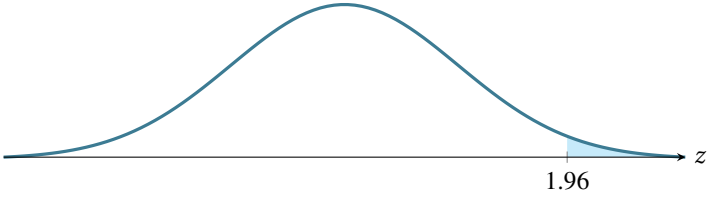
Textbook problem 1.53 asks you to identify rejection region markers. A rejection region marker is the value a test statistic that leads to rejection of the null hypothesis. It marks the edge of a region under the curve corresponding to the level of significance, α. The marker is not a measure of the size of the region. The rejection region marker can vary from $-\infty$ to $+\infty$ while the size of the region is somewhere between zero and one.

This problem invites consultation of textbook table 1.8, showing the relevant values of α and related quantities.

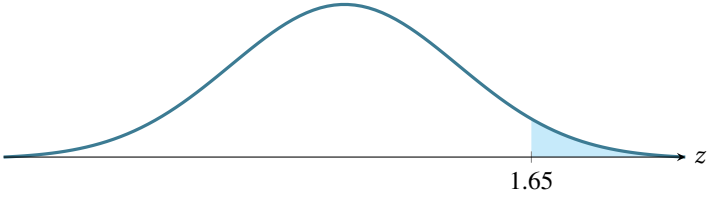
$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	1.645
.95	.05	.025	1.96
.99	.01	.005	2.576

The above table uses the same convention as the textbook tables, using measurements from 0 to $|z|$ to measure probabilities. The above table refers to two-tailed tests, but only parts (e) and (f) of the textbook problem refer to two-tailed tests. In the other four cases, the z-scores refer to z_α rather than $z_{\alpha/2}$.

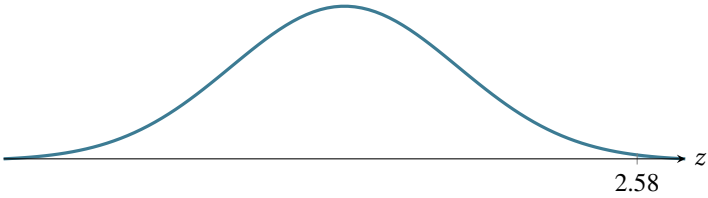
(a) $\alpha = 0.025$, a one-tailed rejection region



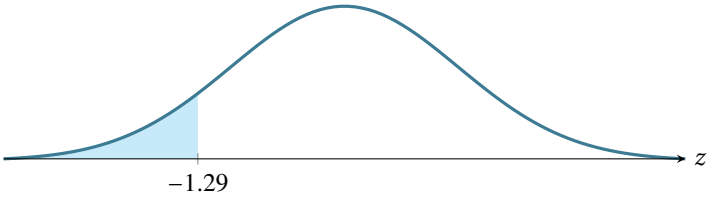
(b) $\alpha = 0.05$, a one-tailed rejection region



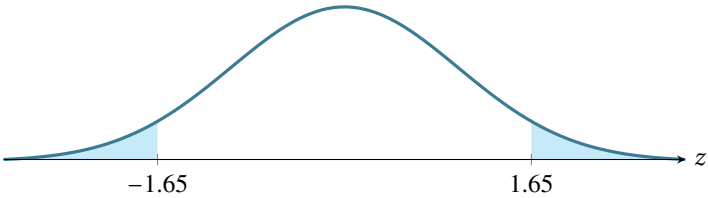
(c) $\alpha = 0.005$, a one-tailed rejection region



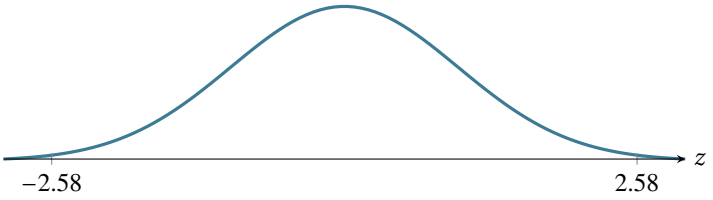
(d) $\alpha = 0.0985$, a one-tailed rejection region (the textbook may have meant this to be $\alpha = 0.1$ or perhaps as an approximation of some kind … or maybe it is simply meant to require consultation with the table in Appendix D).



(e) $\alpha = 0.10$, a two-tailed rejection region



(f) $\alpha = 0.01$, a two-tailed rejection region



XXI. HW 1.61: TEST A HYPOTHESIS ABOUT A MEAN

Textbook problem 1.61 asserts that a bulk vending machine dispenses bags known to contain 15 candies on average. Students claimed to purchase five bags containing 25, 23, 21, 21, and 20 candies. Develop a hypothesis test to consider the possibility that the data are fabricated. Use a level of significance that gives the students the benefit of the doubt.

There are seven steps to solving this problem, as follows.

Step 1. Choose a null hypothesis.

At a high level, we can say that we are trying to choose between two alternatives:

- that the students fabricated their data or
- that the students did not fabricate their data.

We need to reduce this high level view to numbers. The problem states that the machine *is known to* dispense 15 candies per bag, on average. This is equivalent to saying that the true mean is 15 or $\mu = 15$. If the students fabricated their sample, it did not come from this machine with $\mu = 15$. So we could say that one of the hypotheses would be that the sample came from a population with $\mu = 15$ and the other hypothesis would be that the sample did not come from a population with $\mu = 15$. Which should be the null hypothesis?

The null hypothesis represents the status quo, what we would believe if we had no evidence either for or against. Do you believe that students fabricated data (cheated) if there is no evidence either that they did or they didn't? Let's put it another way. Suppose you arrested a person for a crime and then realized that you have no evidence that they did commit the crime and no evidence that they did not commit the crime. Would you imprison them or let them go free? If you let them go free, it means that your null hypothesis is that they are innocent unless proven guilty. This means that if you have no evidence one way or the other, assume the students did not fabricate the data. We can translate this into the null hypothesis $\mu = 15$.

The formal way of writing the null hypothesis is to say $H_0 : \mu = \mu_0$, where $\mu_0 = 15$. Later, when refer to this population mean, we call it μ_0 because it is the population mean associated with the hypothesis H_0. So later we will say $\mu_0 = 15$.

At the end of step 1, we have chosen the null hypothesis: $H_0 : \mu = \mu_0$ with $\mu_0 = 15$.

Step 2. Choose the alternative hypothesis. The appropriate alternative hypothesis can be selected from among the three choices in either of the boxes on pages 45 and 47: $\mu < \mu_0$, $\mu > \mu_0$, or $\mu \neq \mu_0$. The appropriate choice here seems obvious: all the sample values are much larger than μ_0, so if the mean we calculate differs from μ_0 it will have to be larger than μ_0. If all the values in our sample are larger than μ_0, there is just no way their average can be smaller than μ_0.

At the end of step 2, we have determined the alternative hypothesis to be

$H_a : \mu > \mu_0$ with $\mu_0 = 15$.

Step 3. Choose the test statistic. Previously, we have learned two test statistics, z and t. We have learned that the choice between them is predicated on sample size. If $n \geq 30$, use z, otherwise use t. Here $n = 5$ so use t. We can calculate the t-statistic for the sample using the formula

$$t = \frac{\overline{y} - \mu_0}{s/\sqrt{n}}$$

We can calculate the values to use from the following formulas or by using a machine.

$$\overline{y} = \sum y_i/n = 22$$

$$s = \sqrt{\frac{\sum y_i^2 - n(\overline{y})^2}{n-1}} = 2$$

μ_0 was given to us in the problem statement and $\sqrt{n}$ can be determined with the use of a calculator or spreadsheet program. The calculated t-statistic is $t_c = (22 - 15)/(2/\sqrt{5}) = 7.8264$.

At the end of step 3, you have determined and calculated the test statistic, $t_c = 7.8264$.

Step 4. Determine the level of significance, α. You choose the appropriate value of α from the circumstances given in the problem statement. Previously in class, I claimed that there are three common levels of significance

in use as sumarized in the table on page 35: 0.01, 0.05, and 0.10. I gave rules of thumb for these three as 0.01 *life-or-death*, 0.05 *money is riding on it*, and 0.10 *casual / low budget*. In this case, the consequences for the students could be dire if they have been found to have fabricated data. In addition, the admonition to *give the students the benefit of the doubt* suggests that we want to choose the level of significance that, among the three, is most likely to give them the benefit of the doubt. The choice most likely to give them the benefit of the doubt would be the one containing the largest of the three regions supporting the null hypothesis and the smallest of the three regions supporting the alternative hypothesis: 0.01.

At the end of step 4, you have determined $\alpha = 0.01$.

Step 5. Identify the rejection region marker. This is simply a matter of calculating (or reading from a table) an appropriate t-statistic for the α you chose in the previous step. This is $t_{\alpha,\nu} = t_{0.01,4} = 3.7$. Note that ν is the symbol the book uses for df, or degrees of freedom. It is a Greek letter pronounced nyoo. For a single sample t-statistic, df= $\nu = n - 1$.

At the end of step 5, you have calculated the location of the rejection region (but not its size). It is located everywhere between the t curve and the horizontal line to the right of the point $t = 3.7$.

Step 6. Calculate the p-value. This is the size of the region whose location was specified in the previous step, written as $p = P(t > t_c)$. It is the probability of observing a t-statistic greater than the calculated t-statistic if the null hypothesis is true. It is found by a calculator or app or software. It can only be calculated by hand if you know quite a bit more math than is required for this course. In this case $p = P(t > t_c) = 0.0007195$.

At the end of step 6, we have calculated the p-value, $p = P(t > t_c) = 0.0007195$.

Step 7. Form a conclusion from the hypothesis test. We reject the null hypothesis that $\mu = \mu_0$, or in other words, we reject the hypothesis that these five bags came from a vending machine that dispenses an average of 15 candies per bag. Notice we don't conclude that the students are liars. Maybe they made a mistake. Maybe they wrote down the data wrong. Maybe someone else played a trick on them and stuck a bunch of extra candies in bags or stole the machine and replaced it with a fake. Maybe aliens landed and mass hypnotized everyone into thinking that there was more candy than there was. We don't know. We have to limit our conclusion to what we know about the data we see, which is that the data we see did not come from a machine that dispenses an average of 15 candies per bag.

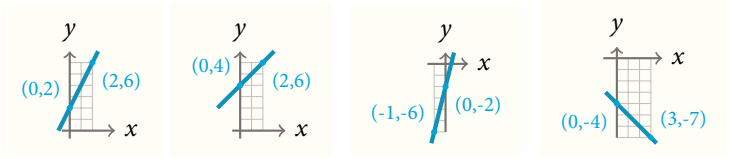
To summarize the answer to textbook problem 1.61, the seven elements of this statistical test of hypothesis are:

1. **null hypothesis** $H_0 : \mu = 15$
2. **alternative hypothesis** $H_a : \mu > 15$
3. **test statistic** $t_c = 7.8264$
4. **level of significance** $\alpha = 0.01$
5. **rejection region marker** $t_{0.01,4} = 3.7$
6. ***p*-value** $P(t > t_c) = 0.0007194883$
7. **conclusion** Reject H_0 : these 5 bags are from a machine dispensing 15 candies per bag average.

XXII. LINEAR REGRESSION INTRODUCTION

This is the 2nd of 4 parts to this course. First we reviewed basic statistics concepts. Now we will learn to find the best straight line through clouds of points. Third we'll use multiple inputs to explain a single output, and fourth we'll learn some regression concepts.

XXIII. HW 3.1: GRAPH LINES



XXIV. HW 3.3: EQUATIONS OF LINES

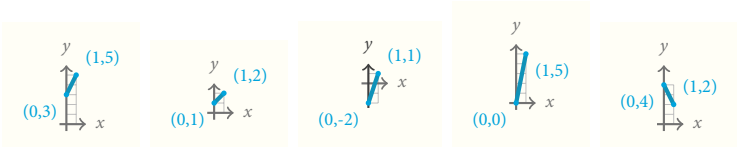
(a) $2 = \beta_0 + \beta_1 0, 2 = \beta_0. \; 6 = \beta_0 + \beta_1 2, 6 = 2 + \beta_1 2, 4 = \beta_1 2, 2 = \beta_1.$

(b) $4 = \beta_0 + \beta_1 0, 4 = \beta_0. \; 6 = \beta_0 + \beta_1 2, 6 = 4 + \beta_1 2, 2 = \beta_1 2, 1 = \beta_1.$

(c) $-2 = \beta_0 + \beta_1 0, -2 = \beta_0. \; -6 = \beta_0 + \beta_1 - 1, -6 = -2 + \beta_1 - 1, -4 = \beta_1 - 1, 4 = \beta_1.$

(d) $-4 = \beta_0 + \beta_1 0, -4 = \beta_0. \; -7 = \beta_0 + \beta_1 3, -7 = -4 + \beta_1 3, -3 = \beta_1 3, -1 = \beta_1.$

XXV. HW 3.4: GRAPHING EQUATIONS OF LINES



XXVI. HW 3.5: FINDING BETA 0 AND BETA 1

(a) 3,2 (b) 1,1 (c) -2, 3 (d) 0,5 (e) 4, -2

XXVII. THE SIMPLEST LINEAR REGRESSION MODEL

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

represents a straight line used to predict y given x with

- y-intercept β_0
- slope β_1
- error or residual ε (the Greek letter epsilon, not a Latin letter e)

This is an idealized model that we will try to estimate, so we need a separate set of symbols for our estimates of y, β_0, and β_1. These symbols are the same except that each has a circumflex or caret or hat on top: $\hat{y}, \hat{\beta}_0, \hat{\beta}_1$ and we pronounce them y hat, beta nought hat, and beta one hat. The equation

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

is our estimate of the real thing and this portion of the course focuses on identifying that estimate and assessing how good it is, which we will do using an extremely simplified example.

Errors in estimation play a central role.

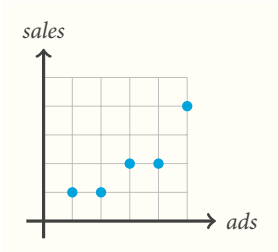
When we estimate the parameters from a set of x, y pairs, we wind up with two values of y. Those values are y_i, the observed value of y in the data, and $\hat{y}_i$, the value of y according to our model. These values will often differ, so we have to keep track of $y_i - \hat{y}_i$ or the ith residual or the ith error. These residuals or errors are the key to determining how good our model is for predicting y given x. The *sum of errors* or the *sum of squared errors* refer to these differences between each y_i and its associated $\hat{y}_i$.

XXVIII. A LINEAR REGRESSION EXAMPLE

zone	advertising (hundreds) (USD)	sales (thousands) (USD)
1	1	1
2	2	1
3	3	2
4	4	2
5	5	4

Consider the above data from textbook table 3.1 on page 93. I've modified the label for the first column. Use this data to see whether and how to predict sales from advertising expenditures. You should be able to see some positive correlation at a glance.

Graph the ad/sales data.



The positive correlation between ads and sales should be even more obvious in this picture.

The least-squares line must meet two criteria.

We're looking for the *best* line that fits these points. That line will have to fit two criteria.

1. $SE = 0$ The sum of errors is zero.
2. *SSE*, the sum of squared errors is smaller than for any other line that meets criteria 1.

These two criteria introduce two terms defined as follows

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

Beginning with the introduction of these two definition on page 96 of the textbook, the authors stop saying

$$\sum_{i=1}^{n}$$

and instead just say $\sum$ to mean exactly the same thing. Bear in mind that any operator like $\sum$ binds to the object immediately to its right and no farther to the right. You must use parentheses if you want to sum an expression like $\sum(p_i - \overline{p})$ and not have it be mistaken for the completely different

$\sum p_i - \overline{p}$ where the $\sum$ symbol would just apply to p_i. I try to use parentheses even around individual symbols for clarification but the textbook from this point on tries to simplify the appearance of the printed page by removing any symbols the authors consider redundant or unneeded. Some parentheses are required for grouping but others are just for clarity and the textbook omits the ones that are only for clarity. Hence, where I might say $\sum(p_i)$, the textbook may just say $\sum p_i$ and assume you understand that the $\sum$ symbol applies only to the p_i.

Here is an example to help you avoid mistakes with the $\sum$ symbol. Let $p_1 = 4, p_2 = 5, p_3 = 6, p_4 = 7, p_5 = 8$. The mean of these values is $\overline{p} = 6$.

Make sure you understand that

$$\sum(p_i - \overline{p})^2 \neq \sum(p_i)^2 - (\overline{p})^2$$

The best way to understand this is to calculate the two quantities.

$$\begin{aligned}
&\sum(p_i - \overline{p})^2 \\
&\quad = (4-6)^2 + (5-6)^2 + (6-6)^2 + (7-6)^2 + (8-6)^2 \\
&\quad = (-2)^2 + (-1)^2 + 0 + (-1)^2 + (-2)^2 \\
&\quad = 4 + 1 + 0 + 1 + 4 = 10
\end{aligned}$$

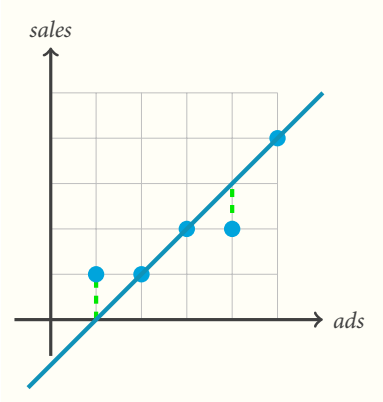
On the other hand,

$$\begin{aligned}
&\sum(p_i)^2 - (\overline{p})^2 \\
&\quad = 4^2 + 5^2 + 6^2 + 7^2 + 8^2 - 6^2 \\
&\quad = 16 + 25 + 36 + 49 + 64 - 36 = 154
\end{aligned}$$

By the way, the square operator, ^{2}, also has a higher precedence than + or −. If I wanted to find −6 squared rather than subtracting 36, I would need to say $(-6)^2$.

Many lines meet the criteria that SE = 0.

Let's look at a line that satisfies the first of the two least squares criteria. The following line meets the criteria $SE = 0$. You can see this by drawing vertical lines from the points not on the line to the line. The first one has length -1 and the second one has length +1. The other three points are on the line, so their distances from the line are all 0. Therefore



$$SE = \sum_{i=1}^{n} (y_1 - \hat{y}_i) = -1 + 0 + 0 + 1 + 0 = 0$$

Another line is demonstrably better.

The least squares line that satisfies both of the criteria has slope

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

and y-intercept

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

In practice, SS_{xy} and SS_{xx} are tedious to calculate but I want you to be able to calculate these values for a conceptual reason. That reason is that the SPSS printout for a regression contains surprisingly little information. Most of the numbers shown in the printout can be derived from combinations of the other numbers on the printout. You could actually blot out most of the numbers on a typical regression printout and a person who understood the relationships between the formulas could recreate all the blotted out numbers.

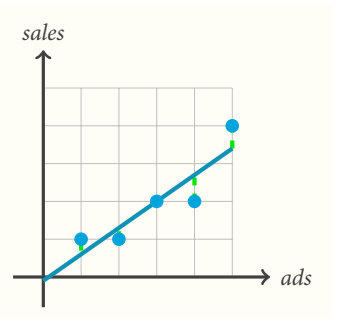
I'm going to ask you to do such calculations so that you understand how each number depends on the others.

$$SS_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y}) = \sum x_i y_i - n\overline{xy}$$
$$SS_{xx} = \sum (x_i - \bar{x})^2 = \sum x_i^2 - n(\bar{x})^2$$

Following are the tedious calculations for the above example.

$$\begin{aligned}
\sum x_i^2 &= 1^2 + 2^2 + 3^2 + 4^2 + 5^2 \\
&= 1 + 4 + 9 + 16 + 25 = 55 \\
SS_{xx} &= 55 - 5(3)^2 = 10 \\
\sum x_i y_i &= 1(1) + 2(1) + 3(2) + 4(2) + 5(4) \\
&= 1 + 2 + 6 + 8 + 20 = 37 \\
SS_{xy} &= 37 - (5)(3)(2) = 7 \\
\hat{\beta}_1 &= \frac{SS_{xy}}{SS_{xx}} = \frac{7}{10} = 0.7 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= 2 - 0.7(3) = -0.1 \\
\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x = -0.1 + 0.7x
\end{aligned}$$

Knowing the y-intercept, -0.1, and slope, 0.7, we can draw the least squares line as follows.



Revisit the errors.

x	y	$\hat{y}$	$(y - \hat{y})$	$(y - \hat{y})^2$
1	1	0.6	0.4	0.16
2	1	1.3	-0.3	0.09
3	2	2.0	0.0	0.00
4	2	2.7	-0.7	0.49
5	4	3.4	0.6	0.36
			0	1.10

Let's revisit the errors, the differences between each actual value of y and its associated predicted value. Since this is the least squares line, it must

satisfy the two criteria mention earlier. We can verify that by calculating *SE* and *SSE*. In the printouts of most statistics programs, by the way, *SSE* is referred to as the sum of squared residuals despite the *SSE* abbreviation.

XXIX. HW 3.6: USE THE METHOD OF LEAST SQUARES

The given data is $x = \{1, 2, 3, 4, 5, 6\}$ $y = \{1, 2, 2, 3, 5, 5\}$

Solution summary: **1.** Find the slope. **2.** Use the slope to find the y-intercept. **3.** Substitute the values for the slope and the y-intercept into the generic regression equation.

Solution in slightly more detail: **1.** Find the slope by finding SS_{xx} and SS_{xy} and using the formula $\hat{\beta}_1 = SS_{xy}/SS_{xx}$. **2.** Find the y-intercept by using the $\hat{\beta}_1$ value you just found in the formula $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$ **3.** Write the regression equation $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$.

Solution in painfully more detail: Step **1.** Find SS_{xx} and SS_{xy}.

x_i	y_i	x_i^2	$x_i y_i$
1	1	1	1
2	2	4	4
3	2	9	6
4	3	16	12
5	5	25	25
6	5	36	30
21	18	91	78

The elements needed for SS_{xx} and SS_{xy} are $\sum x = 21$, $\sum y = 18$, $\overline{x} = \sum x/n = 21/6 = 3.5$, $\overline{y} = \sum y/n = 18/6 = 3$, $\sum x_i^2 = 91$, and $\sum x_i y_i = 78$.

$$\begin{aligned} SS_{xy} &= \sum x_i y_i - n\overline{xy} \\ &= 78 - (6)(3.5)(3) = 15 \\ SS_{xx} &= \sum x_i^2 - n(\overline{x})^2 \\ &= 91 - (6)(3.5)^2 = 17.5 \end{aligned}$$

Step **2.** Substitute those results into the formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$.

$$\begin{aligned} \hat{\beta}_1 &= SS_{xy}/SS_{xx} = 15/17.5 = 0.85714 \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1 \overline{x} = 3 - 0.85714(3.5) = 0.0 \end{aligned}$$

Step **3.** Write the coefficients into the generic regression equation: $y = 0.85714x$. By the way, depending on the rounding policy of your calculator, you may find that the y-intercept is really in the neighborhood of 0.0000000000000001. For most purposes, that's close enough to zero to approximate, although the answer $y = 0.0000000000000001 + 0.85714x$ would also be technically correct (which is the best kind of correct).

XXX. HW 3.7: USE THE METHOD OF LEAST SQUARES

Given the data $x = \{-2, -1, 0, 1, 2\}$ and $y = \{4, 3, 3, 1, -1\}$, estimate the regression model.

Solution summary: Find the slope. Use that to find the y-intercept. Find the slope by finding SS_{xx} and SS_{xy}. Then use the formula $\hat{\beta}_1 = SS_{xy}/SS_{xx}$. Find the y-intercept by using the $\hat{\beta}_1$ value you just found in the formula $\hat{\beta}_0 = \overline{y} - \hat{\beta}_1\overline{x}$

Step 1: Find SS_{xx} and SS_{xy}.

x_i	y_i	x_i^2	x_iy_i
-2	4	4	-8
-1	3	1	-3
0	3	0	0
1	1	1	1
2	-1	4	-2
0	10	10	-12

The elements needed for SS_{xx} and SS_{xy} are $\sum x = 0$, $\sum y = 10$, $\overline{x} = \sum x/n = 0/5 = 0$, $\overline{y} = \sum y/n = 10/5 = 2$, $\sum x_i^2 = 10$, and $\sum x_iy_i = -12$. Substitute the above values into the following formulas.

$$\begin{aligned} SS_{xy} &= \sum x_iy_i - n\overline{xy} \\ &= -12 - (5)(0)(2) = -12 \\ SS_{xx} &= \sum x_i^2 - n(\overline{x})^2 \\ &= 10 - (5)(0)^2 = 10 \end{aligned}$$

Step 2. Substitute those results into the formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$.

$$\begin{aligned} \hat{\beta}_1 &= SS_{xy}/SS_{xx} = -12/10 = -1.2 \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1\overline{x} = 2 - 1.2(0) = 2 \end{aligned}$$

Step 3. Write the generic regression equation: $y = 2 - 1.2x$.

Examine the coefficients section of regression analysis.

Given what we have already learned, we can interpret the output of the **Coefficients** section of the output of any computerized regression analysis. This section is always presented as a table with as many rows as there are coefficients in the regression model. Since we are only considering the simplest case, $y = \beta_0 + \beta_1 x + \varepsilon$, there are only two rows, one for β_0 and one for β_1.

The first column labels each row of the coefficients table.

Some statistics software will number the rows, starting at zero. Almost all will label them with the name of the associated x. Of course, β_0 has no associated x, so it is usually labeled as `Constant` because it does not produce an output that varies with the sample values. For every x and y pair, the value of β_0 is simply added to the rest of the equation. That's not the case for β_1 or any additional coefficients of predictors we learn to use later. In each of those cases, the coefficient is multiplied by the predictor before it is added to the result.

The second column provides the value of the coefficient.

After the label, the next column contains the value of the coefficient. That column is usually labeled b, a Latin letter that corresponds to the Greek letter β. As with other pairs of Greek and Latin letters, β and b refer to the concept and what we see, respectively. In other words, there is some true y-intercept, β_0, and there is an estimated y-intercept from the data, b_0. By the way, the Mendenhall textbook calls this $\hat{\beta}_0$ instead of b_0 but almost all other statistics books and software call it b_0. Because we're using Mendenhall and SPSS, we will use either of those two terms without making any distinction.

The third column provides the standard error of the coefficient.

The estimated standard error of the slope is

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

or, in the terminology of most statistics software

$$s_{b_1} = \frac{s}{\sqrt{SS_{xx}}}$$

because $\hat{\beta}_1 = b_1$.

The standard error of the y-intercept is also given, but may be of little interest unless we have some evidence for the appropriate value of y corresponding to the absence of the predictor. You have to evaluate this on a case-by-case basis.

The fourth column provides a different version of standard error.

We will not use that column in this course.

The fifth column provides a t-statistic for the coefficient.

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

Notice that the statistic is an estimate of the coefficient divided by a measure of the variability of that estimate. If the estimate comes from a sample with high variability, we want to take the estimate less seriously. A small value of the statistic corresponds to that situation since a large denominator diminishes the overall size of the fraction.

The sixth column provides a p-value for the coefficient.

The final column, often labeled *significance* or *Sig* represents the p-value associated with the t-statistic in the previous column. In other words, these two columns provide the same information in two different formats. It is traditional in psychology, management, and marketing to present only the p-value in sentences discussing results of regression analysis. There are also customary shorthand notations for categories of p-values, corresponding to regions of 0.01, 0.001, and 0.0001, often denoted $*$, $**$, and $***$.

This may explain why outputs we look at in class show a p-value of 0.000 or with asterisks replacing the p-values. The custom implemented in many software packages is to regard every region smaller than 0.001 as identical and to call them 0.000 or $***$. This doesn't change the size of the region at all. It only reflects customs in communities.

XXXII. HW 3.11: COMPARE SCATTERPLOTS

Textbook problem 3.11 has five parts.

(a) Does the scatterplot of cooperation and payoff show a linear relationship? No. What we can see from this scatterplot is that most players cooperate between 40 and 50 times and and that the average payoff for doing so ranges from -0.4 to +0.8.

(b) Does the scatterplot of defection and payoff show a linear relationship? No. We can see that most people defect about 15 to 19 times, with payoffs from -0.4 to $+0.8$.

(c) Does the scatterplot of punishment and payoff show a linear relationship? Yes. The dashed line drawn on the picture is meant to emphasize this.

(d) Is the relationship in (c) positive or negative? Negative. For each unit of increase in x, y decreases by an amount less than one.

(e) Do you agree with the author's conclusion that winners don't punish? No. I don't. You don't get any points off for disagreeing with me but I see two issues. First, punishment occurs very rarely, judging by the x axes of the three scatterplots. Therefore, it would be just as accurate to say that losers don't punish, or short people don't punish. The plain fact is that no one punishes. The second issue is the notion of *winners*. How do you define winners? Since punishment happens so infrequently, you could (assuming you have no information about the players) construct different stories about whether players punished because they were losing or lost because they punished. There's no information contributing to cause and effect. Looking at the first two scatterplots, we see no evidence of a winning strategy. Most people chose cooperation most of the time, and with mixed results. Most people chose defection less often, again with mixed results. In my mind, the problem sends us two signals. First it motivates the study of multiple regression as a way to better understand how these three strategies go together in the case of any particular player. Second, it reminds us that correlation is a necessary but not sufficient condition for causation.

XXXIII. CHECK THE 4 MODEL ASSUMPTIONS FOR ANY GIVEN MODEL

The textbook lists four assumptions about ε, the error term in the regression model. A major goal in modeling is to examine errors to identify patterns and to modify the regression model to incorporate any patterns discovered. We will learn how to add additional predictor variables to account for such patterns when we study multiple regression. The four assumptions are as follows:

1. The mean of the probability distribution of ε is 0. If we could take an infinite number of samples, the errors would average 0. This is why we often write the regression equation without the error term. The expected value of the regression equation is $E(y) = \beta_0 + \beta_1 x + 0$.

2. The variance of the probability distribution is a constant for all values of x. In linear regression, we call the true value of that constant σ^2, the variance of ε and we estimate it with the sample variance s^2. We can compute the sample variance using the sum of squared residuals or *SSE*. The

formula for their relationship is $s^2 = SSE/(n - (k + 1))$. Mendenhall says $n - 2$ in Chapter three, then, in Chapter four introduces k and belatedly informs us that $k = 1$ in simple linear regression with a single slope parameter. That explains why the Mendenhall book switches from $n - 2$ to $n - (k + 1)$.

3. The probability distribution of ε is normal. Write this in the notation we used in review as $\varepsilon \sim N(0, s)$ where $s = \sqrt{SSE/(n - (k + 1))}$.

4. The errors are independent of each other. The error associated with one value of y does not affect the error associated with another value of y. One way to understand this might be to consider a process that has measurement error that gets worse and worse over time. Many types of physical measurement processes involve heat and friction and can wear out the measuring device over time. At the beginning of this course, we briefly considered crush testing of cardboard boxes, where the machine that crushes the box measures the amount of force it took to crush the box. Such a machine would have to be very robust to give accurate measurements. Suppose, though, that the machine remains in service for longer than it was intended to last. The relentless heat and friction might eventually cause larger and larger errors until it ceases to function altogether. In that case, the magnitude of one error might become a function of the magnitude of the previous error.

XXXIV. HW 3.17: SUM OF SQUARED RESIDUALS

Problem statement: Suppose you fit a least squares line to nine data points and calculate $SSE = 0.219$. (a) calculate s^2. (b) calculate s.

Solution: Use the formula $s^2 = SSE/(n - (k + 1))$ with $k = 1$ and $s^2 = 0.219/(9 - (1 + 1)) = 0.0312857$. $s = \sqrt{0.0312857} = 0.17687768$

XXXV. HW 3.18: VARIANCE OF EPSILON

Problem statement

Find SSE, s^2, and s for (a) problem 3.6 and (b) problem 3.7, and interpret s.

Solution

First, make sure you understand that s^2 is the variance of ε, the error term in the regression equation. s is the standard deviation of ε, and SSE the sum of squared residuals is a version of that variance but not scaled by sample size and parameters. The formula for the relationship is $s^2 = SSE/(n - (k+1))$ where k is the number of slope parameters. In this chapter we only consider regression equations with one slope parameter, β_1. The 1 in $k + 1$

refers to the y-intercept, β_0. Both β_1 and β_0 are dependent on the sample, so if we estimate them using the sample, we lose two degrees of freedom, just as we lost one degree of freedom when we estimated a mean from the sample.

Mendenhall and SPSS use a variety of synonyms including:

$$
\begin{aligned}
s &= \sqrt{MSE} \\
s &= \text{ROOTMSE} \\
&= \sqrt{SSE/(n-(k+1))} \\
&= \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{(n-(k+1))}} \\
&= \text{standard deviation of } \varepsilon \\
&= \text{standard error of the estimate} \\
&= \text{estimated standard error of the regression model} \\
s^2 &= MSE \\
&= \text{mean squared error} \\
&= \frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{(n-(k+1))} \\
&= SSE/(n-(k+1)) \\
&= \text{variance of } \varepsilon \\
&= \text{Var}(\varepsilon)
\end{aligned}
$$

Use the relationship $SSE = SS_{yy} - \hat{\beta}_1 SS_{xy}$ (found in Mendenhall in the box on page 106) and the fact that you previously calculated some of these quantities for hw 3.6 and hw 3.7.

For (a) problem 3.6, in addition to the previously calculated quantities, you need $SS_{yy} = \sum y_i^2 - n(\overline{y})^2 = 68 - 6(3)^2 = 14$

$SSE = 14 - 0.85714(15) = 1.1429$, $s^2 = SSE/(n-(k+1)) = 1.1429/4 = 0.28$, $s = 0.53$. We expect most (approximately 95%) of the observed y-values to lie within $2s = 1.06$ of their respective least squares predicted values, $\hat{y}$. It is important for you to understand that we are not talking about $\overline{y}$ here. We are actually talking about the individual predicted values of y and how they relate to the observed values of y. When we substitute an observed or planned value of x into our estimated regression model, we call the resulting y a predicted value. We developed the estimated regression model by looking at observed y values. In a table of observed x and y values, we can therefore add a third column of $\hat{y}$ values predicted by the equation.

The difference between each observed y value and its corresponding value of $\hat{y}$ is an error term.

Note that Mendenhall does not define k until the next chapter. Instead, Mendenhall says $n-2$ in this chapter. Students sometimes ask what the 2 stands for. It actually stands for the number of parameters in the regression equation. The parameters are β_0 and β_1. In the next chapter we'll encounter more parameters and use the symbol k to distinguish between slope parameters and the y-intercept parameter.

For (b) problem 3.7, in addition to the previously calculated quantities, you need $SS_{yy} = \sum y_i^2 - n(\bar{y})^2 = 36 - 5(2)^2 = 16$ $SSE = 16 - (-1.2)(-12) = 1.6$, $s^2 = SSE/(n-2) = 1.6/3 = 0.5333$, $s = 0.73$ We expect most (approximately 95%) of the observed y-values to lie within $2s = 1.46$ of their respective least squares predicted values, $\hat{y}$.

XXXVI. HW 3.21: READING REGRESSION ANALYSIS OUTPUT

(a) $y = 119.9 + 0.34560x + \varepsilon$
(b) It's where it says "The regression equation is …"
(c) The first assumption is that the mean of the probability distribution of ε is 0. The second assumption is that the variance of ε is constant for all values of x. The third assumption is that ε is normally distributed. The fourth assumption is that the errors associated with any two observations are independent of each other.
(d) $s = 635.187$
(e) The range is $2s = 1270.374$

XXXVII. HW 3.23 DOES BETA SUB ONE EQUAL ZERO?

Page 113 of Mendenhall:

(a) Test $H_0 : \beta_1 = 0$ at $\alpha = 0.05$ using the data from problem 3.6 as follows:

x	1	2	3	4	5	6
y	1	2	2	3	5	5

The easiest way to solve this problem is to use the `LinRegTTest` function in TI calculators. That function is the same as the linear regression function except that it provides $t_{\hat{\beta}_1}$ and the associated p value. In this case, those values are 6.7 and 0.01285 respectively. Since $p < \alpha$ you conclude that you reject H_0.

You can solve it by hand since there are only six observations. To obtain $t_{\hat{\beta}_1}$, first consider its definition,

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

As for p, it can be calculated from $t_{\hat{\beta}_1}$. You've calculated most of the needed quantities in problem 3.6 except for s and that can be found by calculating

$$s = \sqrt{s^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n-(k+1)}}$$

and that requires that we find

$$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = SS_{yy} - \hat{\beta}_1 SS_{xy}$$

Among these, the only quantity not calculated in 3.6 is $SS_{yy} = \sum y_i^2 - n\bar{y}^2$ for which we need $\sum y_i^2 = 1^2 + 2^2 + 2^2 + 3^2 + 5^2 + 5^2 = 68$. So

$$\begin{aligned}
SS_{yy} &= \sum y_i^2 - n\bar{y}^2 = 68 - 6(3^2) = 68 - 54 = 14 \\
SSE &= SS_{yy} - \hat{\beta}_1 SS_{xy} = 14 - (0.85714)(15) = 1.1429 \\
s &= \sqrt{\frac{SSE}{n-(k+1)}} = \sqrt{\frac{1.1429}{6-(1+1)}} = .5345 \\
t_{\hat{\beta}_1} &= \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}} = \frac{0.85714}{.5345/\sqrt{17.5}} = 6.7121
\end{aligned}$$

(b) This problem is exactly the same as part (a) except with the data from problem 3.7, and can be solved exactly as above, giving $t_{\hat{\beta}_1} = -5.196$ and $p = 0.01385$ and leading to the same conclusion as in part (a), that is, reject H_0.

XXXVIII. HW 3.24 INTERPRET COMPUTERIZED CONF INTVL OUTPUT

(a) Requires you to read the output at the bottom of the page and report the appropriate conclusion for a test of $\beta_1 = 0$ at $\alpha = 0.01$. The appropriate conclusion is to reject H_0, based on $p < 0.001$ which is the smallest p SPSS will display by default.

(b) The 95 percent confidence interval for the slope is (1.335,1.482). The practical interpretation is cumbersome but if you try to condense it

you will misstate it. The cumbersome but true interpretation is that, if we were to take 100 samples each with 76 properties in the sample, we would expect the slope to fall between 1.335 and 1.482 in 95 of those 100 samples.

(c) We are asked how to narrow the confidence interval and the answer provided by Mendenhall at the top of the same page is to use a larger sample size to obtain a smaller confidence interval.

XXXIX. HW 3.31 CONDUCT A SIMPLE LINEAR REGRESSION

The problem is to conduct a simple linear regression from 16 pairs of points where x is an empathy test and y is a level of brainwave activity. Then use the linear regression analysis to answer the question of whether people who score higher in empathy show higher pain-related brain activity.

We do see evidence that the answer to the research question is yes, but with an important caveat. First, a simple linear regression shows a positive slope, 0.03618, indicating that the relationship between x and y is positive. Next, we see that the p-value for the slope is 0.00927, showing evidence that the true slope is non-zero and giving us confidence in the estimated slope. But then we can't help but notice that $R^2 = 0.3937$, indicating that the model explains about 39 percent of the variability in brain activity. So a serious researcher would be advised to look for additional explanatory variables to form a more complete model.

XL. FIND AND INTERPRET THE CORRELATION COEFFICIENT

The Pearson Product-Moment Correlation Coefficient, r, is a calculated sample statistic with a population counterpart called ρ, a Greek letter pronounced roh. The correlation coefficient is a scaled version of the covariance between two variables. r is scaled to be anywhere from -1 to 1.

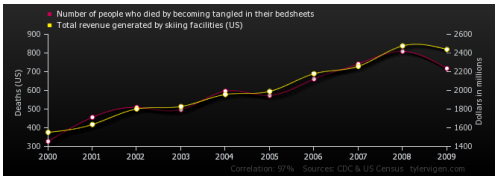
$r = 1$ means that two variables are perfectly positively correlated.

$r = -1$ means that two variables are perfectly negatively correlated.

$r = 0$ means that two variables are not correlated.

An important distinction sometimes missed by students is that, for many purposes, being negatively correlated is the same as being positively correlated. For instance, if y and x are perfectly negatively correlated, then x is just as useful to predict y as if they are perfectly positively correlated. As a result, if I ask which of several data sets are least correlated, I mean the pair with r closest to 0, *not* the pair with r closest to -1.

A second issue is that correlation, either positive or negative (but not 0) is a necessary but not sufficient prerequisite for causation. A website called *Spurious Correlation* produces graphics like this one, chronicling the alarming relationship between deaths by becoming tangled in bedsheets and ski resort revenue.



Here are three datasets, each one corresponding to one of the above three conditions.

Data set 1: $r = 1$

x: 1, 2, 3, 4, 5

y: 1, 2, 3, 4, 5

These two variables are perfectly correlated. I can show this by calculating

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

First, find the sum of the x values: 15.

Second, find the mean of the x values, the above sum divided by the number of values: $15/5 = 3$.

Third, find the sum of the squares of the x values: $1^2+2^2+3^2+4^2+5^2 = 55$. SS_{xx} is this sum minus the number of values times the square of the mean of the values: $55 - (5)(3^2) = 10$.

Since x and y are identical, the formulas for SS_{yy} and SS_{xy} are the same, and $r = 10/\sqrt{10 \cdot 10} = 1$.

Data set 2: $r = -1$

x: 1, 2, 3, 4, 5

y: 5, 4, 3, 2, 1

These two variables are perfectly negatively correlated. I can show this by calculating

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

First, find the sum of the x values: 15.

Second, find the mean of the x values, the above sum divided by the number of values: $15/5 = 3$.

Third, find the sum of the squares of the x values: $1^2+2^2+3^2+4^2+5^2 = 55$. SS_{xx} is this sum minus the number of values times the square of the mean of the values: $55 - (5)(3^2) = 10$.

Since x and y are identical except for the order, the formula for SS_{yy} is the same. However, the fact that the ordre is different causes SS_{xy} to be the different. Calculating $\sum x_i y_i$ gives $(51) + (24) + (33) + (42) + (15) = 35$ and $\sum x_i y_i - n\overline{xy} = 35 - (5)(3)(3) = -10$

So the formula for r with this data is $r = -10/\sqrt{10 \cdot 10} = -1$. Warning: if any statistics software provides $r = 1$ instead of $r = -1$ for the above data, the software is wrong, not the formula.

Data set 3: $r = 0$

x: 1, 2, 3, 4, 5

y: 3, 3, 3, 3, 3

These two variables are not correlated. I can show this by calculating

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

The formula for SS_{xy} requires $\sum x_i y_i = (13)+(23)+(33)+(43)+(53) = 45$ and $\sum x_i y_i - n\overline{xy} = 45 - (5)(3)(3) = 0$ so we are mercifully spared calculating the other two values. There's a technical problem with the denominator anyway since the y values don't vary at all from each other, leading to a denominator of 0. Division by zero is not defined in calculators and computers. Conceptually, you can think of the denominator as approaching 0 and the entire fraction as approaching infinity. Mercifully, we can avoid thinking about that here since the numerator is 0, so we can simply say $r = 0$ and leave it at that.

XLI. FIND AND INTERPRET THE COEFFICIENT OF DETERMINATION

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

XLII. HW 3.34 CORRELATION: POSITIVE, NEGATIVE, WEAK

Correlation may be positive or negative. Strong negative correlation (approaching -1) may be just as good for prediction as strong positive correlation (approaching -1). Weak correlation, approaching 0 from either the positive or negative side, has no predictive value. r^2 is necessarily in the range $[0, 1]$ while r is in the range $[-1, 1]$. It is the absolute value of r that signals the strength of the correlation. You could say that correlation is strong if $r^2 \to 1$ but, unless you know the underlying value of r from which r^2 was calculated, you don't know if it is strong positive or negative correlation.

Describe the slope of the least squares line in the following cases of r

(a) $r = \ \ 0.7$ implies a positive slope
(b) $r = -0.7$ implies a negative slope
(c) $r = \ \ 0$ implies a slope of 0
(d) $r^2 = 0.64$ is equivocal—the slope could be positive or negative

XLIII. HW 3.35: COEFFICIENTS OF CORRELATION AND DETERMINATION

Find r and r^2 for problems 3.6 and 3.7. We can use values calculated in the previous problems.

(a) $15/\sqrt{17.5\ (14)} = .9583$, $r^2 = .9184$
(b) $-12/\sqrt{10\ (16)} = .94868$, $r^2 = .8999$

XLIV. HW 3.48: PREDICTION AND CONFIDENCE INTERVALS

(a) A prediction interval is always larger than the associated confidence interval because it contains two errors, one of which is error accounted for by the confidence interval, the error in estimating the mean value of the predicted quantity. That error is the distance between the least squares line and the true mean. In addition to that error, a prediction interval includes the error associated with predicting a particular value, the distance from that value to its mean.

(b) An easy way to understand why the confidence interval for y grows as x diverges from its mean is to draw a graph showing two lines. First, draw the least squares line. Next, draw another line that intersects that line. Any error in $\hat{\beta}_1$ means that the two lines cross. Any two lines that cross diverge more as we move away from the crossing point. The implication is that it's risky to draw inferences outside the range of the sample.

Given

$$\begin{aligned} n &= 20 & SS_{xx} &= 4.77 \\ \hat{y} &= 2.1 + 3.4x & SS_{yy} &= 59.21 \\ \overline{x} &= 2.5 & SS_{xy} &= 16.22 \\ \overline{y} &= 10.6 \end{aligned}$$

(a) Find SSE and s^2 for which you can use the formulas

$$\begin{aligned} \hat{\beta}_1 &= SS_{xy}/SS_{xx} = 16.22/4.77 = 3.4 \\ SSE &= SS_{yy} - \hat{\beta}_1 SS_{xy} = 59.21 - 3.400419(16.22) = 4.055 \\ s^2 &= SSE/(n-2) = 4.055/(20-2) = 0.23 \end{aligned}$$

For the next three questions, use the formula

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

bearing in mind that $s = \sqrt{MSE}$, not the s derived from the variability of a single variable around its mean.

In all three following cases, the interpretation should be that, if we repeated the experiment 100 times, i.e., took 100 different samples of 20 each, we would find that $\hat{y}$ would lie within the CI in 95 of those 100 times (a total of 2000 observations divided into 100 groups of 20).

(b) Find a 95 percent CI for $E(y)$ if $x = 2.5$. Interpret. $10.6 \pm .233$

(c) Find a 95 percent CI for $E(y)$ if $x = 2.0$. Interpret. $8.9 \pm .319$

(d) Find a 95 percent CI for $E(y)$ if $x = 3.0$. Interpret. $12.3 \pm .319$

(e) What happens in the above three intervals as x diverges more and more from $\overline{x}$? The confidence interval grows wider.

(f) Find a 95 percent PI (not CI) for some y if $x = 3.0$. Interpret. In this case, use a slightly different formula to account for the additional error of a particular observation diverging from its mean.

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

12.3 ± 1.046 The interpretation is as above except that it is just some value of y that lies within the interval in 95 out of 100 experiments, not the mean. Therefore, the interval is inherently wider than it would be if it had to encompass only the mean.

XLVI. HW 3.60: INTERPRET A REGRESSION ANALYSIS

Coefficients

Model	B	StdError	Beta	t	Sig.
(Constant)	250.14203	14.23101		17.58	2e-16
DISTANCE	-0.62944	0.04759	UNR	-13.23	8.48e-16

The regression analysis output for this problem is above. The key features of the output are that the slope of the relationship between distance and accuracy is negative (−0.62944) and the p-value associated with the slope is $8.48e-16$, making it a very credible estimate at any α level you could choose. You would certainly advise the golfer that an overemphasis on distance is linearly related with a decrease in accuracy, although the data don't explore other reasons for the relationship. There could easily be a third variable that is the causal factor and that third variable could be influencing the other two.

XLVII. HW 3.61: CONDUCT A LINEAR REGRESSION ANALYSIS

This problem uses the GMAC dataset, available on blackboard in the datasets.zip file. The textbook prints the first few observations but the entire dataset contains 2,087 observations. If you do a regression with just the first few observations printed in the textbook, you can not hope to approximate the correct answers.

A regression analysis of the data using R gives the following information.

```
> load("GMAC.Rdata")
> attach(GMAC)
> summary(GMAC)
    WLBSCORE         HOURS
 Min.   : 8.54   Min.   :  2.00
 1st Qu.:36.75   1st Qu.: 45.00
 Median :44.51   Median : 50.00
 Mean   :45.07   Mean   : 50.26
 3rd Qu.:54.74   3rd Qu.: 55.00
 Max.   :75.22   Max.   :100.00
> summary(lm(WLBSCORE~HOURS))

Call:
lm(formula = WLBSCORE ~ HOURS)

Residuals:
    Min      1Q  Median      3Q     Max
-35.477  -8.412  -0.652   8.121  33.525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 62.49851    1.41351   44.22   <2e-16 ***
HOURS       -0.34673    0.02761  -12.56   <2e-16 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '. 0.1 ' '  1
```

```
Residual standard error:
12.28 on 2085 degrees of freedom
Multiple R-squared:  0.07033,
Adjusted R-squared:  0.06988
F-statistic: 157.7 on 1 and 2085 DF,
p-value: < 2.2e-16

> anova(lm(WLBSCORE~HOURS))
Analysis of Variance Table

Response: WLBSCORE
            Df Sum Sq Mean Sq F value    Pr(>F)
HOURS        1  23803 23803.1  157.73 < 2.2e-16 ***
Residuals 2085 314647   150.9
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '. 0.1 ' '  1
> length(HOURS)
[1] 2087
```

A professional report would translate these numbers into a narrative. In this course, I will only require that you recognize the elements of this narrative rather than write them yourself. In multiple choice questions, you will have the opportunity to choose between true and false statements about regression analysis as the means to indicate you understand this language.

An informative exercise for this dataset is to do a scatterplot of the work-life balance score with number of hours worked. It should be clear from such a scatterplot that, although there is a negative linear relationship between the two variables, it has no practical value. The result stems from the fact that the small number of people working 80 hours or more per week are uniformly disatisfied with their work-life balance. The vast majority working fewer hours have work-life balance scores all over the place. The hours variable is not a good predictor even though it passes formal tests.

Part of the trouble here is an age-old problem in statistics which is that, with enough data, even a feeble pattern will pass tests. Another part of the trouble, though, is something we can fix by doing what we plan to do in the next chapter. The single variable hours is not enough to model the concept of work-life balance and, in the next chapter, we'll try to model a dependent variable using several indpendent variables that may each tell part of the story.

XLVIII. REVIEW OF PREVIOUS PART OF COURSE

We are now embarking on the third and most important part of DSC 330, multiple regression. It's important to take stock of where we've been so far because we'll use all the concepts from the previous part, the only difference being that matters become more complicated by the *multiple* aspect of multiple regression.

We started the course estimating y using a sample to estimate the expected value of y, $\bar{y}$, and using the sample to estimate the variability of y as s.

Next, we considered the possibility that we might arrive at a better estimate of y by examining another variable, x, that is related to y. We developed a model of the relationship between x and y and we learned to

- create the model
- assess the model
- use the model

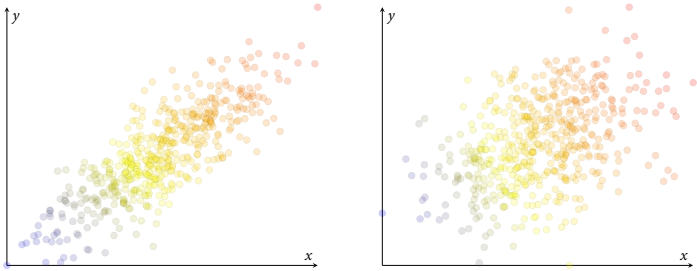
It's worthwhile, going forward, to recall the concepts used for each of these three phases.

Create a linear regression model

For this, the most prevalent tools have been SS_{xx} SS_{yy} and SS_{xy} in the development of the simplest model. From these we can obtain the regression equation (what we'll use) and the statistics about the variability in the model (how we'll assess).

Going forward, we will use computers or calculators for this phase almost exclusively. As the number of variables and parameters increase in multiple regression, the manual calculations become laborious. Yet you need to know that SPSS or whatever software you use is doing these simple but extremely lengthy computations. There is no mystery, just detail.

Assess the created model



We would like to know whether we are looking at the picture on the left or right. The picture on the left shows a stronger relationship between x and y than does the one on the right.

Another way to say this is that the picture on the left shows a relationship between x and y that gives us more confidence about our predictions than does the picture on the right. Notice that, if we predict a value to lie along a line at the center of these two clouds, the samples indicate that the case on the right is more likely to be close to any realized value than the one on the right. We need a number that distinguishes between these two pictures. A

t-statistic for the slope will be *larger* for the picture on the left than for the picture on the right. (These two pictures are of data that differs only in how dispersed it is around $\overline{x}$ and $\overline{y}$.)

Equivalently, the p-value corresponding to the t-statistic for the slope on the left will be *smaller* than that for the p-value on the right.

These two are complementary pieces of information to assess the model, the t-statistic for the slope parameter, and the p-value indexed by that t-statistic. These measures tell us in a single number, how dispersed the sample data is around the slope. Therefore, these two numbers tell us whether we should have confidence in the sample estimate of the slope.

Use the assessed model

To use the assessed model, we either predict or estimate. Predicting requires us to use the regression equation we've developed and to insert an x value into it. The result tells us how much y will change in response to the amount by which we've increased x.

XLIX. MULTIPLE REGRESSION

Now we hope to obtain an even better estimate of y using multiple variables, $x_1, \ldots, x_k$. To do so requires additional tools.

First, consider the tools to create the model. Instead of SS_{xx} and SS_{xy} we have to account for the variability of more than one x and the relationship between more than one x and y. To do so, we'll partition the variability of the regression model into SSR, the variability explained by the model, SSE, the variability in the data for which the model can not account, and SST, the sum of these two quantities.

Next, consider the tools for assessing the model. Instead of t, the main tool is F. Also, we'll use R^2 and, especially, a version of R^2 that is resistant to a flaw of R^2, namely that it is susceptible to manipulation by adding extraneous predictors.

L. THE F DISTRIBUTION

The F distribution plays the role, for multiple predictors, that the t distribution played in the case of a single predictor. The F-statistic is more complicated than the t-statistic and the F tables are lengthier and more cumbersome. The F distribution has three parameters in place of the two parameters of the t distribution, called

- α,
- $\nu_1 = k$, and

- $v_2 = n - (k + 1)$.

The latter two are degrees of freedom and include v_1, pronounced *nyoo one*, referring to numerator degrees of freedom, and v_2, pronounced *nyoo two*, referring to denominator degrees of freedom. There are as many numerator degrees of freedom as there are predictors, identified as k, and the denominator degrees of freedom is equal to the sample size, minus the count of predictors as well as the y-intercept.

It may be necessary to calculate F_{α,v_1,v_2} in any quiz where you are required to use SPSS. Unfortunately, the method for calculating it is cumbersome. Instead I will demonstrate the calculations in R Studio, available on all the lab machines. Open R Studio from the All Programs menu. You will see a prompt like a greater-than sign, >. At this prompt, type any of the following expressions and press the Enter key.

Given α, v_1, and v_2, find F_{α,v_1,v_2}, the value to compare to F_c.

```
qf(0.05,6,12,lower.tail=FALSE)
```

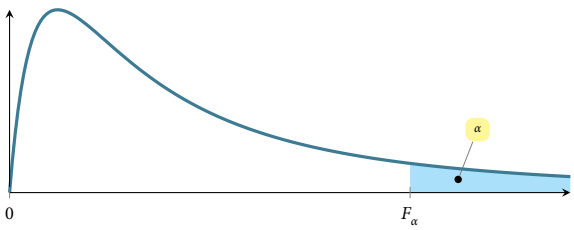
After you press Enter, the F value will be displayed next to a numeral 1 in square brackets.

[1] 2.99612

The above function has the form

qf(α, v_1, v_2, *options*)

The option **lower**.tail=FALSE tells R that you want to know the F corresponding to the following picture.



If, instead, you wanted to know the probability corresponding to the large white area to the left of F_α, you would say **lower**.tail=TRUE instead.

If you want to know the p-value corresponding to a given F, say something like

```
pf(2.99612,6,12,lower.tail=FALSE)
```

giving the result

```
[1] 0.05000002
```

depending on how many digits you have set R to show. You can say

```
options(digits=4)
```

for instance.

This function has the form

pf($F, v_1, v_2,$ *options*)

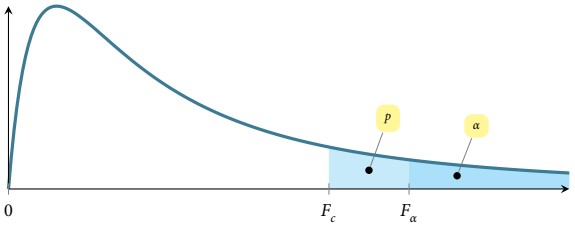
and is the inverse of the function above. You can also do the same thing with *t*, for example

```
> pt(2,60,lower.tail=FALSE)
[1] 0.02501652
> qt(0.025,60,lower.tail=FALSE)
[1] 2.000298
```

(R output always begins with a numeral 1 in square brackets. It will not become obvious why unless you enter a command with a lot of results. Then you will notice that the number in brackets at the beginning of each line is the index number of the first value on that line.)

LI. TEST OVERALL MODEL UTILITY

In multiple regression, the admonition to *test overall model utility* is an admonition to test the null hypothesis that all beta coefficients in the model are equal to zero. We use the same seven step test of hypothesis we learned earlier in the course, just with a new test statistic. The test statistic is what I will call F_c, where the *c* means calculated. Mendenhall simply calls it *F*. In my mind, that leads to occasional confusion about whether it is a value calculated from a sample. We'll compare the test statistic to F_α, the value of the statistic at a given level of significance, identified using a table or calculator.

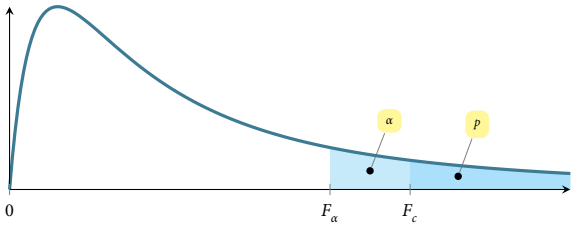


The test leads to two situations. The first, pictured above, is the situation where we fail to reject the null hypothesis and conclude that we have not

seen evidence in the sample that any of the beta coefficients differ from zero.

The above possibility shows the situation where $F_\alpha > F_c$, which is equivalent to saying that $p > \alpha$. The x-axis scale begins at 0, and values of F increase from left to right. At the same time, the shaded regions decrease in size from left to right. Note that the entire shaded region above is p, while only the darker region at the right is α.

The situation pictured below is the reverse. In this situation, we reject the null hypothesis that all beta coefficients are zero and conclude instead that one or more of the beta coefficients, $\beta_1, \ldots, \beta_k$ is not zero. In this case $F_\alpha < F_c$, which is equivalent to saying that $p < \alpha$.



LII. HW 4.1: DEGREES OF FREEDOM

Degrees of Freedom Definition

The number of degrees of freedom in estimating a parameter from a sample is $\nu = n - r$ where n is the sample size and r is the number of parameters estimated. For example, in estimating the mean of a sample, the mean is the parameter subtracted from the sample size. In estimating a simple linear regression with one predictor and one response, there are two parameters to estimate, β_0 and β_1, so we subtract two from the sample size.

One way to think of degrees of freedom is as the number of independent pieces of information involved in the parameter estimation. The mean contains some of the information from the sample—enough so that, if you know the mean and all but one of the sample values, you can name the unknown sample value precisely. Similarly, the slope and y-intercept in the regression equation contain some information from the sample values.

A more detailed definition can be found in an article by Helen Walker, "Degrees of Freedom," *Journal of Educational Psychology*, **31**(4) April 1940. The article is available in the resources folder on blackboard as `Walker1940.pdf`. Walker refers to r as the number of parameters. Note that Mendenhall calls

the number of parameters for the regression model $k + 1$, where k is the number of independent variables and the +1 term refers to the y-intercept.

Problem Statement

How is the number of degrees of freedom available for estimating σ^2, the variance of ε, related to the number of independent variables in a regression model?

Solution

In a previous chapter (p. 142), Mendenhall notes that we only need to estimate β_1 and not β_0 to estimate σ_2. Hence $v = n - 1$ for the straight line model. For each additional parameter, $\beta_2 \ldots \beta_k$, we lose an additional degree of freedom. In other words, we grow more and more constrained as we add parameters to a model. It's tempting to add parameters to a model to improve prediction power, so this is an example of a trade you make in adding parameters and a partial explanation of why regression models don't just grow to arbitrary length.

One easy way to think about this is to suppose that $n = k + 1$. For the single predictor case we studied throughout chapter 3 $k = 1$ so, for the preceding equation to be true, $n = 2$ and we have zero degrees of freedom in estimating the straight line model. This is because, if we know the slope and y-intercept, the two points that constitute n must lie on a straight line. That is, we can draw a straight line between any two points on a plane. If we extend the model to $k = 3$ and the number of dimensions to 3, we can pass a plane through any three points in that 3-dimensional space.

You can visualize the above 2-dimensional case and 3-dimensional case, but the same thing is true for higher dimensions—we just can't draw a picture or a 3d model of it.

Example exam question

How many degrees of freedom are used in the numerator and denominator of the F-statistic for the utility of the auction model where age and number of bidders predict auction price in a sample of 32 clocks?

LIII. ADJUSTED MULTIPLE COEFFICIENT OF DETERMINATION

The discussion of degrees of freedom above may have suggested that you can just add more predictors as you enlarge a sample in order to get a perfect prediction. That's absolutely the case. If the number of predictors, k, is $n - 1$, any model, no matter how uninformed, no matter how useless for future predictions, will perfectly account for any variation in a sample of size n. Hence, it is possible to manipulate R^2 to drive it artificially close to

1 by adding a bunch of extra predictors. The formula

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

is susceptible to manipulation, so it would be nice to have an alternative that explicitly takes into account the relationship between the number of predictors and the sample size. That alternative is R_a^2, the adjusted multiple coefficient of determination, calculated as follows.

$$R_a^2 = 1 - \left(\frac{n-1}{n-(k+1)}\right)(1-R^2)$$

The Mendenhall text advises using it in place of R^2 but R^2 is almost universally reported anyway. There is no universal agreement as to how much it should vary from R^2 before it is used in preference to R^2. In this class, any problems meant to be answered with R_a^2 instead of R^2 will have the property that $R^2 - R_a^2 \geq 0.10$. Be aware that this is purely an artificial cutoff used for grading tests and that, in actual practice, the number will be set by context. In some situations, such as rare diseases or large discretionary purchases, k is often near n. We won't encounter those cases in this class but they represent an area of increasing interest, both due to genome mapping in the first case and the availability of vast amounts of online behavior information about a given consumer on the other hand. A further example can be found in the recent decision by the NBA to instrument all its courts with monitoring equipment that allows vast numbers of measurements to be recorded at every instant of a game.

LIV. HW 4.2: ESTIMATE A MULTIPLE REGRESSION MODEL

(a) Write the equation of the model.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

(b) Test overall model utility at $\alpha = 0.05$.

We can use the same test steps as we first learned on page 43 of Mendenhall.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = 0$

H_a : at least one coefficient is nonzero

Test statistic: $F_c = \frac{(SS_{yy}-SSE)/k}{SSE/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = 4.74$

Level of significance: chosen for you in the problem specification as $\alpha = 0.05$.

Rejection Region Boundary: $F_c > F_\alpha$ where F_α has $k = 4$ numerator degrees of freedom and $n-(k+1) = 198-(4+1) = 193$ denominator degrees of freedom. $F_{.05,4,193} = 2.418$. Note that this value does *not* appear in the F-table on page 761 of Mendenhall. It can be calculated using the R function `qf(0.95,4,193)` giving 2.418445 as the result. The closest values in the table are for 120 denominator degrees of freedom and ∞ denominator degrees of freedom. These two values are 2.45 and 2.37, both of which are much smaller than the computed F-statistic 4.74.

p-value: The rejection region boundary above is equivalent to saying $\alpha >$ p-value, where p-value is $P(F > F_c)$ with F_c the computed value of the test statistic. In this case, $\alpha = 0.05$ and p-value < 0.01. The p-value can also be computed using the R function `pf(4.74,4,193,lower.tail=FALSE)`, giving

0.00113892398018478051867 04

as the result if `options(digits=22)` has been set.

Conclusion: reject H_0 because of the strong evidence that at least one coefficient is nonzero.

(c) Interpret the coefficient of determination, R^2.

The result $R^2 = .13$ suggests that the model explains about thirteen percent of the variation throughout the observed sample of 198 accountants. This is very little explanatory power and suggests that a different model be tried.

(d) Is there sufficient evidence at $\alpha = 0.05$ to say that income is a statistically useful predictor of Mach?

No. The t-statistic is small at 0.52 and the associated p-value is large at $> .10$, suggesting that the estimate of the parameter for income, β_4 contains too much variability for x_4 to be a useful predictor.

Example exam question

Repeat the above for the clock auction example used throughout chapter 4 of Mendenhall.

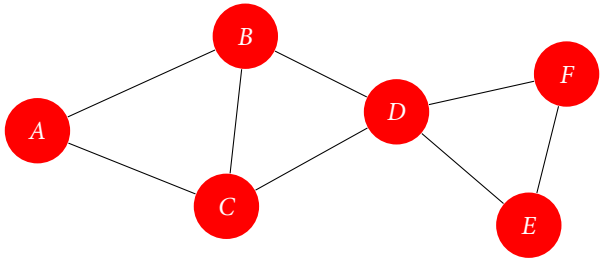
example exam question

Which of the following constitutes a null hypothesis? The lure of gold does not predict evil. (Includes neither a measure of the lure of gold nor of evil.) The bias in a referee's calls does not predict the winner in close contests. (Includes no measure of bias nor of what constitutes a close contest.) The number of time outs remaining does not predict the outcome in football. (This is ambiguous.) The number of cigarettes smoked does not predict a diagnosis of lung cancer. (This is a good example of a testable hypothesis.)

You should pay special attention to the story associated with this problem. It is a problem of social network analysis, a domain of study used increasingly in business. Before solving the textbook problem, it's worthwhile to consider the general natural of the real world problem this represents because you can use this technique to great advantage in business.

The problem concerns the fact that some consumers influence other consumers. These influential consumers are referred to as *lead-users* in the problem statement. The problem studies 362 children to try to determine the influential children among them in terms of game adoption. To see how this is done, consider a simplified example of a social network.

First, imagine you are monitoring communications between children. You put a bowl of candy on the office counter and tell a child about it, then observe who the child communicates with and who gets to the candy bowl first. As you keep monitoring, you record the path of communication among children and the influence of the communication in affecting behavior, such as going to get candy. Now you can draw a graph of the children representing how they communicate.



For the sake of simplicity, consider six children, labeled A–F. The lines between them represent the communication flows discovered by observing some communication channel such as Facebook, verbal communication on the playground, or perhaps some combination of channels. As an example, A only communicates with B and C. It should be clear from the diagram that D enjoys a unique position. That position can be mathematically described by two variables, both of which are used in the problem statement.

First, there are more connections from D to other children than from any other child, 4 in all. This number, 4, is the degree of centrality for D. Another measure used in the problem statement, betweeness centrality, is the number of shortest paths passing through a child. It should be clear that the shortest path to get a message from C to F is to go through D, adding 1 to D's betweeness centrality. The shortest paths for any of $\{A, B, C\}$ to either

of $\{E, F\}$ go through D. Using this method of scoring, the betweeness centrality measure for D is 14. For $\{A, B, C\}$ the betweeness centrality measure is 8. For the other two, the measure is 6. Notice that there is a larger range in the betweeness centrality scores, 6–14, than in the degree of centrality scores, 2–4. Although these measures are related, they measure different things.

Although the above model looks very simple, it can be extended to handle very complicated situations and the results subjected to regression analysis as in the problem. Not only might we add different kinds of communication but we can easily give each line between the children a measure of *length* to account for differences in communication.

The problem statement does not give much detail, almost as if it assumes familiarity with social network analysis. No statistics problem occurs in a vacuum and it will repay you to delve into the domain of each problem rather than to simply do the calculations that follow in questions (a) through (c).

Give two properties of the errors of prediction that result from using the method of least squares to obtain the parameter estimates.

We learned the method of least squares in the previous chapter, where it was claimed that the least squares estimate satisfies two properties. First, the sum of the errors is zero, $SE = 0$. Second, the sum of squares of errors (SSE) is a minimum among all lines that satisfy the first property. These two sums are defined as follows.

$$SE = \sum_{i=1}^{n} (y_1 - \hat{y}_i)$$

$$SSE = \sum_{i=1}^{n} (y_1 - \hat{y}_i)^2$$

Give a practical interpretation of the estimate of beta 4 in the model.

Betweenness centrality measures the number of shortest paths between any two children that pass through child i for $i = 1$ to n. In social network analysis this is a critical measure of the importance of a node (in this case a node is a child) in a network. The value 0.42 indicates that, for every such path that passes through child i, the lead-user rating for that child increases by 0.42.

(c) A test of $H_0 : \beta_4 = 0$ resulted in a two tailed p-value of .002. Make the appropriate conclusion at $\alpha = 0.05$.

Reject H_0 that the slope of betweenness centrality is zero. Instead, conclude that there is strong evidence for the estimate of 0.42 as the true slope of β_4.

LVI. HW 4.6: ESTIMATE ANOTHER MODEL

Mendenhall's description of this problem is a puzzle. Evidently, over 1,000 street vendors were interviewed, but the accompanying data set and the displayed SAS analysis indicate that only information about 15 vendors was used.

Write a first-order model for mean annual earnings.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

where x_1 is age and x_2 is the number of hours worked.

Find the least squares prediction equation.

$$y = -20.35 + 13.35 x_1 + 243.71 x_2 + \varepsilon$$

where x_1 is age and x_2 is hours.

Interpret the estimated beta coefficients.

Each additional year of experience adds 13.35 dollars to earnings and each additional hour worked per day adds 243.71 dollars to earnings.

Test global utility at alpha .01 and interpret the result.

$H_0 : \beta_1 = \beta_2 = 0$

H_a : At least one of age or hours has a nonzero slope.

Test statistic: $F = 8.36$

Level of significance: chosen for you in the problem specification as $\alpha = 0.01$.

Rejection Region Boundary: $F_c > F_\alpha$ where F_α has $k = 2$ numerator degrees of freedom and $n - (k+1) = 15 - (2+1) = 12$ denominator degrees of freedom. $F_{.01,2,12} = 6.93$ is given in the F-table on page 765 of Mendenhall. It can be calculated more precisely using the R function `qf(0.99,2,12)` giving 6.926608 as the result.

p-value: The rejection region boundary above is equivalent to saying $\alpha > p$-value, where p-value is $P(F > F_c)$ with F_c the computed value of the test statistic. In this case, $\alpha = 0.01$ and p-value < 0.0053. It can be calculated more precisely using the R function `pf(8.36,2,12,lower.tail=FALSE)`, giving

0.005320848124948387124211

as the result if `options(digits=22)` has been set.

Conclusion: reject H_0 because of the strong evidence that at least one coefficient is nonzero.

Find and interpret the value of R squared adjusted.

$R_a^2 = 0.5126$ says that the model explains about half the variability in the sample data.

Find and interpret s, the estimated standard deviation of the error term.

The SAS printout refers to s as Root MSE and gives the value 547.73748. This is the square root of Mean Square Error, given earlier in the printout as 300,016.

One interpretation given on page 174 of Mendenhall is that the model can be expected to predict annual earnings for a given input of age and hours to within $\pm 2s = \pm 2(547.74) = \pm 1,095$ dollars, 95 times out of a hundred.

Is age a statistically useful predictor of annual earnings at alpha .01?

No. The computed t-statistic is small at 1.74 and the corresponding p-value is too large at 0.1074 to reject the null hypothesis that the slope for age is zero.

Find a 95 percent confidence interval for beta 2. Interpret the interval in the words of the problem.

A 95 percent confidence interval is given by $\hat{\beta}_2 \pm (t_{\alpha/2}) s_{\hat{\beta}_2}$ with $t_{\alpha/2} = t_{0.025}$ and $n - (k+1) = 15 - (2+1) = 12$ degrees of freedom. $t_{0.025,12} = 2.18$, or use the R function `qt(0.025,12,lower.tail=FALSE)` giving

2.178812829667228889718

as the result if `options(digits=22)` has been set. Using $t_{0.025,12} = 2.18$, a 95 percent ci is $243.71 \pm (2.18)63.51 = 243.71 \pm 138.45 = (105.26, 382.16)$. We are 95 percent confident that β_1 is somewhere between 105.26 and 382.16 dollars and that annual earnings increase between 105.26 and 382.16 for each additional hour per day worked, holding other factors constant.

LVII. HW 4.7: OBTAIN F FROM R SQUARED

Interpret the beta coefficients in the satellite imagery model.

β_1 For each one unit increase in the proportion of a block with low-density residential areas, expect an increase in population density of 2.006.

β_2 For each one unit increase in the proportion of a block with high-density residential areas, expect an increase in population density of 5.006.

Interpret the multiple coefficient of determination, R squared.

The model explains about 68.6 percent of the variation in the observed sample of 125 census blocks.

State the null hypothesis and the alternative hypothesis for a test of model adequacy.

$H_0 : \beta_1 = \beta_2 = 0$

H_a : at least one coefficient has a non-zero slope.

Compute the test statistic for the test of model adequacy.

The expression *test of model adequacy* or *test of model utility* is a code phrase in the Mendenhall book for suggesting an F test of a multiple regression model. Notice that the computed value of F, usually present in the output of computerized regression analysis, is not given here. So we'll have to construct it using what has been given: $R^2 = .686, k = 2, n = 125$. Keep in mind that k is the number of predictor variables. You can also think of it as the number of β coefficients except for the y-intercept. As usual, n is the sample size. The formula for F using these terms is

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{.686/2}{(1-.686)/[125-(2+1)]} = 133.27$$

Draw the appropriate conclusion for the above test at alpha .01.

To do so, we can either compare the given α to the appropriate p-value or compare F_c to F_α. Right now, we only have α and F_c so we need to compute at least one of the other two. We may as well do both. First we'll find F_α which we should really call $F_{\alpha,k,[n-(k+1)]}$. We can look this up in a table or use a calculator after replacing the symbols with the values for this problem: $F_{.01,2,122} = 4.74$. We can also find the p-value in an F distribution calculator. This is a little bit different, since we know $F_{p,2,122} = 133.27$ and we want to find p. It can still be done but it's a very small number. On my iPhone app, it is represented as 0.000000. In a free statistics program called *R* that you can use on a pc or mac, it can be found more precisely using a "probability of f" function, specifically `pf(133.27,df1=2,df2=122,lower.tail=FALSE)`. The result of this function is represented as 2.052909407745667907294e–31 where the e–31 part is scientific notation saying to move the decimal place 31 places to the left. The result would be the number 2 preceded by a decimal point and 30 zeros, a very small number. The conclusion we draw from either of these comparisons, $F_c > F_\alpha$ or $\alpha > p$, is that we reject H_0 because of the strong evidence that at least one slope is non-zero.

LVIII. RELATE EXPLAINED TO UNEXPLAINED VARIANCE

A new model has been suggested to replace a questionable model. It is exactly like the questionable model except that *SSE* has decreased and SS_{yy} has increased. Will it have a higher or lower R^2 than the questionable model? Will it have a higher or lower F than the questionable model? Is it more or less likely for β_1 through k to equal zero?

LIX. HW 4.9: CONFIDENCE INTERVALS FOR BETA 1 TO K

This problem uses some output of a computerized regression analysis of highway crashes. There is an ambiguity in the problem statement in that the sample size is declared to be "over a hundred." The following solution assumes 100 observations each of interstate and non-interstate highways.

Write the least squares equation.

$$\hat{y} = 1.81231 + .10875x_1 + .00017x_2 + \varepsilon$$

Interpret the beta coefficients in the model.

For every mile of interstate highway length, expect the number of crashes every three years to increase by .10875.

For every additional vehicle (average per day) expect the number of crashes every three years to increase by .00017.

Find and interpret a 99 percent confidence interval for beta 1.

This formula is given in the quick summary at the end of the chapter:

$$\hat{\beta}_1 \pm t_{\alpha/2,n-(k+1)} s_{\hat{\beta}_1}$$

We can calculate the t-statistic or read it from a table since we know $\alpha/2 = 0.005$ and $n = 100$ and $k = 2$. $t_{0.005,97} = 2.627$.

The value of $s_{\hat{\beta}_1}$ is given in the coefficients table in the row for x_1 and the column *standard error* as .03166.

$$\hat{\beta}_1 \pm t_{\alpha/2,n-(k+1)} s_{\hat{\beta}_1} = .10875 \pm 2.627(.03166) = (.02453, .19297)$$

The interpretation of this confidence interval assumes that the 100 observations we have so far constitute a single sample. When we say "100 samples" we don't mean 100 observations. We mean replicating the sample 100 times. With 100 observations in each sample, we're talking about a total of 10,000 observations in the following interpretation.

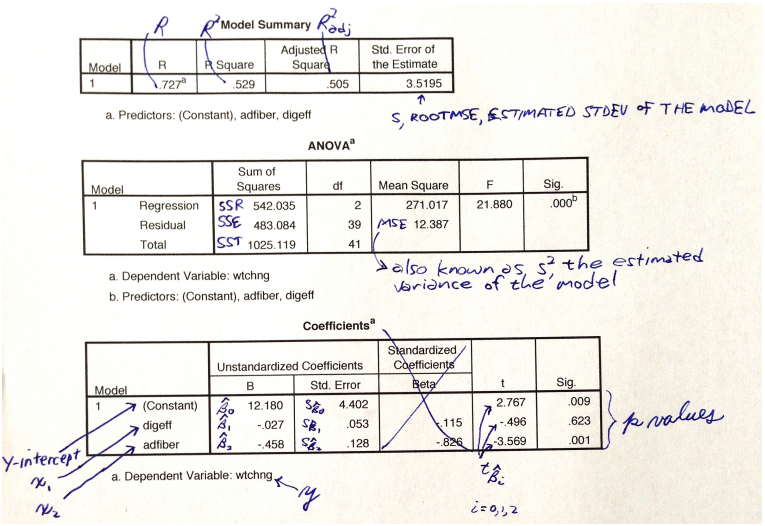
If we sample 100 times, we expect our estimate of the increase in crashes every three years due to a 1 mile increase in road length to be at least .02453 and no more than .19297 in 99 out of the 100 samples.

Find and interpret a 99 percent confidence interval for beta 2.

$$\begin{aligned}\hat{\beta}_2 \pm t_{\alpha/2,n-(k+1)} s_{\hat{\beta}_1} &= .00017 \pm 2.627(.000003) \\ &= (.00009119, .00024881)\end{aligned}$$

We are 99 percent confident that the estimate of the increase in crashes every three years due to an additional average vehicle per day is no less than .00009119 and no more than .00024881. Saying that we're 99 percent confident is the same as saying that we expect a result within this range (from .00009119 to .00024881) to occur in 99 samples out of every 100 samples we collect, assuming each sample contains 100 observations.

LX. HW 4.10: READ COMPUTER OUTPUT



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.727[a]	.529	.505	3.5195

a. Predictors: (Constant), adfiber, digeff

S, ROOTMSE, ESTIMATED STDEV OF THE MODEL

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	SSR 542.035	2	271.017	21.880	.000[b]
	Residual	SSE 483.084	39	MSE 12.387		
	Total	SST 1025.119	41			

a. Dependent Variable: wtchng

b. Predictors: (Constant), adfiber, digeff

also known as s^2 the estimated variance of the model

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	12.180	4.402		2.767	.009
	digeff	-.027	.053	-.115	-.496	.623
	adfiber	-.458	.128	-.826	-3.569	.001

a. Dependent Variable: wtchng

p values

Y-intercept

x_1

x_2

y

$t_{\hat{\beta}_i}$

$i = 0, 1, 2$

This problem is not assigned but we used the dataset in class to identify the three main tables displayed by SPSS.

LXI. HW 4.12: ANOTHER COMPUTER OUTPUT PROBLEM

This is a very timely problem about arsenic in groundwater in Bangladesh, a problem that has worsened in the years since the cited study was pub-

lished. A model based on a sample of 328 wells can be generated using the ASWELLS file.

Write the first order model.

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

Fit the model to the data using least squares.

This can be accomplished with any statistics software to obtain

$$\hat{y} = -86,867 - 2218x_1 + 1542x_2 - .35x_3$$

Interpret the beta estimates.

This is a much more difficult problem. You would probably need to google a little to understand it. Luckily, the day the problem was due in Fall 2013, the subject of arsenic in groundwater wells was the headline of the Daily Star, the largest English-language news site in Bangladesh. The article indicated that there are large arsenic deposits underground in Bangladesh and that the arsenic contaminates the shallow, cheaply dug wells used by the very poor. Hence, the variables of latitude, longitude, and depth are clues to arsenic levels.

You have to know how these variables are measured to interpret the model. First, you must know that Bangladesh is in the Northern Hemisphere and that latitude is measured in degrees starting at zero at the equator, so that positive values of latitude are in the direction of north.

β_1 Walking 69 miles north (1° of latitude), we'll find 2218 fewer micrograms / liter of arsenic in groundwater.

β_2 Walking 1° west (degrees of longitude are harder to translate into miles because the number of miles per degree is a function of distance from the equator), we'll find 1542 fewer micrograms / liter of arsenic in groundwater.

β_3 Drilling a well 1 foot deeper, we'll find 0.35 fewer micrograms / liter of arsenic in groundwater.

Identify and interpret s.

The standard error of the estimate, also known as the standard error of the model, is given in the SPSS summary table as 103.01. It can also be found using other quantities displayed in any computerized regression analysis

$$
\begin{aligned}
s &= \sqrt{SSE/[n-(k+1)}\\
&= \sqrt{MSE}\\
&= \sqrt{(SST-SSR)/[n-(k+1)}
\end{aligned}
$$

s is the variability in units of micrograms / liter of arsenic such that about 95.4 percent of the time we expect the mean concentration of arsenic to be within $2s$ of the estimated mean.

This information raises the very important question: how much is 103 micrograms of arsenic in practical terms? The effect of drinking a liter of water containing some mean level of arsenic plus or minus 103 micrograms of arsenic is not obvious to most people and requires investigation. We'll do that in a later part of this question.

Interpret R squared and R squared adjusted.

$R^2 = .128$ means that about 13 percent of the variability in the data is explained by the model.

$R_a^2 = .12$ means that the value of R_a^2 is very close to the value of R^2, about 94 percent of R^2, suggesting that R^2 has not been artificially manipulated by subtracting rows or adding columns.

Test overall model utility at alpha 0.05.

Testing overall model utility or adequacy means to conduct an F test to determine whether any of the β coefficients is nonzero. If any of them are nonzero, further investigation is warranted. If all are zero, new predictors must be found. Any computerized regression output will provide a value for the F statistic, in this case $F_c = 15.799$. We're given a level of significance for the test, $\alpha = 0.05$. We can use a calculator or table to find the rejection region boundary, $F_{\alpha,v_1,v_2} = F_{\alpha,k,n-(k+1)} = F_{.05,3,324} = 2.6348$. Since $F_c > F_\alpha$, which is the same as saying that $\alpha > p$, we reject the null hypothesis that all β coefficients are zero and instead conclude that we have seen evidence that at least one of the β coefficients is not zero.

Would you use this model?

To make a judgment, you need to learn more about the problem domain. You may already know that nearly every country in the world has a government agency in charge of testing water to determine whether it is safe. The World Health Organization defines the maximum safe level as less than 0.01 milligrams of arsenic / liter. The Bangladesh government defines the

maximum safe level as less than 0.05 milligrams of arsenic / liter. That is equivalent to 50 micrograms, so $s = 103.01$ is in the vicinity of useful information for the estimated 30 million Bangladesh citizens drinking water with more than this level of arsenic.

One reason for using this model is that it is part of an inexpensive kit offered in areas of extreme poverty. It is certainly insufficient if used by itself. An hour of googling convinces me that it would be useful in conjunction with tests of hair samples of children visiting public health clinics in the vicinity of candidate wells. Where resources are scarce, it's worth noting that children are better indicators of current arsenic levels than are adults. Hair may be the least intrusive, cheapest, and easiest sample to handle and forward to labs among the candidates for arsenic detection: urine, stool, nails, and breath.

Previous study of health initiatives in neighboring India convinces me that cultural considerations may outweigh scientific method. For example, I know of a potentially life-saving initiative in rural India that was shut down because it threatened the authority of the mothers-in-law of the targeted population of pregnant wives. The scientists had counted on using village elders as spokespeople for health care methods that contradicted those favored by mothers-in-law. The initiative had limited funding and that funding ran out by the time they realized that mothers-in-law were undermining the program. The hard reality is that statistics is only one of many skills you must learn to really improve the quality of human life.

LXII. HW 4.16: TEST MODEL ADEQUACY

Write the least squares equation estimate of the regression model.

This question does not ask us for the regression model itself, but I want to remind you of the difference between the regression model, which we currently form by writing an equation with the correct number of variables:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \varepsilon$$

and the least squares estimate of that equation, which we form by collecting a sample, analyzing that sample to get $\hat{\beta}_0$ and $\hat{\beta}_i$ for $i = 1, \ldots k$, and then using $\hat{\beta}_0$ to estimate β_0 and $\hat{\beta}_i$ to estimate β_i for $i = 1, \ldots, k$

The least squares estimate of the equation is $y = 0 + 0.110x_1 + 0.0065x_2 + 0.540x_3 - 0.009x_4 - 0.150x_5 - 0.127x_6 + 0$

Give the null hypothesis for a test of overall model adequacy.

$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$

Conduct a test of overall model adequacy (utility).

We're given $\alpha = 0.01$. Conducting an F-test generates an F-statistic that corresponds to a p-value. We can compare α to p or F_c (the F statistic calculated from the sample) to $F_{\alpha,k,n-(k+1)}$.

The following two statements are equivalent:

1. If $F_c > F_\alpha$, then reject H_0.
2. If $p < \alpha$, then reject H_0.

Which of the two we choose may be a matter of fashion or practicality. Keep in mind that you can do arithmetic with p and α because they are in the same units of probability. The difference between an α of 0.10 and a p of 0.08 is exactly the same as the difference between α of 0.05 and a p of 0.03. But F, like t, marks a line under a curve. The difference between an F value of 10 and an F of 12 is a vastly greater difference than the difference between an F value of 20 and an F value of 22. This is because the F curve slopes upward sharply then downward gradually and not as a straight line.

The calculated value of F in this example, $F_c = 32.47$ is enormous as F values go, so practically I already know that I'm going to reject the null hypothesis. I want you to work through the details anyway because it won't always be obvious and because circumstances may require you to do one or the other comparison or perhaps both. $F_{\alpha,k,n-(k+1)} = F_{.01,6,249} = 2.875$ if you are using a calculator. The closest value you will find in the Mendenhall table is 2.87.

As for the p-value corresponding to 32.47, SPSS reports it as 0.000 by default. We can get a better approximation using the free statistics language R, using the function `pf(32.47,6,249,lower.tail=FALSE)` which returns

$$8.769313431206770893936e - 29$$

(if `options(digits=22)` has been set) where the e–29 part is to be read "move the decimal point 29 places to the left."

In any case, $2.87 \ll 32.47$ and $0.01 \gg 0.0$ followed by twenty-six more zeros before getting to a 1, so reject

$$H_0 : \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = 0$$

and instead conclude that at least one member of the set

$$\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6\} \neq 0$$

or, in words, at least one of the true beta values is nonzero.

Interpret R squared and R squared adjusted.

The model explains about 44 percent of the variation in the sample. Since R_a^2 is about 98 percent of R^2, that is, .43/.44 = .98, conclude that R^2 has

not been artificially forced to a higher value and it is okay to use R^2 as a measure of this model. If R_a^2 were substantially smaller than R^2, we would use R_a^2 instead of R^2. The word *substantially* is a matter of judgment and no specific numerical guideline is given.

Give the null hypothesis for a test of contribution of population growth to the model.

$H_0 : \beta_4 = 0$

Test the contribution of population growth to the model.

We're asked to test at $\alpha = 0.01$. The test statistic is $t_{\hat{\beta}_4}$ which is not given. On the other hand, we are given the p-value resulting from the test. Keeping in mind that the t-statistic is a marker of the boundary of the region identified by the p-value, we see that we can identify the t statistic by using the p-value and the appropriate number of degrees of freedom. In this case, we can not use the t-table in the book, because we would need a column for $t_p = t_{.86}$ and we are only given columns for the likely α values. It can be determined by calculator to be -1.082675, although that calculation will not be required on an exam or quiz. We can make the conclusion without knowing the test statistic's exact value because the p-value given in the problem statement is enormous. In general, a parameter estimate is useless if $p > 0.10$ no matter what. Here, the p-value is .86, meaning that, 86 times out of a 100, the result of this test would be a t statistic at least this large if the null hypothesis were true.

The conclusion is that we fail to reject the null hypothesis. We see no evidence for the claim that population growth contributes to this model.

I have not investigated this but my intuition is that the reason for this is that population growth is a rate, unadjusted by size. If my spouse and I are the only people in Gulch County, NV and we have twins, the population growth is tremendous! As a result, population growth is not a very stable number and may be randomly large or small in rural counties, while it is likely to be more stable in urban counties. A second issue is that Nevada has one county that could be considered a very extreme outlier, having an outsized influence on population statistics. The urban Clark county (where Las Vegas is) is extreme, not just among Nevada counties, but among all counties in the USA over the past twenty years. Other urban areas in Nevada are unlikely to resemble Clark county in population growth.

LXIII. HW 4.18: PREDICT AND ESTIMATE

This question revisits the lead users discussed previously in hw problem 4.4, a social network analysis of a group of children playing computer games. Both parts of the question use the estimated model

$$\hat{y} = 3.58 + .01x_1 - .06x_2 - .01x_3 + .42x_4$$

and ask only that you produce a point estimate, not that you add a prediction or confidence interval around that estimate.

Predict the rating for a 10yro female w/5 direct ties and 2 shortest paths.

$$\begin{aligned}\hat{y} &= 3.58 + .01x_1 - .06x_2 - .01x_3 + .42x_4 \\ &= 3.58 + .01(1) - .06(10) - .01(5) + .42(2) \\ &= 3.78\end{aligned}$$

Estimate the mean for 8yro males w/10 direct ties and 4 shortest paths.

$$\begin{aligned}\hat{y} &= 3.58 + .01x_1 - .06x_2 - .01x_3 + .42x_4 \\ &= 3.58 + .01(0) - .06(8) - .01(10) + .42(4) \\ &= 4.68\end{aligned}$$

LXIV. HW 4.20: PREDICTION AND CONFIDENCE INTERVALS

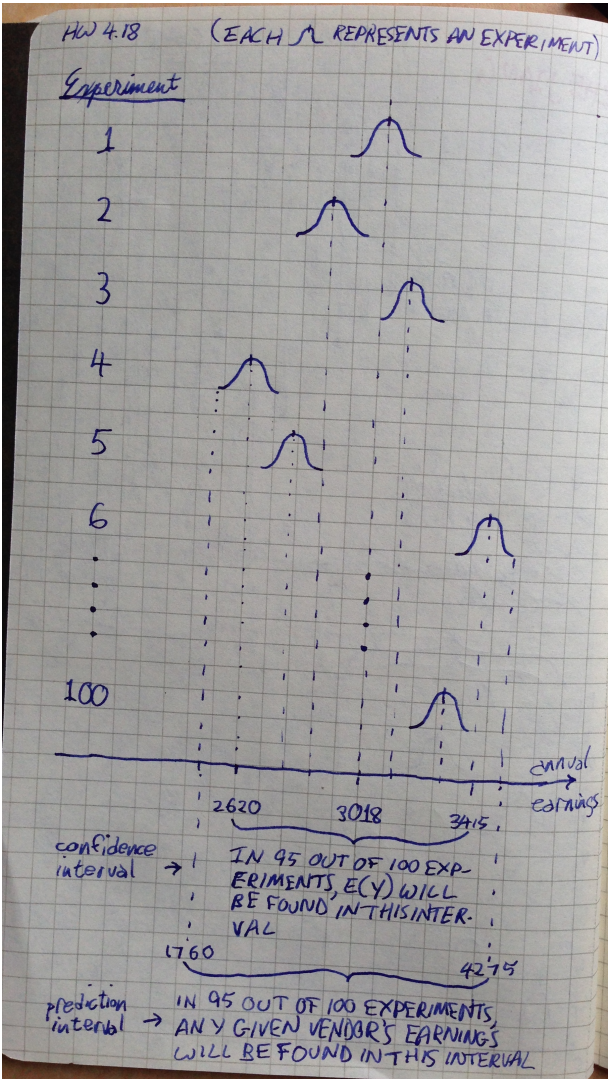
The following illustration shows the difference between the width of prediction and confidence intervals using the numbers from the SAS output for this problem.

As drawn in the picture, the x axis is annual earnings, with a mean of 3018. The intervals around it should be symmetric. That is to say that the prediction interval is 1,258 MXN larger and smaller than 3,018. The confidence interval is 398 MXN larger and smaller than 3,018.

Imagine you conducted an *experiment* over and over. You survey a hundred 45 year old vendors who work 10 hours per day on average. To estimate their annual earnings, you take the expected value of the earnings, that is, the average earnings. Around that, you want to leave enough room that, in 95 of the 100 experiments, the expected value would fall in that range.

Now suppose that you do the same thing except that you want to predict the annual salary of a given 45 year old who works 10 hours. The point estimate

is still the same, the expected value or mean. But the distance around it should be wider because you are measuring both the variability of the mean and the variability of each observation around the mean. So, instead of an interval that captures all the means, you form an interval that captures all the values within, say, a standard deviation (Mendenhall's approach) of each mean. That amount is wider at each side.



The simplest quadratic model is

$$E(y) = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

Even though there is only one x in this model, $k = 2$. This is because we regard x and x^2 as two separate terms.

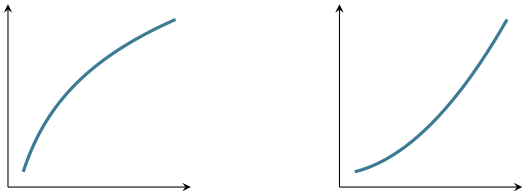
In a quadratic model, β_0 usually has no practical interpretation. This is because the model is typically only valid for a short range of x values.

In a quadratic model, β_1 does not describe the slope. The slope changes with every change in the value of x. The slope is given by

$$\beta_1 + \beta_2 2x$$

which calculus students will recognize as the first derivative of the right hand side of the regression equation.

The sign of β_2 tells which direction the curve takes. A negative β_2 indicates a *diminishing returns* curve (on the left below), while a positive β_2 indicates an *epidemic growth* curve (on the right below).



Diminishing returns occur frequently in business. For example, if a project requires that all workers contributing effort communicate with each other, then adding x workers adds x^2 communication channels, so that each additional worker spends more time trying to communicate instead of contributing effort. A famous book, *The mythical man-month*, by Turing Medal recipient Frederick P. Brooks, Jr., uses this example to caution against expecting software projects (which typically require a lot of communication among contributors) to speed up as more workers are added.

Epidemic growth occurs in business as well, although perhaps not as frequently as investors might like. The *Flappy Bird* phenomenon of January 2014 exemplifies epidemic growth in consumption and the creator of the game removed it from the market because of guilt feelings about its addictive nature.

Unfortunately, expressions like *exponential growth* are often used without regard to the mathematical meaning of the terms employed and mass media may refer to any growth of any kind as epidemic growth. Bear in mind that there is a big difference between quadratic growth, x^2, cubic growth, x^3, and true exponential growth, e^x. Exponential growth is rare outside of sensational news headlines. Epidemic growth is generally not sustainable for long periods of time. Even the bubonic plague of the mid 1300s eventually burned itself out after killing nearly a quarter of the human race, despite a popular belief at the time that all of the human race would be destroyed by it. (William Bowsky, ed, *The black death: a turning point in history?*, Krieger Publishing Co, 1978.)

LXVI. HW 4.34: QUADRATIC MODEL EXAMPLE

Does increasing assertiveness predict increasing leadership ability? Measurements are for $n = 388$ MBAs, with y as leadership score and x as assertiveness score. We don't *merely* want to know if there is a relationship between x and y but also to know whether there is a dynamic relationship, for which we will add a quadratic term.

Test overall model utility.

Test of overall model utility signals an F test. We're given $R^2 = 0.12$ so we can use that along with $k = 2$ and $n = 388$ to calculate the relevant test statistic

$$F = \frac{R^2/k}{(1-R^2)/[n-(k+1)]} = \frac{.12/2}{(1-.12)/[388-(2+1)]} = 26.25$$

We're given the significance level of $\alpha = 0.05$ and the other parameters to identify the rejection region are $k = 2$ and $n = 388$ so we use a calculator or table to find

$$F_{\alpha,\nu_1,\nu_2} = F_{\alpha,k,n-(k+1)} = F_{0.05,2,385} = 3.019$$

Find the above using R Studio by `qf(0.05,2,385,lower.tail=F) [1] 3.019164`

Compare this to the test statistic $F_c = 26.25$ which has a very small p value, found in R by saying `pf(26.25,2,385,lower.tail=F) [1] 2.055485e-11`

Comparison of F_c to F_α or p to α leads us to reject $H_0 : \beta_1 = \beta_2 = 0$ and instead conclude that we've seen evidence that at least one of $\{\beta_1, \beta_2\} \neq 0$

Write the null and alternative hypotheses to test whether leadership increases at a decreasing rate with assertiveness.

$H_0 : \beta_2 = 0, H_a : \beta_2 < 0$

Conduct the above test and give a conclusion in the words of the problem.

$t_c = -3.97, p < 0.01, \alpha = 0.05, \nu_2 = n - (k + 1) = 385$

To say that $p < \alpha$ is the same as saying $|t_c| > |t_\alpha|$ or saying $|-3.97| > |1.64|$

Reject the null hypothesis and conclude instead that we have seen evidence that leadership ability increases at a decreasing rate with assertiveness.

LXVII. HW 4.36: ANOTHER QUADRATIC MODEL

Catalytic converters are devices attached to the exhaust systems of automobiles to reduce air pollution. Air pollution has been a high profile problem for Mexico City for decades and was the subject of a large number of initiatives to reduce pollution in the decade preceding the publication of the study cited in the problem.

The textbook statement appears to contain internal contradictions. If y is a percentage, the y intercept of 325,790 can not be explained. Is it three hundred thousand percent? If x is a year, presumably it is either a year in the Gregorian calendar or a year beginning with 0 as the first year of the study. Either way, it leads to y values that require some arbitrary decimal point somewhere before they can be read as percentages.

I have not yet located the original study on which the problem is based, but a little googling revealed the following facts. (1) At the time the study was published, there were about 2.4m automobiles in Mexico City. (2) In 1998, about 29 percent of cars sold in Mexico City had catalytic converters. They were evidently required in 1992 and 1993 but had been superseded in 1994 by more advanced electronic devices. Some Mexican publications refer to these newer, more effective systems by a different name, but some publications lump them in with the catalytic converters of the 1992–1993 period. (3) The used car market in Mexico received a massive boost beginning in 1987 from the *Hoy No Circula* program. This program prohibits each car in Mexico City from being driven one day per week. Since the day of no driving is determined by the last digit of the automobile's license plate, many auto owners began purchasing second cars to be used on the day prohibited for their main car.

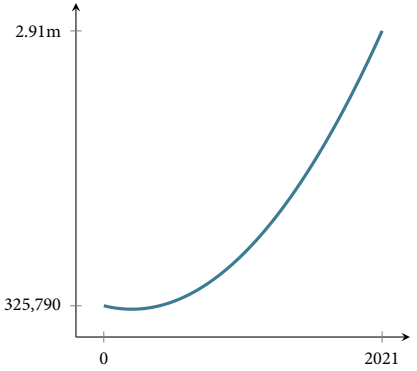
As a result of googling, I can believe that possibly the regression equation describes the raw (not percentage) growth in cars without the 1992–1993 converters. One reason I believe this is that the World Bank reported a study in 2000 on the fraction of cars with each type of emission standard and commented that the worst contributes more than ten times the pollutants of the (then) current standard.

I assume that y is actually the number of vehicles without catalytic converters, not a percentage of vehicles without catalytic converters. Also, I

assume that x refers to a Gregorian year, e.g, the current year would be coded as 2014 in the model, which was published in 2000. The estimated model is

$$\hat{y} = 325,790 - 321.67x + 0.794x^2$$

If these assumptions are true, the following is a graph in the growth of automobiles without catalytic converters from year 0 (the time of the Han Dynasty and the Roman Emperor Augustus) to 2021, the year mentioned in part (d) of the problem. The number of such cars ranges from around 300k to nearly 3m during this time.



Explain why beta 0 has no practical interpretation.

Since β_0 is the value of y when $x = 0$, its literal meaning is that Mexico City had about 325,790 vehicles registered with catalytic converters over 2,000 years ago, even though Mexico City has only existed for 500 years and automobiles for little more than a hundred.

Explain why beta 1 is not the slope.

The slope changes with every value of x. It is not a constant. It is the expression $\beta_1 + \beta_2 2x$, so you can only give the slope for a given value of x.

Determine whether the curvature is upward or downard.

Since the sign of β_2 is positive, the curve is upward.

Comment on the danger of using the model to predict y in 2021.

The model appears to have been constructed from data ranging from 1985 to 2000, so the entire range of the data, 15 years, is smaller than the range from 2000 to 2021. As mentioned above, epidemic growth is never a good long-term bet.

A nested model is one where all the terms in a model that we'll call the *reduced* model also appear in a model with one or more additional terms, called the *complete* model. We'll use the letter g to denote the number of terms in the reduced model and k to denote the number of terms in the complete model. Therefore $k - g$ is the number of additional terms beyond the reduced model. Testing a nested moel involves comparing the unexplained variability of the model with g terms to the unexplained variability of the model with $k - g$ additional terms, i.e., k total terms, to see if the $k - g$ additional terms add any predictive power to the model.

We compare SSE_C, the sum of squares of residuals of the complete model to SSE_R, the sum of squares of residuals of the reduced model, using an F statistic scaled by the number of predictors and the sample size

$$F_C = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/[n-(k+1)]}$$

This formula has the difference between the unexplained variability of the reduced model and complete model in the numerator and the unexplained variability of the complete model in the denominator. Suppose the complete model explains all the variability.

Is it ever possible for $(SSE_R < SSE_C)$ to be true? No, for that to happen, you would have to lose explanatory power by adding more predictor variables. It is possible for additional variables to add nothing but it is not physically possible for additional predictors to subtract explanatory power. To prove this would require math beyond the scope of this course, but you can easily see that it is true by playing around with SPSS or your TI calculator. Simply enter in a few columns of fake data, some of it correlated with the first column you enter, some of it less correlated. Look at what happens to R^2 as you add more and more of the columns as predictors. You will see that R^2 is a *non-decreasing* function of the additional columns. That means that, while it may increase or remain the same, it never decreases.

Another way to find out is to try this very simple thought example. Suppose that, for some y, $SST = 5$. That 5 represents the variability around $\overline{y}$. If we have no predictors of any value, $SSE = 5$ and $SSR = 0$. Any time we add a predictor x_i that has no value, other words any case where $\beta_i = 0$, the relationship between SSE and SSR is unchanged. What happens when we add predictors that do explain some of the variability? Suppose we add five predictors, each of which explains an equal part of the variability and which altogether explain all the variability. Then the following table summarizes the changes:

i	SSE	SSR	SST
1	4	1	5
2	3	2	5
3	2	3	5
4	1	4	5
5	0	5	5

Suppose we regard

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

as the reduced model and

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \varepsilon$$

as the complete model. As long as the reduced model is a proper subset of the complete model, it doesn't really matter how many terms we include in it. *SSE* keeps decreasing as we go through the above table. If, in between the displayed rows, there are rows where we've added a useless predictor, *SSE* does not increase. Instead it simply stays the same. So each such row would be a copy of the row above and we could say that *SSE* is strictly non-increasing in i. The concept *strictly non-increasing* means that *SSE* either gets smaller or stays the same as i increases, but that *SSE* never increases as i increases.

One especially instructive activity, and a great way to spend a weekend, is to add columns that have all 0 values except a single 1. Suppose you have a y column with seven rows. Then, six columns of this type, each with a 1 in a different row, will be sufficient to perfectly predict each value in the y column.

LXIX. HW 4.86: NESTED MODEL

Does prior experience of a chemical fire increase emotional stress of a firefighter responding to a chemical fire?

y is emotional stress

x_1 is experience in years

x_2 is 1 if the firefighter has previous experience of a chemical fire and 0 otherwise. This is called an indicator variable. The model is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_1 x_2 + \beta_5 x_1^2 x_2 + \varepsilon$$

(a) Rate of increase differs on x_2 and the word *rate* signals *quadratic*

(b) means differ on x_2 $\beta_3 = \beta_4 = \beta_5 = 0$

(c) The reduced model is $E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2$

$n = 200$, $SSE_R = 795.23$, $SSE_C = 783.9$, $\alpha = 0.05$

LXX. INTERACTION

Two independent variables interact if the change in the dependent variable for a one-unit change in an independent variable is affected by the value of the other indepdent variable. Graphically, this is the case where the graphs of the slopes of the independent varibles are not parallel.

If two independent variables interact, you can't understand the relationship between one of them and the dependent variable without considering the other.

LXXI. SECOND-ORDER MODELS

The order of a term involving products of quantitative variables is the sum of the exponents of the variables. Thus, x_1^2 can be interpreted as $x_1^1 \times x_1^1$ and therefore a term of order 2. Similarly, $x_1^1 \times x_2^1$ is the product of two variables each to the first power, and is also a second-order term. An interaction model is a second-order model but is not a *complete* second-order model. A *complete* second-order model includes every possible second-order combination. So

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

is a second-order interaction model in two independent variables and

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

is a complete second-order interaction model in two independent variables.

LXXII. HW 5.15: SECOND-ORDER MODEL EXAMPLES

(a) The complete second-order model to relate $E(y)$ to x_1 and x_2 is as above

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

(b) A model with no second order terms is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

(c) A model with only an interaction term as a second-order term is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

(d) The slope of the x_1 line for fixed values of x_2 is the partial derivative of y with respect to x_1 which is $\beta_1 + \beta_3 x_2$.

(e) The slope of the x_2 line for fixed values of x_1 is the partial derivative of y with respect to x_2 which is $\beta_2 + \beta_3 x_1$.

LXXIII. HW 5.16: MORE SECOND-ORDER MODEL EXAMPLES

(a) As in the previous problem, the complete second-order model to relate $E(y)$ to x_1 and x_2 is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_4 x_1^2 + \beta_5 x_2^2$$

(b) The estimated model is found by reading the first column of the coefficients table, specifically the values in the column headed `Coef`:

$$\hat{y} = 606 + 119.68 x_1 - 139.8 x_2 + 2.662 x_1 x_2 - 1.571 x_1^2 + 8.08 x_2^2$$

(c) The question of whether the model is useful is a test of overall model utility, an F-test. The test statistic, $F_c = 5.59$ can be read from the output on page 280, as can the corresponding p value, $p = 0.013$. We're asked to test at $\alpha = 0.05$ so the conclusion is to reject the null hypothesis.

(d) Testing that the second-order terms are not necessary implies that the model without the second-order terms is a reduced model and the model with the second-order terms is a complete model. What is being asked for is a test of a nested model, $H_0 : \beta_4 = \beta_5 = 0$.

(e) The test statistic is

$$F_C = \frac{(SSE_R - SSE_C)/(k - g)}{SSE_C/[n - (k + 1)]}$$

with $SSE_C = 2{,}098{,}183$ and $SSE_R = 3{,}600{,}196$, and $F_c = 2.15$ with the conclusion to reject the null hypothesis.

LXXIV. HW 5:31: QUALITATIVE AND QUANTITATIVE VARIABLES

The problem is to predict a person's intention to leave a situation by observing bullying and organizational support. It includes a quantitative variable, bullying, measured on a scale of 1 to 50, and a qualitative variable, support, measured at three levels: low, medium, and high.

(a) Write a complete model considering a quantitative variable, x_1 for level of bullying and two indicator variables, x_2 and x_3 for level of organizational support. The two indicator variables suffice to track the three levels of organizational support.

$$x_2 = \begin{cases} 1 & \text{if low} \\ 0 & \text{otherwise} \end{cases} \qquad x_3 = \begin{cases} 1 & \text{if medium} \\ 0 & \text{otherwise} \end{cases}$$

The means for each level of organizational support are

$$\begin{aligned} \mu_{\text{high}} &= \beta_0 \\ \mu_{\text{low}} &= \beta_0 + \beta_3 \\ \mu_{\text{medium}} &= \beta_0 + \beta_4 \end{aligned}$$

You can rewrite these expressions in terms of the parameters

$$\begin{aligned} \beta_0 &= \mu_{\text{high}} \\ \beta_3 &= \mu_{\text{low}} - \mu_{\text{high}} \\ \beta_4 &= \mu_{\text{medium}} - \mu_{\text{high}} \end{aligned}$$

and the complete second-order model is

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3 + \beta_7 x_1^2 x_2 + \beta_8 x_1^2 x_3$$

Since x_2 and x_3 are indicator variables and not quantitative variables, don't add their exponents when determining the order of the model. Hence, $x_1^2 x_2$ is a second-order term.

(b) What is the mean value of y if $x_1 = 25$ and $x_2 = 1$?

$$\beta_0 + 25\beta_1 + 25^2\beta_2 + \beta_3 + (0)\beta_4 + 25\beta_5 + (0)25\beta_6 + 25^2\beta_7 + (0)25^2\beta_8$$

(c) Identify the test to determine whether some of the terms (nonlinear relationship between y and x_1) are useful. The preceding sentence tells you to conduct a nested model test. The clue that it is a nested model test is that one model is a proper subset of the other. The particular terms that can establish a nonlinear relationship are the squared terms, hence the null hypothesis for the nested model test is $H_0 : \beta_2 = \beta_7 = \beta_8 = 0$.

(d) A first-order model omits the squared terms:

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_3 x_2 + \beta_4 x_3 + \beta_5 x_1 x_2 + \beta_6 x_1 x_3$$

because the exponents of indicator variables do not contribute to the order of the model. Why? Consider that $1^2 = 1^3 = 1$. Note: the textbook rewrites the model to maintain consecutive numbering of the parameters as

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

This does not matter to me since, in practice, you will use names for the parameters anyway.

(e) Give the slopes for each level of support for the model

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

Note that we cannot use the exact expressions from the beginning of the problem since the indicator variables are now associated with different parameters. Instead, the means for each level of organizational support are

$$\begin{aligned} \mu_{\text{high}} &= \beta_0 \\ \mu_{\text{low}} &= \beta_0 + \beta_2 \\ \mu_{\text{medium}} &= \beta_0 + \beta_3 \end{aligned}$$

and, as a result, the slopes include β_1, a shift parameter for x_1 in the all cases, but only include an additional parameter depending on x_1 in the cases of low and medium but not in the case of high.

$$\text{slope: } \begin{cases} \beta_1 & \text{high} \\ \beta_1 + \beta_4 & \text{low} \\ \beta_1 + \beta_5 & \text{medium} \end{cases}$$

One way to understand combinations of indicator variables is to make a table as in the photo. The column headings are abbreviations for the variables mentioned in the problem statement. The body of the table shows all the possible combinations of ones and zeros. Across the top and down the side are the independent variables and coefficients involved. Notice that we don't need a column for every combination. For example, if both C and G are 0, then the soil type must be sandy and we don't need an explicit column for that.

LXXVI. HW 6.1: STEPWISE REGRESSION PROCESS

This problem reviews the first and second steps in stepwise regression, for which the three steps are listed on Mendenhall page 328. The first step is to fit all one-variable models

$$E(y) = \beta_0 + \beta_1 x_i, \quad i = 1, 2, \ldots, k$$

and conduct a test of the null hypothesis

$$H_0 : \beta_1 = 0$$

against

$$H_a : \beta_1 \neq 0$$

with the test statistic

$$t_{\hat{\beta}_i} = \frac{\hat{\beta}_i - 0}{s_{\hat{\beta}_i}} = \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx_i}}}$$

compared to $t_{\alpha,v}$. This test is only relevant if none of the predictors differs from zero. Otherwise, each t statistic is compared to every other t statistic to find the largest absolute value of a t statistic.

identify the best predictor in the table

The table gives $\hat{\beta}_i$ and $s_{\hat{\beta}_i}$, so the ratio of the two gives the value of t needed to make the comparison. This is an easy case where you can eyeball the table and see that x_2 and x_5 have far larger ratios than the others. Calculating $|-.9/.01| = 90$ is convincing enough not to bother calculating x_5. We conclude that x_2 is the best one-variable predictor of y because it has by far the largest t statistic.

status of best one-variable predictor

The second question is whether x_2 should be included in the model at this stage. Provisionally the answer is yes but the stepwise procedure does not end at this point. If the stepwise procedure results in a one-variable model, it will definitely include x_2 but, because the stepwise procedure will recheck the t statistic in the presence of additional predictors, it remains possible for the procedure to ultimately select a two—or more—variable model that excludes x_2.

step two of stepwise regression

Having selected the best one-variable predictor, the next step in stepwise regression is to fit all two-variable models of the form

$$E(y) = \beta_0 + \beta_1 x_i + \beta_2 x_j, \quad j = 1, 2, \ldots, k, \text{ where } j \neq i$$

and test as above but with one additional constraint: the t statistic for $\hat{\beta}_1$ must be rechecked to see if it is has dropped below the $t_{\alpha,v}$ for the chosen significance level. If so, x_i is dropped from the model and the procedure is to search for another x_k that exhibits the largest t statistic in a model that includes x_j. Then the t statistic for x_j is rechecked as above. Obviously, there is a chance that no two-variable model will be found to surpass the one-variable model in step one. If none is found, the procedure ends here. If a satisfactory two-variable model is found, then Mendenhall describes a step three, where additional variables may be added, one at a time, in the same way.

LXXVII. HW 6.2: STEPWISE REGRESSION EXAMPLE

This problem describes a real situation in which stepwise regression was used to predict the accuracy of estimates of developer effort. Developers are often expensive and estimating the effort they require to produce useful software is a notoriously difficult task, frequently assigned to managers with no understanding of software development. Hence, it could fall to you to read a report like this and make important decisions based on it. The problem includes questions you should be able to answer correctly if you are able to read and make sense of the report.

the number of one-variable models

The number of one-variable models fit to the data is the same as k, the number of predictor variables. In this case, $k = 8$ as you can see by the predictors in the narrative being named $x_1, x_2, \ldots, x_8$.

determination of best one-variable predictor

The determination is made by comparing t statistics of the eight $\hat{\beta}$ values.

the number of two-variable models

First, each remaining variable is added to the one-variable model, so $k-1 = 7$ is the number of models generated initially. However, more are possible if the t statistic for the first variable drops below the chosen significance threshold when the model with the best second variable is tested. In an exam, this might not be the most obvious possibility as written, so I might add some wording to indicate whether or not all t statistics are significant at the chosen level. If they are not, more information is needed to know the number of two-variable models tested. Of course, if absolutely *none* of the t statistics are significant when testing two-variable models, no candidate will be selected at the end of the first seven tests and so the answer would be seven. The only gray area would be if one of the two-variable models emerges as a plausible candidate but, in the process, the t statistic for the first predictor drops below the significance threshold.

interpret the estimates of the resulting coefficients

This is equivalent to interpreting the estimates for any other regression model, where each coefficient $\hat{\beta}_i$ represents the contribution to y from a one unit increase of x_i.

In this case, there is an ambiguity about the nature of y but I have not, at this stage, read the cited publication to try to find out more. The ambiguity is that y is defined as subtracting estimated effort from actual effort and dividing the result by actual effort. Clearly, if estimated effort is smaller than actual effort, a smaller value of y represents a more accurate estimate.

But if the estimated effort is greater than actual effort, a larger value of y (less negative) represents a more accurate estimate.

So does the dataset use the absolute value of y? If so, the interpretation of the coefficients is easy. A developer's prediction improves accuracy (decreases the absolute value of y) by .28 or 28 percent compared to the prediction of a project leader. Counterintuitively, if the previous prediction made by the person offering the estimate was more than 20 percent accurate, the accuracy decreases by .27 or 27 percent. Without having read the article, I would hazard a guess that this result is telling us that the estimates are of no real value and that, when someone gets lucky and makes a good prediction once, they'll, on average, get unlucky and make a worse prediction the next time.

Estimating software effort is, as far as I know, a notoriously unsolved problem for the general case. There are vastly many methods in use but none with clearly useful general results. Of course, my knowledge of software project management is dated and limited to organizations outside Silicon Valley, except Google, where I participated in only one development project. Still, I am quite sure that, ten years ago, when the cited article was published, tarot cards were as good as any other method for the general case.

evaluating the model

Mendenhall disparages stepwise regression and other variable screening methods for four reasons given on pages 337–338 and briefly summarized as follows.

- increased probability of Type I or Type II errors
- no provision in method for higher-order terms or interaction terms
- nonsense terms may survive the screening due to spurious correlation, e.g., sunspots and DJIA
- method may exclude terms that would interact with other predictors but lack independent value

In addition, the method offers no prescription for checking whether the indicator variables interact. At a minimum, we would conduct a nested-model F test with an interaction term before using this model. If we find that do need an interaction term, we can follow the process described in Mendenhall Chapter 5 (see page 296), to enumerate enough interaction terms to generate a separate estimate of $E(y)$ for each of the situations described by the predictors. In this simple case, there would be four predictors so one interaction term plus the two terms in the recommended model would suffice. In longer models, though, the number of interaction terms requires some thought as discussed on Mendenhall page 296.

LXXVIII. HW 6.6: STEPWISE REGRESSION USING SOFTWARE

This exercise requires the use of the CLERICAL dataset, available in the lab on the Students drive as well as on Blackboard in the datasets folder.

conduct a computerized analysis

This can be done in SPSS by choosing `analyze > regression > linear` from the menu and checking the stepwise box in the resulting dialog.

interpret the coefficients

$\hat{\beta}_0$ gives the hours worked if no transactions of any kind are processed.

The other coefficients represent the increase in time required for one unit of each of the noted transaction types.

dangers of using the stepwise model

It is hopelessly naive to imagine that the transaction types in a department store do not interact. Clerical staff may be sharing resources or appealing to limited managerial staff for signatures or encountering other types of bottlenecks that only become noticeable at generally busy times. Customers may wait to visit the department store until they have a number of transactions to process. Customers may change their behavior in the face of an empty or crowded store.

There may also be bottlenecks in the supply chain outside the store itself. For instance, one of the transaction types, check cashing, may require some contact with a bank which may not require linearly more time to process more requests.

No system can be designed to perform equally well under all circumstances, so the department store's setup probably reflects expectations about acceptable service levels and costs and traffic levels. No store can control the fulfillment of these expectations, so out of bounds conditions will arise.

LXXIX. HW 7.2 MULTICOLLINEARITY

This exercise lists general questions about multicollinearity.

describe problems arising from multicollinearity

Rounding errors are more likely throughout a computerized analysis because math. (See Mendenhall Appendix B for math.)

Confusing results may appear. For example, an F test may show that at least one predictor is significant but no predictor's t statistic appears to be

significant. It may be that all predictors are contributing but are redundant, so that they appear insignificant individually.

Signs of coefficients may make no sense. If each predictor is contributing much of the same information, the generated model may use a combination of partial values of the predictors that would make no sense individually.

detecting multicollinearity

Mendenhall lists the symptoms described above as reasons to suspect multicollinearity but also provides a more formal guideline, to calculate the Variance Inflation Factor for each predictor in the model, giving the statistic VIF_i for the ith predictor. The process is conduct a separate regression analysis for each predictor where the ith predictor is interpreted as a dependent variable, modeled by all the other predictors in their role as independent variables. Then VIF_i is calculated by the R_i^2 arising from the ith regression analysis.

$$VIF_i = \frac{1}{1 - R_i^2}, \qquad i = 1, 2, \ldots, k$$

Also, if we assume that neither VIF_i nor R_i^2 takes on a value of 0 or 1, we can take the reciprocal of the above equation to obtain

$$\begin{aligned} VIF_i &= \frac{1}{1 - R_i^2} \\ \frac{1}{VIF_i} &= 1 - R_i^2 \\ R_i^2 &= 1 - \frac{1}{VIF_i} \end{aligned}$$

so we can obtain either statistic given the other. This can be accomplished in SPSS by checking *variance inflation factor* in the options to the dialog box that appears for the menu choice `analyze > regression > linear`.

A common guideline for interpreting VIF_i is given by Mendenhall. Assume that multicollinearity is severe if either of the following equivalent statements is true.

$$R_i^2 > \frac{9}{10} \Leftrightarrow VIF_i > 10$$

The equivalance of these statements can be seen by inserting the boundary values into the equations relating R_i^2 and VIF_i.

remedial measures for multicollinearity

Mendenhall prescribes four remedies for multicollinearity:

- Drop one or more of the correlated predictors from the model.
- Keep all the predictors but make no inferences about the coefficients.
- Code quantitative predictors to reduce the correlation between a predictor and its square.
- Use ridge regression to estimate the model coefficients.

LXXX. HW 7.4: MULTICOLLINEARITY EXAMPLE

This problem describes an observational study of the number of women in managerial positions. It is alleged to be a multicollinearity problem, but strikes me as problematic in other ways.

question the cause and effect relationship

Since $r = .983$, $R^2 = .9662$ for the relationship between y and x_1. Is it safe to say that an increase in x_1 will cause an increase in y?

No, although I disagree with Mendenhall about the reason for this. Since y is the number of females in managerial positions and x_1 is the number of females with a college degree, this correlation only tells us that nearly all female managers have a college degree. That is as far as Mendenhall goes. On the other hand, the model does not separately evaluate sex and education, only the two together. That should be part of the solution. Further, the narrative uses the terms *upper management positions* and *managerial positions* without distinction. External knowledge of other studies tells us that a *glass ceiling* exists for females who may well be able to attain *managerial positions* without being able to attain *upper management positions*. The model should distinguish between these.

identify a potential problem in regression analysis

Mendenhall asks us to identify multicollinearity as a potential problem here. Of course, $r_3 = .722$ so $R_3^2 = .5212$, way below the threshold mentioned in the previous problem. I disagree with Mendenhall's assessment and, in addition to the problems noted in the previous part, would say that the model fails to include any variables about incumbency or attitudinal biases. Also, an observational study suffers from the issue that discrimination is illegal, so anyone practicing it is unlikely to make it easy to prove. Successful studies of the glass ceiling have often been experimental. For example, researchers can ask experimental participants to rate job applicants under controlled conditions, including controlling the applicant qualifications so that they are equal and controlling whether the sex of the applicant is revealed to the participant. Since the participants can not be sued for the bias they show in an experiment, measurement is much easier than in observational studies.

LXXXI. THE REST

That of which we can not speak, we must pass over in silence.

— Ludwig Wittgenstein
in *Tractatus Logico-Philosophicus*

The subject under discussion was never mentioned, of course.

— Flann O'Brien
in *The Hard Life*

This study guide omits much that is important. It is easy to test you on calculations and definitions. It would be very hard to test whether you are a promising analyst. For one thing, every single problem you ever face in statistics will be embedded in a real-life context. Do you know about taxes and insurance and auto repairs and real estate and scams and social network analysis and effective googling and safe drinking water and different markets and the thousands of other things that could have enriched your analysis of the preceding problems? That's the important part.

Appendices

A. PRACTICE FIRST QUIZ

1. Identify a 90 percent confidence interval for the true mean of y if $\overline{y} = 25.5$, $\sum_{i=1}^{n} y^2 = 21755$, and $n = 30$.

A. (21.755,25.5)
B. (25.5,30)
C. (22.85,28.14)
D. (24.92,26.08)
E. (25.08,25.92)

2. If the first few districts have each reported sales of

$$y = [10, 11, 12, 13, 14, 15, 16, 17, 18]$$

, what is the variance of y?

A. 2.739
B. 7.5
C. 9
D. 13
E. 14

3. If y has been reported for $n = 9$ of our dealerships and $\sum_{i=1}^{n} y^2 = 1824$, and $\bar{y} = 14$ what is the standard deviation of y?

A. 2.739
B. 7.5
C. 9
D. 13
E. 14

4. If $y \sim N(50, 8)$ what is $P(20 \leq y \leq 30)$?

A. 0.0001
B. 0.0061
C. 0.08
D. 10 percent
E. 0.5

5. Which would *not* be enough info to find the standard deviation of y?

A. $n, \bar{y}, \sum_{i=1}^{n} y$
B. s^2
C. y
D. $n, \bar{y}, \sum_{i=1}^{n} y^2$
E. $y_1, ..., y_{n-1}, \bar{y}$

6. If $y \sim N(50, 9)$ what is $P(41 \leq y \leq 68)$?

A. 95.4 percent
B. 0.9295
C. 0.8185
D. 0.7175
E. 68.2 percent

B. PRACTICE SECOND QUIZ

7. Suppose you estimate a regression model and find $t_{\hat{\beta}_1}$. Then you receive additional data that leads to a decrease in SS_{xx} and a decrease in s and does not change the slope of the equation. What is the effect on t?

A. t increases.
B. t decreases.
C. t remains approximately the same.
D. t may change but remains normally distributed.
E. There is not enough information to give a valid conclusion.

8. What is closest to the value of SS_{xy} if $\bar{x} = 13, \bar{y} = 15, \sum x_i^2 = 1732, n = 6$, and $\beta_0 = 0.80501$?

A. 13
B. 784
C. 948
D. 1170
E. 1954

9. Given

$$
\begin{aligned}
j_1 &= \{3, 4, 5, 7, 7, 8, 9, 9, 11\} \\
j_2 &= \{11, 10, 10, 8, 6, 6, 4, 3, 3\} \\
j_3 &= \{7, 4, 5, 7, 7, 6, 5, 5, 4\} \\
j_4 &= \{9, 10, 11, 12, 16, 17, 19, 19, 21\} \\
j_5 &= \{3, 22, 5, 51, 7, 64, 9, 79, 11\}
\end{aligned}
$$

With which sample does j_2 have the strongest negative correlation?

A. j_1
B. j_2
C. j_3
D. j_4
E. j_5

10. Given

$$
\begin{aligned}
j_1 &= \{3, 4, 5, 7, 7, 8, 9, 9, 11\} \\
j_2 &= \{11, 10, 10, 8, 6, 6, 4, 3, 3\} \\
j_3 &= \{7, 4, 5, 7, 7, 6, 5, 5, 4\} \\
j_4 &= \{9, 10, 11, 12, 16, 17, 19, 19, 21\} \\
j_5 &= \{3, 22, 5, 51, 7, 64, 9, 79, 11\}
\end{aligned}
$$

Which two samples have the weakest correlation?

A. j_1, j_4
B. j_1, j_5
C. j_2, j_4
D. j_3, j_5
E. j_4, j_5

11. Given $p = \{4, 4, 6, 7, 8, 9\}$ and $q = \{1, 2, 4, 8, 8, 9\}$, give the parameters of the regression model to estimate q given p.

A. $\hat{\beta}_0 = -4.8594, \hat{\beta}_1 = 8.7128$
B. $\hat{\beta}_0 = -4.8594, \hat{\beta}_1 = 1.6094$
C. $\hat{\beta}_0 = -0.8594, \hat{\beta}_1 = 8.7128$
D. $\hat{\beta}_0 = -0.3259, \hat{\beta}_1 = -5.8682$
E. $\hat{\beta}_0 = 0.3259, \hat{\beta}_1 = 1.6094$

12. The motor pool tested a new fuel additive, Bla, promising better gas mileage, on five cars from the company fleet, consisting entirely of the same model and known to have an average gas mileage of 28.8. The five cars tested had gas mileages of 32, 31, 23.5, 26.5, and 32.5. Identify the value of the test statistic used to determine whether the motor pool should purchase Bla for the entire fleet. Assume that this purchase is a business decision with money riding on the outcome.

A. .17
B. 1.645
C. 1.96
D. 3.93
E. 28.8

13. The motor pool tested a new fuel additive, Bla, promising better gas mileage, on five cars from the company fleet, consisting entirely of the same model and known to have an average gas mileage of 28.8. The five cars tested had gas mileages of 32, 31, 23.5, 26.5, and 32.5. Identify the *number that is to be compared to* the test statistic to determine whether the motor pool should purchase Bla for the entire fleet. Assume that this purchase is a business decision with money riding on the outcome.

A. 2.776
B. 2.571
C. 2.132
D. 2.015
E. 1.96

C. PRACTICE THIRD QUIZ

14. The following regression output resulted from trying to predict weight (wt) from height (ht) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-443.767	17.111	-5.908	3.73e-08
ht	9.029	1.016	8.889	1.13e-14

What is the predicted weight for a player with ht = 84?

A. 443.767
B. 434.736
C. 314.669
D. 215.350
E. 206.321

15. Suppose you estimate a regression model as $\hat{y} = 0.3 - 12x + \varepsilon$. Which could be a true statement about the estimate if the model is found to be satisfactory?

A. the errors are negatively correlated
B. the errors are positively correlated
C. the mean of the errors negative
D. the mean of the errors is positive
E. none of the above

16. If $s = 4$ for a linear regression model estimated on two dozen (x, y) pairs, what is the sum of squared residuals for the model?

A. 88
B. 90
C. 160
D. 352
E. 384

17. The following regression output resulted from trying to predict weight (wt) from height (ht) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-443.767	17.111	-5.908	3.73e-08
ht	9.029	1.016	8.889	1.13e-14

Which value below is closest to the appropriate value to compare to the p value in a test of the hypothesis that $\beta_1 = 0$ at $\alpha = 0.05$?

A. 0.05
B. 1.289
C. 1.296
D. 3.1416
E. 8.889

18. The following regression output resulted from trying to predict weight (wt) from height (ht) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-443.767	17.111	-5.908	3.73e-08
ht	9.029	1.016	8.889	1.13e-14

What is the predicted weight for a player with ht = 72?

A. 443.767
B. 434.736
C. 371.676
D. 215.350
E. 206.321

19. The following regression output resulted from trying to predict weight (wt) from height (ht) of 62 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-207.617	56.340	-3.685	0.000493
ht	5.605	0.772	7.258	9.07e-10

What is the best estimate of a 95 percent confidence interval for the weight of a player with ht=74? Assume MSE=296.6 and average height of players is 72.9.

A. (190.292, 295.662)
B. (180.393, 207.617)
C. (155.358, 252.890)
D. (−207.617, 207.617)
E. (202.46, 211.84)

D. PRACTICE MIDTERM EXAM

20. Which of the following is most likely to be tested at a significance level of $\alpha = 0.05$?

A. A test of two alternative web pages, *A* and *B*, will show that viewing page *A* leads consumers to a better understanding of our products.
B. Moving into a new market (which forces us to drop our current market) will lead to a ten percent increase in earnings per share.
C. A new beverage generates more smiles in a sample of loyal customers than our current best-selling beverage.
D. A new vaccine will decrease infant mortality rates if it replaces current vaccines.
E. A new cancer cure is more effective than the standard cure but can only be administered to patients who never take the standard cure.

21. Which of the following makes a better alternative hypothesis than null hypothesis?

A. A brand new cleaner gets clothes as clean as the leading brand.
B. Our 2014 Super Bowl ad for *Crashy* will generate as many pageviews as our 2013 version.
C. A brand new, improved ball will perform just as well as a more expensive existing ball.
D. The expense account of an executive under suspcicion of a crime will not deviate significantly from the average expense accounts of all our executives.
E. A brand new test for pregnancy that is cheaper than the existing test is more accurate than the existing test.

22. You are missing one observation from a sample of five engine temperature readings. The four you know are 150, 153, 160, 161. The mean of the entire sample is 156. What is the missing temperature?

A. 150
B. 153
C. 156
D. 157
E. None of the above

23. You are missing one observation from a sample of five engine temperature readings. The four you know are
140, 146, 150, 151. The mean of the entire sample is 146. What is the missing temperature?

A. 140
B. 143
C. 146
D. 147
E. 156

24. The following regression output resulted from trying to predict weight (wt) from height (ht) of 62 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-207.617	56.340	-3.685	0.000493
ht	5.605	0.772	7.258	9.07e-10

What is the best estimate of a 95 percent confidence interval for the contribution to expected weight of a player for an inch increase in height?

A. $(-207.617, 56.340)$
B. $(-15.512, 26.272)$

C. $(-2.303, 13.519)$
D. $(4.061, 7.149)$
E. $(9.07\text{e}{-10}, 0.000493)$

25. Which of the following samples have the weakest correlation?

$$\begin{array}{cc} p & 8, 2, 5, 3, 8 \\ q & 1, 9, 7, 2, 3 \\ r & 7, 6, 5, 0, 1 \end{array}$$

A. p and q
B. q and r
C. p and r
D. They are equally correlated.
E. An answer can't be determined from the given information

26. In a regression analysis trying to predict weight in pounds (wt) from height in inches (ht) of 115 football players, the following statistics were produced: $SSR = 395.67$ and $SSE = 565.86$. What percentage of the variation in the data is explained by the model?

A. 30
B. 41
C. 59
D. 70
E. It can't be determined from the given information.

27. In a regression analysis trying to predict weight in pounds (wt) from height in inches (ht) of 115 football players, the following statistics were produced: $MSE = 5.0076$ and $SSE = 565.86$ What percentage of the variation in the data is explained by the model?

A. 30
B. 41
C. 59
D. 70
E. It can't be determined from the given information.

28. Which is the shortest list of statistics whose values are required to form a least squares estimate of a simple linear regression model?

A. SS_{xx}, SS_{yy}
B. $SS_{xx}, SS_{yy}, SS_{xy}$
C. $SS_{xx}, SS_{yy}, SS_{xy}, \overline{x}, \overline{y}$
D. $SS_{xx}, SS_{yy}, SS_{xy}, \overline{x}, \overline{y}, SSE$
E. None of the above, i.e., all contain something unnecessary or are missing something necessary

29. Nocturnal Aviation is considering the use of the same ad campaign as it has used in its last nine markets. The costs for those markets are

$$[3, 4, 5, 7, 7, 8, 9, 9, 11]$$

and the corresponding benefits are

$$[9, 10, 11, 12, 16, 17, 19, 19, 21]$$

Identify the appropriate statistics to help management decide *whether* to use the model but do not give any other statistics.

A. $\hat{\beta}_0 = -1.34, \hat{\beta}_1 = 5.60, t_{\hat{\beta}_1} = 9.17$
B. $\hat{\beta}_0 = 3.35, \hat{\beta}_1 = 1.65, t_{\hat{\beta}_1} = 9.17$
C. $\hat{\beta}_0 = -1.34, \hat{\beta}_1 = 5.6, t_{\hat{\beta}_1} = 9.17, R^2 = .92$
D. $t_{\hat{\beta}_1} = 9.17, R^2 = 0.923$
E. $\hat{\beta}_0 = 3.35, \hat{\beta}_1 = 1.65$

30. SludgeCo would like to drive down the cost of toxic waste by using waste processing systems that have operating costs of

$$[11, 10, 10, 8, 6, 6, 4, 3, 3]$$

in each of 9 plants where corresponding toxic waste levels of

$$[3, 4, 5, 7, 7, 8, 9, 9, 11]$$

have been achieved. How much would it cost SludgeCo to achieve a level equaling the average of the other 9 factories?

A. 0.799
B. 5.454
C. 6.75
D. 7
E. 12.4

31. Estimate the parameters of the regression model to predict

$$y = [11, 10, 10, 8, 6, 6, 4, 3, 3]$$

using $x = [9, 10, 11, 12, 16, 17, 19, 19, 21]$.

A. 17, −0.7
B. 0.97, 0.98
C. 24.4, −1.4
D. 34.2, −14.6
E. 23.3, −14.6

32. Which of the following constitutes a null hypothesis?

A. Early to bed and early to rise makes a man healthy, wealthy, and wise.
B. The lure of gold does not predict evil.
C. The bias in a referee's calls does not predict the winner in close contests.
D. The number of time outs remaining does not predict the outcome in football.
E. The number of cigarettes smoked does not predict a diagnosis of lung cancer.

33. Which of the following constitutes a null hypothesis?

A. Websurfing before purchasing does not improve customer satisfaction among auto buyers.
B. A new medicine reduces fever.
C. Progress is the root of all evil.
D. There is no relationship between haste and waste.
E. Thinking deeply does not lead to more correct answers.

34. Which of the following constitutes a null hypothesis?

A. No man is an island, as measured by the number of hermits willing to answer a survey about loneliness.
B. You can't fool all of the people all of the time, as measured by interviewing attendees of a magic show and comparing their explanation of tricks to the magician's.
C. Nothing lasts forever, as measured by the inability of humans to devise a perpetual motion machine.
D. Consumers do not expect the new models to differ from last year's models, as measured by an opinion survey of consumers.
E. Never say never again, as measured by the frequency with which history repeats itself.

35. The following regression output resulted from trying to predict weight (wt) from height (ht) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-73.2220	56.2241	-1.302	0.195
ht	3.7752	0.7704	4.901	3.25e-06

What is the *p*-value for height written out in decimal notation?

A. 0.00000325e
B. 0.00000325
C. 0.000000325

D. 0.0000006
E. 3.25e-06

36. Which of the following are true statements?

A. If $R^2 = 1$, then $SSE = 0$
B. If $R^2 = 0$, then $SSE = 1$
C. If $k + 1 = n$, then $R^2 = 1$
D. If $SSE = SS_{yy}$, then $R^2 = 0$
E. All of the above

37. Let $y \sim N(5, 15)$. Which of the following best approximates the probability that y is between 10 and 20?

A. 0.21
B. 0.56
C. 0.63
D. 0.84
E. 0.95

38. If $y \sim N(50, 8)$ what is $P(60 \leq y \leq 80)$?

A. 0.01
B. 0.11
C. 0.39
D. 0.50
E. 0.89

39. Let $y \sim N(50, 15)$. Which of the following best approximates the probability that y is greater than 50?

A. 0.05
B. 0.21
C. 0.50
D. 0.63
E. 1.67

40. Let $y \sim N(50, 15)$. Which of the following best approximates the probability that y is greater than 70?

A. 0.01
B. 0.02
C. 0.09
D. 0.50
E. 0.65

41. Let $y \sim N(50, 15)$. Which of the following best approximates the probability that y is less than 30?

A. 0.09
B. 0.21
C. 0.50
D. 0.63
E. 1.67

42. Let $y \sim N(50, 15)$. Which of the following best approximates the probability that y is between 30 and 40?

A. 0.02
B. 0.04
C. 0.08
D. 0.12
E. 0.16

43. If $n = 9$, $\sum_{i=1}^{n} y^2 = 13707$, and $\bar{y} = 27.9$ what is the standard deviation of y?

A. 9
B. 27
C. 29
D. 224
E. 13707

44. The following regression output resulted from trying to predict weight (wt) from height (ht) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-443.767	17.111	-5.908	3.73e-08
ht	9.029	1.016	8.889	1.13e-14

Which value in the list is closest to the appropriate value of the test statistic in a test of the hypothesis that $\beta_1 = 0$ at $\alpha = 0.05$?

A. 0.05
B. 1.289
C. 1.296
D. 1.658
E. 8.889

45. Suppose you estimate a regression model and find $t_{\hat{\beta}_1}$. Then you receive additional data that leads to a decrease in SS_{xx} and an increase in s and does not change the slope of the equation. What is the effect on t?

A. t increases.
B. t decreases.
C. t remains approximately the same.
D. t may change but remains normally distributed.

E. There is not enough information to give a valid conclusion.

46. Which of the following is the most plausible test statistic for the test of whether $\bar{y} = 0$ if $y = [3, 4, 5, 7, 7, 8, 9, 9, 11]$?

A. 0
B. 7
C. 8
D. 9
E. 12.1

47. Which is the most appropriate conclusion for a test of whether a sample comes from a population with mean μ, resulting in $p < 0.000007$ given $t_{\alpha/2,\nu} = 1.96$?

A. Reject the null hypothesis and conclude that we have seen evidence that the sample does not come from a population with mean μ.
B. Reject the null hypothesis and conclude that we have not seen evidence that the sample does come from a population with mean μ.
C. Fail to reject the null hypothesis and conclude that we have not seen evidence that the sample comes from a population with mean μ.
D. Fail to reject the null hypothesis and conclude that we have seen evidence that the sample does not come from a population with mean μ.
E. There is not enough information to give a valid conclusion.

48. The Tsunami of 2011 obliterated business records along with everything else. Fujiko Mine is trying to rebuild her father's ink business with sparse records of smudged, illegible paper. She knows he did a regression analysis of advertising and sales and knows that the regression showed that advertising predicted sales. She wants to choose an ad level and project sales based on that level. What is the single most immediately useful number she can generate from knowing $SS_{xx} = 17.5$, $SS_{yy} = 153.3328$, and $SS_{xy} = 50.9999$ to make that projection?

A. $t_{\hat{\beta}_1}$
B. $\hat{\beta}_1$
C. s
D. R^2
E. SSE

49. What is the variance of $y = [3, 22, 5, 51, 7, 64, 9, 79, 11]$?

A. 28

B. 75
C. 29
D. 838
E. 1270

E. PRACTICE FOURTH QUIZ

50. Please conduct a regression of the sara99 dataset, predicting price as a function of livingArea, baths, Bedrooms, fireplace, acres, and age. Next, identify the two least promising predictors, remove them and rerun the analysis. Which is a true statement about the difference between the two models?

A. The proportion of the variability in the data explained by the second model increases appreciably (i.e., more than two percent)
B. The proportion of the variability in the data explained by the second model decreases appreciably (i.e., more than two percent)
C. The proportion of the variability in the data explained by the second model does not change appreciably (i.e., more than two percent)
D. The proportion of the variability in the data explained by the second model is too great to detect any useful information
E. The proportion of the variability in the data explained by the second model is too small to detect any useful information

51. Please conduct a regression of the sara99 dataset, predicting price as a function of livingArea, baths, Bedrooms, fireplace, acres, and age. Which of the following is a true statement about s, the model standard deviation?

A. We can expect a unit increase in one of the predictors to lead to an increase of s in price.
B. The error term in the model is normally distributed and s is the best estimate of its standard deviation.
C. The errors vary by s dollars around the mean.
D. We can expect s dollars to lie within $2s$ of the mean price.
E. We can expect no more than 95 percent of home prices to lie within $2s$ dollars of the mean.

52. Please conduct a regression of the sara99 dataset, predicting price as a function of livingArea, baths, Bedrooms, fireplace, acres, and age. Which of the following best represents the model's estimated standard error?

A. $\sqrt{11{,}284}$
B. $\sqrt{1{,}128{,}400{,}000}$
C. $11{,}284^2$
D. $\sqrt{10{,}382}$

E. $\sqrt{1{,}038{,}200{,}000}$

53. Please conduct a regression of the sara99 dataset, predicting price as a function of livingArea, baths, Bedrooms, fireplace, acres, and age. Suppose you add one more predictor to the model and obtain a new test statistic for the test of overall model utility of 50. What is the corresponding p-value?

A. 2.2×10^{-16}
B. $1.713{,}567 \times 10^{-28}$
C. 0.000,780,8
D. 0.382,95
E. 0.05

54. Please create a scatterplot matrix of all the variables in the sara99 dataset: price, livingArea, baths, bedrooms, fireplace, acres, and age. Which variable appears to be most strongly correlated with price?

A. livingArea
B. baths
C. bedrooms
D. fireplace
E. acres

55. Suppose you have a new sample of 67 house prices. The proportion of variability explained by a model of price with 6 predictors is 95 percent. What is the value of the test statistic for the test of overall model adequacy?

A. 2.76
B. 20.4
C. 21.6
D. 23.25
E. 190

F. PRACTICE FIFTH QUIZ

56. Please conduct a regression of the EXECSAL dataset, predicting SALARY as a function of GENDER, NONSUP, and GEN_SUP. Note that GENDER is 1 if male, 0 if female. NONSUP may be called NUMSUP in the SPSS file. Either way, it is the number of employees supervised. GEN_SUP is the number of employees supervised multiplied by gender. Identify the result of the test of overall model adequacy.

A. Overall, the model is adequate
B. Overall, the model is inadequate
C. The resulting R^2 is adequate
D. The resulting R^2 is inadequate
E. At least one of the coefficients β_1 through β_3 is nonzero

57. Please conduct a regression of the sara99 dataset, predicting price as a function of livingArea, baths, Bedrooms, fireplace, acres, and age. Next, form an interaction term by using the Transpose, Compute Variable menu of SPSS to form a variable called Livebaths by multiplying livingArea by baths. Use the same process to create a variable called BathBed by multiplying baths times Bedrooms. Now conduct a second analysis predicting price as a function of livingArea, baths, Bedrooms, Livebaths, and BathBed. Which is a true statement about the difference between the two models?

A. The proportion of the variability in the data explained by the second model increases appreciably (i.e., more than two percent)
B. The proportion of the variability in the data explained by the second model decreases appreciably (i.e., more than two percent)
C. The proportion of the variability in the data explained by the second model does not change appreciably (i.e., more than two percent)
D. The proportion of the variability in the data explained by the second model is too great to detect any useful information
E. The proportion of the variability in the data explained by the second model is too small to detect any useful information

58. Please conduct a regression of the EXECSAL dataset, predicting SALARY as a function of GENDER, NONSUP, and GEN_SUP. Note that GENDER is 1 if male, 0 if female. NONSUP may be called NUMSUP in the SPSS file. Either way, it is the number of employees supervised. GEN_SUP is the number of employees supervised multiplied by gender.

Next, redo the previous regression analysis, this time dropping NONSUP (or NUMSUP if the name appears that way) from the analysis.

Please redo the analysis yet again, this time *only* using GEN_SUP since the pointy-haired boss is unimpressed with the other two predictors. Which is a true statement about the R^2 and R_a^2 values for the three different analyses you've just conducted?

A. The model with only GEN_SUP explains about 28 percent of the variability in the data
B. The R_a^2 varies by about 1 to 3 percent among the three models
C. The three models don't differ much (more than 3 percent) in the proportion of variability they explain
D. The explanatory power of the three models is similar
E. All of the above

59. Please conduct a regression of the EXECSAL dataset, predicting SALARY as a function of GENDER, NONSUP, and GEN_SUP. Note that GENDER is 1 if male, 0 if female. NONSUP may be called NUMSUP in the SPSS file.

Either way, it is the number of employees supervised. GEN_SUP is the number of employees supervised multiplied by gender. Please interpret the value of β_1, GENDER, in the output.

A. The high p value renders any interpretation meaningless
B. The model predicts that male executives earn $2,676 more than female executives, if all else is held equal
C. The value of 2676 means that adding 1 unit of gender to the input leads to 2676 more units of output
D. The value of 2676 is a kind of salary premium for males
E. All of the above are true

60. Please conduct a regression of the EXECSAL dataset, predicting SALARY as a function of GENDER, NONSUP, and GEN_SUP. Note that GENDER is 1 if male, 0 if female. NONSUP may be called NUMSUP in the SPSS file. Either way, it is the number of employees supervised. GEN_SUP is the number of employees supervised multiplied by gender. Please create a scatterplot of the variables GENDER and NONSUP. Which of the following statements reflects what you see in the scatterplot?

A. There are nearly twice as many male execs as female exacs
B. The male execs are evenly distributed as to the number of employees they supervise
C. The female execs are more sparsely ranged than the males and there are big holes at 100 to 150 and 220 to 300, meaning that, on average, the fewer females supervise more employees than males
D. The scatterplot does not tell us whether GENDER and NONSUP interact
E. All of the above

61. Please conduct a regression of the EXECSAL dataset, predicting SALARY as a function of EXP, EDUC, NONSUP, and GENDER. Note that GENDER is 1 if male, 0 if female. NONSUP may be called NUMSUP in the SPSS file. Either way, it is the number of employees supervised.

Next, conduct a regression of the same dataset but dropping GENDER from the analysis, leaving only EXP, EDUC, and NONSUP to predict SALARY. What is a true statement about the proportion of variability explained by the two models?

A. The model without gender explains much less (at least 10 percent) variability than the model with gender
B. The difference between R^2 and R_a^2 shows that too many variables were used in the first (four predictor) model
C. Neither model explains more than half of the variability in the data

D. The *F* statistic is not closely related to the proportion of variability explained by either model
E. All of the above

G. PRACTICE SIXTH QUIZ

62. Please conduct a regression of the STREETVN dataset available on the Students drive in the lab. Predict EARNINGS as a function of AGE, HOURS, and AGEHOURS. If AGEHOURS does not exist, you can create it in SPSS by using the Transform > Compute Variable menu and multiplying AGE by HOURS, naming the target variable AGEHOURS. Identify the sum of the squares of residuals in the result.

A. 300016
B. 3331000
C. 3600196
D. 5018232
E. 8349232

63. Using the same dataset, form two quadratic terms using the Transform > Compute Variable menu of SPSS. These are AGESQ (the square of AGE) and HOURSSQ (the square of HOURS). Now conduct a regression using the predictors from the previous model as well as the two new terms. Regard this as the Complete model and the previous model as the Reduced model. What is the appropriate value of the test statistic to compare the two models?

A. 3.86
B. 2.15
C. 2.644
D. 0.22
E. 0.05

64. Given the appropriate statistic named in the nested model question, to what other value must it be compared to determine the conclusion of the test, assuming that this is a matter of business with money riding on the outcome?

A. 4.26
B. 2.15
C. 1.76
D. 0.22
E. 0.05

65. How many numerator degrees of freedom are appropriate in the previous question?

A. 1
B. 2
C. 3
D. 4
E. 5

66. Now please rerun the regression after removing the less promising of the two quadratic terms and also removing the interaction term and also removing the linear term that is not the square root of the quadratic term. This leaves two predictors. What is a legitimate statement about the slope of the result?

A. It looks like a diminishing returns curve
B. It looks like an epidemic growth curve
C. It looks linear
D. Its appearance can't be determined
E. All of the above

67. Using the STREETVN data, please test a nested model where the reduced model contains AGE, HOURS, and AGE squared. The complete model adds HOURS squared and an interaction term of AGE and HOURS. What is the value of the test statistic for the test to compare the two models?

A. 10.37
B. 5.594
C. 3.86
D. 0.328
E. 0.22

H. PRACTICE FINAL EXAM

68. KM Knitting

KM Knitting is an online store selling knitted hats in three sizes: small, medium and large, two colors: green and yellow, and two styles: plain and fancy. KM accepts piecework of these hats from knitters. KM has a measure of profitability for each combination of product elements and would like to form a regression model to explain the contribution of each combination to profitability. How many beta coefficients, β_1 through β_k, are needed to form the regression model?

A. 3
B. 4
C. 7
D. 8
E. 11

69. When testing a multiple regression model with 30 observations and 3 predictors for overall model adequacy at $\alpha = 0.05$, to what value would you compare the test statistic?

A. 2.98
B. 2.96
C. 2.92
D. 2.73
E. 2.69

70. When testing a multiple regression model with 31 observations and 2 predictors for overall model adequacy at $\alpha = 0.05$, and assuming that the proportion of variation explained by the model is 0.193, what is the best approximation to the value of the test statistic?

A. 2.69
B. 2.95
C. 3.32
D. 3.34
E. 3.35

71. When testing a multiple regression model with 34 observations and 3 predictors for overall model adequacy at $\alpha = 0.05$, to what value would you compare the test statistic?

A. 2.98
B. 2.96
C. 2.92
D. 2.73
E. 2.69

72.

Following is the SPSS output from two models to predict EARNINGS. First Model:

ANOVA

Model	Sum of Sq	df	Mean Sq	F	Sig.
Regression	6367472	3	?	10.37	0.001547
Residual	2250956	11	?		
Total	8618428	14	?		

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-775.0041	613.9930	-1.262	0.2330
AGE	145.8087	51.9724	2.806	0.0171
HOURS	94.4985	78.2827	1.207	0.2527
AGESQ	-1.6464	0.6412	-2.568	0.0262

For the second model, the only legible number available is the number

6520245, which is in the same location as the number 6367472 above. Additionally, the words HOURSSQ and AGEHOURS are legible in the column called Model below the word AGESQ above. What is the value of the test statistic for the test to compare the two models?

A. 0.22
B. 0.328
C. 3.86
D. 5.594
E. 10.37

73.

Following is the SPSS output from a model to predict EARNINGS.

ANOVA

Model	Sum of Sq	df	Mean Sq	F	Sig.
Regression	6367472	3	?	10.37	0.001547
Residual	2250956	11	?		
Total	8618428	14	?		

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-775.0041	613.9930	-1.262	0.2330
AGE	145.8087	51.9724	2.806	0.0171
HOURS	94.4985	78.2827	1.207	0.2527
AGESQ	-1.6464	0.6412	-2.568	0.0262

How many denominator degrees of freedom does the rejection region marker have in the test for overall model utility?

A. 3
B. 11
C. 14
D. 15
E. 0.05

74.

Following is the SPSS output from a model to predict EARNINGS.

ANOVA

Model	Sum of Sq	df	Mean Sq	F	Sig.
Regression	6367472	3	?	10.37	0.001547
Residual	2250956	11	?		
Total	8618428	14	?		

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-775.0041	613.9930	-1.262	0.2330
AGE	145.8087	51.9724	2.806	0.0171
HOURS	94.4985	78.2827	1.207	0.2527
AGESQ	-1.6464	0.6412	-2.568	0.0262

How many numerator degrees of freedom does the rejection region marker have in the test for overall model utility?

A. 3
B. 11
C. 14
D. 15
E. 0.05

75. Identify R^2 for the model in the following SPSS output.

ANOVA

Model	Sum of Sq	df	Mean Sq	F	Sig.
Regression	6367472	3	?	10.37	0.001547
Residual	2250956	11	?		
Total	8618428	14	?		

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-775.0041	613.9930	-1.262	0.2330
AGE	145.8087	51.9724	2.806	0.0171
HOURS	94.4985	78.2827	1.207	0.2527
AGESQ	-1.6464	0.6412	-2.568	0.0262

A. 0.7388
B. 0.6676
C. 0.3750
D. 0.2996
E. 0.2330

76. What is a legitimate statement about the slope of the regression model in the following SPSS output?

ANOVA

Model	Sum of Sq	df	Mean Sq	F	Sig.
Regression	6367472	3	?	10.37	0.001547
Residual	2250956	11	?		
Total	8618428	14	?		

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-775.0041	613.9930	-1.262	0.2330
AGE	145.8087	51.9724	2.806	0.0171
HOURS	94.4985	78.2827	1.207	0.2527
AGESQ	-1.6464	0.6412	-2.568	0.0262

A. It looks like a diminishing returns curve

B. It looks like an epidemic growth curve
C. It looks linear
D. Its appearance can't be determined
E. All of the above

77. Using the following SPSS output, identify SS_{yy} in a regression of EARNINGS on AGE, HOURS, AGEHOURS, and a new variable, GDP.

ANOVA

Model	Sum of Sq	df	Mean Sq	F	Sig.
Regression	6367472	3	?	10.37	0.001547
Residual	2250956	11	?		
Total	8618428	14	?		

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-775.0041	613.9930	-1.262	0.2330
AGE	145.8087	51.9724	2.806	0.0171
HOURS	94.4985	78.2827	1.207	0.2527
AGESQ	-1.6464	0.6412	-2.568	0.0262

A. 6367472
B. 2250956
C. 8618428
D. 10869384
E. It can not be determined from the given information

78. Which is the most appropriate conclusion for a test of overall model adequacy conducted on the model in the following SPSS output?

ANOVA

Model	Sum of Sq	df	Mean Sq	F	Sig.
Regression	6367472	3	?	10.37	0.001547
Residual	2250956	11	?		
Total	8618428	14	?		

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-775.0041	613.9930	-1.262	0.2330
AGE	145.8087	51.9724	2.806	0.0171
HOURS	94.4985	78.2827	1.207	0.2527
AGESQ	-1.6464	0.6412	-2.568	0.0262

A. Reject the null hypothesis and conclude that we have seen evidence that at least one beta is not 0.
B. Reject the null hypothesis and conclude that we have not seen evidence that at least one beta is 0.
C. Fail to reject the null hypothesis and conclude that we have not seen evidence that at least one beta is 0.
D. Fail to reject the null hypothesis and conclude that we have seen evidence at least one beta is not 0.
E. There is not enough information to give a valid conclusion.

79. Suppose you estimate a regression model as $\hat{y} = 0.3 - 12x_1 + 3x_2 + \varepsilon$. Which could be a true statement about the estimate if the model is found to be satisfactory?

A. the errors are formally distributed
B. the errors are correlated with zero
C. the mean of the errors is zero
D. A, B, and C
E. none of the above

80. Which of the following is not a warning sign that multicollinearity may be an issue?

A. F statistic much larger than F_α but all individual t statistics are smaller than their respective t_α values.
B. Signs of the β coefficients are unexpected, e.g., sales and marketing appear correlated in experience but β coefficient indicates an inverse relationship.
C. R_a^2 is significantly smaller than R^2 (at least 10 percent smaller)
D. *VIF* exceeds 10 for any predictor.
E. Scatterplots of pairs of predictors are tightly packed diagonal ellipses.

81. Suppose you conduct a regression of x_1 on x_2 and x_3, all three being predictors in your model for explaining y. If $R_1^2 = 0.81$ which is a true statement among the following?

A. $VIF = 0.81$
B. x_1 explains 81 percent of the model variability.
C. x_1 is in the danger zone.
D. $VIF = 5.26$
E. none of the above

82. KM Knitting KM Knitting is an online store selling knitted hats in two sizes: small and large, two colors: green and yellow, and two styles: plain and fancy. KM accepts piecework of these hats from knitters. KM has a measure of profitability for each combination of product elements and would like to form a regression model to explain the contribution of each combination to profitability. If KM discovers that color does not matter, how many terms will have an insignificant t statistic?

A. 0
B. 2
C. 4
D. 6
E. 8

83. Football to analyze the relationship between height in inches and weight in pounds of 115 football players, an indicator variable was introduced. This indicator is set to one if the player's position is coded as OL, DL, or DT and zero for all other positions. This is because it was noticed that all the players who weighed over 259 pounds played one of these three positions. The following regression output resulted from trying to predict weight (wt) from height (ht) and the indicator variable (line) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-73.2220	56.2241	-1.302	0.195
ht	3.7752	0.7704	4.901	3.25e-06
line	63.4518	4.9706	12.765	2e-16

What is the p-value for height written out in decimal notation?

A. 0.00000325e
B. 0.00000325
C. 0.000000325
D. 0.0000006
E. 3.25e-06

84. Football To analyze the relationship between height in inches and weight in pounds of 115 football players, an indicator variable was introduced. This indicator is set to one if the player's position is coded as OL, DL, or DT and zero for all other positions. This is because it was noticed that all the players who weighed over 259 pounds played one of these three positions. The following regression output resulted from trying to predict weight (wt) from height (ht) and the indicator variable (line) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-73.2220	56.2241	-1.302	0.195
ht	3.7752	0.7704	4.901	3.25e-06
line	63.4518	4.9706	12.765	2e-16

What is the right hand side of the least squares estimate of the model?

A. $-73.222 + 56.2241x_1$
B. $3.7752x_1 + 63.4518x_2 - 73.222$
C. $-73.222 + 3.7752x_1$
D. $\beta_0 + \beta_1 x_1 + \beta_2 x_2$
E. $63.4518 + 4.9706x_1 + 12.765x_2$

85. Football The following regression output resulted from trying to predict weight (wt) from height (ht) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-443.767	17.111	-5.908	3.73e-08
ht	9.029	1.016	8.889	1.13e-14

Which value in the list is closest to the appropriate value of the test statistic in a test of the hypothesis $\beta_1 = 0$ at $\alpha = 0.05$?

A. 0.05
B. 1.289
C. 1.296
D. 1.658
E. 8.889

86. Football To analyze the relationship between height in inches and weight in pounds of 115 football players, an indicator variable was introduced. This indicator is set to one if the player's position is coded as OL, DL, or DT and zero for all other positions. This is because it was noticed that all the players who weighed over 259 pounds played one of these three positions. The regression output from trying to predict weight (wt) from height (ht) of 115 football players included SSE=112111. The regression output from trying to predict weight (wt) from height (ht) and the indicator variable (line) of 115 football players included SSE=45667. What happened to SSR in the change from the first model to the second?

A. It remained the same.
B. It increased.
C. It decreased.
D. It changed in both directions.
E. It can't be determined from the given information.

87. Football To analyze the relationship between height in inches and weight in pounds of 115 football players, an indicator variable was introduced. This indicator is set to one if the player's position is coded as OL, DL, or DT and zero for all other positions. This is because it was noticed that all the players who weighed over 259 pounds played one of these three positions. The regression output from trying to predict weight (wt) from height (ht) of 115 football players included $SSE = 112111$. The regression output from trying to predict weight (wt) from height (ht) and the indicator variable (line) of 115 football players included $SSE = 45667$. What happened to R^2 in the change from the first model to the second? Assume that SS_{yy} for the first model was 190501.

A. It increased from .41 to .59
B. It increased from .41 to .76
C. It increased from .59 to .76
D. It increased from .59 to .65

E. It increased from .35 to .52

88. Football To analyze the relationship between height in inches and weight in pounds of 30 football players, an indicator variable was introduced. This indicator is set to one if the player's position is coded as OL, DL, or DT and zero for all other positions. This is because it was noticed that all the players who weighed over 259 pounds played one of these three positions. The following regression output resulted from trying to predict weight (wt) from height (ht) and the indicator variable (line) of 30 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-73.2220	56.2241	-1.302	0.195
ht	3.7752	0.7704	4.901	3.25e-06
line	63.4518	4.9706	12.765	2e-16

What is the best estimate of a 95 percent confidence interval for the contribution to expected weight of a player for an inch increase in height?

A. $(0.77, 4.90)$
B. $(2.19, 5.36)$
C. $(2.23, 5.32)$
D. $(2.46, 5.09)$
E. $(4.061, 7.149)$

89. Football The following regression output resulted from trying to predict weight (wt) from height (ht) of 115 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-443.767	17.111	-5.908	3.73e-08
ht	9.029	1.016	8.889	1.13e-14

Which value in the list is closest to the appropriate value to compare to the test statistic in a test of the hypothesis $\beta_1 = 0$ at $\alpha = 0.05$?

A. 0.05
B. 1.289
C. 1.296
D. 1.98
E. 8.889

90. Football In a regression analysis trying to predict weight in pounds (wt) from height in inches (ht) of some football players, only the following statistics are available: $MSE = 5.0076$ and $SSE = 565.86$ What percentage of the variation in the data is explained by the model?

A. 70
B. 59
C. 41

D. 30
E. It can't be determined from the given information.

91. Model 150 on 2 Given a model with 150 observations and 2 predictors, $R^2 = .6928$. Identify the best approximation of a test statistic for a test of overall model utility.

A. 165.8
B. 152
C. 147
D. 128.2
E. 113.7

92. Model 150 on 2 Given a model with 150 observations and 2 predictors, $R^2 = .6928$. Identify the appropriate conclusion if the test for overall model utility is conducted at $\alpha = 0.05$.

A. Reject $H_0 : \beta_1 = \beta_2 = 0$
B. Fail to reject H_0 : Conclude at least one of $\{\beta_1, \beta_2\} \neq 0$
C. Reject $H_a : \beta_1 = \beta_2 = 0$
D. Fail to reject H_a : Conclude at least one of $\{\beta_1, \beta_2\} \neq 0$
E. Insufficient evidence to draw a conclusion

93. A regression model with 500 observations and 2 predictors and $SS_{yy} = 47.2$ was found to have an $R^2 = 0.81$ but the researcher, in the interests of parsimony, decided to recompute, removing one predictor. The result had $SSE = 9.444$. Identify the value that best represents the percentage of variation explained by the first model, before recomputing.

A. 0.94
B. 0.81
C. 0.62
D. 0.35
E. 0.01

94. Small Sample A regression model with 5 observations and 2 predictors and $SS_{yy} = 47.2$ was found to have an $R^2 = 0.81$ but the researcher, in the interests of parsimony, decided to recompute, removing one predictor. The result had $SSE = 9.444$. Identify the value to which the researcher can compare the test statistic to determine the effect of removing one predictor. Assume this is a business situation with money riding on the outcome.

A. 18.5
B. 10.1
C. 7.70

D. 0.10
E. It can't be determined from the given information.

95. Small Sample Assume $\alpha = 0.05$ for this problem. A regression model with 5 observations and 3 predictors and $SS_{yy} = 47.2$ was found to have an $R^2 = 0.81$ but the researcher, in the interests of parsimony, decided to recompute, removing two predictors. The result had $SSE = 9.444$. Identify the value to which the researcher can compare the test statistic to determine the effect of removing two predictors.

A. 9.55
B. 9.28
C. 19.2
D. 199
E. It can't be determined from the given information.

96. Small Sample Assume $\alpha = 0.05$ for this problem. A regression model with 5 observations and 2 predictors and $SS_{yy} = 47.2$ was found to have an $R^2 = 0.81$ but the researcher, in the interests of parsimony, decided to recompute, removing one predictor. The result had $SSE = 9.444$. Identify the value that best represents the percentage of variation explained by the first model, before recomputing.

A. 0.94
B. 0.81
C. 0.62
D. 0.35
E. 0.01

97. KM Knitting

KM Knitting is an online store selling knitted hats in three sizes: small, medium and large, two colors: green and yellow, and two styles: plain and fancy. KM accepts piecework of these hats from knitters. KM has a measure of profitability for each combination of product elements and would like to form a regression model to explain the contribution of each combination to profitability. How many indicator variables are needed in the regression model?

A. 3
B. 4
C. 7
D. 8
E. 11

98. Saratoga

Coefficients

Model	B	StdError	t	Sig.
(Constant)	7400.70	5227.94	1.416	0.157
LivingArea	91.97	2.44	37.687	< 2e−16 ***
Age	−210.96	47.91	−4.404	1.17e−05 ***

Which of the following best represents the least squares estimate of a model for predicting price from Living Area and Age?

A. $E(y) = 7400.70 + 210.96x_1 + 53350x_2 + \varepsilon$
B. $E(y) = 7400.70 + x_1 + x_2 + \varepsilon$
C. $E(y) = 7400.70 + 5227.94x_1 + 1.416x_2 + \varepsilon$
D. $E(y) = 7400.70 + 91.97x_1 - 210.96x_2 + \varepsilon$
E. $E(y) = 7400.70 - 91.97x_1 + 210.96x_2 + \varepsilon$

99. Saratoga For the model of Price predicted by LivingArea and Age, suppose the only legible numbers on the SPSS output are SSR = 4,561,201,000,000 and n = 1063 and s = 53,350. Which value best approximates the test statistic for a test of overall model utility?

A. .6018
B. 201
C. 801
D. 1060
E. 53350

100. Football The following regression output resulted from trying to predict weight (wt) from height (ht) of 62 football players.

Coefficients

Model	B	StdError	t	Sig.
(Constant)	-207.617	56.340	-3.685	0.000493
ht	5.605	0.772	7.258	9.07e-10

What is the best estimate of a 90 percent confidence interval for the weight of a player with ht=80? Assume MSE=296.6 and average height of players is 72.9.

A. (215.350, 443.767)
B. (222.772, 262.340)
C. (230.14, 249.86)
D. (242.6, 246.84)
E. (245.25, 245.981)

I. SOLUTIONS TO ALL QUIZZES AND EXAMS

1. C is the correct answer. To solve this problem you will need s, which

you can find using

$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} \\ &= \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}{n-1}} \\ &= \sqrt{\frac{21755 - 30(25.5)^2}{30-1}} \\ &= 8.803 \end{aligned}$$

Since $n \geq 30$ you may use a large-sample confidence interval.

$$\begin{aligned} \bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}} \\ &= 25.5 \pm z_{.05}\frac{8.803}{\sqrt{30}} \\ &= 25.5 \pm 1.645\frac{8.803}{\sqrt{30}} \\ &(22.85, 28.14) \end{aligned}$$

I checked my calculations with the math program Maxima.

```
25.5-1.645*(8.803/sqrt(30)),numer;
                        22.85615508224464
25.5+1.645*(8.803/sqrt(30)),numer;
                        28.14384491775536
```

If n were less than 30, we would be constrained to use a small-sample confidence interval. The main difference is that we would use a t statistic instead of a z score. The difference would be small but distinctive. You can see in Table 1.8 on page 35 of Mendenhall that $z_{.05} = 1.645$. The table also appears in the study guide under the section on hw 1.53. The corresponding t statistic, $t_{.05,29} = 1.699$ which very close but different enough that, on multiple choice quizzes and exams, you should use z if $n \geq 30$. In this case, using the t statistic would have given the solution (22.77,28.23), which would have been close enough to make the correct selection but that may not always be the case.

The problem was developed using the stat program R as follows.

```
> y<-c(11:40)
> length(y)
[1] 30
> mean(y)
[1] 25.5
> sum(y^2)
[1] 21755
> sd(y)
```

```
[1] 8.803408
> qnorm(0.05,lower.tail=FALSE)
[1] 1.644854
> qt(0.05,29,lower.tail=FALSE)
[1] 1.699127
```

2. B is the correct answer. The 'first few districts' signify a sample of all districts, hinting that you should find sample variance not population variance.

$$\begin{aligned} s^2 &= \frac{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}{n-1} \\ &= \frac{1824 - 9(14)^2}{9-1} \\ &= 7.5 \end{aligned}$$

This problem was developed using R as follows.

```
> y<-c(10:18)
> var(y)
[1] 7.5
> length(y)
[1] 9
> mean(y)
[1] 14
> sum(y^2)
[1] 1824
> sd(y)
[1] 2.738613
```

3. A is the correct answer. The phrase '$n = 9$ of our dealerships' indicates sample not population.

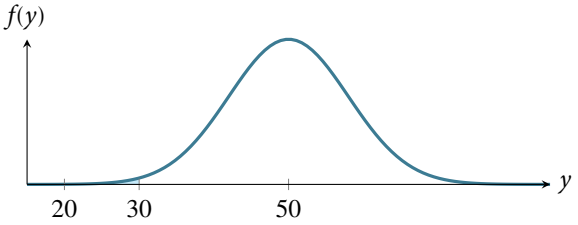
$$\begin{aligned} s &= \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}{n-1}} \\ &= \sqrt{\frac{1824 - 9(14)^2}{9-1}} \\ &= \sqrt{7.5} = 2.739 \end{aligned}$$

This problem was developed using R as follows.

```
> y<-c(10:18)
> var(y)
[1] 7.5
> length(y)
[1] 9
> mean(y)
[1] 14
> sum(y^2)
[1] 1824
> sd(y)
[1] 2.738613
```

4. B is the correct answer. The expression $y \sim N(50, 8)$ means that y has the normal distribution with mean 50 and standard deviation 8. The question asks for the probability that y falls between 20 and 30, well below the mean.

1. Draw the picture of the answer. This should show you that the part you're looking for is a small slice to the left of the mean. Later, you can use this to verify that your answer makes some kind of sense.

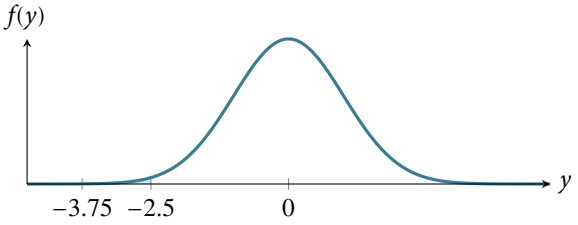


2. Standardize. This means to use the equation

$$(y - \mu)/\sigma$$

on both ends of the slice of probability under consideration, 20 and 30. This gives $(30 - 50)/8 = -2.5$ and $(20 - 50)/8 = -3.75$.

3. Draw the standardized picture. This is the same as the original picture except with standardized numbers.



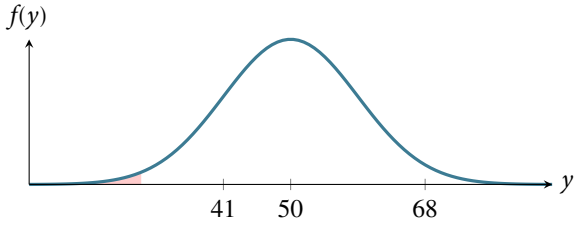
4. Pick a calculator or table from among the three possibilities: (a) reports the probability from 0 to $|z|$, (b) reports the probability from $-\infty$ to z, or (c) from z to $+\infty$.

5. Make the calculations. Suppose we chose the Mendenhall table, which provides method (a). We need two numbers from the table, the probability associated with $z = 2.5$ and $z = 3.75$. The math notation for this would be $P(0 < z < 2.5) = 0.4938$ and $P(0 < z < 3.75) = 0.4999$. Subtracting the smaller from the larger gives us the area, or probability, between them, $0.4999 - 0.4938 = 0.0061$.

5. A is the correct answer. The key here is that there is no way to get $\sum_{i=1}^{n} y^2$ from $\sum_{i=1}^{n} y$. All the other choices contain enough information.

6. C is the correct answer. The expression $y \sim N(50, 9)$ means that y has the normal distribution with mean 50 and standard deviation 9. The question asks for the probability that y falls between 41 and 68, spanning the mean.

1. Draw the picture of the answer. This should show you that the part you're looking for is a large slice, much more than 50 percent. Later, you can use this to verify that your answer makes some kind of sense.

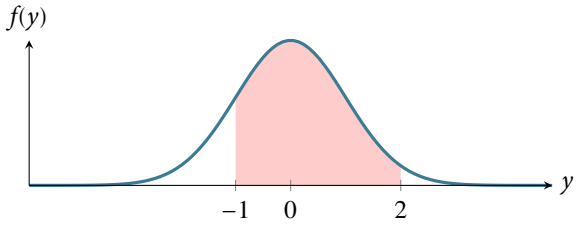


2. Standardize. This means to use the equation

$$(y - \mu)/\sigma$$

on both ends of the slice of probability under consideration, 41 and 68. This gives $(41 - 50)/9 = -1$ and $(68 - 50)/9 = 2$.

3. Draw the standardized picture. This is the same as the original picture except with standardized numbers.



4. Pick a calculator or table from among the three possibilities: (a) reports the probability from 0 to $|z|$, (b) reports the probability from $-\infty$ to z, or (c) from z to $+\infty$.

5. Make the calculations. Suppose we chose the Mendenhall table, which provides method (a). We need two numbers from the table, the probability associated with $z = 1$ and $z = 2$. The math notation for this would be $P(0 < z < 1) = 0.3413$ and $P(0 < z < 2) = 0.4772$. Adding the two together gives us the area, or probability, between them, $0.3413 + 0.4772 = 0.8185$.

7. E is the correct answer. To solve this problem, rewrite the formula for t using the algebraic property given in the term list at the end of the quiz

to simplify stacked fractions:

$$\frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}$$

Using this property, you should be able to see that

$$\begin{aligned} t &= \frac{\hat{\beta}_1 - 0}{\frac{s}{\sqrt{SS_{xx}}}} \\ &= \frac{\hat{\beta}_1}{s/\sqrt{SS_{xx}}} \\ &= \frac{\hat{\beta}_1 \sqrt{SS_{xx}}}{s} \end{aligned}$$

Since the problem says that the slope does not change, you need only be concerned about $\sqrt{SS_{xx}}/s$. For any fraction, changes in the numerator and denominator affect the total fraction as follows.

numerator	*denominator*	*total fraction*
↑	↑	depends on magnitude
↓	↓	depends on magnitude
↑	↓	increases
↓	↑	decreases

In the given case (both shrink) it depends entirely on the comparative magnitude of the decreases in SS_{xx} and s. By decreasing one more than the other, you can make the denominator grow or shrink and the magnitude of entire formula will then shrink or grow in the opposite direction of the denominator. If, on the other hand, they change in opposite directions, there is enough information to answer the question using the table above.

8. B is the correct answer. This problem is meant to emphasize the relationships between the quantities that make up the regression equation. You can often use some of these quantities to identify the other quantities, even if the obvious quantities are not available.

This problem requires some planning to solve. Your first impulse is probably to look up the definition of SS_{xy}, whereupon you will be stopped dead by the fact that $\sum x_i y_i$ is not given, nor is there another formula you could use to find it. Don't panic! Look through the list of terms and see what else is related to the given information!

Since you are given $\sum x_i^2$,n, and $\bar{x}$, you can find SS_{xx}. Will that help? There

is an equation with SS_{xy} and SS_{xx}, $\hat{\beta}_1 = {}^{SS_{xy}}/_{SS_{xx}}$, but that still leaves you with two unknowns.

Since you are given $\hat{\beta}_0$, you should look for formulas containing that quantity and see if they can help.

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

should look good because, if you substitute the formula for $\hat{\beta}_1$ into that equation, you have an equation with one unknown! Your plan is now **1.** Find SS_{xx}. **2.** Substitute SS_{xy}/SS_{xx} into the equation for $\hat{\beta}_0$. **3.** Rewrite the equation to solve for SS_{xy}.

$$\begin{aligned}
SS_{xx} &= \sum (x_i - \bar{x})^2 \\
&= \sum x_i^2 - n\bar{x}^2 \\
&= 1732 - 6(13)^2 \\
&= 718 \\
\hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \\
&= 15 - SS_{xy}/SS_{xx}(13) \\
&= 15 - SS_{xy}/718(13) \\
0.80501 &= 15 - SS_{xy}/718(13) \\
\frac{0.80501}{13} &= \frac{15}{13} - SS_{xy}/718 \\
SS_{xy}/718 &= \frac{14.19499}{13} \\
SS_{xy} &= 718\frac{14.19499}{13} \\
&= 784
\end{aligned}$$

9. D is the correct answer. Individual correlations can be calculated using the formula

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

or via a TI-83 by using the `LinReg` function which produces r as a byproduct if `DiagnosticsOn` is set by pressing `2nd`, `Catalog`, and paging down to `DiagnosticsOn` and pressing `Enter` twice. A quicker way if you have access to an Android phone or jailbroken iPhone is to use the statistics program called *R*. The correlation matrix can be calculated in the *R* statistics program as follows.

```
> j1<-c(3,4,5,7,7,8,9,9,11)
> j2<-c(11,10,10,8,6,6,4,3,3)
> j3<-c(7,4,5,7,7,6,5,5,4)
> j4<-c(9,10,11,12,16,17,19,19,21)
> j5<-c(3,22,5,51,7,64,9,79,11)
> options(digits=3)
```

```
> cor(cbind(j1,j2,j3,j4,j5))
       j1     j2      j3     j4      j5
j1  1.000 -0.958 -0.3114  0.961  0.3257
j2 -0.958  1.000  0.2959 -0.984 -0.3580
j3 -0.311  0.296  1.0000 -0.328  0.0473
j4  0.961 -0.984 -0.3278  1.000  0.2711
j5  0.326 -0.358  0.0473  0.271  1.0000
```

10. D is the correct answer. Individual correlations can be calculated using the formula

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

or via a TI-83 by using the `LinReg` function which produces r as a byproduct if `DiagnosticsOn` is set by pressing `2nd`, `Catalog`, and paging down to `DiagnosticsOn` and pressing `Enter` twice. A quicker way if you have access to an Android phone or jailbroken iPhone is to use the statistics program called *R*. The correlation matrix can be calculated in the *R* statistics program as follows.

```
> j1<-c(3,4,5,7,7,8,9,9,11)
> j2<-c(11,10,10,8,6,6,4,3,3)
> j3<-c(7,4,5,7,7,6,5,5,4)
> j4<-c(9,10,11,12,16,17,19,19,21)
> j5<-c(3,22,5,51,7,64,9,79,11)
> options(digits=3)
> cor(cbind(j1,j2,j3,j4,j5))
       j1     j2      j3     j4      j5
j1  1.000 -0.958 -0.3114  0.961  0.3257
j2 -0.958  1.000  0.2959 -0.984 -0.3580
j3 -0.311  0.296  1.0000 -0.328  0.0473
j4  0.961 -0.984 -0.3278  1.000  0.2711
j5  0.326 -0.358  0.0473  0.271  1.0000
```

11. B is the correct answer. The first issue to consider here is the relationship between p and q. Which one is being used to predict the other? In any single-variable regression problem the variables have a relationship that can be described in words as

regressor	regressand
input	output
known	unknown
independent	dependent
predictor	response
x	y

The phrasing of the question determines which variable is dependent, not the letters themselves, which could be replaced by any letters. This question is phrased *estimate q given p*, suggesting that p is known and q is unknown. Therefore, cast p in the role you'd normally reserve for x and cast q as y.

If you have to do this problem by hand, use the method in hw 3.6 and 3.7:

Step **1.** Find SS_{xx} and SS_{xy}.

x_i	y_i	x_i^2	$x_i y_i$
4	1	16	4
4	2	16	8
6	4	36	24
7	8	49	56
8	8	64	64
9	9	81	81
38	32	262	237

So the elements of the formulas for SS_{xx} and SS_{xy} are $\sum x = 38$, $\sum y = 32$, $\overline{x} = \sum x/n = 38/6 = 6.3333$, $\overline{y} = \sum y/n = 32/6 = 5.3333$, $\sum x_i^2 = 262$, and $\sum x_i y_i = 237$.

Substitute the above values into the following formulas.

$$\begin{aligned} SS_{xy} &= \sum x_i y_i - n\overline{x}\overline{y} \\ &= 237 - (6)(6.3333)(5.3333) = 34.3357 \\ SS_{xx} &= \sum x_i^2 - n(\overline{x})^2 \\ &= 262 - (6)(6.3333)^2 = 21.3359 \end{aligned}$$

Step **2.** Substitute those results into the formulas for $\hat{\beta}_1$ and $\hat{\beta}_0$.

$$\begin{aligned} \hat{\beta}_1 &= SS_{xy}/SS_{xx} = 34.3357/21.3359 = 1.6093 \\ \hat{\beta}_0 &= \overline{y} - \hat{\beta}_1\overline{x} = 5.3333 - 1.6093(6.3333) = -4.8589 \end{aligned}$$

Step **3.** Write the coefficients into the generic regression equation: $q = -4.8589 + 1.6093p$. Here I've substituted the letters from the problem statement to emphasize which is which. The dependent variable is traditionally written on the LHS (left hand side) and the independent variable and coefficients are written on the RHS (right hand side).

The problem was developed using the statistics app R:

```
> p<-c(4,4,6,7,8,9)
> q<-c(1,2,4,8,8,9)
> summary(lm(q~p))

Call:
```

```
lm(formula = q ~ p)

Residuals:
       1        2        3        4        5        6
-0.57813  0.42188 -0.79687  1.59375 -0.01562 -0.62500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -4.8594     1.4446  -3.364  0.02820 *
p             1.6094     0.2186   7.362  0.00181 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '. 0.1 ' '  1

Residual standard error: 1.01 on 4 degrees of freedom
Multiple R-squared:  0.9313,    Adjusted R-squared:  0.9141
F-statistic:  54.2 on 1 and 4 DF,  p-value: 0.001814
```

12. A is the correct answer. To solve this problem you need to calculate t, using the formula

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$
$$t = \frac{29.1 - 28.8}{3.9/\sqrt{5}}$$
$$t = .17$$

I checked my calculation with the math program Maxima.

```
(29.1-28.8)/(3.9/sqrt(5)),numer;
     .1720052290384458
```

The problem was developed using the stat program R as follows.

```
> y<-c(32,31,23.5,26.5,32.5);mean(y);sd(y)
[1] 29.1
[1] 3.927467
> t.test(y,alternative="greater",mu=28.8)

        One Sample t-test

data:  y
t = 0.1708, df = 4, p-value = 0.4363
alternative hypothesis:
true mean is greater than 28.8
95 percent confidence interval:
 25.35559      Inf
sample estimates:
mean of x
     29.1
```

13. C is the correct answer. To solve this problem you need to look up $t_{\alpha,\nu}$ in a table or calculator. The phrase *business decision with money riding on the outcome* should suggest a 95 percent confidence interval, meaning that $0.95 = 1 - \alpha, \alpha = 0.05$. Since a rational person would only purchase Bla if it improves gas mileage, we can conduct a one-tailed test. The Greek letter ν, pronounced nyoo, represents degrees of freedom, sometimes also abbreviated as df. In this case, the degrees of freedom equals the sample

size minus one for having calculated the mean, so $\nu = n - 1 = 5 - 1 = 4$. So $t_{\alpha,\nu} = t_{0.05,4}$ which can be found in the table near the end of the quiz as 2.132.

14. C is the correct answer. The correct calculation is

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_1(84) \\ &= -443.767 + 9.029(84) \\ &= 314.669\end{aligned}$$

15. E is the correct answer. If the model is satisfactory, the errors are not correlated and their mean is 0.

16. D is the correct answer. Use the formula $s^2 = SSE/(n-2)$ with $s = 4$ and $n = 24$ and solve for SSE.

17. A is the correct answer. Compare p to α and t_c to $t_{\alpha/2,df} = t_{0.05,113}$. This is because p and α represent probabilities and can be compared to each other.

18. E is the correct answer. The correct calculation is

$$\begin{aligned}\hat{y} &= \beta_0 + \beta_1(72) \\ &= -443.767 + 9.029(72) \\ &= 206.321\end{aligned}$$

19. E is the correct answer.

The 95 percent confidence interval for the mean of y for a particular $x = x_p$ is the subject of Mendenhall Section 3.9 and requires the formula in the box in the middle of page 129. Don't confuse the *confidence* interval in the middle box with the *prediction* interval in the bottom box.

You must take advantage of the fact that

$$\begin{aligned}s_{\hat{\beta}_1} &= s/\sqrt{SS_{xx}} \\ 0.772 &= \sqrt{296.6}/\sqrt{SS_{xx}} \\ \sqrt{SS_{xx}} &= 17.222/0.772 \\ SS_{xx} &= (0.772/17.222)^2 \\ &= 497.66\end{aligned}$$

You must also realize that MSE and s^2 are the same thing, the mean square error of the model, the estimate of σ^2 for the model, the subject of Mendenhall, section 4.5.

The rest is just lengthy calculation. We can can calculate

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= -207.617 + 5.605(74) \\ &= 207.153\end{aligned}$$

We can look up $t_{\alpha/2,df} = t_{0.025,60} = 2$ in the table on page 758 of Mendenhall. Next, use this knowledge to enter the appropriate information into the expression (from Mendenhall p 129)

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

$$207.153 \pm (2)(17.222)\sqrt{\frac{1}{62} + \frac{(74 - 72.9)^2}{497.66}}$$

$$207.153 \pm 4.69$$

$$(202.46, 211.84)$$

```
2*17.222*sqrt((1/62)+((74-72.9)^2/497.66436682))
  4.69253074480533
```

20. B is the correct answer. This question tests your understanding of the three choices 99 percent for life-or-death situations, 95 percent for situations where money is riding on the outcome and 90 percent for more casual situations. The cancer cure is a harrowing life-or-death situation, as are infant mortality rates. Improving earnings per share is a business issue. The others are more casual.

21. E is the correct answer. We have learned, in the box on page 47 of Mendenhall, as well as in section XXI of the study guide (HW 1.61), that the null hypothesis is couched as $H_0 : \mu = \mu_0$ while the alternative hypothesis is couched as one of the three possibilities $H_a : \mu \neq \mu_0$, $H_a : \mu < \mu_0$, or $H_a : \mu > \mu_0$. All of the choices except one are phrased as null hypotheses, that one thing is the same as another. The only difference is the pregnancy test, where the assertion that the new one is more accurate than the old is a typical alternative hypothesis.

This problem also provides a rationale for why this is standard practice. If we're going to claim that a new test for pregnancy is more accurate, we want to be very, very sure that we see evidence that that is the case. So we surely want the result to have to fall in a small region at the right of the probability density graph for us to pass it. Since the existing test is already in use, any result that falls into the huge region on the left (at least

ninety percent or perhaps as much as 99 percent depending on the significance level we choose) should prevent us from expressing confidence in improved accuracy from the new method.

22. C is the correct answer. Solve $(150 + 153 + 160 + x)/5$ for x.

```
> temper<-c(150,153,156,160,161)
> mean(temper)
[1] 156
```

23. B is the correct answer. Solve $146 = (140 + 146 + 150 + 151 + x)/5$ for x.

```
temper<-c(140,146,150,151,143)
mean(temper)
  [1] 146
```

24. D is the correct answer. This is an easier calculation than for the confidence interval of the mean of y for a given value of x. This question is only asking for a 95 percent confidence interval for the slope of $\hat{\beta}_1$, defined in Mendenhall page 111. The formula given on page 112 of Mendenhall is $\hat{\beta}_1 \pm (t_{\alpha/2,n-2})s_{\hat{\beta}_1}$ and all but t are given in the output. We can look up

$$t_{\alpha/2,n-2} = t_{0.025,60} = 2$$

in the table on page 758 of Mendenhall, so

$$\begin{aligned} \hat{\beta}_1 &\pm (t_{\alpha,df})s_{\hat{\beta}_1} \\ 5.605 &\pm (2)(0.772) \\ 5.605 &\pm 1.544 \\ (4.061, &7.149) \end{aligned}$$

```
> qt(0.975,60)
[1] 2
```

```
(%i1) 2*0.772;
(%o1)                          1.544
(%i2) 5.605-2*0.772;
(%o2)                          4.061
(%i3) 5.605+2*0.772;
(%o3)                  7.149000000000001
```

25. C is the correct answer. You may calculate this by hand if a machine is not allowed

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

or with a correlation function of any spreadsheet or stat software. If the software gives you a choice, select Pearson correlation for compatibility with the textbook.

```
> p<-c(8,2,5,3,8)
> q<-c(1,9,7,2,3)
> r<-c(7,6,5,0,1)
> options(digits=3)
> cor(cbind(p,q,r))
       p      q      r
p  1.0000 -0.614 0.0636
q -0.6137  1.000 0.3599
r  0.0636  0.360 1.0000
```

26. B is the correct answer. The expression *percentage of variation in the data is explained by the model* is the definition of R^2, so the value of R^2 is the answer to the question.

$$\begin{aligned} R^2 &= (SS_{yy} - SSE)/SS_{yy} \\ &= ((395.67 + 565.86) - 565.86)/(395.67 + 565.86) \\ &= .41 \end{aligned}$$

Also, be aware that, for any two integers a and b,

$$\frac{a-b}{a} = \frac{a}{a} - \frac{b}{a} = 1 - \frac{b}{a}$$

so you can also solve this problem by saying

$$\begin{aligned} R^2 &= 1 - \frac{SSE}{SS_{yy}} \\ &= 1 - \frac{565.86}{395.67 + 565.86} \\ &= .41 \end{aligned}$$

27. E is the correct answer. There is no information about the *explained* variation in the model. Both MSE and SSE provide only information about the *unexplained* variation. You would need either information about total variation or explained variation to compare to the unexplained variation to solve this problem.

28. E is the correct answer. SS_{yy} is not necessary. $\bar{x}$, and $\bar{y}$ are necessary. You would also be able to form a least squares estimate from x and y, the raw sample values (if those were given as choices which they are not) but the raw sample values are not statistics.

29. D is the correct answer. To know whether to use the model means to know whether the model is any good. To know whether the model is any good means to know whether the model can actually help predict an output given an input. The statistics $t_{\hat{\beta}_1}$ and R^2 are statistics that tell us whether the model has a fighting chance of predicting anything.

To know whether to use the model we need to assess the slope, $\hat{\beta}_1$, using the t statistic. If we find the slope is not 0, then we need to assess R^2.

The other issue here is that you have to choose cost as the independent variable (the one you can control) and benefit as the dependent variable (the one dependent on the model).

```
> j1<-c(3,4,5,7,7,8,9,9,11)
> j4<-c(9,10,11,12,16,17,19,19,21)
> options(digits=3)
> summary(lm(j4~j1))

Call:
lm(formula = j4 ~ j1)

Residuals:
   Min     1Q Median     3Q    Max
-2.889 -0.481  0.463  0.815  1.111

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)     3.35       1.33    2.52     0.04 *
j1              1.65       0.18    9.17  3.8e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 1.32 on 7 degrees of freedom
Multiple R-squared:  0.923,
Adjusted R-squared:  0.912
F-statistic: 84.1 on 1 and 7 DF,  p-value: 3.77e-05
```

30. C is the correct answer. We need to know the estimated regression equation and $\bar{y}$. We can then solve for the x that would achieve $\bar{y}$. The dependent variable is the *achieved* level (what we can't control directly) so we form the regression model with achieved level as y and operating cost (what we can control) as x. The estimated model is $y = 12.4 - 0.8x$. The mean of y is 7, so we want to know x if $7 = 12.4 - 0.8x$. Rewrite it as $0.8x = 12.4 - 7$, $0.8x = 5.4$, $x = 5.4/0.8$, $x = 6.75$.

```
> options(digits=3)
> x<-c(11,10,10,8,6,6,4,3,3)
> y<-c(3,4,5,7,7,8,9,9,11)
> summary(lm(y~x))

Call:
lm(formula = y ~ x)

Residuals:
   Min     1Q Median     3Q    Max
-1.020 -0.622 -0.221  0.576  0.980

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  12.4183     0.6676   18.60  3.2e-07 ***
x            -0.7994     0.0904   -8.84  4.8e-05 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 0.796 on 7 degrees of freedom
Multiple R-squared:  0.918, Adjusted R-squared: 0.906
F-statistic: 78.2 on 1 and 7 DF, p-value: 4.78e-05
```

31. A is the correct answer.

```
> options(digits=3)
> x<-c(9,10,11,12,16,17,19,19,21)
> y<-c(11,10,10,8,6,6,4,3,3)
> summary(lm(y~x))

Call:
```

```
lm(formula = y ~ x)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9517 -0.1385  0.0483  0.4231  0.6734

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  17.0126     0.7298    23.3  6.8e-08 ***
x            -0.6874     0.0472   -14.6  1.7e-06 ***
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 0.595 on 7 degrees of freedom
Multiple R-squared: 0.968, Adjusted R-squared: 0.964
F-statistic:  212 on 1 and 7 DF, p-value: 1.71e-06
```

32. E is the correct answer. A null hypothesis needs three characteristics:

- It must describe two quantities that can be measured as numbers.
- It must be expressed as an equality, not as an inequality.
- The equality must represent a *status quo* so that, if we mistakenly fail to reject it, we are mistakenly following conventional wisdom rather than mistakenly asserting something that people don't already believe.

The first choice includes some measurable quantities but also includes *wise* as an outcome. Can you measure how wise an early riser is compared to a late riser?

The second choice expresses *the lure of gold* as a cause but can you measure the lure of gold? It's a common expression as a motivation for avarice or greed but is usually spoken in conjunction with some clearly defined extreme situation, like that of Volpone or Midas or Ebeneezer Scrooge. These situations inevitably contain something measurable. Additionally, *evil* may be controversial. Each side in a war may consider the other side evil.

The third choice uses a highly controversial quantity, the bias in a referee's calls. Each side in a close contest often complains that the referee is biased toward the other side, so the interested parties are unlikely to agree on bias.

The fourth choice does specify a quantity, time outs remaining, but does not go far enough since it does not specify an outcome associated with more or fewer time outs remaining.

The fifth choice is typical of a null hypothesis. Bot the diagnosis of lung cancer and the alternative, the diagnosis of lung cancer are measurable. The number of cigarettes smoked is clearly measurable and can be observed for those diagnosed with or without lung cancer. The alternative hypothesis in this case could be that a patient smoking more cigarettes is more

likely to be diagnosed with lung cancer or that the two diagnoses are not equally likely for two groups of patients one smoking more than the other.

33. A is the correct answer. All except the first and second are difficult to measure because they are not defined and different people have conflicting definitions.

Among the first two choices, the difference is that one is written like a null hypothesis and the other is written like an alternative hypothesis.

The first choice says that there is no relationship between two variables. If we have a group of purchasers who websurfed and a group who did not and could hold all else equal, we would find no difference in customer satisfaction, a variable that is among the most frequently measured variables in business.

The second choice claims a relationship between cause and effect and would therefore be a good example of an alternative hypothesis. To convert it to a null hypothesis, it could be rewritten to say that a new medicine does not reduce fever.

34. D is the correct answer. All posit the lack of a relationship between two things except *never say*, which is phrased as a command rather than an assertion. *No man*, *all of the people*, *all of the time*, *nothing*, and *lasts forever* are all untestable, either infinite or, in the case of *no man* and *nothing*, impossible to observe without omniscience. The consumer one, that *consumers expect no difference*, is the safest because it does not specify all people and gives an explicit measurement process.

35. B is the correct answer. It occurred to me that maybe people are having trouble with meaning of scientific notation. The e-06 means to move the decimal place 06 places to the left, which results in five zeros between 3 and the decimal point.

36. E is the correct answer. All except *D* can be found by using the formula for R^2 in the term list. Choice *D* expresses the situation where there are no degrees of freedom, meaning that the values of sample variables are not free to vary at all. This is because the simple linear regression relationship has $n - (k + 1)$ degrees of freedom and $k = 1$. So, if $n = k + 1$ it must follow that $n = 2$ and the simple linear regression relationship has 0 degrees of freedom. You can see this result by placing two points on a graph. You can always draw a line between them, meaning that both points lie on the line and the line completely explains the relationship between the two points.

37. A is the correct answer. Best to use a calculator function like `normalcdf(`*lower bound, upper bound, mean, standard deviation*`)`

to solve this, using the values 10, 20, 5, 15.

38. B is the correct answer. Best to use a calculator function like `normalcdf(`*lower bound, upper bound, mean, standard deviation*`)` to solve this, using the values 60, 80, 50, 8.

39. C is the correct answer. Draw a picture and shade the area you are looking for—it should be clear that it is half of the total area.

40. C is the correct answer.

1. calculate $z = (70 - 50)/15 = 4/3 \sim 1.33$

2. look up $z = 1.33 \Rightarrow p(y > 70) = 0.09175914$

```
> pnorm(-1.33)
[1] 0.09175914
```

41. A is the correct answer. 1. calculate $z = (30-50)/15 = -4/3 \sim -1.33$

2. look up $z = 1.33 \Rightarrow P(y < 30) = 0.09175914$

```
> pnorm(-1.33)
[1] 0.09175914
```

42. E is the correct answer.

1. calculate $z = (30 - 50)/15 = -4/3 \sim -1.33$

2. calculate $z = (40 - 50)/15 = -2/3 \sim -0.67$

3. look up $z = -1.33 \Rightarrow P(y < 30) = 0.09175914$

3. look up $z = -0.67 \Rightarrow P(y < 40) = 0.2514289$

4. subtract $0.2514289 - 0.09175914 = 0.1596698$

```
> pnorm(-1.33)
[1] 0.09175914
> pnorm(-0.67)
[1] 0.2514289
> pnorm(-0.67)-pnorm(-1.33)
[1] 0.1596698
```

43. C is the correct answer.

```
> y<-c(3,22,5,51,7,64,9,79,11)
> mean(y)
[1] 27.9
> sum(y^2)
[1] 13707
> sd(y)
[1] 29
(%i49) sqrt((13707-9*27.9^2)/8),numer;
(%o49)                          28.94242128779139
```

44. E is the correct answer. The test statistic is t_c as reported in the Coefficients table.

45. B is the correct answer. Look at the formula for t and enter any values. Then change them in accordance with the problem statement and you will see that the denominator must grow while the numerator must shrink.

$$\frac{\beta}{s/\sqrt{SSxx}} = \frac{\beta\sqrt{SSxx}}{s}$$

46. C is the correct answer. Since the sample size is $n = 9$ which is less than 30, you should use the formula for a small sample test statistic, given in the term list:

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}} = \frac{7 - 0}{2.598/\sqrt{9}} = 8.08314$$

```
> j1<-c(3,4,5,7,7,8,9,9,11)
> sd(j1)
[1] 2.598076
> t.test(j1)

        One Sample t-test

data:  j1
t = 8.0829, df = 8, p-value = 4.054e-05
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 5.002942 8.997058
sample estimates:
mean of x
        7
```

47. A is the correct answer. This question tests two things. First, if you did HW 1.61, you encountered exactly these numbers in exactly this context. When we discussed the problem in class, I mentioned that the given p value was vastly smaller than any of the three possible rejection regions, 0.1, 0.05, or 0.01. So the comparison of the calculated p value with the tabulated t-statistic should have been easy to do without consulting a table or calculator. If you did the homework and went over it with us in class and this is still not obvious, you should to do more homework problems of the same type, or redo that problem, making up different numbers for practice.

The second thing that is being tested is the statement of the conclusion. When we fail to reject the null hypothesis, it means that the test result has not provided any evidence to contradict the status quo but we admit that there might be other evidence that has not been presented. Thus, the answer beginning with *Fail to reject ... conclude we have seen ...* is not ever

a conclusion we would draw from a statistical test. Similarly, we would never reject the null hypothesis because we have not seen evidence. The rejection of the null hypothesis indicates that we have seen evidence and, in this case, the extraordinarily small p value means that we have seen overwhelming evidence.

48. B is the correct answer. The regression process has three possible steps:

1. Form a model
2. Assess a model
3. Use a model to predict or estimate

The problem statement indicates that the model has already been formed and assessed and is ready to be used. To use the model to predict a value of y given x, use the estimate of the slope, $\hat{\beta}_1$ to predict the effect on y of a one-unit increase in x.

49. D is the correct answer.

```
> y<-c(3,22,5,51,7,64,9,79,11)
> var(y)
[1] 838.3611
> sd(y)
[1] 28.95447
```

50. C is the correct answer. R regression of complete model:

```
> summary(lm(price~livingArea+baths+bedrooms+fireplace+acres+age))

Call:
lm(formula = price ~ livingArea + baths + bedrooms + fireplace +
    acres + age)

Residuals:
    Min      1Q  Median      3Q     Max
-115606  -16784     708   16191   78257

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -27328.87   18096.75  -1.510  0.13443
livingArea      50.46      10.92   4.620 1.25e-05 ***
baths        21544.11    8486.36   2.539  0.01281 *
bedrooms      9701.31    6184.51   1.569  0.12016
fireplace    14563.26    7230.15   2.014  0.04690 *
acres        13297.77    4455.01   2.985  0.00363 **
age            -71.55      75.66  -0.946  0.34682
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 33590 on 92 degrees of freedom
Multiple R-squared:  0.6287,    Adjusted R-squared:  0.6045
F-statistic: 25.97 on 6 and 92 DF,  p-value: < 2.2e-16
```

The two least promising predictors are those whose t statistics have the smallest absolute value: bedroom and age.

R regression of reduced model:

```
> summary(lm(price~livingArea+baths+fireplace+acres))
```

```
Call:
lm(formula = price ~ livingArea + baths + fireplace + acres)

Residuals:
    Min      1Q  Median      3Q     Max
-113959  -17786     798   16797   84835

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -11122.49   12688.65  -0.877  0.38295
livingArea      52.72      10.15   5.195 1.19e-06 ***
baths        25101.54    7659.47   3.277  0.00147 **
fireplace    14438.38    7261.44   1.988  0.04968 *
acres        13938.26    4455.27   3.128  0.00234 **
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '  1

Residual standard error: 33760 on 94 degrees of freedom
Multiple R-squared:  0.6168,    Adjusted R-squared:  0.6005
F-statistic: 37.83 on 4 and 94 DF,  p-value: < 2.2e-16
```

The difference is $0.6287 - 0.6198 = 0.0089 < 0.02$

51. B is the correct answer. The correct choice above is the definition of *s*.

52. B is the correct answer. One way to find *s* is by taking the square root of *MSE*.

53. B is the correct answer. The test statistic is $F_c = 50$. The corresponding p-value can be found using RStudio. The appropriate function is

```
pf(50,7,91,lower.tail=FALSE)
```

where $F_\alpha = 50, v_1 = 7, v_2 = 91$. The result is $p = 1.713{,}567 \times 10^{-28}$ or 17 preceded by 27 zeros in decimal notation.

54. A is the correct answer. This question tests whether you can read a scatterplot matrix. This is a symmetric matrix. That means that the entry in the upper left corner is a mirror image of the entry in the lower right corner.

55. E is the correct answer. The test statistic for the test of overall model adequacy is the *F* statistic, found by

$$F_c = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

$$F_c = \frac{0.95/6}{(1-0.95)/[67-(6+1)]}$$

$$F_c = 190$$

```
(0.95/6)/((1-0.95)/(67-(6+1)));
             189.9999999999998
```

56. E is the correct answer. The result of the test of overall model adequacy is either to reject the null hypothesis and conclude that we have seen

evidence that at least one β_1 through β_3 coefficient is nonzero (this case) or to fail to reject the null hypothesis and conclude that all the betas equal zero.

57. A is the correct answer.

```
> summary(lm(price ~ livingArea+baths+Bedrooms
         +I(livingArea*baths)+I(baths*Bedrooms)))

Call:
lm(formula = price ~ livingArea + baths
            + Bedrooms + I(livingArea * baths)
            + I(baths * Bedrooms))
```

The difference is $0.658 - 0.6209 = 0.0371$ or nearly 4 percent.

58. E is the correct answer. All the statements are roughly true. You might try to argue about the word *much* and say that even one percent is massive but I can envision no argument you'd win. The only real difference between these three models that I can see is not a difference of explanatory power but a difference in the application of the principle known as Occam's Razor: if two explanations offer similar results, choose the simpler. That principle would argue for the use of only GEN_SUP. And yet, GEN_SUP is an artificial variable constructed from the other two.

59. E is the correct answer. The choices B through D are different ways of stating the meaning of the coefficient, with choice B being the best way of stating it. The problem with these statements is that the small t statistic and large p value tell us that the salaries fluctuate so much that, alone, GENDER tells us nothing meaningful. Even though these are the interpretations of the beta for gender, they are not useful interpretations.

60. E is the correct answer. The key to this whole problem is the unequal sample sizes of male and female executives. With so few females, any evidence of a pattern becomes exaggerated. The absence of female executives at the lower numbers of employees creates the impression of a pattern. The interaction term essentially removes the skewed subsample of female executives from the sample and makes it easier to see the relationship between salary and number of employees in the remainder of the sample.

61. A is the correct answer. The proportion of variability explained is R^2 which is about 91 percent in the first model and 74 percent in the second model. There is less than a one percent difference between either R^2 and the corresponding R_a^2 value. Both models explain more than half of the variability. The F statistic can be computed from R^2 if you also know the sample size and number of predictors. One formula for F is a ratio with R^2 in the numerator and $1 - R^2$ in the denominator so they are closely related.

62. B is the correct answer. This question is asking for *SSE*.

63. C is the correct answer. The appropriate statistic is the F statistic for comparing a reduced model to a complete model.

$$F_c = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/[n-(k+1)]}$$

Begin by reading SSE_R from the SPSS reduced model output and SSE_C from the SPSS complete model output. You can find n in either the SPSS data view where the observations are numbered 1 to 15 or in the SPSS output, where the numerator and denominator degrees of freedom are given in the ANOVA table: 2 and 9. Since denominator degrees of freedom are $n-(k+1)$ and numerator degrees of freedom are $k = 2, 9 = n-(2+1)$, $n = 15$. Substitute these values into the above equation for F_c:

$$\begin{aligned} F_c &= \frac{(3331000 - 2098183)/(5-3)}{2098183/[15-(5+1)]} \\ &= \frac{(1232817)/2}{2098183/9} = \frac{(1232817)\cdot 9}{2098183 \cdot 2} \\ &= 2.644 \end{aligned}$$

Note that I used the following identity:

$$\frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}$$

You could use a calculator as follows:

```
((3331000-2098183)/(5-3))/(2098183/(15-(5+1))),numer;
                          2.644038437066738
```

64. A is the correct answer. The F statistic should be compared to the rejection region boundary, $F_{\alpha,v_1,v_2} = F_{0.05,k-g,n-(k+1)}$ which can be found by a table or calculator.

```
> qf(0.05,2,9,lower.tail=FALSE)
[1] 4.256495
```

65. B is the correct answer. There are $k-g$ numerator degrees of freedom, where $k-g = 5-3 = 2$.

66. A is the correct answer. The sign of β_2 determines the shape of the curve in a quadratic regression. The less promising of the two quadratic terms is the one with the smallest absolute value of the t statistic, which is equivalent to saying the one with the largest p value. In this case that is HOURSSQ with $t = 0.778$ and $p > 0.76$. So the model should now have only AGE as x_1 and AGESQ as x_2 and $\beta_2 = -2.22$ so the slope looks like a diminishing returns curve.

If $\beta_2 > 0$ you would have epidemic growth and if $t_{\beta_2} < t_{\alpha/2,n-(k+1))}$, then you would have a linear relationship. This is because such a small t_c would indicate that the quadratic term is contributing nothing to the predictive power of the model.

67. D is the correct answer. The appropriate statistic is the F statistic for comparing a reduced model to a complete model.

$$F_c = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/[n-(k+1)]}$$

Begin by reading SSE_R from the SPSS reduced model output and SSE_C from the SPSS complete model output. You can find n in either the SPSS data view where the observations are numbered 1 to 15 or in the SPSS printout, where the numerator and denominator degrees of freedom are given in the ANOVA table: 2 and 9. Since denominator degrees of freedom are $n-(k+1)$ and numerator degrees of freedom are $k = 2, 9 = n-(2+1)$, $n = 15$. Substitute these values into the above equation for F_c:

$$\begin{aligned} F_c &= \frac{(2250956 - 2098183)/(5-3)}{2098183/[15-(5+1)]} \\ &= \frac{(152773)/2}{2098183/9} = \frac{(152773)\cdot 9}{2098183 \cdot 2} \\ &= 0.328 \end{aligned}$$

Note that I used the following identity:

$$\frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}$$

You could use a calculator as follows:

```
((2250956-2098183)/(5-3))/(2098183/(15-(5+1))),numer;
                        0.3276542131930342
```

68. E is the correct answer. First, distinguish between terms and indicator variables. The indicator variables are $x_1, x_2, \ldots$. The terms are the terms of a linear equation. Each term of a linear equation is what is added together to form the sum. So β_0 is a term by itself and every other combination between a pair of plus signs is a term. It is customary to treat β_0 separately because it represents a constant so it does not participate in the list of 11 terms we will discover as we solve this problem.

You need almost as many terms as there are combinations of characteristics, but one term may be omitted and the value of β_0 used to account for one combination of characteristics.

You need enough indicator variables to represent the characteristics. For each characteristic, if the value is set to 1 it means one value, and set to 0 it means the other value. In this case, there are three sizes, times two colors, times two styles, so $3 \times 2 \times 2 = 12$ is the number of cases to be represented. Since each indicator variable can take on only the values $\{0, 1\}$, the number of indicator variables is given by the smallest power of 2 that can represent all the combinations. For example,

$$2^1 = 2, \quad 2^2 = 3, \quad 2^3 = 8, \quad 2^4 = 16$$

The first three powers of two are insufficient to describe the 12 combinations mentioned above. The smallest power of 2 that will suffice is 4, so we need four indicator variables, x_1, x_2, x_3, x_4.

In a simple case like this, we can take advantage of the fact that we can represent the size *large* by setting $x_1 = x_2 = 0$. If $x_1 = 1$ it means the size is small. If $x_2 = 1$ it means the size is medium. We can never have $x_1 = x_2 = 1$ because that would mean that the size is both small and medium, so we will not try to account for that case in the table below. Similarly, we can set $x_3 = 1$ if the color is green and 0 if the color is yellow. Finally, we set $x_4 = 1$ if the style is plain and 0 if it is fancy.

The next step is to make a table of the combinations we can use to write the regression model. It's not really strictly necessary in such a simple case but it should help you to understand and it may prevent mistakes in other, more complicated cases. The following table shows all possible combinations using four indicator variables.

The column headings are s: small, m: medium, g: green, p: plain.

```
          s m g p
no attributes
  beta  0 0 0 0 0
all individual attributes
  beta  1 1
  beta  2   1
  beta  3     1
  beta  4       1
combinations of two attributes
  beta  5 1   1
  beta  6 1     1
  beta  7   1 1
  beta  8   1   1
  beta  9     1 1
combination of three attributes
  beta 10 1   1 1
```

```
beta 11   1 1 1
```

Notice that I organized this table to reduce the chances that I have missed any cases. Note also that the number of rows differs from the solution if a hat could be small and medium at the same time.

Now that we have constructed this table, we can write the regression equation by creating terms composed of the 11 beta coefficients and, for each row, the x_i variables that will accompany that beta coefficient. Hence, the regression model is

$$\begin{aligned} E(y) =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_1 x_3 \\ &+ \beta_6 x_1 x_4 + \beta_7 x_2 x_3 + \beta_8 x_2 x_4 + \beta_9 x_3 x_4 \\ &+ \beta_{10} x_1 x_3 x_4 + \beta_{11} x_2 x_3 x_4 + \varepsilon \end{aligned}$$

You might wonder why we need eleven terms when there are twelve combinations. This is because, the way we've written it, β_0 now represents the contribution of a large, yellow, fancy hat. This is the one combination where every $x_i = 0$, so $E(y) = \beta_0$. Each other beta, $\beta_1, \ldots, \beta_k$ represents the contribution of exatly one possible type of hat.

69. A is the correct answer. Since $k = 3$ and $n = 30$, $n - (k + 1) = 26$ and $F_{0.05,3,26} = 2.98$.

```
qf(0.05,3,26,lower.tail=F)
[1] 2.975154
```

70. E is the correct answer.

$$\begin{aligned} F_c &= \frac{R^2/k}{(1 - R^2)/[n - (k + 1)]} \\ &= \frac{0.193/2}{(1 - 0.193)/[31 - (2 + 1)]} \\ &= 3.348 \end{aligned}$$

```
(%i4) solve((3.35=((x/2)/((1-x)/(31-3)))),x),numer;
(%o4)                        [x = 0.1930835734870317]
```

71. C is the correct answer. Since $k = 3$ and $n = 34$, $n - (k + 1) = 30$ and $F_{0.05,3,30} = 2.92$.

```
qf(0.05,3,30,lower.tail=F)
[1] 2.922277
```

72. B is the correct answer. The appropriate statistic is the F statistic for comparing a reduced model to a complete model.

$$F_c = \frac{(SSE_R - SSE_C)/(k - g)}{SSE_C/[n - (k + 1)]}$$

Begin by reading SSE_R from the SPSS reduced model output and SSE_C from the SPSS complete model output. You can find n in either the SPSS data view where the observations are numbered 1 to 15 or in the SPSS output, where the numerator and denominator degrees of freedom are given in the ANOVA table: 2 and 9. Since denominator degrees of freedom are $n - (k + 1)$ and numerator degrees of freedom are $k = 2, 9 = n - (2 + 1)$, $n = 15$. Substitute these values into the above equation for F_c:

$$\begin{aligned} F_c &= \frac{(2250956 - 2098183)/(5 - 3)}{2098183/[15 - (5 + 1)]} \\ &= \frac{(152773)/2}{2098183/9} = \frac{(152773) \cdot 9}{2098183 \cdot 2} \\ &= 0.328 \end{aligned}$$

Note that I used the following identity:

$$\frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}$$

You could use a calculator as follows:

```
((2250956-2098183)/(5-3))/(2098183/(15-(5+1))),numer;
                        0.3276542131930342
```

73. B is the correct answer. The appropriate formula is $n - (k + 1) = 15 - (3 + 1) = 11$

74. A is the correct answer. The expression "numerator degrees of freedom" refers to k, the number of predictors, in this case 3, AGE, HOURS, and AGESQ.

75. A is the correct answer. Use the formula for F_c along with k and n to find R^2.

76. A is the correct answer. The sign of the quadratic term determines the curve's rise like epidemic growth or flattening like diminishing returns.

Since it is negative (−1.6464), we see a diminishing returns curve. If it were positive, say, 1.6464 without the minus sign, we'd see the appearance of epidemic growth.

Of course, this explanation assumes that the quadratic term is meaningful. If I were to change the p value to some value such that $p > 0.1$ in the cell that contains 0.0262 above, it would signal that there is no contribution to the model from the quadratic term.

77. C is the correct answer. This total variability in y is not changed by the addition of a new explanatory variable.

78. A is the correct answer. As it stands, you would NEVER select B or D in any statistics class that I have ever heard of. At the beginning of this course, and probably throughout MATH 243, we said that we may either reject H_0 because we *have* seen evidence or fail to reject H_0 because we have not seen evidence. The opposite cases are not possible because H_0 is supposed to be phrased as the status quo and H_a challenges the status quo. If the challenge fails, it does not mean that all challenges would fail, just that we have not yet seen any evidence to that effect.

79. C is the correct answer. If the model is satisfactory, the errors are not correlated, their mean is 0, and they are normally distributed.

80. C is the correct answer. These are the reasons given in Mendenhall to suspect the presence of multicollinearity, except for the answer about R_a^2 which is an indication of the use of too many predictors for the given number of observations.

81. D is the correct answer.

$$\begin{aligned} VIF_1 &= \frac{1}{1 - R_1^2} \\ VIF_1 &= \frac{1}{1 - 0.81} \\ VIF_1 &= \frac{1}{0.19} \\ VIF_1 &= 5.26 \end{aligned}$$

82. C is the correct answer. This is an extremely simple case that could be solved by what you may remember as the *multiplication principle* from Math 243, where $2 \times 2 = 4$ is the number of unique combinations of size and style.

It may make sense to think about the problem in general terms rather than to memorize the above answer, since the problem could be easily made more complicated.

The general question here requires you first to form a model, then figure out what happens if that model turns out to describe, among other things, something that has no significant effect on profitability. So you have to first construct the model, then subtract from it those things that do not matter.

In this very, very simple case, you can just use what you know of combinations from prior classes, that $2 \times 2 \times 2 = 8$ possible combinations. Then, if you remove color, only $2 \times 2 = 4$ possible combinations remain.

Since you can not be guaranteed to see such a simple problem, it may be helpful to see how each combination is used in the regression model. The easiest way to do so is to create a table as for the example shown in class.

You will need almost as many terms as there are combinations of characteristics, but one term may be omitted and the value of β_0 used to account for one combination of characteristics.

You need enough indicator variables to represent the characteristics. For each characteristic, if the value is set to 1 it means one value, and set to 0 it means the other value. In this case, there are two sizes, times two colors, times two styles, so $2 \times 2 \times 2 = 8$ is the number of cases to be represented. Since each indicator variable can take on only the values $\{0, 1\}$, the number of indicator variables is given by the smallest power of 2 that can represent all the combinations. For example,

$$2^1 = 2, \quad 2^2 = 3, \quad 2^3 = 8$$

The first two powers of two are insufficient to describe the 8 combinations mentioned above. The smallest power of 2 that will suffice is 3, so we need three indicator variables, x_1, x_2, x_3. The table will therefore have three columns, one for each indicator variable.

The column headings are s: small, g: green, p: plain.

```
         s g p
no attributes
  beta 0 0 0 0
all individual attributes
  beta 1 1
  beta 2   1
  beta 3     1
combinations of size and color
  beta 4 1 1
  beta 5 1   1
  beta 6   1 1
combinations of type and pattern
  beta 7 1 1 1
```

You could then count the rows with a 1 for color. Alternatively you could count the rows with a 0 for color.

Now that we have constructed this table, we can write the regression model by creating terms composed of the 7 beta coefficients and, for each row, the x_i variables that will accompany that beta coefficient. Hence, the regression model is

$$\begin{aligned} E(y) =& \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 \\ &+ \beta_5 x_1 x_3 + \beta_6 x_2 x_3 + \beta_7 x_1 x_2 x_3 + \varepsilon \end{aligned}$$

If color does not matter, the terms containing x_2 will have insignificant t statistics. There are four such terms. We can alternatively choose to remove the four terms that describe yellow hats, i.e., terms with $x_i = 0$ instead. If we do that instead, we will remove the terms with coefficients β_2, β_4, and β_6. What, wut? That's only three terms? What about the eighth combination, not included in the terms with coefficients $\beta_1, \ldots, \beta_7$? That combination is a large, yellow, fancy hat and its contribution is modeled by β_0. We never remove β_0 from a regression model but instead we make it represent the contribution where all x_i indicator variables are zero, in this case a large, fancy hat.

83. B is the correct answer. It occurred to me that maybe people are having trouble with meaning of scientific notation. The e-14 means to move the decimal place 14 places to the left, which results in thirteen zeros between 113 and the decimal point.

84. B is the correct answer. The expression *right hand side* refers to the right hand side of the equals sign. You should have encountered this expression in many math classes.

The expression *least squares estimate* refers to the parameter estimates generated by hand or, as in this case, by computerized regression analysis.

This estimated regression equation differs from the general regression equation, such as is expressed by choice D. The general regression equation expresses the model without taking a sample and analyzing the sample. The general regression equation for any linear model with two predictors looks the same. The estimates, generated using the method of least squares on samples, may by unique and, in any case, replace the Greek letters with numerical estimates.

The column headed B in the coefficients table gives the beta estimates for the estimated regression equation.

Don't be confused by the order of the terms. Any linear combination of terms may be reordered in any way and you will still have $a + b = b + a$ for any pair of real numbers.

85. E is the correct answer. The test statistic is t_c while the value to compare to it is $t_{\alpha,\nu} = t_{0.025,113}$.

86. B is the correct answer. The total variability, SST, does not change just because we try to explain it in a different way. When SSE decreases and SST does not change, SSR must inevitably increase because $SST = SSR + SSE$.

87. B is the correct answer. You must know that $SS_{yy} = SST$ and that $R^2 = (SST - SSE)/SST$ to answer this question.

```
(190501-112111)/190501,numer;
```

```
              .4114939029191448
(190501-45667)/190501,numer;
              .7602794735985637
```

88. B is the correct answer. The relevant formula is

$$\widehat{\beta}_i \pm t_{\alpha/2,n-(k+1)} s_{\widehat{\beta}_i}$$

$t_{0.025,27} = 2.052$ If you did not divide α by 2, you may have said $t_{0.05,27} = 1.703$ which leads to the incorrect result in E.

If you did not use the formula above but instead relied on the crude rule of thumb presented on page 21 of Mendenhall, you would obtain the incorrect result given in C because you substituted 2 for 2.052.

```
      /* yes */
3.7752+2.052*0.7704;
      5.3560608
3.7752-2.052*0.7704;
      2.1943392
      /* no */
ev(3.7752+1.703*0.7704,numer);
         5.087191199999999
ev(3.7752-1.703*0.7704,numer);
         2.4632088
      /* also no */
3.7752-2*0.7704;
              2.2344
3.7752+2*0.7704;
              5.316
```

89. D is the correct answer. Compare t_c, the test statistic, to

$t_{\alpha,v} = t_{0.025,113} = 1.98118.$

```
qt(0.025,113,lower.tail=F)
[1] 1.98118
```

90. E is the correct answer. There is no information about the *explained* variation in the model. Both MSEand SSEprovide only information about the *unexplained* variation. You would need either information about total variation or explained variation to compare to the unexplained variation to solve this problem.

91. A is the correct answer.

$$\begin{aligned} F_c &= \frac{(SS_{yy} - SSE)/k}{SSE/[n-(k+1)]} \\ &= \frac{R^2/k}{(1-R^2)/[n-(k+1)]} \\ &= \frac{.6928/2}{(1-.6928)/[150-(2+1)]} \\ &= 165.7 \end{aligned}$$

```
(.6928/2)/((1-.6928)/(150-(2+1)));
        165.7578125
```

92. A is the correct answer. You would NEVER select B, C, or D in any statistics class that I have ever heard of. At the beginning of this course, and probably throughout MATH 243, we said that we always couch the conclusion to a test in terms of H_0. Either we reject H_0 or we fail to reject H_0. We never frame our conclusion in terms of H_a ever. Any statistician who does so is violating the Code Of The West and can never again hold up his or her head in a confab of statisticians. Therefore, cross C and D off the list before looking beyond the part that says H_a. As for B, it contains an error of a different type. When writing H_0 :, you should always follow the colon with the words that express H_0. Here, the opposite words have been used. This may seem like a trivial point but, if you want to say H_0 and follow it with the words of the alternative, you should say something like "H_0 and conclude that" instead of saying something like "H_0 : Conclude that" so that it is clear that you are not writing out the words that express H_0 but are instead using the conjunction "and" to separate H_0 from the alternative.

93. B is the correct answer. This differs from the small sample version in that the $n = 500$ overwhelms the number of predictors in the expression for R^2_a, which turns out to be almost identical to R^2.

$$\begin{aligned} R^2_a &= 1 - \left(\frac{n-1}{n-(k+1)}\right)(1-R^2) \\ &= 1 - \left(\frac{500-1}{500-(2+1)}\right)(1-.81) \\ &= .809 \end{aligned}$$

```
1-((500-1)/(500-(2+1)))*(1-.81);
       .8092354124748491
```

94. A is the correct answer. The key to solving this problem is to recognize it is a nested model and that you need to compare it to

$$F_{\alpha,k-g,n-(k+1)} = F_{0.05,2-1,5-(2+1)} = 18.5$$

a very high bar because of the very small sample size.

95. D is the correct answer. The model with 1 predictor is nested within the model with three predictors, so we compare the test statistic, F_c to an F statistic having the appropriate numerator and denominator degrees of

freedom for a nested model, $F_{\alpha,k-g,n-(k+1)}$. One predictor is not controversial, hence, $g = 1$. There are a total of three predictors, hence $k = 3$. Since $k - g = 2$ and $n = 5$, $n - (k + 1) = 5 - (3 + 1) = 1$ and $F_{0.05,2,1} = 199$. With so few observations, adding more predictors sets a high bar for the F test. Adding two predictors to a model based on five observations creates a situation where R^2 is likely to be artificially inflated.

96. C is the correct answer. Because of the ratio of predictors to observations, use R_a^2 instead of R^2, dropping the percentage of explained variation by 19%. The relevant formula is

$$R_a^2 = 1 - \left(\frac{n-1}{n-(k+1)}\right)(1 - R^2)$$

```
1-(5-1)/(5-(2+1))*(1-0.81);
             .62
```

97. B is the correct answer. First, distinguish between terms and indicator variables. The indicator variables are $x_1, x_2, \ldots$. The terms are the terms of a linear equation. Each term of a linear equation is what is added together to form the sum. So β_0 is a term by itself and every other combination between a pair of plus signs is a term.

You need enough indicator variables to represent the characteristics. For each characteristic, if the value is set to 1 it means one value, and set to 0 it means the other value. In this case, there are three sizes, times two colors, times two styles, so $3 \times 2 \times 2 = 12$ is the number of cases to be represented. Since each indicator variable can take on only the values $\{0, 1\}$, the number of indicator variables is given by the smallest power of 2 that can represent all the combinations. For example,

$$2^1 = 2, \quad 2^2 = 3, \quad 2^3 = 8, \quad 2^4 = 16$$

The first three powers of two are insufficient to describe the 12 combinations mentioned above. The smallest power of 2 that will suffice is 4, so we need four indicator variables, x_1, x_2, x_3, x_4.

In a simple case like this, we can take advantage of the fact that we can represent the size *large* by setting $x_1 = x_2 = 0$. If $x_1 = 1$ it means the size is small. If $x_2 = 1$ it means the size is medium. We can never have $x_1 = x_2 = 1$ because that would mean that the size is both small and medium, so we will not try to account for that case in the table below. Similarly, we can set $x_3 = 1$ if the color is green and 0 if the color is yellow. Finally, we set $x_4 = 1$ if the style is plain and 0 if it is fancy.

98. D is the correct answer. Read the column with heading `B` in the coefficients table for the estimated coefficients in order, starting with the y-intercept.

99. C is the correct answer. At a minimum, the F statistic requires that you know n and k and either R^2 or any two of the three quantities SSE, SSR, and SS_{yy}, also known as SST. You can use the following relationship to obtain SSE.

$$\begin{aligned} s^2 = MSE &= \frac{SSE}{n-(k+1)} \\ 53350^2 &= \frac{SSE}{1063-(2+1)} \\ 2846222500 &= \frac{SSE}{1063-(2+1)} \\ SSE &= 3016995850000 \end{aligned}$$

Then use that number along with the given value of SSR, bearing in mind that $SSR = SS_{yy} - SSE$.

$$\begin{aligned} F_c &= \frac{(SS_{yy}-SSE)/k}{SSE/[n-(k+1)]} \\ &= \frac{(4561201000000/2}{3016995850000/[1063-(2+1)]} \\ &= 801.27 \end{aligned}$$

```
  /* s^2 */
53350^2;
    2846222500
  /* find SSE  */
solve(2846222500=SSE/1060),numer;
    [SSE = 3016995850000]
  /* find F_c */
(4561201000000/2)/(3016995850000/1060),numer;
    801.272739569728
```

100. C is the correct answer.

The 90 percent confidence interval for the mean of y for a particular $x = x_p$ is the subject of Mendenhall Section 3.9 and requires the formula in the box in the middle of page 129. Don't confuse the *confidence* interval in the middle box with the *prediction* interval in the bottom box.

You must take advantage of the fact that

$$\begin{aligned} s_{\hat{\beta}_1} &= s/\sqrt{SS_{xx}} \\ 0.772 &= \sqrt{296.6}/\sqrt{SS_{xx}} \\ \sqrt{SS_{xx}} &= 17.222/0.772 \\ SS_{xx} &= (0.772/17.222)^2 \\ &= 497.66 \end{aligned}$$

You must also realize that MSE and s^2 are the same thing, the mean square error of the model, the estimate of σ^2 for the model, the subject of Mendenhall, section 4.5.

The rest is just lengthy calculation. We can can calculate

$$\begin{aligned}\hat{y} &= \hat{\beta}_0 + \hat{\beta}_1 x \\ &= -207.617 + 5.605(80) \\ &= 240.783\end{aligned}$$

We can look up $t_{\alpha/2,df} = t_{0.05,60} = 1.671$ in the table on page 758 of Mendenhall. Next, use this knowledge to enter the appropriate information into the expression (from Mendenhall p 129)

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

$$263.203 \pm (1.671)(17.222)\sqrt{\frac{1}{62} + \frac{(80 - 72.9)^2}{497.66}}$$

$$240.783 \pm 9.861359335935934$$

$$(230.14, 249.86)$$

```
1.671*17.222*sqrt((1/62)+((80-72.9)^2/497.66))
  9.861359335935934
```

J. TERMINOLOGY

sum of y values

$$\sum_{i=1}^{n} y_i = y_1 + y_2 + \cdots + y_n$$

synonyms for above $\sum_i y_i$ or $\sum y$

sample mean $\overline{y}$

$$\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n} = \frac{y_1 + y_2 + \cdots + y_n}{n}$$

population mean expected value of y

$$E(y) = \mu$$

range: maximum value minus the minimum

population variance σ^2

sample variance s^2

population standard deviation σ

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \mu)^2}{n}}$$

sample standard deviation s and calculation shortcut

$$s = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}} = \sqrt{\frac{\sum_{i=1}^{n} y_i^2 - n(\bar{y})^2}{n-1}}$$

stdev rules of thumb (if too little info to calc a ci) **(a)** any data set: at least 3/4ths of measurements within $2s$ of $\bar{y}$ **(b)** most data sets w/ 25+ obs and mound-shaped distr, about 95 % of obs within $2s$ of $\bar{y}$

z score standardizes y if $y \sim N(\mu, \sigma)$ meaning that y is drawn from a normally distributed pop with mean μ and stdev σ

$$z = \frac{y - \mu}{\sigma}$$

central limit theorem definition For large sample sizes, the sample mean $\bar{y}$ from a population with mean μ and standard deviation σ has a sampling distribution that is approximately normal, regardless of the probability distribution of the sampled population.

elements of a hypothesis test **(a)** Null hypothesis **(b)** item Alternative hypothesis **(c)** item Test statistic **(d)** Level of significance **(e)** Rejection region boundary **(f)** p-value **(g)** Conclusion

standard error of an estimate (not stdev)

$$\sigma_{\bar{y}} = \frac{\sigma}{\sqrt{n}}$$

large-sample $100(1-\alpha)\%$ conf interval for μ

$$\bar{y} \pm z_{\alpha/2}\sigma_{\bar{y}} \approx \bar{y} \pm z_{\alpha/2}\frac{s}{\sqrt{n}}$$

small-sample $100(1-\alpha)\%$ conf interval for μ

$$\bar{y} \pm t_{\alpha/2,n-1}s_{\bar{y}} \approx \bar{y} \pm t_{\alpha/2,n-1}\frac{s}{\sqrt{n}}$$

test statistic for hypothesis regarding μ

$$t = \frac{\bar{y} - \mu_0}{s/\sqrt{n}}$$

α selection driven by criticality: $\alpha = 0.01$ cases where life could be threatened by failure $\alpha = 0.05$ cases where money riding on outcome $\alpha = 0.10$ cases where consequences of failing to capture true mean not severe

$1-\alpha$	α	$\alpha/2$	$z_{\alpha/2}$
.90	.10	.05	1.645
.95	.05	.025	1.96
.99	.01	.005	2.576

Type I error imprisoning an innocent person is like Type I error: we rejected the null hypothesis when it was true.

Type II error letting a guilty person go free is like Type II error: we failed to reject the null hypothesis when it was false.

elements of hypothesis test, e.g. hw 1.61 **1.** null hypothesis, e.g., $H_0 : \mu = 15$ **2.** alternative hypothesis, e.g., $H_a : \mu > 15$ **3.** test statistic, e.g., $t_c = 7.8264$ **4.** level of significance, e.g., $\alpha = 0.01$ **5.** rejection region marker, e.g., $t_{0.01,4} = 3.746947$ **6.** p-value, e.g., $P(t > t_c) = 0.0007194883$ **7.** conclusion, e.g., Reject the null hypothesis that these five bags came from a machine that dispenses an average of 15 candies per bag.

simplest linear regression model

$$E(y) = \beta_0 + \beta_1 x + \varepsilon$$

straight line used to predict y given x with **(a)** y-intercept β_0 **(b)** slope β_1 **(c)** error or residual ε (Greek letter epsilon, not Latin letter e)

estimate of the model

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

least-squares line criteria **1.** $SE = 0$ sum of errors is zero **2.** SSE, the sum of squared errors, smaller than for any other line that meets criteria 1

sum of errors

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

sum of squares of errors also known as sum of squares of residuals

$$\begin{aligned} SSE &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= SS_{yy} - \hat{\beta}_1 SS_{xy} \end{aligned}$$

estimator of σ^2, which is the variance of ε

$$s^2 = MSE = \frac{SSE}{n - (k+1)}$$

est std error of the model

$$s = \sqrt{s^2} = \sqrt{MSE} = \sqrt{\frac{SSE}{n - (k+1)}}$$

warning

$$\sum (p_i - \overline{p})^2 \neq \sum (p_i)^2 - (\overline{p})^2$$

slope of least squares line that satisfies both above criteria

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}}$$

y-intercept of least squares line that satisfies both above criteria

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

sums of squares with one predictor since these formulas use x values explicitly, they are only used for models with one predictor—for multiple regression, use the formulas in the next section

	concept version	*calculation shortcut*
sum of squares of deviations of x from $\overline{x}$	$SS_{xx} = \sum (x_i - \overline{x})^2$	$= \sum x_i^2 - n\overline{x}^2$
sum of squares of deviations of y from $\overline{y}$	$SS_{yy} = \sum (y_i - \overline{y})^2$	$= \sum y_i^2 - n\overline{y}^2$
sum of product of deviations of x from $\overline{x}$ and deviations of y from $\overline{y}$	$SS_{xy} = \sum (x_i - \overline{x})(y_i - \overline{y})$	$= \sum x_i y_i - n\overline{xy}$

est std err of a β coefficient (Mendenhall)

$$s_{\hat{\beta}_1} = \frac{s}{\sqrt{SS_{xx}}}$$

where s is the estimated std error of the model, not the stdev of a single variable from its mean

t-statistic for a β coefficient using the above est std error

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

where the multiple regression model will have subscripts called i instead of 1 and and SS_{xx} will vary from x_1 to x_k

calc t_α using RStudio

```
qt(0.05,60,lower.tail=FALSE)
```

where $\alpha = 0.05, \nu = 60$.

calc p-value knowing t_α using RStudio

```
pt(1.67,60,lower.tail=FALSE)
```

where $t_\alpha = 1.67, v = 60$.

p-value for a β coefficient is known as Sig. in SPSS

rejection region boundary $t_{\alpha,v}$ distinguish it from t_c calculated from sample data

degrees of freedom $v = df = n - 2$ for single variable regression $n - (k + 1)$ for multiple regression

coefficient of correlation

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

above formula for r is only good for correlation between two variables, x and y — for correlation between one y and multiple x values, use square root of following formula, for pairwise correlations, add a subscript, e.g., r_{y,x_2} to be specific about which SS_{xx} is being used

coefficient of determination

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

100(1 − α)% CI for mean of y for $x = x_p$, $s = \sqrt{MSE}$

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

100(1 − α)% PI for individual y for $x = x_p$, $s = \sqrt{MSE}$

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

multiple regression model with $k = 4$

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$$

estimate of the model

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x + \hat{\beta}_2 x + \hat{\beta}_3 x + \hat{\beta}_4 x$$

multiple regression sums of squares

total sum of squares for the model, equivalent to the sum of squares of deviations of y from $\bar{y}$

$$SST = SS_{yy} = \sum (y_i - \bar{y})^2$$

sum of squares of regression SS(Model) in Mendenhall is equivalent to SSR in most stat software and refers to the square of the differences between $\hat{y}$ and $\bar{y}$, which is the variation in the sample explained by the model

$$SSR = SS_{yy} - SSE = \sum (\hat{y}_i - \bar{y})^2$$

relationship of sums of squares

$$SS_{yy} = SST = SSR + SSE$$

test overall model utility / adequacy

null hypothesis $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$

alternative hypothesis H_a : at least one coefficient is nonzero

test statistic for overall model utility

$$F_c = \frac{(SS_{yy} - SSE)/k}{SSE/[n-(k+1)]} = \frac{R^2/k}{(1-R^2)/[n-(k+1)]}$$

level of significance α

rejection region boundary $F_c > F_\alpha$ where F_α has k numerator degrees of freedom and $n-(k+1)$ denominator degrees of freedom

calc F_α using RStudio

```
qf(0.05,6,12,lower.tail=FALSE)
```

where $\alpha = 0.05, v_1 = 6, v_2 = 12$.

calc p-value knowing F_α using RStudio

```
pf(2.99612,6,12,lower.tail=FALSE)
```

where $F_\alpha = 2.99612, v_1 = 6, v_2 = 12$.

p-value rejection region boundary above $\Leftrightarrow \alpha > p$-value, where p-value is $P(F_\alpha > F_c)$ and $_c$ means computed value

$100(1-\alpha)\%$ conf interval for β_i

$$\hat{\beta}_i \pm t_{\alpha/2,n-(k+1)} s_{\hat{\beta}_i}$$

multiple coefficient of determination same as r^2 above but can not be calculated by squaring the r obtained by the formula using SS_{xx} when there are multiple x variables

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

adjusted multiple coefficient of determination

$$R_a^2 = 1 - \left(\frac{n-1}{n-(k+1)}\right)(1-R^2)$$

Use R_a^2 in preference to R^2 as the proportion of variability explained by the model if $R^2 - R_a^2 > 0.1$

nested models run two regression analyses where one is a subset of the other and compare the values of SSE

reduced model

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g + \varepsilon$$

complete model

$$E(y) = \beta_0 + \beta_1 x_1 + \cdots + \beta_g x_g \beta_{g+1} x_{g+1} + \cdots + \beta_k x_k + \varepsilon$$

null hypothesis for nested models

$$H_0 : \beta_{g+1} = \beta_{g+2} = \cdots = \beta_k = 0$$

alternative hypothesis for nested models

H_a : at least one of $\beta_{g+1} \cdots \beta_k \neq 0$

test statistic for nested models

$$F_c = \frac{(SSE_R - SSE_C)/(k-g)}{SSE_C/[n-(k+1)]}$$

simplify stacked fractions

$$\frac{a/b}{c/d} = \frac{a \cdot d}{b \cdot c}$$

degrees of freedom for nested models numerator $k - g$ where k is the total number of parameters excluding β_0 in the complete model and g is the number of parameters excluding β_0 in the reduced model

indicator term *i* introduces a new parameter only for cases falling into some specific category

$$x_i = \begin{cases} 1 & \text{if a given variable falls into a specific category} \\ 0 & \text{otherwise} \end{cases}$$

multicollinearity

$$VIF_i = \frac{1}{1 - R_i^2}, i = 1, 2, \ldots, k$$

K. TABLES

Critical Values of the Standard Normal Distribution, $z \sim N(0, 1)$

z	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0753
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2257	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2517	0.2549
0.7	0.2580	0.2611	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2995	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4515	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4893	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000

Critical Values of the Student's t Distribution

ν	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$	$t_{.001}$
1	3.078	6.314	12.706	31.821	63.657	318.31
2	1.886	2.920	4.303	6.965	9.925	22.327
3	1.638	2.353	3.182	4.541	5.841	10.215
4	1.533	2.132	2.776	3.747	4.604	7.173
5	1.476	2.015	2.571	3.365	4.032	5.893
6	1.440	1.943	2.447	3.143	3.707	5.208
7	1.415	1.895	2.365	2.998	3.499	4.785
8	1.397	1.860	2.306	2.896	3.355	4.501
9	1.383	1.833	2.262	2.821	3.250	4.297
10	1.372	1.812	2.228	2.764	3.169	4.144
11	1.363	1.796	2.201	2.718	3.106	4.025
12	1.356	1.782	2.179	2.681	3.055	3.930
13	1.350	1.771	2.160	2.650	3.012	3.852
14	1.345	1.761	2.145	2.624	2.977	3.787
15	1.341	1.753	2.131	2.602	2.947	3.733
16	1.337	1.746	2.120	2.583	2.921	3.686
17	1.333	1.740	2.110	2.567	2.898	3.646
18	1.330	1.734	2.101	2.552	2.878	3.610
19	1.328	1.729	2.093	2.539	2.861	3.579
20	1.325	1.725	2.086	2.528	2.845	3.552
21	1.323	1.721	2.080	2.518	2.831	3.527
22	1.321	1.717	2.074	2.508	2.819	3.505
23	1.319	1.714	2.069	2.500	2.807	3.485
24	1.318	1.711	2.064	2.492	2.797	3.467
25	1.316	1.708	2.060	2.485	2.787	3.450
26	1.315	1.706	2.056	2.479	2.779	3.435
27	1.314	1.703	2.052	2.473	2.771	3.421
28	1.313	1.701	2.048	2.467	2.763	3.408
29	1.311	1.699	2.045	2.462	2.756	3.396
30	1.310	1.697	2.042	2.457	2.750	3.385
35	1.306	1.690	2.030	2.438	2.724	3.340
40	1.303	1.684	2.021	2.423	2.704	3.307
45	1.301	1.679	2.014	2.412	2.690	3.281
50	1.299	1.676	2.009	2.403	2.678	3.261
55	1.297	1.673	2.004	2.396	2.668	3.245
60	1.296	1.671	2.000	2.390	2.660	3.232
∞	1.282	1.645	1.960	2.326	2.576	3.090

Critical Values of the F Distribution, $\alpha = 0.05$

$\nu_2 \backslash \nu_1$	1	2	3	4	5	6	7	8	10	12	15	20	30	50	∞
1	161.45	199.5	216	225	230	234	237	239	242	244	246	248	250	252	254
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.79	8.74	8.70	8.66	8.62	8.58	8.53
4	7.70	6.94	6.59	6.39	6.26	6.16	6.09	6.04	5.96	5.91	5.86	5.80	5.75	5.70	5.63
5	6.60	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.74	4.68	4.62	4.56	4.50	4.44	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.06	4.00	3.94	3.87	3.81	3.75	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.64	3.57	3.51	3.44	3.38	3.32	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.35	3.28	3.22	3.15	3.08	3.02	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.14	3.07	3.01	2.94	2.86	2.80	2.71
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	2.98	2.91	2.84	2.77	2.70	2.64	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.85	2.79	2.72	2.65	2.57	2.51	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.75	2.69	2.62	2.54	2.47	2.40	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.67	2.60	2.53	2.46	2.38	2.31	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.60	2.53	2.46	2.39	2.31	2.24	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.54	2.48	2.40	2.33	2.25	2.18	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.49	2.42	2.35	2.28	2.19	2.12	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.45	2.38	2.31	2.23	2.15	2.08	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.41	2.34	2.27	2.19	2.11	2.04	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.38	2.31	2.23	2.16	2.07	2.00	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.35	2.28	2.20	2.12	2.04	1.97	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.32	2.25	2.18	2.10	2.01	1.94	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.30	2.23	2.15	2.07	1.98	1.91	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.27	2.20	2.13	2.05	1.96	1.88	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.25	2.18	2.11	2.03	1.94	1.86	1.73
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.24	2.16	2.09	2.01	1.92	1.84	1.71
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.22	2.15	2.07	1.99	1.90	1.82	1.69
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.20	2.13	2.06	1.97	1.88	1.81	1.67
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.19	2.12	2.04	1.96	1.87	1.79	1.65
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.18	2.10	2.03	1.94	1.85	1.77	1.64
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.16	2.09	2.01	1.93	1.84	1.76	1.62
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	1.99	1.92	1.84	1.75	1.65	1.56	1.39
80	3.96	3.11	2.72	2.49	2.33	2.21	2.13	2.06	1.95	1.88	1.79	1.70	1.60	1.51	1.32
100	3.94	3.09	2.70	2.46	2.31	2.19	2.10	2.03	1.93	1.85	1.77	1.68	1.57	1.48	1.28
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.91	1.83	1.75	1.66	1.55	1.46	1.25
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.83	1.75	1.67	1.57	1.46	1.35	1.00