From IPS 3 e

2.40

Joan is concerned about the amount of energy she uses to heat her home in the Midwest. She keeps a record of the natural gas she consumes each month over one year's heating season. Because the months are not all the same length, she divides each month's consumption by the number of days in the month to get the average number of cubic feet of gas used per day. Demand for heating is strongly influenced by the outside temperature. From local weather records, Joan obtains the average number of heating degree-days per day for each month. (One heating degree-day is accumulated for each degree a day's average temperature falls below 65°F. An average temperature of 20°F, for example, corresponds to 45 degree-days.) Here are Joan's data (provided by Robert Dale, Purdue University):

Month	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
Degree-days	15.6	26.8	37.8	36.4	35.5	18.6	15.3	7.9	0.0
Gas consumed	520	610	870	850	880	490	450	250	110

(a) Make a scatterplot of these data. There is a strongly linear pattern with no outliers.

(b) Find the equation of the least-squares regression line for predicting gas use from degree-days. Draw this line on your graph. Explain in simple language what the slope of the regression line tells us about how gas use responds to outdoor temperature.

(c) Joan adds insulation in her attic during the summer, hoping to reduce her gas consumption. The next February, there are an average of 40 degree-days per day and her gas consumption is 870 cubic feet per day. Predict from the regression equation how much gas the house would have used at 40 degree-days per day last winter before the extra insulation. Did the insulation reduce gas consumption?

18.6

Utility companies need to estimate the amount of energy that will be used by their customers. The consumption of natural gas required for heating homes depends on the outdoor temperature. When the weather is cold, more gas will be consumed. A study of one home recorded the average daily gas consumption y (in hundreds of cubic feet) for each month during one heating season. The explanatory variable x is the average number of heating degree-days per day during the month. One heating degree-day is accumulated for each degree a day's average temperature falls below 65° F. An average temperature of 50°, for example, corresponds to 15 degree-days. The data for October through June are given in the following table. (Data provided by Professor Robert Dale of Purdue University.)

	Oct.	Nov.	Dec.	Jan.	Feb.	Mar.	Apr.	May	June
x	15.6	26.8	37.8	36.4	35.5	18.6	15.3	7.9	0.0
y	5.2	6.1	8.7	8.5	8.8	4.9	4.5	2.5	1.1

(a) Find the equation of the least-squares line.

(b) Test the null hypothesis that the slope is zero and describe your conclusion.

(c) Give a 95% confidence interval for the slope.

(d) The parameter β_0 corresponds to natural gas consumption for cooking, hot water, and other uses when there is no demand for heating. Give a 95% confidence interval for this parameter.

From IPS 3 e

10.11

Exercise 10.6 (page 697) demonstrates that there is a strong linear relationship between household consumption of natural gas and outdoor temperature, measured by heating degree-days. The slope and intercept depend on the particular house and on the habits of the household living there. Data for two heating seasons (18 months) for another household produce the least-squares line $\hat{y} = 2.405 + 0.26896x$ for predicting average daily gas consumption y from average degree-days per day x. The standard error of the slope is $\mathrm{SE}_{b_1} = 0.00815$.

(a) Explain briefly what the slope β_1 of the population regression line represents. Then give a 95% confidence interval for β_1.

(b) This interval is based on twice as many observations as the one calculated in Exercise 10.6 for a different household, and the two standard errors are of similar size. How would you expect the margins of error of the two intervals to be related? Check your answer by comparing the two margins of error.

10.12

The standard error of the intercept in the regression of gas consumption on degree-days for the household in the previous exercise is $\mathrm{SE}_{b_0} = 0.20351$.

(a) Explain briefly what the intercept represents in this setting. Find a 95% confidence interval for the intercept.

(b) Compare the width of your interval with the one calculated for a different household in Exercise 10.6. Explain why it is shorter.

10.13

Exercise 10.6 gives information about the regression of natural gas consumption on degree-days for a particular household.

(a) What is the t statistic for testing H_0: $\beta_1 = 0$?

(b) For the alternative H_a: $\beta_1 > 0$, what critical value would you use for a test at the $\alpha = 0.05$ significance level? Do you reject H_0 at this level?

(c) How would you report the P-value for this test?

2.43

In Professor Smith's economics course the correlation between the students' total scores before the final examination and their final examination scores is $r = 0.6$. The pre-exam totals for all students in the course have mean 280 and standard deviation 30. The final exam scores have mean 75 and standard deviation 8. Professor Smith has lost Julie's final exam but knows that her total before the exam was 300. He decides to predict her final exam score from her pre-exam total.

(a) What is the slope of the least-squares regression line of final exam scores on pre-exam total scores in this course? What is the intercept?

(b) Use the regression line to predict Julie's final exam score.

(c) Julie doesn't think this method accurately predicts how well she did on the final exam. Calculate r^2 and use the value you get to argue that her actual score could have been much higher or much lower than the predicted value.

2.46

The mean height of American women in their early twenties is about 64.5 inches and the standard deviation is about 2.5 inches. The mean height of men the same age is about 68.5 inches, with standard deviation about 2.7 inches. If the correlation between the heights of husbands and wives is about $r = 0.5$, what is the equation of the regression line of the husband's height on the wife's height in young couples? Draw a graph of this regression line. Predict the height of the husband of a woman who is 67 inches tall.

Homegrown Exercises around M&M Chapter 10

If proceeding to Course 621 in Jan 2002, use SAS (INSIGHT or PROC REG) or SPSS for the following analyses. If not, use software/calculator of your choice).

To set up the SAS dataset, you can either (i) download the already created sas file directly to your sasuser directory or (ii) or download the sas program that contains the data, bring it into the Program Editor, and run the data step from there to create it. No matter which route you take, you can then perform analyses (a) from INSIGHT or (b) by running PROCedure steps from within the program Editor. To get help on the syntax for a procedure (e.g. the CORR procedure), type HELP CORR in the command box)

-1- 1970 Draft Lottery (data in Excel sheet under Resources for Ch 10)

Run the "new" lottery 30 times (use F9 key) and make a stem & leaf plot (by hand, on paper, is sufficient) of the 30 (a) correlations (b) p-values. Comment on the shapes and ranges of the distributions of (a) and (b), and on how many of the 30 correlations exceed the magnitude of the one observed in the 1970 lottery.

~~**-2- Correlations in twins of bone density measures at different anatomical sites** (data under Resources for Ch 10)~~ **OMIT this Q**

Determine the point and (by hand) the 95% interval estimate for the correlation of the heights of dizygotic twin pairs (Interpolation using the CI nomogram, rather than calculating it from the formula, is sufficient. Despite M&M's comment at the bottom of p690 of IPS4e, the calculation isn't *that* "tedious" — they could easily have provided a nomogram or formula)

Are the correlations of the heights of dizygotic and monozygotic twin pairs significantly different from each other at the alpha=0.05 level (2-sided)? Use a direct test on the difference, rather than comparing for overlap of the 2 CI's [cf notes Ch 2]

Determine the correlations — for dizygotic twin 1—between (a) tea and coffee consumption (b) the bone density at the 3 sites (spine, and 2 femoral) within the same twin. Interpret these coefficients in words. (If you have time, check whether the same patterns hold up for twin2, and for twin1 and twin2 in the monozygotic pairs)

-3- Correlations: heights of parents (100 from Galton's dataset) (data under Resources for Ch 10)

(i) Determine the (Pearson) correlation, and its associated 95% CI, between fathers' and mothers' heights. (ii) Contrast this with M&M's assumed modern day value for persons in their early twenties (cf q 2.46 in IPS3e or q2.49 in IPS4e). (iii) Likewise, contrast the Galton and modern day means and s.d.'s. (iv) Predict the height of a husband in Galton's dataset, married to a woman reported to be 67 inches tall. Why the big difference between the answer for then and now? [If interested, see the 1902 article by Pearson and Lee, under Resources for Ch 10]

-4- Variability of, and trends in, proportions (SAS program and data, and data in an Excel sheet, are available under Resources for Ch 10, or you can use the two-variable calculator—available via link in Resources for Ch 10)

Refer again to the data on the proportion of Canadian adults responding YES to the question "Have you yourself smoked any cigarettes in the past week?" in Gallup Polls for the years 1974 to 1985.

a Fit a linear regression to these data (regress Rate on Year).

b Identify and interpret the 2 regression coefficients (parameter estimates)

c Calculate a 95% CI to accompany each coefficient. *[Can use respective SE's, together with appropriate t value from the $t_{(n-2)df}$ table, to construct them]*

d Regress Rate on (Year minus1974) *[a new variable already set up in SAS program... this new variable would also be easy to create "after the fact" within INSIGHT: EditMenu->Variables->Other... Apply the a+bY transformation to "Y"=Year, using a= –1974 and b=1. Notice that the use of the names "X" and "Y" within the "Edit Variables" dialog box bears no relation to "X" and "Y" used in the regression. The transformation will be applied to whatever you designate as "X" and "Y", but this designation is local and is forgotten once the variables are created.]*

e Identify and interpret the 2 coefficients of this new equation.

f Compare the coefficients of the new fitted equation with those you obtained under the original equation. From this, state a general rule about the effect on the regression coefficients of 'shifting' the X Variable.

g Why do you think the SE for the intercept is much smaller under the new formulation? Why hasn't the SE for the slope (the coefficient of Year or "Year minus 1974") changed from one formulation to the other?

h Identify and interpret a measure of residual variation from the fitted line (Since the "y" variable is on a percentage scale, make sure you measure the residual variation in this same scale)

Note that this measure of residual variation (which is a mix of sampling variation and any inaccuracies in specifying the form of the curve) does not use the n's (1050 or so) from which these proportions were estimated, or their stated margins of error. Yet it comes close to the value on which the stated margins of error are based.