Course manual Statistics II, SSC3018

I. Course Information

II. Course Material: the Weekly program

1.1 How's life: Introduction to Statistics II

Statistics has the reputation to be an uninteresting, calculus oriented academic course. And statistics (data) producing organizations like the OECD share that reputation: producers of 'cold numbers of GDP and economic statistics' (in their own words), without having an open eye for 'what matters most to people around the world, for well-being, material living conditions and quality of life' (again, their own words). In order to improve that image, the OECD started the 'OECD Better Life Initiative' two years ago, with the goal to develop statistics to capture aspects of life that matter to people and shape the quality of their lives. You will find its website at: http://www.oecdbetterlifeindex.org/. And beyond providing you with this new type of data, the OECD even calls for bringing in new ideas to improve quality of life: 'Directed at young researchers and PhD-level students, we are looking for fresh ideas that challenge our thinking on how government policies in the areas of competition, corporate governance, capital markets and financial services, international investment and foreign bribery can have an impact on our well-being as defined by the OECD's Better Life Initiative.' (http://www.oecd.org/statistics/better-life-initiative.htm). To contribute to this noble goal of OECD (and to improve our own reputation), we will spend the next seven weeks in analysing this new OECD index data, doing our own methodological research in well-being, trying to develop models that help us better understand the factors that influence quality of life.

Beyond this adventurous focus, the course Statistics II provides an advanced introduction to research methods commonly used in social sciences and humanities. Emphasis is on issues of inferential statistics, regression modelling, multivariate statistics and on computing skills needed to apply these statistical tools.
In Statistics II, we resume the thread of Statistics I: a discussion of the basic tools of inferential statistics: confidence intervals and hypothesis tests (which in turn involved concepts like null and alternative hypotheses, Type I and Type II errors, rejection points and p-values), all these concepts illustrated in the context of the one-sample tests. In this course, you will encounter a whole battery of additional tests, enabling you to examine a large array of questions that may occur in social sciences. In the first weeks, we discuss amongst others the two-sample t-test (allowing you to compare the mean of a quantitative variable between two populations), oneway-ANOVA (ditto, for more than two populations), the paired-sample t-test and the chi-square test (allowing you to establish relationships between qualitative variables, using contingency tables). But the main dish of that course is obviously regression analysis, a very flexible technique which allows you to relate a dependent variable to a number of independent or explanatory variables.
In Statistics II we will use SPSS rather than applets or EXCEL, the software packages that were used in Statistics I. SPSS is a leading statistical package in social sciences, widely used in academia and in professional practice (e.g. in marketing research). Another difference with Statistics I is the strong focus on actively applying our statistical tools, using SPSS, to solve case studies based on real-life datasets.
All information in this course manual as well as a lot of additional information can be found in our electronic learning environment "Eleum". Among the features you will find online will be computer support, hints for doing the case studies, copies of trial exams and announcements regarding your group and the entire course. It is our aim to update the site frequently, and we strongly advise you to check the site regularly. Ideally, you should check the site daily, but minimally you should drop at least once a week.

1.2 Goals Statistics II

The course Statistics II is the successor of Statistics I. Succeeding Statistics I has several dimensions. First of all, to be able to participate in Statistics II, students should have passed Statistics I, since Statistics I constitutes the required prior knowledge of Statistics II. Second, in the design of Statistics II a similar focus is chosen as in Statistics I: a focus on getting familiarized with quantitative research methods. This implies that you will develop the abilities to read, understand and criticize scientific articles in the domain of your concentration, as a passive use of your knowledge of quantitative techniques. On top of that, you will gain

experience in actively performing such a quantitative analysis yourself, making use of the (more advanced features of the) tool SPSS.

1.3 Structure Statistics II

We have two different kinds of tutorial sessions: the first session in the week, taking place in a regular TG room, and the second session, taking place in the computer lab. Those physical differences represent different educational goals. Every week, a new statistical tool will be covered. We start with reading on this new tool with the help of the textbook, and discussing its properties in the first tutorial session, using some discussion exercises from the textbook. The lecture should finalise that 'begin of the week new learning stage'. That is followed by an application stage in the second part of the week. You will work on a case study, based on the 'OECD Better-Life project' data set, in each of the six weeks, and in your overarching student project. Doing so will raise you skills to actively use statistics, beyond the passively using it.

1.3.1 Literature

We strongly recommend that you turn to the weekly literature before the first session of the week. You best start up the weekly learning cycle by reading the textbook chapters as the preparation of the first tutorial session.

1.3.2 The computer lab/SPSS session

This session is reserved for working on the six weekly case studies. A short explanation on the statistical tools you need to use in order to solve the case studies will be provided in the weekly lectures. In the lab session itself, support is provided in solving the case studies.

1.3.3 The tutorial group session

The weekly program, contained in the second section of the course manual, counts several 'discussion tasks': these are all exercises from the textbook (nearly always even numbered exercises), you do not need to prepare, but they are used as discussion material. We will discuss them, along with problems you might encounter when reading the text or preparing the uneven numbered exercises: exercises we leave to you for solving as part of your preparation of the first tutorial session.

1.4 Position in Curriculum

Statistics II is the 3000 level course in SS preparing for quantitative research. The course prepares for other, more scientific oriented courses in your Bachelor, and will be a prerequisite for any Master study Social Sciences.

1.5 Grading policies Statistics II

To determine your Statistics II grade, a portfolio of different assessment instruments with different weights will be used, as described in the table:

Assessment instruments	Weight in grading
Six case studies & reviews	30%
Final exam	50%
Student project	20%
Total grade	100%

1.5.1 Six case studies and six case study reviews

The applied part of the course consists of six weekly case studies in which you do an application of the statistical tools studied the same week, using the OECD data set. For doing these case studies, the end of the week lab sessions are reserved: if you work efficiently (and are well-prepared), the two hour lab sessions will offer sufficient time and support to finish the case studies in class. The outcome of your case study is a statistical report. These reports will contain 6-8 pages, mostly filled with statistical output copy/pasted from SPSS (tables, graphs); your own text with interpretations of the statistical output is at least 2 pages. You will need to upload your case study report at the end of the week (deadline Sunday, 24.00). The second step in doing the case studies includes reviewing two other case study reports, and providing short feedback (deadline: Wednesday, 24.00). Both steps are graded: every case study counts for max 5 credit points, max 4 reserved for your own case study report, max 1 for your two reviews (except for case study Week6: in this last case study, there is no feedback round, and all 5 credits are for the study report). The StudentPortal will provide you with detailed information on how to upload your case reports and reviews: you will need to upload digital versions only, no paper versions. The StudentPortal will also contain exact deadlines: at the time of writing this course manual, it is not clear how the many holidays will impact the sequence of our sessions.

1.5.2 Final exam and Resit

The final exam consists of two parts. In the first part, you will be supplied with several short problem descriptions, and are asked to identify the statistical techniques that are appropriate in solving those problems. The second part consists of a combination of multiple choice and short answers questions. The exam counts for max 50 credit points.

This year final exam may be organized as a digital exam. If that is the case, the second part of the exam is somewhat similar to your weekly case studies: you receive some data and a problem description, are supposed to solve that problem using SPSS, and as your exam answers provide relevant pieces of SPSS output, together with your interpretations. If the final exam is a more traditional pencil and paper type of exam, you will be provided with SPSS output, and the exam questions are about the interpretation of this output. You will be informed about this issue later in the course.

Students who fail the course in the first sit, but meet the attendance requirement, or are allowed to make up for it by means of an additional assignment, can resit the final exam, and/or the final case study. If in the resit of the final exam you obtain a score that is higher than your score for the six case studies, that higher score will overwrite your previous case studies score.

1.5.3 Project study

In week 8, you have to hand in your project study through StudentPortal. Reports should contain at least 5 pages written text (and many more pages statistical output). The project study is rewarded for max 20 credit points.

1.5.4 Attendance & participation requirement

For the 'begin of the week' tutorial group sessions (weekly, so in total: 6), the compulsory attendance requirement is 85%; that is 5 meetings (so you can miss at most 1). Students who have not met the attendance requirement, but who have not missed more than 30% of the group meetings, will be given a provisional overall grade, but will not receive credits for the course until they have successfully completed an additional assignment.

1.6 Schedule

See MyUM.

1.7 Literature Statistics II

The literature for Statistics II is the 3rd edition of the De Veaux book:

- ISBN 9780321753724: STATS DATA & MODELS, 3rd Edition, by DEVEAUX, VELLEMAN, & BOCK

There are a lot of copies around, given the large number of students who took the course in last years, so it should not be problematic to borrow/buy 2nd hand (also the 2nd edition will do). Additional readers will be provided through Eleum, and can be found in this course manual.

1.8 Using SPSS

The default UM SPSS version is 24: however, in terms of the functionality that we need for Stats 2 and the "looks and feel" of the different windows and menus, older versions are just as adequate. There are three ways to do so:

SPSS is installed on the PC's in the Computing Rooms of UCM.

- SPSS is installed (under Citrix) on the PC's in the university library.
- For home use, you can order a CD-rom via www.surfspot.nl. Delivery by mail will take two working days (after payment) and is possible only to addresses within the Netherlands.

If you prefer to work at home rather than in the library, we urge you to buy and install SPSS as soon as possible, preferably even before the start of the course, in order to pre-empt any technical difficulties.

1.9 Planning Group Statistics II

The planning group consists of: dr. Dirk Tempelaar, Tongersestraat 53, Room A2.20, tel. 043-3883858, e-mail: **D.Tempelaar@MaastrichtUniversity.nl**

II. Course Material: weekly program

II.1: Program week 1: Statistical Inference for one or two variables

Literature: De Veaux, Velleman, & Bock, chapters 19, 20, 21, 22, 23, 24, 25.

This is a lot of reading, but: all these chapters are covered in Statistics 1, so it is in fact rehearsing what you did in the previous Statistics course. It is however crucial that you master these topics: the topics of later weeks build on this. And the topics are included in the final exam.

Discussion tasks (for discussion in the tutorial class session, 1st session of the week):

Chapter 23: exercises 4, 6, 8, 10, 12, 22, 24, 26, 28, 40.
Chapter 24: exercises 2, 4, 6, 8, 12, 16, 22, 24, 26.
Chapter 25: exercises 2, 4, 6, 10, 12

SPSS Case Study I: we will do this case study in the lab session, the 2nd session of the week. Finish outside, if you cannot finish within the lab time. The case study assignment can be found in the StudentPortal.

Discussion tasks:

Ex23.4. t-models, part IV (last one!). Describe how the critical value of t for a 95% confidence interval changes as the number of degrees of freedom increases.

Ex23.6. Teachers. Software analysis of the salaries of a random sample of 288 Nevada teachers produced the confidence interval shown below. Which conclusion is correct? What's wrong with the others?
t-Interval for μ: with 90.00% Confidence, 38944 < μ(TchPay) < 42893
a) If we took many random samples of 288 Nevada teachers, about 9 out of 10 of them would produce this confidence interval.
b) If we took many random samples of Nevada teachers, about 9 out of 10 of them would produce a confidence interval that contained the mean salary of all Nevada teachers.
c) About 9 out of 10 Nevada teachers earn between $38,944 and $42,893.
d) About 9 out of 10 of the teachers surveyed earn between $38,944 and $42,893.
e) We are 90% confident that the average teacher salary in the United States is between $38,944 and $42,893.

Ex23.8. Snow. Based on meteorological data for the past century, a local TV weather forecaster estimates that the region's average winter snowfall is 23"with a margin of error of ±2 inches. Assuming he used a 95% confidence interval, how should viewers interpret this news? Comment on each of these statements:
a) During 95 of the last 100 winters, the region got between 21"and 25"of snow.
b) There's a 95% chance the region will get between 21"and 25"of snow this winter.
c) There will be between 21"and 25"of snow on the ground for 95% of the winter days.
d) Residents can be 95% sure that the area's average snowfall is between 21"and 25".
e) Residents can be 95% confident that the average snowfall during the last century was between 21"and 25" per winter.

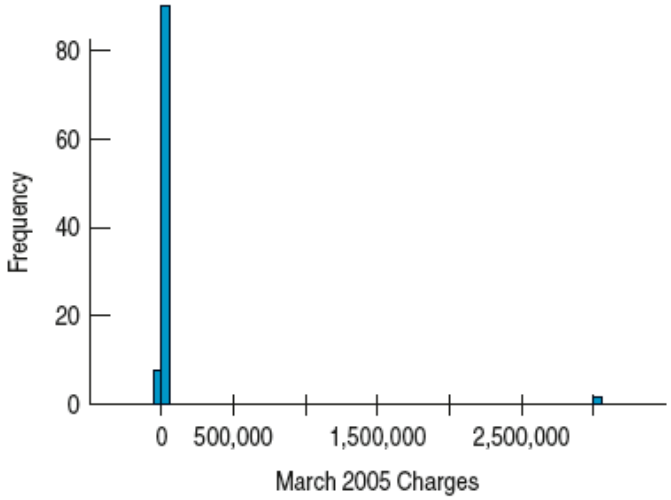
Ex23.10. Crawling. Data collected by child development scientists produced this confidence interval for the average age (in weeks) at which babies begin to crawl:

```
t-Interval for μ (95.00% Confidence): 29.202 < μ(age) < 31.844
```

a) Explain carefully what the software output means.
b) What is the margin of error for this interval?

c) If the researcher had calculated a 90% confidence interval, would the margin of error be larger or smaller? Explain.

Ex23.12. Credit card charges. A credit card company takes a random sample of 100 cardholders to see how much they charged on their card last month. Here's a histogram.



A computer program found that the resulting 95% confidence interval for the mean amount spent in March 2005 is (-$28366.84, $90691.49). Explain why the analysts didn't find the confidence interval useful, and explain what went wrong.

Ex23.22. Hot dogs. A nutrition lab tested 40 hot dogs to see if their mean sodium content was less than the 325-mg upper limit set by regulations for "reduced sodium" franks. The lab failed to reject the hypothesis that the hot dogs did not meet this requirement, with a P-value of 0.142. A 90% confidence interval estimated the mean sodium content for this kind of hot dog at 317.2 to 326.8 mg. Explain how these two results are consistent. Your explanation should discuss the confidence level, the P-value, and the decision.

Ex23.24. Golf balls. The United States Golf Association (USGA) sets performance standards for golf balls. For example, the initial velocity of the ball may not exceed 250 feet per second when measured by an apparatus approved by the USGA. Suppose a manufacturer introduces a new kind of ball and provides a sample for testing. Based on the mean speed in the test, the USGA comes up with a P-value of 0.34. Explain in this context what the "34%" represents.

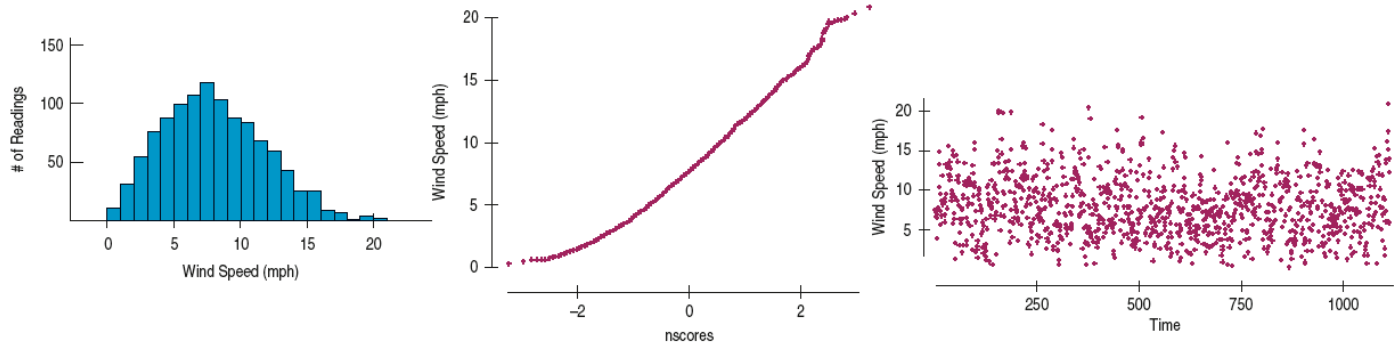
Ex23.26. Catheters. During an angiogram, heart problems can be examined via a small tube (a catheter) threaded into the heart from a vein in the patient's leg. It's important that the company that manufactures the catheter maintain a diameter of 2.00 mm. (The standard deviation is quite small.) Each day, quality control personnel make several measurements to test H_0: $\mu = 2.00$ against H_A: $\mu \neq 2.00$ at a significance level of $\alpha = 0.05$. If they discover a problem, they will stop the manufacturing process until it is corrected.
a) Is this a one-sided or two-sided test? In the context of the problem, why do you think this is important?
b) Explain in this context what happens if the quality control people commit a Type I error.
c) Explain in this context what happens if the quality control people commit a Type II error.

Ex23.28. Catheters, again. The catheter company in Exercise 26 is reviewing its testing procedure.
a) Suppose the significance level is changed to $\alpha = 0.01$. Will the probability of a Type II error increase, decrease, or remain the same?
b) What is meant by the power of the test the company conducts?
c) Suppose the manufacturing process is slipping out of proper adjustment. As the actual mean diameter of the catheters produced gets farther and farther above the desired 2.00 mm, will the power of the quality control test increase, decrease, or remain the same?
d) What could they do to improve the power of the test?

Ex23.40. Wind power. Should you generate electricity with your own personal wind turbine? That depends on whether you have enough wind on your site. To produce enough energy, your site should have an

annual average wind speed above 8 miles per hour, according to the Wind Energy Association. One candidate site was monitored for a year, with wind speeds recorded every 6 hours. A total of 1114 readings of wind speed averaged 8.019 mph with a standard deviation of 3.813 mph. You've been asked to make a statistical report to help the landowner decide whether to place a wind turbine at this site.

a) Discuss the assumptions and conditions for using Student's t inference methods with these data. Here are some plots that may help you decide whether the methods can be used:



b) What would you tell the landowner about whether this site is suitable for a small wind turbine? Explain.

Ex24.2. Dogs and sodium. The Consumer Reports article described in Exercise 1 also listed the sodium content (in mg) for the various hot dogs tested. A test of the null hypothesis that beef hot dogs and meat hot dogs don't differ in the mean amounts of sodium yields a P-value of 0.11. Would a 95% confidence interval for $\mu_{Meat} - \mu_{Beef}$ include 0? Explain.

Ex24.4. Washers. In the June 2007 issue, Consumer Reports examined top-loading and front-loading washing machines, testing samples of several different brands of each type. One of the variables the article reported was "cycle time," the number of minutes it took each machine to wash a load of clothes. Among the machines rated good to excellent, the 98% confidence interval for the difference in mean cycle time ($\mu_{Top} - \mu_{Front}$) is (–40, –22).

a) The endpoints of this confidence interval are negative numbers. What does that indicate?

b) What does the fact that the confidence interval does not contain 0 indicate?

c) If we use this confidence interval to test the hypothesis that $\mu_{Top} - \mu_{Front} = 0$, what's the corresponding alpha level?

Ex24.6. Second load of wash. In Exercise 4, we saw a 98% confidence interval of (–40, –22) minutes for $\mu_{Top} - \mu_{Front}$ the difference in time it takes top-loading and front-loading washers to do a load of clothes. Explain why you think each of the following statements is true or false:

a) 98% of top loaders are 22 to 40 minutes faster than front loaders.

b) If I choose the laundromat's top loader, there's a 98% chance that my clothes will be done faster than if I had chosen the front loader.

c) If I tried more samples of both kinds of washing machines, in about 98% of these samples I'd expect the top loaders to be an average of 22 to 40 minutes faster.

d) If I tried more samples, I'd expect about 98% of the resulting confidence intervals to include the true difference in mean cycle time for the two types of washing machines.

e) I'm 98% confident that top loaders wash clothes an average of 22 to 40 minutes faster than front-loading machines.

Ex24.8. Stereograms. Stereograms appear to be composed entirely of random dots. However, they contain separate images that a viewer can "fuse" into a three-dimensional (3D) image by staring at the dots while defocusing the eyes. An experiment was performed to determine whether knowledge of the form of the embedded image affected the time required for subjects to fuse the images. One group of subjects (group

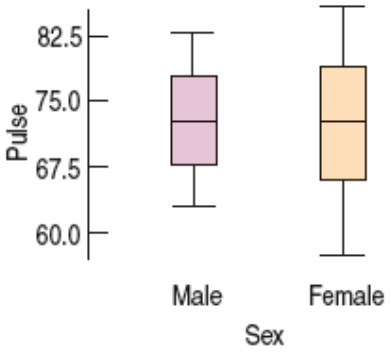
NV) received no information or just verbal information about the shape of the embedded object. A second group (group VV) received both verbal information and visual information (specifically, a drawing of the object). The experimenters measured how many seconds it took for the subject to report that he or she saw the 3D image.

2-Sample t-Interval for $\mu_1 - \mu_2$
Conf level 90% df 70 μ(NV) − μ(VV) interval: (0.55, 5.47)

a) Interpret your interval in context.
b) Does it appear that viewing a picture of the image helps people "see" the 3D image in a stereogram?
c) What's the margin of error for this interval?
d) Explain carefully what the 90% confidence level means.
e) Would you expect a 99% confidence level to be wider or narrower? Explain.
f) Might that change your conclusion in part b? Explain.

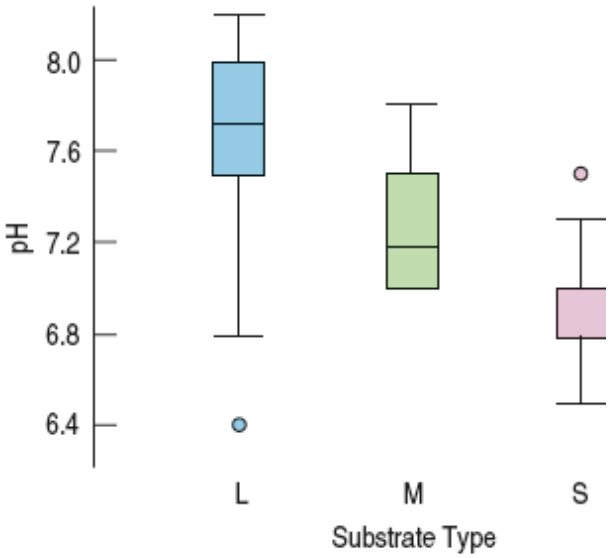
Ex24.12. Pulse rates. A researcher wanted to see whether there is a significant difference in resting pulse rates for men and women. The data she collected are summarized and displayed in the boxplots below.

	Sex	
	Male	Female
Count	28	24
Mean	72.75	72.625
Median	73	73
StdDev	5.37225	7.69987
Range	20	29
IQR	9	12.5



a) What do the boxplots suggest about differences between male and female pulse rates?
b) Is it appropriate to analyze these data using the methods of inference discussed in this chapter? Explain.
c) Create a 90% confidence interval for the difference in mean pulse rates.
d) Does the confidence interval confirm your answer to part a? Explain.

Ex24.16. Streams. Researchers collected samples of water from streams in the Adirondack Mountains to investigate the effects of acid rain. They measured the pH (acidity) of the water and classified the streams with respect to the kind of substrate (type of rock over which they flow). A lower pH means the water is more acidic. Here is a plot of the pH of the streams by substrate (limestone, mixed, or shale). And here are selected parts of a software analysis comparing the pH of streams with limestone and shale substrates.



2-Sample t-Test of $\mu 1 - \mu 2$
Difference Between Means = 0.735
t-Statistic = 16.30 w/133 df
$p \leq 0.0001$

a) State the null and alternative hypotheses for this test.
b) From the information you have, do the assumptions and conditions appear to be met?
c) What conclusion would you draw?

Ex24.22. Summer school. Having done poorly on their math final exams in June, six students repeat the course in summer school, then take another

exam in August. If we consider these students representative of all students who might attend this summer school in other years, do these results provide evidence that the program is worthwhile?

June	54	49	68	66	62	62
Aug.	50	65	74	64	68	72

Ex24.24. Ad campaign. You are a consultant to the marketing department of a business preparing to launch an ad campaign for a new product. The company can afford to run ads during one TV show, and has decided not to sponsor a show with sexual content. You read the study described in Exercise 23, then use a computer to create a confidence interval for the difference in mean number of brand names remembered between the groups watching violent shows and those watching neutral shows.

```
TWO-SAMPLE T   95% CI FOR MUviol - MUneut: (-1.578, -0.602)
```

a) At the meeting of the marketing staff, you have to explain what this output means. What will you say?
b) What advice would you give the company about the upcoming ad campaign?

Ex24.26. Ad recall. In a research we see the number of advertised brand names people recalled immediately after watching TV shows and 24 hours later. Strangely enough, it appears that they remembered more about the ads the next day. Should we conclude this is true in general about people's memory of TV ads?
a) Suppose one analyst conducts a two-sample hypothesis test to see if memory of brands advertised during violent TV shows is higher 24 hours later. If his P-value is 0.00013, what might he conclude?
b) Explain why his procedure was inappropriate. Which of the assumptions for inference was violated?
c) How might the design of this experiment have tainted the results?
d) Suggest a design that could compare immediate brand-name recall with recall one day later.

Ex25.2. MTV. Some students do homework with the TV on. (Anyone come to mind?) Some researchers want to see if people can work as effectively with as without distraction. The researchers will time some volunteers to see how long it takes them to complete some relatively easy crossword puzzles. During some of the trials, the room will be quiet; during other trials in the same room, a TV will be on, tuned to MTV.
a) Design an experiment that will require a two-sample t procedure to analyze the results.
b) Design an experiment that will require a matched pairs t procedure to analyze the results.
c) Which experiment would you consider the stronger design? Why?

Ex25.4. Freshman 15? Many people believe that students gain weight as freshmen. Suppose we plan to conduct a study to see if this is true.
a) Describe a study design that would require a matched-pairs t procedure to analyze the results.
b) Describe a study design that would require a two-sample t procedure to analyze the results.

Ex25.6. Cloud seeding. Simpson, Alsen, and Eden (Technometrics 1975) report the results of trials in which clouds were seeded and the amount of rainfall recorded. The authors report on 26 seeded and 26 unseeded clouds in order of the amount of rainfall, largest amount first. Here are two possible tests to study the question of whether cloud seeding works. Which test is appropriate for these data? Explain your choice. Using the test you select, state your conclusion.

Paired t-Test of $\mu(1 - 2)$ Mean of Paired Differences = – 277.39615
t-Statistic = -3.641 w/25 df p = 0.0012

and

2-Sample t-Test of $\mu_1 - \mu_2$ Difference Between Means = –277.4 t-Statistic = -1.998 w/33 df p= 0.0538

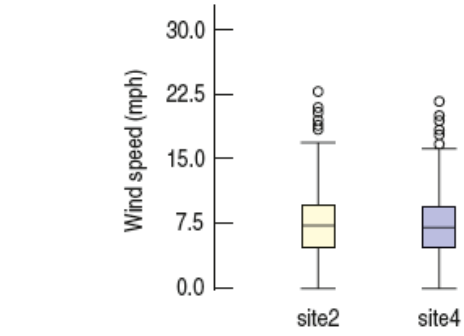
a) Which of these tests is appropriate for these data? Explain.
b) Using the test you selected, state your conclusion.

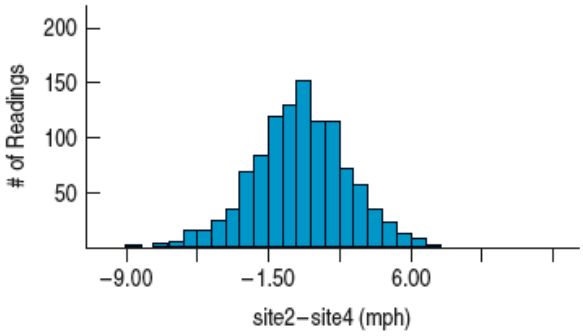
Ex25.10. Wind speed, part I. To select the site for an electricity generating wind turbine, wind speeds were recorded at several potential sites every 6 hours for a year. Two sites not far from each other looked good. Each had a mean wind speed high enough to qualify, but we should choose the site with a higher average daily wind speed. Because the sites are near each other and the wind speeds were recorded at the same times, we should view the speeds as paired. Here are the summaries of the speeds (in miles per hour):

Variable	Count	Mean	StdDev
site2	1114	7.452	3.586
site4	1114	7.248	3.421
site2 – site4	1114	0.204	2.551

Is there a mistake in this output? Why doesn't the Pythagorean Theorem of Statistics work here? In other words, shouldn't $SD(site2 - site4) = \sqrt{SD^2(site2) + SD^2(site4)}$. But $\sqrt{3.586^2 + 3.421^2} = 4.956$, not 2.551 as given by the software. Explain why this happened.

Ex25.12 Wind speed, part II. In Exercise 10, we saw summary statistics for wind speeds at two sites near each other, both being considered as locations for an electricity generating wind turbine. The data, recorded every 6 hours for a year, showed each of the sites had a mean wind speed high enough to qualify, but how can we tell which site is best? Here are some displays:





a) The boxplots show outliers for each site, yet the histogram shows none. Discuss why.

b) Which of the summaries would you use to select between these sites? Why?

c) Using the information you have, discuss the assumptions and conditions for paired t inference for these data. (Hint: Think hard about the independence assumption in particular.)

II.2: Program week 2: Comparing counts and analysis of variance

Literature: De Veaux, Velleman, & Bock, chapters 25 & 28.

Less reading, but we are still in the rehearsing mode: these topics have been covered in the Statistics 1 course. But are part of Statistics II content too.

Discussion tasks (for discussion in the tutorial class session, 1st session of the week):

Chapter 26: exercises 2, 4, 6, 14, 16, 18, 20, 22, 33, 35.

Chapter 28: exercises 2, 4, 6, 8, 10, 12, 14, 16, 18, 20.

SPSS Case Study II: we will do this case study in the lab session, the 2nd session of the week. Finish outside call, if you cannot finish within the lab time. The case study assignment can be found in the StudentPortal.

Discussion tasks:

Ex26.2. Which test, again? For each of the following situations, state whether you'd use a chi-square goodness-of-fit test, a chi-square test of homogeneity, a chi-square test of independence, or some other statistical test:

a) Is the quality of a car affected by what day it was built? A car manufacturer examines a random sample of the warranty claims filed over the past two years to test whether defects are randomly distributed across days of the workweek.

b) A medical researcher wants to know if blood cholesterol level is related to heart disease. She examines a database of 10,000 patients, testing whether the cholesterol level (in milligrams) is related to whether or not a person has heart disease.

c) A student wants to find out whether political leaning (liberal, moderate, or conservative) is related to choice of major. He surveys 500 randomly chosen students and performs a test.

Ex26.4. M&M's. As noted in an earlier chapter, the Masterfoods Company says that until very recently yellow candies made up 20% of its milk chocolate M&M's, red another 20%, and orange, blue, and green 10% each. The rest are brown. On his way home from work the day he was writing these exercises, one of the authors bought a bag of plain M&M's. He got 29 yellow ones, 23 red, 12 orange, 14 blue, 8 green, and 20 brown. Is this sample consistent with the company's stated proportions? Test an appropriate hypothesis and state your conclusion.

a) If the M&M's are packaged in the stated proportions, how many of each color should the author have expected to get in his bag?

b) To see if his bag was unusual, should he test goodness-of-fit, homogeneity, or independence?

c) State the hypotheses.

d) Check the conditions.

e) How many degrees of freedom are there?

f) Find χ^2 and the P-value.

g) State a conclusion.

Ex26.6. Mileage. A salesman who is on the road visiting clients thinks that, on average, he drives the same distance each day of the week. He keeps track of his mileage for several weeks and discovers that he averages 122 miles on Mondays, 203 miles on Tuesdays, 176 miles on Wednesdays, 181 miles on Thursdays, and 108 miles on Fridays. He wonders if this evidence contradicts his belief in a uniform distribution of miles across the days of the week. Explain why it is not appropriate to test his hypothesis using the chi-square goodness-of-fit test**.**

Ex26.14. Does your doctor know? A survey7 of articles from the New England Journal of Medicine (NEJM) classified them according to the principal statistics methods used. The articles recorded were all noneditorial articles appearing during the indicated years. Let's just look at whether these articles used statistics at all. Has there been a change in the use of Statistics?

	Publication Year			
	1978–79	1989	2004–05	Total
No stats	90	14	40	144
Stats	242	101	271	614
Total	332	115	311	758

a) What kind of test would be appropriate?
b) State the null and alternative hypotheses.

Ex26.16. Does your doctor know? (part 2). The table in Exercise 14 shows whether NEJM medical articles during various time periods included statistics or not. We're planning to do a chi-square test.
a) How many degrees of freedom are there?
b) The smallest expected count will be in the 1989/No cell. What is it?
c) Check the assumptions and conditions for inference.

Ex26.18. Does your doctor know? (part 3). In Exercises 14 and 16, we've begun to examine whether the use of statistics in NEJM medical articles has changed over time.
a) Calculate the component of chi-square for the 1989/No cell.
b) For this test, $\chi^2 = 25.28$. What's the P-value?
c) State your conclusion.

Ex26.20. Does your doctor know? (part 4). In Exercises 14, 16, and 18, we've tested a hypothesis about whether the use of statistics in NEJM medical articles has changed over time. The table shows the test's residuals.

	1978–79	1989	2004–05
No stats	3.39	−1.68	−2.48
Stats	−1.64	0.81	1.20

a) Show how the residual for the 1989/No cell was calculated.
b) What can you conclude from the patterns in the standardized residuals?

Ex26.22. Does your doctor know? (part 5). In Exercises 14, 16, 18, and 20, we considered data on articles in the NEJM. The original study listed 23 different Statistics methods. (The list read: t-tests, contingency tables, linear regression, ...) Why would it not be appropriate to use a chi-square test on the 23 * 3 table with a row for each method?

Ex26.33. Grades. Two different professors teach an introductory Statistics course. The table shows the distribution of final grades they reported. We wonder whether one of these professors is an "easier" grader.

	Prof. Alpha	Prof. Beta
A	3	9
B	11	12
C	14	8
D	9	2
F	3	1

a) Will you test goodness-of-fit, homogeneity, or independence?
b) Write appropriate hypotheses.
c) Find the expected counts for each cell, and explain why the chi-square procedures are not appropriate.

Ex26.35. Grades, again. In some situations where the expected cell counts are too small, as in the case of the grades given by Professors Alpha and Beta in Exercise 33, we can complete an analysis anyway. We can often proceed after combining cells in some way that makes sense and also produces a table in which the conditions are satisfied. Here we create a new table displaying the same data, but calling D's and F's "Below C":a) Find the expected counts for each cell in this new table, and explain why a chi-square procedure is now appropriate.

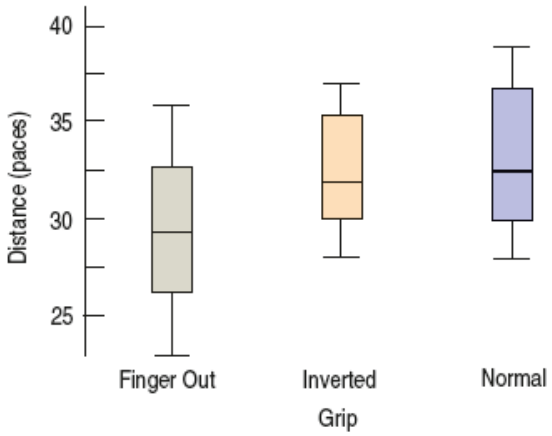
	Prof. Alpha	Prof. Beta
A	3	9
B	11	12
C	14	8
Below C	12	3

b) With this change in the table, what has happened to the number of degrees of freedom?
c) Test your hypothesis about the two professors, and state an appropriate conclusion.

Ex28.2. Skating. A figure skater tried various approaches to her Salchow jump in a designed experiment using 5 different places for her focus (arms, free leg, midsection, takeoff leg, and free). She tried each jump 6 times in random order, using two of her skating partners to judge the jumps on a scale from 0 to 6. After collecting the data and analyzing the results, she reports that the F-ratio is 7.43.
a) What are the null and alternative hypotheses?
b) How many degrees of freedom does the treatment sum of squares have? How about the error sum of squares?
c) Assuming that the conditions are satisfied, what is the P-value? What would you conclude?
d) What else about the data would you like to see in order to check the assumptions and conditions?

Ex28.4. Darts. A student interested in improving her dartthrowing technique designs an experiment to test 4 different stances to see whether they affect her accuracy. After warming up for several minutes, she randomizes the order of the 4 stances, throws a dart at a target using each stance, and measures the distance of the dart in centimeters from the center of the bull's-eye. She replicates this procedure 10 times. After analyzing the data she reports that the F-ratio is 1.41.
a) What are the null and alternative hypotheses?
b) How many degrees of freedom does the treatment sum of squares have? How about the error sum of squares?
c) What would you conclude?
d) What else about the data would you like to see in order to check the assumptions and conditions?
e) If your conclusion in part c is wrong, what type of error have you made?

Ex28.6. Frisbee throws. A student performed an experiment with three different grips to see what effect it might have on the distance of a backhanded Frisbee throw. She tried it with her normal grip, with one finger out, and with the Frisbee inverted. She measured in paces how far her throw went. The boxplots and the ANOVA table for the three grips are shown below:a) State the hypotheses about the grips.
b) Assuming that the assumptions for inference are satisfied, perform the hypothesis test and state your conclusion. Be sure to state it in terms of Frisbee grips and distance thrown.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Grip	2	58.58333	29.2917	2.0453	0.1543
Error	21	300.75000	14.3214		
Total	23	359.33333			

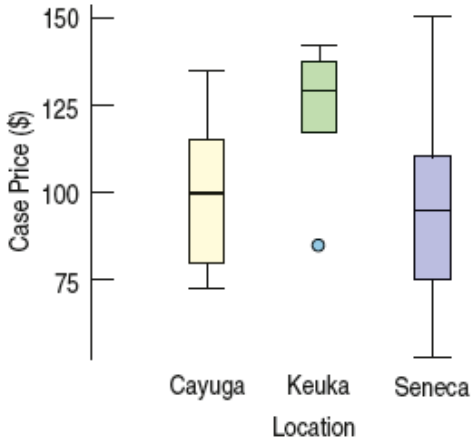
c) Would it be appropriate to follow up this study with multiple comparisons to see which grips differ in their mean distance thrown? Explain.

Ex28.8. Zip codes, revisited. The intern from the marketing department at the Holes R Us online piercing salon (Chapter 4, Exercise 49) has recently finished a study of the company's 500 customers. He wanted to know whether people's zip codes vary by the last product they bought. They have 16 different products, and the ANOVA table of zip code by product showed the following: (Nine customers were not included because of missing zip code or product information.) What criticisms of the analysis might you make? What alternative analysis might you suggest?

ANOVA table

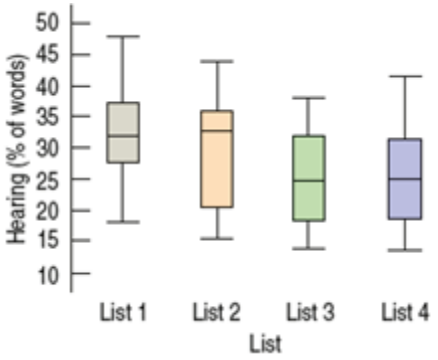
Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Product	15	3.836e10	2.55734e9	4.9422	<0.0001
Error	475	2.45787e11	517445573		
Total	490	2.84147e11			

Ex28.10. Wines, revisited. This boxplots displays case prices (in dollars) of wines produced by wineries along three of the Finger Lakes.



a) What are the null and alternative hypotheses? Talk about prices and location, not symbols.
b) Do the conditions for an ANOVA seem to be met here? Why or why not?

Ex28.12. Hearing. A researcher investigated four different word lists for use in hearing assessment. She wanted to know whether the lists were equally difficult to understand in the presence of a noisy background. To find out, she tested 96 subjects with normal hearing randomly assigning 24 to each of the four word lists and measured the number of words perceived correctly in the presence of background noise. Here are the boxplots of the four lists:
a) What are the null and alternative hypotheses?
b) What do you conclude?
c) Would it be appropriate to run a multiple comparisons test (for example, a Bonferroni test) to see which lists differ from each other in terms of mean percent correct? Explain.



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
List	3	920.4583	306.819	4.9192	0.0033
Error	92	5738.1667	62.371		
Total	95	6658.6250			

Ex28.14. Smokestack scrubbers. Particulate matter is a serious form of air pollution often arising from industrial production. One way to reduce the pollution is to put a filter, or scrubber, at the end of the

An incomplete ANOVA Table for the Smokestack Data

Source	Sum of Squares	Degrees of Freedom	Mean Square	F-ratio
Treatment	81.2			
Residual	30.8			
Total	112.0			

smokestack to trapthe particulates. An experiment to determine which smokestack scrubber design is best was run by placing four scrubbers of different designs on an industrial stack in random order. Each scrubber was tested 5 times. For each run, the same material was produced, and the particulate emissions coming out of the scrubber were measured (in parts per billion). A partially complete Analysis of Variance table of the data is shown below.

a) Calculate the mean square of the treatments and the mean square of the error.

b) Form the F-statistic by dividing the two mean squares.

c) The P-value of this F-statistic turns out to be 0.0000949. What does this say about the null hypothesis of equal means?

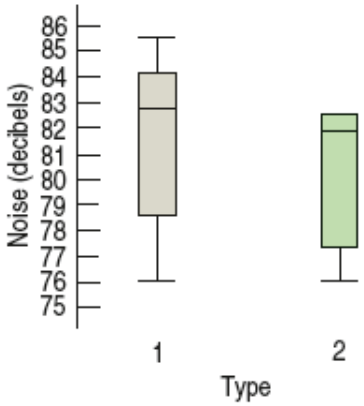
d) What assumptions have you made in order to answer part c?

e) What would you like to see in order to justify the conclusions of the F-test?

f) What is the average size of the error standard deviation in particulate emissions?

Ex28.16. Auto noise filters. In a statement to a Senate Public Works Committee, a senior executive of Texaco, Inc., cited a study on the effectiveness of auto filters on reducing noise. Because of concerns about performance, two types of filters were studied, a standard silencer and a new device developed by the Associated Octel Company.

Here are the boxplots from the data on noise reduction (in decibels) of the two filters. Type 1 = Standard; Type 2 = Octel.



a) What are the null and alternative hypotheses?

b) What do you conclude from the ANOVA table?

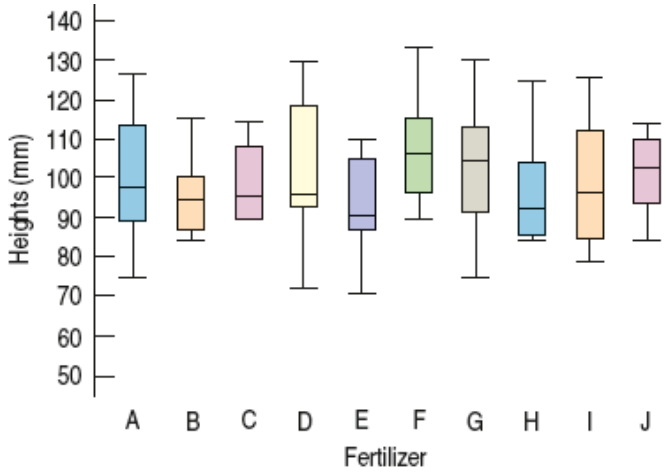
c) Do the assumptions for the test seem to be reasonable?

d) Perform a two-sample pooled t-test of the difference. What P-value do you get? Show that the square of the t-statistic is the same (to rounding error) as the F-ratio.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Type	1	6.31	6.31	0.7673	0.3874
Error	33	271.47	8.22		
Total	34	2.77			

Level	*n*	Mean	StdDev
Standard	18	81.5556	3.2166
Octel	17	80.7059	2.43708



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Fertilizer	9	2073.708	230.412	1.1882	0.3097
Error	110	21331.083	193.919		
Total	119	23404.791			

Ex28.18. Fertilizers. A biology student is studying the effect of 10 different fertilizers on the growth of mung bean sprouts. She sprouts 12 beans in each of 10 different petri dishes, and adds the same amount of fertilizer to each dish. After one week she measures the heights of the 120 sprouts in millimeters. Here are boxplots and an ANOVA table of the data:

a) What are the null and alternative hypotheses?

b) What does the ANOVA table say about the null hypothesis? (Be sure to report this in terms of heights and fertilizers).

c) Her lab partner looks at the same data and says that he did t-tests of every fertilizer against every other fertilizer and finds that several of the fertilizers seem to have significantly higher mean heights. Does this match your finding in part b? Give an explanation for the difference, if any, between the two results.

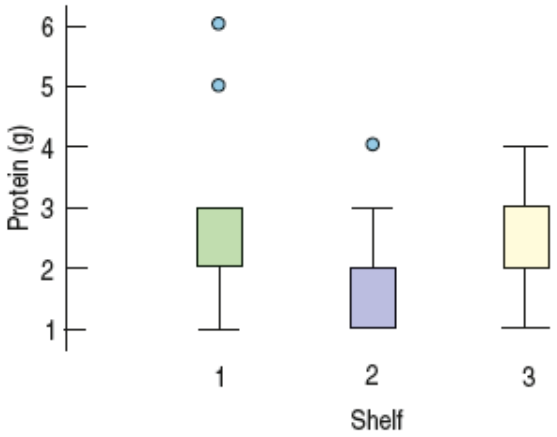
Ex28.20. Cereals, redux. We also have data on the protein content of cereals by their shelf number. Here are the boxplot and ANOVA table:

a) What are the null and alternative hypotheses?

b) What does the ANOVA table say about the null hypothesis? (Be sure to report this in terms of protein content and shelves.)

c) Can we conclude that cereals on shelf 2 have a lower mean protein content than cereals on shelf 3? Can we conclude that cereals on shelf 2 have a lower mean protein content than cereals on shelf 1? What can we conclude?

d) To check for significant differences between the shelf means we can use a Bonferroni test, whose results are shown below. For each pair of shelves, the difference is shown along with its standard error and significance



Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Shelf	2	12.4258	6.2129	5.8445	0.0044
Error	74	78.6650	1.0630		
Total	76	91.0909			

Means and Std Deviations

Level	*n*	Mean	StdDev
1	20	2.65000	1.46089
2	21	1.90476	0.99523
3	36	2.86111	0.72320

Dependent Variable: PROTEIN Bonferroni

(I) SHELF	(J) SHELF	Mean Difference (I-J)	Std. Error	P-value	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	0.75	0.322	0.070	−0.04	1.53
	3	−0.21	0.288	1.000	−0.92	0.49
2	1	−0.75	0.322	0.070	−1.53	0.04
	3	−0.96(*)	0.283	0.004	−1.65	−0.26
3	1	0.21	0.288	1.000	−0.49	0.92
	2	0.96(*)	0.283	0.004	0.26	1.65

*The mean difference is significant at the 0.05 level.

II.3: Program week 3: Multivariate ANOVA

Literature: De Veaux, Velleman, & Bock, chapter 29.

This is for the first time real new stuff.

Discussion tasks (for discussion in the tutorial class session, 1st session of the week):

Chapter 29: exercises 2, 4, 6, 8, 12, 14, 16, 19, 20, 21.

SPSS Case Study III: we will do this case study in the lab session, the 2nd session of the week. Finish outside call, if you cannot finish within the lab time. The case study assignment can be found in the StudentPortal.

Discussion tasks:

Ex29.2. Gas mileage revisited. A student runs an experiment to study the effect of *Tire Pressure* and *Acceleration* on gas mileage. He devises a system so that his Jeep Wagoneer uses gasoline from a one-liter container. He uses 3 levels of *Tire Pressure* (*low*, *medium*, and *full*) and 2 levels of *Acceleration*, either holding the pedal *steady* or *pumping* it every few seconds. He randomizes the trials, performing 4 runs under each treatment condition, carefully recording the number of miles he can go in his Jeep Wagoneer on one liter of gas.

a) What are the null and alternative hypotheses for the main effects?

b) How many degrees of freedom does each treatment sum of squares have? How about the error sum of squares?

c) Should he consider fitting an interaction term to the model? Why might it be a good idea?

d) If he fits an interaction term, how many degrees of freedom would it have?

Ex29.4. Gas mileage again. Refer to the experiment in Exercise 2. After analyzing his data the student reports that the F-ratio for *Tire Pressure* is 4.29 with a P-value of 0.030, the F-ratio for *Acceleration* is 2.35 with a P-value of 0.143, and the F-ratio for the Interaction effect is 1.54 with a P-value of 0.241.

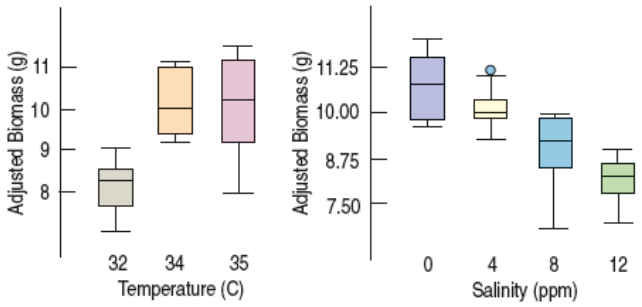
a) What would you conclude?

b) What else about the data would you like to see in order to check the assumptions and conditions?

c) If your conclusion about the *Acceleration* factor in part a is wrong, what type of error have you made?

Ex29.6. Sprouts. An experiment on mung beans was performed to investigate the environmental effects of salinity and water temperature on sprouting. Forty beans were randomly allocated to each of 36 petri dishes that were subject to one of four levels of Salinity (0, 4, 8 and 12 ppm) and one of three Temperatures (32°, 34°, or 36° C). After 48 hours, the biomass of the sprouts was measured.

Here are partial boxplots of Biomass on Salinity and Temperature:

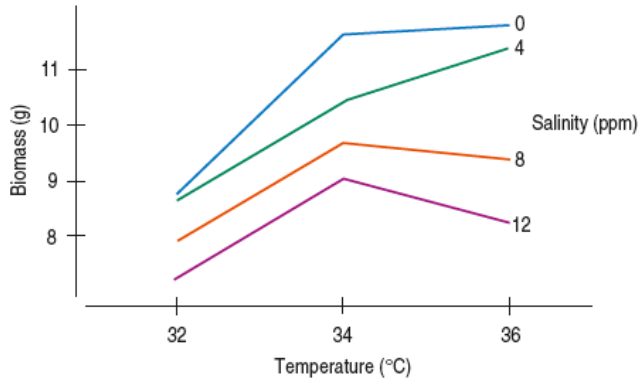


A two-way ANOVAmodel is fit, and the following ANOVA table results:

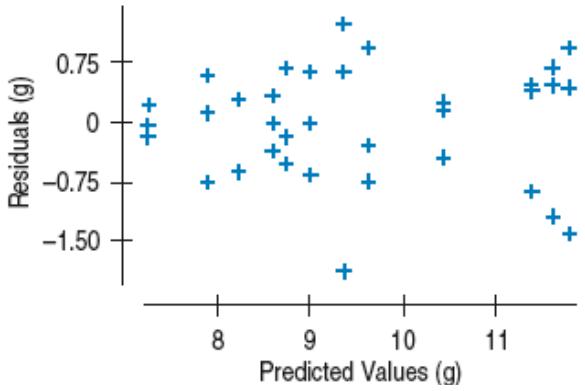
Analysis of Variance for Biomass (g)

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Salinity	3	36.4701	12.1567	16.981	<0.0001
Temp	2	34.7168	17.3584	24.247	<0.0001
Salinity × Temp	6	5.2972	0.8829	1.233	0.3244
Error	24	17.1816	0.7159		
Total	35	93.6656			

The interaction plot shows:



The plot of residuals vs. predicted values shows:



a) State the hypotheses about the factors (both numerically and in words).
b) Are the conditions for two-way ANOVA met?
c) Perform the hypothesis tests and state your conclusions. Be sure to state your conclusions in terms of biomass, salinity, and water temperature.

Ex29. 8. Fish and prostate. In the Chapter 3 Step-By-Step, we looked at a Swedish study that asked 6272 men how much fish they ate and whether or not they had prostate cancer. Here are the data:

Eat Fish?	Prostate Cancer? No	Prostate Cancer? Yes
Never/seldom	110	14
Small part of diet	2420	201
Moderate part	2769	209
Large part	507	42

Armed with the methods of this chapter, a student performs a two-way ANOVA on the data. Here is her ANOVA table:

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Fish	3	3110203.0	1036734.3	1.3599	0.4033
Prostate cancer	1	3564450.0	3564450.0	4.6756	0.1193
Error	3	2287051.0	762350.0		

a) Comment on her analysis. What problems, if any, do you find with the analysis?
b) What sort of analysis might you do instead?

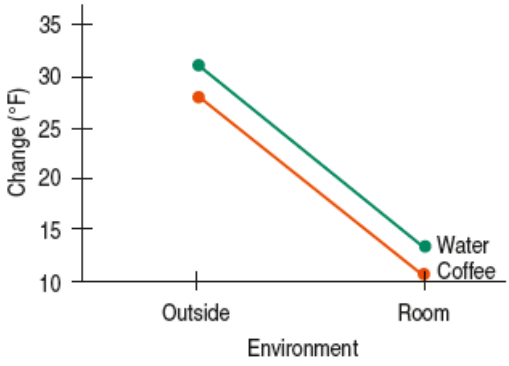
Ex29.12. Washing. For his final project, Jonathan examined the effects of two factors on how well stains are removed when washing clothes. On each of 16 new white handkerchiefs, he spread a teaspoon of dirty motor oil (obtained from a local garage). He chose 4 *Temperature* settings (each of which is a combination of wash and rinse: cold-cold, cold-warm, warm-hot, and hot-hot) and 4 *Cycle* lengths (short, med short, med long, and long). After its washing, each handkerchief was dried in a dryer for 20 minutes and hung up. He rounded up 10 family members to judge the handkerchiefs for cleanliness on a scale of 1 to 10 and used the average score as his response. Here are the data:

Temp	Cycle	Scor
Cold-cold	Med long	3.7
Warm-hot	Med long	6.5
Cold-warm	Med long	4.9
Hot-hot	Med long	6.5
Cold-cold	Long	4.6
Warm-hot	Long	8.3
Cold-warm	Long	4.7
Hot-hot	Long	9.1
Cold-cold	Short	3.4
Warm-hot	Short	5.6
Cold-warm	Short	3.8
Hot-hot	Short	7.1
Cold-cold	Med short	3.1
Warm-hot	Med short	6.3
Cold-warm	Med short	5
Hot-hot	Med short	6.1

You may assume, as Jonathan did, that interactions between *Temperatue* and *Cycle* are negligible. Write a report showing what you found about washing factors and stain removal.

Ex29.14. Containers revisited. Building on the cup experiment of the Chapter 4 Step-By-Step, a student selects one type of container and designs an experiment to see whether the type of Liquid stored and the outside Environment affect the ability of a cup to maintain temperature. He randomly chooses an experimental condition and runs each twice. After fitting a two-way ANOVA-model, he obtains the following interaction plot, ANOVA-table, and effects table.

Liquid	Environment	Change in Temperature
Water	Room	13
Water	Room	14
Water	Outside	31
Water	Outside	31
Coffee	Room	11
Coffee	Room	11
Coffee	Outside	27
Coffee	Outside	29

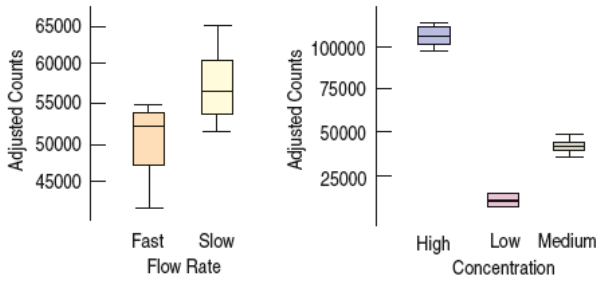


Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Liquid	1	15.125	15.125	24.2	0.0079
Envirn	1	595.125	595.125	952.2	<0.0001
Interaction	1	0.125	0.125	0.2	0.6779
Error	4	2.500	0.625		
Total	7	612.875			

Term	Estimate
Overall mean	20.875
Liquid II [Coffee]	−1.375
Liquid II [Water]	1.375
Environment II [Outside]	8.625
Environment II [Room]	−8.625

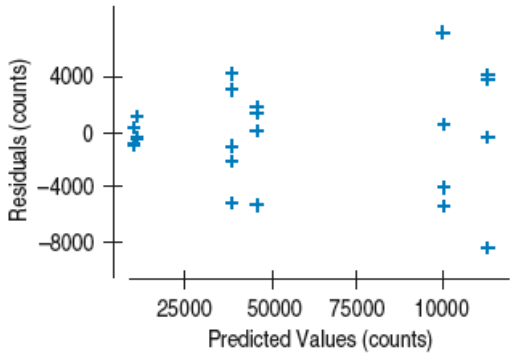
a) State the null and alternative hypotheses.
b) Test the hypotheses at $\alpha = 0.05$
c) Perform a residual analysis.
d) Summarize your findings

Ex29.16. Chromatography. A gas chromatograph is an instrument that measures the amounts of various compounds contained in a sample by separating the various constituents. Because different components are flushed through the system at different rates, chromatographers are able to both measure and distinguish the various constituents of the sample. A counter is placed somewhere along the instrument that records how much material is passing at various times. By looking at the counts at various times, the chemist is able to reconstruct the amounts of various compounds present. The total number of counts is proportional to the amount of the compound present.
An experiment was performed to see whether slowing down the flow rate would increase total counts. A mixture was produced with three different Concentration levels: low, medium, and high. The two *Flow Rates* used were slow and fast. Each mixture was run 5 times and the total counts recorded each time. Partial boxplots for *Concentration* and *Flow Rate* show:
What conclusions about the effect of flow rate do you draw? Do you see any potential problems with the analysis?



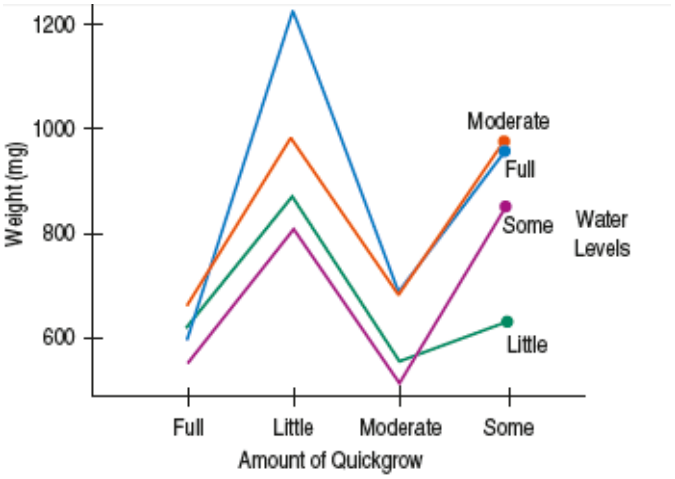
A two-way ANOVA with interaction model was run, and the following ANOVA table resulted:

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Conc	2	483655E5	241828E5	1969.44	<0.0001
Flow rate	1	364008E3	364008E3	29.65	<0.0001
Interaction	2	203032E3	101516E3	8.27	0.0019
Error	24	294698E3	122791E3		
Total	29	492272E5			



Ex29.20. Peas. In an experiment on growing sweet peas, a team of students selected 2 factors at 4 levels each and recorded *Weight*, *Stem Length*, and *Root Length* after days of growth. They grew plants using various amounts of *Water* and *Quickgrow* solution, a fertilizer designed to help plants grow faster. Each factor was run at 4 levels: little, some, moderate, and full. They grew 2 plants under each of the 16 conditions.

An interaction plot of Weight in mg (x-axis is *Quickgrow*—levels are water) shows:



Because of this, a two-way ANOVAwith interaction model was fit to the data, resulting in the following ANOVA table:

Analysis of Variance for Weight

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Water	3	246255	82084.8	2.5195	0.0948
QS	3	827552	275851	8.4669	0.0013
Water × QS	8	176441	22055.1	0.6770	0.7130
Error	16	521275	32579.7		
Total	31	1771523			

Residuals plots showed no violations of the variance or Normality conditions. A table of effects for *Quickgrow* Is shown left, a table of effects for Water shown right.

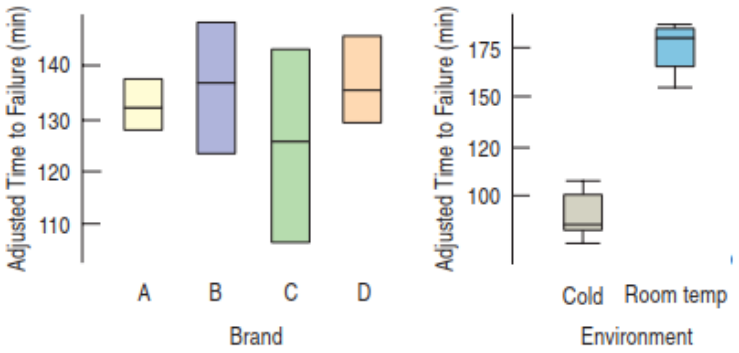
Level of Quickgrow	Effect
Little	213.8
Some	97.1
Moderate	−155.5
Full	−155.5

Level of Water	Effect
Little	−91.1
Some	−81.9
Moderate	66.4
Full	106.0

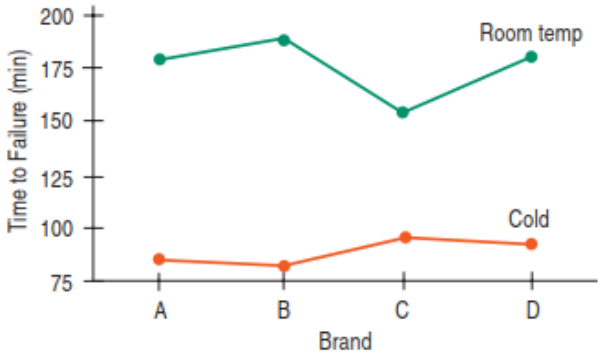
Summarize what the students have learned about the effects of *Water* and *Quickgrow* solution on the early stages of sweet pea growth as measured by *Weight*.

Ex 29.19 Batteries. A student experiment was run to test the performance of 4 brands of batteries under 2 different Environments (room temperature and cold). For each of the 8 treatments, 2 batteries of a particular brand were put into a flashlight. The flashlight was then turned on and allowed to run until the light went out. The number of minutes the flashlight stayed on was recorded. Each treatment condition was run twice.

Partial boxplots showed:



An interaction plot showed:



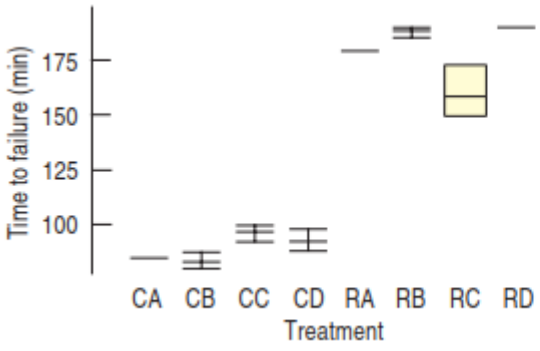
a) What are the main effect null and alternative hypotheses?
b) From the partial boxplots, do you think that the Brand has an effect on the time the batteries last? How about the condition?
c) Do the conclusions of the ANOVA table match your intuition?
d) What does the interaction plot say about the performance of the brands?

An ANOVA table showed:

Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Envir	1	30363.1	30363.1	789.93	<0.0001
Brand	3	338.187	112.729	2.9328	0.0994
Interaction	3	1278.19	426.063	11.085	0.0032
Error	8	307.5	38.4375		
Total	15	32286.9			

e) Why might you be uncomfortable with a recommendation to go with the cheapest battery (brand C)?

Ex 29.21 Batteries once more. Another student analyzed the battery data from Exercise 19, using a one-way ANOVA. He considered the experimental factor to be an 8-level factor consisting of the 8 possible combinations of Brand and Environment. Here are the boxplots for the 8 treatments and a one-way ANOVA:



Source	DF	Sum of Squares	Mean Square	F-ratio	P-value
Treatment	7	31895.8	4556.54	93.189	<0.0001
Error	8	391.167	48.8958		
Total	15	32286.9			

Compare this analysis with the one performed in Exercise 19. Which one provides a better understanding of the data? Explain.

II.4: Program week 4: Inference for regression

Literature: De Veaux, Velleman, & Bock, chapters 7, 8, 9, 10 & 27.

We covered correlation and regression in Statistics 1, so not much new here. Be it that we have been short on inference for regression, so the last chapter will be relatively new, and topic of all discussion tasks. All chapters are however part of the course content.

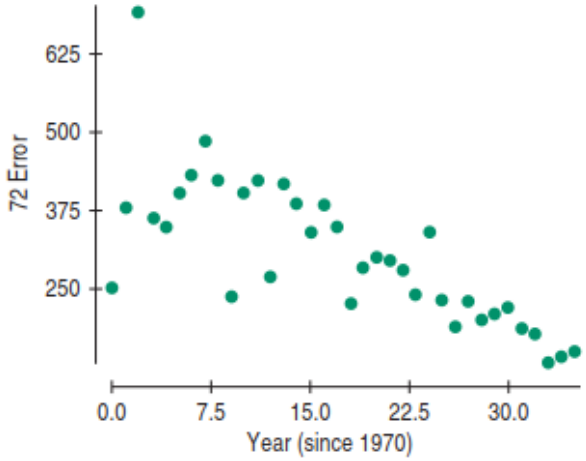
Discussion tasks (for discussion in the tutorial class session, 1st session of the week):

Chapter 27: exercises 1, 2, 4, 8, 10, 12, 18, 24, 26, 28, 40, 42, 43.

SPSS Case Study IV: we will do this case study in the lab session, the 2nd session of the week. Finish outside class, if you cannot finish within the lab time. The case study assignment can be found in the StudentPortal.

Discussion tasks:

Ex27.1. Hurricane predictions. Let's look at data from the National Oceanic and Atmospheric Administration about their success in predicting hurricane tracks. Here is a scatterplot of the error (in nautical miles) for predicting hurricane locations 72 hours in the future vs. the year in which the prediction (and the hurricane) occurred, and the outcome of the regression analysis:



Dependent variable is: 72Error
R squared = 58.5%
s = 75.38 with 36 − 2 = 34 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	453.223	24.61	18.4	≤0.0001
Year since 1970	−8.37084	1.209	−6.92	≤0.0001

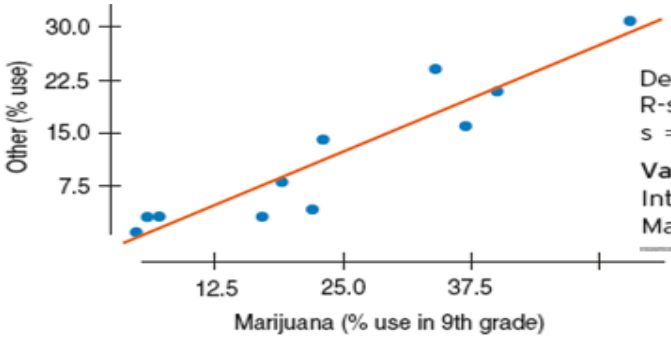
a) Explain in words and numbers what the regression says.

b) State the hypothesis about the slope (both numerically and in words) that describes how hurricane prediction quality has changed.

c) Assuming that the assumptions for inference are satisfied, perform the hypothesis test and state your conclusion. Be sure to state it in terms of prediction errors and years.

d) Explain what the R-squared means in terms of this regression.

Ex27.2. Drug use. The European School Study Project on Alcohol and Other Drugs, investigated the use of marijuana and other drugs. Data from 11 countries are summarized in the following scatterplot and regression analysis. They show the association between the percentage of a country's ninth graders who report having smoked marijuana and who have used other drugs such as LSD, amphetamines, and cocaine.

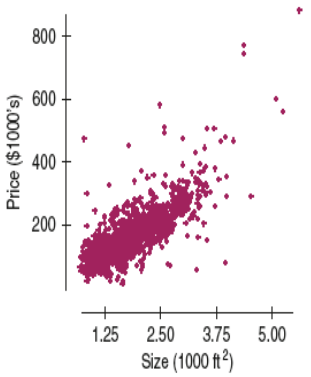


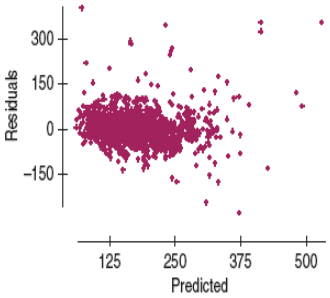
Dependent variable is: Other
R-squared = 87.3%
s = 3.853 with 11 − 2 = 9 degrees of freedom

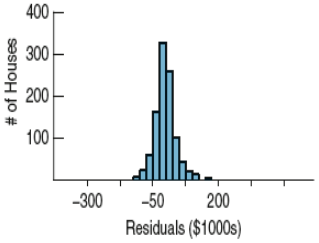
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−3.06780	2.204	−1.39	0.1974
Marijuana	0.615003	0.0784	7.85	<0.0001

a) Explain in context what the regression says.
b) State the hypothesis about the slope (both numerically and in words) that describes how use of marijuana is associated with other drugs.
c) Assuming that the assumptions for inference are satisfied, perform the hypothesis test and state your conclusion in context.
d) Explain what R-squared means in context.
e) Do these results indicate that marijuana use leads to the use of harder drugs? Explain.

Ex27.4. Saratoga house prices. How does the price of a house depend on its size? Data from Saratoga, New York, on 1064 randomly selected houses that had been sold include data on price ($1000's) and size (1000's ft^2), producing the following graphs and computer output:







Dependent variable is: Price
R squared = 59.5%
s = 53.79 with 1064 − 2 = 1062 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−3.11686	4.688	−0.665	0.5063
Size	94.4539	2.393	39.5	≤0.0001

a) Explain in context what the regression says.
b) The intercept is negative. Discuss its value, taking note of its P-value.
c) The output reports Explain what that means in this context.
d) What's the value of the standard error of the slope of the regression line?
e) Explain what that means in this context.

Ex 27.8. Cholesterol 2007. Does a person's cholesterol level tend to change with age? Data collected from 1406 adults aged 45 to 62 produced the regression analysis shown. Assuming that the data satisfy the conditions for inference, examine the association between age and cholesterol level.

Dependent variable is: Chol
s = 46.16

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	194.232	13.55	14.3	≤0.0001
Age	0.771639	0.2574	3.00	0.0056

a) State the appropriate hypothesis for the slope.
b) Test your hypothesis and state your conclusion in the proper context.

Ex27.10. More cholesterol. Look again at Exercise 8's regression output for age and cholesterol level.

a) The output reports *s* = 46.16. Explain what that means in this context.
b) What's the value of the standard error of the slope of the regression line?
c) Explain what that means in this context.

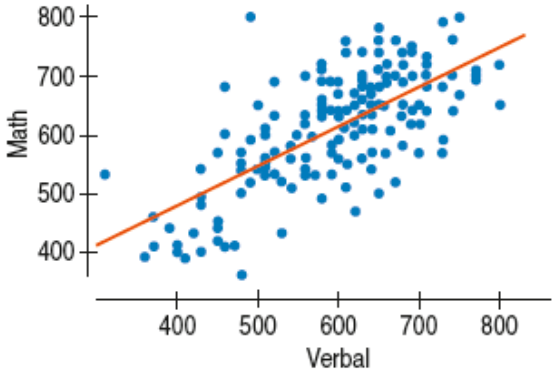
Ex27.12. Cholesterol, finis. Based on the regression output seen in Exercise 8, create a 95% confidence interval for the slope of the regression line and interpret it in context.

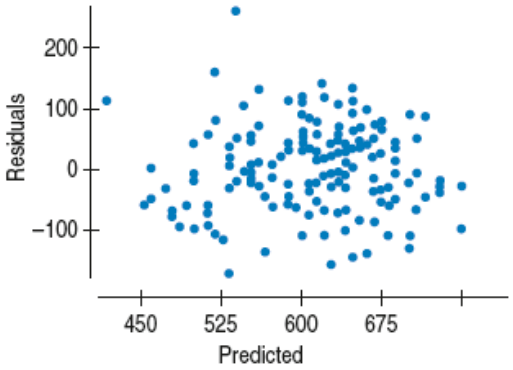
Ex27.18. SAT scores. How strong was the association between student scores on the Math and Verbal sections of the old SAT? Scores on each ranged from 200 to 800 and were widely used by college admissions offices. Here are summaries and plots of the scores for a graduating class at Ithaca High School:
a) Is there evidence of an association between Math and Verbal scores? Write an appropriate hypothesis.
b) Discuss the assumptions for inference.
c) Test your hypothesis and state an appropriate conclusion.

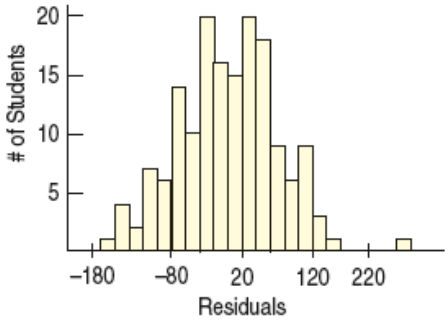
Variable	Count	Mean	Median	StdDev	Range	IntQRange
Verbal	162	596.296	610	99.5199	490	140
Math	162	612.099	630	98.1343	440	150

Dependent variable is: Math
R-squared = 46.9%
s = 71.75 with 162 − 2 = 160 df

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	209.554	34.35	6.10	≤0.0001
Verbal	0.675075	0.0568	11.9	≤0.0001



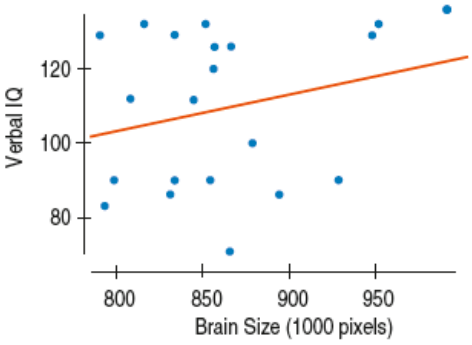




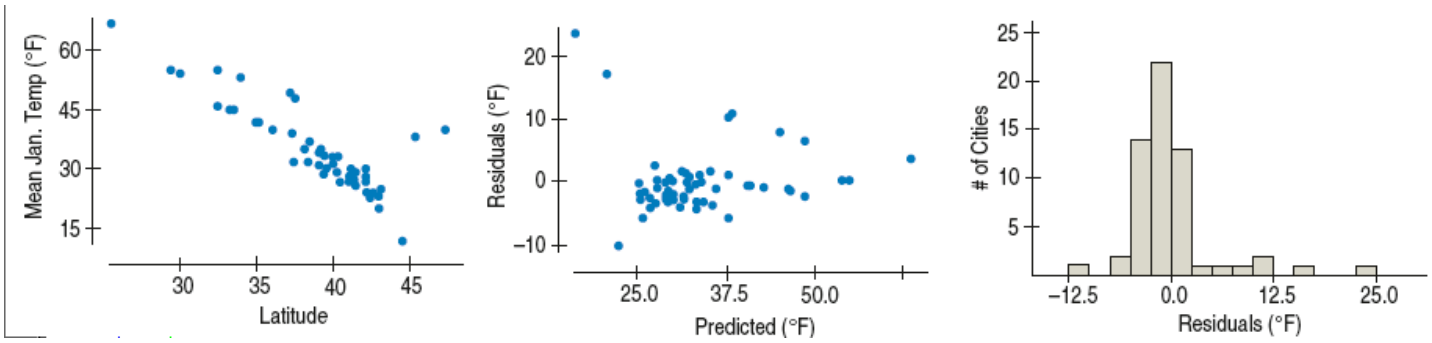
Ex27.24. Brain size. Does your IQ depend on the size of your brain? A group of female college students took a test that measured their verbal IQs and also underwent an MRI
scan to measure the size of their brains (in 1000s of pixels).
The scatterplot and regression analysis are shown, and the assumptions for inference were satisfied.
a) Test an appropriate hypothesis about the association between brain size and IQ.
b) State your conclusion about the strength of this association.

Dependent variable is: IQ_Verbal
R-squared = 6.5%

Variable	Coefficient	SE(Coeff)
Intercept	24.1835	76.38
Size	0.098842	0.0884



Ex27.26. Winter. The output shows an attempt to model the association between average January Temperature (in degrees Fahrenheit) and Latitude (in degrees north of the equator) for 59 U.S. cities. Which of the assumptions for inference do you think are violated? Explain.



Ex27.28. Climate change and CO_2. Concern over the weather associated with El Niño has increased interest in the possibility that the climate on earth is getting warmer. The most common theory relates an increase in atmospheric levels of carbon dioxide a greenhouse gas, to increases in temperature.

Here is part of a regression analysis of the mean annual concentration in the atmosphere, measured in parts per million (ppm), at the top of Mauna Loa in Hawaii and the mean annual air temperature over both land and sea across the globe, in degrees Celsius. The scatterplots and residuals plots indicated that the data were appropriate for inference.

Dependent variable is: Annual Temp
R-squared = 67.8%
s = 0.0985 with 29 − 2 = 27 degrees of freedom

Variable	Coefficient	SE(Coeff)
Intercept	10.7071	0.4810
CO2	0.010062	0.0013

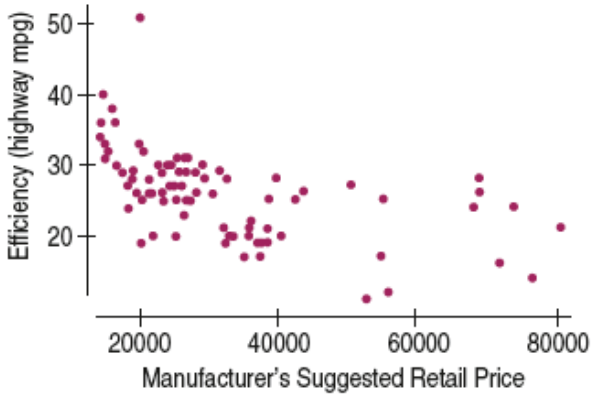
a) Write the equation of the regression line.
b) Is there evidence of an association between level and global temperature?
c) Do you think predictions made by this regression will be very accurate? Explain.

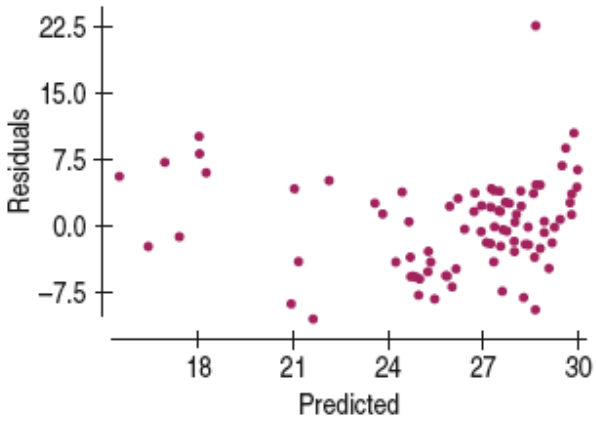
Ex27.40. All the efficiency money can buy. A sample of 84 model-2004 cars from an online information service was examined to see how fuel efficiency (as highway mpg) relates to the cost (Manufacturer's Suggested Retail Price in dollars) of cars. Here are displays and computer output:

Dependent variable is: Highway MPG
R squared = 30.1%
s = 5.298 with 84 − 2 = 82 degrees of freedom

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Constant	33.0581	1.299	25.5	≤0.0001
MSRP	−2.16543e−4	0.0000	−5.95	≤0.0001

a) State what you want to know, identify the variables, and give the appropriate hypotheses.
b) Check the assumptions and conditions.
c) If the conditions are met, complete the analysis.



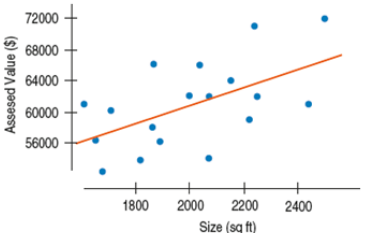


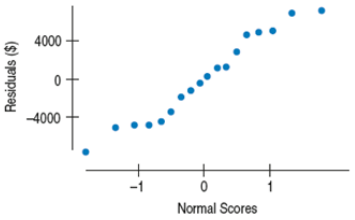
Ex27.42. Property assessments. The following software outputs provide information about the Size (in square feet) of 18 homes in Ithaca, New York, and the city's assessed Value of those homes.

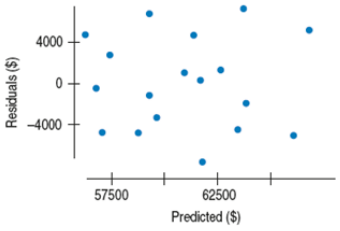
Variable	Count	Mean	StdDev	Range
Size	18	2003.39	264.727	890
Value	18	60946.7	5527.62	19710

Dependent variable is: Value
R-squared = 32.5%
s = 4682 with 18 − 2 = 16 degrees of freedom

Variable	Coefficient	SE(Coeff)
Intercept	37108.8	8664
Size	11.8987	4.290







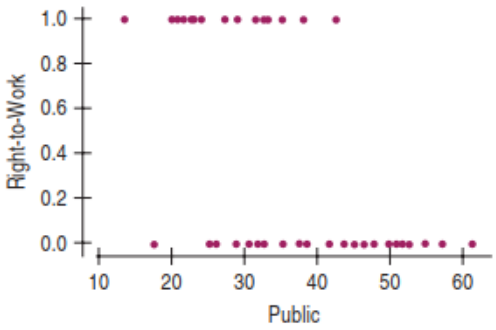
a) Explain why inference for linear regression is appropriate with these data.
b) Is there a significant association between the Size of a home and its assessed Value? Test an appropriate hypothesis and state your conclusion.
c) What percentage of the variability in assessed Value is explained by this regression?
d) Give a 90% confidence interval for the slope of the true regression line, and explain its meaning in the proper context.
e) From this analysis, can we conclude that adding a room to your house will increase its assessed Value? Why or why not?
f) The owner of a home measuring 2100 square feet files an appeal, claiming that the $70,200 assessed Value is too high. Do you agree? Explain your reasoning.

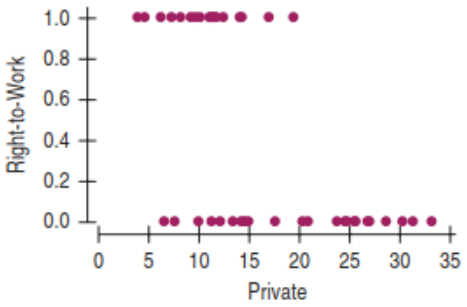
Ex27.43. Right-to-work laws. Are state right-to-work laws related to the percent of public sector employees in unions and the percent of private sector employees in unions? This data set looks at these percentages for the states in the United States in 1982. The dependent variable is whether the state had a right-to-work law or not. The computer output for the logistic regression is given here. (Source: N. M. Meltz, "Interstate and Interprovincial Differences in Union Density," Industrial Relations, 28:2 [Spring 1989], 142–158 by way of DASL.)

Logistic Regression Table

Predictor	Coeff	SE(Coeff)	z	P
Intercept	6.19951	1.78724	3.47	0.001
publ	−0.106155	0.0474897	−2.24	0.025
pvt	−0.222957	0.0811253	−2.75	0.006

a) Write out the estimated regression equation.
b) The following are scatterplots of the response variable against each of the explanatory variables. Examine them for the conditions required by logistic regression. Does logistic regression seem appropriate here? Explain.





II.5: Program week 5: Multiple Regression

Literature: De Veaux, Velleman, & Bock, chapters 30 & 31.

Discussion tasks (for discussion in the tutorial class session, 1st session of the week):

Chapter 30: exercises 2, 4, 6, 8, 10, 12, 16.

Chapter 31: exercises 1, 2, 4, 6, 8, 10, 11, 12.

SPSS Case Study V: we will do this case study in the lab session, the 2nd session of the week. Finish outside class, if you cannot finish within the lab time. The case study assignment can be found in the StudentPortal.

Discussion tasks:

Ex30.2. More interpretations. A household appliance manufacturer wants to analyze the relationship between total sales and the company's three primary means of advertising (television, magazines, and radio). All values were in millions of dollars. They found the regression equation

$$\widehat{Sales} = 250 + 6.75\,TV + 3.5\,Radio + 2.3\,Magazines.$$

One of the interpretations below is correct. Which is it? Explain what's wrong with the others.

a) If they did no advertising, their income would be $250 million.

b) Every million dollars spent on radio makes sales increase $3.5 million, all other things being equal.

c) Every million dollars spent on magazines increases TV spending $2.3 million.

d) Sales increase on average about $6.75 million for each million spent on TV, after allowing for the effects of the other kinds of advertising.

Ex30.4. Scottish hill races 2008. Hill running—races up and down hills—has a written history in Scotland dating back to the year 1040. Races are held throughout the year at different locations around Scotland. A recent compilation of information for 91 races (for which full information was available and omitting two unusual races) includes the *Distance* (km), the *Climb* (m), and the *Record Time* (minutes). A regression to predict the men's records as of 2008 looks like this:

Dependent variable is: Time (mins)
R-squared = 98.0% R-squared (adjusted) = 98.0%
s = 6.623 with 90 − 3 = 87 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	189204	2	94602.1	2157
Residual	3815.92	87	43.8612	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−10.3723	1.245	−8.33	≤0.0001
Climb (m)	0.034227	0.0022	15.7	≤0.0001
Distance (km)	4.04204	0.1448	27.9	≤0.0001

a) Write the regression equation. Give a brief report on what it says about men's record times in hill races.

b) Interpret the value of in this regression.

c) What does the coefficient of *Climb* mean in this regression?

Ex30.6. More hill races 2008. Here is the regression for the women's records for the same Scottish hill races

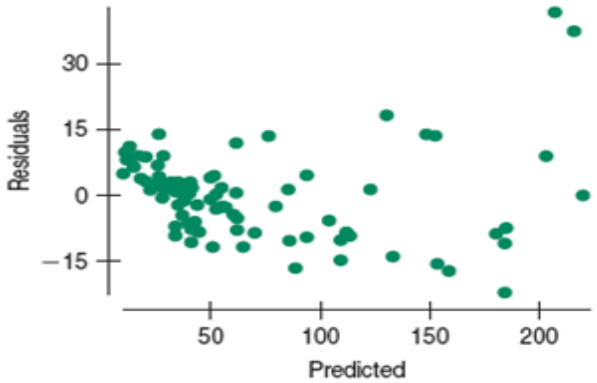
a) Compare the regression model for the women's records with that found for the men's records.

Here's a scatterplot of the residuals for this regression:

Dependent variable is: Women's Time (mins)
R-squared = 96.7% R-squared (adjusted) = 96.7%
s = 10.06 with 90 − 3 = 87 degrees of freedom

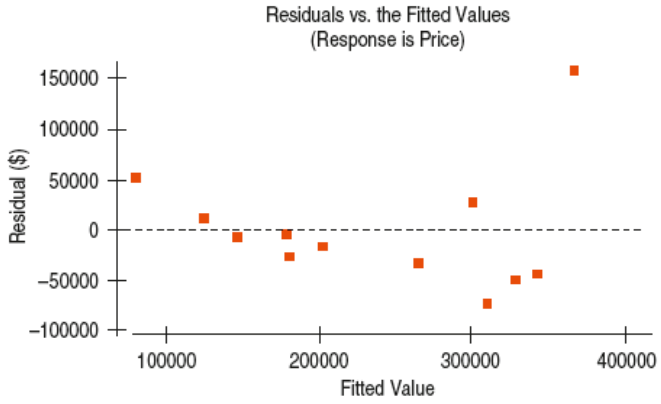
Source	Sum of Squares	df	Mean Square	F-ratio
Regression	261029	2	130515	1288
Residual	8813.02	87	101.299	

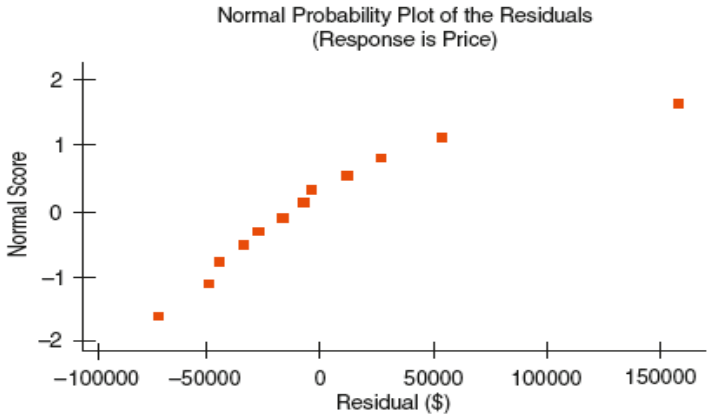
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−11.6545	1.891	−6.16	<0.0001
Climb (m)	0.045195	0.0033	13.7	<0.0001
Distance	4.43427	0.2200	20.2	<0.0001

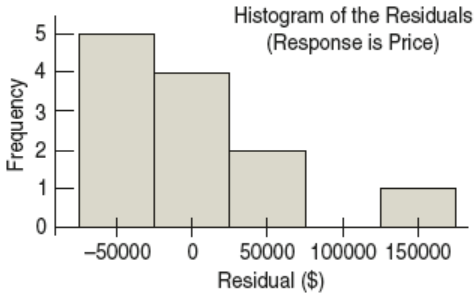


b) Discuss the residuals and what they say about the assumptions and conditions for this regression.

Ex30.8. Home prices II. Here are some diagnostic plots for the home prices data. These were generated by a computer package and may look different from the plots generated by the packages you use. (In particular, note that the axes of the Normal probability plot are swapped relative to the plots we've made in the text. We only care about the pattern of this plot, so it shouldn't affect your interpretation.) Examine these plots and discuss whether the assumptions and conditions for the multiple regression seem reasonable.







Ex30.10. GPA and SATs. A large section of Stat 101 was asked to fill out a survey on grade point average and SAT scores. A regression was run to find out how well Math and Verbal SAT scores could predict academic performance as measured by GPA. The regression was run on a computer package with the following output:

Response: GPA

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	0.574968	0.253874	2.26	0.0249
SAT Verbal	0.001394	0.000519	2.69	0.0080
SAT Math	0.001978	0.000526	3.76	0.0002

a) What is the regression equation?

b) From this model, what is the predicted GPA of a student with an SAT Verbal score of 500 and an SAT Math score of 550?

c) What else would you want to know about this regression before writing a report about the relationship between SAT scores and grade point averages? Why would these be important to know?

Ex30.12. Breakfast cereals. We saw in Chapter 8 that the calorie content of a breakfast cereal is linearly associated with its sugar content. Is that the whole story? Here's the output of a regression model that regresses Calories for each serving on its Protein(g), Fat(g), Fiber(g), Carbohydrate(g), and Sugars(g) content.

Assuming that the conditions for multiple regression are met,

a) What is the regression equation?

b) Do you think this model would do a reasonably good job at predicting calories? Explain.

c) To check the conditions, what plots of the data might you want to examine?

d) What does the coefficient of Fat mean in this model?

Dependent variable is: Calories
R-squared = 93.6% R-squared (adjusted) = 93.1%
s = 5.113 with 77 − 6 = 71 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ratio
Regression	26995.9	5	5399.18	207
Residual	1856.03	71	26.1412	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−0.879994	4.383	−0.201	0.8414
Protein	3.60495	0.6977	5.17	≤0.0001
Fat	8.56877	0.6625	12.9	≤0.0001
Fiber	0.309180	0.3337	0.927	3.572
Carbo	4.13996	0.2049	20.2	≤0.0001
Sugars	4.00677	0.1719	23.3	≤0.0001

Ex30.16. Breakfast cereals again. We saw in Chapter 8 that the calorie count of a breakfast cereal is linearly associated with its sugar content. Can we predict the calories of a serving from its vitamin and mineral content? Here's a multiple regression model of Calories per serving on its Sodium (mg), Potassium (mg), and Sugars (g):

Assuming that the conditions for multiple regression are met,

a) What is the regression equation?

b) Do you think this model would do a reasonably good job at predicting calories? Explain.

c) Would you consider removing any of these predictor variables from the model? Why or why not?

d) To check the conditions, what plots of the data might you want to examine?

Dependent variable is: Calories
R-squared = 38.4% R-squared (adjusted) = 35.9%
s = 15.60 with 77 − 4 = 73 degrees of freedom

Source	Sum of Squares	dF	Mean Square	F-ratio	P-value
Regression	11091.8	3	3697.28	15.2	<0.0001
Residual	17760.1	73	243.289		

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	83.0469	5.198	16.0	<0.0001
Sodium	0.05721	0.0215	2.67	0.0094
Potass	−0.01933	0.0251	−0.769	0.4441
Sugars	2.38757	0.4066	5.87	<0.0001

Ex31.1. Climate change 2009. Recent concern with the rise in global temperatures has focused attention on the level of carbon dioxide (CO2) in the atmosphere. The National Oceanic and Atmospheric Administration (NOAA) records the CO2 levels in the atmosphere atop the Mauna Loa volcano in Hawaii, far from any industrial contamination, and calculates the annual overall temperature of the atmosphere and the oceans using an established method. (See http://www.esrl.noaa.gov/gmd/ccgg/trends/ for the CO2 levels and http://www.ncdc.noaa.gov/oa/climate/research/anomalies/index.php#means for the temperatures.) Here is a regression predicting Mean Annual Temperature Anomaly (°C away from the 20th-century mean) from annual CO2 levels (parts per million). We'll examine the data from 1959 to 2009.

a) Comment on the distribution of the Studentized residuals.

b) It is widely understood that global temperatures have been rising consistently during this period. But the coefficient of Year is negative and its P-value is small.

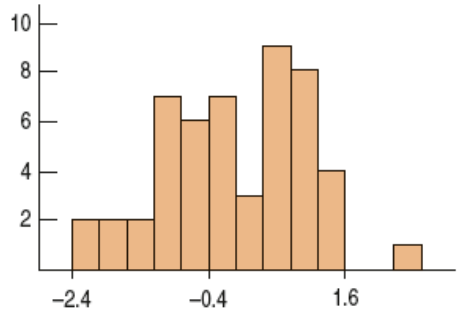
Does this contradict the common wisdom?

Dependent variable is: Global Temperature Anomaly
R-squared = 85.3% R-squared (adjusted) = 84.7%
s = 0.0850 with 51 − 3 = 48 degrees of freedom

Source	Sum of Squares	df	Mean Square	F-ra
Regression	2.00798	2	1.00399	13
Residual	0.346811	48	0.007225	

Variable	Coefficient	SE(Coeff)	t-ratio	P-valu
Intercept	30.9595	12.52	2.47	0.017
Year	−0.019404	0.0072	−2.71	0.009
CO_2	0.022413	0.0049	4.55	≤0.000

A histogram of the externally Studentized resid looks like this:



Ex31.2. Pizza. Consumers' Union rated frozen pizzas. Their report includes the number of Calories, Fat content, and Type (cheese or pepperoni, represented here as an indicator variable that is 1 for cheese and 0 for pepperoni). Here's a regression model to predict the "Score" awarded each pizza from these variables:

a) What is the interpretation of the coefficient of cheese in this regression?

b) What displays would you like to see to check assumptions and conditions for this model?

Dependent variable is: Score
R-squared = 28.7%
R-squared (adjusted) = 20.2%
s = 19.79 with 29 − 4 = 25 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	3947.34	3	1315.78	3.36
Residual	9791.35	25	391.654	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−148.817	77.99	−1.91	0.0679
Calories	0.743023	0.3066	2.42	0.0229
Fat	−3.89135	2.138	−1.82	0.0807
Type	15.6344	8.103	1.93	0.0651

Ex31.4. Fifty states. In Exercise 15 of Chapter 30 we looked at data from the 50 states. Here's an analysis of the same data from a few years earlier. The Murder rate is per 100,000, HS Graduation rate is in %, Income is per capita income in dollars, Illiteracy rate is per 1000, and Life Expectancy is in years. We are trying to find a regression model for Life Expectancy.

Here's the result of a regression on all the available predictors:

Dependent variable is: Lifeexp
R-squared = 67.0% R-squared (adjusted) = 64.0%
s = 0.8049 with 50 − 5 = 45 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	59.1430	4	14.7858	22.8
Residual	29.1560	45	0.6479	

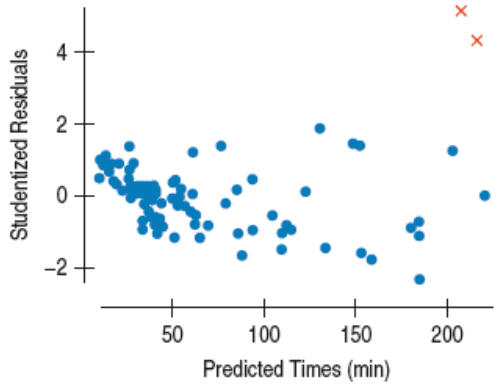
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	69.4833	1.325	52.4	≤0.0001
Murder	−0.261940	0.0445	−5.89	≤0.0001
HS grad	0.046144	0.0218	2.11	0.0403
Income	1.24948e-4	0.0002	0.516	0.6084
Illiteracy	0.276077	0.3105	0.889	0.3787

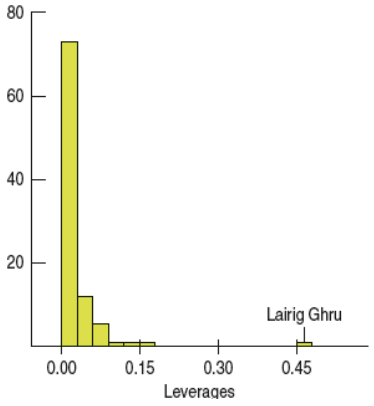
Ex31.6. Scottish hill races 2008. In Chapter 30, Exercises 4 and 6, we considered data on hill races in Scotland. These are overland races that climb and descend hills—sometimes several hills in the course of one race. Here is a regression analysis to predict the Women's Record times from the Distance and total vertical Climb of the races:

Dependent variable is: Women's record
R-squared = 96.7% R-squared (adjusted) = 96.7%
s = 10.06 with 90 − 3 = 87 degrees of freedom

Source	Sum of Squares	dF	Mean Square	F-ratio
Regression	261029	2	130515	1288
Residual	8813.02	87	101.299	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−11.6545	1.891	−6.16	≤0.0001
Distance	4.43427	0.2200	20.2	≤0.0001
Climb	0.045195	0.0033	13.7	≤0.0001





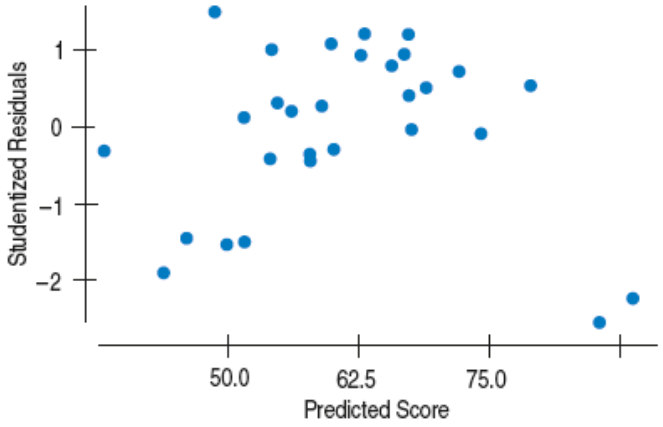
Here is the scatterplot of externally Studentized residuals against predicted values, as well as a histogram of leverages for this regression:

a) Comment on what these diagnostic displays indicate.

b) The two races with the largest Studentized residuals are the Arochar Alps race and the Glenshee 9. Both are relatively new races, having been run only one or two times with relatively few participants. What effects can you be reasonably sure they have had on the regression? What displays would you want to see to investigate other effects? Explain.

c) If you have access to a suitable statistics package, make the diagnostic plots you would like to see and discuss what you find.

Ex31.8. Gourmet pizza. Here's a plot of the Studentized residuals against the predicted values for the regression model found in Exercise 2:



The two extraordinary cases in the plot of residuals are Reggio's and Michelina's, two gourmet pizzas.

a) Interpret these residuals. What do they say about these two brands of frozen pizza? Be specific—that is, talk about the Scores they received and might have been expected to receive.

We can create indicator variables to isolate these cases. Adding them to the model results in the following model:

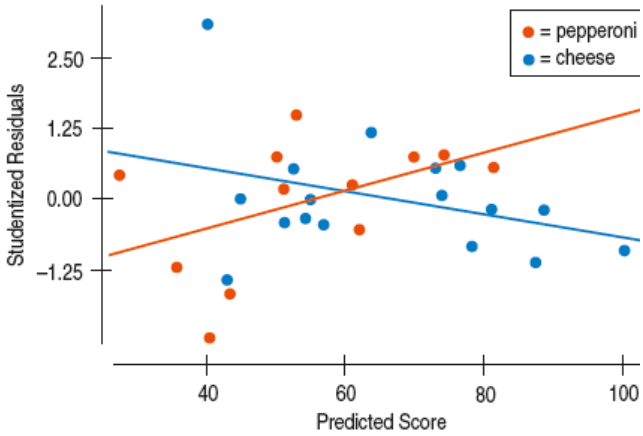
b) What does the coefficient of Michelina's mean in this regression model? Do you think that Michelina's pizza is an outlier for this model for these data? Explain.

Dependent variable is: Score
R-squared = 65.2% R-squared (adjusted) = 57.7%
s = 14.41 with 29 − 6 = 23 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	8964.13	5	1792.83	8.64
Residual	4774.56	23	207.590	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−363.109	72.15	−5.03	≤0.0001
Calories	1.56772	0.2824	5.55	≤0.0001
Fat	−8.82748	1.887	−4.68	0.0001
Cheese	25.1540	6.214	4.05	0.0005
Reggio's	−67.6401	17.86	−3.79	0.0010
Michelina's	−67.0036	16.62	−4.03	0.0005

Ex31.10. Another slice of pizza. A plot of Studentized residual against predicted values for the regression model found in Exercise 8 now looks like this. It has been colored according to Type of pizza and separate regression lines fitted for each type:



a) Comment on this diagnostic plot. What does it say about how the regression model deals with cheese and pepperoni pizzas?

Based on this plot, we constructed yet another variable consisting of the indicator cheese multiplied by Calories:

b) Interpret the coefficient of Cheese*cals in this regression model.

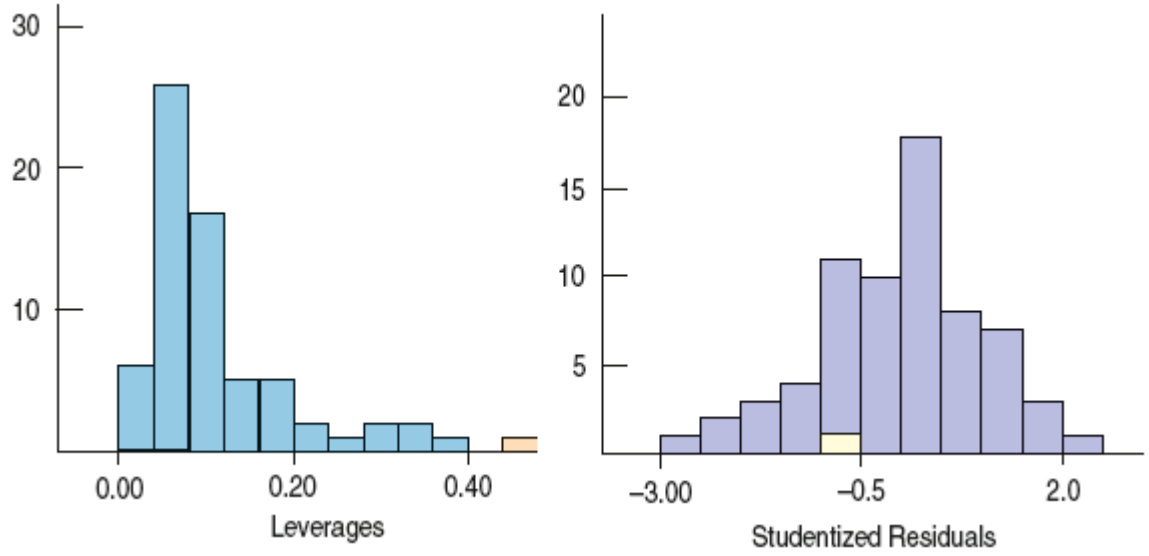
c) Would you prefer this regression model to the model of Exercise 8? Explain.

Dependent variable is: Score
R-squared = 73.7% R-squared (adjusted) = 66.5%
s = 12.82 with 29 − 7 = 22 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	10121.4	6	1686.90	10.3
Residual	3617.32	22	164.424	

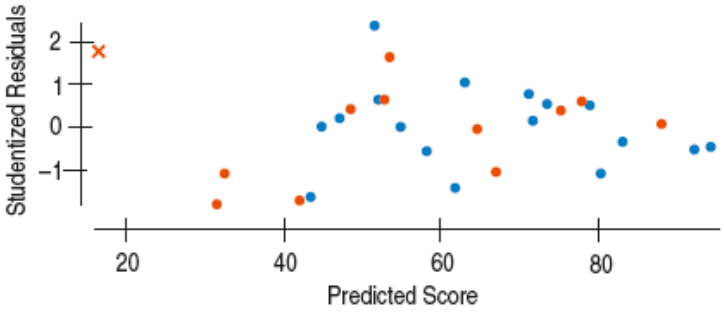
Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−464.498	74.73	−6.22	≤0.0001
Calories	1.92005	0.2842	6.76	≤0.0001
Fat	−10.3847	1.779	−5.84	≤0.0001
Cheese	183.634	59.99	3.06	0.0057
Cheese*cals	−0.461496	0.1740	−2.65	0.0145
Reggio's	−64.4237	15.94	−4.04	0.0005
Michelina's	−51.4966	15.90	−3.24	0.0038

Ex31.11. Influential traffic? Here are histograms of the leverage and Studentized residuals for the regression model of Exercise 9.



The city with the highest leverage is Colorado Springs, CO. It's highlighted in both displays. Do you think Colorado Springs is an influential case? Explain your reasoning.

Ex31.12. The final slice. Here's the residual plot corresponding to the regression model of Exercise 10:



The extreme case this time is Weight Watchers Pepperoni (makes sense, doesn't it?). We can make one more indicator for Weight Watchers. Here's the model:

a) Compare this model with the others we've seen for these data. In what ways does this model seem better or worse than the others?

b) Do you think the indicator for Weight Watchers should be in the model? (Consider the effect that including it has had on the other coefficients also.)

c) What do the Consumers' Union tasters seem to think makes for a really good pizza?

Dependent variable is: Score
R-squared = 77.1% R-squared (adjusted) = 69.4%
s = 12.25 with 29 − 8 = 21 degrees of freedom

Source	Sum of Squares	DF	Mean Square	F-ratio
Regression	10586.8	7	1512.41	10.1
Residual	3151.85	21	150.088	

Variable	Coefficient	SE(Coeff)	t-ratio	P-value
Intercept	−525.063	79.25	−6.63	≤0.0001
Calories	2.10223	0.2906	7.23	≤0.0001
Fat	−10.8658	1.721	−6.31	≤0.0001
Cheese	231.335	63.40	3.65	0.0015
Cheese*cals	−0.586007	0.1806	−3.24	0.0039
Reggio's	−66.4706	15.27	−4.35	0.0003
Michelina's	−52.2137	15.20	−3.44	0.0025
Weight W...	28.3265	16.09	1.76	0.0928

II.6: Program week 6: Your free choice of an advanced topic

In this last week of class sessions, we will let you choose the topic of study. We have collected materials for four different topics:

- Quality control
- Nonparametric models
- Decision making and risk
- Data mining.

Quality control is a typical business oriented topic, as may be decision making and risk (but one can imagine applications in other areas). Nonparametric models are relevant for application areas where we cannot just assume all our data is nicely normally distributed. Data mining extends the last chapters of the textbook: large data sets that are analysed by regression analysis methods, typically in an 'automated way', to discover patterns in the data.
Together with choosing your topic, you will also need to choose a case study in that same context. In the reading materials provided with these four topics, case studies are described at the end of each chapter. However, if you can come up with your own case study, that is well appreciated.
We will use the tutorial session in order to discuss both choices: what topic and what case study. We will use the computer lab session to provide support in doing your case study. You will have somewhat more time doing this case study: hand in at the end of Week7. We won't do the feedback round for this last case study, so all 5 credits are for your report.
As with the topics in the first five weeks: the topic of this week is part of the final exam.

II.7: Program week 7: Student project

The Student project for Stats II is of the kind of 'reverse engineering'. The normal case study is one for which you receive a problem set, and based on the characteristics of that problem, you pick an appropriate statistical method to investigate the problem. In our case study, you receive a data set that allows many different issues to be investigated, and a set of statistical methods. That last set consists of:

1. Multiple regression, including correlation analysis and simple regression as intermediate techniques (10 points);

2. Multivariate (2-way) analysis of variance (5 points);

3. Hypothesis testing: means, independent and-or paired samples and-or two proportions (5 points).

For each of these methods, formulate a statistical problem based on the provided data set, provide a reasoning why the method is appropriate for that problem set, and do the statistical analysis. Finalize your investigation with writing a statistical report, with at least 2 pages clean text per problem set (and many more pages with graphs, tables, …). In assessing all your partial case studies, the quality of the statistical investigation (including the check of relevant conditions, doing the proper descriptive analysis before starting the inferential step), and the match between statistical method and problem definition.

In your study, you need several types of categorical data. The data set contains no categorical variables from itself (except for year). However, it is not difficult to create more categorical variables yourself, out of the quantitative variables. E.g., by applying a median split (into 2 categories) or a quartile split (into 4 categories) of any variable.

The provided SPSS files contain regional data very similar in kind as the Better Life data we have been analyzing in the weekly SPSS assignments. There are many more regions than countries, so in that respect

the data is much richer, but less variables (10 and 12, rather than 24). Please take care that the data set contains both country and region data; typically, you would want to exclude the country data from your analysis, when investigating the effects of regions.

Please be certain you pick a problem set that is different from any other student. You will cooperate in the sense that some of you will discuss how to solve the case study and to interpret the outcomes, but cooperation should stop at that point. All students are required to come up with a personal problem set.

II.8: Program week 8: Exam week

Final exam