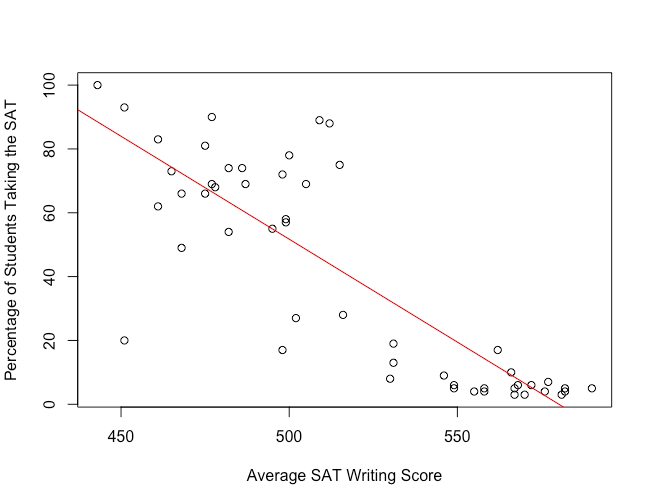
**CHAPTER 5 SOLUTIONS**

* 1. The relationship has a linear component, although from the look of the scatterplot, the relationship appears to be better described by a simple curve rather than a line. For satw averages that are between 520 and 575, almost all of the points are below the fit line, whereas for satw averages that are above 575, all of the points are above the line. The R commands for creating this graph are:

**plot(States$pertak~States$satw, xlab = "Average SAT Writing Score", ylab = "Percentage of Students Taking the SAT")**

**abline(lm(States$pertak~States$satw), col = "red")**

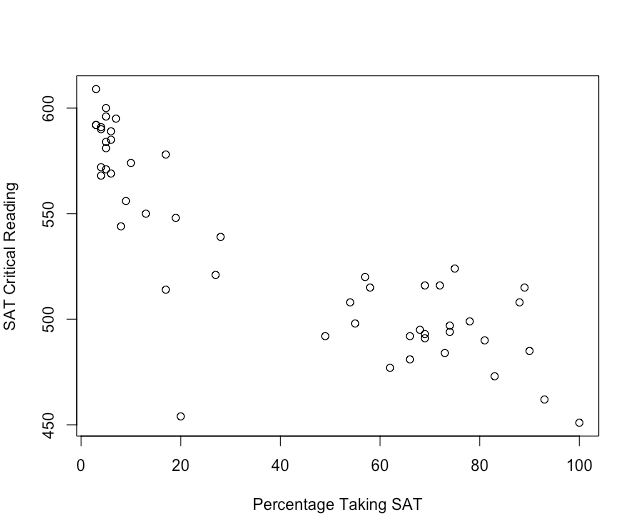
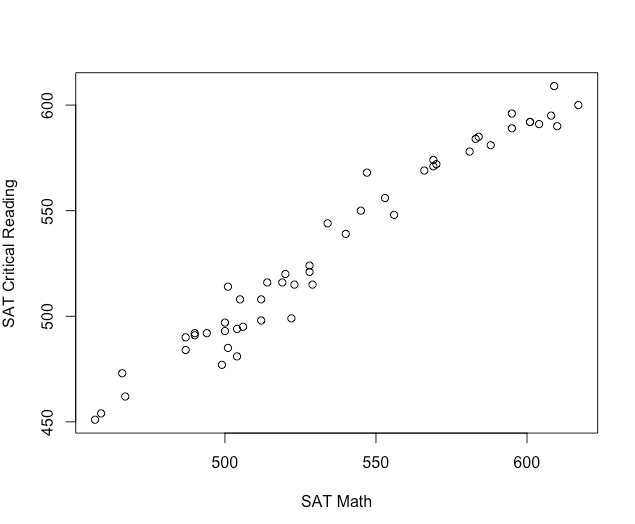


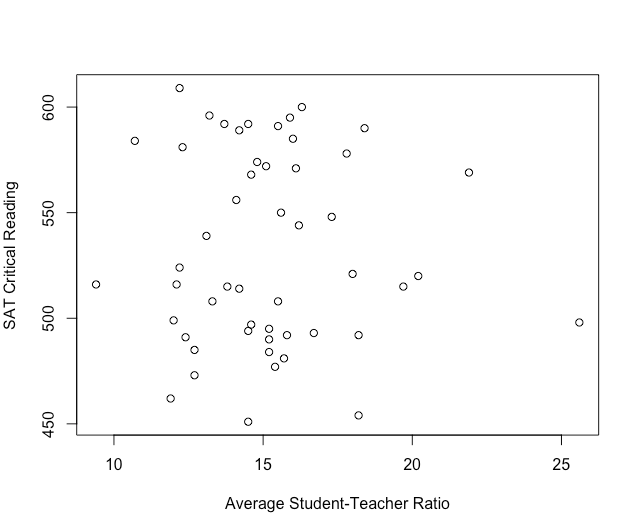
* 1. There appears to be little or no relationship (linear or non-linear) between average score on the SAT Math and average teacher salary because the scatterplot has a cloud like shape and because there is very little difference in the average SAT Math score for different salary levels. In particular, it appears that states that pay their teachers better do not tend to have better scores on the SAT Math, on average.



The R command to generate the first plot is:

**plot(States$satcr~States$satm, xlab = "SAT Math", ylab = "SAT Critical Reading")**

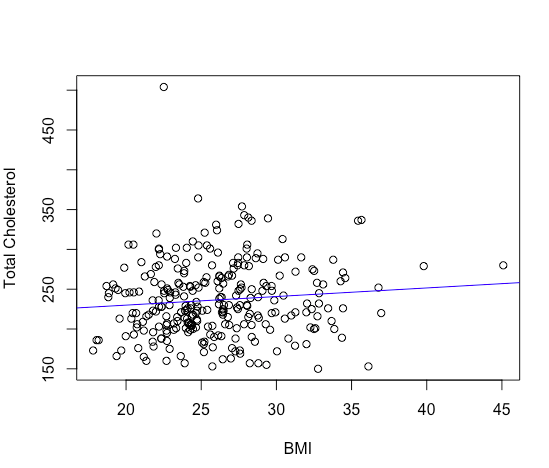




1. The strongest linear relationship is with satm, followed by pertak, with stuteach having the weakest linear relationship to satcr.
2. It appears that states with higher average SAT Critical Reading scores tend also to have higher SAT Math scores; states with lower SAT Critical Reading scores tend also to have lower SAT Math scores.
3. It appears that states with higher average SAT Critical Reading scores tend to have lower percentages of students taking the SAT; states with lower SAT Critical Reading scores tend to have higher percentages of students taking the SAT. In this instance, a simple curve may slightly fit the data better than a line, although the line does capture that part of the relationship that is linear.
4. There appears to be little or no linear relationship between these two variables.
   1. Using the complete cases in the dataset, *r* = .0975, indicating that adults with relatively high body mass index have relatively high total cholesterol levels as well. However, this correlation would be considered weak by Cohen’s (1988) guidelines. The R command used to generate this result is **cor(Framingham$TOTCHOL3, Framingham$BMI3, use = "complete.obs")**.
   2. The scatterplot is created using the commands:

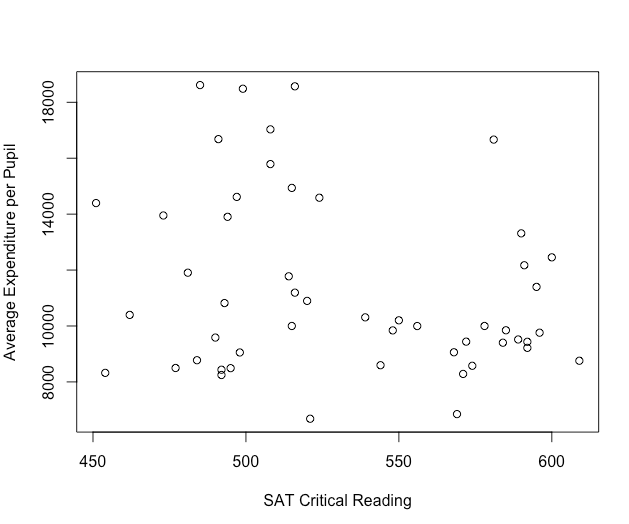
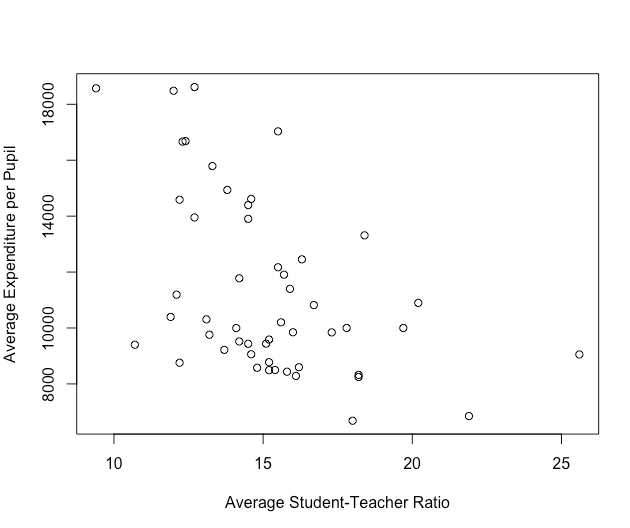
**plot(Framingham$TOTCHOL3~Framingham$BMI3, xlab = "BMI", ylab = "Total Cholesterol")**

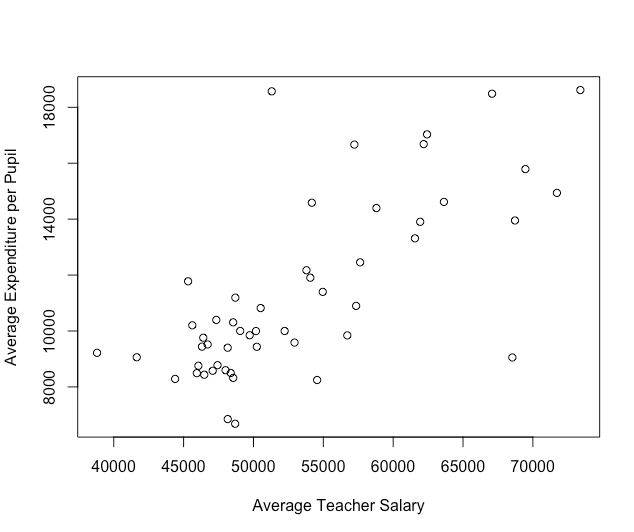
**abline(lm(Framingham$TOTCHOL3~Framingham$BMI3), col="blue")**



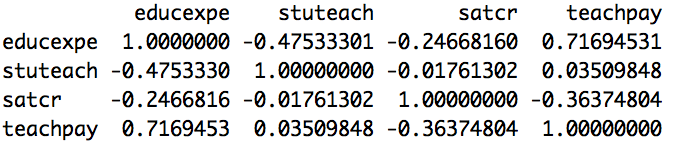
* 1. The most extreme person in the dataset has ID 205. For this person, the total cholesterol is a lot higher than one would expect given the body mass index. The R command to find the ID number of the person with the highest cholesterol is **which(Framingham$TOTCHOL3==max(Framingham$TOTCHOL3, na.rm=T))**.
  2. When omitting ID 205, the correlation is slightly larger at *r* =.1220. This is not unexpected as there are too many data points in this dataset for this one outlier to drastically affect the value of the correlation. Also, as we will see in Chapter 16, because the person is not very unusual in terms of BMI, we do not expect a big change in the correlation when the score is omitted.
  3. A linear model appears to be an appropriate way to represent these relationships in all cases because none of the scatterplots would be better fit by a simple curve. The R command to generate the first plot is:

**plot(States$educexpe~States$stuteach, xlab = "Average Student-Teacher Ratio", ylab = "Average Expenditure per Pupil")**





* 1. The relevant correlations can be found in the first row or first column.

****

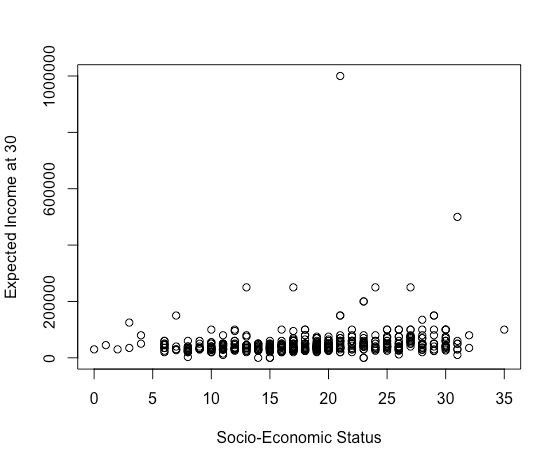
* 1. Yes, since the correlation between these two variables is *r* = .72.
  2. No, since the correlation between these two variables is negative, *r* = -.25, states in which expenditures per pupil are higher tend to have lower SAT Critical Reading scores, on average.
  3. Yes, since the correlation between these two variables is negative, *r* = -.48.
  4. Teacher salary, because this correlation has the greatest magnitude.

a) It is likely that students who have higher socio-economic status tend to have higher math achievement, and so one should expect a positive correlation between these two variables. Using R, the actual correlation is moderate and positive, *r* = .32.

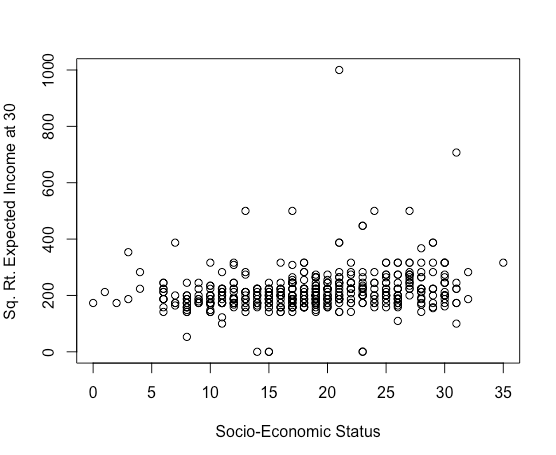
b) It is likely that students who tend to have higher math achievement tend to perceive that their teachers are interested in them, and so one should expect a negative correlation between these two variables. Using R, the actual correlation is weak and negative, *r* = -.18.

c) There is likely to be no relationship between family size and math achievement, and so one should expect a near zero correlation between these variables. Using R, the actual correlation is near zero, *r* = .02.

1. The Pearson correlation between schattrt and slfcnc08 for the students in the NELS dataset is *r* = -.09, indicating that for these data, there is virtually no linear relationship between these two variables.
2. The correlation between schattpp and slfcnc08 remains at *r* = -.09 because in converting from a percentage to a proportion, we multiply by the positive constant, .01. The linear transformation, in this case, does not involve reflection, so the sign of the correlation unchanged.
3. Without the most extreme value of expected income at age 30 of $1,000,000, the pattern of points in the scatterplot indicates a slight positive correlation between ses and expected income. The extreme value at $1,000,000, belonging to someone in the middle range of ses, serves to reduce the systematic positive orientation of points, as indicated by the correlation value, *r* = .16.



1. The new value of the correlation is now *r* = .22, suggesting that the influence of the extreme value of $1,000,000 has been somewhat reduced by the square root transformation. While weak, the correlation suggests that students with lower ses tend to have lower expected incomes, while those with higher ses tend to have higher expected incomes. The R command for generating the square root of expinc30 is **NELS$expincsq = sqrt(NELS$expinc30)**.



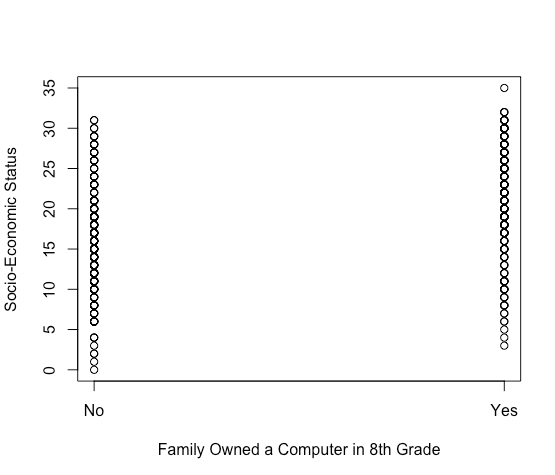
1. The correlation is *r* = .90, which is less than one. The correlation would be equal to one in the case of a linear transformation (without reflection). A square root transformation is nonlinear, and so we expect the correlation between the original and transformed variable to be less than one.
2. *r* = .21
3. *r* = -.35. The command to create the new self-concept variable is **NELS$slfcnclg = log(-1\*NELS$slfcnc12+44)**.
4. The correlation between the transformed variables is stronger than the one between the original variables. By transforming the variables in this case, they have each become more symmetric and the relationship between them more symmetric.
5. The signs are different because the self-esteem variable was reflected.
6. The interpretation remains the same because the self-esteem variable was reflected in the process of it being transformed.
7. According to the scatterplot, the socio-economic status of students whose families owned a computer when they were in eighth grade appears to be higher overall than that of those who did not. That is, the mean ses for those families who owned a computer is higher than the mean of ses for those families who did not own a computer. The scatterplot was created using the R commands:

**plot(NELS$ses~as.numeric(NELS$computer), axes = F, xlab = "Family Owned a Computer in 8th Grade", ylab = "Socio-Economic Status")**

**box()**

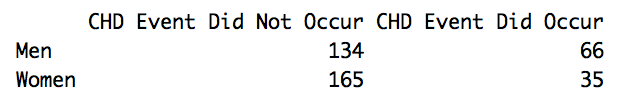
**axis(2)**

**axis(1, at = c(1,2), labels = c("No", "Yes"))**

****

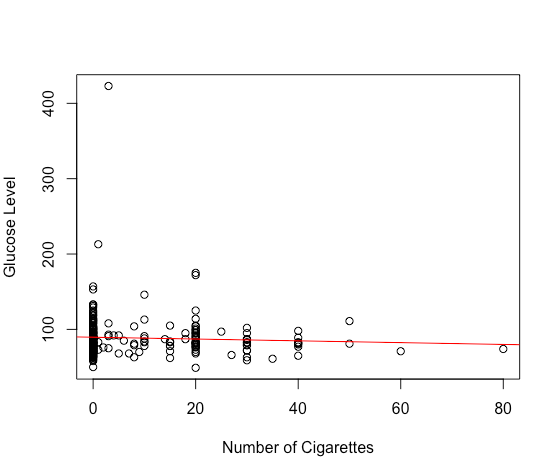
1. As shown by the scatterplot, the correlation between computer and ses is positive and moderate (*r* = .35), indicating that students whose families owned a computer when they were in eighth grade tend to have higher ses than those whose families did not own a computer when they were in eighth grade.
2. There is a small to moderate positive correlation between reading comprehension and intellectual ability, *r* = .29, suggesting that these students with higher reading comprehension scores also tend to have higher intellectual ability and that students with lower reading comprehension scores also tend to have lower intellectual ability.
3. Because the correlation between these two variablesis negative (*r = -*.44), and because resource room students are coded 1 and self-contained classroom students are coded 2, these resource room students, on average, have the higher level of reading comprehension.
4. Because the correlation is so small, *r* = -.07, there is only a very weak linear relationship between type of placement and grade level. That is, the typical grade level is about the same for the two placement types.
5. Because the correlation between grade level and reading comprehension is negative, *r = -*.32, the students in the higher grades tend to perform worse in reading comprehension relative to their peers than those in the lower grades. That is, as these students with learning differences advance in grades, they appear to fall farther behind relative to their peers.
   1. *r* = .18. This relationship is positive, yet weak, suggesting preliminarily that adults taking blood pressure medication have slightly higher blood pressure, on average, than those who do not. The mean (81.00) and median (80.75) initial diastolic blood pressure of those not taking antihypertensive medication is in the normal range. The mean (91.54) and median (92.50) initial diastolic blood pressure of those taking antihypertensive medication is slightly above the normal range.
   2. The point-biserial correlation.
   3. People with higher blood pressure often take such medication to lower their blood pressure. The positive relationship, however weak, suggests either that the medication is not wholly effective in lowering blood pressure to acceptable levels or that it has not yet been taken for long enough to be effective.
   4. The correlation between SEX and ANYCHD4 is *r* = -.18. Because women are coded as higher on SEX and developing CHD is coded higher on ANYCHD4, women are somewhat less likely than men to develop CHD. A similar conclusion is reached using a cross-tabulation approach as noted below. The cross-tabulation was generated using the R command:

**table(Framingham$SEX,Framingham$ANYCHD4)**

****

Men are more likely to develop CHD (66/200 \* 100 = 33%) than women (35/200 \* 100 = 17.5%).

* 1. The phi-coefficient.
  2. Because the scatterplot indicates that the relationship is non-linear. A simple curve would represent the relationship better than a line.



According to the correlations, those with high LDL cholesterol tend to be those who developed CHD during the course of the study (*r* = .30), while those with high HDL tend to be those who did not develop CHD during the course of the study (*r* = -.25).



1. The Pearson correlation between ses and computer for the students in the NELS dataset is *r* = .35, suggesting that students who owned computers in eighth grade have higher ses, on average, than those who do not.
2. **NELS$comp1 = as.numeric(NELS$computer) - 1**

The transformation does not include a reflection of the variable computer because the lower score on comp1 still represents “No”.

1. *r* = 1
2. **NELS$comp2 = -1\*as.numeric(NELS$computer) + 2**

The transformation includes a reflection of the variable computer since we are multiplying computer by -1. As a result, the lower score on comp2 now represents “Yes” when it originally represented “No”.

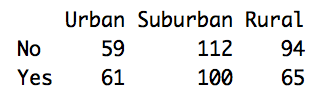
1. *r* = -1
2. The correlation between ses and comp1 is the same as the correlation between ses and computer because the linear transformation of computer did not involve reflection: *r* = .35. The correlation between ses and comp2 is the negative of the correlation between ses and computer because the linear transformation of computer involved reflection: *r* = -.35.
3. The Pearson correlation between ses and slfcnc12 for the students in NELS is *r* = .11. The interpretation is that students with a higher socioeconomic status tend to have a higher self-concept, while students with a lower socioeconomic status tend to have a lower self-concept. However, the relationship is weak.
4. The Pearson correlation between sesdi and slfcnc12di is *r* = .008, indicating a near zero linear relationship between the two variables.

After finding the medians of ses and slfcnc12, we create the respective median splits using the following R commands:

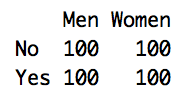
**NELS$sesdi = ifelse(NELS$ses<=19, 1, 2)**

**NELS$slfcnc12di = ifelse(NELS$slfcnc12<=31, 1, 2)**

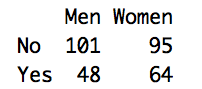
1. In this case, there was a reduction in the magnitude of the correlation based on the original continuous variables. The loss of information when transforming the multi-valued continuous variables into two dichotomous variables, each split at its median, resulted in a near zero correlation. As noted in Chapter 4, other types of non-linear transformations that retain the multi-valued nature of the continuous variable are to be preferred to the median split.
2. According to the value of Spearman’s *rho* (-.20), there is a weak negative relationship between cigarette smoking and participation in extracurricular activities in twelfth grade. In particular, students who had ever tried smoking were less likely to participate in extracurricular activities in twelfth grade.
3. According to the value of Spearman’s *rho* (-.08), there is virtually no relationship between missing school and participation in extracurricular activities in twelfth grade.
4. According to the value of Spearman’s *rho* (.12), there is a weak positive relationship between socioeconomic status and participation in extracurricular activities in twelfth grade.
   1. The contingency table is generated using the R command **table(NELS$advmath8,NELS$urban)**.

****

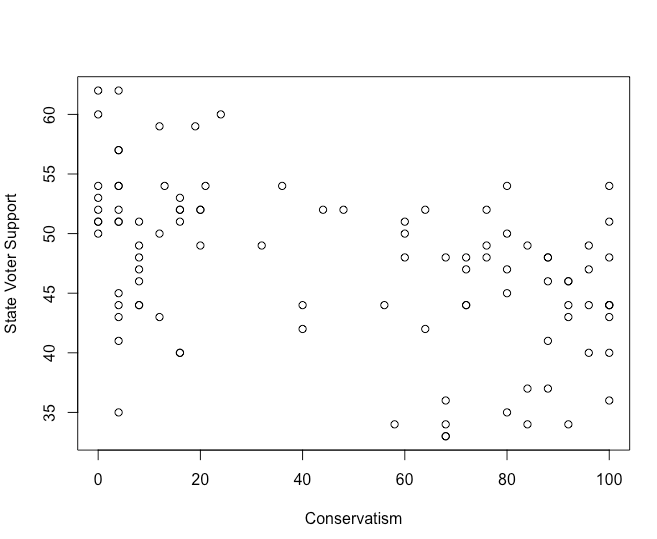
1. Urban: 61/120 = .51; Suburban: 100/212 = .47; Rural: 65/159 = .41
2. 226/491 = .46
3. Urban with .51 advanced math takers.
4. There appears to be some relationship. If there were no relationship, each of the three environments would have the same proportion of students who took advanced math in eighth grade as the overall proportion (.46). Because the proportions given in part (c) are not all equal to .46, there appears to be a relationship. Namely, students from urban environments appear to be most likely to have taken advanced math in eighth grade, followed by those from suburban environments, and then finally those from rural environments.
5. According to the contingency table, there is no relationship whatsoever between these two variables. The proportions of men and women who smoked at time 1 respectively equal the proportion of all individuals who smoked at time 1. Each proportion equals one-half. The sample for the dataset was selected to have this property. The contingency table is generated using the R command **table(Framingham$CURSMOKE1,Framingham$SEX)**.

****

1. According to the contingency table at time 3, after being enrolled in the study for 12 years, women were more slightly more likely to smoke (40.3%) than men (32.2%). Overall, 36.4% were smokers at time 3. The patterns of difference from 36.4% is slight.

****

1. The correlation between the two variables is *r* = 0, as expected.
2. The correlation between the two variables is quite small (*r* = .08), as expected.
   1. There can be more than one correct analysis, although only one is given.
3. The correlation between years of math taken in high school and math achievement in twelfth grade is .42, which suggests that in general, students who take more math tend to score higher on math achievement.
4. According to Spearman’s *rho* (*rho* = .43) students who came to school late also had a tendency to cut or skip classes.
5. Using the crosstabs procedure, we find that 52 percent of public school students did not to take AP classes, while only 38 percent of private religious school students did not take AP classes, and only 30 percent of private non-religious school students did not take AP classes. Thus, public school students appear to be less likely to take AP classes than private school students.
6. The correlation between these two variables is *r* = .02, suggesting that there is virtually no linear relationship between family size and twelfth grade self-concept.
7. The correlation between these two variables is *r* = .32, suggesting that among the students in the NELS dataset, a higher socioeconomic status is associated with a higher science achievement.
8. According to a contingency table analysis using crosstabs, region appears to make a difference as to whether a student took advanced math in eighth grade. While 62 percent of students from the West took advanced math in eighth grade, 48 percent from the South did, 41 percent from the Northeast did, and 38 percent from the North Central region did.
9. According to Spearman’s Correlation, *rho* = -.006, there is virtually no relationship between the number of classes a student cut in twelfth grade and whether or not that student took advanced math in eighth grade.
10. According to a contingency table analysis using crosstabs, there appears to be no relationship between urban and cigarett among students in the NELS dataset. Specifically, students overwhelmingly report that they have not smoked cigarettes, regardless of how urban the area is from which they come. In particular, 85 percent of urban students report not having smoked, 86 percent of suburban students report not having smoked, and 87 percent of the rural students report not having smoked. The uniformity of response across types of area suggests little relationship between these two variables.
11. According to the phi-coefficient, *r* = .27, students that were enrolled in advanced math in eighth grade (advmath8 = 2) tended also to take AP classes (approg = 2) in high school.
12. According to the Spearman Correlation Coefficient, *rho* = .09, there is little or no relationship between the two variables.
13. According to the scatterplot, and as corroborated by a Pearson Product-Moment Correlation Coefficient of *r* = .11, there is a weak positive linear relationship between these two variables. In particular, students with higher self-concept tend to have higher socioeconomic status, while those with lower self-concept tend to have lower socioeconomic status.
14. According to the Point-Biserial Correlation Coefficient, *r* = .38, suggesting that students who attended nursery school (nursery = 2) tend to have a higher socioeconomic status than those who did not (nursery = 1).
15. Because none of the three distributions of ses by urbanicity is severely skewed, this analysis is based on a comparison of means. On average, urban students have the highest ses (mean = 20.33), followed by suburban students (mean = 19.23), and then rural students (mean = 15.94). Because the distributions are not skewed, a comparison of medians yields the same ordering.
16. Because a simple curve does not give a better fit than a line to the points in the scatterplot, it is appropriate to compute the correlation between the two variables.



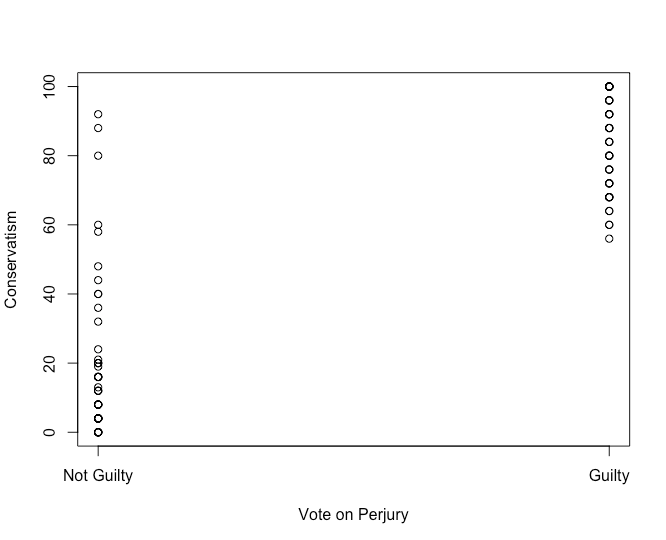
1. The correlation between these two variables is *r* = -.45, a moderate to strong negative correlation, indicating that conservative senators tended to come from states with little voter support for Clinton.
2. Conservative senators were much more likely to vote guilty on perjury. The value of *r* (*r* = .866) is positive and very strong, indicating that high scores on conserva are associated with a high score on vote1, which represents a guilty vote.
3. Senators from states that supported Clinton in 1996 were more likely to vote not guilty on perjury. The value of *r* (*r = -*.429) is negative and moderate to strong, indicating that high scores on supportc are associated with a low score on vote1, which represents a not guilty vote.
4. First-term senators were slightly more likely to vote guilty on perjury. The value of *r* (*r* = .203) is positive, yet weak, indicating that a high score on newbie (the senator is first-term) has a slight tendency to be associated with a high score on vote1 (guilty on perjury).
5. conserva. Its correlation with vote1 is the strongest of the three.
6. Because region is nominal with more than two categories, it has no inherent ordering, a requirement when computing a correlation. In this case, a more appropriate analysis would be based on a contingency table.
   1. The R commands used to create the scatterplot for the vote on perjury is given below. A similar set of commands were used to produce the scatterplot for the vote on obstruction of justice.

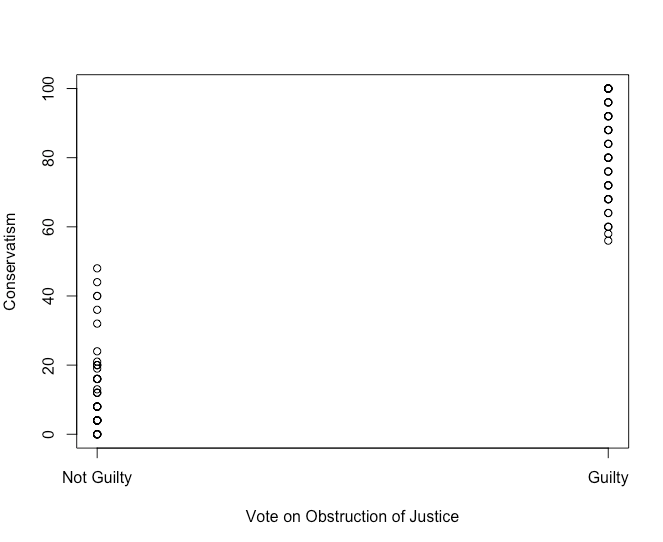
**plot(Impeach$conserva~as.numeric(Impeach$vote1), axes = F, xlab = "Vote on Perjury", ylab = "Conservatism")**

**box()**

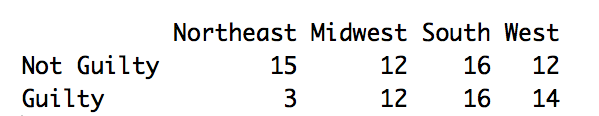
**axis(2)**

**axis(1, at = c(1,2), labels = c("Not Guilty", "Guilty"))**

****

****

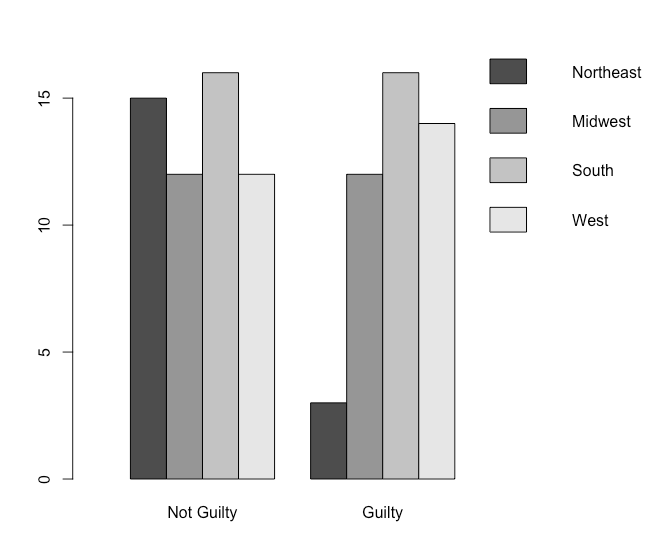
1. Those who voted not guilty on obstruction of justice tended to be less conservative, and those who voted guilty on obstruction of justice tended to be more conservative, on average, *r* = .94 .
2. The vote was along party lines. Democrats tend to be less conservative and tended to support the democratic president by voting not guilty.
3. The correlation between conservatism and the vote on obstruction of justice is the stronger of the two. Every senator with a conservatism score below 50 voted not guilty on obstruction of justice, while every senator with a conservatism score above 50 voted guilty. This clear pattern of separation between voting groups is not evident in the vote on perjury.
   1. The R command used to create the contingency table is **table(Impeach$vote1, Impeach$region)**.

****

1. 55.
2. 55 percent.
3. 29 percent.
4. 50 percent.
5. Yes. Senators from the Northeast had a great tendency to vote not guilty (83.33 percent voted not guilty), while those from the other regions were much more evenly split on guilty/not guilty. In the Midwest, South, and West, 50, 50, and 46.15 percent of the senators, respectively, voted not guilty.
6. Clustered bar graph. The R command to create the clustered bar graph is:

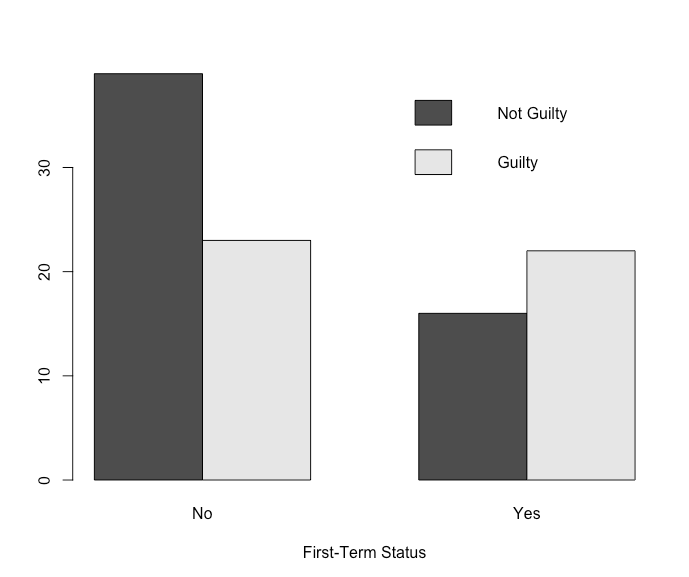
**barplot(table(Impeach$region, Impeach$vote1), beside = T, xlim=c(0,15), legend.text = c("Northeast", "Midwest", "South", "West"), args.legend = list(x = 20, y = 18, bty = "n"))**

Note that the x-axis was extended using the **xlim** argument to make more room for the legend.



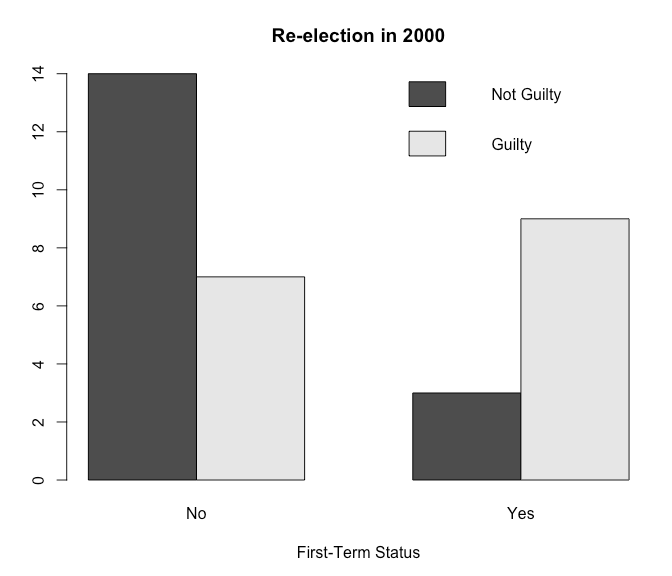
1. A contingency table is appropriate for variables with a small number of values. While region is fine with only four categories, conserva is not. A contingency table that included the variable conserva would probably stretch over more than a page.
   1. The R command to create the clustered bar graph is:

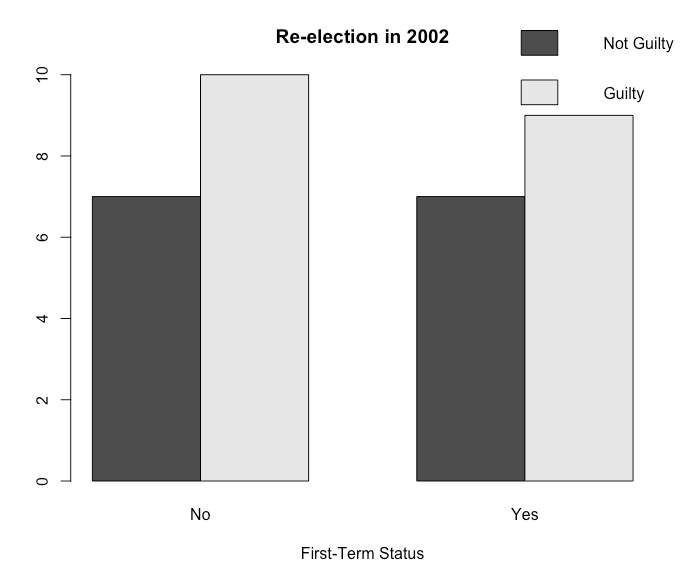
**barplot(table(Impeach$vote1, Impeach$newbie), beside = T, xlab = "First-Term Status", legend.text = c("Not Guilty", "Guilty"), args.legend = list(x = 7, y = 40, bty = "n"))**



1. Yes. Not guilty votes tend to come from senior senators while guilty votes are more evenly split between first-term and senior senators.
2. Positive. Low scores on vote1 (1 = Not Guilty) correspond to low scores on newbie (1 = No, indicating a senior senator).
   1. The graph for re-election in 2000 can be generated using the R command shown below. This can be easily adapted to produce the graph for re-election in 2002.

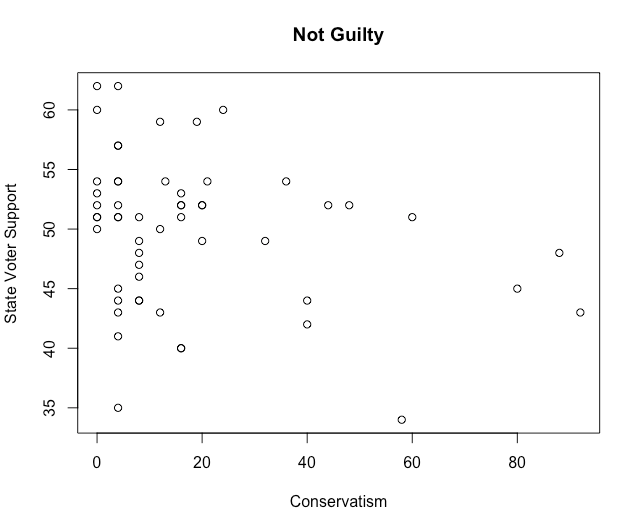
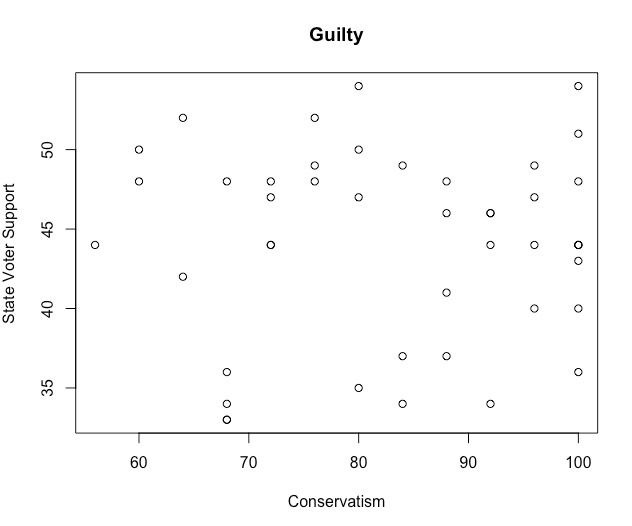
**barplot(table(Impeach$vote2[Impeach$reelect==2000], Impeach$newbie[Impeach$reelect==2000]), beside = T, xlab = "First-Term Status"** **, main = "Re-election in 2000", legend.text = c("Not Guilty", "Guilty"), args.legend = list(x = 7, y = 15, bty = "n"))**



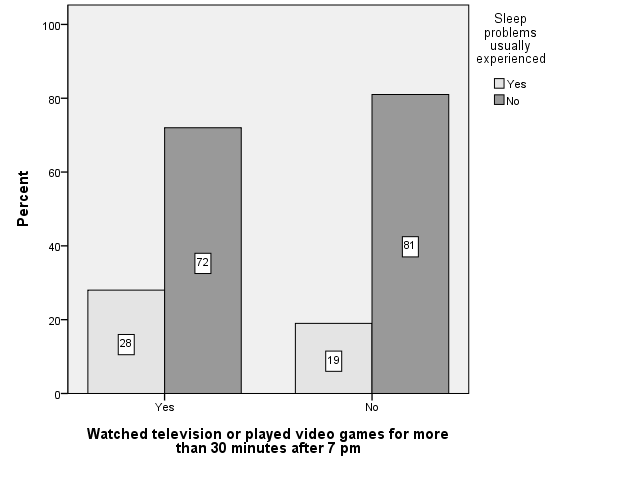


1. There is a relationship. First-term senators tended to vote guilty while more experienced, non-first-term senators tended to vote not guilty.
2. Positive. First-term senators scored 2 (relatively high) on newbie and tended to score 2 (relatively high) on vote2. More experienced senators scored 1 (relatively low) on newbie and tended to score 1 (relatively low) on vote2.
3. It is stronger for senators up for re-election in 2000. Among senators up for re-election in 2002, there is little or no relationship between whether or not the senator was first term and his or her vote on obstruction of justice because both first-term and experienced senators tended to vote guilty.
   1. The R command used to create the guilty scatterplot is:

**plot(Impeach$supportc[Impeach$vote1=="Guilty"]~Impeach$conserva[Impeach$vote1=="Guilty"], xlab = "Conservatism", ylab = "State Voter Support", main = "Guilty")**



1. According to the cloud-like circular shape of the points on the scatterplot for the senators who voted guilty, there appears to be little or no relationship between state voter support for Clinton and conservatism among those senators who voted guilty. This graphical impression is corroborated by the value of the correlation. In this case, *r* = .02, indicating little or no relationship between the two variables.
2. Because the points on the scatterplot for senators who voted not guilty may be described by a line with negative slope, it appears that senators from states with high voter support for Clinton who voted not guilty tended to be less conservative. This graphical impression is corroborated by the value of the correlation. In this case, *r* = -.26, indicating that that senators from states with high voter support for Clinton who voted not guilty tended to be less conservative.
   1. The relationship between manual dexterity and age is nonlinear.
   2. The statement is false because the magnitude of the correlation measures the strength of a relationship on an ordinal, not ratio scale. A correct statement is “A Pearson Correlation Coefficient value of *r* = 0.8 represents a stronger linear relationship than a Pearson Correlation Coefficient value of *r* = 0.4.”
3. Not necessarily. While the two variables are related, they may not be causally related, so changing one may not cause the other variable to change. It may be that the relationship between tv-watching/video-game-playing and sleep could be due to a third variable, e.g., parents who allow their children to watch tv, etc. a lot may not want to spend time with their children, creating sleep-reducing anxiety in their children.
4. One possible answer:



1. It is not fair to conclude this. The article excerpt reports percentages, not raw numbers. So if the number of children allowed to watch TV after 7 p.m. is small, the number of those children with sleep problems could still be smaller than the number of children who don’t watch TV after 7 p.m. and have sleep problems.
2. Positive
3. No. Correlation does not imply causation. A confounding factor might be the severity of the injury. People with more serious injuries are the ones who need the painkillers and with or without them might be out of work longer.
   1. The value of *r* ranges between -1 and 1 inclusive. The value *r* = 1.05 is impossible.
   2. Region is a nominal-leveled variable with more than two categories. Correlation is not meaningful when at least one of the variables is nominal with more than two categories, as is the case for region.
   3. The statement is false because reflection, a type of linear transformation, of one of the variables (for example, multiplying the variable *X* by -1) will change the correlation to *r* = -.50.
   4. The correlation does not provide information about the level of the two variables (how they compare), only about the relationship between them (how they are associated).
   5. Nonlinear relationships are not measured well by the Pearson Product-Moment Correlation Coefficient since this correlation coefficient is designed to measure linear relationships only. An example of such a scatterplot is the following:



Group 1: c

Group 2: d

Group 3: e

Group 4: f