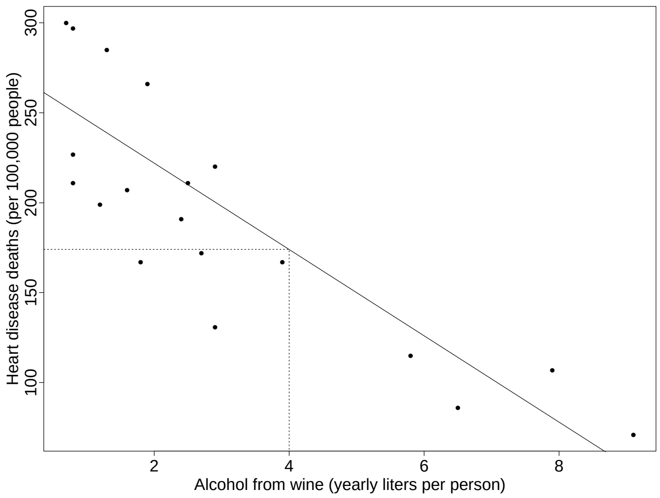
36-201 Spring 1999
Solutions to Homework 4

1. Moore, 5.55 (p.346).

(a)-(b) In this data set, Wine Consumption is the explanatory variable for the number of heart disease deaths. Then, the scatter plot of the data is



We drew a regression line by eye in the scatter plot.

(c) Reading from the plot, a consumption of 4 liters of alcohol from wine each year would yield to approximately 175 heart disease deaths per 100,000 people in that country.

(d) We check that the correlation between these two variables is $r = -0.843$. The correlation is negative because as the consumption of alcohol from wine increases, the number of heart disease deaths decreases, so there is an inverse relationship. About $(-0.843)^2 \times 100\% = 71\%$ of the variation in heart disease rates is explained by the straight-line relationship with wine consumption.

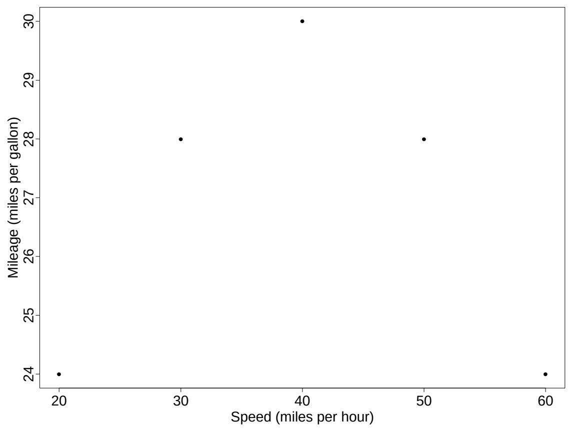
Moore, 5.57 (p.347). We are told that the least-squares equation is $y = 260.56 - 22.969\ x$, where x represents alcohol from wine and y represents number of heart disease deaths. So we will predict that a country with 4 yearly liters of alcohol consumption will have a heart disease death rate of

$$260.56 - 22.969 \times 4 = 169 \text{ deaths}$$

per 100,000 people.

The strong negative correlation that we found tells us that those countries where there is a higher consumption of alcohol per person tend to have lower deaths rates due to heart disease. However there may be other explanatory variables to be taken into account (as life style or others) to conclude causality.

2. Moore, 5.25 (p.316). The plot of Mileage vs. Speed is



The correlation coefficient is

$$r = \frac{1}{n-1} \text{sum} \, \frac{(x_i - \bar{x})}{S_x} \, \frac{(y_i - \bar{y})}{S_y}$$

.

Speed	Mileage	$\bar{x}$	$\bar{y}$	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})\ (y_i - \bar{y})$
20	24	40	26.8	-20	-2.80	56
30	28	40	26.8	-10	1.20	-12
40	30	40	26.8	0	3.20	0
50	28	40	26.8	10	1.20	12
60	24	40	26.8	20	- 2.80	-56
					sum	0.0

So the correlation coefficient is

$$r = \frac{1}{n-1} \, \frac{0}{S_x \, S_y} = 0.$$

Although there is a strong association between *Mileage* and *Speed*, their correlation is zero because their association is not linear (seems to be quadratic).

Moore, 5.63 (p.350).

We know that class attendance explains 16% of the variation of grades. That is, $r^2 \times 100\% = 16\%$. We also know that the correlation is positive because students who attend higher percent of their classes earn higher grades. So the correlation between class attendance and grades is $r = +\sqrt{16/100} = 0.4$.

3. Moore, 5.27 (p.317-318).

(a) The plot shows a strong positive correlation, so it may describe a correlation of $r = 0.9$.

(b) The data looks like scattered at random, so we think that it describes $r = 0$.

(c) The data shows a mild positive correlation, so it may represent a correlation of $r = 0.7$.

(d) This data shows a very mild negative correlation, so we may think of $r = -0.3$.

(e) The plot shows a strong negative correlation, so it may describe a correlation of $r = -0.9$.

Moore, 5.34 (p.320).

(a) *Sex* is a qualitative variable so we cannot measure linear correlation. We may say instead "There is a strong *association* between *Sex* and *Income* of American workers".

(b) The reported correlation is $r = 1.21 > 1$, but correlations cannot be greater than 1.

(c) We may need more information about other variables affecting the yield. Were all the fields exposed to the same conditions (water, temperature, etc)? We can conclude association between the *Amount of Fertilizer* and *Yield*, but not causality.

4. Moore, 5.11 (p.297).

(a) To understand this data we need to know the definition of "Poverty Level".

(b) The total number of people below poverty level in 1993 was

$$9,752+5,125+850+3,274+1,264+316+10,261+3,785+881+2,939+702+114 = 39,263$$

thousands of people.

(c) The total number of people below the poverty level that were 65 or older was $2,939 + 702 + 114 = 3,755$ thousands of people. So the percent is

$$\frac{3,755}{39,263} \times 100\% = 9.56\%.$$

(d) The total number of black people below the poverty level was $5,125 + 1,264 + 3,785 + 702 = 10,876$ thousands of people. So the percent is

$$\frac{10,876}{39,263} \times 100\% = 27.7\%.$$

(e) There were $9,752 + 3,274 + 10,261 + 2,939 = 26,226$ thousands of whites below the poverty level. Among them, there were 2,939 thousands of people 65 years or older, so the percent is

$$\frac{2,939}{26,226} \times 100\% = 11.2\%.$$

(f) There were $9,752 + 5,125 + 850 = 15,727$ thousands of children under 18 below the poverty level. Among them, there were $5,125$ thousands of black children. So the percent is

$$\frac{5,125}{15,727} \times 100\% = 32.6\%.$$

(g) We cannot read what percent of all the people 65 and older were below the poverty level because we are not given the total population of people 65 and older.

Moore, 5.12 (p.297-298).

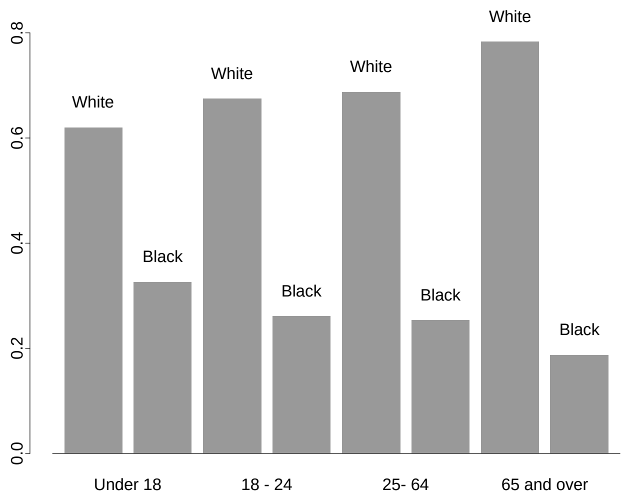
(a) The book does not state this question clearly. We can either compute row percentages that is, Race percents within Age Group, or column percentages that is, Age Group percentages within Race. In the first case, we compute the percentage of Whites in each each Age Group as the ratio between number of White people in that Age Group and the total number of people in the Age Group. Similarly for the percentages of Black people.

	White		Black		Total	
Age Group	#	%	#	%	#	%
Under 18	9,752	62%	5,125	33%	15,727	100%
18 to 24 years	3,274	67%	1,264	26%	4,854	100%
25 to 64 years	10,261	69%	3,785	25%	14,927	100%
65 years and over	2,939	78%	702	19%	3,755	100%

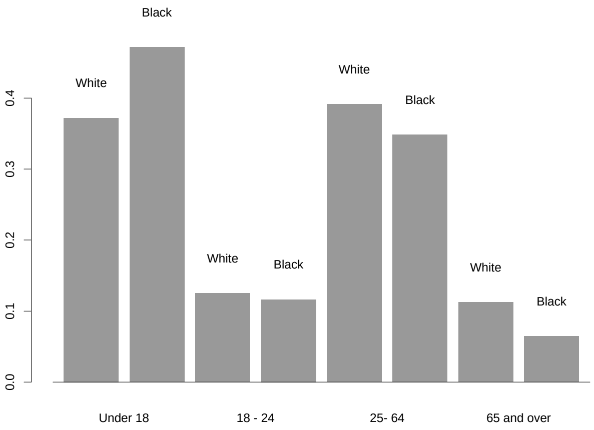
In the second case, we compute the percentages of each Age Group within Whites as the ratio between the number of White people in that Age Group and the total number of White people in the sample. Similarly for Black people. We get

	White		Black	
Age Group	#	%	#	%
Under 18	9,752	37%	5,125	47%
18 to 24 years	3,274	13%	1,264	12%
25 to 64 years	10,261	39%	3,785	35%
65 years and over	2,939	11%	702	6%
Total	26,226	100%	10,876	100%

(b) The bar graphs of the percentage of White and Black people by Age Group is as follows



The bar graphs of Age Group percentages, by Race is as follows



(c) From the last chart, we can see that nearly half of under-poverty-level Blacks are in the Age Group "Under 18" - about 10 percentage points greater than the number for Whites. The percentages for Blacks in Age Groups "25 - 64" and "65 and Older" are each about 5 points lower than Whites' numbers.

5. `Table 5-3 p.297 in Moore`.

(a) The table of expected cell frequencies is: `(row*col/total)`

Age Group	White	Black	Other
Under 18	10,505	4,356	866
18 to 24	3,242	1,345	267
25 to 64	9,971	4,135	822
65 and over	2,508	1,040	207

(b) The table of standardized residuals is: `((O - E)/sqrt(E))`

Age Group	White	Black	Other
Under 18	-7.35	11.64	-0.53
18 to 24	0.56	-2.20	2.99
25 to 64	2.91	-5.44	2.07
65 and over	8.60	-10.48	-6.45

(c) If we look at the column labeled "White" we see an increasing pattern in the residuals. That means that there are less White people that expected in the lower age categories while more White people than expected in the upper age categories. The opposite occurs in the column labeled "Black". The conclusion is that the percentage of White increases as Age increases, while the percentage of Black decreases as Age increases, which is consistent with our findings in `Problem 5.12 (c)`.