1. (a) (3 pts) A certain set of numerical data has average $\bar{x} = 30$ and standard deviation $SD_x = 4$. Assuming that the data has an approximately normal distribution, what percentage of the data lies between 30 and 38?

If $30 \leq x_j \leq 38$, then $0 \leq z_j = (x_j - 30)/4 \leq 2$. The standard units for this data follow the normal curve (approximately), so the percentage of the data lying between 30 and 38 is approximately equal to the area under the normal curve between 0 and 2, which is half the area under the normal curve between $-2$ and 2. I.e., the percentage we want is approximately $95%/2 = 47.5%$.

(b) (3 pts) What percentage of the data from part (a) lies below 26?

If $x_j \leq 26$, then $z_j = (x_j - 30)/4 \leq -1$ so as above, the percentage of the data that lies below 26 is approximately equal to the area under the normal curve below $-1$. The area under the normal curve between $-1$ and 1 is about 68%, so the area between $-1$ and 0 is about 34% (half of 68%). This means that percentage of the data that lies below 26 is approximately is $50% - 34% = 16%$.

(c) (2 pts) A different set of data has average $\bar{y} = 15$, standard deviation $SD_y = 5$ and the smallest value in the data is $y_{min} = 11$. Does this data have an (approximately) normal distribution? Answer yes or no and explain your choice briefly.

No, this data does not have an approximately normal distribution. The smallest value in the data is $0.8 = 4/5$ standard deviations below average, which means that 0% of the data lies more than 1 standard deviation below average. But in a normally distributed data set, approximately 16% of the data lies more than 1 standard deviation below average, as we saw in (b).

2. (a) (3 pts) For a representative sample of cars, would the correlation between the income of the car’s owner ($1000s$) and the car’s fuel efficiency (miles-per-gallon) be positive or negative? Why?

The correlation will be positive because people with higher incomes are more likely to have newer cars than people with lower incomes, and newer cars tend to be more fuel efficient than older cars. I.e., the fuel efficiency of the car tends to rise with the income of the car’s owner.

Comment: Some students suggested that the correlation would be negative because (very) rich people would be buying luxury cars that are not fuel efficient (e.g., Ferraris and Bentleys), and furthermore that these people would not be concerned with saving money on gas. This might be true for (some of) the very rich, but these people represent a tiny fraction of the population. For most of the people, a (somewhat) higher income allows them to buy a newer car, not a luxury car, and saving money on gas is still a priority.

(b) (3 pts) Cross-sectional data from the 2005 Current Population Survey was used to summarize the relationship between age and educational level of U.S. women age 25 and above with the following statistics.

average age $\approx 50$ years, $SD_{age} \approx 16$ years
average ed. level $\approx 13.2$ years, $SD_{ed} \approx 3$ years, $r \approx -0.2$

Does this mean that as women age, they become less educated? If not, what accounts for the negative correlation?

No. Neither women nor men become less educated as they age, if only because ‘education’ is measured here in years of schooling. This study was cross-sectional and included women of many different ages. The negative correlation is explained by the fact that the younger women in the study tended to have more years of schooling on average than the older women, due largely to generational differences. E.g., women born in the 1940s and 1950s were less likely to attend college than women born in the 1970s and 1980s.
3. In a study of the stability of IQ scores, a large group of individuals is tested at age 18 and again at age 35. The following results are obtained.

<table>
<thead>
<tr>
<th>Age</th>
<th>Average Score</th>
<th>SD</th>
</tr>
</thead>
<tbody>
<tr>
<td>18</td>
<td>~100</td>
<td>~15</td>
</tr>
<tr>
<td>35</td>
<td>~100</td>
<td>~15</td>
</tr>
</tbody>
</table>

A score of 120 at age 18 is $(120 - 100)/15 = 4/3$ standard deviations above average. The predicted average IQ score at 35 of people who scored 120 at 18 is therefore $0.6 \times (4/3)$ standard deviations above the age-35 average, which is

$$100 + 0.6 \times \frac{4}{3} \times 15 = 112.$$ 

Alternatively, the regression equation that predicts IQ at 35 from IQ at 18 is given by

$$\hat{IQ}_{35} = \beta_0 + \beta_1 IQ_{18},$$

where

$$\beta_1 = r \cdot \frac{SD_{IQ_{35}}}{SD_{IQ_{18}}} = 0.6 \cdot \frac{15}{15} = 0.6 \quad \text{and} \quad \beta_0 = \overline{IQ}_{35} - \beta_1 \cdot \overline{IQ}_{18} = 100 - 0.6 \cdot 100 = 40.$$

This gives the predicted average IQ at 35 of people who scored 120 at age 18 as

$$\hat{IQ}_{35}(120) = 40 + 0.6 \cdot 120 = 112.$$ 

(b) (4 pts) Predict the score at age 35 of an individual who scores 110 at age 18. Include a margin of error (so your answer should look like this: score $\approx A \pm B$. for appropriate $A$ and $B$).

The predicted average score at 35 of people who score 110 at 18 is

$$\hat{IQ}_{35}(110) = 40 + 0.6 \cdot 110 = 106.$$ 

The predicted score at 35 of an individual who scored 110 at age 18 is predicted to be the average of all like people (namely 106) with a margin of error equal to the R.M.S. error of regression, which is equal to

$$\sqrt{1 - r^2} \times SD_{IQ_{135}} = \sqrt{1 - 0.36} \times 15 = 0.8 \times 15 = 12.$$ 

I.e., the predicted score at 35 of an individual who scored 110 at age 18 is predicted to be 106 ± 12.

Comment: This estimate for the margin of error assumes that the IQ data is homoskedastic.