1 Here is a graph of the kids feet data showing a model of footwidth as a function of footlength and sex. Both the length and width variables are measured in cm.

The model values are solid symbols, the measured data are hollow symbols.

Judging from the graph, what is the model value for a boy with a footlength of 22 cm?

- A 8.0cm
- B 8.5cm
- C 9.0cm
- D 9.5cm
- E Can’t tell from this graph.

According to the model, after adjusting for the difference in foot length, what is the typical difference between the width of a boy’s foot and a girl’s foot?

- A no difference
- B 0.25cm
- C 0.50cm
- D 0.75cm
- E 1.00cm
- F Can’t tell from this graph.

Judging from the graph, what is a typical size of a residual from the model?

- A 0.10cm
- B 0.50cm
- C 1.00cm
- D 1.50cm
- E Can’t tell from this graph.

2 The graphs below show models values for different models of the Old Faithful geyser, located in Yellowstone National Park in the US. The geyser blows water and steam high in the air in periodic eruptions. These eruptions are fairly regularly spaced, but there is still variation in the time that elapses from one eruption to the next.

The variables are

- waiting The time from the previous eruption to the current one
- duration The duration of the previous eruption
- biggerThan3 A categorical variable constructed from duration, which depicts simply whether the duration was greater or less than 3 minutes.

In each case, judge from the shape of the graph which model is being presented.

- (A) waiting ~ duration
- (B) waiting ~ duration + biggerThan3
- (C) waiting ~ duration*biggerThan3
- (D) waiting ~ biggerThan3
- (E) waiting ~ poly(duration,2)
- (F) waiting ~ poly(duration,2)*biggerThan3
The file \texttt{diamonds.csv} contains several variables relating to diamonds: their price, their weight (in carats), their color (which falls into several classes — D, E, F, G, H, I), and so on. The following several graphs show different models fitted to the data: price is the response variable and weight and color are the explanatory variables.

Which model corresponds to which graph?

a) \texttt{> lm(price ~ carat + color, data = diamonds)}
Which graph?  \text{Graph 1}  \text{Graph 2}  \text{Graph 3}  \text{Graph 4}

b) \texttt{> lm(price ~ carat * color, data = diamonds)}
Which graph?  \text{Graph 1}  \text{Graph 2}  \text{Graph 3}  \text{Graph 4}

c) \texttt{> lm(price ~ poly(carat, 2) + color, data = diamonds)}
Error: object 'diamonds' not found
Which graph?  \text{Graph 1}  \text{Graph 2}  \text{Graph 3}  \text{Graph 4}

d) \texttt{> lm(price ~ poly(carat, 2) * color, data = diamonds)}
Which graph?  \text{Graph 1}  \text{Graph 2}  \text{Graph 3}  \text{Graph 4}
The graph shows data on three variables, score, age, and species. The score and age are quantitative. species is categorical with levels x and y.

```r
> age <- rep(seq(0, 8, length.out = 10), times = 2)
> species <- rep(c("x", "y"), each = 10)
> score = 10 + 2.7 * age + 2 * age^2 - 33 * (species == "x") + 4.7 * age * (species == "x") # + rnorm(20,0,1)
> xyplot(score ~ age, groups = species, type = "p", + scales = list(draw = FALSE), auto.key = TRUE)
```

Explain which of the following models is a plausible candidate to describe the data. (Don’t do any detailed calculations; you can’t because the axes aren’t marked with a scale.) Note speciesx means that the case has a level of x for variable species. For each model explain in what ways it agrees or disagrees with the graphed data.

4. a) > score = 10 - 2.7 * age + 1.3 * speciesx
   b) > score = 10 + 5 * age - 2 * age^2 - 1.3 * speciesx
   c) > score = 10 + 5 * age + 2 * age^2 - 1.3 * speciesx
   d) > score = 10 + 2.7 * age + 2 * age^2 - 1.3 * speciesx + 0.7 * age * speciesx

Enter your answers for all four models here:

5. Here is a news article summarizing a research study by Bingham et al., “Drinking Behavior from High School to Young Adulthood: Differences by College Education,” Alcoholism: Clinical & Experimental Research; Dec. 2005; vol. 29; No. 12.

After reading the article, answer these questions:

   a) The article headline is about “drinking behavior.” Specifically, how are they measuring drinking behavior?
   b) What explanatory variables are being studied?
   c) Are any interactions reported?
   d) Imagine that the study was done using a single numerical indicator of drinking behavior, a number that would be low for people who drink little and don’t binge drink, and would be high for heavy and binge drinkers.
      For a model with this numerical index of drinking behavior as the output, what structure of model is implied by the article?
   e) For the model you wrote down, indicate which coefficients are positive and which negative.

Binge Drinking Is Age-Related Phenomenon

By Katrina Woznicki, MedPage Today Staff Writer December 14, 2005

ANN ARBOR, Mich., Dec. 14 - Animal House notwithstanding, going to college isn’t an excess risk factor for binge drinking any more than being 18 to 24 years old, according to researchers here.

The risks of college drinking may get more publicity, but the college students are just late starters, Raymond Bingham, Ph.D., of the University of Michigan and colleagues reported in the December issue of Alcoholism: Clinical & Experimental Research.

Young adults in the work force or in technical schools are more likely to have started binge drinking in high school and kept it up, they said.

The investigators said the findings indicated that it’s incorrect to assume, as some do, that young adults who don’t attend college are at a lower risk for alcohol misuse than college students.

“The ones who don’t go on to a college education don’t change their at-risk alcohol consumption,” Dr. Bingham said. “They don’t change their binge-drinking and rates of drunkenness.”

In their study comparing young adults who went to college with those who did not, they found that men with only a high school education were 91% more likely to have greater alcohol consumption than college students in high school. Men with only a postsecondary education (such as technical school) were 49% more likely to binge drink compared with college students.
There were similar results with females. Women with only a high school education were 88% more likely to have greater alcohol consumption than college students.

The quantity and frequency of alcohol consumption increased significantly from the time of high school graduation at the 12th grade to age 24 \((p < 0.001)\), investigators reported in the December issue of Alcoholism: Clinical & Experimental Research College students drank, too, but their alcohol use peaked later than their non-college peers. By age 24, there was little difference between the two groups, the research team reported.

“In essence,” said Dr. Bingham, “men and women who did not complete more than a high school education had high alcohol-related risk, as measured by drunkenness and heavy episodic drinking while in the 12th grade, and remained at the same level into young adulthood, while levels for the other groups increased.”

The problem, Dr. Bingham said, is that while it’s easier for clinicians to target college students, a homogenous population conveniently located on concentrated college campuses, providing interventions for at-risk young adults who don’t go on to college is going to be trickier.

“The kids who don’t complete college are everywhere,” Dr. Bingham said. “They’re in the work force, they’re in the military, they’re in technical schools.”

Dr. Bingham and his team surveyed 1,987 young adults who were part of the 1996 Alcohol Misuse Prevention Study. All participants had attended six school districts in southeastern Michigan. They were interviewed when they were in 12th grade and then again at age 24. All were unmarried and had no children at the end of the study. Fifty-one percent were male and 84.3% were Caucasian.

The 1,987 participants were divided into one of three education status groups: high school or less; post-secondary education such as technical or trade school or community college, but not a four-year degree college; and college completion.

The investigators looked at several factors, including quantity and frequency of alcohol consumption, frequency of drunkenness, frequency of binge-drinking, alcohol use at young adulthood, cigarette smoking and marijuana use.

Overall, the men tend to drink more than the women regardless of education status. The study also showed while lesser-educated young adults may have started heavier drinking earlier on, college students quickly caught up.

For example, the frequency of drunkenness increased between 12th grade and age 24 for all groups except for men and women with only a high school education \((p < 0.001)\).

“The general pattern of change was for lower-education groups to have higher levels of drunkenness in the 12th grade, and to remain at nearly the same level while college-completed men and women showed the greatest increases in drunkenness,” the authors wrote.

Lesser-educated young adults also started binge-drinking earlier, but college students, again, caught up. High school-educated women were 27% more likely to binge drink than college women, for example. High-school-educated men were 25% more likely to binge drink than men with post-secondary education.

But binge-drinking frequency increased 21% more for college-educated men than post-secondary educated men. And college women were 48% more likely to have an increase in binge-drinking frequency than high school-educated women.

The study also found post-secondary educated men had the highest frequency of drunken-driving. High school educated men and women reported the highest frequencies of smoking in the 12th grade and at age 24 and also showed the greater increase in smoking prevalence over this period whereas college-educated men and women had the lowest levels of smoking.

Then at age 24, the investigators compared those who were students to those who were working and found those who were working were 1.5 times more likely to binge drink \((p < 0.003)\), 1.3 times more likely to be in the high drunkenness group \((p < 0.018)\), and were 1.5 times more likely to have a greater quantity and frequency of alcohol consumption \((p < 0.005)\).

“The transition from being a student to working, and the transition from residing with one’s family of origin to another location could both partially explain differences in patterns,” the authors wrote.

Dr. Bingham said the findings reveal that non-college attending young adults “experience levels of risk that equal those of their college-graduating age mates.”

Consider this model of a child’s height as a function of the father’s height, the mother’s height, and the sex of the child:

\[
\text{height} \sim \text{father*sex} + \text{mother*sex}
\]

Use the Galton data galton.csv to fit the model and examine the coefficients. Based on the coefficients, answer the following:

a) There are two boys, Bill and Charley. Bill’s father is 1 inch taller than Charley’s father. According to the model, and assuming that their mothers are the same height, how much taller should Bill be than Charley?

- A. They should be the same height.
- B. 0.01 inches
- C. 0.03 inches
- D. 0.31 inches
- E. 0.33 inches
- F. 0.40 inches
- G. 0.41 inches

b) Now imagine that Bill and Charley’s fathers are the same
height, but that Charley’s mother is 1 inch taller than Bill’s mother. According to the model, how much taller should Charley be than Bill?
- A) They should be the same height.
- B) 0.01 inches
- C) 0.03 inches
- D) 0.31 inches
- E) 0.33 inches
- F) 0.40 inches
- G) 0.41 inches

c) Now put the two parts together. Bill’s father is one inch taller than Charley’s, but Charley’s mother is one inch taller than Bill’s. How much taller is Bill than Charley?
- A) They should be the same height.
- B) 0.03 inches
- C) 0.08 inches
- D) 0.13 inches
- E) 0.25 inches

7 The graph shows schematically a possible relationship between used car price, mileage, and the car model year.

Consider the model \( \text{price} \sim \text{mileage} \times \text{year} \).

In your answers, treat \text{year} \ as a simple categorical variable, and use year 2005 as the reference group when thinking about coefficients.

a) What will be the sign of the coefficient on \text{mileage}?
- A) Negative
- B) Zero
- C) Positive
- D) No way to tell from the information given

b) What will be the sign of the coefficient on \text{model year}?
- A) Negative
- B) Zero
- C) Positive
- D) No way to tell from the information given

c) What will be the sign of the interaction coefficient?
- A) Negative
- B) Zero
- C) Positive
- D) There is no interaction coefficient.
- E) No way to tell from the information given

8 The graph shows schematically a hypothesized relationship between how fast a person runs and the person’s age and sex.

Consider the model \( \text{speed} \sim \text{age} \times \text{sex} \). 

a) What will be the sign of the coefficient on \text{age}?
- A) Negative
- B) Zero
- C) Positive
- D) No way to tell, even roughly, from the information given

b) What will be the sign of the coefficient on \text{sex}? (Assume that the \text{sex} \ variable is an indicator for women.)
- A) Negative
- B) Zero
- C) Positive

C) What will be the sign of the interaction coefficient? (Again, assume that the \text{sex} \ variable is an indicator for women.)
- A) Negative
- B) Zero
- C) Positive
- D) There is no interaction coefficient.
- E) No way to tell, even roughly, from the information given
For the simple model $A \sim G$, where $G$ is a categorical variable, the coefficients will be group means. More precisely, there will be an intercept that is the mean of one of the groups and the other coefficients will show how the mean of the other groups each differ from the reference group.

Similarly, when there are two grouping variables, $G$ and $H$, the model $A \sim G + H + G:H$ (which can be abbreviated $A \sim G*H$) will have coefficients that are the group-wise means of the crossed groups. Perhaps “subgroup-wise means” is more appropriate, since there will be a separate mean for each subgroup of $G$ divided along the lines of $H$. The interaction term $G:H$ allows the model to account for the influence of $H$ separately for each level of $G$.

However, the model $A \sim G + H$ does not produce coefficients that are group means. Because no interaction term has been included, this model cannot reflect how the effect of $H$ differs depending on the level of $G$. Instead, the model coefficients reflect the influence of $H$ as if it were the same for all levels of $G$.

To illustrate these different models, consider some simple data.

Suppose that you found in the literature an article about the price of small pine trees (either Red Pine or White Pine) of different heights in standard case/variable format, which would look like this:

<table>
<thead>
<tr>
<th>Case #</th>
<th>Color</th>
<th>Height</th>
<th>Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Red</td>
<td>Short</td>
<td>11</td>
</tr>
<tr>
<td>2</td>
<td>Red</td>
<td>Short</td>
<td>13</td>
</tr>
<tr>
<td>3</td>
<td>White</td>
<td>Tall</td>
<td>37</td>
</tr>
<tr>
<td>4</td>
<td>White</td>
<td>Tall</td>
<td>35</td>
</tr>
</tbody>
</table>

and so on ...

Commonly in published papers, the raw case-by-case data isn’t reported. Rather some summary of the raw data is presented. For example, there might be a summary table like this:

<table>
<thead>
<tr>
<th>SUMMARY TABLE</th>
<th>Mean Price</th>
</tr>
</thead>
<tbody>
<tr>
<td>Color</td>
<td>Both Colors</td>
</tr>
<tr>
<td>Height</td>
<td>Red</td>
</tr>
<tr>
<td>Short</td>
<td>$20$</td>
</tr>
<tr>
<td>Tall</td>
<td></td>
</tr>
<tr>
<td>Both Heights</td>
<td>$16$</td>
</tr>
</tbody>
</table>

The table gives the mean price of a sample of 10 trees in each of the four overall categories (Tall and Red, Tall and White, Short and Red, Short and White). So, the ten Tall and Red pines averaged $20$, the ten Short and White pines averaged $18$, and so on. The margins show averages over larger groups. For instance, the 20 white pines, averaged $26$, while the 20 short pines averaged $15$.

The average price of all 40 trees in the sample was $21$.

Based on the summary table, answer these questions:

a) In the model $\text{price} \sim \text{color}$, which involves the coefficients “intercept” and “colorWhite”, what will be the values of the coefficients?

- Intercept $12$ $15$ $16$ $18$ $20$ $21$ $26$ $27$ $34$
- colorWhite $10$ $8$ $0$ $5$ $8$ $10$

b) In the model $\text{price} \sim \text{height}$, which involves the coefficients “intercept” and “heightTall”, what will be the values of the coefficients?

- Intercept $0$ $4$ $8$ $12$ $15$ $16$ $18$ $20$ $21$ $26$ $27$ $34$
- heightTall $0$ $4$ $8$ $12$ $15$ $16$ $18$ $20$ $21$ $26$ $27$ $34$

c) The model $\text{price} \sim \text{height} \times \text{color}$, with an interaction between height and color, has four coefficients and therefore can produce an exact match to the prices of the four different kinds of trees. But they are in a different format: not just one coefficient for each kind of tree. What are the values of these coefficients from the model? (Hint: Start with the kind of tree that corresponds to the intercept term.)

- Intercept $0$ $4$ $8$ $12$ $16$
- heightTall $0$ $4$ $8$ $12$ $16$
- colorWhite $0$ $4$ $8$ $12$ $16$
- heightTall:colorWhite $0$ $4$ $8$ $10$ $12$ $16$

d) The model $\text{price} \sim \text{height} + \text{color}$ gives these three coefficients:

- Intercept : 10
- heightTall : 12
- colorWhite : 10

It would be hard to figure out these coefficients by hand because they can’t be read off from the summary table of Mean Price.

According to the model, what are the fitted model values for these trees:

- Short Red $10$ $12$ $15$ $16$ $20$ $22$ $32$ $34$
- Short White $10$ $12$ $15$ $16$ $20$ $22$ $32$ $34$
- Tall Red $10$ $12$ $15$ $16$ $20$ $22$ $32$ $34$
- Tall White $10$ $12$ $15$ $16$ $20$ $22$ $32$ $34$

Notice that the fitted model values aren’t a perfect match to the numbers in the table. That’s because a model with three coefficients can’t exactly reproduce a set of four numbers.
In the *SwimRecords* data, the variables are

- **time**: World record time (in seconds)
- **year**: The year in which the record was set
- **sex**: Whether the record is for men or women.

Here are the coefficients from several different fitted models.

```r
> lm(time ~ year, data = SwimRecords)
Call:
  lm(formula = time ~ year, data = SwimRecords)
Coefficients:
(Intercept)       year
    567.24      -0.26

> lm(time ~ year + sex, data = SwimRecords)
Call:
  lm(formula = time ~ year + sex, data = SwimRecords)
Coefficients:
(Intercept)       year       sexM
     555.72       -0.251      -9.798

> lm(time ~ year * sex, data = SwimRecords)
Call:
  lm(formula = time ~ year * sex, data = SwimRecords)
Coefficients:
(Intercept)       year       sexM
     697.30       -0.324     -302.464 year:sexM
          0.150

> lm(time ~ sex, data = SwimRecords)
Call:
  lm(formula = time ~ sex, data = SwimRecords)
Coefficients:
(Intercept)       sexM
     65.2        -10.5
```

For each of the following, pick the appropriate model from the set above and use its coefficients to answer the question.

**a)** How does the world record time typically change from one year to the next for both men and women taken together?

<table>
<thead>
<tr>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>-302.4</td>
</tr>
<tr>
<td>-10.5</td>
</tr>
<tr>
<td>-9.790</td>
</tr>
<tr>
<td>-0.2599</td>
</tr>
<tr>
<td>-0.2515</td>
</tr>
<tr>
<td>-0.324</td>
</tr>
<tr>
<td>-0.174</td>
</tr>
</tbody>
</table>

**b)** How does the world record time change from one year to the next for women only?

<table>
<thead>
<tr>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>-302.4</td>
</tr>
<tr>
<td>-10.5</td>
</tr>
<tr>
<td>-9.790</td>
</tr>
<tr>
<td>-0.2599</td>
</tr>
<tr>
<td>-0.2515</td>
</tr>
<tr>
<td>-0.324</td>
</tr>
<tr>
<td>-0.174</td>
</tr>
</tbody>
</table>

**c)** How does the world record time change from one year to the next for men only?

<table>
<thead>
<tr>
<th>Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>-302.4</td>
</tr>
<tr>
<td>-10.5</td>
</tr>
<tr>
<td>-9.790</td>
</tr>
<tr>
<td>-0.2599</td>
</tr>
<tr>
<td>-0.2515</td>
</tr>
<tr>
<td>-0.324</td>
</tr>
<tr>
<td>-0.174</td>
</tr>
</tbody>
</table>

Here is a report from the New York *Times*:

It has long been said that regular physical activity and better sleep go hand in hand. But only recently have scientists sought to find out precisely to what extent. One extensive study published this year looked for answers by having healthy children wear actigraphs — devices that measure movement — and then seeing whether more movement and activity during the day meant improved sleep at night.

The study found that sleep onset latency — the time it takes to fall asleep once in bed — ranged from as little as roughly 10 minutes for some children to more than 40 minutes for others. But physical activity during the day and sleep onset at night were closely linked: every hour of sedentary activity during the day resulted in an additional three minutes in the time it took to fall asleep at night. And the children who fell asleep faster ultimately slept longer, getting an extra hour of sleep for every 10-minute reduction in the time it took them to drift off. (Anahad O’Connor, Dec. 1, 2009 — the complete article is at [http://www.nytimes.com/2009/12/01/health/01really.html](http://www.nytimes.com/2009/12/01/health/01really.html))

There are two models described here with two different response variables: sleep onset latency and duration of sleep.

**a)** In the model with sleep onset latency as the response variable, what is the explanatory variable?

- [ ] Time to fall asleep.
- [ ] Hours of sedentary activity.
- [ ] Duration of sleep.

**b)** In the model with duration of sleep as the response variable, what is the explanatory variable?

- [ ] Time to fall asleep.
- [ ] Hours of sedentary activity.
- [ ] Duration of sleep.
c) Suppose you are comparing two groups of children. Group A has 3 hour of sedentary activity each day, Group B has 8 hours of sedentary activity. Which of these statements is best supported by the article?

A) The children in Group A will take, on average, 3 minutes less time to fall asleep.
B) The children in Group B will have, on average, 10 minutes less sleep each night.
C) The children in Group A will take, on average, 15 minutes less time to fall asleep.
D) The children in Group B will have, on average, 45 minutes less sleep each night.

d) Again comparing the two groups of children, which of these statements is supported by the article?

A) The children in Group A will get, on average, about an hour and a half hours of extra sleep compared to the Group B children.
B) The children in Group A will get, on average, about 15 minutes more sleep than the Group B children.
C) The two groups will get about the same amount of sleep.

11 In the SAT data sat.csv, the variables have these units:

- sat has units of “points.”
- expend has units of “dollars.”
- ratio has units of “students.”
- frac has units of “percentage points.”

Consider the model formula

\[ sat = 994 + 12.29 \text{expend} - 2.85 \text{frac} \]

What are the units of the coefficient 994?

A) points
B) dollars
C) students
D) percentage points
E) points per dollar
F) students per point
G) points per percentage points

b) What are the units of the coefficient 12.29?

A) points
B) dollars
C) students
D) percentage points
E) points per dollar
F) students per point
G) points per percentage points

12 There is a correspondence between the model formula and the coefficients found when fitting a model.

For each of the following model formulas, tell what the coefficient is:

a) \[ 3 - 7x + 4y + 17z \]

- Intercept: \([-7, 3, 4, 17]\)
- x coef: \([-7, 3, 4, 17]\)
- y coef: \([-7, 3, 4, 17]\)
- z coef: \([-7, 3, 4, 17]\)

b) \[ 1.22 + 0.12age + 0.27educ - 0.04age:educ \]

- Intercept: \([-0.04, 0.12, 0.27, 1.22]\)
- educ coef: \([-0.04, 0.12, 0.27, 1.22]\)
- age coef: \([-0.04, 0.12, 0.27, 1.22]\)
- age:educ coef: \([-0.04, 0.12, 0.27, 1.22]\)

c) \[ 8 + 3\textcolor{Red}{\text{color}} - 4\textcolor{Blue}{\text{color}} \]

- Intercept: \([-4, 3, 8]\)
- colorRed coef: \([-4, 3, 8]\)
- colorBlue coef: \([-4, 3, 8]\)

13 For each of the following coefficient reports, tell what the corresponding model formula is:

<table>
<thead>
<tr>
<th>term</th>
<th>coef</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>10</td>
</tr>
<tr>
<td>x</td>
<td>3</td>
</tr>
<tr>
<td>y</td>
<td>5</td>
</tr>
</tbody>
</table>

a) \[ \text{x + y} \]
B) \[ 1 + x + y \]
C) \[ 10 + 3 + 5 \]
D) \[ 10 + 3x + 5y \]
E) \[ 10x + 5y + 3 \]
b) The model $A \sim \text{color}$ gave these coefficients:

<table>
<thead>
<tr>
<th>term</th>
<th>coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>10</td>
</tr>
<tr>
<td>colorBlue</td>
<td>5</td>
</tr>
<tr>
<td>colorGreen</td>
<td>12</td>
</tr>
</tbody>
</table>

- What is the mean of $A$ for those cases that are Blue: 5 10 12 15 17 22 27 unknown
- What is the mean of $A$ for those cases that are Green: 5 10 12 15 17 22 27 unknown
- What is the mean of $A$ for those cases that are Red: 5 10 12 15 17 22 27 unknown
- What is the grand mean of $A$ for all cases: 5 10 12 15 17 22 27 unknown

b) The model $B \sim \text{color} - 1$ gave these coefficients:

<table>
<thead>
<tr>
<th>term</th>
<th>coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>colorRed</td>
<td>100</td>
</tr>
<tr>
<td>colorBlue</td>
<td>-40</td>
</tr>
<tr>
<td>colorGreen</td>
<td>35</td>
</tr>
</tbody>
</table>

- What is the group mean of $B$ for those cases that are Blue: -40 -5 0 35 60 65 100 135 unknown
- What is the group mean of $B$ for those cases that are Green: -40 -5 0 35 60 65 100 135 unknown
- What is the group mean of $B$ for those cases that are Red: -40 -5 0 35 60 65 100 135 unknown
- What is the grand mean of $B$ for all cases: -40 -5 0 35 60 65 100 135 unknown

c) The model $C \sim 1$ gave this coefficient:

<table>
<thead>
<tr>
<th>term</th>
<th>coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>4.7</td>
</tr>
</tbody>
</table>

- What is the group mean of $C$ for those cases that are Blue: 0.0 4.7 unknown
- What is the grand mean of $C$ for all cases: 0.0 4.7 unknown

15 Using the appropriate data set and correct modeling statements, compute each of these quantities and give the model statement you used (e.g., age $\sim$ sex)

a) From the \texttt{cps.csv} data, what is the mean age of single people? (Pick the closest answer.)

28 31 32 35 39

What was your model expression?

b) From the \texttt{CPS} data, what is the difference between the mean ages of married and single people? (Pick the closest answer.)

- Single people are, on average, 5 years younger.
- Single people are, on average, 5 years older.
- Single people are, on average, 7 years younger.
- Single people are, on average, 7 years older.

What was your model expression?

c) From the \texttt{SwimRecords} data, what is the mean swimming time for women? (Pick the closest.)

55 60 65 70 75 80

What is your model expression?

d) From the \texttt{utilities.csv} data, what is the mean CCF for November? (Pick the closest.) (Hint: use \texttt{as.factor(month)} to convert the month number to a categorical variable.)

-150 -93 42 150 192

What is your model expression?

16 Car prices vary. They vary according to the model of car, the optional features in the car, the geographical location, and the respective bargaining abilities of the buyer and the seller.

In this project, you are going to investigate the influence of at least three variables on the asking price for used cars:

- Model year
- Mileage
- Geographical location
These variables are relatively easy to measure and code. There are web sites that allow us quickly to collect a lot of cases. One site that seems easy to use is www.cars.com. Pick a particular model of car that is of interest to you. Also, pick a few scattered geographical locations. (At www.cars.com you can specify a zip code, and restrict your search to cars within a certain distance of that zip code.)

For each location, use the web site to find prices and the other variables for 50-100 cars. Record these in a spreadsheet with five variables: price, model year, mileage, location, model name. (The model name will be the same for all your data. Recording it in the spreadsheet will help in combining data for different types of cars.) You may also choose to record some other variables of interest to you.

Using your data, build models make a series of claims about the patterns seen in used-car prices. Some basic claims that you should make are in this form:

• Looking just at price versus mileage, the price of car model XXX falls by 12 cents per mile driven.
• Looking just at price versus age, the price of car model XXX falls by 1000 dollars per year of age driven.
• Considering both age and mileage, the price of car model XXX falls by ...
• Looking at price versus location, the price differs ...

You may also want to look at interaction terms, for example whether the effect of mileage is modulated by age or location. Note whether there are outliers in your data and indicate whether these are having a strong influence on the coefficients you find.

### Table

<table>
<thead>
<tr>
<th>Year</th>
<th>Vehicle</th>
<th>Price</th>
<th>Mileage</th>
</tr>
</thead>
<tbody>
<tr>
<td>2006</td>
<td>Mazda Miata MX-5</td>
<td>$23,995</td>
<td>9,989</td>
</tr>
<tr>
<td>2004</td>
<td>Mazda Miata MX-5</td>
<td>$19,977</td>
<td>10,198</td>
</tr>
<tr>
<td>2004</td>
<td>Mazda Miata MX-5</td>
<td>$15,999</td>
<td>20,360</td>
</tr>
<tr>
<td>2002</td>
<td>Mazda Miata MX-5 SE</td>
<td>$15,995</td>
<td>18,826</td>
</tr>
<tr>
<td>2003</td>
<td>Mazda Miata MX-5</td>
<td>$15,995</td>
<td>15,947</td>
</tr>
<tr>
<td>2002</td>
<td>Mazda Miata MX-5</td>
<td>$14,995</td>
<td>44,215</td>
</tr>
<tr>
<td>2002</td>
<td>Mazda Miata MX-5 LS</td>
<td>$14,500</td>
<td>18,000</td>
</tr>
<tr>
<td>2002</td>
<td>Mazda Miata MX-5</td>
<td>$14,500</td>
<td>1,800</td>
</tr>
<tr>
<td>2001</td>
<td>Mazda Miata MX-5</td>
<td>$13,995</td>
<td>70,136</td>
</tr>
<tr>
<td>2002</td>
<td>Mazda Miata MX-5</td>
<td>$12,995</td>
<td>43,978</td>
</tr>
<tr>
<td>1999</td>
<td>Mazda Miata MX-5</td>
<td>$9,500</td>
<td>63,000</td>
</tr>
<tr>
<td>1999</td>
<td>Mazda Miata MX-5</td>
<td>$9,500</td>
<td>63,000</td>
</tr>
<tr>
<td>1991</td>
<td>Mazda Miata MX-5</td>
<td>$4,200</td>
<td>50,000</td>
</tr>
</tbody>
</table>