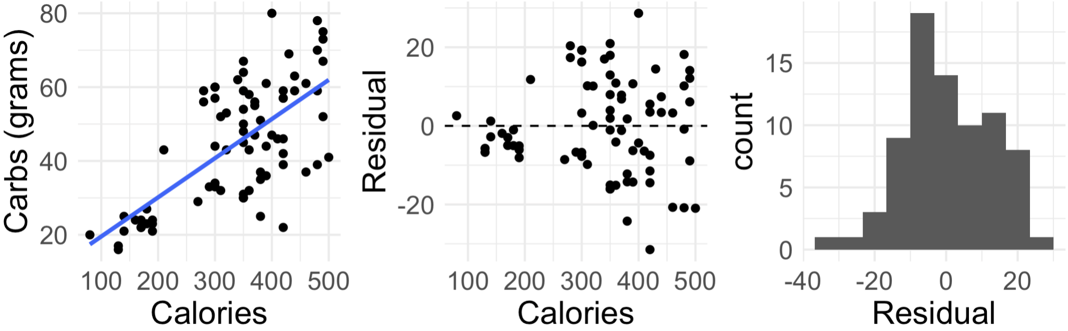
This problem set covers material from Week 10, dates 11/10 – 11/13.

Instructions: Write or type complete solutions to the following problems and submit answers to the corresponding Canvas assignment. Your solutions should be neatly-written, show all work and computations, include figures or graphs where appropriate, and include some written explanation of your method or process (enough that I can understand your reasoning without having to guess or make assumptions). A general rubric for homework problems appears on the final page of this assignment.

Monday 11/10

1. We will re-visit the `starbucks` data from `openintro`. Since Starbucks only lists the number of calories on the display items in stores, we may be interested in predicting the amount of carbs a menu item has based on its calorie content. The scatterplot below shows the relationship between the number of calories and amount of carbohydrates (in grams) Starbucks food menu items contain.



Blue line = least squares line

   (a) Describe the relationship (strength, direction, linearity) between number of calories and amount of carbohydrates (in grams) that Starbucks food menu items contain.

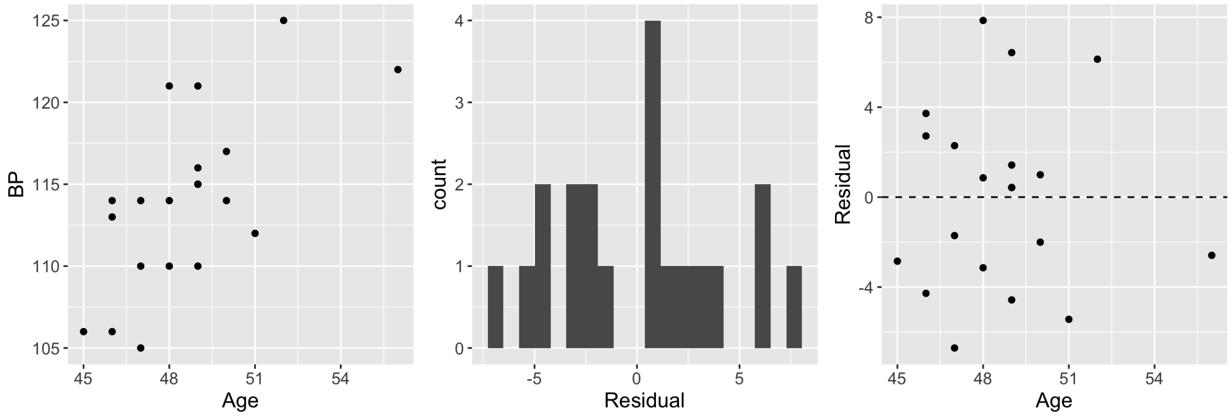
   (b) The least-squares line shown in the first plot is obtained from estimates $b_0 = 8.944$ and $b_1 = 0.106$. Write out two equations *in context*: 1) the linear regression for these data and 2) the fitted model for these data.

   (c) The menu item Morning Bun has 350 calories, 16 grams of fat, and 45 grams of carbohydrates. Based on your model in (c), obtain the residual for the Morning Bun and explain the meaning of this residual value in context.

   (d) Do these data meet the conditions required for fitting a least squares line? Check if each condition is met or not, providing brief justification where appropriate.

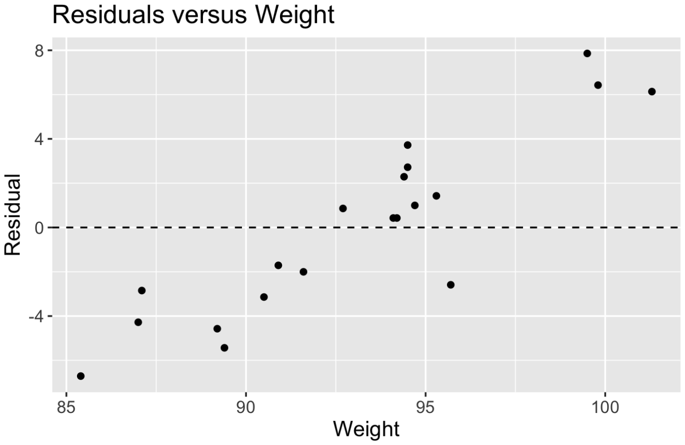
2. A researcher is interested in determining if a person's age and/or weight are good predictors of their diastolic blood pressure (BP) (specifically for individuals with high blood pressure). They have data on 20 randomly sampled men with high BP, and

for each individual they have recorded the BP, Age, and Weight. The researchers fit a simple linear regression model for BP using Age as the explanatory variable. A scatterplot of the data is shown below, as are the histogram of the residuals and the residual plot.



(a) Do these data meet the conditions required for fitting a least squares line? Check if each condition is met or not, providing brief justification where appropriate.

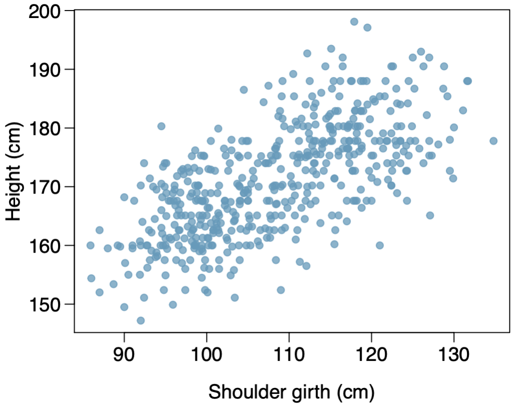
(b) Next, the researchers would like to consider the Weight of each individual as a predictor for BP. They take the residuals from the model above (where Age was the explanatory), and plot those residuals against the individuals' Weight values (shown below). Interpret what this plot is telling you, in context of the data and the fitted model.



(c) Based on your response in (b), do you think it would be worth considering a model that *also* includes Weight as an explanatory variable, in addition to Age (we don't know how to do that yet). Why or why not?

Wednesday 11/12

3. Researchers studying anthropometry collected body girth measurements and skeletal diameter measurements, as well as age, weight, height and gender for 507 physically active individuals. They are interested in the relationship between height (cm) and shoulder girth (cm). They would like create linear regression model for height using shoulder girth as the predictor.



The mean shoulder girth is 107.20 cm with a standard deviation of 10.37 cm. The mean height is 171.14 cm with a standard deviation of 9.41 cm. The correlation between height and shoulder girth is 0.67. We will assume that all conditions of LINE are met.

(a) Write the equation of the fitted regression line for predicting height.

(b) Interpret the slope and intercept in context.

(c) Calculate the R^2 of the regression line and interpret it in context.

(d) A one year old has a shoulder girth of 56 cm. Would it be appropriate to use this linear model to predict the height of this child? If so, obtain the predicted height. If not, explain why not.

4. This exercise explores why fitting a SLR model is perhaps unnecessary when we have a singe categorical predictor variable. We have data about the life expectancy (at birth, in years) in 2007 from a sample of American (both South and North America) and European countries. We'd like to fit a model for life expectancy using the continent that the country belongs to as predicator. Some summary statistics of the data are:

continent	mean	median	s
Americas	73.6081	72.8990	4.4409
Europe	77.6486	78.6085	2.9798

Table 1: Summary statistics of life expectancy.

We fit the linear model, and the output from `R` is as follows. We will pretend we checked the LINE conditions.

term	estimate	std.error	statistic	p.value
(Intercept)	73.6081	0.7427	99.112	0e+00
continentEurope	4.0405	1.0056	4.018	2e-04

Table 2: Fitted model.

(a) Write out the indicator variable that is implied by the table above.

(b) Write out the fitted model using your indicator variable.

(c) Interpret the slope and intercept of the model in context.

(d) In 2007, what was the estimated life expectancy for a country for someone from Europe?

(e) Look again at Tables 1 and 2 above. How exactly do the sample means (and their relationship) compare to the point estimates b_0 and b_1?

(f) Suppose instead that we change the base level to be the other continent, yielding a different indicator. Using what you discovered in (e), write out the equation of the fitted line under this new indicator.

5. We will now see the case where x is a categorical explanatory variable with three levels (this generalizes to the general k-level case). Let x be our explanatory variable with the following three levels: A, B, and C. Without loss of generality, we will let A be the baseline level.

(a) We might think to create the following variable to make a "legal" SLR:

$$x\_new = \begin{cases} 0 & \text{if } x = A \\ 1 & \text{if } x = B \\ 2 & \text{if } x = C \end{cases}$$

And then have the linear regression model:

$$y = \beta_0 + \beta_1 x\_new + \epsilon$$

What assumption does this model make about the how relationship between the different levels of x relate to y? Why is this bad?

(b) Instead, we will make a series of new indicator variables:

$$x_B = \begin{cases} 0 & \text{if } x \neq B \\ 1 & \text{if } x = B \end{cases} \qquad x_C = \begin{cases} 0 & \text{if } x \neq C \\ 1 & \text{if } x = C \end{cases}$$

With these new variables, our new linear regression model and corresponding fitted model are:

$$y = \beta_0 + \beta_1 x_B + \beta_2 x_C + \epsilon$$
$$\hat{y} = b_0 + b_1 x_B + b_2 x_C$$

In this model, what are the estimated responded values if $x = A$? $x = B$? $x = C$?

(c) Given what you learned in (b), provide a general interpretation of b_0, b_1, and b_2 in this model.

Thursday 11/13

6. Problems in the associated `.qmd`. Problems 1-3 will be graded as one problem, as well as 4-5, and 6-8.

General rubric

Points	Criteria
5	The solution is correct *and* well-written. The author leaves no doubt as to why the solution is valid.
4.5	The solution is well-written, and is correct except for some minor arithmetic or calculation mistake.
4	The solution is technically correct, but author has omitted some key justification for why the solution is valid. Alternatively, the solution is well-written, but is missing a small, but essential component.
3	The solution is well-written, but either overlooks a significant component of the problem or makes a significant mistake. Alternatively, in a multi-part problem, a majority of the solutions are correct and well-written, but one part is missing or is significantly incorrect.
2	The solution is either correct but not adequately written, or it is adequately written but overlooks a significant component of the problem or makes a significant mistake.
1	The solution is rudimentary, but contains some relevant ideas. Alternatively, the solution briefly indicates the correct answer, but provides no further justification.
0	Either the solution is missing entirely, or the author makes no non-trivial progress toward a solution (i.e. just writes the statement of the problem and/or restates given information).
Notes:	For problems with multiple parts, the score represents a holistic review of the entire problem. Additionally, half-points may be used if the solution falls between two point values above.
Notes:	For problems with code, well-written means only having lines of code that are necessary to solving the problem, as well as presenting the solution for the reader to easily see. It might also be worth adding comments to your code.