3 Life Expectancy

Note: for this question, you do not need to do the arithmetic (adding, subtracting, multiplying, dividing), but rather write down what you would enter into a calculator if you had one. Let's consider the `gapminder` development data, but only for the year 2007.

```
gapminder2007 <- gapminder %>%
  filter(year == 2007) %>%
  select(country, continent, lifeExp)
```

Let's look at a random sample of 5 out of the 142 rows of this dataset:

country	continent	lifeExp
Namibia	Africa	53
Portugal	Europe	78
Iran	Asia	71
Brazil	Americas	72
Italy	Europe	81

We are interested in modeling the relationship between the outcome variable y = life expectancy in years and the categorical explanatory variable x = continent. You fit a following regression and obtain the following regression table rounded to the nearest integer:

term	estimate	std.error	statistic	p.value
(Intercept)	55	1.0	53.5	0
continentAmericas	19	1.8	10.4	0
continentAsia	16	1.6	9.7	0
continentEurope	23	1.7	13.5	0
continentOceania	26	5.3	4.9	0

a) What is the fitted value $\widehat{y}$ for any given country in:

1. Africa
2. Asia
3. Europe

b) What is the residual for the following three countries?

1. Nambia
2. Iran
3. Italy

c) What is the mean age for countries in the following continents:

1. Africa
2. Asia
3. Europe

1 Short Answer

a) For each of these five regression scenarios, name an appropriate visualization (along with any distinguishing features) that graphically summarizes the relationship between the outcome variable y and the explanatory/predictor variable(s).

1. Simple linear regression with one numerical predictor
2. Simple linear regression with one categorical predictor
3. Multiple regression with two numerical predictors
4. Multiple regression with one categorical and one numerical predictor
5. Multiple regression with two categorical predictors

b) Consider the following hypothetical study. Say you collect two variables of information from a population of interest: y = life expectancy and x = annual income out of college measured in units of thousands of dollars. You find that

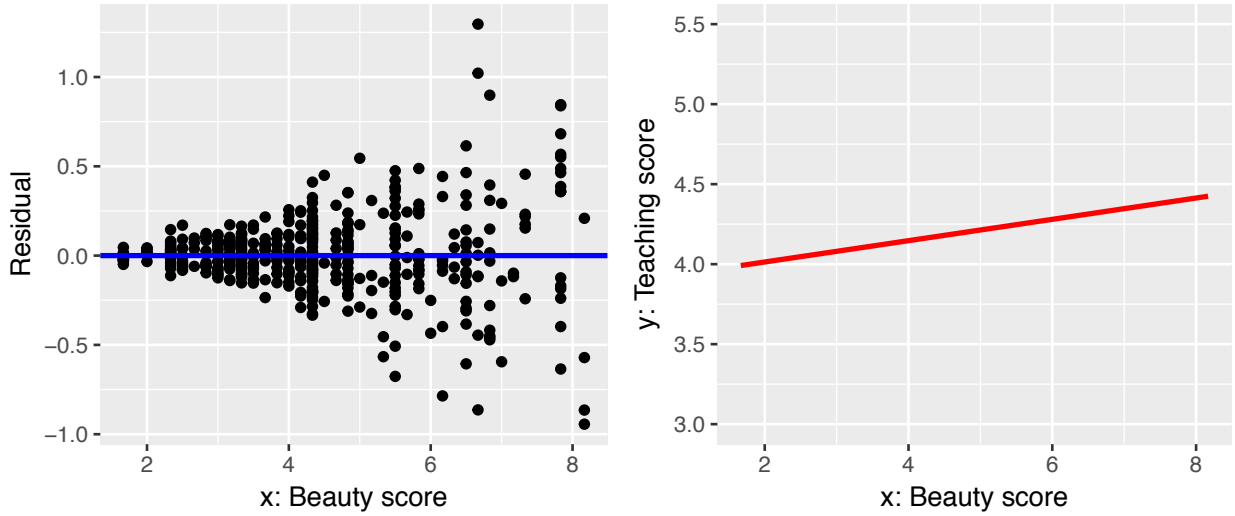
1. the correlation coefficient is 0.25
2. the fitted regression line $\widehat{y} = 45 + 0.5x$

Write down what the following two quantities would be if x was not measured in units of thousands of dollars, but measured in units of dollars:

1. the correlation coefficient
2. the fitted slope b_1 of the regression line $\widehat{y} = b_0 + b_1 x$

c) Name one situation when doing data analysis/modeling where log 10-transformations are useful? Answer in **20 words** or less.

d) Say we perform a regression to model an instructors' teaching score as a function of their beauty score, and obtain the following residual plot on the left. Draw a *rough* sketch of what the scatterplot of x and y would look like given that the red line is the fitted regression line.



e) BONUS What is the name of the situation we have when residuals form the pattern exhibited on the left? Hint: it's one word that's ancient Greek in origin.

2 Teaching evaluations

Recall the teaching evaluation data from class. Let's look at a random sample of 5 out of the 463 rows of this dataset:

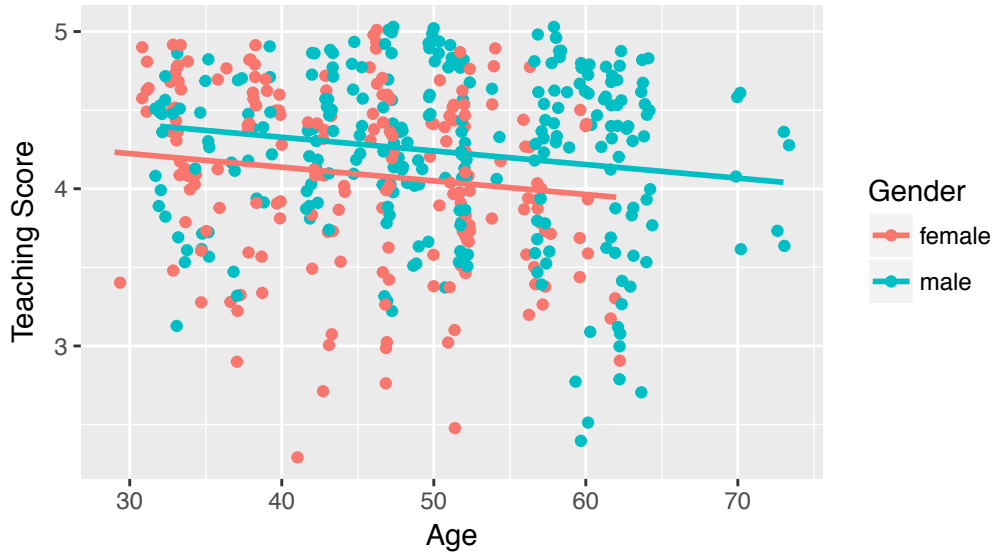
	score	gender	age
341	4.9	male	43
108	5.0	female	46
187	4.3	female	47
206	4.1	male	62
176	4.7	male	39

We are interested in modeling the outcome variable y = teacher evaluation score as a function of two explanatory variables:

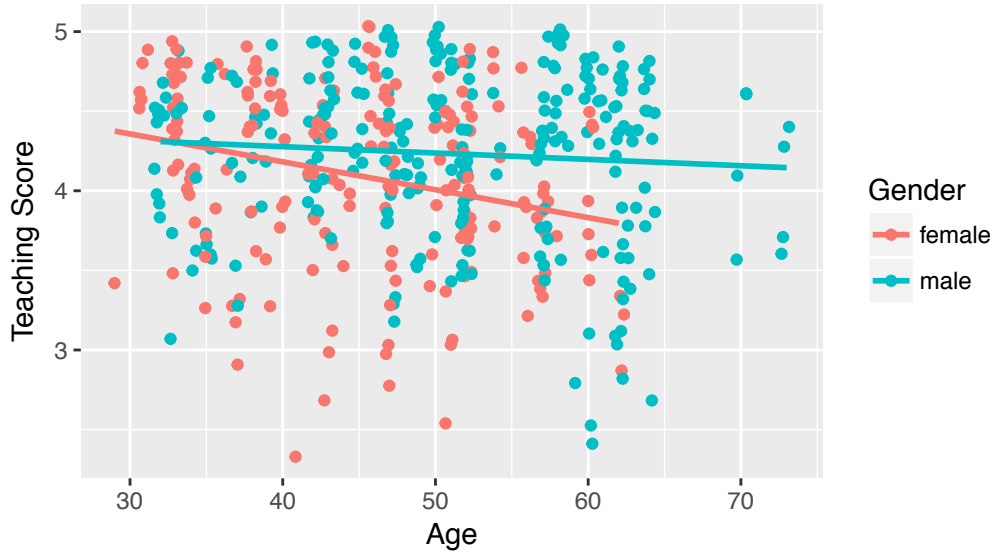
1. x_1: numerical explanatory/predictor variable of age
2. x_2: categorical explanatory/predictor variable of gender

You decide two fit regression models, one without an interaction term and the other with an interaction term, and compare the results.

a) Does this plot represent a visualization of the regression model with or without the interaction term? Circle either "with" or "without."



b) Does this plot represent a visualization of the regression model with or without the interaction term? Circle either "with" or "without."



c) You fit the regression model which includes the interaction term and obtain the following regression table output:

term	estimate	std.error	statistic	p.value
(Intercept)	4.88	0.21	23.8	0.00
age	-0.02	0.00	-3.9	0.00
gendermale	-0.45	0.27	-1.7	0.09
age:gendermale	0.01	0.01	2.5	0.01

Rewrite the equation for the line $\widehat{y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$, but specifically for this data. In other words, put appropriate subscripts on the b's and write down what each of the three's x and the $\widehat{y}$ correspond to.

d) Interpret the 4 resulting fitted coefficients of the line (the intercept b_0 and the three slopes b_1, b_1, and b_3) not in abstract mathematical terms, but in the context of our outcome variable: teaching score.

e) If a faculty member at UT Austin is female and aged 50, what do you predict their teaching score to be? Do not perform any arithmetic, but write down what you would enter into your calculator if you had one.