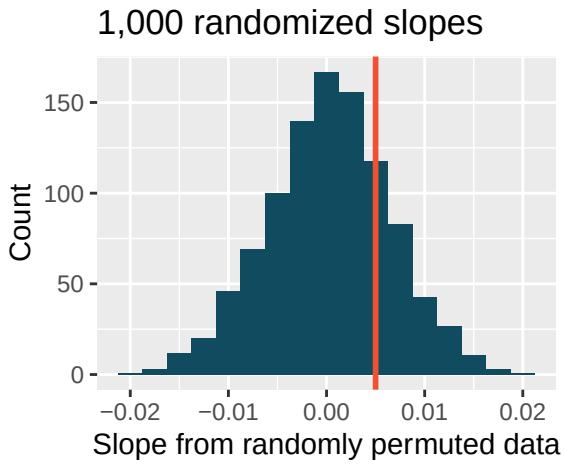
Problem Set 12

Multiple Linear Regression

1. **Baby's weight and father's age, randomization test.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1000 births from 2014. Here, we study the relationship between the father's age and the weight of the baby.

   Below are two items. The first is the standard linear model output for predicting baby's weight (in pounds) from father's age (in years). The second is a histogram of slopes from 1000 randomized datasets (1000 times, `weight` was permuted and regressed against `fage`). The red vertical line is drawn at the observed slope value which was produced in the linear model output.

term	estimate	std.error	statistic	p.value
`(Intercept)`	7.101	0.199	35.674	<0.0001
`fage`	0.005	0.006	0.757	0.4495

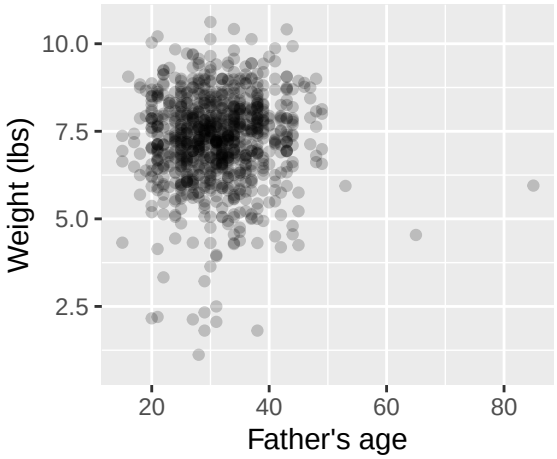


   a. What are the null and alternative hypotheses for evaluating whether the slope of the model for predicting baby's weight from father's age is different than 0?

   b. Using the histogram which describes the distribution of slopes when the null hypothesis is true, find the p-value and conclude the hypothesis test in the context of the problem (use words like father's age and weight of baby). What does the conclusion of your test say about whether the father's age is a useful predictor of baby's weight?

   c. Is the conclusion based on the histogram of randomized slopes consistent with the conclusion which would have been obtained using the mathematical model? Explain.

2. **Baby's weight and father's age, mathematical test.** Is the father's age useful in predicting the baby's weight? The scatterplot and least squares summary below show the relationship between baby's weight (measured in pounds) and father's age for a random sample of babies.



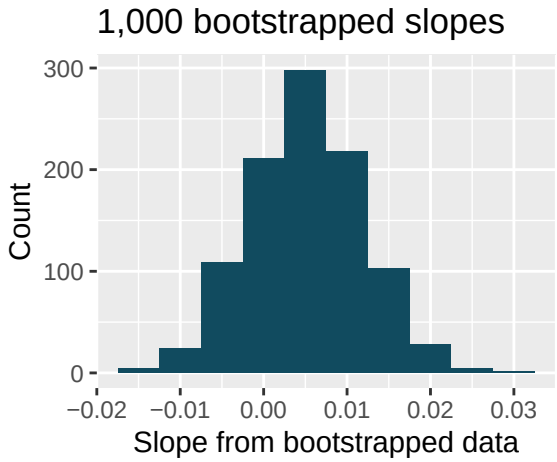
term	estimate	std.error	statistic	p.value
`(Intercept)`	7.1042	0.1936	36.6980	<0.0001
`fage`	0.0047	0.0061	0.7794	0.4359

   a. What is the predicted weight of a baby whose father is 30 years old.

   b. Do the data provide convincing evidence that the model for predicting baby weights from father's age has a slope different than 0? State the null and alternative hypotheses, report the p-value (using a mathematical model), and state your conclusion.

   c. Based on your conclusion, is father's age a useful predictor of baby's weight?

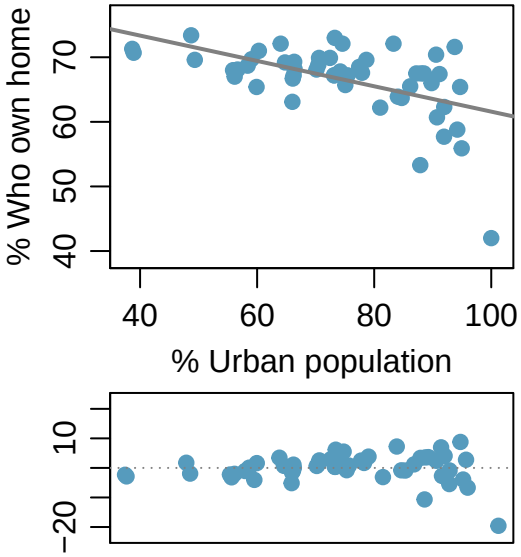
3. **Baby's weight and father's age, bootstrap percentile interval.** US Department of Health and Human Services, Centers for Disease Control and Prevention collect information on births recorded in the country. The data used here are a random sample of 1000 births from 2014. Here, we study the relationship between the father's age and the weight of the baby. Below is the bootstrap distribution of the slope statistic from 1,000 different bootstrap samples of the data.



   a. Using the bootstrap percentile method and the histogram above, find a 95% confidence interval for the slope parameter.

   b. Interpret the confidence interval in the context of the problem.

4. **Urban homeowners, conditions.** The scatterplot below shows the percent of families who own their home vs. the percent of the population living in urban areas. There are 52 observations, each corresponding to a state in the US. Puerto Rico and District of Columbia are also included.



   a. For these data, R^2 is 29.16%. What is the value of the correlation coefficient? How can you tell if it is positive or negative?

   b. Examine the residual plot. What do you observe? Is a simple least squares fit appropriate for these data? Which of the LINE conditions are met or not met?

5. **Training for the 5K.** Nico signs up for a 5K (a 5,000 metre running race) 30 days prior to the race. They decide to run a 5K every day to train for it, and each day they record the following information: `days_since_start` (number of days since starting training), `days_till_race` (number of days left until the race), `mood` (poor, good, awesome), `tiredness` (1-not tired to 10-very tired), and `time` (time it takes to run 5K, recorded as mm:ss). Top few rows of the data they collect is shown below.

days_since_start	days_till_race	mood	tiredness	time
1	29	good	3	25:45
2	28	poor	5	27:13
3	27	awesome	4	24:13
...	...	...	...	...

   Using these data Nico wants to build a model predicting `time` from the other variables. Should they include all variables shown above in their model? Why or why not?

6. **Movie returns, prediction.** A model was fit to predict return-on-investment (ROI) on movies based on release year and genre (Adventure, Action, Drama, Horror, and Comedy). The model output is shown below.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-156.04	169.15	-0.92	0.3565
`release_year`	0.08	0.08	0.94	0.348
`genreAdventure`	0.30	0.74	0.40	0.6914
`genreComedy`	0.57	0.69	0.83	0.4091
`genreDrama`	0.37	0.62	0.61	0.5438
`genreHorror`	8.61	0.86	9.97	<0.0001

a. For a given release year, which genre of movies are predicted, on average, to have the highest predicted return on investment?

b. If you were to plot this model on a single scatterplot of the data, what would the model look like? A single line? A curved line? Multiple lines? Parallel or skewed from one another?

c. The adjusted R^2 of this model is 10.71%. Adding the production budget of the movie to the model increases the adjusted R^2 to 10.84%. Should production budget be added to the model?

7. **Palmer penguins, predicting body mass.** Researchers studying a community of Antarctic penguins collected body measurement (bill length, bill depth, and flipper length measured in millimeters and body mass, measured in grams), species (Adelie, Chinstrap, or Gentoo), and sex (female or male) data on 344 penguins living on three islands (Torgersen, Biscoe, and Dream) in the Palmer Archipelago, Antarctica. The summary table below shows the results of a linear regression model for predicting body mass (which is more difficult to measure) from the other variables in the dataset.

term	estimate	std.error	statistic	p.value
`(Intercept)`	-1461.0	571.3	-2.6	0.011
`bill_length_mm`	18.2	7.1	2.6	0.0109
`bill_depth_mm`	67.2	19.7	3.4	7e-04
`flipper_length_mm`	16.0	2.9	5.5	<0.0001
`sexmale`	389.9	47.8	8.1	<0.0001
`speciesChinstrap`	-251.5	81.1	-3.1	0.0021
`speciesGentoo`	1014.6	129.6	7.8	<0.0001

a. Write the equation of the regression model.

b. Interpret each one of the slopes in this context.

c. Calculate the residual for a male Adelie penguin that weighs 3750 grams with the following body measurements: `bill_length_mm` = 39.1, `bill_depth_mm` = 18.7, `flipper_length_mm` = 181. Does the model overpredict or underpredict this penguin's weight?

d. The R^2 of this model is 87.5%. Interpret this value in context of the data and the model.