Week 2 Quiz

Quiz, 8 questions

✖ Try again once you are ready.

Required to pass: 80% or higher

You can retake this quiz up to 3 times every 8 hours.

Back to Week 2

Retake

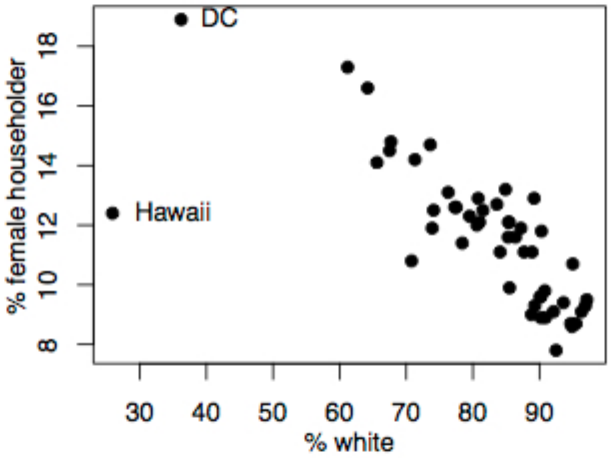
Week 2 Quiz ✓

1 / 1
point

Quiz, 8 questions

1.

The scatterplot on the right shows the relationship between percentage of white residents and percentage of households with a female head in all 50 US States and the District of Columbia (DC). Which of the below **best** describes the two points marked as DC and Hawaii?



○ Hawaii has higher leverage and is more influential than DC.

Correct

Hawaii has higher leverage than DC because it is farther away from the bulk of the data in the x direction.

This question refers to the following learning objective(s):

- Define a leverage point as a point that lies away from the center of the data in the horizontal direction.
- Define an influential point as a point that influences (changes) the slope of the regression line.

1. This is usually a leverage point that is away from the trajectory of the rest of the data.

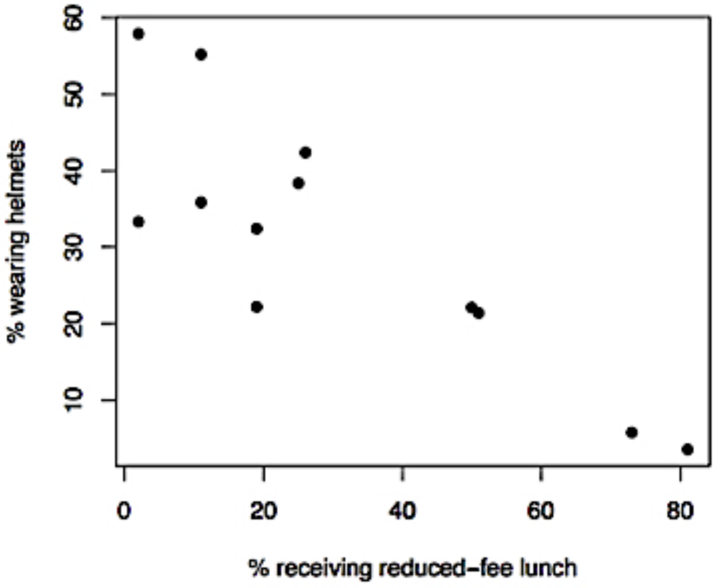
✖ 0 / 1 point

2.

The scatterplot below shows the relationship between socioeconomic status measured as the percentage of children in a neighborhood receiving reduced-fee lunches at school (lunch) and bike helmet use measured as the percentage of bike riders in the neighborhood wearing helmets (helmet). The equation of the regression line is

$$helmet = 47.49 - 0.54 \quad lunch$$

and the R^2 is 72%. Which of the following is **true**?



○ 72% of the percentage of children receiving reduced-fee lunches at school can be accurately predicted by the model.

○ Decreasing the percentage of children receiving reduced-fee lunches at school by 5% will increase the percentage of bike riders wearing helmets in that neighborhood by 2.7%.

This should not be selected

We cannot make such a causal statement based on these observational data.

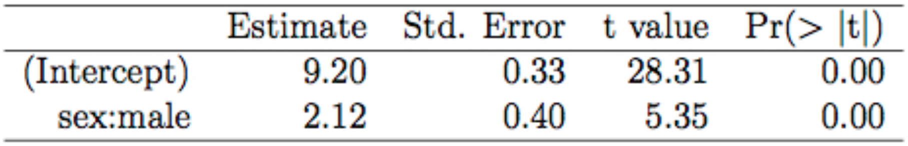
Week 2 Quiz

Quiz, 8 questions

✓ 1 / 1 point

3.

The model below is for predicting the heart weight (in g) of cats from their gender (female and male). The coefficients are estimated using a dataset of 144 domestic cats. Which of the following is **false**?



	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	9.20	0.33	28.31	0.00
sex:male	2.12	0.40	5.35	0.00

○ The expected heart weight for male cats is, on average, 11.32 grams.

○ If the regression equation is written $\hat{y} = b_0 + b_1 x$, then plugging in $x = 0$ would give you the predicted heart weight for a female cat.

○ The intercept is meaningless.

Correct

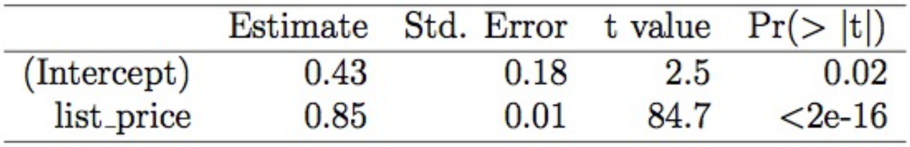
For a categorical explanatory variable like we have here (gender), a value of 0 for the explanatory variable corresponds to the baseline level.

○ Female cats on average are expected to have hearts that weigh 2.12 grams less than those of male cats.

✓ 1 / 1 point

4\.

We fit a linear regression model for predicting the best used price of 23 GMC pickup trucks from their list price, both measured in thousands. Which of the following is **false** based on this model output?



	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	0.43	0.18	2.5	0.02
list_price	0.85	0.01	84.7	<2e-16

○ For each additional $1,000 in the list price of a GMC pickup truck we would expect the best used price to be higher on average by $850.

○ The 95% confidence interval for the slope can be calculated as $0.85 \pm 84.7 \times 0.01$.

Correct

False. We need the critical t score, not the observed t score, in calculation of the margin of error.

This question refers to the following learning objective(s):

- Calculate a confidence interval for the slope as

$b_1 \pm t^\star_{df} SE_{b_1}$,

where $df = n - 2$ and $t^\star_{df}$ is the critical score associated with the given confidence level at the desired degrees of freedom.

- Note that the standard error of the slope estimate SE_{b_1} can be found on the regression output.

○ The linear model is $\widehat{best\_used\_price} = 0.43 + 0.85\ list\_price.$

Week 2 Quiz

Quiz, 8 questions

1 / 1 point

5.

Answer Question 5, 6 and 7 based on the information below:

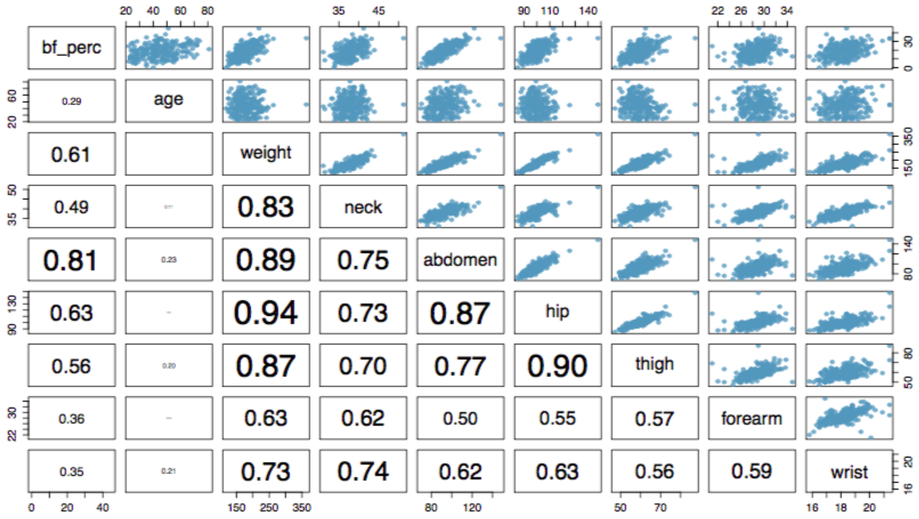
Body fat percentage can be complicated to estimate, while variables such age, height, weight, and measurements of various body parts are easy to measure. Based on data on body fat percentage and other various easy to obtain measurements, we develop a model to predict body fat percentage based on the following variables:

-age (years) - abdomen circumference (cm) - forearm circumference (cm)

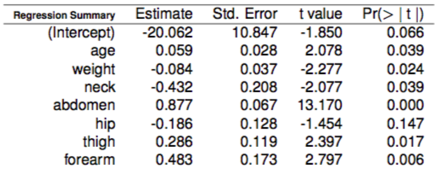
-wight (pounds) - hip circumference (cm) - wrist circumference (cm)

-neck circumference (cm) - thigh circumference (cm)

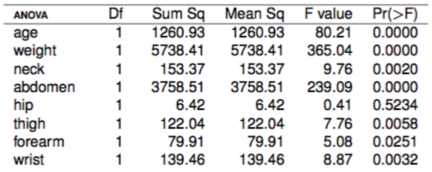
The plot below shows the relationship between each of these variables and body fat percentage (the response variable) as well as the correlation coefficients between these variables:



And the following are the model outputs associated with this analysis:



Regression Summary	Estimate	Std. Error	t value	Pr(> | t |)
(Intercept)	-20.062	10.847	-1.850	0.066
age	0.059	0.028	2.078	0.039
weight	-0.084	0.037	-2.277	0.024
neck	-0.432	0.208	-2.077	0.039
abdomen	0.877	0.067	13.170	0.000
hip	-0.186	0.128	-1.454	0.147
thigh	0.286	0.119	2.397	0.017
forearm	0.483	0.173	2.797	0.006



ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
age	1	1260.93	1260.93	80.21	0.0000
weight	1	5738.41	5738.41	365.04	0.0000
neck	1	153.37	153.37	9.76	0.0020
abdomen	1	3758.51	3758.51	239.09	0.0000
hip	1	6.42	6.42	0.41	0.5234
thigh	1	122.04	122.04	7.76	0.0058
forearm	1	79.91	79.91	5.08	0.0251
wrist	1	139.46	139.46	8.87	0.0032

✖ 0 / 1 point

6.

Do these data provide convincing evidence that age and body fat percentage are significantly **positively** associated? Why or why not? Use quantitative information based on the model output to support your answer, and make sure to note the p-value you use to make this decision.

○ Yes, the p-value for testing for a positive correlation between age and body fat percentage is 0.039 / 2 = 0.0195. Since the p-value is small we reject the null hypothesis of no relationship.

○ Yes, the p-value for testing for a positive correlation between age and body fat percentage is 2e^{-16}. Since the p-value is small we reject the null hypothesis of no relationship.

This should not be selected

Recall that for a predictor i, the p-value given in the Regression Summary output refers to the following t-test:

- H_0: $\beta_i = 0$; with all other predictors in the model, predictor i does not explain a significant portion of the variance in body fat once all other predictors are included in the model, so the coefficient is not significantly different from 0.
- H_A: $\beta_i \neq 0$; predictor i does explain a significant portion of the variance in body fat even when all other predictors are included in the model, so the coefficient is significantly different from 0.

This question refers to the following learning objective:

Determine whether an explanatory variable is a significant predictor for the response variable using the t-test and the associated p-value in the regression output.

○ Yes, the p-value for testing for a positive correlation between age and body fat percentage is 0.000. Since the p-value is small we reject the null hypothesis of no relationship

Week 2 Quiz ✓

Quiz, 8 questions

1 / 1 point

7.

Construct a 95% confidence interval for the slope of abdomen circumference and interpret it in context of the data.

○ (0.00539, 0.88239); All else held constant, for each additional cm in abdomen circumference, body fat percentage is expected to be higher by 0.00539 to 0.88239 percentage points.

○ (-0.00539, 1.75); All else held constant, for each additional cm in abdomen circumference, body fat percentage is expected to change by -0.00539 to 1.75 percentage points.

○ (0.745, 1.009); All else held constant, for each additional cm in abdomen circumference, body fat percentage is expected to be higher by 0.745 to 1.009 percentage points.

Correct

We recall that this confidence interval is supposed to capture $\beta_{abdomen}$, ie the impact of increasing abdomen circumference by 1 cm on the response of body fat percentage.

This question refers to the following learning objective:

Calculate a confidence interval for the slope as $b_1 \pm t^*_{df} SE_{b_1}$ where $df = n - 2$ and t^*_{df} is the critical score associated with the given confidence level at the desired degrees of freedom. Note that the standard error of the slope estimate SE_{b_1} can be found on the regression output.

○ (0.745, 1.009); All else held constant, for each additional percentage point increase in body fat, abdomen circumference is expected to be higher by 0.745 to 1.009 cm.

Week 2 Quiz

Quiz, 8 questions

✓ 1 / 1 point

8.

True/False: Outliers should always be removed from the data set prior to final analysis.

○ False; we only remove outliers after checking to make sure doing so drastically improves model fit.

○ True; outliers distort model fit and must be removed to assure reliable results.

○ False; we only remove outliers if we have very good justification that suggests that removing the outlier is appropriate.

Correct

Outliers can sometimes prove to be the most interesting data points in the analysis! It is very important that you do not remove outliers arbitrarily, even if their removal really improves model fit. Check with the data supplier to see if there is justification that suggests that the outlier might be an error or should be removed.

The question refers to the following learning objective:

Do not remove outliers from an analysis without good reason.