1. We modeled the gas mileage of 398 cars built in the 1970's and early 1980's using engine displacement (in cubic inches), year of manufacture in relation to 1970 (e.g. 4 means the car was built in 1974; 12 means built in 1982, etc.), and manufacturing site (domestic to the USA = 0; foreign to the USA = 1). The regression output is provided below. Note that domestic is the reference level for manufacturing site.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	26.86	0.87	30.75	0.0000
displacement	-0.04	0.00	-16.42	0.0000
year	0.72	0.06	12.48	0.0000
site:foreign	2.21	0.54	4.08	0.0001

Which of the following is the degrees of freedom associated with the p-value for site?

- ○ 2
- ○ 398
- ◉ 394
- ○ 395
- ○ 3

2. We modeled the prices of 93 cars (in $1,000s) using its city MPG (miles per gallon) and its manufacturing site (foreign or domestic). The regression output is provided below. Note that domestic is the reference level for manufacturing site. Data are outdated so the prices may seem low.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	42.56	3.17	13.42	0.0000
city MPG	-1.14	0.14	-8.03	0.0000
site:foreign	5.26	1.59	3.30	0.0014

Which of the following is the correct predicted price (in $1,000s) **of a foreign car that gets 26 MPG**?

- 42.56+(1.14x26)+5.26
- 42.56−(1.14x26)
- 42.56-(1.14x26)+5.26
- (−1.14x26)+5.26

✓ **Correct**

This question refers to the following learning objective(s): Define the multiple linear regression model as

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$

where there are k predictors (explanatory variables).

3. The data in this question come from the Second International Mathematics Study on 8th graders from randomly sampled classrooms in the US who completed mathematics achievement tests at the beginning and at the end of the academic year. Students also answered questions regarding their attitudes toward mathematics. The linear model output below is for predicting the gain score in this test (posttest - pretest score) using the following explanatory variables:

    • pretest: score on the exam taken at the beginning of the semester

    • gender: male or female

    • more_ed: expected number of years for continued education (up to 2 years, 2 to 5 years, 5 to 6 years, 8 or more years)

    • useful: Math is useful in everyday life (strongly disagree, disagree, undecided, agree, strongly agree)

    • ethnic: ethnicity of student (African American, Anglo, Other)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.9529	6.8446	-0.43	0.666
pretest	-0.1629	0.0372	-4.38	1.5e-05
gender:male	0.5586	1.1910	0.47	0.639
more_ed:2 to 5 years	0.0731	6.2351	0.01	0.991
more_ed:5 to 6 years	5.8558	6.1974	0.94	0.345
more_ed:8 or more years	6.6138	6.2726	1.05	0.292
useful:disagree	7.2809	4.3065	1.69	0.092
useful:undecided	7.7716	3.8461	2.02	0.044
useful:agree	8.5578	3.6693	2.33	0.020
useful:strongly agree	9.2262	3.7946	2.43	0.015
ethnic:Anglo	6.4974	2.1779	2.98	0.003
ethnic:Other	5.3995	2.8049	1.92	0.055

What does the **intercept** in this model represent?

- ○ Any student who scored 0 on the pretest.
- ◉ An African American female student who scored 0 on the pretest, expects to continue their education for up to 2 years, who strongly disagrees with the statement on usefulness of math.
- ○ An African American male student who scored 0 on the pretest, expects to continue their education for up to 2 years, who strongly disagrees with the statement on usefulness of math.
- ○ A student who scored 0 on the pretest and did not answer the other questions on expected years of education, usefulness of math, and ethnicity.

✓ **Correct**

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

- Interpret the estimate for the intercept (b_0) as the expected value of y when all predictors are equal to 0, on average.

- Interpret the estimate for a slope (say b_1) as "All else held constant, for each unit increase in x_1, we would expect y to be higher/lower on average by b_1."

4. We modeled the prices of 93 cars (in $1,000s) using its city MPG (miles per gallon) and its manufacturing site (foreign or domestic). The regression output is provided below. Note that domestic is the reference level for manufacturing site. Data are outdated so the prices may seem low.

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	42.56	3.17	13.42	0.0000
city MPG	-1.14	0.14	-8.03	0.0000
site:foreign	5.26	1.59	3.30	0.0014

Which of the following is **false**?

- ○ Manufacturing site is a significant predictor of car price, given information on the city MPG of the car.
- ○ If we add another variable to the model, say highway MPG, the p-values associated with city MPG and manufacturing site may change.
- ◉ The 95% confidence interval for the slope of city MPG can be calculated as $-1.14 \pm (-8.03 * 0.14)$.
- ○ City MPG is a significant predictor of car price, given information on the manufacturing site of the car.

5. Researchers investigating academic performance of high school students examined data from 2011 from all 50 states and DC ($n = 51$). The data included information on

• SAT math (response variable) - average SAT math score of all students in the state who took the exam

• per_ppl_sp - average number of dollars per pupil spent on education by the state

• perc_take - percentage of high school seniors in the state that took the exam

The output of the model fit by one of theresearchers is shown below:

Linear model	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	553.7294	12.5089	44.27	0.0000
per_ppl_sp	0.0029	0.0013	2.28	0.0274
perc_take	-1.2250	0.1111	-11.03	0.0000

ANOVA	Df	Sum Sq	Mean Sq	F value	Pr(>F)
per_ppl_sp	1	7003	7003	14.75	0.0004
perc_take	1	57737	57737	121.61	0.0000
Residuals	48	22790	475		

What proportion of the variability in the response variable is explained by the model? Choose the closest answer.

- 31%
- 66%
- 74%
- 8%

Correct

$R^2 = SSR/SST = \frac{7003+57737}{7003+57737+22790} = 0.7396 \approx 74$

This question refers to the following learning objective(s): Note that R^2 will increase with each explanatory variable added to the model, regardless of whether or not the added variable is a meaningful predictor of the response variable. Therefore we use adjusted R^2, which applies a penalty for the number of predictors included in the model, to better assess the strength of a multiple linear regression model:

$$R^2_{adj} = 1 - \frac{SSE/(n-k-1)}{SST/(n-1)}$$

where n is the number of cases and k is the number of predictors.

- Note that R^2_{adj} will only increase if the added variable has a meaningful contribution to the amount of explained variability in y, i.e. if the gains from adding the variable exceeds the penalty.

6. The data in this question come from the Second International Mathematics Study on 8th graders from randomly sampled classrooms in the US who completed mathematics achievement tests at the beginning and at the end of the academic year. Students also answered questions regarding their attitudes toward mathematics. The linear model output below is for predicting the gain score in this test (posttest - pretest score) using the following explanatory variables:

    • pretest: score on the exam taken at the beginning of the semester

    • gender: male or female

    • more_ed: expected number of years for continued education (up to 2 years, 2 to 5 years, 5 to 6 years, 8 or more years)

    • useful: Math is useful in everyday life (strongly disagree, disagree, undecided, agree, strongly agree)

    • ethnic: ethnicity of student (African American, Anglo, Other)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-2.9529	6.8446	-0.43	0.666
pretest	-0.1629	0.0372	-4.38	1.5e-05
gender:male	0.5586	1.1910	0.47	0.639
more_ed:2 to 5 years	0.0731	6.2351	0.01	0.991
more_ed:5 to 6 years	5.8558	6.1974	0.94	0.345
more_ed:8 or more years	6.6138	6.2726	1.05	0.292
useful:disagree	7.2809	4.3065	1.69	0.092
useful:undecided	7.7716	3.8461	2.02	0.044
useful:agree	8.5578	3.6693	2.33	0.020
useful:strongly agree	9.2262	3.7946	2.43	0.015
ethnic:Anglo	6.4974	2.1779	2.98	0.003
ethnic:Other	5.3995	2.8049	1.92	0.055

In model selection using backwards elimination based on p-values, which of the following variables would be dropped first?

- ○ pre-test
- ○ gender
- ○ ethnic
- ○ useful
- ◉ more ed

! **Incorrect**
Lowest p-value associated with one of the levels of this variable is lower than others.

7. Which of the following is **false** about conditions for multiple linear regression?

- ○ Residuals should be normally distributed around 0.
- ○ Explanatory variables should have linear relationship with the response variable.
- ○ Residuals should have constant variance.
- ◉ Explanatory variables should have strong relationships with each other.

8. Which of the following is the **best** definition of a parsimonious model?

- ○ The model with the most number of predictors.
- ◉ The simplest model with the highest predictive power.
- ○ The model that includes all the predictors that your audience is interested in.
- ○ The model with the highest R^2.

✓ **Correct**
This question refers to the following learning objective(s): Note that we usually prefer simple (parsimonious) models over more complicated ones.

9. A high correlation between two explanatory variables such that the two variables contribute redundant information to the model is known as

- ○ multiple correlation
- ○ heteroscedasticity
- ◉ collinearity
- ○ homogeneity
- ○ multiple interaction
- ○ heterogeneity
- ○ homoscedasticity
- ○ adjusted R^2

✓ **Correct**

This question refers to the following learning objective(s): Define collinearity as a high correlation between two independent variables such that the two variables con- tribute redundant information to the model – which is something we want to avoid in multiple linear regression.

10. A model selection method where we start with an empty model and add variables one at a time until no other important variables are found is called

- ◉ forward selection
- ○ randomization
- ○ multiple testing
- ○ bootstrapping
- ○ ANOVA