# Intro to Statistical Modeling

Question 1

* A statistical unit is one member among the set of entities being studied.
* True

Question 2

* Suppose that researchers would like to predict whether a patient's cancer will return based on several features of the patient, including the age, gender, and weight of the patient, and on characteristics of the patient's original tumor, including, its radius, perimeter, smoothness, and symmetry.

In this modeling scenario, the statistical unit is:

* Each individual patient

Question 3

* Suppose that researchers would like to predict whether a patient's cancer will return based on several features of the patient, including the age, gender, and weight of the patient, and on characteristics of the patient's original tumor, including, its radius, perimeter, smoothness, and symmetry.
* In this modeling scenario, the "output" variable is:
* Whether or not a patient will develop another tumor

Question 4

* Suppose that researchers would like to predict whether a patient's cancer will return based on several features of the patient, including the age, gender, and weight of the patient, and on characteristics of the patient's original tumor, including, its radius, perimeter, smoothness, and symmetry.
* Which of the following is most plausibly the the population for this modeling scenario?
* The collection of all cancer patients that might visit the hospital being observed

Question 5

* To *operationalize* a concept means to derive a set of steps that would allow one to measure the concept.
* True

Question 6

* Consider a study meant to predict whether an individual will default on a mortgage loan based on the individual's socioeconomic status. Researchers identify socioeconomic status with income.
* In this study:
* socioeconomic status has been operationalized using income

Question 7

* Suppose that researchers at a given hospital measure the severity of each patient's condition, along with how satisfied each patient is with their own care. Researchers hypothesize that the more severe the condition, the lower the care satisfaction score.
* A valid measure of the satisfaction of a patient's care:
* Should be carefully operationalized using relevant research on satisfaction scores.

Question 8

* Researchers should decide which interpretation of a concept would be most reasonable for their purposes only after collecting data to analyze.
* False

# The Linear Regression Model

Question 1

* In the age of "big data", it is much more common for data scientists to specify their hypotheses *before* collecting data.
* False

Question 2

* Circular data analyses, or double dipping, is:
* The process of exploring a dataset in an attempt to discover what relationships exist, and then test hypotheses related to that exploration on the same dataset.
* Something to be avoided.

Question 3

* Consider a dataset containing three variables, each measured in dollars:
* Sales of a product (response, Y)
* YouTube advertising budget (predictor, x, start subscript, 1, end subscript)
* Facebook advertising budget (predictor, x, start subscript, 2, end subscript)
* Suppose that the relationship between the predictors and the response is linear, i.e., Y, equals, beta, start subscript, 0, end subscript, plus, beta, start subscript, 1, end subscript, x, start subscript, 1, end subscript, plus, beta, start subscript, 2, end subscript, x, start subscript, 2, end subscript, plus, \varepsilon. What is the interpretation of beta, start subscript, 2, end subscript?
* beta, start subscript, 2, end subscript is the average change in sales for a one dollar increase in the Facebook advertising budget, assuming that the YouTube advertising budget is held constant.

Question 4

* What is the correct interpretation of beta, start subscript, 0, end subscript?
* beta, start subscript, 0, end subscript is the mean/expected sales for the sample average Facebook and YouTube advertising budgets.

Question 5

* Which of the following most plausibly describes the statistical units in the dataset?
* Individual companies that sell the product whose sales are quantified in Y.

Question 6

* Which of the following most plausibly describes the population of interest?
* The set of all companies that sell the product whose sales are quantified in Y.

Question 7

* Which of the following are assumptions of the linear regression model?
* The existence of a linear relationship between the response variable and the model parameters.
* Distinct random error terms must be independent.
* Each response measurement is assumed to have the same variance.

Question 8

* What is the correct interpretation of beta, start subscript, 1, end subscript?
* beta, start subscript, 1, end subscript is the average change in sales for a one standard deviation increase in the YouTube advertising budget, assuming that the Facebook advertising budget is held constant.

# Least Squares

Question 1

* Let y, start subscript, i, end subscript, equals, beta, start subscript, 0, end subscript, plus, beta, start subscript, 1, end subscript, x, start subscript, i, end subscript, plus, \varepsilon, start subscript, i, end subscript where E, left parenthesis, \varepsilon, start subscript, i, end subscript, right parenthesis, equals, 0 for i, equals, 1, comma, point, point, point, comma, n. Fix x, start subscript, i, end subscript and y, start subscript, i, end subscript and allow beta, start subscript, 0, end subscript and beta, start subscript, 1, end subscript to vary. Then the least squares estimator is \min, start subscript, beta, start subscript, 0, end subscript, comma, beta, start subscript, 1, end subscript, end subscript, sum, start subscript, i, equals, 1, end subscript, start superscript, n, end superscript, left parenthesis, y, start subscript, i, end subscript, minus, beta, start subscript, 0, end subscript, minus, beta, start subscript, 1, end subscript, x, start subscript, i, end subscript, right parenthesis, squared, point
* True

Question 2

* Under the definitions and assumptions given in "Lesson: Introduction to least squares estimation", the least squares estimator is the orthogonal projection of the vector Y onto the column space of X.
* True

Question 3

* The equation X, start superscript, T, end superscript, X, beta, equals, X, start superscript, T, end superscript, Y always has a unique solution.
* False

Question 4

* In order to use least squares in the linear regression context, we must assume that:
* The relationship between the expected value of the response and the parameters is linear (or can be transformed to linearity).
* The error term has zero mean.
* The variance is constant across error terms.

Question 5

* The Gauss-Markov theorem states that the least squares estimator has the lowest variance among all esimators of beta (given that the least squares assumptions are met).
* False

Question 6

* In the context of linear regression, for any error distribution, the least squares estimator is equivalent to the maximum likelihood estimator.
* False

Question 7

* The fitted values of a regression model are defined as y, with, \widehat, on top, equals, X, beta, with, \widehat, on top, point
* True

Question 8

* Let H be the hat matrix, as defined in "Lesson: Deriving the least squares solution". Then H, Y, equals, Y, with, \widehat, on top.
* True

Question 9

* Let H be the hat matrix, as defined in "Lesson: Deriving the least squares solution". Then H, Y, with, \widehat, on top, equals, point, point, point

# Variability and Identifiability in Regression Models

Question 1

* The residual sum of squares, as defined in "Lesson: Sums of squares and estimating the error variance" is a measure of the amount of unexplained variability in the data.
* True

Question 2

* The explained sum of squares, as defined in "Lesson: Sums of squares and estimating the error variance" is a measure of the overall variability in the data.
* False

Question 3

* The total sum of squares, as defined in "Lesson: Sums of squares and estimating the error variance" is a measure of variability that is independent of any predictor values.
* True

Question 4

* The explained sum of squares, as defined in "Lesson: Sums of squares and estimating the error variance" is a measure of variability that is independent of any linear regression parameter values.
* False

Question 5

* If our linear regression assumptions are met, then we can estimate the error variance using start fraction, T, S, S, divided by, n, minus, left parenthesis, p, plus, 1, right parenthesis, end fraction
* False

Question 6

* The higher the coefficient of determination, the stronger the causal relationship is between the response and predictors.
* False

Question 7

* A fitted model with a high coefficient of determination is a well-fitting model.
* True

Question 8

* Which of the following imply that a linear regression model will be non-identifiable?
* One predictor value is a linear combination of several other predictor variables.
* One predictor variable is a constant multiple of another.

# Statistical Inference: Intro and T-Tests

Question 1

* Researchers and government officials in Boulder, Colorado are studying the impact of universal basic income (UBI) on the personal savings of city residents. The researchers hypothesize that providing all city residents of Boulder with $1,000 per month will lead to a "significant increase" in personal savings. Data were collected that were relevant to this hypothesis, and a regression was performed.
* The regression parameter associated with the UBI payment was estimated to be *β*ˆ*UBI*=$1, and was (correctly) interpreted as follows: controlling for other variables in the regression (e.g., demographic information, initial wealth), city residents who received the UBI payments saved $1 more per year, on average, than those who did not receive the payments.
* The p-value associated with beta, with, \widehat, on top, start subscript, U, B, I, end subscript was 0, point, 01. Which of the following statements is correct?
* Researchers found a statistically significant increase in savings over the course of the year at the alpha, equals, 0, point, 05 level (controlling for other variables).

Question 2

* Statistical significance implies practical significance.
* False

Question 3

* Practical significance implies statistical significance.
* False

Question 4

* Estimators of regression parameters have sampling distributions.
* True

Question 5

* In order to perform t-tests and confidence intervals in individal regression parameters, we must assume that the error terms are normally distributed.
* True

Question 6

* The amount of variability in beta, with, \widehat, on top depends on:
* The response variance.
* The error variance.
* The values of the predictors.

Question 7

* The distribution of beta, with, \widehat, on top depends on:
* The response variance.
* The values of the predictors.
* The value of beta.
* The error variance.

Question 8

* The t-test for an individual regression parameter beta, start subscript, j, end subscript can be used to test which of the following null and alternative hypotheses?
* H, start subscript, 0, end subscript, colon, beta, start subscript, j, end subscript, equals, 5 vs H, start subscript, 1, end subscript, colon, beta, start subscript, j, end subscript, does not equal, 5
* H, start subscript, 0, end subscript, colon, beta, start subscript, j, end subscript, equals, 0 vs H, start subscript, 1, end subscript, colon, beta, start subscript, j, end subscript, is greater than, 0

Question 9

* A regression was performed in R, and the following output was obtained.
* Call: lm(formula = y ~ x) Residuals: Min 1Q Median 3Q Max -2.83464 -0.53756 -0.04512 0.59881 1.80620 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 0.9514 0.1163 8.18 6.47e-12 \*\*\* x 3.9163 0.1988 **BLANK** < 2e-16 \*\*\* --- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 1.007 on 73 degrees of freedom Multiple R-squared: 0.8417, Adjusted R-squared: 0.8395 F-statistic: 388.1 on 1 and 73 DF, p-value: < 2.2e-16
* Find **BLANK**, rounded tothe hundreths place.
* 19.70

Question 10

* Call: lm(formula = y ~ x) Residuals: Min 1Q Median 3Q Max -3.7070 -1.3931 0.1937 0.7267 4.2754 Coefficients: Estimate Std. Error t value Pr(>|t|) (Intercept) 1.1721 0.5228 2.242 0.043 \* x 1.0692 0.8470 1.262 **BLANK** --- Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1 Residual standard error: 2.025 on 13 degrees of freedom Multiple R-squared: 0.1092, Adjusted R-squared: 0.04066 F-statistic: 1.593 on 1 and 13 DF, p-value: 0.229
* Find **BLANK**, rounded tothe hundreths place. (HINT: the degrees of freedom provided in the output may be useful!)
* 0.23

# Statistical Inference: The F-Tests and Confidence Intervals

Question 1

* Consider a multiple linear regression with 20 predictor variables. Let alpha, equals, 0, point, 05. The probability of a type I error for any single predictor is 0, point, 05.
* True

Question 2

* Consider a multiple linear regression with 20 predictor variables. Let alpha, equals, 0, point, 05. The probability of a type I error for at least one of the predictors is 0, point, 05.
* False

Question 3

* Researchers studied the bedtime habits of 200 individuals. Among them, 100 individuals read for at least 30 minutes before bed, and 100 did not. Over 20 variables were measured on these individuals, including their sleep duration, quality, blood pressure, blood sugar levels, anxiety and depression levels, etc.
* Let alpha, equals, 0, point, 01. It is more likely than not that researchers will observe a false positive.
* False

Question 4

* The F-distribution is a special case of the Gamma distribution.
* False

Question 5

* Under the standard regression assumptions, the partial F-test tests the null hypothesis that E, left parenthesis, Y, right parenthesis, equals, beta, start subscript, 0, end subscript.
* False

Question 6

* Under the standard regression assumptions, the standard error for a response y, with, \widehat, on top, start subscript, i, end subscript, start superscript, times, end superscript, equals, x, start superscript, times, end superscript, beta, with, \widehat, on top at a new set of predictors is
* True

**3 points**

**7.**

Question 7

* Consider a well-fitting simple linear regression model in R. The predict() function produces the following output (at the default alpha, equals, 0, point, 05):
* fit lwr upr 1.32209 -0.1830998 2.82728
* There is no statistical evidence that the mean response at the new value of x is different from zero.

Question 8

* Researchers and government officials in Boulder, Colorado are studying the impact of universal basic income (UBI) on the personal savings of city residents. The researchers hypothesize that providing all city residents of Boulder with $1,000 per month will lead to a "significant increase" in personal savings. Data were collected that were relevant to this hypothesis, and a regression was performed.
* The regression parameter associated with the UBI payment was estimated to be *β*ˆ*UBI*=$1, and was (correctly) interpreted as follows: controlling for other variables in the regression (e.g., demographic information, initial wealth), city residents who received the UBI payments saved $1 more per year, on average, than those who did not receive the payments.
* The associated confidence interval for beta, with, \widehat, on top, start subscript, U, B, I, end subscript is given as (−$10,$15). Which of the following statements is correct?
* There is not enough information to provide a best point estimate for the mean personal savings held by a city resident.
* There is no statistical evidence that, when controlling for other variables in the regression, UBI payments increase mean personal savings.

Question 9

* Under the standard linear regression assumptions, identify the expectation of the point estimate for the average response at a new set of predictors, x, start superscript, times, end superscript. That is, what is E, left parenthesis, y, with, \widehat, on top, start superscript, times, end superscript, right parenthesis?

# Ethics in Statistical Practice and Communication: Five Recommendations

Gelman says, "audiences interpret the statistics (and the paragraphs) they read in the context of their understanding of the world and their expectations of the author, who in turn has various goals of exposition and persuasion – and all of this is happening within a competitive publishing environment, in which authors of scientific papers and policy reports have incentives to make dramatic claims," which results in scientists "not communicating their work to one another, let alone to general audiences, in terms appropriately geared to enlarging knowledge."

The major point Gelman makes is that statistics is "dependent on the methods used to analyze the data," as well as the data itself. Therefore, we need to be aware of not only "arbitrary decisions in data analysis," but also "researcher freedom in data coding, exclusion, and cleaning more generally. Open data and open methods imply a replicable “paper trail” leading from raw data, through processing and statistical analysis, to published conclusions."

The biggest limitation Gelman cites is when "the data do not provide strong evidence for the authors’ claims," pointing to the motivation for authors to overstate their claims or make them out to have stronger evidence (statistically) than they really do.

# Prediction

Question 1

* Which of the following scenarios can be classified under explanatory modeling?
* Observational studies that seek to describe how changes in some variables are caused by changes in others.
* Data analysis problems where the goal is to learn about regression parameters that have a physical meaning.
* Experimental studies where regression parameters are given a causal interpretation.
* Data analysis problems where researchers pay special attention to the interpretation of parameter values.

Question 2

* It can be shown that prediction and explanation are equivalent.
* False

Question 3

* The best point prediction for the response at a new set of predictors is equivalent to the best point estimate for the mean response at those same predictors.
* True

Question 4

* All else equal (e.g., the confidence level, sample size), a prediction interval for the response at a new set of predictors is equivalent to a confidence interval for the mean response at those same predictors.
* False

Question 5

* The mean squared prediction error may be used for comparing linear regression models.
* True

Question 6

* Consider a dataset that has been split into a training set and a test set. The mean squared prediction error is computed using the:
* test set data.

Question 7

* All else equal (e.g., the confidence level, sample size), a prediction interval for the response at a new set of predictors will always be two units wider than a confidence interval for the mean response at those same predictors.
* False

Question 8

* Prediction intervals account for both the variation in estimating the population mean (regression line/surface), and also the random variation of the individual response values.
* True

Question 9

* A 90, percent prediction interval can be interpreted as follows: for a fixed set of predictors, if one were to resample the response and refit a model many times, and use each of those models to make a prediction interval at a fixed set of new predictors, 90, percent of those intervals would contain the true value of the new response.
* True

Question 10

* In cities with 100,000 to 500,000 residents, the sales (y) of a product are recorded, along with TV (x, start subscript, 1, end subscript), radio (x, start subscript, 2, end subscript), and newspaper advertising (x, start subscript, 3, end subscript).  A linear regression model is fit with y as the response and the remaining variables as predictors. Diagnostics are performed and there is no evidence of any violations of assumptions.
* Suppose that a market opens for the same product in a city with 200,000 people. The market has advertising characteristics x, start subscript, 1, end subscript, start superscript, times, end superscript, x, start subscript, 2, end subscript, start superscript, times, end superscript, and x, start subscript, 3, end subscript, start superscript, times, end superscript, which are within the range of values in the original dataset. Predicting sales of this product in this new market, y, start superscript, times, end superscript, using x, start subscript, 1, end subscript, start superscript, times, end superscript, x, start subscript, 2, end subscript, start superscript, times, end superscript, and x, start subscript, 3, end subscript, start superscript, times, end superscript and the model fit in the previous paragraph is an example of a quantitative extrapolation.
* False