Exam: final (cumulative)

This is a preview of the published version of the quiz

Started: Dec 15 at 10:21am

Quiz Instructions

Note carefully the time limit/deadline. You may only submit once; after you submit, you may not change your responses.

You may use any physical notes you want, as well as the digital version of our textbook (access your version of choice from **my website here**) including the linked YouTube videos, **Wikipedia**, our Canvas site, and any PDF/.txt/etc. files on your computer that you can view through your web browser by entering a "URL" (location) starting with file:/// such as file:///C:/Users/kaplandm/Papers/Risk_mgmt/Acerbi_2002.pdf. But, **you may not consult with any other person or artificial intelligence** (whether in person, by text/chat, email, etc.), nor may you use any search engine.

If you prefer to work on a paper version, then (I think/hope!) you can print **this linked PDF version**. The formatting is unpleasant, but the substance seems correct.

As you know, Honorlock will be used to record your screen and you (via webcam/mic) and your web traffic, and I plan to review each recording even if unflagged.

If you have any technical difficulties, I suggest sending me a very short message in Canvas (like just "Hi, I'm having some technical difficulties but trying to resolve them") for documentation, and then trying to resolve them through the Honorlock tech support (chat) or MU Canvas tech support (855-675-0755) depending on the problem. And keep me updated as needed.

There is no penalty for guessing.

Good luck!

Question 1 1 pts

I hereby pledge to follow the exam instructions and certify that my responses represent my efforts alone: I have not communicated with any other humans about this exam, and I have not used any artificial intelligence. (Per instructions, textbook/notes/Wikipedia/etc. are still fine to consult.)

- ○ True
- ○ False

Question 2 1 pts

Let X be a discrete random variable representing the number of times a student's laptop will die during college (i.e., how many new laptops they'll need to buy). X=0 means their original laptop never dies, and X=4 means they have four laptops die. It is also possible to observe X=1, X=2, or X=3, but no other values. The probability distribution is given by

$P(X=m) = (4-m)/10$.

What is the mean of X?

- ○ 1
- ○ 2
- ○ 2.1
- ○ [not enough information]

Question 3 1 pts

You have a dataset of 200 individuals in Missouri, in which you observe their change in annual income ($/year) over the past 5 years. The sample average of these income changes is $7,142.85/year. Even though this estimate is a specific number/value, the frequentist framework treats the sample average as a "random variable" in the sense that

○ it's not a well-defined quantity

○ the estimates almost never equal the true values

○ the researcher is at the mercy of the dataset she has available

○ if we drew a different random sample from the population, the estimate would be different (the "before sampling" perspective)

Question 4 1 pts

Consider estimators G and H of parameter p. The sampling distribution of G is P(G=p-7)=P(G=p+7)=1/2; the distribution of H is P(H=p+3)=1. The mean squared error difference is MSE(G)-MSE(H)=

Question 5 1 pts

An economics journal's editorial board decides that it's tired of publishing 95% confidence intervals only to learn later that the true value is actually outside the 95% CI. To address this, one board member suggests changing the required confidence level from 95% to 99%. Somebody else notes that even with 99%, roughly 1 in 100 empirical studies will have a CI that fails to include the true value. Consequently, the board decides to get tough and not permit any such errors: in order to get published, a paper must report a 100% CI. As a result:

○ p-values also increase to 100%.

○ usually the CI includes the estimated value, unless the estimated value is zero but the true value is strictly positive (or negative).

○ the CI must be so wide that it is not helpful (because it also contains so many wrong values).

○ zero papers are published because it is mathematically impossible to construct a CI with 100% coverage probability (in finite samples).

Question 6 1 pts

Your company wants to learn the effect on worker productivity of blocking Facebook on work computers. They form pairs of similar workers and randomly pick 1 of the 2 workers in the pair to block from Facebook. Imagine they can measure Y, the productivity difference in hours per day; i.e., the blocked worker's productive work time minus the non-blocked worker's productive work time. From their sample, their estimate of E(Y) is 0.98, and the CI is [-1.02, 2.98]. Among the following, the most reasonable interpretation is:

○ we do not know if the true change is positive or negative, but we are very sure that it's not significant economically.

○ our best guess is that average productivity increases a lot, but we also have a lot of uncertainty and wouldn't be surprised if there's actually zero or even negative change.

○ we can see that the estimator is upward biased because 0.98>0, so we should not rely on these results.

○ the estimate is not economically significant, but the biggest values in the CI are.

Question 7 1 pts

SUTVA requires

○ the treatment status of one individual cannot affect the outcome of another individual

○ any "type" of individual in the population has treatment probability strictly between 0 and 1

○
[both of these]
○
[none of these]

Question 8 1 pts

Consider the question, "What is the mean daily return (%) of a certain stock on the day after the overall market (S&P 500 index) increases, minus the mean daily stock return the day after an overall decrease?" This question primarily concerns

○
description
○
prediction
○
causality
○
[none of these]

Question 9 1 pts

Consider completion of this econometrics class as the "treatment" and annual salary 5 years from today as the outcome. Let's assume you indeed complete this class. If we observe your salary in 5 years, and subtract your current salary, that's called your

○
untreated potential outcome minus treated potential outcome
○
treatment effect
○
average treatment effect
○
[none of these]

Question 10 1 pts

Assume that the true causal effect of a 3rd-grade child in Kenya having an insecticide-treated bed net (ITN; to prevent mosquito bites, to reduce incidence of malaria) is attending 10 more days of school per year than with no bed net. If we take a random (iid) sample of children in Kenya who just completed 3rd grade, and we subtract the average school attendance (in days) of children with no bed net from the average attendance of children with an insecticide-treated bed net, the expected value of this difference (in days) is

○
-10
○
0
○
10
○
[not enough information]

Question 11 1 pts

Heteroskedasticity means that

○
homogeneous treatment effects should not be assumed for the model
○
the structural error cannot have conditional mean zero for literally every possible value of X
○
the error terms for different observations may be different: $U_i \neq U_k$ for $i \neq k$
○
the variance of Y depends on X

Question 12 1 pts

Let W=1 if a household has more than zero children (under age 18), and W=0 otherwise. Let Z be the number of children in the household. Conditional on a household having a non-zero number of children, the probability of having (exactly) two children is

○
P(Z=2,W=1)/P(W=1)

○
P(Z=1,W=2)/P(W=2)
○
P(W=1 | Z=2)
○
[none of these]

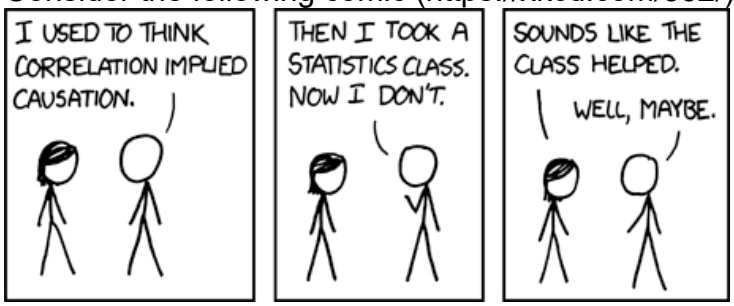
Question 13 1 pts

Let Y be a student's score in points out of 100 possible on an introductory physics final exam. Let X be the number of minutes a student was allowed to spend on the exam. The physics professor randomly assigned X=90 or X=120 to see the effect on score. The estimated regression has intercept 49pts and slope 0.24pts/min. If students are instead given 5 hours for the test, we'd expect

○
a mean score of 121, with some students below and some above
○
each student will score 121, since no error term is estimated
○
the class average will be strictly below 100
○
the original data seem to be flawed (graded incorrectly, cheating, etc.)

Question 14 1 pts

Consider the following comic (https://xkcd.com/552/):



The joke is essentially that

○
the main character has become overly skeptical of the identifying assumptions required to interpret his change in belief as a causal effect of the class
○
since the data are time series and iid sampling is violated, the data technically can't be used when discussing causality
○
the protagonist has become so reliant on statistical significance that he won't infer a causal effect unless the p-value is less than 0.05 (which implicitly is not true in this case)
○
the woman naively interprets the facts stated in the first two panels in terms of causality, but she doesn't understand things as well as the man who has taken statistics

Question 15 1 pts

Imagine the true structural/causal model is Y=U, with no effect of X; Y is fully determined by U, i.e., if you change somebody's X without changing their U, their Y does not change. Random variables X, Y, and U are all binary (0 or 1). Also, X=U. What's the slope of CMF E(Y|X=x)? [Hint: graph Y vs. X.]

○
-1
○
0
○
1
○
[not enough information]

Question 16 1 pts

Let A represent the statement "E(Y)≤0" and B represent "E(Y)=0". The logical relationship between A and B is

○
B ⇒ A

○
A is stronger than B
○
A is true only if B is true
○
[none of these]

Question 17 1 pts

Because the linear projection slope coefficient can be written as Cov(X,Y)/Var(X), it cannot be
○
bigger than 1
○
bigger than 1 in absolute value
○
nonpositive (i.e., strictly negative)
○
[none of these]

Question 18 1 pts

Let Y denote a firm's profit margin as a decimal/proportion, where typical values in the industry you're studying are from Y=0.00 to Y=0.15; for example, Y=0.08 means 8% profit margin. Let X denote the firm's share of employees who have an economics degree as a decimal/proportion; for example, X=0.15 means 15%. You run OLS, computing a slope estimate of 0.0002 and 95% CI of [-2.0202, 2.0202]. Among the following, the most reasonable interpretation is:
○
we feel fairly sure that the statistical relationship is not economically significant (although we are unsure whether it is positive or negative).
○
the estimate suggests that firms can increase their profit margins greatly by hiring more economics-trained employees, but that assumes the slope can be interpreted as a causal effect.
○
if a firm hires one additional employee with an economics degree, then our best guess (prediction) is that the firm's profit margin will increase by 0.0002.
○
our best guess is that the statistical relationship is not economically significant, but we are not very certain about that because the CI contains economically significant values.

Question 19 1 pts

Consider linear-log regression of Y on X, using the model , where Y is the average 4th-grade math test score in a school district (in units of points), and X is the average annual income among individuals in the district, in units of thousands of dollars per year. You estimate the function 557.8 + 36.42 ln(X). Given this, without worrying about misspecification, a 1% increase in income is associated with a change in test score of (approximately)
○
0.36 points
○
36 points
○
557.8 points
○
36 percentage points

Question 20 1 pts

Let Y be hourly wage ($/hr) and X=years of experience. You want to estimate a statistical relationship between wage and experience, not worrying about causality. You've heard that initially more experience is associated with higher wage, but that at very large values of experience, more experience is actually associated with lower wage. You've also heard that each year of experience is associated with a percentage change in wage. Given all this, the best model to try first would be
○
$\ln(Y) = a + b \ln(X) + U$
○
$Y = a + b X + c X^2 + U$

○
$Y = a + b \ln(X) + c \ln(X)^2 + U$
○
$\ln(Y) = a + b X + c X^2 + U$

Question 21 1 pts

Consider the example of regressing a school's 4th-grade reading test score (S) on its 4th-grade student-teacher ratio (R), i.e., R = (total # students) / (total # teachers). We hope to learn the structural slope coefficient on R, but we strongly suspect a large OVB. Still, using our statistical software, we regress S on R and construct a 95% confidence interval for the slope coefficient. Assuming there is indeed large OVB, such a CI

○
contains the true structural slope in close to 95% of (large) samples
○
contains the true structural slope with probability less than 95%
○
is usually very wide because it is centered at the OLS estimate but incorporates uncertainty about identification/OVB
○
should be multiplied by 2 (or technically 1.96)

Question 22 1 pts

Consider the CMF model $Y=\beta_0+\beta_1 D+\beta_2 F+\beta_3 DF+V$, where Y is annual earnings in dollars, D=1 if the individual has a college degree (and D=0 otherwise), and F=1 if the individual is female (and F=0 otherwise). The interpretation of β_2 is

○
the difference between mean earnings for females without a college degree and the unconditional population mean earnings
○
the mean earnings in dollars for the subpopulation that both is female and does not hold a college degree
○
the mean earnings difference between females and non-females within the subpopulation of individuals without a college degree
○
the mean earnings difference between females and non-females

Question 23 1 pts

Let Y be hourly wage, D=1 if female (and D=0 otherwise), and X=age. You want to estimate the CMF of log wage, E[ln(Y) | D=d, X=x]. One reason the true CMF is probably not linear-in-variables is

○
there is omitted variable bias due to self-selection (by D) into different occupations
○
30-year-old make much higher wages than 20-year-olds, but 55- and 65-year-olds have very similar wages
○
individuals in urban areas tend to have steeper wage "profiles" where wage increases more rapidly with age (compared to individuals in rural areas)
○
the variables age and sex are not the most important economic determinants of wage

Question 24 1 pts

Consider a regression where Y=annual income (in tens of thousands of dollars), X=years of education beyond kindergarten, and D=1 for males and D=0 otherwise. Consider the specification $Y=\beta_0+\beta_1 X+\beta_2 D+\beta_3 XD+U$. The estimated regression function computed with OLS is 10+2X+5D-0.8XD. The positive (+5) coefficient on D suggests that

○
we have severe omitted variable bias
○
all else held equal, male incomes are higher
○
all else held equal, male incomes are lower
○
[none of these]

Question 25 1 pts

Which of the following does NOT raise concern about a threat to external validity? The goal is to learn the effects on employment of a job training program targeting all unemployed adults in Missouri.

○
Study participants were recruited from a prison in Springfield, MO by offering them early parole

○
Due to unexpected funding cuts, 1/10 of the treatment group was randomly removed from the program before the study ended

○
The trial program in the study makes treated individuals look better than others (which raises their employment probability), but if everyone did the training, then the overall state unemployment would not change

○
The funding per participant available for the large-scale program is 1/3 of that in the study

Question 26 1 pts

You have a cross-sectional dataset from 2014 with one observation per country. The Y variable is annual GDP growth (%), and the X variable is a measure of democracy, where 1=most democratic and 0=least democratic. You are interested in the slope coefficient in the linear projection of Y onto (1,X), i.e., the parameter b in LP(Y | 1,X) = a + bX. You've heard rumors that the less-democratic countries sometimes intentionally report GDP growth that is better than reality, but you have no way to actually test that hypothesis because in your dataset you only observe the GDP growth reported by each country, not the true GDP growth (unless they are identical). Let Y* be the true annual GDP growth, Y the observed/reported value, and M = Y - Y* the measurement error. If the rumor is true, then the OLS slope estimator has _____ asymptotic bias. (Hint: draw a picture.)

○
upward / positive

○
attenuation

○
downward / negative

○
zero

Question 27 1 pts

Descriptively, you're curious about the difference in college attendance rates between students who take at least one "advanced placement" (AP) class in high school (X=1) vs. students who never take any AP class (X=0). You have data on "number of AP classes taken" for every single student who graduated from high school in Missouri in Spring 2020. You also have data from every single college/university in Missouri on every student enrolled for Fall 2020. For each student in your first dataset, you assign X=0 or X=1 depending on the number of AP classes, and then you generate Y=1 if they appear in your Fall 2020 enrollment dataset or else Y=0.

Consider two points:
A. your first dataset excludes students who should have graduated from high school in Spring 2020 but dropped out (i.e., will not even graduate high school); they all should have Y=0 and (almost all) X=0, but instead they do not appear in your sample.
B. the very best Missouri high school students (who all have X=1) often attend college in other states (like Stanford in CA, Harvard in MA, etc.), so they are incorrectly coded as Y=0.

For your estimator of the mean difference E(Y|X=1) - E(Y|X=0), point A causes _____ bias, and point B causes ____ bias.

○
positive / positive

○
positive / negative

○
negative / positive

○
negative / negative

Question 28 1 pts

Descriptively, you're curious about the difference in college attendance rates between students who take at least one "advanced placement" (AP) class in high school (X=1) vs. students who never take any AP class (X=0). You have data on "number of AP classes taken" for every single student who graduated from high school in Missouri in Spring 2020. You also have data from every single college/university in Missouri on every student enrolled for Fall 2020. For each student in your first dataset, you assign X=0 or X=1 depending on the number of AP classes, and then you generate Y=1 if they appear in your Fall 2020 enrollment dataset or else Y=0.

Consider two potential problems:
A. your first dataset excludes students who should have graduated from high school in Spring 2020 but dropped out (i.e., will not even graduate high school); they all should have Y=0 and (almost all) X=0, but instead they do not appear in your sample.
B. the very best Missouri high school students (who all have X=1) often attend college in other states (like Stanford in CA, Harvard in MA, etc.), so they are incorrectly coded as Y=0.

These are best categorized as

○
A=misspecification, B=OVB
○
A=sample selection, B=measurement error
○
A=missing data, B=reverse causality / simultaneity
○
A=internal validity, B=external validity

Question 29 1 pts

Consider the time series of how many hours of sleep you got each day over the past month. At day t, the first lag is

○
Y_{t+1}
○
how much sleep you got the day before t
○
–
○
[none of these]

Question 30 1 pts

Strict stationarity implies

○
$Cov(Y_t,Y_{t+1})=Cov(Y_s,Y_{s+1})$
○
$E(Y_t \mid Y_{t-1}, Y_{t-2}, Y_{t-3}, ...) = Y_{t-1}$
○
$E(Y_t-Y_{t-1} \mid Y_{t-1}, Y_{t-2}, Y_{t-3}, ...)=0$
○
[none of these]

Question 31 1 pts

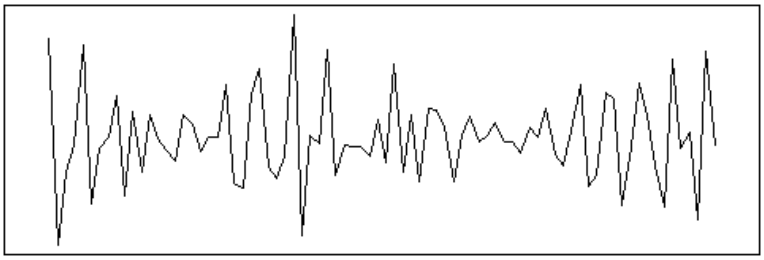
Consider a strictly stationary time series with 0.8 first autocorrelation and mean zero. One way to conceptualize this is:
A. If we could observe the series infinitely far into the future and/or past, we'd see that positive values tend to follow positive values and that negative values tend to follow negative values.
B. If we could observe an infinite number of universes with the same process generating the time series, then looking across the universes we'd see that positive values tend to follow positive values and that negative values tend to follow negative values.

○
A only
○
B only

○
both A and B
○
neither

Question 32 1 pts

The following graph shows one component of a decomposed time series (Johnson & Johnson's earnings per share); which component does it most look like?



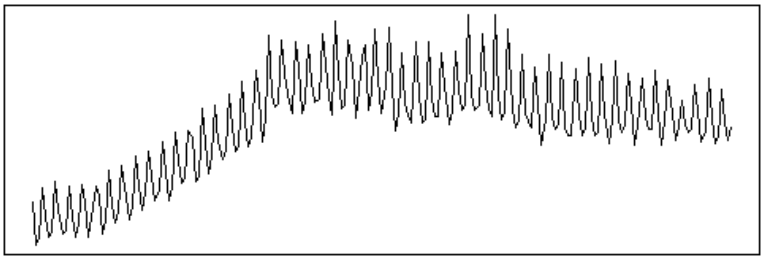
○
trend
○
seasonal
○
remainder
○
classical

Question 33 1 pts

Consider two annual time series, named A and B, both with mean zero and strictly stationary. When Series A has a negative value in year t, it often switches to be positive in year t+1. Series B tends to remain negative in t+1 if it was negative in t. In terms of their autocorrelations (e.g., A=B means same autocorrelation), this suggests

○
B > A
○
A > B
○
|A| > |B|
○
A > B for most t, but not all t

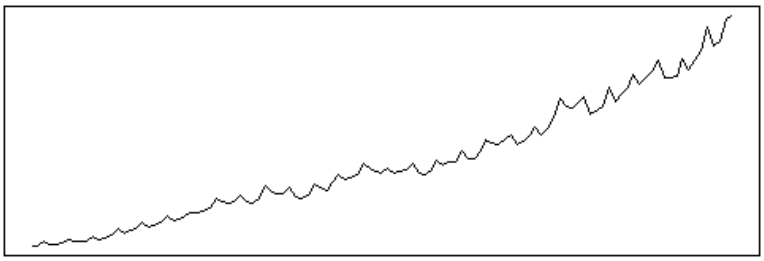
Question 34 1 pts



The most obvious nonstationary feature of the time series graphed above is

○
seasonality
○
negative covariances
○
white noise
○
[no obvious nonstationarity]

Question 35 1 pts



The most obvious nonstationary feature of the time series graphed above is

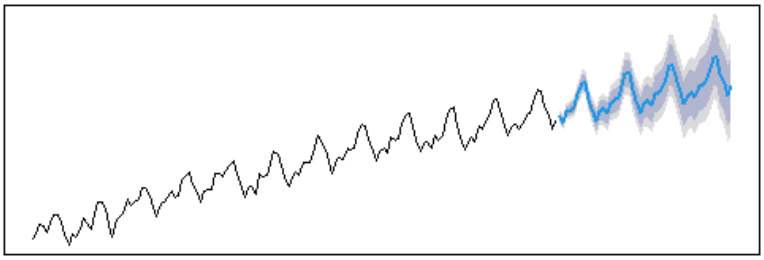
○
upward trend
○
decreasing variance
○
positive autocorrelation
○
positive autocovariance

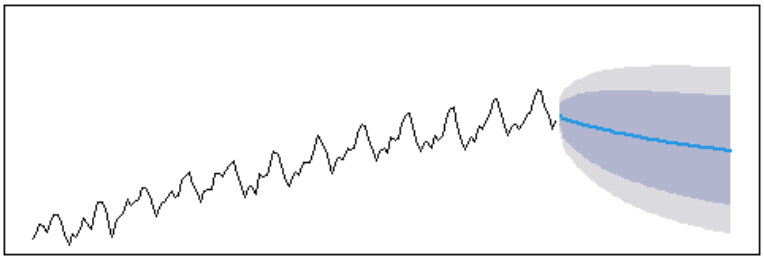
Question 36 1 pts

You have the time series data Y_t of your company's sales (in millions of dollars) in each year t. You want to model the sales growth rate. You should model

○
Y_t
○
$Corr(Y_t, Y_{t-1})$
○
$ln(Y_t)$
○
$ln(Y_t) - ln(Y_{t-1})$

Question 37 1 pts





The first (top) forecast is clearly better because

○
it assumes a random walk
○
it incorporates trend and seasonality
○
it uses an AR(1) model but estimates the parameters from the data
○
[it's not better]

Question 38 1 pts

You fit an AR(1) model to T=48 months of data (2008.m1-2011.m12) on housing starts, i.e., how many new, private, residential housing units began construction in a particular month t. You are planning to

use it to forecast next month's housing starts given this month's value. This

○
should work well since the AR(1) model implies strict stationarity

○
will probably forecast too low because housing starts were extraordinarily low during/after the Great Recession (Dec. 2007-June 2009), so the sample mean is well below the true mean

○
won't work because you need to estimate next month's error term, which is not observed in your dataset

○
[none of these]

Question 39 1 pts

You compute the AIC for AR(p) models for lag lengths p=0,1,2: AIC(0) = -3.4, AIC(1) = -3.5, AIC(2) = -3.4. According to AIC the best among these models is

○
AR(0)

○
AR(1)

○
AR(2)

○
[none of these]

Question 40 1 pts

You want to forecast a time series that is well known to have the Markov property, so you know you should only use one lag (otherwise you just add noise), and you try an AR(1) model. The R-squared is 0.34 and the AIC is 0.13. But, even though additional lags of Y aren't useful, you realize you could use other variables to improve your forecast, so you try an ADL model with the first lag of Y (like before) along with the first lags of 30 other variables in your dataset. Now the R-squared is 0.96 and the AIC is 0.75. This suggests

○
the ADL model is worse for forecasting, due to overfitting

○
R-squared and AIC agree the ADL model is better

○
you could probably improve your forecasts further by collecting more variables in order to increase the R-squared closer to 1.00 and increase the AIC further

○
the simple AR(1) is better because it has lower R-squared