Portfolio Credit Risk

University of Chicago Masters in Financial Mathematics 36702

https://uchicago.instructure.com/courses/48373

Lecture 5

Thursday 20 April 2023

Regression forecasts of the default rate

Jon Frye

JonFrye@UChicago.edu

Questions?

Are there any questions, comments, objections, whatever?

Next week: final exam

Next week: final exam

The final exam is Thursday 27 April 2023 at 6:00pm.

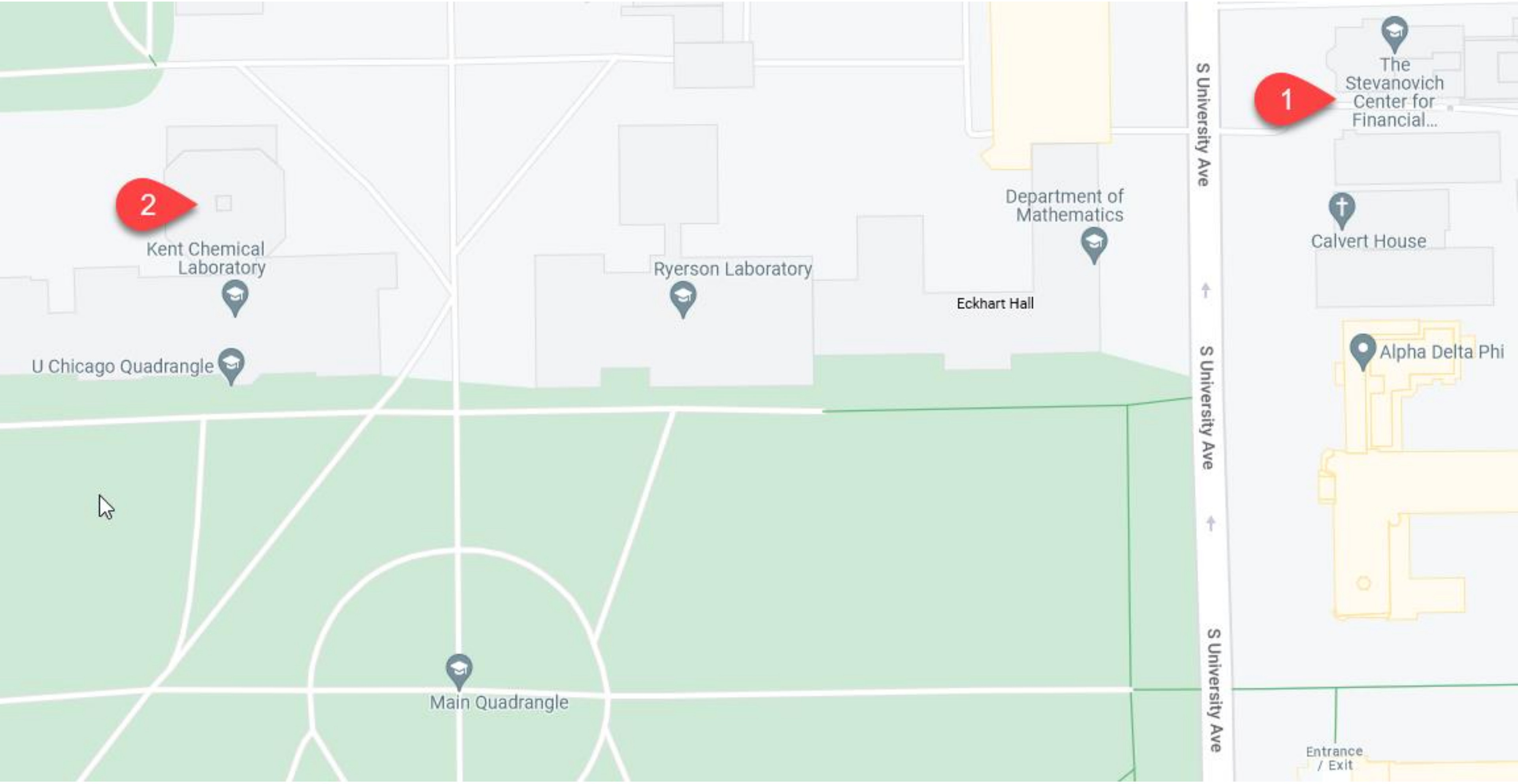
- **Location: Kent Chem Lab Room 107**
  - **Kent is on the north side of the main quadrangle.**
  - **It is about a block west of Stevanovich.**

You must bring your U of C photo ID to the exam.

There are 21 numerical-answer questions.

- **You cannot use books or written notes of any kind.**
- **You cannot use any electronic gear.**
  - **This includes everything: No laptops, tablets, cell phones, calculators…**
- **Any violations will be treated as serious.**

Please arrive at Kent 107 by 6:00



Please sit in a ***<u>designated</u>*** seat

Only certain seats in Kent 107 are OK.

After everyone is seated in a designated seat, we will distribute the exam papers.

Wait for my signal to start the exam.

Exam first page, top half:

- Write your student ID number at the top of this page
- Print your name in the space below
- Sign the agreement in the space provided

By presenting this exam for credit at the University of Chicago, you agree that:

The work submitted in this exam is my own.
I did not use any reference materials such as notes or books.
I did not use any electronic equipment such as cell phones, laptops, or tablets.
I promise not to store copies of these questions after the exam.

Print your name: ______________________________

Your signature: ________________________________

Exam first page, bottom half:

After the exam begins:

- You may not leave the exam room.
- You will have 75 minutes to complete 21 questions.
- Time is short; identify easy questions and answer them first
- If you have questions, please go to the aisle and I will meet you there

Scores:

- In the space on page 2, enter a decimal number to answer each question.
  (You can use the spaces on the other pages to perform calculations.)
- Only your first two digits affect your score.
- If both digits are correct, you score four points.
- If only the first digit is correct, you score two points.
- There is no penalty for other answers.

To pass this course:

- ***You must return every page of this exam.***
- Your student ID number must appear at the top of every page.
- You must print your name on this page and sign the agreement.

I’ll signal when to open the exam

A timer will count down from 75:00.

Write all your answers on Page 2

For each question, the correct answer is a decimal number that is accurate to two significant digits.

Answer each question with a decimal number (not a formula for a number).
Following possible leading zeros, only your first two digits affect your score.
If both digits are correct, you score four points.
If only the first digit is correct, you score two points.

Question	Answer
1	
2	
3	
4	

"two significant digits"

Example question: What is 2 / 30?

Here's how we grade answers:

- **1 / 15: Zero points. Answers must be decimal numbers.**
- **0.067: 4 points, because correct**
  - **The first digit, 6, is correct. 2 points.**
  - **The second digit, 7, is correct. 2 more points. 4 points on this question.**

    2 / 30 is approximately 0.06666666666666666... Rounded to two significant digits, this is 0.067.
- **0.67: 4 points**
  - **We grade only the digits, not the order of magnitude.**

    Please do not conclude that orders of magnitude are not important in financial math.
- **0.0666: 2 points. *Round* to two significant digits.**
- **0.06: 2 points.**
  - **To get credit for the second digit, you must write the second digit.**

$$\mathbf{0.23} = \int_{-\infty}^{\Phi^{-1}[0.5]} \int_{-\infty}^{\Phi^{-1}[0.4]} \phi_2[z_1, z_2, 0.2]\, dz_1\, dz_2$$

Value	Value	Value of PD1								
of ρ	of PD2	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.1	0.01	0.03	0.04	0.05	0.06	0.07	0.08	0.08	0.09
	0.2	0.03	0.05	0.07	0.09	0.11	0.13	0.15	0.17	0.18
	0.3	0.04	0.07	0.10	0.13	0.16	0.19	0.22	0.25	0.28
	0.4	0.05	0.09	0.13	0.17	0.22	0.25	0.29	0.33	0.37
	0.5	0.06	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.36	0.41	0.46
	0.6	0.07	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.43	0.49	0.55
	0.7	0.08	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.50	0.57	0.64
	0.8	0.08	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.57	0.65	0.73
	0.9	0.09	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.64	0.73	0.81
0.2	0.1	0.02	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.08	0.09	0.10
	0.2	0.03	0.06	0.08	0.10	0.12	0.14	0.16	0.17	0.19
	0.3	0.04	0.08	0.11	0.15	0.18	0.21	0.23	0.26	0.28
	0.4	0.05	0.10	0.15	0.19	0.23	0.27	0.31	0.34	0.37
	0.5	0.06	0.12	0.18	0.23	0.28	0.33	0.38	0.42	0.46
	0.6	0.07	0.14	0.21	0.27	0.33	0.39	0.45	0.50	0.55
	0.7	0.08	0.16	0.23	0.31	0.38	0.45	0.51	0.58	0.64
	0.8	0.09	0.17	0.26	0.34	0.42	0.50	0.58	0.66	0.73
	0.9	0.10	0.19	0.28	0.37	0.46	0.55	0.64	0.73	0.82
0.3	0.1	0.02	0.04	0.05	0.06	0.07	0.08	0.09	0.09	0.10

The exam contains a table of integral values that are treated as exact.

Exam questions

Some questions ask for PD, EL, and so forth given data:

State:	A	B	C	D
Probability of state:				
cPD:				
cLGD:				

Some questions give a PDF of cDR and an LGD function.

- **You find a distribution of cLGD or cLoss using change of variable and state a numerical fact about the distribution.**

Two ask about default correlation; know the expression.

- **Better yet: know how to derive it.**

Exam tip: Bring a pencil and a good eraser.

Time management

You have 75 minutes to complete the 21-question exam.

Expect that you will run out of time; manage time wisely.

- **Minute 0-20: Read every question and answer any easy ones.**
- **20-50: Work priorities from easy questions to hard questions.**
- **50-60: Stop solving and *<u>check</u>* your answers so far.**
  - **Find an error; provisionally give up on problems that you don’t think are easy.**
- **60-72: Work on the easiest questions that remain.**
- **72-75: Guess. There are only so many digits and no penalties.**

Don't worry afterwards. Hard exam! I grade on a curve.

- **The number correct ranges from about 1 to about 21.**

Any questions about the exam?

Or anything else?

Main topic tonight is forecasting

An additional perspective on ELGD

Regression forecasts of the default rate

A forecasting plan

Measures of forecast error

Statistical significance and bias

Model specification search and bias

Serial dependence and forecast dispersion

Forecasting with confidence

Additional perspective on ELGD

Time-weighted LGD

I use "time-weighted LGD" to illustrate how *not* to calculate a loan's expected LGD = E[LGD] = ELGD.

Example:

- **If a loan defaults in State 1, it certainly produces an LGD of 0.20. ($cLGD_1 = .2$).**
- **If the loan defaults in State 2, it certainly produces an LGD of 0.60. ($cLGD_2 = .6$).**

If $Pr_1 = Pr_2 = 0.50$, then time-weighted LGD = 0.40.

- **"Time weighted" LGD depends on the *time* that is spent in each state.**
- **But a credit risk manager cares about credit loss, and that depends on *default*.**

Expected LGD

- **ELGD places other weights on 0.20 and 0.60.**
  - **The weights involve Pr_1 and Pr_2.**
    - **If Pr_2 is high, then ELGD tends to be closer to 0.60.**
  - **The weights also involve cPD_1 and cPD_2.**
    - **If cPD_2 is high, then ELGD tends to be closer to 0.60.**

- **The weight on 0.20 is the probability that a loan is in State 1 *given* that there is a default.**
  - **We're calculating the LGD expected for a *loan*.**
  - **The weight on a value of cLGD is the probability that the loan is in that state when default happens.**
    - **The probability that the loan is in that state, period, produces time-weighted LGD.**

The weight on 0.20

$$Pr[State\ 1|D] = \frac{Pr[State\ 1 \cap D]}{Pr[D]}$$

$$= \frac{Pr[State\ 1]\ \ Pr[D|State\ 1]}{Pr[D]}$$

$$= \frac{Pr_1\ \ cPD_1}{PD}$$

ELGD written out

$$ELGD = \frac{Pr_1\ cPD_1}{\sum Pr_i\ cPD_i} 0.20 + \frac{Pr_2\ cPD_2}{\sum Pr_i\ cPD_i} 0.60$$

- **Each denominator is the sum of all the numerators.**
  - **So, the weights add up to 1.0.**
  - **Each denominator is $\sum Pr_i\ cPD_i$ = PD.**

- **Each numerator is a *product:***
  - **The default rate in a state times the probability of being in that state, as shown in the previous slide.**

- **EL = PD * ELGD.**
  - **The lender must cover EL in the loan spread.**
  - **So, a lender *wants* to know ELGD.**
    - **It also shows up in the F-J LGD function.**

Questions?

Are there any questions, comments, objections, whatever?

Regression forecasts of DR

Tonight is different

In Weeks 1-4, we developed a model of credit *<u>risk</u>*.

- **The inputs to the model are random numbers.**
- **Repeated draws produce the distribution of loss.**
  - **Comparing bank capital to the distribution quantifies the *<u>risk</u>* of failure.**

Tonight, we regress a default rate on freely chosen data.

- **There's no *<u>theory</u>*; this is "ad-hoc modeling."**
- **Such models are easy to create.**
  - **You might have imagined we would study them more.**
- **A regression produces point forecasts, not a distribution.**

Key idea: Forecasting is different from estimation.

- **To obtain good forecasts you must analyze forecasts.**

I am particularly interested in your feedback on Lec. 5.

Why this?

Expect to see a lot of ad-hoc modeling in your career.

- In a few cases, good theories exist.
  - If there's no friction or risk aversion, then the price of an option is the expected cost of hedging it.
  - If conditional credit default and conditional credit loss are comonotonic, then cLGD is a function of cPD and no other random variables.
- But in most cases the subject is people, and we lack theories.

Ad-hoc forecasts of ***credit*** ***loss*** are especially poor.

- Reason is, the modeling environment is loaded with problems.
  - There are only a few thousand "wholesale" (large) borrowers.
  - They default at a low rate, on average.
  - A PD can change materially over time. So can cPD.
  - There are only a few dozen carefully observed quarters of data.
  - Those quarters are highly serially dependent.
  - The computer treats the records as if they were independent and reports to the user a high level of confidence in the results….

The easy way to a bad forecast

A forecast plan reduces a data set to a calibrated model. Call the following the "significant regression plan."

Given historical data on DR and 3 macro variables, estimate some linear regressions with OLS.

Select the regressions having only significant slopes.

- Among them, use the one with greatest R^2.
  - If there are none, use the "null" forecast that DR equals the average rate of default contained in the data sample.

Use the preferred regression to make conditional forecasts of DR given values of the macro variables.

- Or use the null forecast and look like a dummy.

The trouble with the easy way

- The easy way uses methods good for one task,
  - namely, to calibrate a model that can be treated as correct,
- to perform an entirely different task,
  - which is to choose, among several calibrated models, the one that is likely to make the best forecasts.

- There is no warrant for this, obviously.
  - A masonry hammer can be used for dry wall.
    - But don’t expect outstanding results, and check the results are OK.
  - A better choice is a *dry wall hammer*.
  - Pro Tip: Use a tool that is made for the job you face.
    - Don’t expect a tool that is designed for model calibration to do a good job with model selection.
    - BTW, there is <u>*no good tool for model selection*</u>.

We offer a proof

Tonight, we follow the significant regression plan and make forecasts. We find that they are *reliably* poor.

- **"Reliably": The regression forecasts are poor on average over all possible data sets that might be drawn.**
- **"Poor": Better forecasts are made a different way.**

We compare two contending forecast plans.

- **One contender is the significant regression plan.**
- **The other contender is the "null" forecast plan. It predicts the default rate equals its average value in historical data.**
  - **No one thinks that the historical average is a good forecast.**
  - **That is because the it ignores economic conditions, and people know that credit loss *depends* on economic conditions.**
  - **And yet, the null plan makes better forecasts than the regression plan.**

Back to your career

You won't see many models as simple as tonight.

But you can often implement these *methods*.

- **You might be able to use them to compare a complicated model to a model that is a bit less complicated.**
  - **Depending on the assumptions you make, this might easy, or messy, or even impossible. It depends.**
  - **We keep the assumptions as simple as we can. Gauss copula!**

Don't assume that a regression forecast is good.

- **A modeler must check and see.**
  - **Tonight's methods are better than nothing.**

Already, an interim summary

The last step of a forecast plan should be to check the quality of the forecasts that result.

- **A bad forecast plan checks only *inside* the data sample.**
- **It *assumes* that forecasts are good outside the sample.**
- ***There is no warrant for this assumption.***

In other parts of life, people do check.

- **When people make shoes, they try walking in them.**

Why don't bad forecasters check for bad forecasts?

- **In the past, checking forecasts was hard to do.**
  - **Like, when the first caveperson squared the first forecast error.**
  - **But today, people have *computers*…**
- **And even now, journal referees, managers, and bank supervisors often let modelers get away with it.**

Questions?

Are there any questions, comments, objections, whatever?

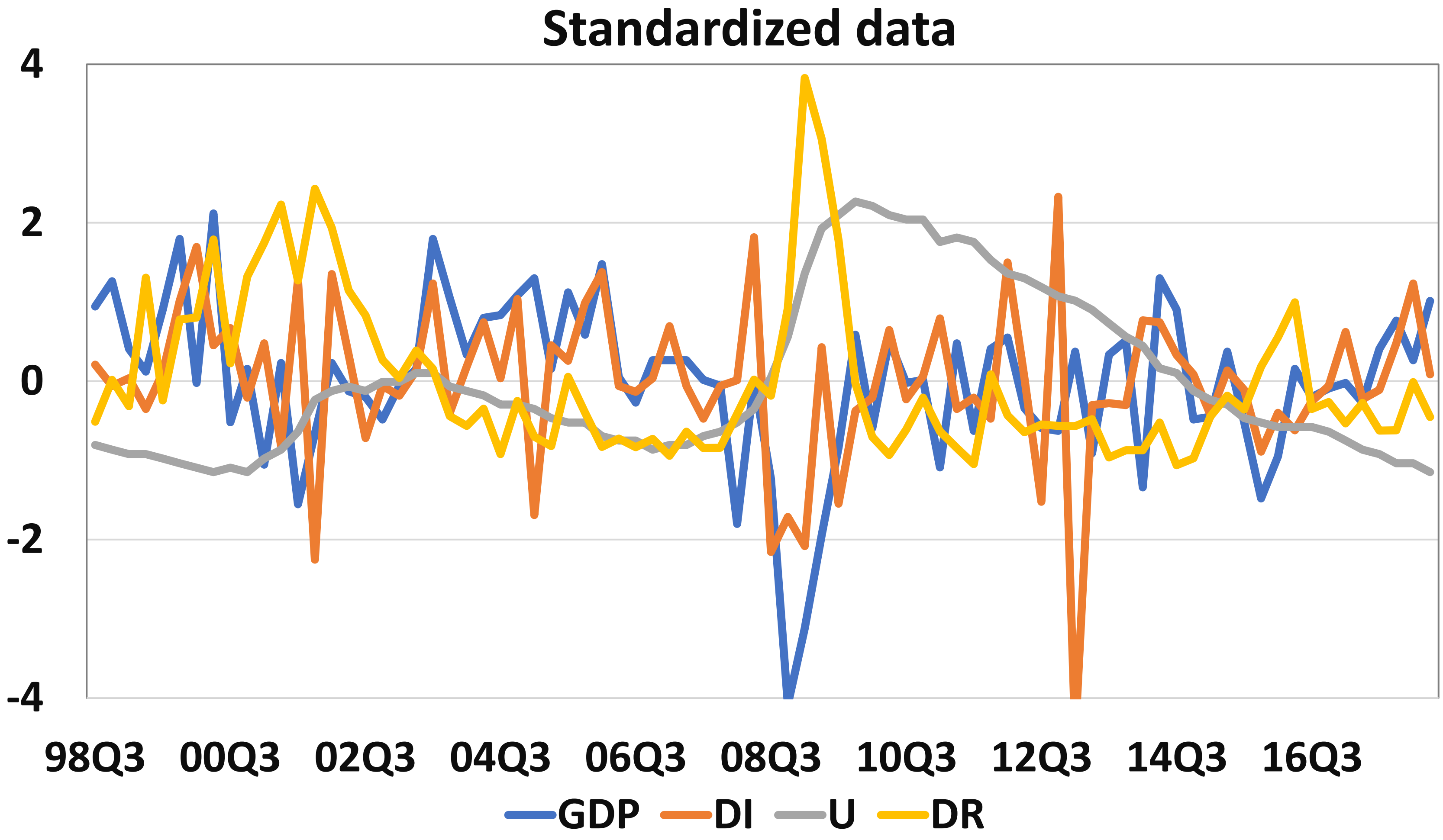
Following the forecast plan

Data set

- **The sample is N = 80 data records, 1998Q3 − 2018Q2.**
  - **(In 2018, I assembled these data for another purpose.**
    - **Neither the sample period nor the set of variables nor the definintion of the default rate were influenced by tonight's contest.)**

- **The value to be forecasted is a default rate, DR.**
  - **For US firms with a Moody's rating, each quarter I found:**
    - **The set of firms vulnerable to default at the start of the quarter.**
    - **The number of vulnerable firms is the denominator of quarterly DR.**
    - **The number that actually default is the numerator of DR.**

- **The possible explanatory variables are**
  - **GDP: Gross Domestic Product, quarterly growth**
  - **DI: Disposable Income, quarterly growth**
  - **U: Unemployment rate**


Standardized data
4
2
0
-2
-4
98Q3
00Q3
02Q3
04Q3
06Q3
08Q3
10Q3
12Q3
14Q3
16Q3
GDP
DI
U
DR

Seven models to be estimated

Estimated model	R^2
$DR = .030 - 0.217^* \, GDP$	0.116
$DR = 0.024 - 0.070 \, DI$	0.026
$DR = 0.022 - 0.031 \, U$	0.001
$DR = 0.031 - 0.205^* \, GDP - 0.028 \, DI$	0.120
$DR = 0.039 - 0.241^* \, GDP - 0.138 \, U$	0.133
$DR = 0.028 - 0.075 \, DI - 0.066 \, U$	0.029
$DR = 0.041 - 0.227^* \, GDP - 0.036 \, DI - 0.148 \, U$	0.139

A star* indicates statistical significance at 5% level.

The only regression showing significance in all its coefficients is the first one.

That regression becomes the forecast equation.

The easy way stops here

For the modeler, it is time to kick back and relax.

But we'll see that the regression forecasts are worse than those of the null forecast plan.

- **This seems crazy because the default rate is surely *<u>not</u>* independent of macro variables. But it is true.**

We show this using simulation.

- **Simulation reveals problems not seen by the modeler.**
- **It is better to find out problems in simulation than to depend on forecasts that turn out to be faulty. Yes?**

Questions at this point?

Measures of forecast error

Expected Squared Error, ESE

Expected Squared Error $= \boldsymbol{ESE} = \boldsymbol{E_{X,Y}}\ [\ (\boldsymbol{Y} - \boldsymbol{f}[\boldsymbol{X}])^{\mathbf{2}}\ ]$**,**
where $\boldsymbol{Y}$ **is the variable of interest and**

$\boldsymbol{X}$ **is a vector of explanatory variables.**

$\boldsymbol{X}$ **and** $\boldsymbol{Y}$ **are jointly normal**

The forecasting plan calls for a linear function $\boldsymbol{f}$**:**

$$\boldsymbol{ESE} = \boldsymbol{E_{X,Y}}\ [\ (\boldsymbol{Y} - (\boldsymbol{a} + \boldsymbol{b_1 X_1} + \boldsymbol{b_2 X_2} + \boldsymbol{b_3 X_3}))^{\mathbf{2}}\ ].$$

This quadratic expansion has 15 terms. Each term can be simply expressed. Example:

$$\boldsymbol{E}\ [\mathbf{2}\boldsymbol{b_1 b_2 X_1 X_2}\ ] = \mathbf{2}\boldsymbol{b_1 b_2}\left(\ \boldsymbol{\mu_1 \mu_2} + \boldsymbol{\rho_{1,2} \sigma_1 \sigma_2}\ \right)^{*}$$

So, when $\boldsymbol{f}$ **is linear and the population is jointly normal, ESE reduces to high-school math.**

*

Let X and Y be two random variables.

Their covariance is defined this way:

$$\boldsymbol{Cov[X,Y] = E[\,(X - E[X])\,(Y - E[Y])\,]}$$

$$\boldsymbol{= E[\,XY - X\,E[Y] - Y\,E[X] + E[X]\,E[Y]\,]}$$

$$\boldsymbol{= E[\,XY\,] - E[X]\,E[Y]}$$

Therefore, $\boldsymbol{E[\,XY\,] = Cov[X,Y] + E[X]\,E[Y]}$

If X and Y are jointly *<u>standard</u>* normal, then $\boldsymbol{E[X\,Y] = \rho}$**.**

$$ESE = E_{X,Y}\left[\,(Y - f[X])^2\,\right]$$

ESE depends on

- **a forecast function, f, and the distribution of the variables.**
  - **ESE is *not* a function of data; it is an "out-of-sample" error measure.**
  - **Its value is unknown to a modeler working on a practical problem because the modeler does not know the distribution of the variables.**

Suppose that values of DR and GDP were drawn from a joint standard normal distribution and $Corr = 0.5$. Then, the ESE of the preferred regression is this:

$$ESE = E_{DR,GDP}\left[\,(DR - (.030 - 0.217\ GDP))^2\,\right]$$

$$= E[.0009 - .06\ DR + DR^2 - .1302\ GDP + .434\ DR\ GDP + .47089\ GDP^2]$$

$$= .0009 - 0 + 1 - 0 + 0.434 * 0.5 + 0.47089 = 1.6879.$$

Note: this ESE is huge. The regression line points the *wrong way*.

ESE and MSE

ESE is the population equivalent of MSE.

- **MSE is the *average* squared error on a *finite* set of points.**
  - **The linear f that minimizes MSE on a data set is the OLS estimate.**
- **ESE is the *expected* squared error on the joint *distribution*.**
  - **We call the linear f that minimizes ESE the "population regression".**

ESE is the weighted average of every possible forecast.

- **A model can have high ESE and still make good forecasts.**
  - **By luck, that is.**

Questions at this point?

Out-of-sample R^2

Closely related to ESE is out-of-sample R^2:

$$R^2 = 1 - \frac{MSE[regression\ forecast\ function]}{MSE[null\ forecast\ function]}$$

$$oosR^2 = 1 - \frac{ESE[regression\ forecast\ function]}{ESE[null\ forecast\ function]}$$

Keep in mind:

- When a regression is optimized to a data set,
  - the regression always does better than the null forecast, and
  - R^2 is always non-negative.
- But $oosR^2$ can be negative; then, $ESE[null] < ESE[regression]$.
  - That is, iff $oosR^2$ is negative then the null forecast beats regression.

Population regression

Let (X, Y) be bivariate normal:

$$ESE[\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho, a, b] = E_{X,Y}\,[\,(Y - a - b\,X)^2\,]$$

$$= a^2 + 2ab\mu_X + b^2\mu_X^2 - 2a\mu_Y - 2b\mu_X\mu_Y + \mu_Y^2 + b^2\sigma_X^2 - 2b\rho\sigma_X\sigma_Y + \sigma_Y^2$$

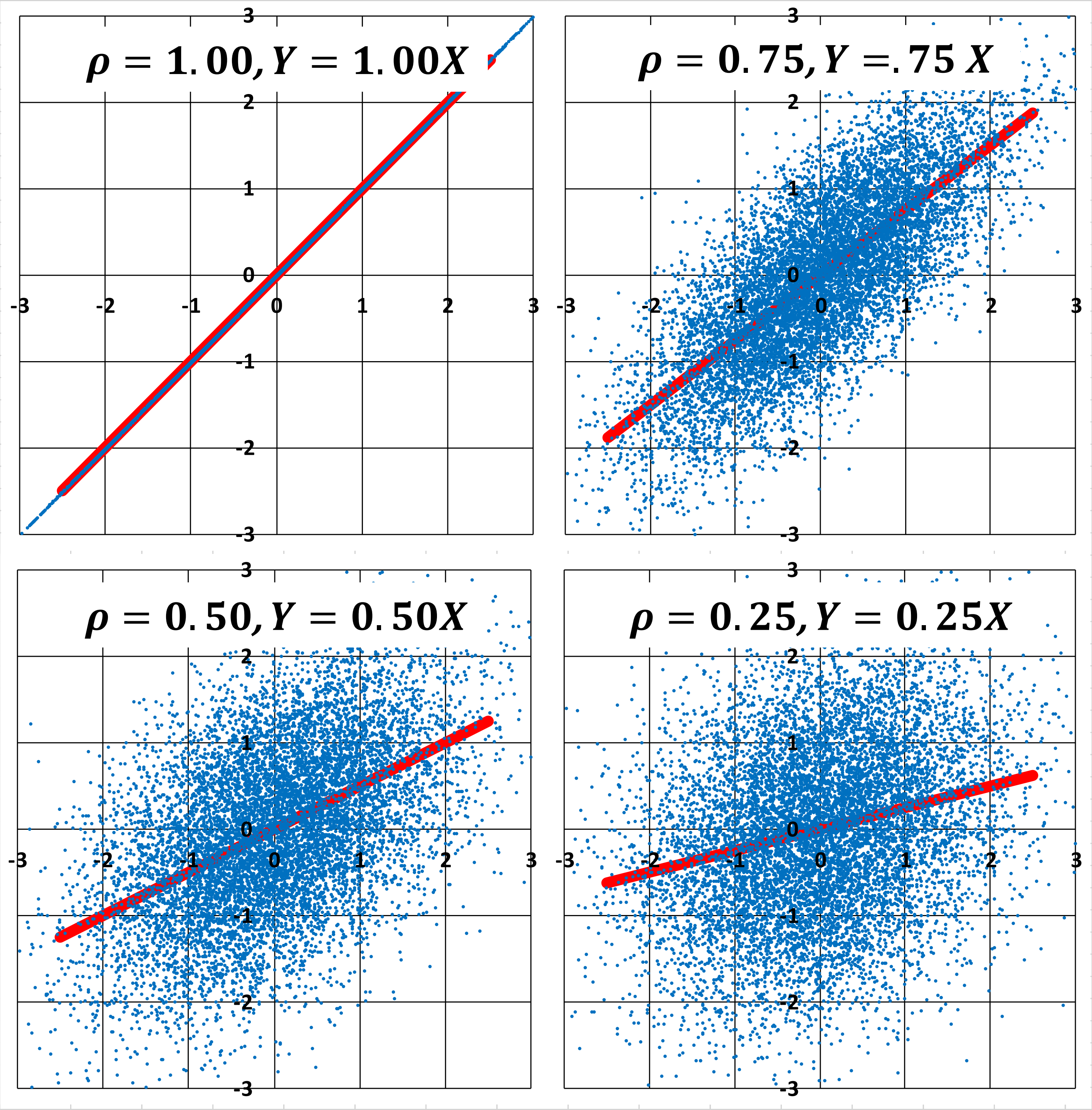
The terms with a^2 or b^2 have positive signs; there's a *minimum*.
We call the minimizing coefficients the "population regression."

Special Case: If (X, Y) is bivariate *standard* normal, then

$$ESE[0, 0, 1, 1, \rho, a, b]$$

$$= a^2 + b^2 - 2b\rho + 1 = a^2 + (b - \rho)^2 - \rho^2 + 1$$

This is minimized at $a = 0$, $b = \rho$; therefore, for standard normal variables the population regression is $Y = 0 + \rho X$.



X, Y are joint standard normal variables.

The principal axis of a data swarm is 45°.

The slope of the population regression of Y on X equals ρ.

Forecast ***plan*** ***ESE***

ESE depends on a forecast *equation*, $f[\cdot]$.

Plan ESE depends on a forecast *plan*.

- **To estimate *plan* ESE, take the average of this simulation run:**
  - **Draw a random simulated data sample, apply the forecast plan to find the forecast equation, find its value of ESE, and keep track.**

The modeler chooses between forecast *plans*.

- **We assume the modeler prefers the plan with least plan ESE.**
  - **For an assumed population, simulation reveals which plan does this.**

Questions?

ESE and plan ESE are the main tools.

Are there any questions, comments, objections, whatever?

The three contests

We simulate the search for significance and R^2.

- **Search produces biased estimates that harm forecasts.**
  - **How important is that?**

We simulate the effect of serially dependent data.

- **Serial dependence increases the variance of estimates.**
  - **The extra variance increases plan ESE.**
  - **How important is that?**

We allow the population parameter values to differ from their MLEs.

- **How confident are we that the significant regression plan can clear this hurdle?**

The search for significance

In Lecture 3, scientists searched for significance in tests of the effects of jellybean color on acne.

- **Among 20 colors of jellybeans, the green ones were found to have a significant relation to acne.**
- **The true effect − zero − of the green ones was *overstated*.**

The search for significance produces an *overstatement* of effects and a bias in the associated coefficient.

Although regression parameters are estimated by OLS, and although the OLS estimator is unbiased, the search for significance introduces *bias*…

Significance and bias

An estimate that is significantly different from zero tends to be further from zero than other estimates.

Suppose an unbiased estimator is used repeatedly on a large collection of independent data samples.

- The average of all estimates is accurate, but:
  - The average estimate with significance is too far from zero.
  - The average estimate without significance is too close to zero.

If the modeler chooses models with significance, then the chosen models have biased estimates.

QED.

But it can be confusing

The ad-hoc modeler chooses $DR = .030 - 0.217^{*}\ GDP$.

The regression contains a slope estimate, -0.217.

The slope estimate is calculated by OLS.

OLS is unbiased.

The estimate, -0.217, is statistically significant.

Note: The test does _not**_ change the estimate.**

Is the estimate unbiased?

Trick question!

Bias is *<u>not</u>* a property of an estimate!

- **An estimate might be high or low.**
- **An estimate might be accurate or inaccurate.**
- **An estimate might be significant or not significant.**
- **But an estimate *<u>cannot</u>* be biased or unbiased.**

We can ask, “Is the *<u>estimator</u>* unbiased?” Answer:

- **If the estimator is OLS (and assumptions fulfilled), *<u>yes</u>*.**
- **If the estimator is OLS plus some other condition,**
  - **such as in this case, the estimate must be significant,**
- **then the estimator is *<u>not</u>* OLS.**
  - **The not-OLS estimator should *<u>not</u>* be expected to be unbiased.**
  - **The bias of the significant-regression estimator is away from zero.**

See for yourself

Suppose that Y and X are jointly standard normal:

$$\begin{bmatrix} Y \\ X \end{bmatrix} \sim N\left[\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}\right]$$

Using a finite data set, estimate the model with OLS:

$$Y = \widehat{a} + \widehat{b}X$$

Keep track whether the $\widehat{b}$ estimate displays significance.

Repeat 1,000 times to get 1,000 estimates of b.

- **Partition those that are significantly different from zero from those not significantly different from zero. Take averages.**
- **The first average tends to be further from zero than ρ.**

If the modeler chooses models with significance, then the models that are chosen have biased estimates.

Questions?

I had to find out for myself that the requirement of statistical significance introduces estimation bias.

The model that fits best

- **For some data samples, more than one of the seven regressions have all slopes significant.**
  - **Then, the significant regression plan chooses the regression with the greatest value of R^2.**
    - **This second stage of search introduces more bias.**

- **The search for high R^2 introduces more bias. Reason:**
  - **R^2 measures the fraction of variance explained by an equation.**
  - **If a coefficient estimate is close to zero, that limits the equation's power to explain variance.**
    - **Such a model is not likely to have greater R^2 than all other models.**
    - **$\therefore$, the model with greatest R^2 tends to have estimates *<u>not</u>* close to zero.**
    - **The estimates in *<u>chosen</u>* models are *<u>further</u>* *<u>biased</u>* away from zero.**

See for yourself

Simulate a finite sample from jointly standard normal variables with a 3 equal correlations.

Regress Y on each of the two X's using OLS.

- Place the regression with greater R^2 into Collection A, and place the regression with lesser R^2 into Collection B.

Repeat that step many times.

- The average slope in Collection A is apt to be further from 0.
  - You can add a step that checks for significance in advance, if you like.

Neither collection uses the OLS estimator.

- Even though every estimate is calculated by OLS.

Questions?

The assumed population

We quantify the effect of these biases on forecasts.

We assume that data are drawn from a *multinormal* population with parameter values equal to the estimated values.

- **This is nearly a best-case assumption for the modeler.**

			Correlations			
	Mean	SD	GDP	DI	U	DR
GDP	0.043	0.028	1	0.297	-0.277	-0.341
DI	0.045	0.041	0.297	1	-0.197	-0.160
U	0.059	0.018	-0.277	-0.197	1	-0.030
DR	0.021	0.018	-0.341	-0.160	-0.030	1

First simulation experiment

We draw 1,000 independent 80-record samples from the assumed population. For each sample we find:

- **The regression chosen by the significant regression plan**
  - **and the ESE of the regression forecast.**
- **The sample average value of DR**
  - **and the ESE of the null forecast.**

The next slide summarizes the experiment.

- **For each of the model specifications it shows:**
  - **The population regression coefficients**
  - **The average estimated coefficients**
  - **Average measures of goodness of fit: $\boldsymbol{R^2}$ and $\boldsymbol{oosR^2}$.**

Ad-hoc model, independent records

	Population slopes			Average estimates			Average fit	
Cases	GDP	DI	U	GDP	DI	U	R^2	$oosR^2$
Significant regression forecast plan								
87							0.00	0.00
587	-0.22			-0.24			0.14	0.11
42		-0.07			-0.13		0.10	0.00
6			-0.03			-0.30	0.08	-0.07
40	-0.21	-0.03		-0.21	-0.09		0.21	0.06
201	-0.24		-0.14	-0.27		-0.28	0.20	0.10
3		-0.08	-0.07		-0.12	-0.31	0.13	-0.03
34	-0.23	.-04	-0.15	-0.21	-0.11	-0.30	0.23	0.07
1000	Significant regression plan averages:						0.15	0.09

Slope estimates tend further from zero than true values.

In-sample values of R^2 exaggerate out-of-sample R^2.

Regression $oosR^2$ is <u>*positive*</u>; model outperforms null.

Significance and bias review

Estimation bias is introduced by the search for significance and the search for elevated R^2.

Bias is one reason that the "easy way" of modeling a default rate produces bad forecasts.

- **Still, the $oosR_2$ of the significant regression plan is positive; we haven't shown the plan is *bad*.**

Are there any questions, comments, whatever?

Questions?

Serial dependence

Forgetting to flip the coin

Flip a fair coin.

Make a record of the face that shows.

Make an additional record of the face that shows.

- **Repeat until you have a data set of 100 records.**
- **All 100 records agree; either all H or all T.**

The MLE of the probability of Heads is either 0 or 1.

This estimator is *unbiased*: the expected estimate is ½.

The trouble is, the 100 records are *dependent*.

- **Records reflect other records instead of reflecting the *coin*.**
- **This elevates the *variance* of the estimator.**
- **Highly variable *estimates* imply poor *forecasts*.**

Serially dependent data

- **Record the quarterly default rate.**
- **Next quarter, record the quarterly default rate…**
  - **Repeat 100 times. Although a default rate is not completely unchanging like a tossed coin, it changes *slowly*.**
    - **The default rate has (positive) serial dependence.**
    - **It reflects the past to a great degree.**
    - **The data does *not* reflect only the distribution being estimated.**
- **The estimator of PD is *unbiased* but has high *variance,***
  - **compared to when the data sample is a random sample.**
  - **Forecasts are poor because estimates are highly variable.**
- **The effect of serially dependent data is much like a reduction in the number of data records.**

Serial dependence preview

- **Serially dependent data degrade forecast quality.**
  - **The problem is the *variance* of the parameter estimates.**

- **We analyze the effect in two steps.**
  - **We measure serial dependence out to three lags.**
    - **There is no super-good reason for 3 rather than something else.**
  - **Then we simulate multinormal data having the same pattern of dependence as the real data.**
    - **We run the contest as before.**

Questions?

Measure serial dependence

	Correlation with spot and once-lagged values							
	GDP	DI	U	DR	GDP_{-1}	DI_{-1}	U_{-1}	DR_{-1}
GDP	1	0.30	-0.26	-0.34	0.42	0.42	-0.15	-0.21
DI	0.30	1	-0.20	-0.16	0.33	-0.18	-0.14	-0.02
U	-0.26	-0.20	1	-0.04	-0.35	-0.24	0.98	0.05
DR	-0.34	-0.16	-0.04	1	-0.56	-0.16	-0.16	0.75
GDP_{-1}	0.42	0.33	-0.35	-0.56	1	0.30	-0.25	-0.34
DI_{-1}	0.42	-0.18	-0.24	-0.16	0.30	1	-0.20	-0.16
U_{-1}	-0.15	-0.14	0.98	-0.16	-0.25	-0.20	1	-0.04
DR_{-1}	-0.21	-0.02	0.05	0.75	-0.34	-0.16	-0.04	1

DR has stronger correlation with its own past than with any macro variable, spot or lagged.

The urge to change the contest

An ad hoc modeler would now get inspired:

- "Hey! There is a strong correlation between DR and lagged DR! I'm putting lagged DR on the right-hand side!"

Even if this is a good idea, tonight so far we have worked with a well-defined forecast plan. Let's stick to it.

- The plan doesn't allow for the modeler to change the plan.

BTW, a ***huge*** number of models depend on the past.

- The matrix has DR and 15 candidate variables.
  - A forecasting plan must choose between a set of candidates based in some way on the data sample. This allows the plan to be ***simulated***.
  - It is hard to avoid the urge to tinker with the model, based on data.
  - The range of possible tinkering should be built into the forecast plan.
  - The plan must not depend on forecasts; the plan is to be ***judged*** by them.

Simulate serially dependent data

We continue to run the contest defined before, this time using data that mimic the observed pattern of dependence.

- **There are sixteen variables all together: spot and 3 lagged values of each of four random variables.**
  - **There are 120 unique correlations all told.**

Define the column vectors $\boldsymbol{X_1}$ and $\boldsymbol{X_2}$:

$$\boldsymbol{X_1 = [GDP, DI, U, DR]'}$$
$$\boldsymbol{X_2 = [GDP_{-1}, DI_{-1}, U_{-1}, DR_{-1}, GDP_{-2}, \ldots, DR_{-3},]'}$$

When the values in $\boldsymbol{X_1}$ are determined in the current quarter, they depend on the past values in $\boldsymbol{X_2}$.

- **We want $\boldsymbol{PDF_{X_1|X_2}[X_1|X_2]}$.**
  - **$\boldsymbol{X_1}$ is vector of 4 joint normals conditioned on a vector of 12 joint normals.**

The conditional PDF

Symbolize the means and variances:

$$\boldsymbol{E}\begin{bmatrix}\boldsymbol{X_1}\\ \boldsymbol{X_2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\mu_1}\\ \boldsymbol{\mu_2}\end{bmatrix}; \quad \boldsymbol{Var}\begin{bmatrix}\boldsymbol{Z_1}\\ \boldsymbol{Z_2}\end{bmatrix} = \begin{bmatrix}\boldsymbol{\Sigma_{1,1}} & \boldsymbol{\Sigma_{1,2}}\\ \boldsymbol{\Sigma_{1,2}}' & \boldsymbol{\Sigma_{2,2}}\end{bmatrix}$$

A standard result is (see D. F. Morrison's textbook):

$$\boldsymbol{E}[\boldsymbol{X_1}|\boldsymbol{X_2}] = \boldsymbol{\mu_1} + \boldsymbol{\Sigma_{1,2}}\,\boldsymbol{\Sigma_{2,2}^{-1}}\,(\boldsymbol{X_2} - \boldsymbol{\mu_2})$$

The unconditional expectation, $\boldsymbol{\mu_1}$, must be adjusted for the departure of past values from past expected values.

$$\boldsymbol{Var}[\boldsymbol{X_1}|\boldsymbol{X_2}] = \boldsymbol{\Sigma_{1,1}} - \boldsymbol{\Sigma_{1,2}}\,\boldsymbol{\Sigma_{2,2}^{-1}}\,\boldsymbol{\Sigma_{1,2}'}$$

The conditional variance of $\boldsymbol{X_1}$ does not depend on $\boldsymbol{X_2}$.

- **Similarly, in a regression of Y on X the variance of the error does not depend on X (unless variables are heteroscedastic).**

Questions?

Hey, why not just randomize the sequence of data records and forget about serial dependence?

Simulate serially dependent data

Replace the μ's and Σ's in Morrison by their MLEs.

- We don't bother to impose stationarity.

Simulate X_1 one quarter at a time

- In quarter 1,
  - Define $X_2 = \mu_2$ to initialize.
  - Draw X_1 from the joint normal distribution shown on the previous slide.
- In quarter 2,
  - Redefine X_2 to contain the simulated values from quarter 1 and the two deeper-lagged initialization values of μ_2.
  - Draw X_1 from the joint normal distribution shown on the previous slide.
- …, and so forth for 100,100 quarters.

Discard the first 100 quarters because they depend a bit on the initialization.

Second simulation experiment

First, check that all is well. Estimate the variance-covariance matrix of the 100,000 records together.

- Each 4x4 principal submatrix should look like the VCV of the real data. If not, there is probably an error somewhere.

Partition the 100,000 records into 1,000 sets of 100.

Discard the first 20 records of each set.

- This reduces the dependence between data sets.
- The result is 1,000 data sets each with 80 dependent records.

Re-run the contest using the sets of ***dependent*** records.

Modeling with dependent records

	Ind. Records		Dep. Records	
Cases	R^2	$oosR^2$	R^2	$oosR^2$
87	0.00	0.00	0.00	0.00
587	0.14	0.11	0.18	0.07
42	0.10	0.00	0.06	0.02
6	0.08	-0.07	0.11	-0.24
40	0.21	0.06	0.29	0.01
201	0.20	0.10	0.25	-0.13
3	0.13	-0.03	0.17	-0.10
34	0.23	0.07	0.32	-0.07
Avgs.	0.15	0.09	0.19	-0.001

R^2 tends to be *greater* than with independent data.

$oosR^2$ tends to be less than with independent data.

Plan $oosR^2$ is negative: regression loses by a little.

Questions?

Cumulative review

The modeler employs the significant regression plan using 80 serially dependent records of DR, GDP, DI, and U.

We simulate this repeatedly.

- **We assume that the population parameters have the values estimated from the modeler's data.**
  - **This includes three lagged values of DR, GDP, DI, and U.**

The forecasts of the significant regression strategy are outperformed by the sample average.

- **Any suggestions for what the modeler might say, given this?**

Almost surely…

Almost surely, the parameters of the population are *<u>not</u>* equal to the values estimated from the data sample.

- **We started with a real-world data sample. It might be drawn from a population that is different from its MLEs.**
  - **Maybe a lot different!**

On average, this departure should harm the forecasts of each forecast plan.

- **But I imagine that the significant regression plan is harmed *<u>more</u>* because it is more carefully fitted to the data.**

Don't have this in the deck yet. Come back next year!

Thank you!

I love teaching this course.

After the exam, contact me anytime about anything.

Best of luck at U of C and in your career.