Applied Probability and Statistics

- Unit 1: Introduction to Probability
- Unit 2: Probability Distributions
- Unit 3: Statistical Inference
- Unit 4: Introduction to Linear Regression
- Unit 5: Regression Analysis
- **Unit 6: Regression Modeling**

Announcements

- Past:

  Unit 5 Team Assignment solutions available and scores posted.

- Present:

  There is **no assignment due**. (Practice exam problems released later today.)

- Future:

  Final Exam Period: **December 2 - December 11**

Modeling categorical (non-numerical) data:

- dummy variables: each category represented by its own 0-1 variable
  - if N categories, use at most $N-1$ dummy variables (to avoid perfect multicolinearity)
  - dummy variable coefficient provides an adjustment for the intercept coefficient
- slope dummies: interaction effect (dummy)$\times$(ind.variable)
  - slope dummy coefficient provides an adjustment for the slope coefficient
- segmentation (stratification): segment the data in subsamples (issues with interpretation of SE)

Modeling non-linear relations:

- could transform any ind. variable X_j so that Y and $f(X_j)$ are linearly related
  - regression gives optimal slope coefficient for $f(X_j)$
- could transform dependent variable Y to ensure error assumptions are satisfied
  - requires caution with interpretation of forecasts
- logarithmic transformation measures orders of magnitude

Time series forecasting:

- trend: use time as an independent variable (e.g., use quarter number); time as a proxy
- extrapolation: use lagged variables as independent variables (e.g., use sales from k periods ago)
- modeling exponential growth: forecast logarithm of the dependent variable (e.g., Y=ln(Sales))
- modeling seasonality: use season dummies or appropriately lagged variable(s)

Unit 5 Team Assignment revisited

- Based on the available data, which of the experts would you suggest the Coffee Team not to retain? Why? How can regression analysis be used to justify your answer?

300 310 330

$$P = w_0 + w_{MI} MI + w_{GR} GR + w_{NS} NS$$

Unit 5 Team Assignment revisited

- Based on the available data, which of the experts would you suggest the Coffee Team not to retain? Why? How can regression analysis be used to justify your answer?

	All Three
Std. Err. Reg.	**3.439**
Regression Coefficients: Beta (p-value)	
Intercept	6.317 (0.006)
Gabriela Rodrigues	0.759 (0.000)
Neil Stevens	-0.047 (0.552)
Mario Illy	0.255 (0.000)

Unit 5 Team Assignment revisited

- Based on the available data, which of the experts would you suggest the Coffee Team not to retain? Why? How can regression analysis be used to justify your answer?

	All Three	**Rodrigues**	**Stevens**	**Illy**
Std. Err. Reg.	**3.439**	**7.881**	**8.104**	**19.345**
Regression Coefficients: Beta (p-value)				
Intercept	6.317 (0.006)	11.343 (0.024)	10.295 (0.047)	37.835 (0.001)
Gabriela Rodrigues	0.759 (0.000)	0.934 (0.000)		
Neil Stevens	-0.047 (0.552)		0.940 (0.000)	
Mario Illy	0.255 (0.000)			0.769 (0.000)

Unit 5 Team Assignment revisited

- Based on the available data, which of the experts would you suggest the Coffee Team not to retain? Why? How can regression analysis be used to justify your answer?

	All Three	**Rodrigues**	**Stevens**	**Illy**	**Rodrigues and Stevens**	**Rodrigues and Illy**	**Stevens and Illy**
Std. Err. Reg.	**3.439**	**7.881**	**8.104**	**19.345**	**6.899**	**3.409**	**7.197**
Regression Coefficients: Beta (p-value)							
Intercept	6.317 (0.006)	11.343 (0.024)	10.295 (0.047)	37.835 (0.001)	9.268 (0.036)	6.275 (0.005)	8.819 (0.056)
Gabriela Rodrigues	0.759 (0.000)	0.934 (0.000)			0.496 (0.000)	0.721 (0.000)	
Neil Stevens	-0.047 (0.552)		0.940 (0.000)		0.451 (0.001)		0.794 (0.000)
Mario Illy	0.255 (0.000)			0.769 (0.000)		0.247 (0.000)	0.156 (0.002)

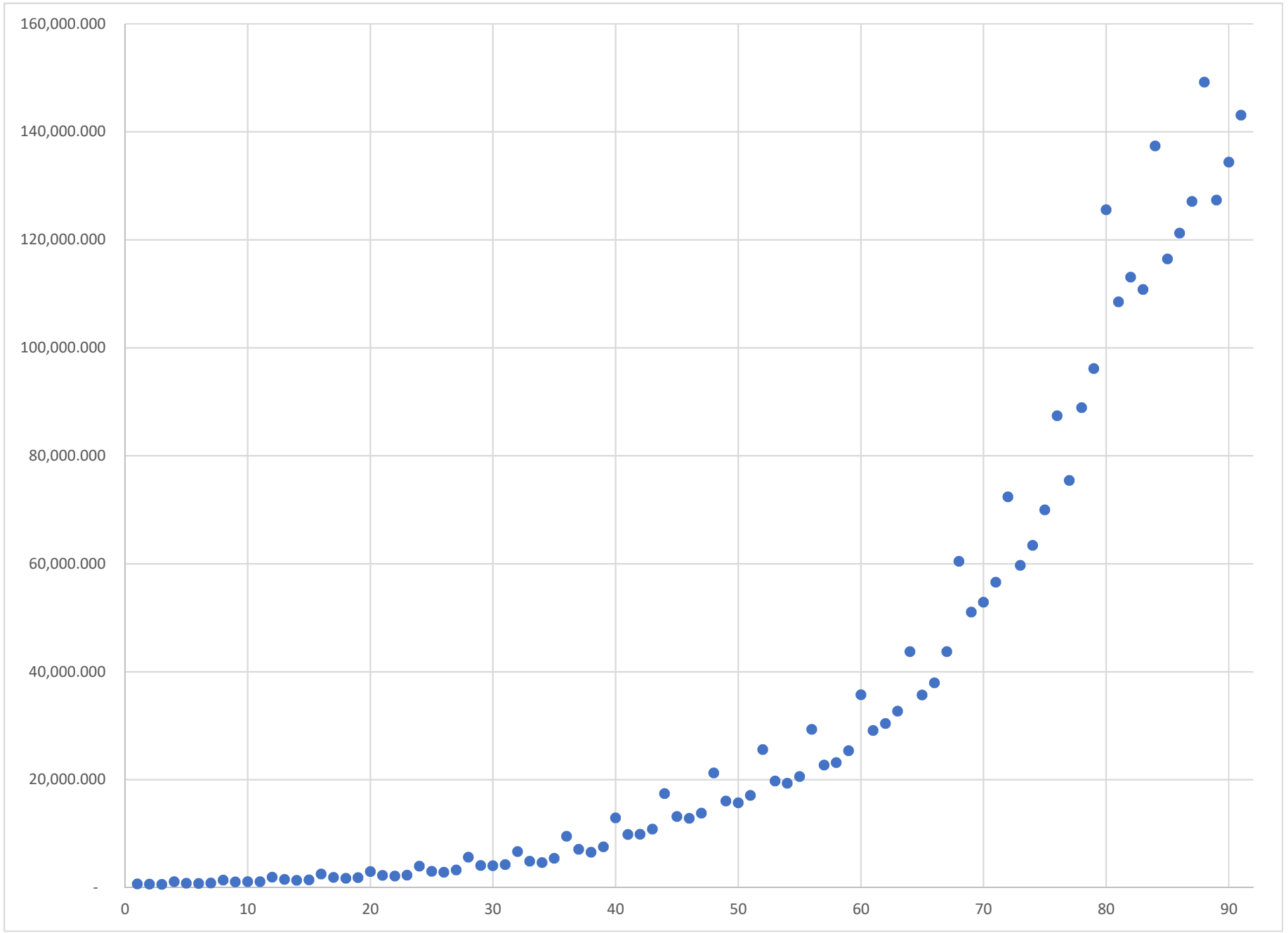
Combining forecasts

- Conceptual advantages of using multiple approaches
  - Look at problem from different angles
  - Many times there is no "true model"
- Tangible advantages in combining forecasts
  - Improved accuracy (like including more information)
  - Risk reduction
- Disadvantages
  - Costs of acquiring different perspectives
  - More sophisticated analysis needs to be conducted/communicated

Amazon's quarterly revenue

Amazon's quarterly revenue (variable QR, millions of US dollars) is in the file *AmazonQuarterlyRevenue_Data*.

Goal: Forecast Amazon's revenue for Q4 2023. (Will be reported at the end of January 2024)


160,000.000
140,000.000
120,000.000
100,000.000
80,000.000
60,000.000
40,000.000
20,000.000
-
0
10
20
30
40
50
60
70
80
90

$$QR = \beta_0 + \beta_1 LAG4QR$$

$$LNQR = \beta_0 + \beta_1 LAG4LNQR$$

Amazon's quarterly revenue

Amazon's quarterly revenue (variable QR, millions of US dollars) is in the file *AmazonQuarterlyRevenue_Data*.

Goal: Forecast Amazon's revenue for Q4 2023. (Will be reported at the end of January 2024)

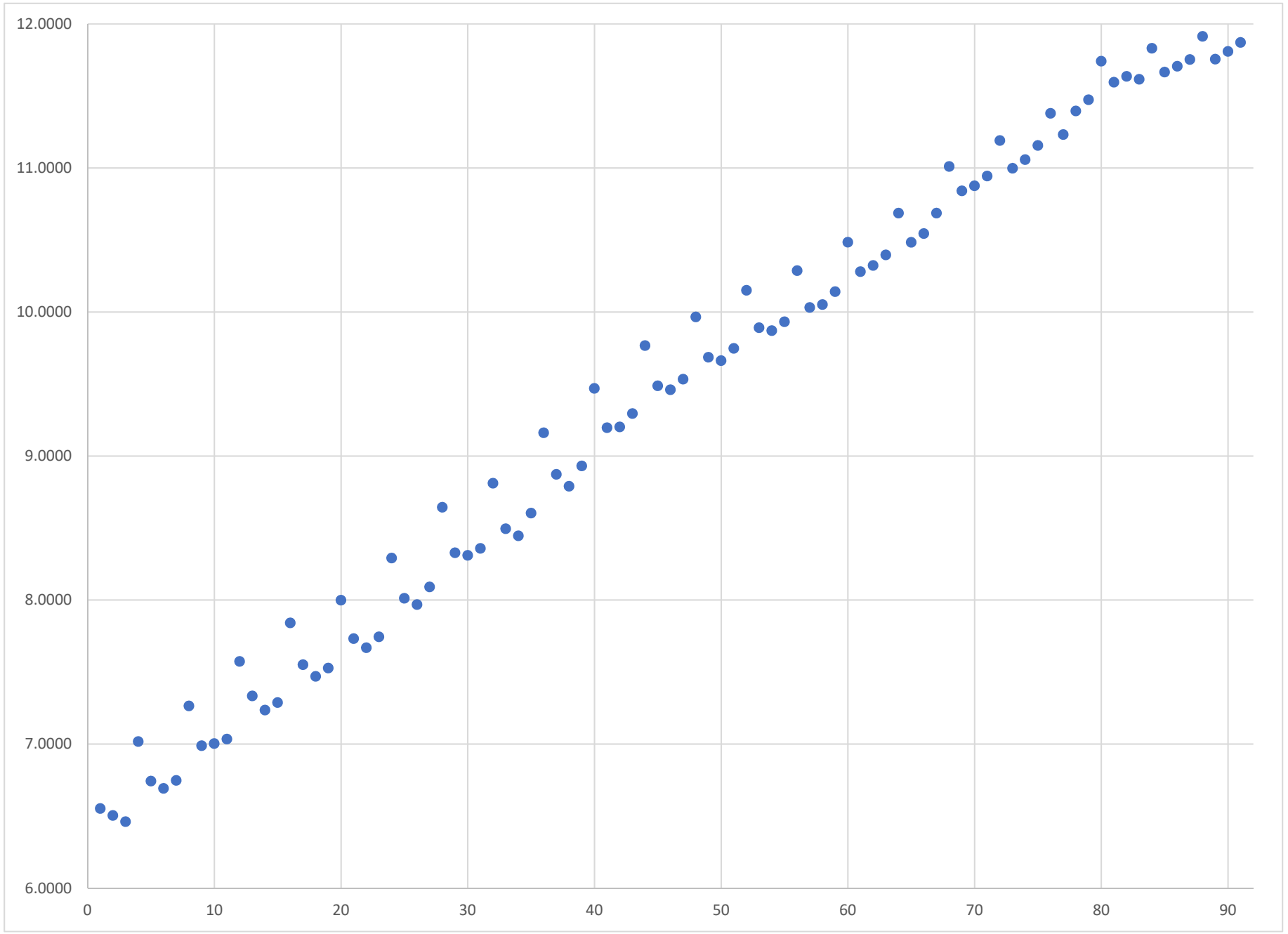
Modeling choices with the objective of obtaining an accurate forecast

- Is the Amazon's revenue growth linear or exponential?

  Use QR or LnQR as the dependent variable?

- Are there any seasonality effects?

  Use dummy variables and/or lagged variables (which lag)?


12.0000
11.0000
10.0000
9.0000
8.0000
7.0000
6.0000
0
10
20
30
40
50
60
70
80
90

Preview: exam practice questions

Build a regression model to predict Amazon's quarterly revenues for the fourth quarter of 2023 (Period=92).

Use the logarithm of quarterly revenues (variable name: LnQR) as the dependent variable, and select independent variables only among those provided in the *Data4Regressions* worksheet. (In other words, you should not create any new variables.)

- Explain your modeling process and provide the argumentation for your model choice (e.g., why you decided to include/exclude any variables in your model; 250 words max.)
- What is the Standard Error of Regression for your model?
- According to your model, what is the forecast for the Amazon's revenue in the fourth quarter of 2023 (Period=92)? (Hint: Need to exponentiate the forecast value.)
- According to your model, what is the standard error of forecast for LnQR when Period=92?

Final exam (logistics)

- **Available during the Final Exam Period:** Saturday, December 2 - Monday, December 11. Must be submitted by Monday, December 11, 23:59 EDT (Durham local time)
- **Flexibility:** You may start the exam at any point during the Final Exam period
- **12 hour window:** You have 12 hours to complete the exam, from the moment you start taking it
  - designed to be a 3-4 hours long exam
  - can work on the exam in multiple blocks of time, but must be completed within the 12 hour window
  - one submission only (Your answers should be saved/visible if/when you resume taking the exam)
- **Individual assignment:** Consultation with or assistance from any other person or source is prohibited. You may neither give nor receive any help (Honor code applies)
- **Open book, notes, and class materials, but communication prohibited:** Applies from the moment you start the exam until the end of the Final Exam Period (December 11, 23:59 EDT)
  - Allowed: books (offline), your personally prepared notes, all course materials on the course site
  - Prohibited: any materials or solutions obtained from other Fuqua students or any other source; no access to any online sources except the course site
  - You are not allowed to communicate with anyone regarding the exam
- **Variety of question formats:** numeric, multiple choice, multiple answer, short essay.

The road we traveled

- Unit 1: Introduction to Probability
- Unit 2: Probability Distributions
- Unit 3: Statistical Inference
- Unit 4: Introduction to Linear Regression
- Unit 5: Regression Analysis
- Unit 6: Regression Modeling

Thank you!