- 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 4 Individual Assignment solutions available and scores posted.

- Present:

  Unit 5 Team Assignment **due Monday, 23:59 ET** (Durham local time).

- Future:

  Unit 6 materials available.

Multiple regression:

Population model: $\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X_1} + \ldots + \beta_k \mathbf{X_k} + \epsilon$, where $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma_\epsilon^2)$

- conceptual description of how Y linearly depends on $X_1, \ldots, X_k$

Sample model: $\widehat{\mathbf{Y}} = \widehat{\beta_0} + \widehat{\beta_1} \mathbf{X_1} + \ldots + \widehat{\beta_k} \mathbf{X_k} + \epsilon$, where $\epsilon \sim \mathbf{N}(\mathbf{0}, \mathbf{SE}_{\mathbf{reg}}^2)$

- use sample data to estimate (a) coefficients, and (b) the standard deviation of the error term

Multiple regression:

Population model: $\mathbf{Y} = \beta_0 + \beta_1 \mathbf{X_1} + \ldots + \beta_k \mathbf{X_k} + \epsilon$, where $\epsilon \sim \mathbf{N}(\mathbf{0}, \sigma_\epsilon^2)$

- conceptual description of how Y linearly depends on $X_1, \ldots, X_k$

Sample model: $\widehat{\mathbf{Y}} = \widehat{\beta}_0 + \widehat{\beta}_1 \mathbf{X_1} + \ldots + \widehat{\beta}_k \mathbf{X_k} + \epsilon$, where $\epsilon \sim \mathbf{N}(\mathbf{0}, \mathbf{SE}_{\mathbf{reg}}^2)$

- use sample data to estimate (a) coefficients, and (b) the standard deviation of the error term

Evaluating regression models:

- **In terms of underlying assumptions**
  - Linearity: Each X_i should be linearly related to Y
  - Slope significance (t-stats or p-values)
  - Error term ϵ and residuals assumptions (normality, independence, homoskedasticity)
- **In terms of accuracy**
  - General forecasting accuracy: Standard error of regression (SE_{reg})
  - Accuracy of an individual forecast of interest: Standard error of forecast (SE_{fcst})
- **In terms of descriptive fit:**
  - R-squared (R^2)
  - Adjusted R-squared

$SE_{FCST} \geq SE_{REG}$

Consumption of durable goods

INC	SIZE	CONS
85	5	8,120
85	7	7,020
55	1	2,340
55	2	1,840
55	1	1,950
90	5	8,840
65	4	3,560
65	3	3,840
65	3	3,650
55	2	2,810
55	1	1,240
80	4	6,850
75	6	4,850
70	4	4,850
90	4	8,380
65	5	4,520
...	...	...

In a market-segmentation study, data on household income (INC, in thousands of U.S. dollars), household size (SIZE), and expenditure on consumer durable goods (CONS, in U.S. dollars) were collected from randomly-selected households. The company sponsoring the study was interested in isolating the effects of income and family size on the dollar amounts spent for consumer durable goods.

Can you help them?

Consumption of durable goods

Dependent Variable:	CONS	(Annual consumption of durable goods)
Independent Variables:	INC, SIZE	(Household annual income and size)

Regression Statistics

R Square	Adj.RSqr	Std.Err.Reg.	# Cases	# Missing	t(2.5%,37)
0.957	0.954	463.397	40	0	2.026

Summary Table CONS = -8,974.884 + 206.5492 INC - 175.9621 SIZE

Variable	Coeff	Std.Err.	t-Stat.	P-value	Lower95%	Upper95%
Intercept	-8,974.884	533.103	-16.835	0.000	-10,055.053	-7,894.715
INC	206.549	10.251	20.150	0.000	185.780	227.319
SIZE	-175.962	75.254	-2.338	0.025	-328.440	-23.484

Forecasted : CONS CONS = -8,974.884 + 206.5492 INC - 175.9621 SIZE

	INC	SIZE	Forecast	StErrFst	Lower95%	Upper95%
Fcst# 1	80.000	1	7,373.091	554.030	6,250.520	8,495.662
Fcst# 2	80.000	4	6,845.205	478.513	5,875.645	7,814.764

Consumption of durable goods

Dependent Variable:	CONS	(Annual consumption of durable goods)
Independent Variables:	INC	(Annual household income)

Regression Statistics

R Square	**Adj.RSqr**	**Std.Err.Reg.**	**# Cases**	**# Missing**	**t(2.5%,38)**
0.950	0.949	489.880	40	0	2.024

Summary Table $\quad \text{CONS} = -8,353.5514 + 188.2363\ \text{INC}$

Variable	**Coeff**	**Std.Err.**	**t-Stat.**	**P-value**	**Lower95%**	**Upper95%**
Intercept	-8,353.551	488.569	-17.098	0.000	-9,342.607	-7,364.496
INC	188.236	6.991	26.925	0.000	174.083	202.389

Forecasted : CONS $\quad \text{CONS} = -8,353.5514 + 188.2363\ \text{INC}$

	INC	**Forecast**	**StErrFst**	**Lower95%**	**Upper95%**
Fcst# 1	80.000	6,705.349	501.892	5,689.321	7,721.376

Consumption of durable goods

Dependent Variable:	CONS	(Annual consumption of durable goods)
Independent Variables:	SIZE	(Household size)

Regression Statistics

R Square	Adj.RSqr	Std.Err.Reg.	# Cases	# Missing	t(2.5%,38)
0.480	0.467	1582.256	40	0	2.024

Summary Table $\quad \text{CONS} = 1,048.1943 + 982.618\ \text{SIZE}$

Variable	Coeff	Std.Err.	t-Stat.	P-value	Lower95%	Upper95%
Intercept	1,048.194	654.756	1.601	0.118	-277.291	2,373.680
SIZE	982.618	165.774	5.927	0.000	647.025	1,318.211

Forecasted : CONS $\quad \text{CONS} = 1,048.1943 + 982.618\ \text{SIZE}$

	SIZE	Forecast	StErrFst	Lower95%	Upper95%
Fcst# 1	1	2,030.812	1,661.056	-1,331.821	5,393.445
Fcst# 2	4	4,978.666	1,602.962	1,733.639	8,223.694

Proxy effects

Key insights:

- Slope of included variable captures some of the impact of the missing variable (omitted variable bias)
- "Too few variables" in the regression model

In applications, proxy effect can play a major role:

- Relevant in interpretations
- Relevant if we do not observe all the variables

Evaluating regression models

- In terms of regression model assumptions
  - Linearity: Each X_i should be linearly related to Y
  - Slope significance (t-stats or p-values)
  - Error term ϵ and residuals assumptions: normality, independence, homoskedasticity (residual plots and statistics)
- In terms of accuracy
  - General forecasting accuracy: Standard error of regression (SE_{reg})
  - Accuracy of an individual forecast of interest: Standard error of forecast (SE_{fcst})
    (Error term assumptions need to be satisfied)
- In terms of descriptive fit of the (sample) data
  - R-squared (R^2)
  - Adjusted R-squared
    (Not suitable for inference out of the sample)

Consumption of durable goods: evaluating regression models

	Model 1 CONS $\sim$ INC + SIZE	**Model 2** CONS $\sim$ INC	**Model 3** CONS $\sim$ SIZE
Linearity	✓ ✓	✓	✓
Slope significance	✓ ✓	✓	✓
Error assumptions	✓	✓	✗
General forecasting (Std.Err.Reg.)	463.397	489.880	illegible
Individual forecast (INC = 80, SIZE = 1)	554.030	501.892	1661.056
Individual forecast (INC = 80, SIZE = 4)	478.513		1602.962
R^2	0.957	0.950	0.480
Adjusted R^2	0.954	0.949	0.467

Unit 5 Team Assignment

- Due Monday, 23:59 ET (Durham local time).
- Three questions with different point allocations per question
- Need to provide support for your answers. Communicating effectively is important.
- One submission per team: two-page pdf document (no other documents)

Assignment questions:

- Question 1: analytics supported business decision
- Question 2: a specific forecast of interest
- Question 3: qualitative response that should provide a generic advice

Final exam (logistics)

- **Available during the Final Exam Period:** Saturday, December 2 - Monday, December 11. Must be submitted by Monday, December 11, 23:59 ET (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, short essay