

UCLA-ALIGNED COURSE FIELD NOTES

ECON 104

Data Science for Economists

Multivariate regression, specification, instruments, binary outcomes, and real-data analysis

How these notes are written. I wrote each chapter the way I wish the material had been explained the first time: the real idea, the point where I usually get stuck, the simplest mental model that still stays honest, and a worked decision instead of a polished answer appearing from nowhere. Every chapter closes with practice and a fully written solution. If a sentence sounds impressive but does not help me solve or explain something, it does not belong here.

Daniel Mastick

Curriculum map, working notes, practice, and solutions

Course map

Week	Core thread	What should be solid by the end
1	Multivariate regression as conditional comparison	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
2	Multiple-regression assumptions and inference	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
3	Specification, interactions, and flexible controls	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
4	Model diagnostics and sensitivity	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
5	Endogeneity and instrumental variables	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
6	Binary outcome models	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
7	Prediction, validation, and causal restraint	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
8	Real-data pipeline, communication, and ethics	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
9	Integration studio	Combine several chapters in one case, preserve their assumptions, and reconcile the result across representations.
10	Cumulative review	Retrieve the full course map from memory and solve mixed problems without chapter labels.

Scope anchor: <https://catalog.registrar.ucla.edu/course/2025/econ104?siteYear=current>. The sequence is aligned to the official catalog description and expanded into a practical ten-week study plan.

How I would use this packet

Treat each numbered section as one chapter. Read the formal explanation, pause at the student interpretation, and see whether the easy version still says the same thing. Rework the example without looking, then cover the solution in the chapter practice box. The cumulative set at the end is deliberately mixed: knowledge becomes durable when I have to choose the tool instead of being told which chapter I am in.

Contents

1	Multivariate regression as conditional comparison	5
1.1	Rebuild the chapter from first principles	5
1.2	Details I would refuse to skip	5
1.3	How this chapter connects	6
1.4	The full working model	6
2	Multiple-regression assumptions and inference	7
2.1	Rebuild the chapter from first principles	7
2.2	Details I would refuse to skip	8
2.3	How this chapter connects	8
2.4	The full working model	8
3	Specification, interactions, and flexible controls	10
3.1	Rebuild the chapter from first principles	10
3.2	Details I would refuse to skip	10
3.3	How this chapter connects	11
3.4	The full working model	11
4	Model diagnostics and sensitivity	12
4.1	Rebuild the chapter from first principles	12
4.2	Details I would refuse to skip	13
4.3	How this chapter connects	13
4.4	The full working model	13
5	Endogeneity and instrumental variables	15
5.1	Rebuild the chapter from first principles	15
5.2	Details I would refuse to skip	15
5.3	How this chapter connects	16
5.4	The full working model	16
6	Binary outcome models	17
6.1	Rebuild the chapter from first principles	17
6.2	Details I would refuse to skip	18
6.3	How this chapter connects	18

6.4	The full working model	19
7	Prediction, validation, and causal restraint	20
7.1	Rebuild the chapter from first principles	20
7.2	Details I would refuse to skip	20
7.3	How this chapter connects	21
7.4	The full working model	21
8	Real-data pipeline, communication, and ethics	22
8.1	Rebuild the chapter from first principles	22
8.2	Details I would refuse to skip	23
8.3	How this chapter connects	23
8.4	The full working model	24

CHAPTER 1 Multivariate regression as conditional comparison

The idea

Multiple regression estimates the relationship between one regressor and the outcome holding included covariates fixed. The Frisch-Waugh-Lovell theorem interprets a coefficient through residualized variation.

Rebuild the chapter from first principles

The claim I need to own is this: Multiple regression estimates the relationship between one regressor and the outcome holding included covariates fixed. The Frisch-Waugh-Lovell theorem interprets a coefficient through residualized variation. I do not count that as learned just because the sentence looks familiar. I should be able to name the objects, state what changes and what stays fixed, recover the rule from the definitions, and explain why the result is allowed.

The move

My working sequence

1. **Translate.** Rewrite the prompt in the language of multivariate regression as conditional comparison. List the known quantities, the unknown, the units or types, and any hidden constraint.
2. **Choose.** State the theorem, model, algorithm, accounting rule, or physical law before substituting values. This is where I make the assumption visible.
3. **Execute.** Work one justified step at a time and keep intermediate state readable enough that I can audit it later.
4. **Verify.** Check a boundary case, sign, unit, invariant, order of magnitude, or alternative derivation. Then translate the result back into an ordinary sentence.

Details I would refuse to skip

- **Definitions:** I should be able to define every technical noun in the chapter without using that same noun in the definition.
- **Assumptions:** I write the condition that makes the method legal beside the method itself. A correct procedure under the wrong assumptions is still a wrong answer.
- **Representation:** I move freely among words, equations or code, a diagram, and a small numerical example. If one representation disagrees with another, that disagreement is useful evidence.
- **Failure modes:** I keep the nearest counterexample in mind: the boundary where the rule changes, the invariant breaks, or the model stops matching reality.
- **Communication:** My final answer names what the result means and what it does not establish. I do not let notation do the explanatory work for me.

Sanity check

Closed-note check. Can I teach multivariate regression as conditional comparison on a blank page, derive or reconstruct its main rule, work a fresh example, and diagnose a deliberately broken solution? If I can only recognize the finished work, I am not done.

How this chapter connects

I read this chapter as the bridge from the course foundations to multiple-regression assumptions and inference. The earlier material supplies the language and constraints; this chapter adds a new reasoning move; the later material asks me to use that move without being told its name. That connection matters because exams and real projects mix chapters on purpose.

The full working model

- **Core claim.** Multiple regression estimates the relationship between one regressor and the outcome holding included covariates fixed.
- **What moves.** The Frisch-Waugh-Lovell theorem interprets a coefficient through residualized variation.
- **What keeps the answer honest.** The model becomes useful only when I can apply multivariate regression as conditional comparison to an unfamiliar case and defend the assumptions.
- **Mastery standard.** I can define the objects, rebuild the rule, work a fresh case, compare alternatives, and diagnose a broken solution without relying on recognition.

How I actually think about it

I ask what variation remains after controls; 'holding fixed' is a mathematical comparison, not a physical experiment.

The easy version

Multiple regression estimates the relationship between one regressor and the outcome holding included covariates fixed.

Worked example

Question. What does the coefficient on x mean in y regressed on x and z ?

Walkthrough. It is the change in fitted y per unit x comparing observations with the same modeled z , equivalently the simple regression of y residualized on z against x residualized on z .

Common miss

The fastest way to get this chapter wrong is to use the visible procedure without naming its assumption, units, time period, reference group, or boundary. I put that condition beside the work and verify the result independently.

Solved chapter practice

Practice 1. What does the coefficient on x mean in y regressed on x and z ?

Solution. It is the change in fitted y per unit x comparing observations with the same modeled z , equivalently the simple regression of y residualized on z against x residualized on z .

Practice 2. Explain multivariate regression as conditional comparison to a classmate using one definition, one concrete example, one assumption, and one failure mode.

Solution. Begin with the core claim above, use the worked case as the concrete example, state the condition highlighted by the warning, and choose a failure where that condition does not hold. A complete explanation connects all four pieces instead of listing them.

Mastery practice A. Reconstruct multivariate regression as conditional comparison from a blank page. Include the central definition, the governing rule, one required assumption, and one boundary case.

Solution. Start from the blue idea box, but rewrite it in my own words. Identify the objects and state, name the rule before using it, copy the relevant assumption from the details audit, and choose a boundary where that assumption becomes equality or fails. A complete solution explains why each part belongs; a list of vocabulary is not enough.

Mastery practice B. A classmate gets a polished-looking answer to a multivariate regression as conditional comparison problem but never states a model or checks the result. Write the shortest useful review of that solution.

Solution. I would ask four concrete questions: What exactly are the inputs and desired output? Which rule licenses the main step? Which assumption could invalidate it? Which independent check supports the conclusion? If the answer cannot survive those four questions, I treat the numerical or verbal result as unverified.

CHAPTER 2 Multiple-regression assumptions and inference

The idea

Unbiasedness relies on zero conditional mean given all regressors; multicollinearity affects precision; heteroskedastic and clustered dependence require appropriate variance estimators. Joint F tests evaluate restrictions collectively.

Rebuild the chapter from first principles

The claim I need to own is this: Unbiasedness relies on zero conditional mean given all regressors; multicollinearity affects precision; heteroskedastic and clustered dependence require appropriate variance estimators. Joint F tests evaluate restrictions collectively. I do not count that as learned just because the sentence looks familiar. I should be able to name the objects, state what changes and what stays fixed, recover the rule from the definitions, and explain why the result is allowed.

The move

My working sequence

1. **Translate.** Rewrite the prompt in the language of multiple-regression assumptions and inference. List the known quantities, the unknown, the units or types, and any hidden constraint.
2. **Choose.** State the theorem, model, algorithm, accounting rule, or physical law before substituting values. This is where I make the assumption visible.
3. **Execute.** Work one justified step at a time and keep intermediate state readable enough that I can audit it later.
4. **Verify.** Check a boundary case, sign, unit, invariant, order of magnitude, or alternative derivation. Then translate the result back into an ordinary sentence.

Details I would refuse to skip

- **Definitions:** I should be able to define every technical noun in the chapter without using that same noun in the definition.
- **Assumptions:** I write the condition that makes the method legal beside the method itself. A correct procedure under the wrong assumptions is still a wrong answer.
- **Representation:** I move freely among words, equations or code, a diagram, and a small numerical example. If one representation disagrees with another, that disagreement is useful evidence.
- **Failure modes:** I keep the nearest counterexample in mind: the boundary where the rule changes, the invariant breaks, or the model stops matching reality.
- **Communication:** My final answer names what the result means and what it does not establish. I do not let notation do the explanatory work for me.

Sanity check

Closed-note check. Can I teach multiple-regression assumptions and inference on a blank page, derive or reconstruct its main rule, work a fresh example, and diagnose a deliberately broken solution? If I can only recognize the finished work, I am not done.

How this chapter connects

I read this chapter as the bridge from multivariate regression as conditional comparison to specification, interactions, and flexible controls. The earlier material supplies the language and constraints; this chapter adds a new reasoning move; the later material asks me to use that move without being told its name. That connection matters because exams and real projects mix chapters on purpose.

The full working model

- **Core claim.** Unbiasedness relies on zero conditional mean given all regressors; multicollinearity affects precision; heteroskedastic and clustered dependence require appropriate variance estimators.
- **What moves.** Joint F tests evaluate restrictions collectively.

- **What keeps the answer honest.** The model becomes useful only when I can apply multiple-regression assumptions and inference to an unfamiliar case and defend the assumptions.
- **Mastery standard.** I can define the objects, rebuild the rule, work a fresh case, compare alternatives, and diagnose a broken solution without relying on recognition.

How I actually think about it

My shorthand for this chapter is: my shorthand for this chapter is: a high R-squared does not validate identification, and a low one does not invalidate a causal coefficient.

The easy version

Unbiasedness relies on zero conditional mean given all regressors; multicollinearity affects precision; heteroskedastic and clustered dependence require appropriate variance estimators.

Worked example

Question. Why can two individually insignificant variables be jointly significant?

Walkthrough. Their estimates may be correlated and the null concerns a combination. An F test uses the joint covariance structure rather than two separate t-test decisions.

Common miss

The fastest way to get this chapter wrong is to use the visible procedure without naming its assumption, units, time period, reference group, or boundary. I put that condition beside the work and verify the result independently.

Solved chapter practice

Practice 1. Why can two individually insignificant variables be jointly significant?

Solution. Their estimates may be correlated and the null concerns a combination. An F test uses the joint covariance structure rather than two separate t-test decisions.

Practice 2. Explain multiple-regression assumptions and inference to a classmate using one definition, one concrete example, one assumption, and one failure mode.

Solution. Begin with the core claim above, use the worked case as the concrete example, state the condition highlighted by the warning, and choose a failure where that condition does not hold. A complete explanation connects all four pieces instead of listing them.

Mastery practice A. Reconstruct multiple-regression assumptions and inference from a blank page. Include the central definition, the governing rule, one required assumption, and one boundary case.

Solution. Start from the blue idea box, but rewrite it in my own words. Identify the objects and state, name the rule before using it, copy the relevant assumption from the details audit, and choose a boundary where that assumption becomes equality or fails. A complete solution explains why each part belongs; a list of vocabulary is not enough.

Mastery practice B. A classmate gets a polished-looking answer to a multiple-regression assumptions and inference problem but never states a model or checks the result. Write the shortest useful review of that solution.

Solution. I would ask four concrete questions: What exactly are the inputs and desired output? Which rule licenses the main step? Which assumption could invalidate it? Which independent check supports the conclusion? If the answer cannot survive those four questions, I treat the numerical or verbal result as unverified.

CHAPTER 3 Specification, interactions, and flexible controls

The idea

Polynomials, splines, interactions, fixed effects, and transformations capture heterogeneity and nonlinear patterns. Specification should follow the estimand and data support; indiscriminate controls can add bias by conditioning on colliders or mediators.

Rebuild the chapter from first principles

The claim I need to own is this: Polynomials, splines, interactions, fixed effects, and transformations capture heterogeneity and nonlinear patterns. Specification should follow the estimand and data support; indiscriminate controls can add bias by conditioning on colliders or mediators. I do not count that as learned just because the sentence looks familiar. I should be able to name the objects, state what changes and what stays fixed, recover the rule from the definitions, and explain why the result is allowed.

The move

My working sequence

1. **Translate.** Rewrite the prompt in the language of specification, interactions, and flexible controls. List the known quantities, the unknown, the units or types, and any hidden constraint.
2. **Choose.** State the theorem, model, algorithm, accounting rule, or physical law before substituting values. This is where I make the assumption visible.
3. **Execute.** Work one justified step at a time and keep intermediate state readable enough that I can audit it later.
4. **Verify.** Check a boundary case, sign, unit, invariant, order of magnitude, or alternative derivation. Then translate the result back into an ordinary sentence.

Details I would refuse to skip

- **Definitions:** I should be able to define every technical noun in the chapter without using that same noun in the definition.
- **Assumptions:** I write the condition that makes the method legal beside the method itself. A correct procedure under the wrong assumptions is still a wrong answer.
- **Representation:** I move freely among words, equations or code, a diagram, and a small numerical example. If one representation disagrees with another, that disagreement is useful evidence.

- **Failure modes:** I keep the nearest counterexample in mind: the boundary where the rule changes, the invariant breaks, or the model stops matching reality.
- **Communication:** My final answer names what the result means and what it does not establish. I do not let notation do the explanatory work for me.

Sanity check

Closed-note check. Can I teach specification, interactions, and flexible controls on a blank page, derive or reconstruct its main rule, work a fresh example, and diagnose a deliberately broken solution? If I can only recognize the finished work, I am not done.

How this chapter connects

I read this chapter as the bridge from multiple-regression assumptions and inference to model diagnostics and sensitivity. The earlier material supplies the language and constraints; this chapter adds a new reasoning move; the later material asks me to use that move without being told its name. That connection matters because exams and real projects mix chapters on purpose.

The full working model

- **Core claim.** Polynomials, splines, interactions, fixed effects, and transformations capture heterogeneity and nonlinear patterns.
- **What moves.** Specification should follow the estimand and data support; indiscriminate controls can add bias by conditioning on colliders or mediators.
- **What keeps the answer honest.** The model becomes useful only when I can apply specification, interactions, and flexible controls to an unfamiliar case and defend the assumptions.
- **Mastery standard.** I can define the objects, rebuild the rule, work a fresh case, compare alternatives, and diagnose a broken solution without relying on recognition.

How I actually think about it

More controls is not automatically more credible; I draw the causal timing before choosing them.

The easy version

Polynomials, splines, interactions, fixed effects, and transformations capture heterogeneity and nonlinear patterns.

Worked example

Question. Interpret an interaction between treatment and a female indicator.

Walkthrough. The treatment coefficient is the effect for the reference group; the interaction is the additional effect for women. Their sum gives the modeled effect for women.

Common miss

The fastest way to get this chapter wrong is to use the visible procedure without naming its assumption, units, time period, reference group, or boundary. I put that condition beside the work and verify the result independently.

Solved chapter practice

Practice 1. Interpret an interaction between treatment and a female indicator.

Solution. The treatment coefficient is the effect for the reference group; the interaction is the additional effect for women. Their sum gives the modeled effect for women.

Practice 2. Explain specification, interactions, and flexible controls to a classmate using one definition, one concrete example, one assumption, and one failure mode.

Solution. Begin with the core claim above, use the worked case as the concrete example, state the condition highlighted by the warning, and choose a failure where that condition does not hold. A complete explanation connects all four pieces instead of listing them.

Mastery practice A. Reconstruct specification, interactions, and flexible controls from a blank page. Include the central definition, the governing rule, one required assumption, and one boundary case.

Solution. Start from the blue idea box, but rewrite it in my own words. Identify the objects and state, name the rule before using it, copy the relevant assumption from the details audit, and choose a boundary where that assumption becomes equality or fails. A complete solution explains why each part belongs; a list of vocabulary is not enough.

Mastery practice B. A classmate gets a polished-looking answer to a specification, interactions, and flexible controls problem but never states a model or checks the result. Write the shortest useful review of that solution.

Solution. I would ask four concrete questions: What exactly are the inputs and desired output? Which rule licenses the main step? Which assumption could invalidate it? Which independent check supports the conclusion? If the answer cannot survive those four questions, I treat the numerical or verbal result as unverified.

CHAPTER 4 Model diagnostics and sensitivity

The idea

Residual plots, leverage, influence, outliers, functional-form tests, multicollinearity diagnostics, and sample sensitivity reveal where conclusions depend on a few observations or rigid assumptions.

Rebuild the chapter from first principles

The claim I need to own is this: Residual plots, leverage, influence, outliers, functional-form tests, multicollinearity diagnostics, and sample sensitivity reveal where conclusions depend on a few observations or rigid assumptions. I do not count that as learned just because the sentence looks familiar. I should be able to name the objects, state what changes and what stays fixed, recover the rule from the definitions,

and explain why the result is allowed.

The move

My working sequence

1. **Translate.** Rewrite the prompt in the language of model diagnostics and sensitivity. List the known quantities, the unknown, the units or types, and any hidden constraint.
2. **Choose.** State the theorem, model, algorithm, accounting rule, or physical law before substituting values. This is where I make the assumption visible.
3. **Execute.** Work one justified step at a time and keep intermediate state readable enough that I can audit it later.
4. **Verify.** Check a boundary case, sign, unit, invariant, order of magnitude, or alternative derivation. Then translate the result back into an ordinary sentence.

Details I would refuse to skip

- **Definitions:** I should be able to define every technical noun in the chapter without using that same noun in the definition.
- **Assumptions:** I write the condition that makes the method legal beside the method itself. A correct procedure under the wrong assumptions is still a wrong answer.
- **Representation:** I move freely among words, equations or code, a diagram, and a small numerical example. If one representation disagrees with another, that disagreement is useful evidence.
- **Failure modes:** I keep the nearest counterexample in mind: the boundary where the rule changes, the invariant breaks, or the model stops matching reality.
- **Communication:** My final answer names what the result means and what it does not establish. I do not let notation do the explanatory work for me.

Sanity check

Closed-note check. Can I teach model diagnostics and sensitivity on a blank page, derive or reconstruct its main rule, work a fresh example, and diagnose a deliberately broken solution? If I can only recognize the finished work, I am not done.

How this chapter connects

I read this chapter as the bridge from specification, interactions, and flexible controls to endogeneity and instrumental variables. The earlier material supplies the language and constraints; this chapter adds a new reasoning move; the later material asks me to use that move without being told its name. That connection matters because exams and real projects mix chapters on purpose.

The full working model

- **Core claim.** Residual plots, leverage, influence, outliers, functional-form tests, multicollinearity diagnostics, and sample sensitivity reveal where conclusions depend on a few observations or rigid assumptions.

- **What moves.** The model becomes useful only when I can apply model diagnostics and sensitivity to an unfamiliar case and defend the assumptions.
- **What keeps the answer honest.** The model becomes useful only when I can apply model diagnostics and sensitivity to an unfamiliar case and defend the assumptions.
- **Mastery standard.** I can define the objects, rebuild the rule, work a fresh case, compare alternatives, and diagnose a broken solution without relying on recognition.

How I actually think about it

I do not delete an outlier because it is inconvenient; I investigate provenance and report sensitivity with and without a justified correction.

The easy version

Residual plots, leverage, influence, outliers, functional-form tests, multicollinearity diagnostics, and sample sensitivity reveal where conclusions depend on a few observations or rigid assumptions.

Worked example

Question. What does a high-leverage observation mean?

Walkthrough. Its regressor values are unusual, giving it potential to move the fitted line. Influence also depends on its residual; leverage alone is not proof of error.

Common miss

The fastest way to get this chapter wrong is to use the visible procedure without naming its assumption, units, time period, reference group, or boundary. I put that condition beside the work and verify the result independently.

Solved chapter practice

Practice 1. What does a high-leverage observation mean?

Solution. Its regressor values are unusual, giving it potential to move the fitted line. Influence also depends on its residual; leverage alone is not proof of error.

Practice 2. Explain model diagnostics and sensitivity to a classmate using one definition, one concrete example, one assumption, and one failure mode.

Solution. Begin with the core claim above, use the worked case as the concrete example, state the condition highlighted by the warning, and choose a failure where that condition does not hold. A complete explanation connects all four pieces instead of listing them.

Mastery practice A. Reconstruct model diagnostics and sensitivity from a blank page. Include the central definition, the governing rule, one required assumption, and one boundary case.

Solution. Start from the blue idea box, but rewrite it in my own words. Identify the objects and state, name the rule before using it, copy the relevant assumption from the details audit, and choose a boundary where that assumption becomes equality or fails. A complete solution explains why each part belongs; a list of vocabulary is not enough.

Mastery practice B. A classmate gets a polished-looking answer to a model diagnostics and

sensitivity problem but never states a model or checks the result. Write the shortest useful review of that solution.

Solution. I would ask four concrete questions: What exactly are the inputs and desired output? Which rule licenses the main step? Which assumption could invalidate it? Which independent check supports the conclusion? If the answer cannot survive those four questions, I treat the numerical or verbal result as unverified.

CHAPTER 5 Endogeneity and instrumental variables

The idea

Endogeneity arises from omitted causes, simultaneity, measurement error, or selection. A valid instrument must shift the endogenous regressor, be independent of unobserved outcome determinants, and affect the outcome only through that regressor.

Rebuild the chapter from first principles

The claim I need to own is this: Endogeneity arises from omitted causes, simultaneity, measurement error, or selection. A valid instrument must shift the endogenous regressor, be independent of unobserved outcome determinants, and affect the outcome only through that regressor. I do not count that as learned just because the sentence looks familiar. I should be able to name the objects, state what changes and what stays fixed, recover the rule from the definitions, and explain why the result is allowed.

The move

My working sequence

1. **Translate.** Rewrite the prompt in the language of endogeneity and instrumental variables. List the known quantities, the unknown, the units or types, and any hidden constraint.
2. **Choose.** State the theorem, model, algorithm, accounting rule, or physical law before substituting values. This is where I make the assumption visible.
3. **Execute.** Work one justified step at a time and keep intermediate state readable enough that I can audit it later.
4. **Verify.** Check a boundary case, sign, unit, invariant, order of magnitude, or alternative derivation. Then translate the result back into an ordinary sentence.

Details I would refuse to skip

- **Definitions:** I should be able to define every technical noun in the chapter without using that same noun in the definition.
- **Assumptions:** I write the condition that makes the method legal beside the method itself. A correct procedure under the wrong assumptions is still a wrong answer.

- **Representation:** I move freely among words, equations or code, a diagram, and a small numerical example. If one representation disagrees with another, that disagreement is useful evidence.
- **Failure modes:** I keep the nearest counterexample in mind: the boundary where the rule changes, the invariant breaks, or the model stops matching reality.
- **Communication:** My final answer names what the result means and what it does not establish. I do not let notation do the explanatory work for me.

Sanity check

Closed-note check. Can I teach endogeneity and instrumental variables on a blank page, derive or reconstruct its main rule, work a fresh example, and diagnose a deliberately broken solution? If I can only recognize the finished work, I am not done.

How this chapter connects

I read this chapter as the bridge from model diagnostics and sensitivity to binary outcome models. The earlier material supplies the language and constraints; this chapter adds a new reasoning move; the later material asks me to use that move without being told its name. That connection matters because exams and real projects mix chapters on purpose.

The full working model

- **Core claim.** Endogeneity arises from omitted causes, simultaneity, measurement error, or selection.
- **What moves.** A valid instrument must shift the endogenous regressor, be independent of unobserved outcome determinants, and affect the outcome only through that regressor.
- **What keeps the answer honest.** The model becomes useful only when I can apply endogeneity and instrumental variables to an unfamiliar case and defend the assumptions.
- **Mastery standard.** I can define the objects, rebuild the rule, work a fresh case, compare alternatives, and diagnose a broken solution without relying on recognition.

How I actually think about it

Relevance can be tested imperfectly; exclusion is an economic argument I have to defend.

The easy version

Endogeneity arises from omitted causes, simultaneity, measurement error, or selection.

Worked example

Question. Why is a weak instrument dangerous?

Walkthrough. It provides little exogenous variation, producing imprecise estimates and finite-sample bias toward ordinary least squares while making standard inference unreliable.

Common miss

The fastest way to get this chapter wrong is to use the visible procedure without naming its assumption, units, time period, reference group, or boundary. I put that condition beside the work and verify the result independently.

Solved chapter practice

Practice 1. Why is a weak instrument dangerous?

Solution. It provides little exogenous variation, producing imprecise estimates and finite-sample bias toward ordinary least squares while making standard inference unreliable.

Practice 2. Explain endogeneity and instrumental variables to a classmate using one definition, one concrete example, one assumption, and one failure mode.

Solution. Begin with the core claim above, use the worked case as the concrete example, state the condition highlighted by the warning, and choose a failure where that condition does not hold. A complete explanation connects all four pieces instead of listing them.

Mastery practice A. Reconstruct endogeneity and instrumental variables from a blank page. Include the central definition, the governing rule, one required assumption, and one boundary case.

Solution. Start from the blue idea box, but rewrite it in my own words. Identify the objects and state, name the rule before using it, copy the relevant assumption from the details audit, and choose a boundary where that assumption becomes equality or fails. A complete solution explains why each part belongs; a list of vocabulary is not enough.

Mastery practice B. A classmate gets a polished-looking answer to a endogeneity and instrumental variables problem but never states a model or checks the result. Write the shortest useful review of that solution.

Solution. I would ask four concrete questions: What exactly are the inputs and desired output? Which rule licenses the main step? Which assumption could invalidate it? Which independent check supports the conclusion? If the answer cannot survive those four questions, I treat the numerical or verbal result as unverified.

CHAPTER 6 Binary outcome models

The idea

Linear probability, logit, and probit models describe conditional event probabilities. Nonlinear coefficients are not probability changes by themselves; marginal effects depend on covariates, and classification thresholds encode costs.

Rebuild the chapter from first principles

The claim I need to own is this: Linear probability, logit, and probit models describe conditional event probabilities. Nonlinear coefficients are not probability changes by themselves; marginal effects depend on covariates, and classification thresholds encode costs. I do not count that as learned just because the

sentence looks familiar. I should be able to name the objects, state what changes and what stays fixed, recover the rule from the definitions, and explain why the result is allowed.

The move

My working sequence

1. **Translate.** Rewrite the prompt in the language of binary outcome models. List the known quantities, the unknown, the units or types, and any hidden constraint.
2. **Choose.** State the theorem, model, algorithm, accounting rule, or physical law before substituting values. This is where I make the assumption visible.
3. **Execute.** Work one justified step at a time and keep intermediate state readable enough that I can audit it later.
4. **Verify.** Check a boundary case, sign, unit, invariant, order of magnitude, or alternative derivation. Then translate the result back into an ordinary sentence.

Details I would refuse to skip

- **Definitions:** I should be able to define every technical noun in the chapter without using that same noun in the definition.
- **Assumptions:** I write the condition that makes the method legal beside the method itself. A correct procedure under the wrong assumptions is still a wrong answer.
- **Representation:** I move freely among words, equations or code, a diagram, and a small numerical example. If one representation disagrees with another, that disagreement is useful evidence.
- **Failure modes:** I keep the nearest counterexample in mind: the boundary where the rule changes, the invariant breaks, or the model stops matching reality.
- **Communication:** My final answer names what the result means and what it does not establish. I do not let notation do the explanatory work for me.

Sanity check

Closed-note check. Can I teach binary outcome models on a blank page, derive or reconstruct its main rule, work a fresh example, and diagnose a deliberately broken solution? If I can only recognize the finished work, I am not done.

How this chapter connects

I read this chapter as the bridge from endogeneity and instrumental variables to prediction, validation, and causal restraint. The earlier material supplies the language and constraints; this chapter adds a new reasoning move; the later material asks me to use that move without being told its name. That connection matters because exams and real projects mix chapters on purpose.

The full working model

- **Core claim.** Linear probability, logit, and probit models describe conditional event probabilities.
- **What moves.** Nonlinear coefficients are not probability changes by themselves; marginal effects depend on covariates, and classification thresholds encode costs.
- **What keeps the answer honest.** The model becomes useful only when I can apply binary outcome models to an unfamiliar case and defend the assumptions.
- **Mastery standard.** I can define the objects, rebuild the rule, work a fresh case, compare alternatives, and diagnose a broken solution without relying on recognition.

How I actually think about it

I separate estimating a probability from choosing an action threshold.

The easy version

Linear probability, logit, and probit models describe conditional event probabilities.

Worked example

Question. A logit coefficient is positive. What can be concluded immediately?

Walkthrough. The regressor raises modeled log odds and, holding other variables fixed, raises probability, but the size of the probability change depends on the starting covariates.

Common miss

The fastest way to get this chapter wrong is to use the visible procedure without naming its assumption, units, time period, reference group, or boundary. I put that condition beside the work and verify the result independently.

Solved chapter practice

Practice 1. A logit coefficient is positive. What can be concluded immediately?

Solution. The regressor raises modeled log odds and, holding other variables fixed, raises probability, but the size of the probability change depends on the starting covariates.

Practice 2. Explain binary outcome models to a classmate using one definition, one concrete example, one assumption, and one failure mode.

Solution. Begin with the core claim above, use the worked case as the concrete example, state the condition highlighted by the warning, and choose a failure where that condition does not hold. A complete explanation connects all four pieces instead of listing them.

Mastery practice A. Reconstruct binary outcome models from a blank page. Include the central definition, the governing rule, one required assumption, and one boundary case.

Solution. Start from the blue idea box, but rewrite it in my own words. Identify the objects and state, name the rule before using it, copy the relevant assumption from the details audit, and choose a boundary where that assumption becomes equality or fails. A complete solution explains why each part belongs; a list of vocabulary is not enough.

Mastery practice B. A classmate gets a polished-looking answer to a binary outcome models problem but never states a model or checks the result. Write the shortest useful review of that solution.

Solution. I would ask four concrete questions: What exactly are the inputs and desired output? Which rule licenses the main step? Which assumption could invalidate it? Which independent check supports the conclusion? If the answer cannot survive those four questions, I treat the numerical or verbal result as unverified.

CHAPTER 7 Prediction, validation, and causal restraint

The idea

Prediction evaluates out-of-sample loss using properly separated training, validation, and test data. Regularization and cross-validation manage variance, while causal interpretation still needs design assumptions that prediction accuracy cannot supply.

Rebuild the chapter from first principles

The claim I need to own is this: Prediction evaluates out-of-sample loss using properly separated training, validation, and test data. Regularization and cross-validation manage variance, while causal interpretation still needs design assumptions that prediction accuracy cannot supply. I do not count that as learned just because the sentence looks familiar. I should be able to name the objects, state what changes and what stays fixed, recover the rule from the definitions, and explain why the result is allowed.

The move

My working sequence

1. **Translate.** Rewrite the prompt in the language of prediction, validation, and causal restraint. List the known quantities, the unknown, the units or types, and any hidden constraint.
2. **Choose.** State the theorem, model, algorithm, accounting rule, or physical law before substituting values. This is where I make the assumption visible.
3. **Execute.** Work one justified step at a time and keep intermediate state readable enough that I can audit it later.
4. **Verify.** Check a boundary case, sign, unit, invariant, order of magnitude, or alternative derivation. Then translate the result back into an ordinary sentence.

Details I would refuse to skip

- **Definitions:** I should be able to define every technical noun in the chapter without using that same noun in the definition.
- **Assumptions:** I write the condition that makes the method legal beside the method itself. A correct procedure under the wrong assumptions is still a wrong answer.

- **Representation:** I move freely among words, equations or code, a diagram, and a small numerical example. If one representation disagrees with another, that disagreement is useful evidence.
- **Failure modes:** I keep the nearest counterexample in mind: the boundary where the rule changes, the invariant breaks, or the model stops matching reality.
- **Communication:** My final answer names what the result means and what it does not establish. I do not let notation do the explanatory work for me.

Sanity check

Closed-note check. Can I teach prediction, validation, and causal restraint on a blank page, derive or reconstruct its main rule, work a fresh example, and diagnose a deliberately broken solution? If I can only recognize the finished work, I am not done.

How this chapter connects

I read this chapter as the bridge from binary outcome models to real-data pipeline, communication, and ethics. The earlier material supplies the language and constraints; this chapter adds a new reasoning move; the later material asks me to use that move without being told its name. That connection matters because exams and real projects mix chapters on purpose.

The full working model

- **Core claim.** Prediction evaluates out-of-sample loss using properly separated training, validation, and test data.
- **What moves.** Regularization and cross-validation manage variance, while causal interpretation still needs design assumptions that prediction accuracy cannot supply.
- **What keeps the answer honest.** The model becomes useful only when I can apply prediction, validation, and causal restraint to an unfamiliar case and defend the assumptions.
- **Mastery standard.** I can define the objects, rebuild the rule, work a fresh case, compare alternatives, and diagnose a broken solution without relying on recognition.

How I actually think about it

My shorthand for this chapter is: my shorthand for this chapter is: a model can predict well for the wrong reason, especially when the future data pipeline changes.

The easy version

Prediction evaluates out-of-sample loss using properly separated training, validation, and test data.

Worked example

Question. Why must preprocessing be fit within each training fold?

Walkthrough. Using the full dataset leaks validation information into training. Fit imputation, scaling, and feature selection on the fold's training data, then apply them to its validation data.

Common miss

The fastest way to get this chapter wrong is to use the visible procedure without naming its assumption, units, time period, reference group, or boundary. I put that condition beside the work and verify the result independently.

Solved chapter practice

Practice 1. Why must preprocessing be fit within each training fold?

Solution. Using the full dataset leaks validation information into training. Fit imputation, scaling, and feature selection on the fold's training data, then apply them to its validation data.

Practice 2. Explain prediction, validation, and causal restraint to a classmate using one definition, one concrete example, one assumption, and one failure mode.

Solution. Begin with the core claim above, use the worked case as the concrete example, state the condition highlighted by the warning, and choose a failure where that condition does not hold. A complete explanation connects all four pieces instead of listing them.

Mastery practice A. Reconstruct prediction, validation, and causal restraint from a blank page. Include the central definition, the governing rule, one required assumption, and one boundary case.

Solution. Start from the blue idea box, but rewrite it in my own words. Identify the objects and state, name the rule before using it, copy the relevant assumption from the details audit, and choose a boundary where that assumption becomes equality or fails. A complete solution explains why each part belongs; a list of vocabulary is not enough.

Mastery practice B. A classmate gets a polished-looking answer to a prediction, validation, and causal restraint problem but never states a model or checks the result. Write the shortest useful review of that solution.

Solution. I would ask four concrete questions: What exactly are the inputs and desired output? Which rule licenses the main step? Which assumption could invalidate it? Which independent check supports the conclusion? If the answer cannot survive those four questions, I treat the numerical or verbal result as unverified.

CHAPTER 8 Real-data pipeline, communication, and ethics

The idea

Economic data work requires documented collection, identifiers, units, missingness, joins, reproducible scripts, privacy, subgroup checks, uncertainty, and a decision-linked narrative. A result is incomplete until its scope and failure conditions are stated.

Rebuild the chapter from first principles

The claim I need to own is this: Economic data work requires documented collection, identifiers, units, missingness, joins, reproducible scripts, privacy, subgroup checks, uncertainty, and a decision-linked narrative. A result is incomplete until its scope and failure conditions are stated. I do not count that

as learned just because the sentence looks familiar. I should be able to name the objects, state what changes and what stays fixed, recover the rule from the definitions, and explain why the result is allowed.

The move

My working sequence

1. **Translate.** Rewrite the prompt in the language of real-data pipeline, communication, and ethics. List the known quantities, the unknown, the units or types, and any hidden constraint.
2. **Choose.** State the theorem, model, algorithm, accounting rule, or physical law before substituting values. This is where I make the assumption visible.
3. **Execute.** Work one justified step at a time and keep intermediate state readable enough that I can audit it later.
4. **Verify.** Check a boundary case, sign, unit, invariant, order of magnitude, or alternative derivation. Then translate the result back into an ordinary sentence.

Details I would refuse to skip

- **Definitions:** I should be able to define every technical noun in the chapter without using that same noun in the definition.
- **Assumptions:** I write the condition that makes the method legal beside the method itself. A correct procedure under the wrong assumptions is still a wrong answer.
- **Representation:** I move freely among words, equations or code, a diagram, and a small numerical example. If one representation disagrees with another, that disagreement is useful evidence.
- **Failure modes:** I keep the nearest counterexample in mind: the boundary where the rule changes, the invariant breaks, or the model stops matching reality.
- **Communication:** My final answer names what the result means and what it does not establish. I do not let notation do the explanatory work for me.

Sanity check

Closed-note check. Can I teach real-data pipeline, communication, and ethics on a blank page, derive or reconstruct its main rule, work a fresh example, and diagnose a deliberately broken solution? If I can only recognize the finished work, I am not done.

How this chapter connects

I read this chapter as the bridge from prediction, validation, and causal restraint to the cumulative course model. The earlier material supplies the language and constraints; this chapter adds a new reasoning move; the later material asks me to use that move without being told its name. That connection matters because exams and real projects mix chapters on purpose.

The full working model

- **Core claim.** Economic data work requires documented collection, identifiers, units, missingness, joins, reproducible scripts, privacy, subgroup checks, uncertainty, and a decision-linked narrative.
- **What moves.** A result is incomplete until its scope and failure conditions are stated.
- **What keeps the answer honest.** The model becomes useful only when I can apply real-data pipeline, communication, and ethics to an unfamiliar case and defend the assumptions.
- **Mastery standard.** I can define the objects, rebuild the rule, work a fresh case, compare alternatives, and diagnose a broken solution without relying on recognition.

How I actually think about it

I want the reader to know exactly what would make me change my conclusion.

The easy version

Economic data work requires documented collection, identifiers, units, missingness, joins, reproducible scripts, privacy, subgroup checks, uncertainty, and a decision-linked narrative.

Worked example

Question. What should a final empirical notebook prove?

Walkthrough. That raw inputs are preserved, transformations are auditable, estimates reproduce, assumptions and diagnostics are explicit, plots and tables match the code, uncertainty is visible, and claims do not outrun the design.

Common miss

The fastest way to get this chapter wrong is to use the visible procedure without naming its assumption, units, time period, reference group, or boundary. I put that condition beside the work and verify the result independently.

Solved chapter practice

Practice 1. What should a final empirical notebook prove?

Solution. That raw inputs are preserved, transformations are auditable, estimates reproduce, assumptions and diagnostics are explicit, plots and tables match the code, uncertainty is visible, and claims do not outrun the design.

Practice 2. Explain real-data pipeline, communication, and ethics to a classmate using one definition, one concrete example, one assumption, and one failure mode.

Solution. Begin with the core claim above, use the worked case as the concrete example, state the condition highlighted by the warning, and choose a failure where that condition does not hold. A complete explanation connects all four pieces instead of listing them.

Mastery practice A. Reconstruct real-data pipeline, communication, and ethics from a blank page. Include the central definition, the governing rule, one required assumption, and one boundary case.

Solution. Start from the blue idea box, but rewrite it in my own words. Identify the objects and state, name the rule before using it, copy the relevant assumption from the details audit, and choose a boundary where that assumption becomes equality or fails. A complete solution explains why each part belongs; a list of vocabulary is not enough.

Mastery practice B. A classmate gets a polished-looking answer to a real-data pipeline, communication, and ethics problem but never states a model or checks the result. Write the shortest useful review of that solution.

Solution. I would ask four concrete questions: What exactly are the inputs and desired output? Which rule licenses the main step? Which assumption could invalidate it? Which independent check supports the conclusion? If the answer cannot survive those four questions, I treat the numerical or verbal result as unverified.

Cumulative study plan

I use this packet in three passes. First I reconstruct each model and work its examples without looking. Second I mix chapters so the tool is no longer named for me. Third I explain a complete case aloud, including assumptions, units, uncertainty, failure modes, and what evidence would change my conclusion.