UCLA-ALIGNED COURSE FIELD NOTES

STATS 101A

Introduction to Data Analysis and Regression

Multiple regression, generalized linear models, diagnostics, resampling, and defensible applied 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	**Exploratory analysis and the regression question**	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
2	**Simple least squares and projection**	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
3	**Multiple regression and partial effects**	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
4	**Categorical predictors, interactions, and nonlinear terms**	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
5	**Inference for coefficients and predictions**	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
6	**Residual diagnostics and influential observations**	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
7	**Logistic and generalized linear models**	Reconstruct the model, work an application, test its assumptions, and explain the result in ordinary language.
8	**Bootstrap, validation, and the complete analysis**	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/STATS101A. 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

CHAPTER 1 Exploratory analysis and the regression question

The idea

Applied regression begins with the estimand, data provenance, missingness, units, distributions, and relationships. Graphics expose nonlinearities, groups, and influential cases before a formula hardens assumptions.

Rebuild the chapter from first principles

The claim I need to own is this: Applied regression begins with the estimand, data provenance, missingness, units, distributions, and relationships. Graphics expose nonlinearities, groups, and influential cases before a formula hardens 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 exploratory analysis and the regression question. 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 exploratory analysis and the regression question 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 simple least squares and projection. 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.** Applied regression begins with the estimand, data provenance, missingness, units, distributions, and relationships.
- **What moves.** Graphics expose nonlinearities, groups, and influential cases before a formula hardens assumptions.
- **What keeps the answer honest.** The model becomes useful only when I can apply exploratory analysis and the regression question 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 decide what comparison I want the coefficient to represent before I decide which columns enter the model.

The easy version

Applied regression begins with the estimand, data provenance, missingness, units, distributions, and relationships.

Worked example

Question. Why inspect a scatterplot matrix before fitting a multiple regression?
Walkthrough. It can reveal transformations, collinearity, group structure, missingness patterns, and outliers that a coefficient table hides, guiding a model that matches the 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 inspect a scatterplot matrix before fitting a multiple regression?
Solution. It can reveal transformations, collinearity, group structure, missingness patterns, and outliers that a coefficient table hides, guiding a model that matches the data.

Practice 2. Explain exploratory analysis and the regression question 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 exploratory analysis and the regression question 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 exploratory analysis and the regression question 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 Simple least squares and projection

The idea

Least squares chooses the line minimizing squared residuals. The slope is covariance divided by predictor variance, residuals sum to zero with an intercept, and R-squared describes in-sample variation explained rather than causal validity.

Rebuild the chapter from first principles

The claim I need to own is this: Least squares chooses the line minimizing squared residuals. The slope is covariance divided by predictor variance, residuals sum to zero with an intercept, and R-squared describes in-sample variation explained rather than causal validity. 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 simple least squares and projection. 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 simple least squares and projection 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 exploratory analysis and the regression question to multiple regression and partial effects. 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.** Least squares chooses the line minimizing squared residuals.
- **What moves.** The slope is covariance divided by predictor variance, residuals sum to zero with an intercept, and R-squared describes in-sample variation explained rather than causal validity.

- **What keeps the answer honest.** The model becomes useful only when I can apply simple least squares and projection 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 read the fitted line as a compact description of conditional averages, not as a law governing every observation.

The easy version

Least squares chooses the line minimizing squared residuals.

Worked example

Question. If covariance of x and y is 12 and variance of x is 4, find the least-squares slope.
Walkthrough. The slope is 12/4=3. The intercept would then place the line through the point of sample means.

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. If covariance of x and y is 12 and variance of x is 4, find the least-squares slope.
Solution. The slope is 12/4=3. The intercept would then place the line through the point of sample means.

Practice 2. Explain simple least squares and projection 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 simple least squares and projection 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 simple least squares and projection 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 Multiple regression and partial effects

The idea

Multiple regression projects an outcome onto several predictors. A coefficient compares observations after linearly adjusting for other included predictors; coding, scale, collinearity, omitted causes, and post-treatment controls shape interpretation.

Rebuild the chapter from first principles

The claim I need to own is this: Multiple regression projects an outcome onto several predictors. A coefficient compares observations after linearly adjusting for other included predictors; coding, scale, collinearity, omitted causes, and post-treatment controls shape interpretation. 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 and partial effects. 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 and partial effects 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 simple least squares and projection to categorical predictors, interactions, and nonlinear terms. 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 projects an outcome onto several predictors.
- **What moves.** A coefficient compares observations after linearly adjusting for other included predictors; coding, scale, collinearity, omitted causes, and post-treatment controls shape interpretation.
- **What keeps the answer honest.** The model becomes useful only when I can apply multiple regression and partial effects 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

Holding variables fixed is a model-based comparison. I ask whether that comparison exists in the data and makes substantive sense.

The easy version

Multiple regression projects an outcome onto several predictors.

Worked example

Question. In a wage model with education and experience, what does the education coefficient mean?
Walkthrough. It is the fitted wage difference per additional education unit among observations with the same modeled experience, conditional on the model form and included 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. In a wage model with education and experience, what does the education coefficient mean?
Solution. It is the fitted wage difference per additional education unit among observations with the same modeled experience, conditional on the model form and included covariates.

Practice 2. Explain multiple regression and partial effects 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 and partial effects 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 and partial effects 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 Categorical predictors, interactions, and nonlinear terms

The idea

Indicators encode group comparisons, interactions allow effects to vary, and polynomial or spline bases model curvature while remaining linear in coefficients. Reference levels and centering determine what lower-order terms mean.

Rebuild the chapter from first principles

The claim I need to own is this: Indicators encode group comparisons, interactions allow effects to vary, and polynomial or spline bases model curvature while remaining linear in coefficients. Reference levels and centering determine what lower-order terms mean. 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 categorical predictors, interactions, and nonlinear terms. 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 categorical predictors, interactions, and nonlinear terms 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 and partial effects to inference for coefficients and predictions. 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.** Indicators encode group comparisons, interactions allow effects to vary, and polynomial or spline bases model curvature while remaining linear in coefficients.
- **What moves.** Reference levels and centering determine what lower-order terms mean.

- **What keeps the answer honest.** The model becomes useful only when I can apply categorical predictors, interactions, and nonlinear terms 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 compute two concrete predictions from an interaction model; that is faster and safer than interpreting symbols in isolation.

The easy version

Indicators encode group comparisons, interactions allow effects to vary, and polynomial or spline bases model curvature while remaining linear in coefficients.

Worked example

Question. For y=b0+b1x+b2g+b3xg, give the x slope when g=0 and g=1.
Walkthrough. The slope is b1 when g=0 and b1+b3 when g=1. The interaction b3 is the difference between those slopes.

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. For y=b0+b1x+b2g+b3xg, give the x slope when g=0 and g=1.
Solution. The slope is b1 when g=0 and b1+b3 when g=1. The interaction b3 is the difference between those slopes.

Practice 2. Explain categorical predictors, interactions, and nonlinear terms 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 categorical predictors, interactions, and nonlinear terms 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 categorical predictors, interactions, and nonlinear terms 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 Inference for coefficients and predictions

The idea

Standard errors, t tests, confidence intervals, F tests, mean-response intervals, and prediction intervals quantify different uncertainties. Their calibration depends on model assumptions, design, and the variance estimator.

Rebuild the chapter from first principles

The claim I need to own is this: Standard errors, t tests, confidence intervals, F tests, mean-response intervals, and prediction intervals quantify different uncertainties. Their calibration depends on model assumptions, design, and the variance estimator. 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 inference for coefficients and predictions. 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 inference for coefficients and predictions 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 categorical predictors, interactions, and nonlinear terms to residual diagnostics and influential observations. 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.** Standard errors, t tests, confidence intervals, F tests, mean-response intervals, and prediction intervals quantify different uncertainties.
- **What moves.** Their calibration depends on model assumptions, design, and the variance estimator.
- **What keeps the answer honest.** The model becomes useful only when I can apply inference for coefficients and predictions 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

A narrow interval is not useful if the model targets the wrong comparison, so I audit identification before celebrating precision.

The easy version

Standard errors, t tests, confidence intervals, F tests, mean-response intervals, and prediction intervals quantify different uncertainties.

Worked example

Question. Why is a prediction interval wider than a confidence interval for the mean response?
Walkthrough. It includes uncertainty in the estimated mean plus irreducible variation of a new observation around that mean.

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 prediction interval wider than a confidence interval for the mean response?
Solution. It includes uncertainty in the estimated mean plus irreducible variation of a new observation around that mean.

Practice 2. Explain inference for coefficients and predictions 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 inference for coefficients and predictions 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 inference for coefficients and predictions 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 Residual diagnostics and influential observations

The idea

Residuals diagnose nonlinearity, heteroskedasticity, dependence, and tail behavior. Leverage measures unusual predictors; Cook's distance and deletion diagnostics measure influence. Investigation and sensitivity analysis replace automatic deletion.

Rebuild the chapter from first principles

The claim I need to own is this: Residuals diagnose nonlinearity, heteroskedasticity, dependence, and tail behavior. Leverage measures unusual predictors; Cook's distance and deletion diagnostics measure influence. Investigation and sensitivity analysis replace automatic deletion. 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 residual diagnostics and influential observations. 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 residual diagnostics and influential observations 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 inference for coefficients and predictions to logistic and generalized linear 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.** Residuals diagnose nonlinearity, heteroskedasticity, dependence, and tail behavior.
- **What moves.** Leverage measures unusual predictors; Cook's distance and deletion diagnostics measure influence.
- **What keeps the answer honest.** Investigation and sensitivity analysis replace automatic deletion.
- **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: an outlier is a question about the data-generating process, not a license to erase a row.

The easy version

Residuals diagnose nonlinearity, heteroskedasticity, dependence, and tail behavior.

Worked example

Question. Can a point have high leverage but little influence?
Walkthrough. Yes. If its response aligns with the fitted pattern, deleting it may barely change coefficients even though its predictor values are unusual.

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. Can a point have high leverage but little influence?
Solution. Yes. If its response aligns with the fitted pattern, deleting it may barely change coefficients even though its predictor values are unusual.

Practice 2. Explain residual diagnostics and influential observations 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 residual diagnostics and influential observations 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 residual diagnostics and influential observations 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 Logistic and generalized linear models

The idea

Logistic regression models log odds for binary outcomes; generalized linear models combine a response distribution, linear predictor, and link. Coefficients live on the link scale, while marginal effects and predictions translate results for readers.

Rebuild the chapter from first principles

The claim I need to own is this: Logistic regression models log odds for binary outcomes; generalized linear models combine a response distribution, linear predictor, and link. Coefficients live on the link scale, while marginal effects and predictions translate results for readers. 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 logistic and generalized linear 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 logistic and generalized linear 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 residual diagnostics and influential observations to bootstrap, validation, and the complete analysis. 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.** Logistic regression models log odds for binary outcomes; generalized linear models combine a response distribution, linear predictor, and link.
- **What moves.** Coefficients live on the link scale, while marginal effects and predictions translate results for readers.
- **What keeps the answer honest.** The model becomes useful only when I can apply logistic and generalized linear 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 convert log-odds coefficients into predicted probabilities at realistic covariate values before I explain them.

The easy version

Logistic regression models log odds for binary outcomes; generalized linear models combine a response distribution, linear predictor, and link.

Worked example

Question. A logistic coefficient is log(2). Interpret its odds ratio.
Walkthrough. Holding other predictors fixed, a one-unit increase multiplies modeled odds by

exp(log 2)=2. The probability change is not constant and depends on baseline risk.

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 logistic coefficient is log(2). Interpret its odds ratio.
Solution. Holding other predictors fixed, a one-unit increase multiplies modeled odds by exp(log 2)=2. The probability change is not constant and depends on baseline risk.

Practice 2. Explain logistic and generalized linear 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 logistic and generalized linear 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 logistic and generalized linear 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 8 Bootstrap, validation, and the complete analysis

The idea

Bootstrap resampling approximates estimator uncertainty; cross-validation estimates predictive performance; both must reproduce the whole fitting pipeline. A complete analysis documents cleaning, assumptions, diagnostics, sensitivity, uncertainty, and scope.

Rebuild the chapter from first principles

The claim I need to own is this: Bootstrap resampling approximates estimator uncertainty; cross-validation estimates predictive performance; both must reproduce the whole fitting pipeline. A complete analysis documents cleaning, assumptions, diagnostics, sensitivity, uncertainty, and scope. 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 bootstrap, validation, and the complete analysis. 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 bootstrap, validation, and the complete analysis 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 logistic and generalized linear models 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.** Bootstrap resampling approximates estimator uncertainty; cross-validation estimates predictive performance; both must reproduce the whole fitting pipeline.
- **What moves.** A complete analysis documents cleaning, assumptions, diagnostics, sensitivity, uncertainty, and scope.
- **What keeps the answer honest.** The model becomes useful only when I can apply bootstrap, validation, and the complete analysis 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 another student to rerun every choice and see where my conclusion would change.

The easy version

Bootstrap resampling approximates estimator uncertainty; cross-validation estimates predictive performance; both must reproduce the whole fitting pipeline.

Worked example

Question. Why must variable selection occur inside each cross-validation training fold?
Walkthrough. Selecting variables on the full data leaks validation information into the fitted pipeline and makes performance look better than it will be on genuinely new 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 variable selection occur inside each cross-validation training fold?
Solution. Selecting variables on the full data leaks validation information into the fitted pipeline and makes performance look better than it will be on genuinely new data.

Practice 2. Explain bootstrap, validation, and the complete analysis 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 bootstrap, validation, and the complete analysis 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 bootstrap, validation, and the complete analysis 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.