

Models and Theory - Student Handout

Models and Theory

Testing Models Theoretically - Presentation Handout

Overview

Summary to Date

Review of linear regression model

Partialing out variance

Combining similar sources of variance

Observable & latent variables

Model fit

Summary to Date

Descriptives vs. Inference

Descriptives good

Simple & intuitive

Can efficiently describe the sample

Robust

Because they make no assumptions about the population

Mean & SD

SD as average distance from the mean

SD as a standard unit of measurement

Standardized (z) scores

Why correlation is so popular

And covariance isn't

Student notes

Descriptive statistics summarize the observations we actually have. Inferential statistics go further: they use a sample, a model, and assumptions about the sampling process to say something about a larger population. A mean or standard deviation can be calculated without assuming that the population is normally distributed, but the measurement still needs to be meaningful and missing data or selection can make even an accurate sample description misleading if we generalize it.

Remember that the standard deviation is the square root of the *average squared* deviation from the mean (with the sample variance usually dividing by $n - 1$); it is not literally the mean absolute distance. A *z* score measures distance from the mean in SD units. Correlation standardizes *covariance* by the two standard deviations, which is why correlations are unit-free and covariances are not.

Descriptives vs. Inference (cont.)

Inference

Making assumptions about the population

Most importantly the distribution
Often assume approximates a normal distribution
But we know we're wrong
Assumptions most **robust** against:
Kurtosis (& skew a bit)
Non-independence of measures ("multicollinearity")
Changes in variance over time ("heteroscedasticity")

Student notes

In ordinary linear regression, normality is an assumption about conditional *errors* when we use the usual exact small-sample tests, not an assumption that every observed variable or the whole population must be normal. Many tests become reasonably accurate with large samples even when errors are somewhat non-normal, but that does not make every assumption harmless.

Multicollinearity means predictors overlap substantially; it can make their *individual coefficients* imprecise, but it is not the same problem as correlation among participants' errors.

Heteroscedasticity is non-constant error variance, which can affect standard errors and inference; heteroscedasticity-robust standard errors or an appropriate variance model may help. Robustness always depends on the estimator, sample size, and severity of the violation.

Descriptives vs. Inference (end)

Inference (cont.)
Assumptions **not robust** against:
Non-independence of participants
Multi-modality (more than one "hump")
Selection bias, even with a large sample
Precision and generalizability:
Larger samples generally improve **precision**
Representative sampling supports **generalizability**
Repeated hypothetical samples describe **sampling variability**

Student notes

Dependence among observations is especially relevant in health research: measurements from the same person, hospital, clinic, household, or neighborhood can resemble one another. Treating these observations as independent can make confidence intervals too narrow. Cluster-robust standard errors, generalized estimating equations, and multilevel models are possible remedies, depending on the research question.

A distribution with several peaks is a prompt to investigate subgroups or a misspecified model; multimodality by itself is not a universal violation of inference. Larger samples generally reduce *random sampling variability*, whereas representative sampling and careful attention to nonresponse are central to *generalizability*. Repeated samples help us understand sampling variability; they do not automatically remove a systematic bias.

Regression Assumptions: Which Problem Are We Solving?

Correlated observations
Standard errors may be incorrect
Clustering, GEE, multilevel models
Heteroscedasticity
Conventional SEs may be incorrect
Robust SEs, variance models
Strong multicollinearity
Individual coefficients become imprecise
Theory-driven grouping or careful retention
Wrong functional form
Predictions and effects can be biased
Residual checks, flexible terms

Student notes

Normality receives a great deal of attention in elementary statistics, but it is far from our only concern. Linear regression's usual exact small-sample coefficient tests rely on assumptions about the errors; the observed predictors and outcomes need not each be normally distributed. Dependence within hospitals and repeated measurements from the same patient can make naive standard errors too small. Heteroscedasticity affects conventional standard errors even when coefficient estimates are still useful under the other required conditions. Multicollinearity is a different problem: it makes it difficult to separate the associations of related predictors. Finally, a model that forces a curved relationship to be straight can misrepresent the phenomenon regardless of sample size. Remedies should follow a diagnosis of the problem, not a checklist applied mechanically.

Central Limit Theorem

Imagine repeatedly sampling from the same population
Each sample gives a slightly different mean
The distribution of those means is the **sampling distribution**
(Under suitable conditions) as sample size increases:

The sampling distribution of the mean approaches normality

The standard error decreases as sample size increases

$$SE(\bar{X}) = \frac{\sigma}{\sqrt{n}}$$

The theorem does **not** say that every statistic or observation is normally distributed

Student notes

The central limit theorem concerns the *sampling distribution* of a statistic, most familiarly the sample mean. Imagine drawing many independent samples of the same size, calculating one mean from each, and looking at the distribution of those means. Under suitable conditions, that distribution approaches a normal distribution as the sample size increases, even when individual observations are not normally distributed.

For independent observations with finite population variance, the standard error of the sample mean is $\sigma/\sqrt{n}$. A bigger sample reduces that standard error, but does not correct measurement error, selection bias, or confounding. The theorem does not say that every statistic has an approximately normal sampling distribution, nor that every sample estimate has nonzero error in every realized sample. The crucial distinction is between the distribution of *individual observations* and the distribution of *estimates across hypothetical samples*.

Variance & Covariance

Variance = Information

Seek to understand that information

The more we understand, the better

Often quantify “how much we understand” as a **signal-to-noise** ratio

$$\text{Variance understood} = \frac{\text{Variance accounted for}}{\text{Variance } not \text{ accounted for}}$$

So, if “accounting” for the effect of one variable on another:

$$\text{Variance understood} = \frac{\text{Covariance}}{\text{Unshared variance}}$$

Which is a correlation

(When it's standardized)

Partialing Out Variance

Ways to increase the size of the signal to the size of the noise:

Increase size of signal

Bigger effects

Greater range of measurements of effects

Decrease size of noise

Greater precision of measurement

Remove the noise

Partial out the variance that is unshared between those variables

But *is* accounted for by some third variable

Student notes

There are two different goals here. Increasing signal may mean improving a study's ability to detect a real relationship, while reducing noise may mean measuring variables more precisely or accounting for relevant background differences. Merely adding predictors does not necessarily improve the validity of an analysis.

When we adjust for a third variable, the coefficient for our predictor describes its association with the outcome *holding that third variable constant*, under the model. We usually do this to address confounding or improve precision, but controlling for a mediator or a collider can change the question or introduce bias. Ask why a covariate belongs in the model before adjusting for it. The best adjustment set should follow the research question and the proposed causal structure, not simply statistical significance.

Ways of Communicating

Merging visuals with text

What is best described where

Visuals as “conversation pieces”

Text as highlighting what to focus on in visuals

Especially vis-à-vis theory

Efficiency & simplicity

“Information-to-ink ratio”

Strong organization

Clear guideposts & structure

Common—but not colloquial—language

Following writing conventions

Avoiding jargon *and* acronyms

Student notes

When presenting statistical results, decide what readers should learn from the graphic and what they need in the surrounding text. A graph can show the shape, pattern, and uncertainty of an association, while the caption and text supply variable definitions, units, sample size, and the substantive interpretation.

Do not treat a sparse-looking figure as permission to omit important information. Report the estimated outcome (the “estimand”), confidence intervals, reference groups, and limitations where needed. Make figures readable in grayscale when possible, use labels rather than color alone, and provide alternative text for students using screen readers. The goal is not merely fewer marks on the page; it is a higher ratio of useful information to visual burden.

Review of Linear Regression

Basic Strategy

Specify how the predictors relate to the outcome

In ordinary linear regression, the expected outcome is linear in the coefficients

Estimate the coefficients

Ordinary least squares or maximum likelihood under a specified error model

Evaluate model fit and assumptions

For example, R^2 , residual plots, and influential observations

Test appropriate hypotheses and interpret coefficients

Report effect sizes and confidence intervals, not just p -values

Student notes

For ordinary least squares, “best” means minimizing the sum of squared residuals, not finding a line that literally passes through or explains all observations. Maximum likelihood instead selects parameter values that maximize the probability density or likelihood assigned to the observed data under a specified model. For normally distributed errors with constant variance, least squares and maximum likelihood give the same slope and intercept estimates.

R^2 is descriptive model fit, not a hypothesis test. An overall regression F -test can evaluate whether a group of slopes is zero, and coefficient tests or confidence intervals address individual terms. A predictor can improve fit without being causally important; conversely, a clinically important predictor can be imprecisely estimated in a small sample. Inspect residuals and model assumptions before interpreting either kind of result.

Partialing Out Variance

We distinguish associations represented by the predictor(s) from unexplained variation

The intercept is the predicted outcome when all predictors equal zero

$$Y = b_0 + b_1 X_1 + e$$

Adding other predictors separates out—partials out—the variance associated with each:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e$$

Student notes

In the first equation, b_0 is the fitted intercept: the expected outcome when the predictor equals zero. That need not describe a real person’s “baseline” unless zero has a meaningful interpretation. b_1 is the expected change in the outcome for a one-unit increase in X_1 under a linear specification, and e represents what the model does not explain for an observation.

When we add X_2 , the interpretation of b_1 changes to a *conditional* association, comparing observations with the same value of X_2 . The predictors may share explanatory variance, so their individual contributions cannot always be summed as if they were entirely distinct. If variables have different units, remember that unstandardized coefficients are expressed in the outcome’s units per unit of the predictor.

Partialing Out Variance

Error can also be separated

E.g., if we know the sources of those errors (same hospital, neighborhood, etc.)

$$Y = b_0 + b_1 X_1 + b_2 X_2 + e_1 + e_2$$

We can also separate out error and effects of predictors over time

Assume that events will be more similar to each other at one time point

And that values *within* a person will tend to be more similar than between people

Different ways to do this,

But I suggest using multilevel models of change for most longitudinal analyses (more later)

Student notes

Repeated observations from the same hospital or patient create levels of variation. For example, a multilevel formulation could include a hospital-specific random intercept u_j and an individual-level residual e_{ij} : $Y_{ij} = \beta_0 + \beta_1 X_{ij} + u_j + e_{ij}$. These are distinct components with specified levels and statistical assumptions; simply writing two errors, without saying what they represent, is not yet a complete model.

For longitudinal data, repeated outcomes within a person can be correlated, and we may allow both starting values and rates of change to vary. If time is a predictor, a model may include a fixed effect of time and patient-specific random intercepts or slopes. Multilevel models are often suitable, but generalized estimating equations or other repeated-measures approaches may be preferable for some questions. Choose the approach to match the estimand and data structure.

Combining Similar Sources of Variance

Common Sources of Variance

Similar predictors may share too much variance

If left un-addressed, can lead to “multicollinearity”X

Which leads to unstable models terms

E.g., terms will flip from being significant to not & back depending on what other terms are added to the model

Usually addressed by removing one of the multicollinear terms

But we can also combine or group those variables...

Student notes

Multicollinearity is chiefly a problem of separating associations among predictors that carry overlapping information. If two predictors are highly correlated, a small change in the sample or adjustment set can alter their individual coefficients, standard errors, or p values, even when their *joint* contribution is stable.

We should not automatically remove one variable. First consider the substantive question, examine the correlation matrix and variance inflation factors, and distinguish problems with interpreting individual coefficients from problems with prediction. If two measures are intended to represent the same construct, a defensible summary score, a latent-variable model, or a prespecified composite may be more informative than mechanically deleting one.

Combining Sources of Variance (cont.)

We do this all the time, in fact

Adding up responses to items on the same survey

Taking the average of a results from a blood draw

But what if we know two variables are related, but not really easily combined?

E.g., distance to fresh-produce retailers and meals skipped because of insufficient money

A mean distance of 2 km + 3 days of skipped meals $\neq$ a meaningful score of 5

Student notes

Combining variables is reasonable when we have a substantive and measurement-based argument for doing so. For survey items, we should check that they are coded in the same direction, represent a coherent construct, and have sufficient reliability; averaging unrelated items may hide meaningful differences. Repeated measurements of the same biological quantity can sometimes be averaged to improve precision, but systematic shifts or clinically distinct measurements should not be collapsed without justification.

Distance to retailers selling fresh fruits and vegetables and skipping meals because of insufficient money are quite different observations. The first is a geographic measure of potential access, expressed here as the mean distance in kilometers to the three nearest fresh-produce retailers; the second is a count of days on which someone skipped meals. Adding distances to counts has no interpretable meaning. Even standardizing both variables would not make them interchangeable or establish that they measure a single construct.

Observable and Latent Variables

Observable and Latent Variables

Some Things We Can See...

Distance to fresh-produce retailers & skipped meals

Blood pressure & adipose tissue

Smiles and cortisol levels

Some Things We Can't See...

"Food insecurity"

"Health"

"Stress"

Student notes

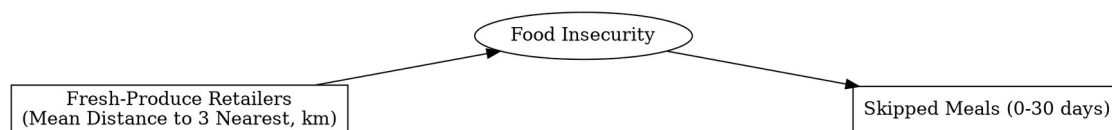
An observed variable is something we have actually recorded, even when the measurement is imperfect. A latent construct is inferred from a set of measures and a theoretical account of how they relate. Blood pressure is measured with an instrument; general health or perceived stress cannot be captured completely by one reading.

Latent does not mean imaginary, nor does observed mean perfectly measured. A study must define what each indicator records, on what scale, and why it should reflect or contribute to the underlying construct. Validity is a question about the interpretation and intended use of scores, not something a statistical model can establish automatically.

Observable and Latent Variables (cont.)

Things we can observe empirically are sometimes called, well, **observable**

While underlying constructs that help explain them may be **latent**



Conceptual pathway relating food access, food insecurity, and skipped meals.

Student notes

Food insecurity is a social determinant of health that cannot be completely captured with one measurement. The two rectangles show quite different kinds of observed variables: the mean distance in kilometers to the three nearest retailers offering fresh fruits and vegetables, and the number of days in the past 30 on which a person skipped meals because of insufficient money. The former concerns the local food environment; the latter concerns an experience of material hardship. Distance should ideally be measured along accessible travel routes, with an explicit definition of which retailers qualify. Distance alone says little about affordability, opening hours, transportation, or the quality of available food.

Notice the **direction of the arrows**. Greater distance may contribute to difficulties accessing food, but food insecurity does not cause a retailer to be farther away. By contrast, skipping meals for financial reasons may be a consequence of food insecurity. The diagram is therefore a simplified *conceptual pathway*, not a two-indicator reflective measurement model. Neither arrow proves causality. An estimable latent-variable model would require an explicit measurement specification and usually additional validated indicators or identifying constraints.

Model Testing

Model Fit

We can enter theoretically related predictors as a **block**

Reduced model: baseline predictors

Full model: baseline predictors **plus** the new block

Compare explained variance using the **same participants**:

$$\Delta R^2 = R^2_{full} - R^2_{reduced}$$

In ordinary nested linear regression:

ΔR^2 describes the added explained variance

A **partial F test** can test whether the block improves fit

Improvement does not, by itself, establish causation

Student notes

Adding a *block* of theoretically related predictors is often called hierarchical or blockwise regression. Fit a reduced model with the baseline predictors, then a full model that adds the prespecified block, using the same participants in each comparison. The question is whether the additional block improves prediction or fit sufficiently to justify its complexity.

The equations displayed for R^2 should be read as shorthand for the *models from which R^2 is calculated*; R^2 itself is a number, not a regression equation. If the full model contains everything in the reduced model plus the new block, the incremental explained variance is $\Delta R^2 = R^2_{full} - R^2_{reduced}$, which is nonnegative in ordinary least squares on the same data.

A nested-model F test can assess whether that increase is larger than expected from sampling variation.

Model Selection Beyond R^2

For models fitted with a comparable likelihood, we can use **information criteria**

Especially useful when candidate models are not nested

Lower values are preferred among the models compared

Two common criteria:

Akaike Information Criterion (AIC)

Bayesian Information Criterion (BIC)

Both penalize additional parameters

BIC generally penalizes complexity more strongly in larger samples

Neither provides a p -value or replaces diagnostics

Student notes

AIC and BIC help us compare candidate models by balancing fit to the observed data against the number of estimated parameters. They are based on a likelihood (or another explicitly compatible likelihood-derived quantity). Compare models fitted to the same outcome and observations, with comparable likelihoods and estimation conventions; a lower value is favored.

Ordinary least squares slopes are also maximum-likelihood slopes under the usual independent, normally distributed, constant-variance error model. Maximum likelihood is not inherently “more robust” than least squares; both inherit assumptions from the model being fit. BIC uses a stronger complexity penalty than AIC for sufficiently large samples, but the choice between them should follow whether prediction, explanatory model selection, or another inferential goal is most important. Neither criterion is a conventional p value.

Model Testing (cont.)

The **-2 log likelihood (-2LL)** is a measure of **misfit**

So larger numbers indicate poorer fit

Thus want to see that the model with the “block” of interest has a smaller -2LL

I.e.:

Improvement in Model Fit = Δ

$$(-2\ell)_{\text{Model without Block}} - (-2\ell)_{\text{Model with Block}}$$

For example, if the reduced model has -2LL = 2010 and the full model -2LL = 2000:

$$\text{Improvement in Model Fit} = 2010 - 2000 = 10$$

Student notes

The -2 log likelihood is the negative of twice the maximized log likelihood. Smaller values indicate better fit to the observed data, when models use the same outcome, observations, and comparable likelihoods. The numbers 2010 and 2000 are meaningful for their difference rather than in isolation.

Adding a block of predictors cannot worsen the maximized likelihood in a properly fitted nested model. Here, the full model improves -2LL by 10 units. Unlike an AIC or BIC difference, this difference does not already include a penalty for additional parameters. For nested maximum-likelihood models that meet the likelihood-ratio test assumptions, this difference can be compared with a chi-square distribution using the number of added parameters as the test degrees of freedom.

Model Testing (cont.)

For nested models fitted by maximum likelihood to the same data,

Test the -2LL difference against a χ^2 distribution

With degrees of freedom equal to the number of added parameters

E.g.:

$$b_1 X_1$$

$$b_1 X_1 + (b_{\text{Distance}} X_{\text{Distance}} + b_{\text{Skipped}} X_{\text{Skipped}})$$

The second model estimates 2 additional continuous parameters

$$\text{So } \chi^2_{LR} = 2010 - 2000 = 10 \text{ with } df = 2$$

Significant at $\alpha = .05$: critical $\chi^2 \approx 5.99$; $p \approx .0067$

Model Testing (cont.)

The full model estimates **2 additional parameters** (cont.)

Assuming each added predictor requires one coefficient (cont.)

Test degrees of freedom: $\Delta df = 2$

$$\text{Likelihood-ratio statistic: } \chi^2_{LR} = 2(\ell_{\text{full}} - \ell_{\text{reduced}})$$

If the -2LLs are **2010** and **2000**, respectively:

$$\chi^2_{LR} = (-2\ell)_{\text{reduced}} - (-2\ell)_{\text{full}} = 2010 - 2000 = 10$$

Compare $\chi^2 = 10$ with a chi-square distribution with 2 *df*

1 *df* to estimate effect of distance to fresh produce

1 *df* to estimate number of meals skipped

Critical $\chi^2 \approx 5.99$ at $\alpha = .05$; $p \approx .0067$

Student notes

The likelihood-ratio test determines whether the improvement in model fit is greater than we would expect under the reduced model, assuming the null hypothesis and the test's regularity conditions hold.

The full model includes two additional coefficients. Consequently, the likelihood-ratio test has two degrees of freedom. This is different from residual degrees of freedom, which ordinarily decrease when we add predictors.

The likelihood-ratio statistic is the reduced model's -2 log likelihood minus the full model's -2 log likelihood. Here, 2010 minus 2000 equals 10. Because we are already using -2LL rather than an information criterion, we do not add a parameter penalty to this difference.

The chi-square test uses the upper tail of the chi-square distribution. The critical value at the 5% significance level with two degrees of freedom is approximately 5.99. Our statistic of 10 exceeds that threshold ($p \approx .0067$), indicating that the additional predictors collectively improve fit according to this test.

This approximation is appropriate for regular nested maximum-likelihood models, such as many logistic regression models. For conventional nested linear regression, we generally use a partial F test instead.

Finally, mean distance to the three nearest fresh-produce retailers is a continuous geographic measure, whereas skipped-meal days is a count. Entering each as one numeric predictor gives one coefficient per variable, but assumes its effect is appropriately linear on the model's link scale. For example, representing a nonlinear relationship with several spline terms for distance would add parameters and change the test degrees of freedom. The earlier food-insecurity diagram is a conceptual pathway: distance may affect food insecurity, and food insecurity may contribute to skipped meals. Estimating that pathway as a latent-variable model is a separate statistical task.

Which Model Comparison Should We Use?

Did an OLS predictor block add explanatory value?

Nested-model partial F test

$$\Delta R^2, F, \text{df}, p$$

Did nested ML models differ in fit?

Likelihood-ratio test

$$\chi^2_{LR}, \text{df}, p$$

Which candidate model balances fit and complexity?

AIC or BIC

Scores, differences, rationale

Student notes

The method follows the question, not the availability of a statistic in the software output. A partial F test addresses a specified set of additional coefficients in ordinary linear regression. A likelihood-ratio test compares nested likelihood-based models under the relevant assumptions; logistic regression is a common example. AIC and BIC support relative comparison of candidate models, including some nonnested models, but do not provide a significance test. For every method, use compatible likelihoods, the same outcome and the same observations. When comparing different fixed effects in a linear mixed model, use maximum likelihood rather than restricted maximum likelihood for the model-comparison step. None of these procedures replaces substantive theory, residual diagnostics, or validation.