Methods of Applied Statistics[1]

STA442/2101 Fall 2017

[1] See last slide for copyright information.

Goal of statistical analysis

The goal of statistical analysis is to draw reasonable conclusions from noisy numerical data.

Steps in the process of statistical analysis

One approach

- Consider a fairly realistic example or problem.
- Decide on a statistical model.
- Perhaps decide sample size.
- Acquire data.
- Examine and clean the data; generate displays and descriptive statistics.
- Estimate model parameters, for example by maximum likelihood.
- Carry out tests, compute confidence intervals, or both.
- Perhaps re-consider the model and go back to estimation.
- Based on the results of estimation and inference, draw conclusions about the example or problem.

Two domains

- The messy outside world of reality and data.
- The pure world of the statistical model.
- Applied Statistics navigates the interface.

What is a statistical model?

You should always be able to state the model

A *statistical model* is a set of assertions that partly specify the probability distribution of the observable data. The specification may be direct or indirect.

- Let $X_1, \ldots, X_n$ be a random sample from a Poisson distribution with expected value λ.
- For $i = 1, \ldots, n$, let $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$, where

  $\beta_0, \ldots, \beta_{p-1}$ are unknown constants (parameters).
  $x_{i,j}$ are known constants.
  $\epsilon_1, \ldots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.
  σ^2 is an unknown constant (parameter).
  $Y_1, \ldots, Y_n$ are observable random variables.

Is the model the same thing as the *truth*?

Parameter Space

The *parameter space* is the set of values that can be taken on by the parameter.

- Let $X_1, \dots, X_n$ be a random sample from a normal distribution with expected value μ and variance σ^2. The parameter space is $\{(\mu, \sigma^2) : -\infty < \mu < \infty, \sigma^2 > 0\}$.
- For $i = 1, \dots, n$, let $Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$, where

  $\beta_0, \dots, \beta_{p-1}$ are unknown constants.
  $x_{i,j}$ are known constants.
  $\epsilon_1, \dots, \epsilon_n$ are independent $N(0, \sigma^2)$ random variables.
  σ^2 is an unknown constant.
  $Y_1, \dots, Y_n$ are observable random variables.

  The parameter space is $\{(\beta_0, \dots, \beta_{p-1}, \sigma^2) : -\infty < \beta_j < \infty, \sigma^2 > 0\}$.

The SENIC data

Study of the

Efficacy of

Nosocomial

Infection

Control

Background

The Study on the Efficacy of Nosocomial Infection Control (SENIC) data are from a study of infections acquired in hospital. That is, patients are admitted to hospital for something, and while in hospital they get infections (such as pneumonia and urinary tract infections) that are unrelated to why they were admitted, and require treatment. This is a partial reconstructed data set based on one in Kutner et al.'s *Applied Linear Statistical Models*. In this aggregated data set, the cases are 100 U.S. hospitals.

Variables in the openSENIC data set

	Hospital identification number
region	Geographic region of U.S.
mdschl	Medical school affiliation (Yes or No)
census	Number of patients
nbeds	Number of beds in the hospital
nurses	Number of nurses
lngstay	Mean length of stay in days
age	Mean age of patient in years
xratio	Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, × 100
culratio	Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, × 100
infpercent	Percentage of patients who acquired an infection while in hospital

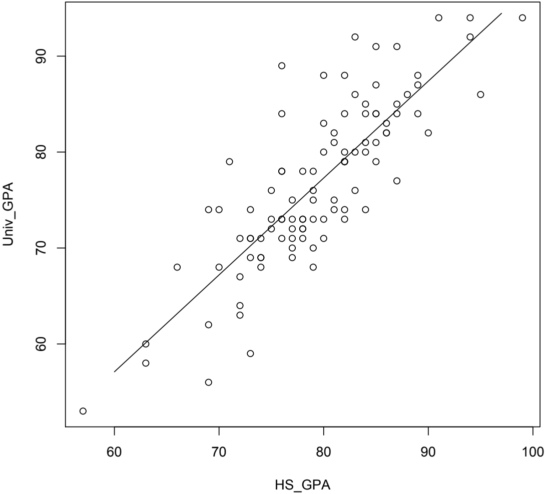
Xratio and Culratio

Two strange variables

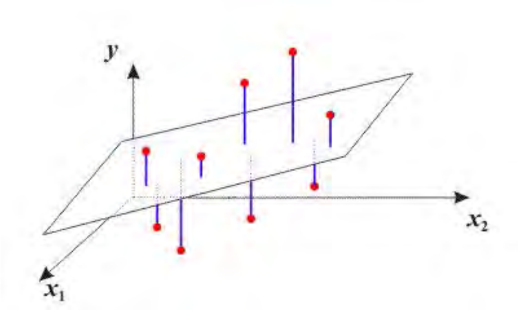
- **xratio**: Ratio of number of X-rays performed to number of patients without signs or symptoms of pneumonia, × 100
- **culratio**: Ratio of number of cultures performed to number of patients without signs or symptoms of hospital-acquired infection, × 100

Sometimes mini-epidemics spread through a hospital. The variables xratio and culratio represent special efforts to monitor the health of patients who show no signs of having gotten sick in hospital, yet. They are a kind of early warning system, intended to detect outbreaks of disease in the hospital so they can be dealt with before they get established. My guess is that xratio is primarily for pneumonia, and culratio is primarily for urinary tract infections.

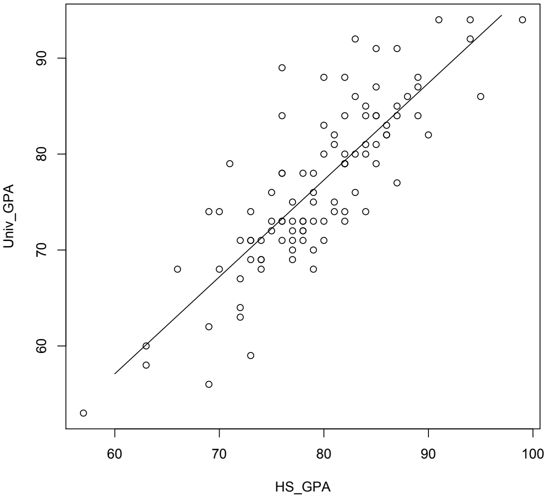
Let’s do “Regression”



Least squares Plane



Regression?



Francis Galton (1822-1911) studied “Hereditary Genius” (1869) and other traits

- Heights of fathers and sons
    - Sons of the tallest fathers tended to be taller than average, but shorter than their fathers.
    - Sons of the shortest fathers tended to be shorter than average, but taller than their fathers.
- This kind of thing was observed for many other traits.
- Galton was deeply concerned about “regression to mediocrity.”

Measure the same thing twice, with error

Suppose that's all it is

$Y_1 = X + e_1$
$Y_2 = X + e_2$
$X \sim N(\mu, \sigma^2_x)$
e_1 and e_2 independent $N(0, \sigma^2_e)$

Then,

$$\left(\begin{array}{c} Y_1 \\ Y_2 \end{array} \right) \sim N\left(\left[\begin{array}{c} \mu \\ \mu \end{array} \right], \left[\begin{array}{cc} \sigma^2_x + \sigma^2_e & \sigma^2_x \\ \sigma^2_x & \sigma^2_x + \sigma^2_e \end{array} \right] \right)$$

Conditional distribution of Y_2 given $Y_1 = y_1$

For the bivariate normal

$$\begin{aligned} & N\left(\mu_2 + \frac{\sigma_2}{\sigma_1}\,\rho(y_1 - \mu_1), (1-\rho^2)\sigma_2^2\right) \\ = \quad & N\left(\mu + \rho(y_1 - \mu), (1-\rho^2)(\sigma_x^2 + \sigma_e^2)\right) \text{ where} \\ \rho \quad = \quad & \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}. \end{aligned}$$

Have

$$\begin{aligned} E(Y_2 | Y_1 = y_1) &= \mu + \rho(y_1 - \mu) \\ &= \mu(1-\rho) + \rho\, y_1 \\ &= \qquad \beta_0 \quad + \beta_1\, y_1 \end{aligned}$$

Regression.

$E(Y_2|Y_1 = y_1) = \mu + \rho(y_1 - \mu)$

Where $\rho = \frac{\sigma_x^2}{\sigma_x^2 + \sigma_e^2}$

- If y_1 is above the mean, average y_2 will also be above the mean.
- But only a fraction ρ as far above as y_1.
- If y_1 is below the mean, average y_2 will also be below the mean.
- But only a fraction ρ as far below as y_1.
- This exactly the “regression toward the mean” that Galton observed.

Regression toward the mean

- Does not imply systematic change over time.
- Is a characteristic of the bivariate normal and other joint distributions.
- Can produce very misleading results, especially in the evaluation of social and educational programs.

Regression Artifact

- Measure something important, like performance in school or blood pressure.
- Select an extreme group, usually those who do worst on the baseline measure.
- Do something useless to help them, and measure again.

$$E(Y_2|Y_1 = y_1) = \mu + \rho(y_1 - \mu)$$

Even if the treatment does nothing, they are expected to do worse than average, but still better than they did the first time - completely artificial!
And the apparent effect for those who do better than average on baseline will be negative.

General Mixed Linear Model

With both fixed and random effects

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\mathbf{b} + \boldsymbol{\epsilon}$$

- $\mathbf{X}$ is an $n \times p$ matrix of known constants.
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants.
- $\mathbf{Z}$ is an $n \times q$ matrix of known constants.
- $\mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\Sigma}_b$ unknown.
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.

Fixed Effects Linear Regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- $\mathbf{X}$ is an $n \times p$ matrix of known constants.
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants.
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.

- $\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$
- $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$
- $\mathbf{e} = (\mathbf{y} - \widehat{\mathbf{y}})$

Comparing scalar and matrix form

Scalar form is $y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$

$$\begin{array}{ccccc} \mathbf{y} & = & \mathbf{X} & \boldsymbol{\beta} & + & \boldsymbol{\epsilon} \\ \begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} & = & \begin{pmatrix} 1 & 14.2 & \cdots & 1 \\ 1 & 11.9 & \cdots & 0 \\ 1 & 3.7 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6.2 & \cdots & 1 \end{pmatrix} & \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} & + & \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix} \end{array}$$

Vocabulary

- Explanatory variables are x
- Response variable is y
- For the SENIC data, what's the response variable?

Meaning of the statistical model

$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$

- There are $p - 1$ explanatory variables.
- For each *combination* of explanatory variables, the conditional distribution of the response variable y is normal, with constant variance σ^2.
- The conditional population mean of y depends on the x values, as follows:

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$$

"Control" means hold constant

- Regression model with four explanatory variables.
- Hold x_1, x_2 and x_4 constant at some fixed values.

$$\begin{aligned} E(Y|\boldsymbol{X}=\boldsymbol{x}) &= \beta_0+\beta_1x_1+\beta_2x_2+\beta_3x_3+\beta_4x_4 \\ &= (\beta_0+\beta_1x_1+\beta_2x_2+\beta_4x_4)+\beta_3x_3 \end{aligned}$$

- The equation of a straight line with slope β_3.
- Values of x_1, x_2 and x_4 affect only the intercept.
- So β_3 is the rate at which $E(Y|\mathbf{x})$ changes as a function of x_3 with all other variables held constant at fixed levels.
- *According to the model.*

It's model-based control

- To "hold x_1 constant" at some particular value, like $x_1 = 14$, you don't even need data at that value.
- Ordinarily, to estimate $E(Y|X_1 = 14, X_2 = x)$, you would need a lot of data at $X_1 = 14$.
- But look: $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot 14 + \widehat{\beta}_2 x$
- Of course extrapolation is still dangerous.

Three Meanings of “Control”

- Procedural
- Sub-division
- Model-based

More vocabulary

$E(Y|\boldsymbol{X} = \boldsymbol{x}) = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4) + \beta_3 x_3$

- If $\beta_3 > 0$, describe the relationship between x_3 and (expected) y as "positive," controlling for the other variables. If $\beta_3 < 0$, negative.
- Useful ways of saying "controlling for" or "holding constant" include
  - Allowing for
  - Correcting for
  - Taking into account

Categorical Explanatory Variables

Unordered categories

- $X = 1$ means Drug, $X = 0$ means Placebo.
- Population mean is $E(Y|X = x) = \beta_0 + \beta_1 x$.
- For patients getting the drug, mean response is $E(Y|X = 1) = \beta_0 + \beta_1$
- For patients getting the placebo, mean response is $E(Y|X = 0) = \beta_0$
- And β_1 is the difference between means, the average treatment effect.

Similar interpretation for sample regression coefficients

Think of a scatterplot

- $X = 1$ means Drug, $X = 0$ means Placebo.
- Predicted response is $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 x$.
- For patients getting the drug, predicted response is $\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 = \overline{Y}_1$.
- For patients getting the placebo, predicted response is $\widehat{Y} = \widehat{\beta}_0 = \overline{Y}_0$.
- And $\widehat{\beta}_1$ is the difference between sample means, the estimated average treatment effect.

Regression test of $H_0 : \beta_1 = 0$

With a binary explanatory variable

- Same as an independent t-test.
- Same as a one-way ANOVA with 2 categories.
- Same t, same F, same p-value.

Testing Statistical Hypotheses

- Parameters, parameter space
- Null hypothesis
- Significance level α.
- p-value.
- Function in the discourse of Science: Skeptic in a box.

Plain language is important

- If you can only be understood by mathematicians and statisticians, your knowledge is much less valuable.
- Often a question will say "Give the answer in plain, non-statistical language."
- This means if x is income and y is credit card debt, you make a statement about income and average or predicted credit card debt, like "customers with higher incomes tend to have less credit card debt."
- If you use mathematical notation or words like null hypothesis, unbiased estimator, p-value or statistically significant, you will lose a lot of marks even if the statement is correct. Even avoid "positive relationship," and so on.
- If the study is about fish, talk about fish.
- If the study is about blood pressure, talk about blood pressure.
- If the study is about breaking strength of yarn, talk about breaking strength of yarn.
- Assume you are talking to your boss, a former Commerce major who got a D+ in ECO220 and does not like to feel stupid.

We will be guided by hypothesis tests with $\alpha = 0.05$

For plain-language conclusions

- If we do not reject a null hypothesis like $H_0 : \beta_1 = 0$, we will not draw a definite conclusion.
- Instead, say things like:
  - There is no evidence of a connection between blood sugar level and mood.
  - These results are not strong enough for us to conclude that attractiveness is related to mark in first-year Computer Science.
  - These results are consistent with no effect of dosage level on bone density.
- If the null hypothesis is not rejected, please do *not* claim that the drug has no effect, etc..
- In this we are taking Fisher's side in a historical fight between Fisher on one side and Neyman & Pearson on the other.
- Though we are guided by $\alpha = 0.05$, we *never* mention it when plain language is required.

A technical issue

- In this class we will mostly avoid one-tailed tests.
- Why? Ask what would happen if the results were strong and in the opposite direction to what was predicted (dental example).
- But when H_0 is rejected, we still draw directional conclusions.
- For example, if x is income and y is credit card debt, we test $H_0 : \beta_1 = 0$ with a two-sided t-test.
- Say $p = 0.0021$ and $\widehat{\beta}_1 = 1.27$. We say “Consumers with higher incomes tend to have more credit card debt.”
- Is this justified? We’d better hope so, or all we can say is “There is a connection between income and average credit card debt.”
- Then they ask: “What’s the connection? Do people with lower income have more debt?”
- And you have to say “Sorry, I don’t know.”
- It’s a good way to get fired, or at least look silly.

The technical resolution

- Decompose the two-sided test into a set of two one-sided tests with significance level $\alpha/2$, equivalent to the two-sided test (explain and draw a picture).
- In practice, just look at the sign of the regression coefficient.
- Under the surface you are decomposing the two-sided test, but you never mention it.
- *Marking rule*: If the question asks for plain language and you draw a non-directional conclusion when a directional conclusion is possible, you get half marks at most.

Correlation-causation

More on how to talk (and think) about the results

- Suppose the two conditions were standard treatment versus new treatment.
- $x = 0$ means standard treatment, $x = 1$ means new treatment.
- We could collect data on people who were treated for the disease, observe whether they got the standard treatment or the new treatment, and also observe y to see how they did.
- Suppose $H_0 : \mu_1 = \mu_2$ is rejected, and patents receiving the new treatment did better on average.
- Is the new treatment better?
- Maybe, but it's also possible that those receiving the new treatment were more motivated, or more educated, or healthier in the first place (so they have energy to pursue non-standard options).
- Controlling for those possibilities is a good idea, but will you think of everything?
- The standard saying is "Correlation does not imply causation."
- Correlation means association between variables.
- Causation means influence, not absolute determination.

More examples

- Wearing a hat and baldness.
- Exercise and arthritis pain.
- The Mozart effect.
- Alchohol consumption and health.

In general

If A and B are related

- A could be influencing B.
- B could be influencing A.
- C could be influencing both A and B.
- C is called a *confounding variable.*

Confounding variable

- Is related to both the explanatory variable and response variable
- Causing an apparent relationship.
- A and B are related only because they are both related to C.
- Exercise and health. You'd better control for age.
- Controlling for age may not be enough.

The discourse of science

- These conversations follow a pattern.
- “Canadians who eat more fresh (less processed) food tend to have better health. Eat fresh vegetables!”
- Reply: It’s possible that people who eat fresh food tend to have higher income, and they are healthier for reasons connected to income rather than diet. Consider Aboriginals on reserve? Did you control for income?
- “We couldn’t. Information about income was not available. But we controlled for education. And we had a large sample.”
- Reply: Sorry, you lose.

To shoot a study down on the basis of confounding variables

All you need to do is

- Name a specific *potential* confounding variable that is not in the data.
- Make a plausible case for why it *might* be related to the explanatory variable.
- Make a plausible case for why it might be related to the response variable.
- Of course if you can cite specific published results your argument is even stronger.
- Sometimes this is all it takes to show that person-years of effort and hundreds of thousands of dollars in grant money were wasted.

The solution: Random assignment

Again, $x = 0$ means standard treatment and $x = 1$ means new treatment

- What if patients were randomly assigned to treatment?
- In an *experimental study*, subjects are randomly assignent to treatment conditions — values of a categorical explanatory variable — and values of the response variable are observed.
- In an *observational study*, values of the explanatory and response variables are just observed.
- In a well-designed experimental study, confounding variables are ruled out.
- $B \rightarrow A$ is ruled out too.
- Thank you, Mr. Fisher.

Talking about the results of a purely observational study

Avoid language that implies causality or influence.

- Don't say "Music lessons led to better academic performance."
- Say "Students who had private music lessons tended to have better academic performance."
- A good follow-up might be "Music lessons may stimulate cognitive development, but it's also possible that students who had private music lessons were different in other ways, such as average income or parents' education."
- Don't say "Solving puzzles on a regular basis tended to provide protection against the development of dementia."
- Say "Participants who solved puzzles on a regular basis tended to develop dementia later in life than those who did not solve puzzles on a regular basis."
- It is okay to follow up with "Solving puzzles may provide mental simulation that slows the onset of dementia."
- But then say "Or, it is possible that early stages of dementia that are difficult to detect may lead to decreased interest in solving puzzles."

More than Two Categories

We need this for the SENIC study

Suppose a study has 3 treatment conditions. For example Group 1 gets Drug 1, Group 2 gets Drug 2, and Group 3 gets a placebo, so that the Explanatory Variable is Group (taking values 1,2,3) and there is some Response Variable Y (maybe response to drug again).

Why is $E[Y|X = x] = \beta_0 + \beta_1 x$ (with x = Group) a silly model?

Indicator Dummy Variables

With intercept

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.
- Fill in the table.

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A			$\mu_1 =$
B			$\mu_2 =$
Placebo			$\mu_3 =$

Answer

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$.

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

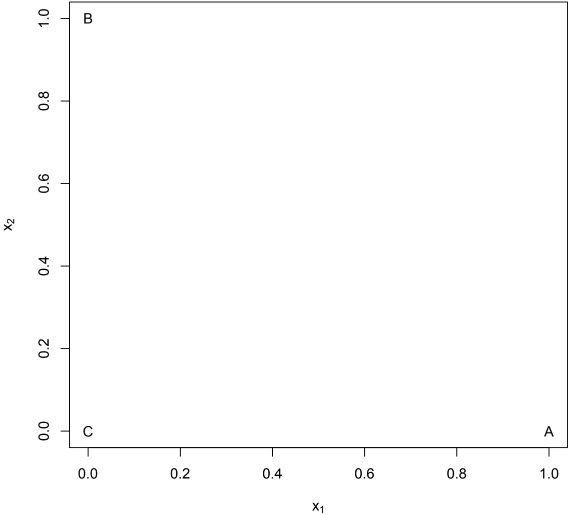
Regression coefficients are contrasts with the category that has no indicator – the *reference category*.

Indicator dummy variable coding with intercept

- With an intercept in the model, need $p-1$ indicators to represent a categorical explanatory variable with p categories.
- If you use p dummy variables and an intercept, trouble.
- Regression coefficients are contrasts with the category that has no indicator.
- Call this the *reference category*.

$x_1 = 1$ if Drug A, zero o.w., $x_2 = 1$ if Drug B, zero o.w.

Recall $\sum_{i=1}^{n}(y_i - m)^2$ is minimized at $m = \overline{y}$



What null hypotheses would you test?

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

- Is the effect of Drug A different from the placebo? $H_0 : \beta_1 = 0$
- Is Drug A better than the placebo? $H_0 : \beta_1 = 0$
- Did Drug B work? $H_0 : \beta_2 = 0$
- Did experimental treatment have an effect? $H_0 : \beta_1 = \beta_2 = 0$
- Is there a difference between the effects of Drug A and Drug B? $H_0 : \beta_1 = \beta_2$

Now add a quantitative explanatory variable (covariate)

Covariates often come first in the regression equation

- $x_1 = 1$ if Drug A, zero otherwise
- $x_2 = 1$ if Drug B, zero otherwise
- $x_3 =$ Age
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$.

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = (\beta_0 + \beta_1) + \beta_3 x_3$
B	0	1	$\mu_2 = (\beta_0 + \beta_2) + \beta_3 x_3$
Placebo	0	0	$\mu_3 = \quad \beta_0 \quad + \beta_3 x_3$

Parallel regression lines.

More comments

Drug	x_1	x_2	$E(Y\|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = (\beta_0 + \beta_1) + \beta_3 x_3$
B	0	1	$\mu_2 = (\beta_0 + \beta_2) + \beta_3 x_3$
Placebo	0	0	$\mu_3 = \quad \beta_0 \quad + \beta_3 x_3$

- If more than one covariate, parallel regression planes.
- Non-parallel (interaction) is testable.
- "Controlling" interpretation holds.
- In an experimental study, quantitative covariates are usually just observed.
- Could age be related to drug?
- Good covariates reduce MSE, make testing of categorical variables more sensitive.

Cell means coding: p indicators and no intercept

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

Drug	x_1	x_2	x_3	$E(Y\|\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	0	$\mu_1 = \beta_1$
B	0	1	0	$\mu_2 = \beta_2$
Placebo	0	0	1	$\mu_3 = \beta_3$

- This model is equivalent to the one with $p - 1$ dummy variables and the intercept.
- If you have p dummy variables and the intercept, the model is over-parameterized.

Add a covariate: x_4

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$$

Drug	x_1	x_2	x_3	$E(Y\|\mathbf{x}) = \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
A	1	0	0	$\beta_1 + \beta_4 x_4$
B	0	1	0	$\beta_2 + \beta_4 x_4$
Placebo	0	0	1	$\beta_3 + \beta_4 x_4$

This model is equivalent to the one with the intercept.

Key to the equivalence of dummy variable coding schemes

Drug	x_0	x_1	x_2	$\beta_0 x_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	1	0	$(\beta_0 + \beta_1) + \beta_3 x_3$
B	1	0	1	$(\beta_0 + \beta_2) + \beta_3 x_3$
Placebo	1	0	0	$\beta_0 \qquad + \beta_3 x_3$

Drug	x_1	x_2	x_3	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4$
A	1	0	0	$\beta_1 + \beta_4 x_4$
B	0	1	0	$\beta_2 + \beta_4 x_4$
Placebo	0	0	1	$\beta_3 + \beta_4 x_4$

$$
\begin{aligned}
& \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \\
\Leftrightarrow \quad & \mathbf{y} = (\mathbf{X}\mathbf{A})(\mathbf{A}^{-1}\boldsymbol{\beta}) + \boldsymbol{\epsilon}
\end{aligned}
$$

Partitioning sums of squares

- Variation to explain: **Total Sum of Squares**

$$\text{SSTO} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

- Variation that is still unexplained: **Error (or Residual) Sum of Squares**

$$\text{SSE} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

- Variation that is explained: **Regression (or Model) Sum of Squares**

$$\text{SSR} = \text{SSTO} - \text{SSR} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

R^2: Proportion of variation in y that is explained

- $SSTO = SSR + SSE$
- Proportion of variation in the response variable that is explained by the explanatory variable

$$R^2 = \frac{SSR}{SSTO}$$

- For a simple regression, same as the square of the correlation coefficient: $r^2 = R^2$

Hypothesis Testing

Standard tests when errors are normal

- Overall F-test for all the explanatory variables at once $H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$
- t-tests for each regression coefficient: Controlling for all the others, does that explanatory variable matter? $H_0 : \beta_j = 0$
- Test a collection of explanatory variables controlling for another collection $H_0 : \beta_2 = \beta_3 = \beta_5 = 0$
- Example: Controlling for mother's education and father's education, are (any of) total family income, assessed value of home and total market value of all vehicles owned by the family related to High School GPA?
- Most general: Testing whether sets of linear combinations of regression coefficients differ from specified constants. $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$.

Full versus Restricted Model

Restricted by H_0

- You have 2 sets of variables, A and B. Want to test B controlling for A.
- Fit a model with both A and B: Call it the *Full Model.*
- Fit a model with just A: Call it the *Restricted Model.* $R^2_F \geq R^2_R$.
- The F-test is a likelihood ratio test (exact).

When you add the r more explanatory variables in set B, R^2 can only go up

By how much? Basis of the F test.

$$\begin{aligned} F &= \frac{(R^2_F - R^2_R)/r}{(1-R^2_F)/(n-p)} \\ &= \frac{(SSR_F - SSR_R)/r}{MSE_F} \overset{H_0}{\sim} F(r, n-p) \end{aligned}$$

General Linear Test of $H_0 : \mathbf{L}\boldsymbol{\beta} = \mathbf{h}$

$\mathbf{L}$ is $r \times p$, rows linearly independent

$$F = \frac{(\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})^\top (\mathbf{L}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{L}^\top)^{-1} (\mathbf{L}\widehat{\boldsymbol{\beta}} - \mathbf{h})}{r \, MSE_F}$$

$$\overset{H_0}{\sim} \; F(r, n-p)$$

Equal to full-restricted formula.

Are the x values really constants?

$Y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{p-1} x_{i,p-1} + \epsilon_i$

- In the general linear regression model, the $\mathbf{X}$ matrix is supposed to be full of fixed constants.
- This is convenient mathematically. Think of $E(\widehat{\boldsymbol{\beta}})$.
- But in any non-experimental study, if you selected another sample, you'd get different $\mathbf{X}$ values, because of random sampling.
- So $\mathbf{X}$ should be at least partly random variables, not fixed.
- View the usual model as *conditional* on $\mathbf{X} = \mathbf{x}$.
- All the usual probabilities and expected values are *conditional* probabilities and *conditional* expected values.
- But this would seem to mean that the *conclusions* are also conditional on $\mathbf{X} = \mathbf{x}$.

$\widehat{\boldsymbol{\beta}}$ is (conditionally) unbiased

$$E(\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x})=\boldsymbol{\beta} \text{ for } \textit{any} \text{ fixed } \mathbf{x}.$$

It's *unconditionally* unbiased too.

$$E\{\widehat{\boldsymbol{\beta}}\}=E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}\}\}=E\{\boldsymbol{\beta}\}=\boldsymbol{\beta}$$

Perhaps Clearer

$$\begin{aligned}
E\{\widehat{\boldsymbol{\beta}}\} &= E\{E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}\}\} \\
&= \int\cdots\int E\{\widehat{\boldsymbol{\beta}}|\mathbf{X}=\mathbf{x}\}\,f(\mathbf{x})\,d\mathbf{x} \\
&= \int\cdots\int \boldsymbol{\beta}\,f(\mathbf{x})\,d\mathbf{x} \\
&= \boldsymbol{\beta}\int\cdots\int f(\mathbf{x})\,d\mathbf{x} \\
&= \boldsymbol{\beta}\cdot 1 = \boldsymbol{\beta}.
\end{aligned}$$

Conditional size α test, Critical region A

$$Pr\{F \in A|\mathbf{X} = \mathbf{x}\} = \alpha$$

$$\begin{aligned}
Pr\{F \in A\} &= \int \cdots \int Pr\{F \in A|\mathbf{X} = \mathbf{x}\} f(\mathbf{x})\, d\mathbf{x} \\
&= \int \cdots \int \alpha f(\mathbf{x})\, d\mathbf{x} \\
&= \alpha \int \cdots \int f(\mathbf{x})\, d\mathbf{x} \\
&= \alpha
\end{aligned}$$

The moral of the story

- Don't worry.
- Even though X variables are often random, we can apply the usual fixed-x model without fear.
- Estimators are still unbiased.
- Tests have the right Type I error probability.
- Similar arguments apply to confidence intervals and prediction intervals.
- And it's all distribution-free with respect to X.

Copyright Information