Some Comments on Linear Regression[1]

STA442/2101 Fall 2018

[1] See last slide for copyright information.

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

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

$$\begin{pmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & 14.2 & \cdots & 1 \\ 1 & 11.9 & \cdots & 0 \\ 1 & 3.7 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 6.2 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_{p-1} \end{pmatrix} + \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \epsilon_3 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Vocabulary

- Explanatory variables are x
- Response variable is y.

"Control" means hold constant

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

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

- 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.*

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.

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.

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 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."

Three ways to think about a regression model for observational data

$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$

- Literally a model for how y is *produced* from the x values. In this case it's a causal model.
- A convenient way to say that y might be related to the x values, by specifying a rough model of the conditional distribution of y given $x_1, \ldots, x_{p-1}$.
- Pure prediction. In this case all the correlation-causation business is irrelevant, but it's not Science.

More than Two Categories

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	$\mu_3 = \beta_0$

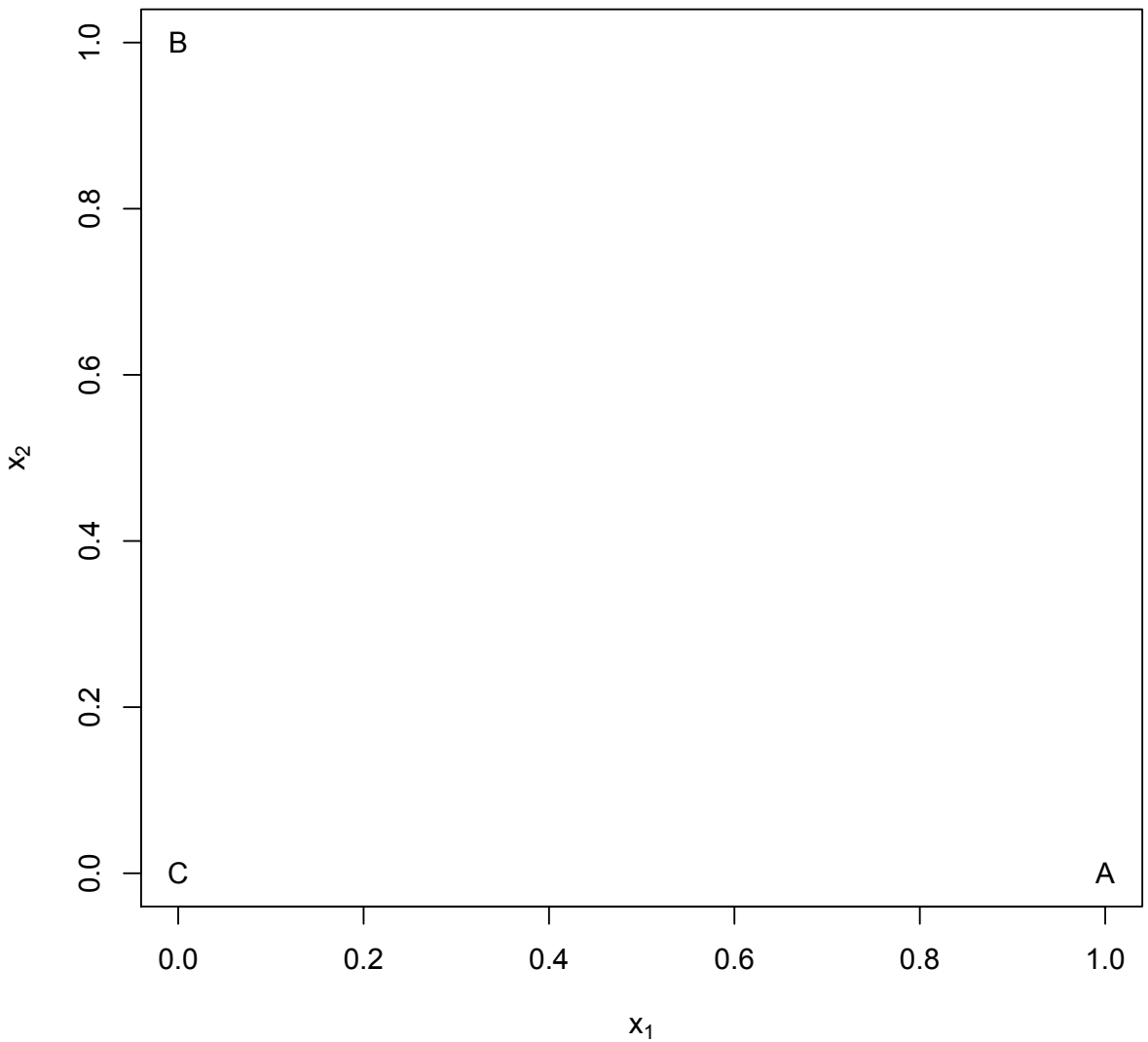
Regression coefficients are contrasts with the category that has no indicator – the *reference category*.

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	$\mu_3 = \beta_0 + \beta_3 x_3$

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.

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.

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

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