STA441: Spring 2016

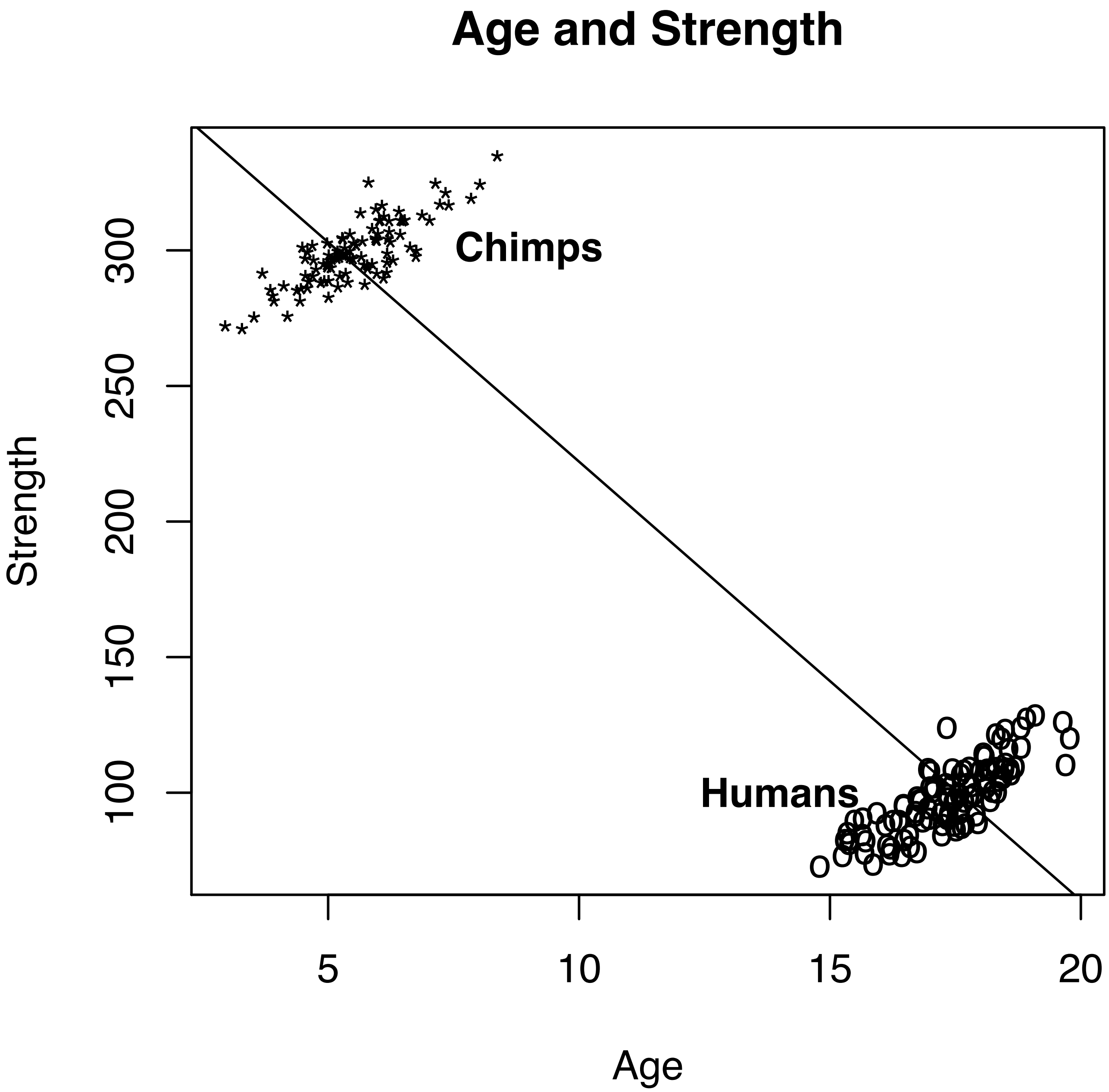
Multiple Regression

More than one explanatory
variable at the same time

This slide show is a free open source document.
See the last slide for copyright information.

One Explanatory Variable at a Time Can Produce Misleading Results

- The standard elementary tests all have a single explanatory variable, so they should be used with caution in practice.
- Example: Artificial and extreme, to make a point
- Suppose the correlation between Age and Strength is $r = -0.96$


Age and Strength
Chimps
Humans
Strength
300
250
200
150
100
Age
5
10
15
20

Comments

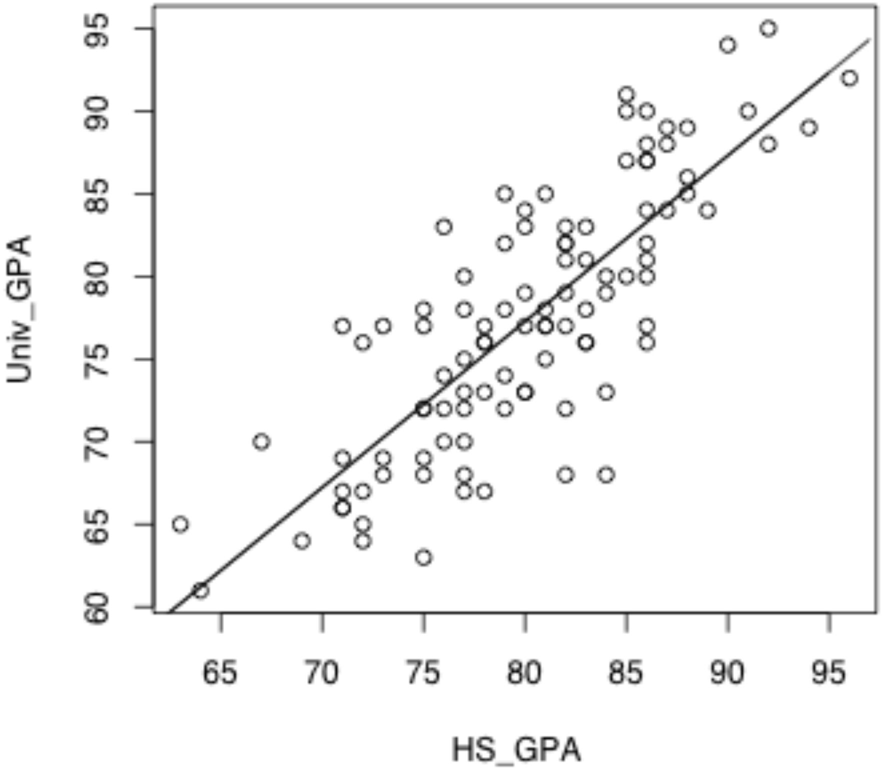
- This is an example of *Simpson's paradox*: The overall relationship between variables is clear, but it is reversed when examined separately for the values of another variable.
- Recall the Berkeley data (lecture and Chapter 4).
- Can be hard to see when there are lots of variables.
- In the example, species is a *confounding variable* (2 criteria).
- Need a systematic way to allow (control) for potential confounding variables by including them in the analysis.

Three Meanings of Control

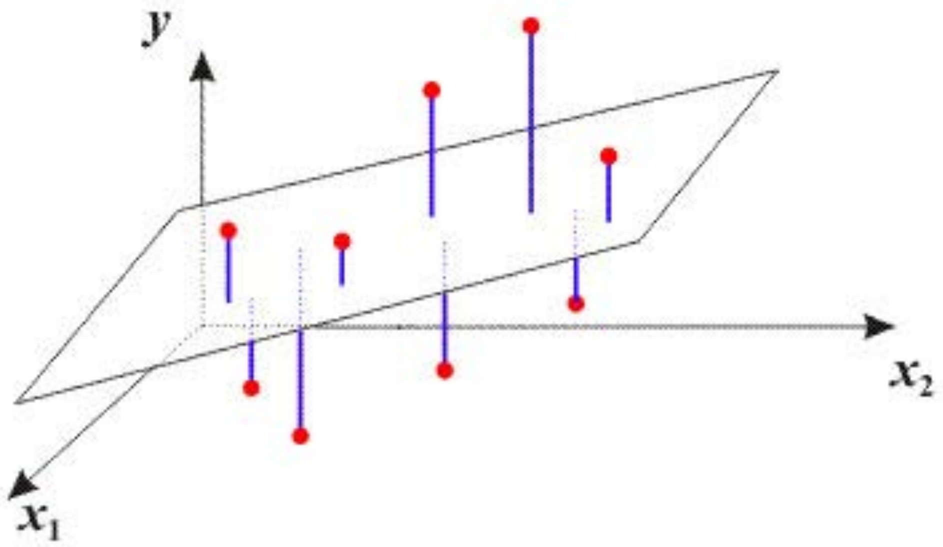
- Experimental
- Sub-division
- Model-based

Multiple regression is the prime example of model-based Control.

Least Squares Line



Least Squares Plane



$$\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2$$

What is b_2?

- $\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2$
- Hold x_1 constant at some fixed level
- What is predicted Y, as a function of x_2?
- $\widehat{Y} = (b_0 + b_1 x_1) + b_2 x_2$
- b_1x_1 is now part of the intercept,
- And b_2 is the slope.

$$\widehat{Y} = (b_0 + b_1 x_1) + b_2 x_2$$

- b_2 is the slope
- It’ s the rate at which predicted Y changes as a function of x_2, with x_1 held constant.
- Say “controlling” for x_1.

$$\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4$$

Control for x_1, x_3 and x_4

$$\widehat{Y} = (b_0 + b_1 x_1 + b_3 x_3 + b_4 x_4) + b_2 x_2$$

Significance tests for the Regression Coefficients

- Test for b_k tells you whether, x_k makes a meaningful contribution to predicting Y, controlling for the other explanatory variables
- “Allowing for”
- “Holding constant”

Statistical **MODEL**

- 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
- The conditional population mean of Y depends on the X values, as follows:

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

Conditional Distributions are normal

- Same variance, and population mean $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$
- This means the *only* way Y can be related to any x is through the β values.
- Because the (conditional) expected value has a simple structure, it is possible to draw conclusions about the conditional distribution of Y, holding the explanatory variables constant at sets of x values where *there are no data*!

High School Calculus and University Calculus

$$\widehat{Y} = -84.85 + 1.79x$$

- With the sub-division approach, you need a lot of data at a particular value to give a good estimate of the conditional population mean.
- Here, we can easily give a good estimate of university calculus mark for a HS Calculus mark of 59, (estimate is 20.76) even though there was just one person with a 59 in the data and he dropped the course.
- We can do this because of the *assumption* (model)

  $E(Y|x) = \beta_0 + \beta_1 x$.
- The more data you have, the less you need to assume.

Statistics b estimate parameters β

$$
\begin{aligned}
E[Y|\boldsymbol{X} = \boldsymbol{x}] &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 \\
\widehat{Y} &= b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3 + b_4 x_4
\end{aligned}
$$

Think of “control” in terms of conditional distributions

- For every combination of control variable values, there is a joint distribution of the explanatory variable and response variable, and a possible relationship between them.
- $H_{0:}$ There is no relationship between x_k and Y for *any* combination of control variable values.

$$H_0 : \beta_k = 0$$

- This is the test of b_k

Categorical explanatory variables

- 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, population mean response is $E[Y|X=1] = \beta_0 + \beta_1$
- For patients getting the placebo, mean response is $E[Y|X=0] = \beta_0$

Sample regression coefficients for a binary explanatory variable

- X=1 means Drug, X=0 means Placebo

- Predicted response is $\widehat{Y} = b_0 + b_1 x$

- For patients getting the drug, predicted response is

$$\widehat{Y} = b_0 + b_1 = \overline{Y}_1$$

- For patients getting the placebo, predicted response is

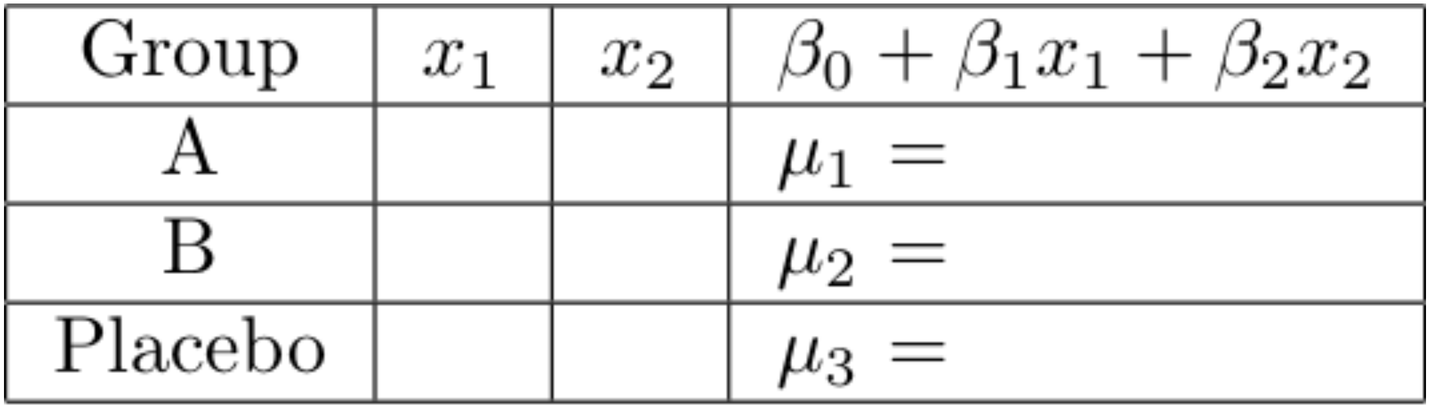
$$\widehat{Y} = b_0 = \overline{Y}_0$$

Regression test of b_1

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

Drug A, Drug B, Placebo

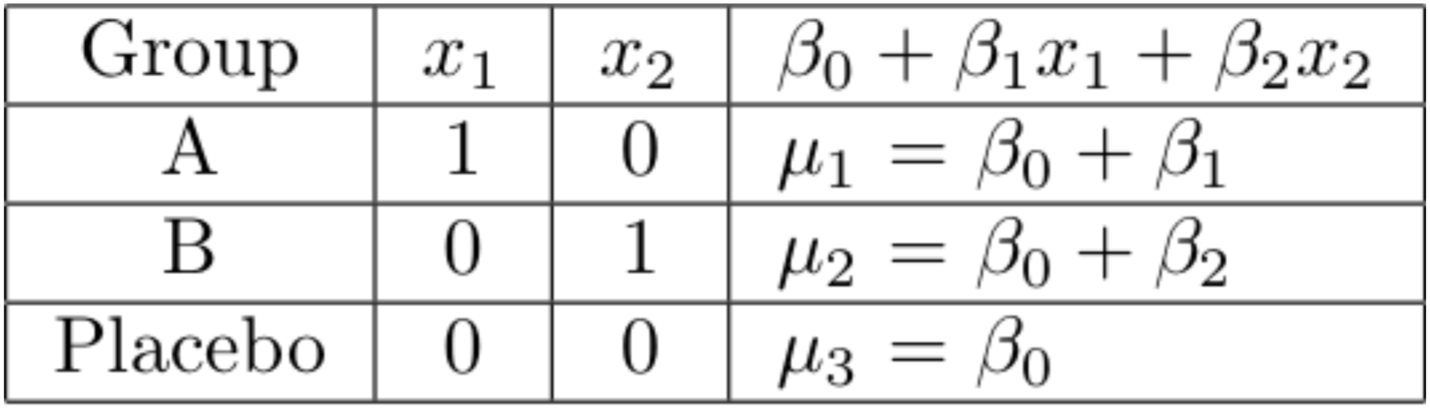
- x_1 = 1 if Drug A, Zero otherwise
- x_2 = 1 if Drug B, Zero otherwise
- $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
- Fill in the table



Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
A			$\mu_1 =$
B			$\mu_2 =$
Placebo			$\mu_3 =$

Drug A, Drug B, Placebo

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



Group	x_1	x_2	$\beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	0	0	$\mu_3 = \beta_0$

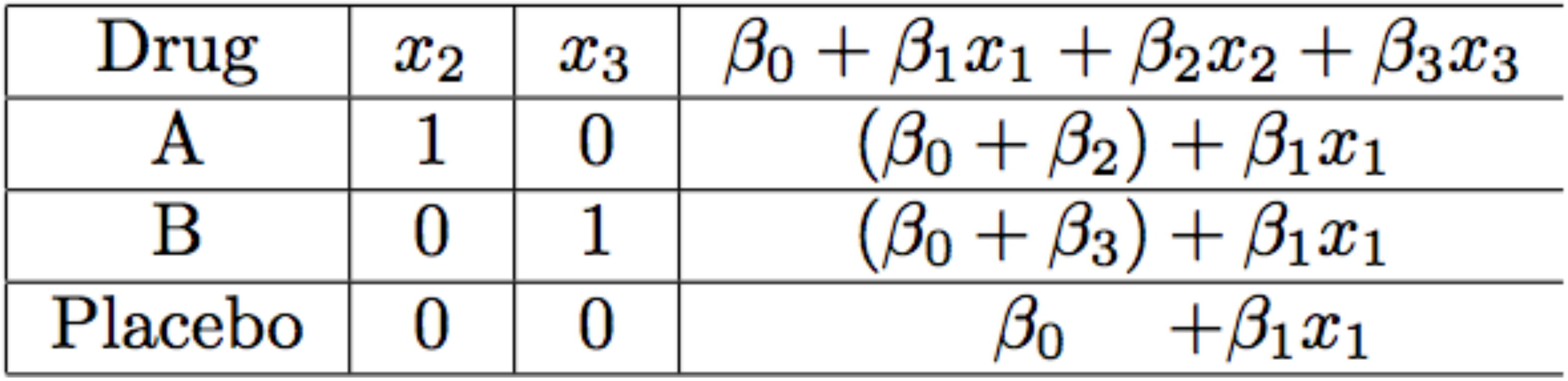
Regression coefficients are CONTRASTS with the category that has no indicator - The REFERENCE category

Indicator dummy variable coding with intercept

- Need p-1 indicators to represent a categorical explanatory variable with p categories
- If you use p dummy variables, trouble
- Regression coefficients are **contrasts** with the category that has no indicator
- Call this the **reference category**

Now add a quantitative variable (covariate)

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

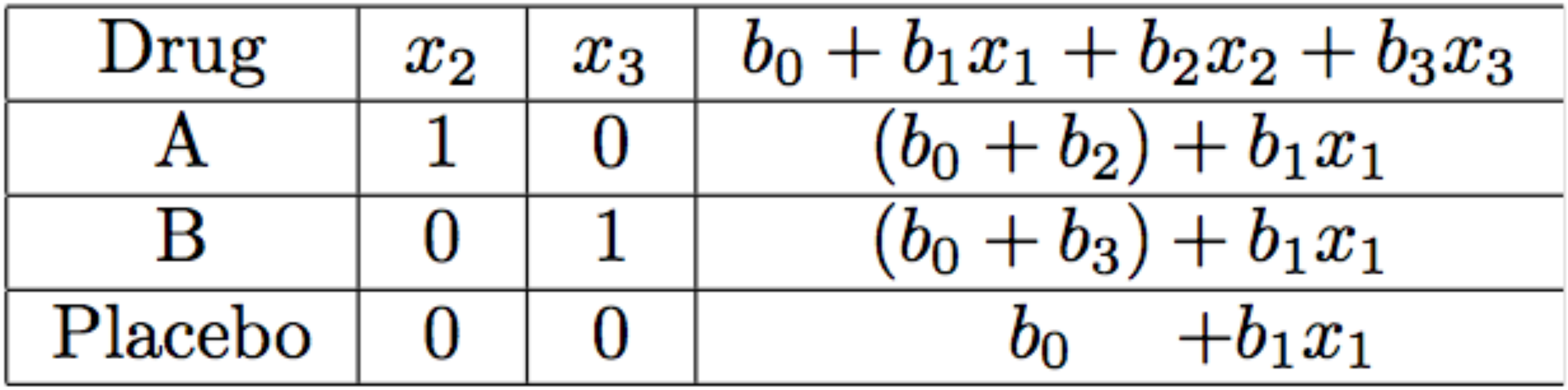


Drug	x_2	x_3	$\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$(\beta_0 + \beta_2) + \beta_1 x_1$
B	0	1	$(\beta_0 + \beta_3) + \beta_1 x_1$
Placebo	0	0	$\beta_0 \quad +\beta_1 x_1$

Parallel regression lines (equal slopes): ANCOVA

What do you report?

- x_1 = Age
- x_2 = 1 if Drug A, Zero otherwise
- x_3 = 1 if Drug B, Zero otherwise
- $\widehat{Y} = b_0 + b_1x_1 + b_2x_2 + b_3x_3$



Drug	x_2	x_3	$b_0 + b_1x_1 + b_2x_2 + b_3x_3$
A	1	0	$(b_0 + b_2) + b_1x_1$
B	0	1	$(b_0 + b_3) + b_1x_1$
Placebo	0	0	$b_0 \quad +b_1x_1$

Set all covariates to their sample mean values

- And compute Y-hat for each group
- Call it an “adjusted” mean, or something like “average university GPA adjusted for High School GPA.”
- SAS calls it a **least squares mean** (`lsmeans`)

Analysis of Variance

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

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

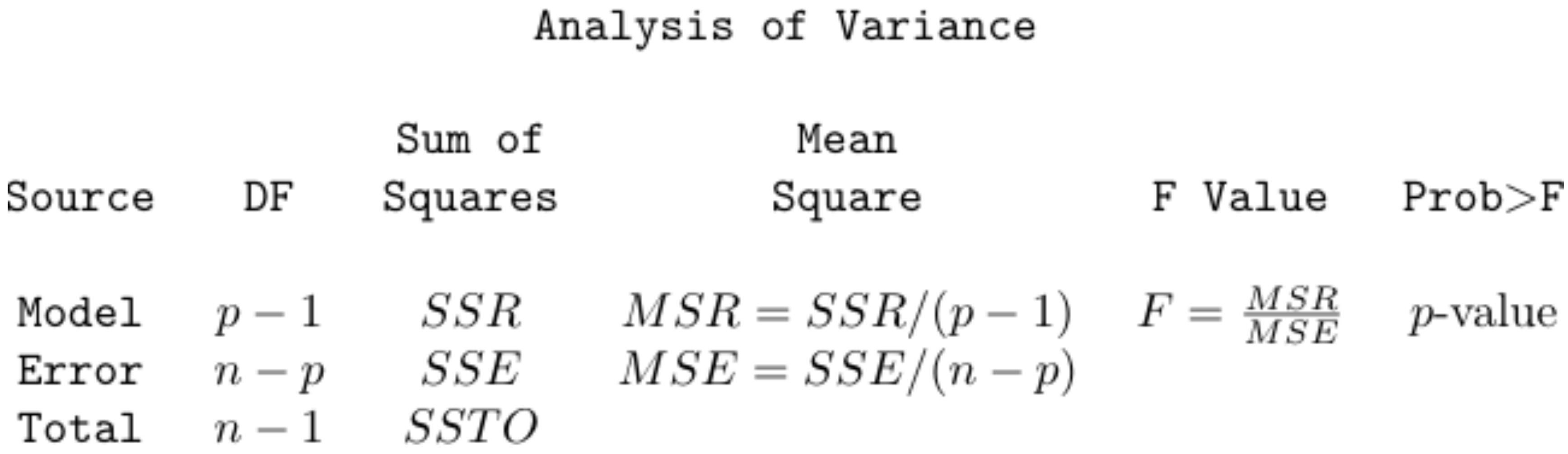
- Variation that is still unexplained: **Error Sum of Squares**

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

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

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

ANOVA Summary Table



Analysis of Variance

		Sum of	Mean		
Source	DF	Squares	Square	F Value	Prob>F
Model	$p-1$	SSR	$MSR = SSR/(p-1)$	$F = \frac{MSR}{MSE}$	p-value
Error	$n-p$	SSE	$MSE = SSE/(n-p)$		
Total	$n-1$	$SSTO$			

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

Proportion of variation in the response variable that is explained by the explanatory variables

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

Significance Testing

- Overall F test for all the explanatory variables at once,
- T-tests for each regression coefficient: Controlling for all the others, does that explanatory variable matter?
- Test a collection of explanatory variables controlling for another collection,
- Most general: Testing whether sets of linear combinations of regression coefficients differ from specified constants.

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?

$$E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \cdots + \beta_5 x_5$$

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0$$

Full vs. Reduced Model

- 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 **Reduced Model**

$$R_F^2 \geq R_R^2$$

When you add explanatory variables, R^2 can only go up

- By how much? Basis of F test.
- Same as testing H_0: All betas in set B (there are *s* of them) equal zero

$$F = \frac{(SSR_F - SSR_R)/s}{MSE_F}$$

F test is based not just on change in R^2, but upon

$$a = \frac{R_F^2 - R_R^2}{1 - R_R^2}$$

Increase in explained variation expressed as a fraction of the variation that the reduced model does *not* explain.

$$F = \left(\frac{n-p}{s}\right)\left(\frac{a}{1-a}\right)$$

- For any given sample size, the bigger *a* is, the bigger *F* becomes.
- For any a ≠0, *F* increases as a function of *n.*
- So you can get a large *F* from strong results and a small sample, or from weak results and a large sample.

Can express *a* in terms of *F*

$$a = \frac{sF}{n - p + sF}$$

- Often, scientific journals just report *F*, numerator df = *s*, denominator df = *(n-p),* and a *p*-value.
- You can tell if it's significant, but how strong are the results? Now you can calculate it.
- This formula is less subject to rounding error than the one in terms of R-squared values

When you add explanatory variables to a model

- Statistical significance can appear when it was not present originally
- Statistical significance that was originally present can disappear
- Even the signs of the b coefficients can change, reversing the interpretation of how their variables are related to the response variable.
- This is consistent with what happened in the age and strength example.

Watch out for measurement error in the explanatory variables

- When we test for a relationship “controlling” for some set of variables, we are seeking it in the conditional distributions - conditional on the values of the variables for which we are controlling.
- If the control variables are measured with error, the conditional distributions given the observed variables need not be the same as the conditional distributions given the true variables .

Suppose you are testing the relationship of age to BMI, controlling for exercise and calorie intake.

- Questionnaire measures are known to be inaccurate. People mis-report, and *not* by a constant amount.
- And, age is related to both explanatory variables, especially exercise
- Can't see the control variables clearly to hold them constant
- So even if age is unrelated to BMI for every combination of *true* exercise and *true* calorie intake, a relationship can exist conditionally upon *observed* exercise and *observed* calorie intake.

Want to test B controlling for A: The poison combination

- A is related to the response variable
- A and B are related to each other, and
- A is measured with error

- Estimation of B's relationship with Y is biased
- Type I error is badly inflated (Brunner and Austin, 2009)
- Large sample size makes it *worse*!
- For observational studies, all three conditions usually are present.

Especially a problem in observational medical research

- Seek to assess potential risk factors, controlling for known risk factors
- The known risk factors do matter
- Known and potential risk factors are correlated
- Known risk factors are difficult to measure without error
- Experimental research is essential to confirm findings - and it often does not.

But all is not lost

- As long as you are interested in prediction rather than interpretation, there is no problem. Test for whether age is a useful predictor is still valid, even if its usefulness comes from its correlation with true exercise.
- The problem comes from trying to use regression as a *causal* model for observational data.
- If one or more categorical explanatory variables are experimentally manipulated, analysis of covariance can help reduce MSE and makes the analysis more precise, even if the covariates (control variables) are measured with error.
- No inflation of Type I error rate for ANCOVA - because random assignment breaks up the association between A and B.

If it’s an observational study, just ask

- How did you control for ____?
- How did you take measurement error into account? (There are ways, but if it were easy people would do it more often. Nature of data collection is involved, not just statistical analysis.)

- If they say “Oh, there was just a little measurement error,” observe that if the sample is large enough, no amount of measurement error is safe. Brunner and Austin (2009) give a proof.
- If they say “Well, it’s the best we could do,” you could ask whether it’s better to say something incorrect, or to be silent.

In this course

- We will carry out classical regression analysis on observational data *only* when our primary purpose is prediction.
- We will be very careful about the way we describe the results.
- We will use regression methods extensively on experimental data.

Copyright Information