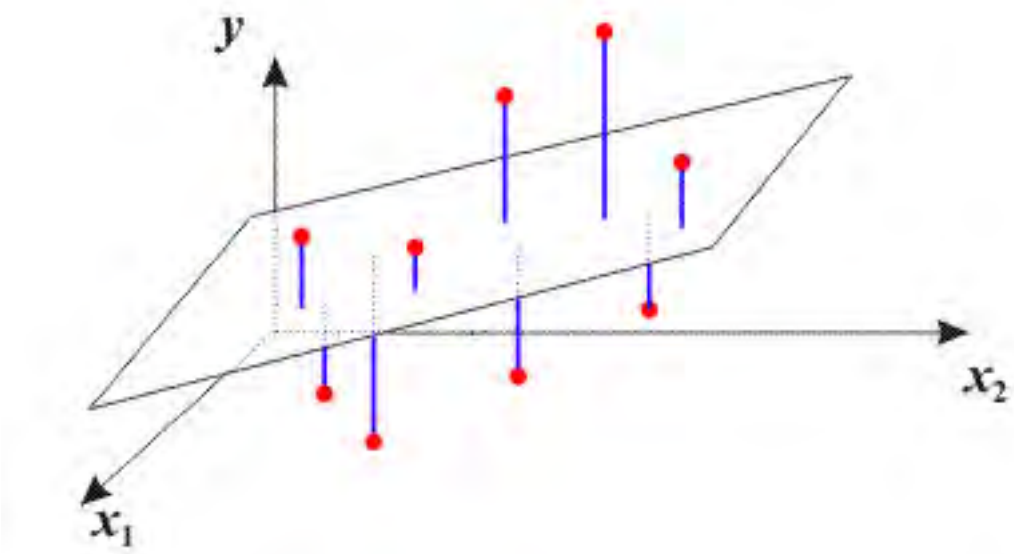
STA441: Spring 2018

Multiple Regression

Least Squares Plane



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

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.

Correlation and causation

- Model says that *y* values have the same variance, and population mean that depends on *x* in a particular way:

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

- Nobody said how it got to be this way.
- It could be because of unobserved variables that are not in the model.
- It could be direct or indirect influence of *x* on *y*.
- It could be influence of *y* on *x*.

- **The regression model is a model of <u>relationship</u>.**

Traditional way to write the model

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1} x_{p-1} + \epsilon$$

- Implies that for any combination of x values,

  $E(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$

  with constant variance.
- Can be viewed as a technically convenient way to produce the model of relationship.
- Or you can take it literally as a statement of how x produces y.
- If you take it literally, you must live with the correlation-causation issue.

Summary

- The regression model is a model of relationship.
- Null hypothesis will usually say there is no relationship.
- If you reject the null hypothesis, you conclude there is a relationship.
- How you talk about it depends on the design of the study.

Back to the model

- Same variance, and population mean $E[Y|\boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \ldots + \beta_{p-1} x_{p-1}$

- 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*!

Statistics b estimate parameters β

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

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

Three Meanings of Control

- Procedural
- Sub-division
- Model-based

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

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?
- $\widehat{Y} = (b_0 + b_1 x_1) + b_2 x_2$
- $b_1 x_1$ 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"
- "Correcting for"

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.

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 = 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	$\beta_0 \quad + \beta_1 x_1$

What do you report?

- x_1 = Age
- x_2 = 1 if Drug A, Zero otherwise
- x_3 = 1 if Drug B, Zero otherwise
- $\widehat{Y} = b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$

Drug	x_2	x_3	$b_0 + b_1 x_1 + b_2 x_2 + b_3 x_3$
A	1	0	$(b_0 + b_2) + b_1 x_1$
B	0	1	$(b_0 + b_3) + b_1 x_1$
Placebo	0	0	$b_0 \quad + b_1 x_1$

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

Source	DF	Sum of Squares	Mean 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

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

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.

Interpretation

- Null hypotheses always have = signs. There are no one-sided t-tests or z-tests in this course.
- Draw directional conclusions for
  - t-tests
  - z-tests
  - F-test with numerator df=1
  - Chi-squared tests with df=1
- Avoid causal conclusions from observational data.

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. Restricted 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 **Restricted 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

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

Looking at the formula

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

- Numerator is *average* improvement in explained SS.
- Anything that reduces MSE increases 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

More about Dummy Variables

- Indicator dummy variables with intercept
- Indicator dummy variables without intercept (Cell means coding)
- Effect coding

Recall indicators with intercept

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

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	$\beta_0 \quad + \beta_1 x_1$

Can test contrasts *controlling* for covariates

- Valuable
- Sometimes very easy, sometimes can require a bit of algebra
- An easy example: Are responses to Drug A and B different, controlling for age?

Are responses to Drug A and B different, controlling for age?

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	$\beta_0 \quad + \beta_1 x_1$

$$H_0 : \beta_2 = \beta_3$$

Test whether the average response to Drug A and Drug B is different from response to the placebo, controlling for age. What is the null hypothesis?

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	$\beta_0 \quad + \beta_1 x_1$

$$H_0 : \beta_2 + \beta_3 = 0$$

Show your work

$$\tfrac{1}{2}\left[\,(\beta_0+\beta_2+\beta_1 x_1)+(\beta_0+\beta_3+\beta_1 x_1)\,\right]=\beta_0+\beta_1 x_1$$

$$\iff \quad \beta_0+\beta_2+\beta_1 x_1+\beta_0+\beta_3+\beta_1 x_1=2\beta_0+2\beta_1 x_1$$

$$\iff \quad 2\beta_0+\beta_2+\beta_3+2\beta_1 x_1=2\beta_0+2\beta_1 x_1$$

$$\iff \quad \beta_2+\beta_3=0$$

We want to avoid this kind of thing

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	$\beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
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 the intercepts.)

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

- Parallel regression lines
- Equivalent to the model with intercept
- Regression coefficients for the dummy vars are the intercepts
- Easy to specify contrasts

Effect coding

- *p-1* dummy variables for *p* categories
- Include an intercept
- Last category gets -1 instead of zero
- What do the regression coefficients mean?

Group	x_1	x_2	$E[Y\|\boldsymbol{X}=\boldsymbol{x}]=\beta_0+\beta_1 x_1+\beta_2 x_2$
A	1	0	$\mu_1=\beta_0+\beta_1$
B	0	1	$\mu_2=\beta_0+\beta_2$
Placebo	-1	-1	$\mu_3=\beta_0-\beta_1-\beta_2$

Meaning of the regression coefficients

Group	x_1	x_2	$E[Y \mid \boldsymbol{X} = \boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2$
A	1	0	$\mu_1 = \beta_0 + \beta_1$
B	0	1	$\mu_2 = \beta_0 + \beta_2$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_1 - \beta_2$

$$\mu = \frac{1}{3}(\mu_1 + \mu_2 + \mu_3) = \beta_0$$

With effect coding

- Intercept is the *Grand Mean*
- Regression coefficients are deviations of group means from the grand mean
- Equal population means is equivalent to zero coefficients for all the dummy variables
- Last category is not a reference category

Group	x_1	x_2	$E[Y\|\boldsymbol{X}=\boldsymbol{x}]=\beta_0+\beta_1x_1+\beta_2x_2$
A	1	0	$\mu_1=\beta_0+\beta_1$
B	0	1	$\mu_2=\beta_0+\beta_2$
Placebo	-1	-1	$\mu_3=\beta_0-\beta_1-\beta_2$

Add a covariate: Age = x_1

Group	x_2	x_3	$E[Y\|\boldsymbol{X}=\boldsymbol{x}] = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$
A	1	0	$\mu_1 = \beta_0 + \beta_2 \quad + \beta_1 x_1$
B	0	1	$\mu_2 = \beta_0 + \beta_3 \quad + \beta_1 x_1$
Placebo	-1	-1	$\mu_3 = \beta_0 - \beta_2 - \beta_3 + \beta_1 x_1$

Regression coefficients are deviations from the average conditional population mean (conditional on x_1).

So if the regression coefficients for all the dummy variables equal zero, the categorical explanatory variable is unrelated to the response variable, controlling for the covariates.

We will see later that effect coding is very useful when

- There is more than one categorical explanatory variable and
- We are interested in *interactions* --- ways in which the relationship of an explanatory variable with the response variable depends on the value of another explanatory variable.

What dummy variable coding scheme should you use?

- Whichever is most convenient, and gives you the information you want most directly
- They are all equivalent, if done correctly.
- Same test statistics, same conclusions

Interactions

- Interaction between explanatory variables means “It depends.”
- Relationship between one explanatory variable and the response variable *depends* on the value of another explanatory variable.
- Note that an interaction is ***not*** a relationship between explanatory variables (in this course).

Interactions between explanatory variables can be

- Quantitative by quantitative
- Quantitative by categorical
- Categorical by categorical

General principle

- Interaction between A and B means
  - Relationship of A to Y depends on value of B
  - Relationship of B to Y depends on value of A
- The two statements are formally equivalent

Quantitative by Quantitative

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \\ E(Y|\mathbf{x}) &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 \end{aligned}$$

For fixed x_2

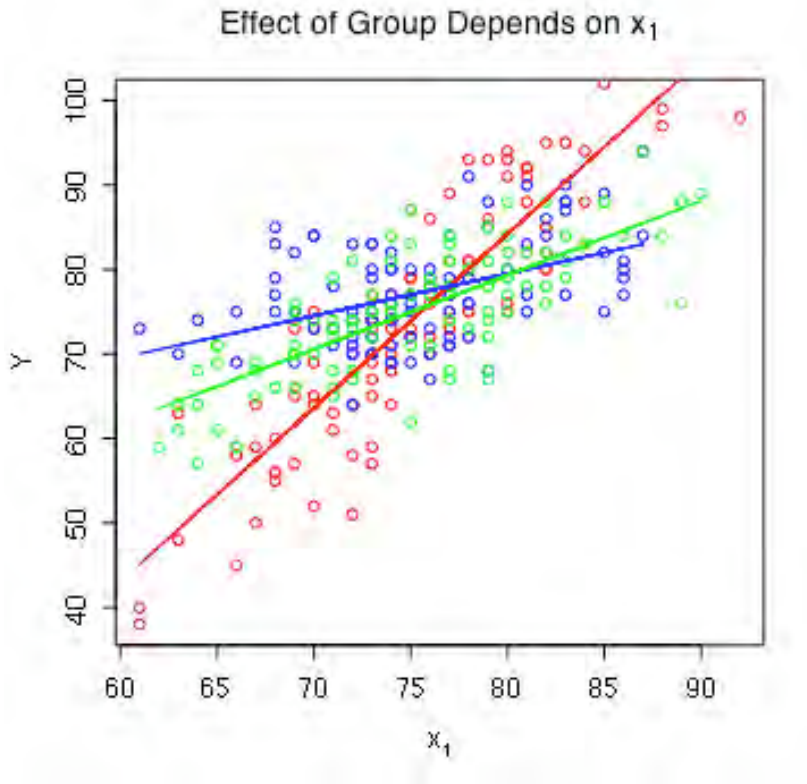
$$E(Y|\mathbf{x}) = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

Both slope and intercept depend on value of x_2

And for fixed x_1, slope and intercept relating x_2 to E(Y) depend on the value of x_1

Quantitative by Categorical

- Separate regression line for each value of the categorical explanatory variable.
- Interaction means slopes of regression lines are not equal.



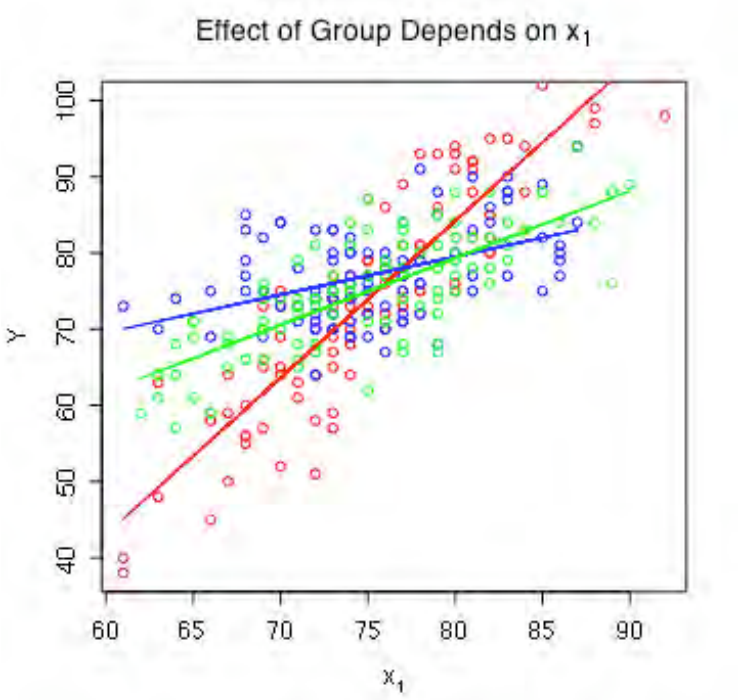
One regression Model

- Form a product of quantitative variable times each dummy variable for the categorical variable
- For example, three treatments and one covariate: x_1 is the covariate and x_2, x_3 are dummy variables

$$\begin{aligned} Y &= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 \\ &\quad + \beta_4 x_1 x_2 + \beta_5 x_1 x_3 + \epsilon \end{aligned}$$

$$E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0 + \beta_2) + (\beta_1 + \beta_4)x_1$
2	0	1	$(\beta_0 + \beta_3) + (\beta_1 + \beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$



Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0+\beta_2)+(\beta_1+\beta_4)x_1$
2	0	1	$(\beta_0+\beta_3)+(\beta_1+\beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

What null hypothesis would you test for

- Equal slopes
- Compare slopes for group one vs three
- Compare slopes for group one vs two
- Equal regressions
- Interaction between group and x_1

Equal regressions = Conditional independence

$$E(Y|\mathbf{x}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$$

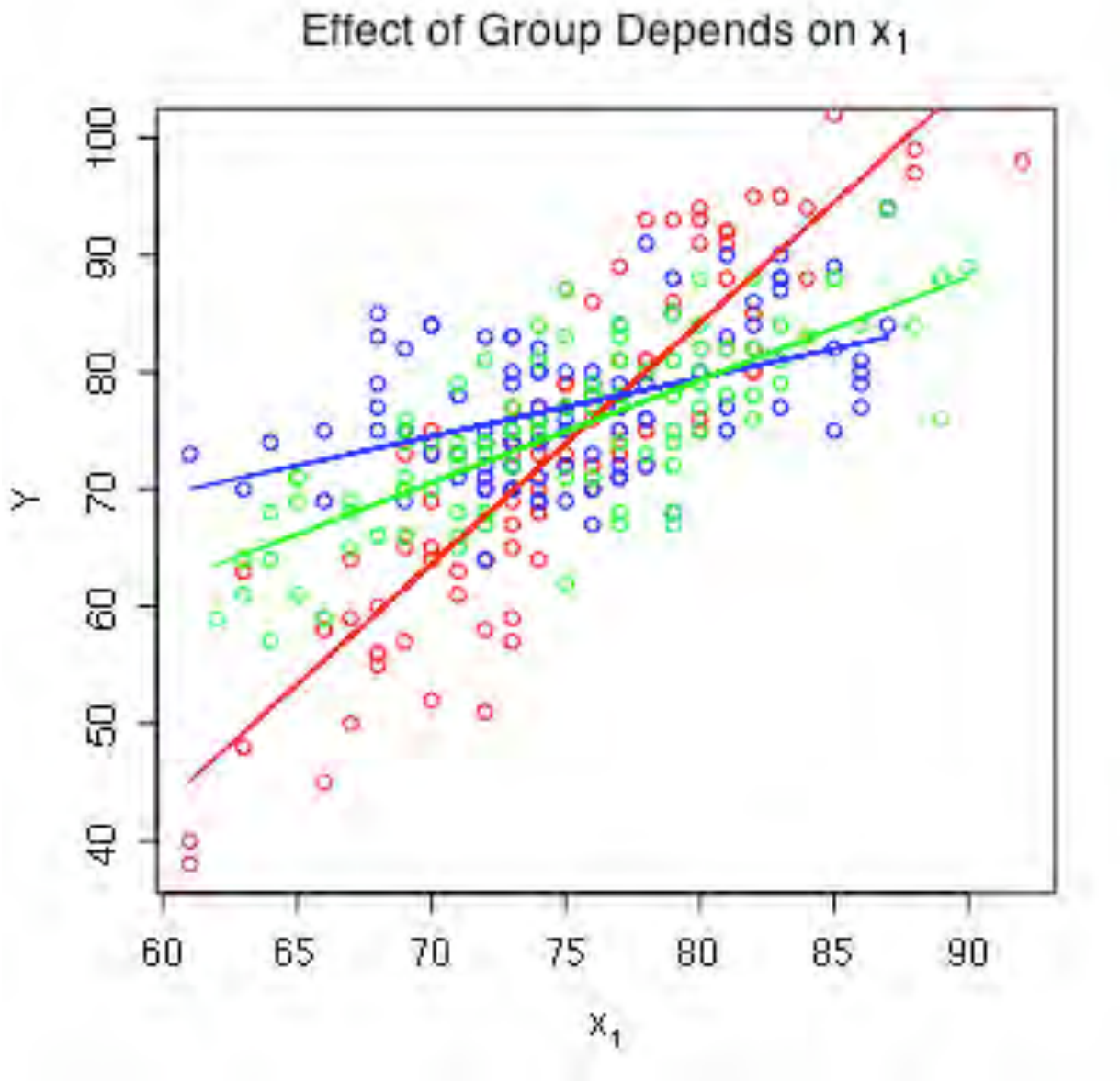
- H_0: $\beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$
- For any fixed value of x_1, E(y) is the same for all three groups.
- And the variances are always the same.
- So the distributions are the same.
- Conditionally on x_1, Y is independent of Group.
- Allowing for x_1, Group does not matter **at all**.

- It's not a very standard null hypothesis, but it's meaningful.

What to do if H_0: $\beta_4=\beta_5=0$ is rejected

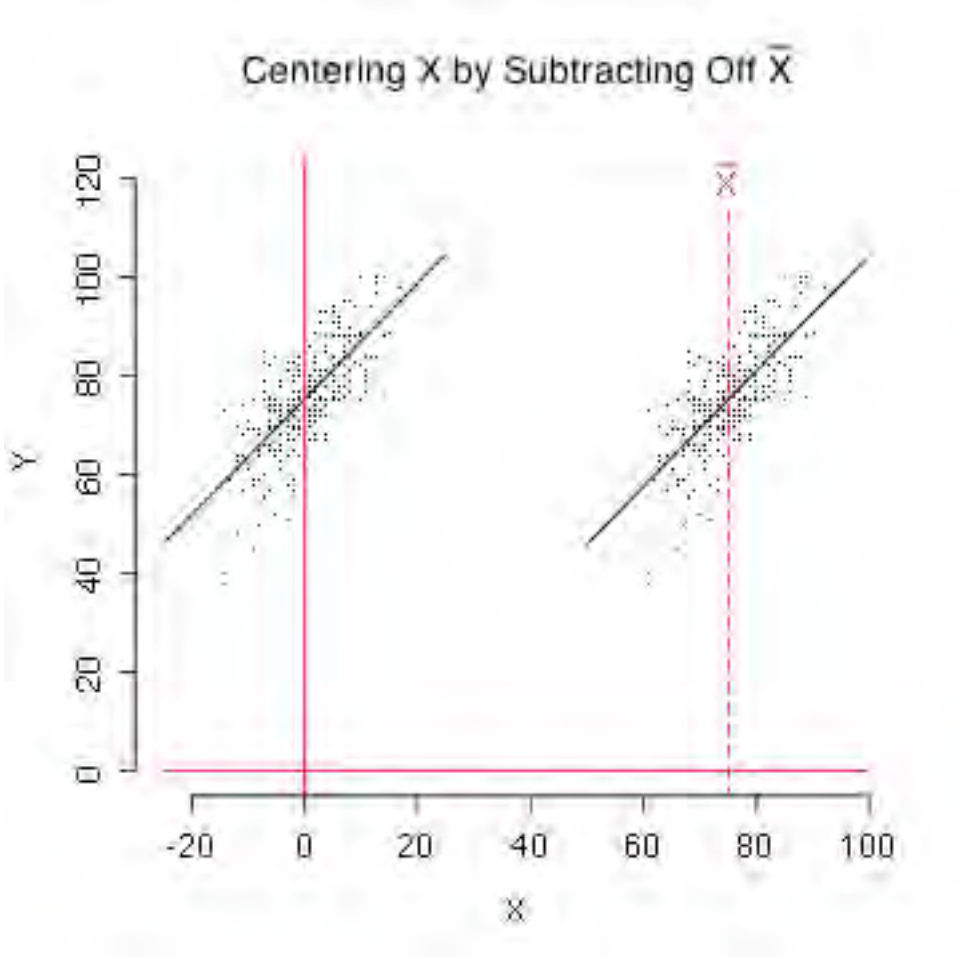
- How do you test Group "controlling" for x_1?
- A popular choice is to set x_1 to its sample mean, and compare treatments at that point. SAS calls the estimates (Y-hat values) "Least Squares Means."
- Or, test equal regressions, in which mean response is the same for all values of the covariate(s).

Test for differences at mean of x_1?



"Centering" the explanatory variables

- Subtract mean (for entire sample) from each quantitative explanatory variable.



Properties of Centering

- When explanatory variables are centered, estimates and tests for *intercepts* are affected.
- Relationships between explanatory variables and response variables are **unaffected**.
- Estimates and tests for slopes are **unaffected**.
- R^2 is **unaffected**.
- Predictions and prediction intervals are **unaffected**.

More Properties

- Suppose a regression model has an intercept.
- Then the residuals add up to zero. But there are models *without* intercepts where the sum of residuals *is* zero. These are often equivalent to models with intercepts.
- Suppose the residuals do add to zero. Then if each explanatory variable is set to its sample mean value, Y-hat equals Y-bar, the sample mean of all the Y values.
- In this case, if *all* explanatory variables are centered by subtracting off their means, then the intercept equals Y-bar, exactly.

Comments

- Often, X=0 is outside the range of explanatory variable values, and it is hard to say what the intercept *means* in terms of the data.
- When explanatory variables are centered, the intercept is the average Y value for average value(s) of X.
- If there are both quantitative variables and categorical variables (represented by dummy variables), it can help to center just the quantitative variables.

"Centering" just the quantitative explanatory variables

- Subtract mean (for entire sample) from each quantitative explanatory variable.
- Then, comparing intercepts is the same as comparing expected values for "average" X values. It's more convenient than testing linear combinations.

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0+\beta_2)+(\beta_1+\beta_4)x_1$
2	0	1	$(\beta_0+\beta_3)+(\beta_1+\beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

For Example

Group	x_2	x_3	$E(Y\|\mathbf{x})$
1	1	0	$(\beta_0+\beta_2)+(\beta_1+\beta_4)x_1$
2	0	1	$(\beta_0+\beta_3)+(\beta_1+\beta_5)x_1$
3	0	0	$\beta_0 + \beta_1 x_1$

- Suppose you want to test for differences among population mean Y values when x_1 equals its sample mean value.
- You could test H_0: $\beta_2+\beta_4\overline{x}_1=\beta_3+\beta_5\overline{x}_1=0$
- Or, center x_1 and test H_0: $\beta_2=\beta_3=0$

Categorical by Categorical

- Soon
- But first, an example.

Copyright Information