Regression Part 2

- Residuals
- More about dummy variables
- Interactions

$$e_i = Y_i - \widehat{Y}_i$$

Analysis of Residuals

Data = Fit + Residual

$$Y_i = b_0 + b_1 X_{i,1} + \ldots + b_{p-1} X_{i,p-1} + e_i$$

Mean residual equals zero
(usually)

- Suppose a regression model has an intercept.
- Then the residuals add up to zero. Having an intercept in the model is a sufficient but not a necessary condition for the sum of residuals to be zero.
- That is, there are some models without intercepts for which the residuals still add to zero.
- Often these are equivalent to models with intercepts.

Residual means left over

- Vertical distance of Y_i from the regression hyper-plane
- An error of “prediction”
- Big residuals merit further investigation
- Big compared to what?
- They are normally distributed
- Consider standardizing
- Maybe detect outliers

Residuals are like estimated error terms

$$Y_i = \beta_0 + \beta_1 X_{i,1} + \ldots + \beta_{p-1} X_{i,p-1} + \varepsilon_i$$

$$Y_i = b_0 + b_1 X_{i,1} + \ldots + b_{p-1} X_{i,p-1} + e_i$$

Normal distribution of ε_i implies normal distribution of e_i

Standardized Residuals

- Could divide by square root of sample variance of $e_1, \ldots, e_n$
- "Semi-Studentized" (Kutner et al.)

$$e_i^* = \frac{e_i}{\sqrt{MSE}}$$

- Studentized: Estimate $Var(e_i)$ (not all the same) and divide by square root of that

Studentized *deleted* residuals

- An outlier will make MSE big
- In that case, the Studentized residual will be too small – less noticeable
- So calculate Y-hat for each observation based on all the other observations, but not that one
- Basically, predict each observed Y based on all the others, and assess error of prediction (divide by standard error).

Deleted residual

$$
\begin{aligned}
d_i &= Y_i - \widehat{Y}_{i(i)} \\
s^2\{d_i\} &= \ldots
\end{aligned}
$$

Studentized deleted residual is
$t_i = \frac{d_i}{s\{d_i\}} \sim t(n - p - 1)$

Is it too big? Use a t-test.

Prediction interval

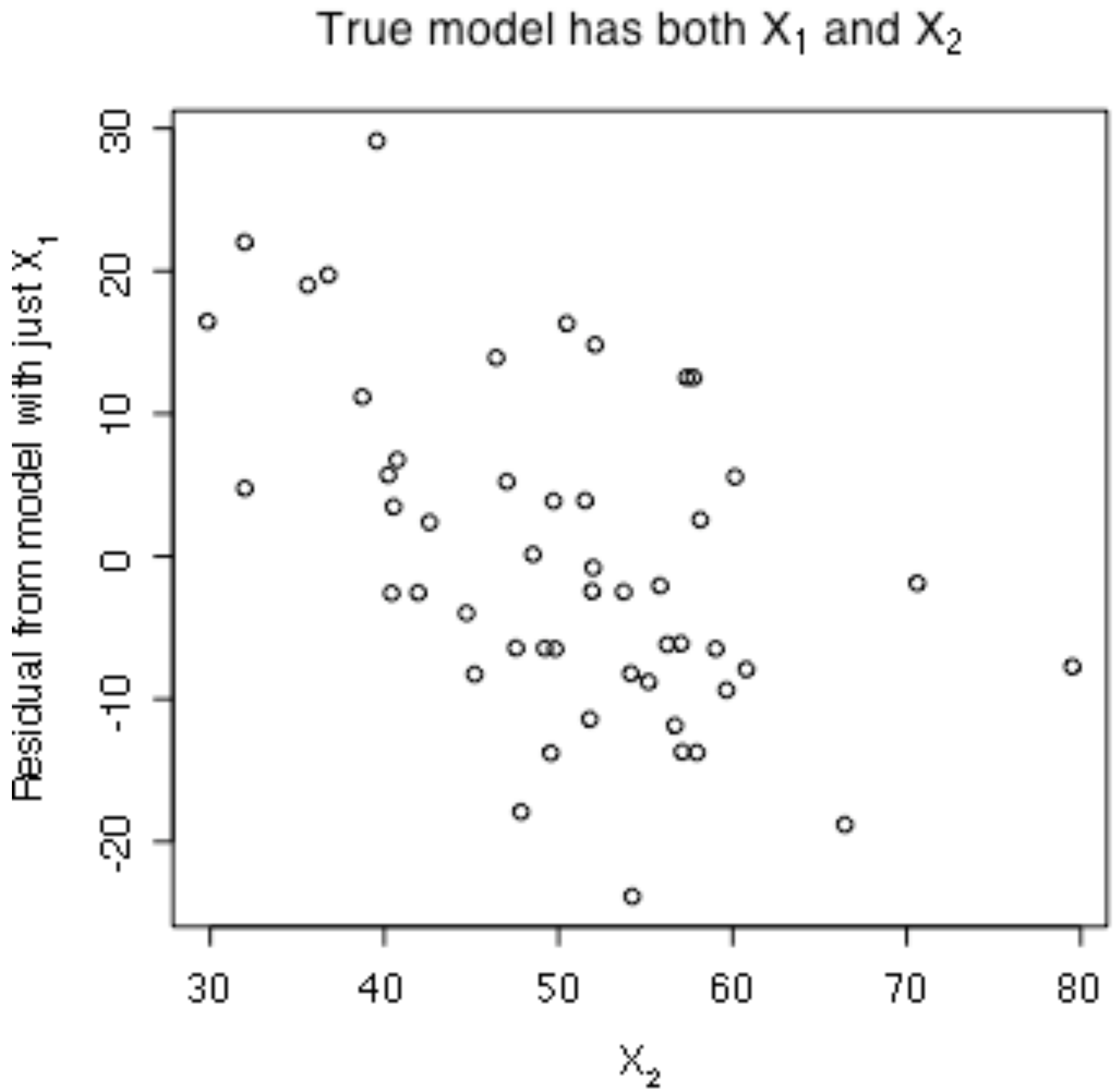
- Apply the same technology
- Think of Studentized deleted residual for case n+1
- So $$t_{n+1} = \frac{d_{n+1}}{s\{d_{n+1}\}} \sim t(n-p)$$

$$\begin{aligned}
1-\alpha &= Pr\left\{-t_{\alpha/2}(n-p) < \frac{Y_{n+1}-\widehat{Y}_{n+1}}{s\{d_{n+1}\}} < t_{\alpha/2}(n-p)\right\} \\
&= Pr\left\{-t_{\alpha/2}\, s\{d_{n+1}\} < Y_{n+1}-\widehat{Y}_{n+1} < t_{\alpha/2}\, s\{d_{n+1}\}\right\} \\
&= Pr\left\{\widehat{Y}_{n+1} - t_{\alpha/2}\, s\{d_{n+1}\} < Y_{n+1} < \widehat{Y}_{n+1} + t_{\alpha/2}\, s\{d_{n+1}\}\right\}
\end{aligned}$$

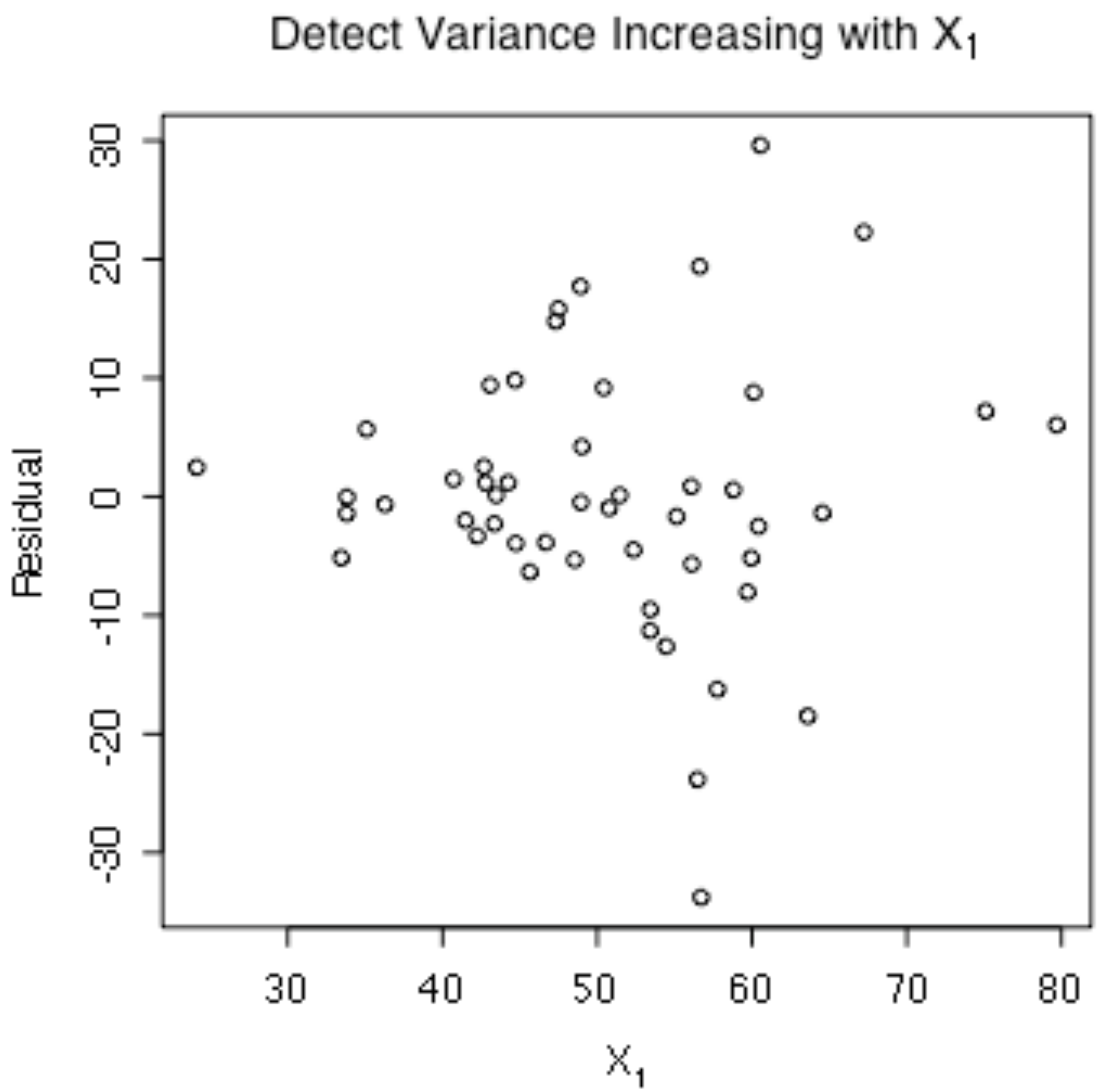
Plotting residuals

- Against independent variables not in the equation
- Against independent variables in the equation
- Test for approximate normality

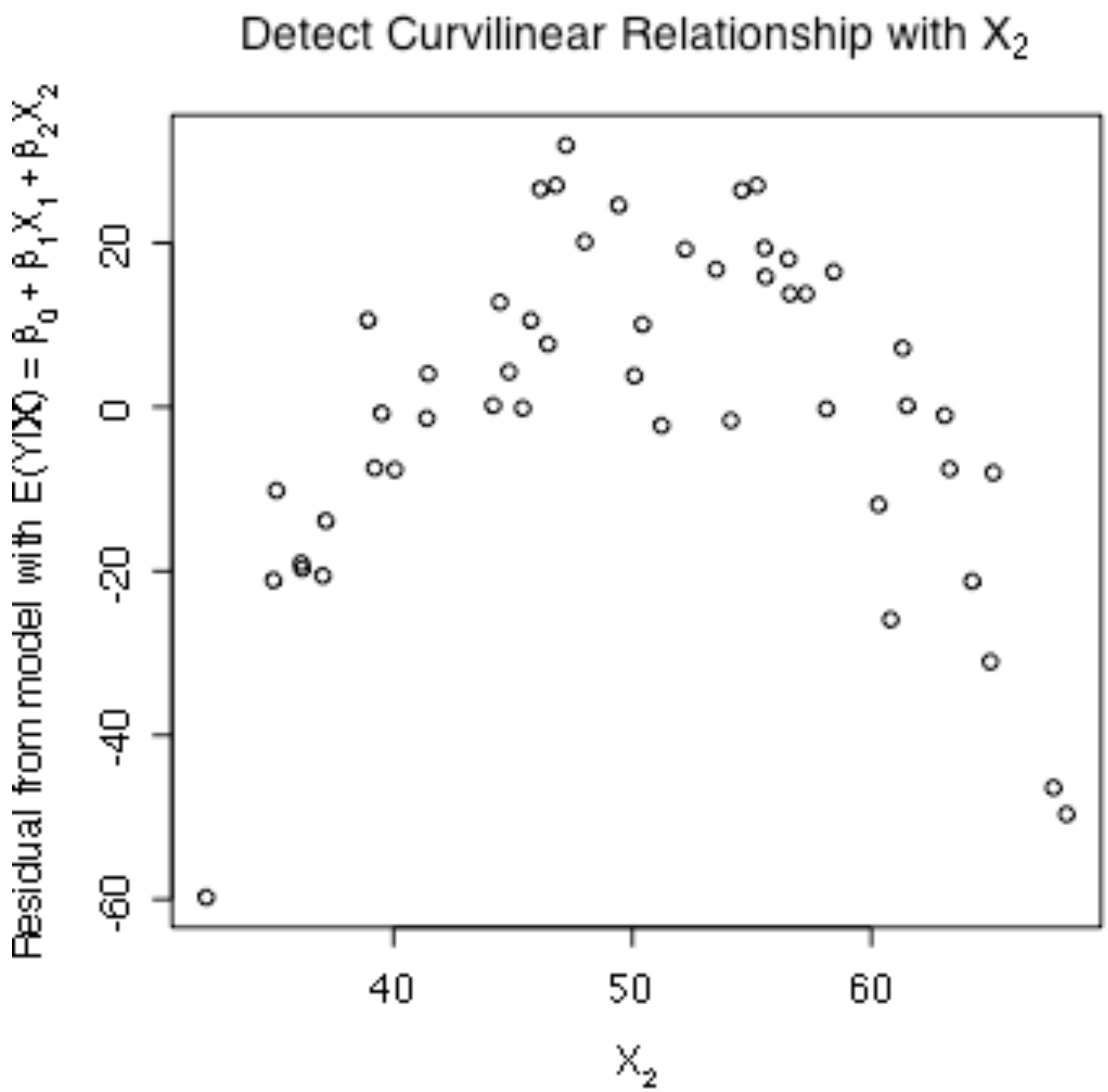
Plot Residuals Against Independent Variables Not in the Equation



Plot Residuals Against Independent Variables in the Equation: $E(Y|\mathbf{X})=\beta_0+\beta_1X_1+\beta_2X_2$



Plot Residuals Against Independent Variables in the Equation



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

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

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

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

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

We want to avoid this kind of thing

A common error

- Categorical IV with *p* categories
- *p* dummy variables (rather than *p-1*)
- And an intercept

- There are *p* population means represented by *p+1* regression coefficients – representation is not unique

But suppose you leave off the intercept

- Now there are *p* regression coefficients and *p* population means
- The correspondence is unique, and the model can be handy -- less algebra
- Called **cell means coding**

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

Sometimes speak of the “main effect” of a categorical variable

- More than one categorical IV (factor)
- Marginal means are average group mean, averaging across the other factors
- This is loose speech: There are actually *p* main effects for a variable, not one
- Blends the “effect” of an experimental variable with the technical statistical meaning of effect.
- It’s harmless

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	$\mu_1 = \beta_0 + \beta_2 \qquad + \beta_1 x_1$
B	0	1	$\mu_2 = \beta_0 + \beta_3 \qquad + \beta_1 x_1$
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 IV is unrelated to the DV, controlling for the covariates.

We will see later that effect coding is very useful when there is more than one categorical independent variable and we are interested in *interactions* --- ways in which the relationship of an independent variable with the dependent variable depends on the value of another independent 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 independent variables means "It depends."
- Relationship between one IV and the DV *depends* on the value of another IV.
- Can have
  - 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

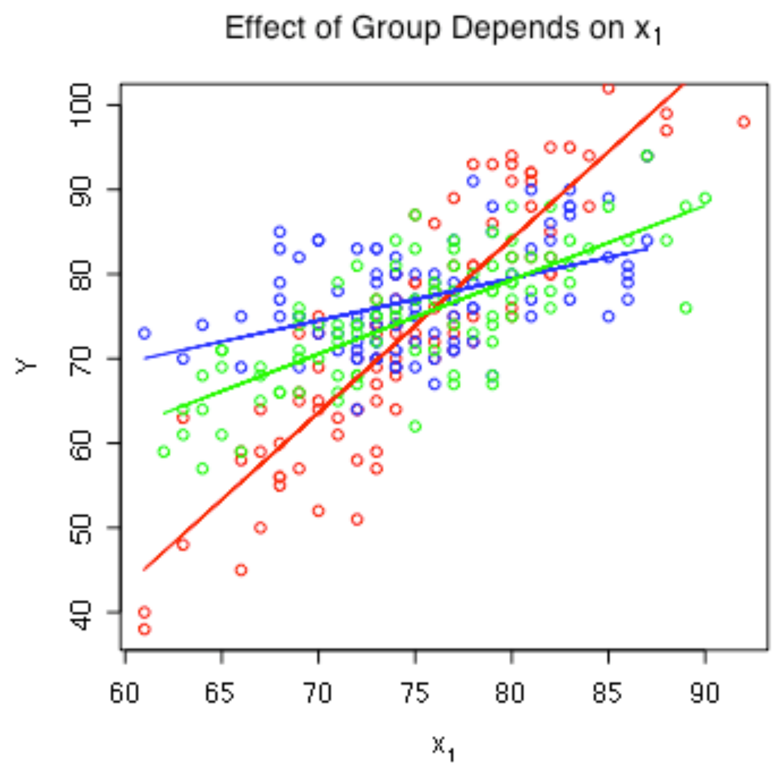
$$E(Y|\mathbf{x}) = (\beta_0 + \beta_2 x_2) + (\beta_1 + \beta_3 x_2) x_1$$

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 independent variable.
- Interaction means slopes of regression lines are not equal.



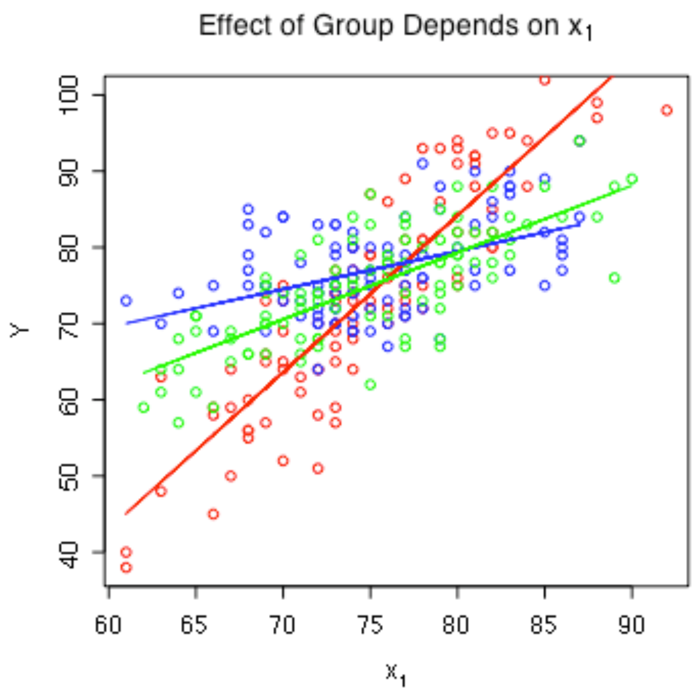
One regression Model

- 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	$\beta_0 + \beta_1 x_1$

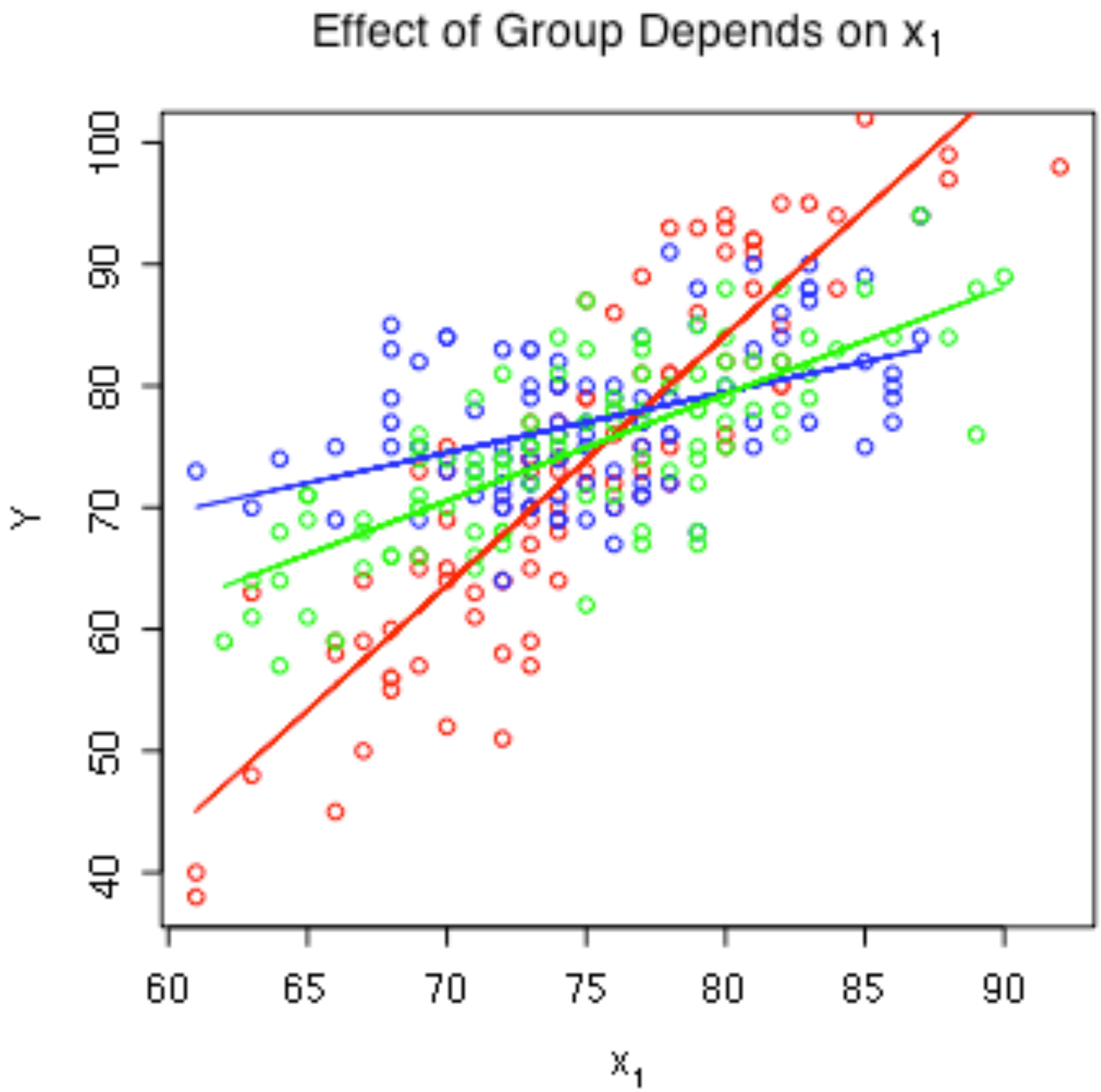
What null hypothesis would you test for

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

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

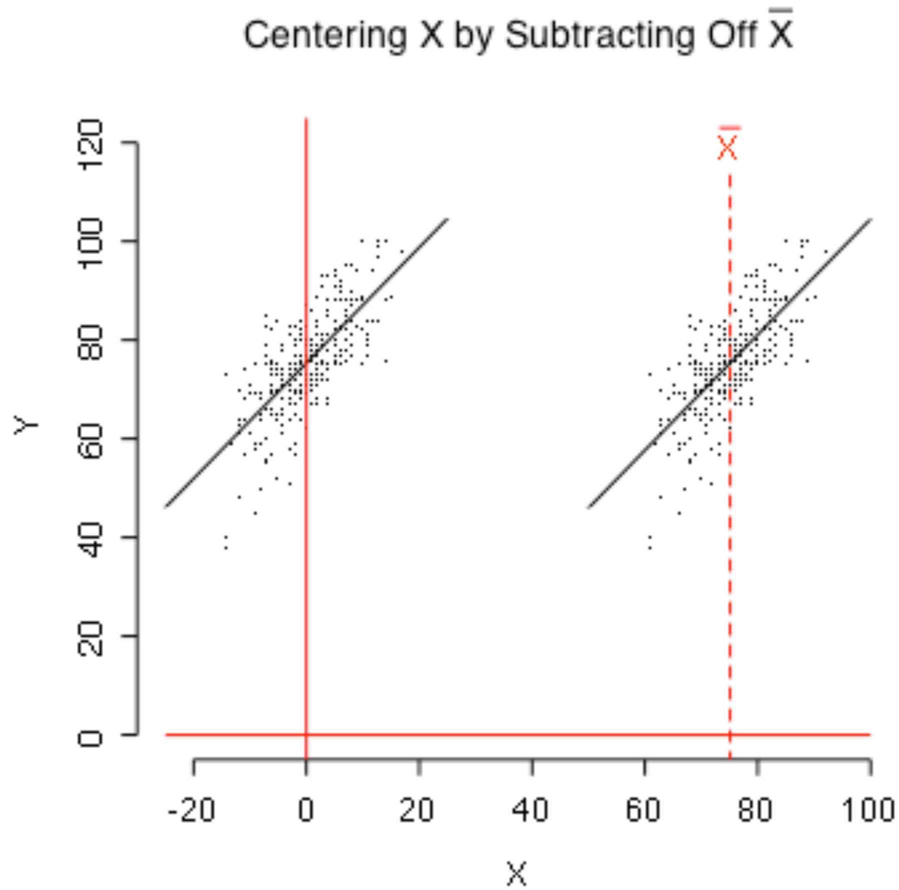
- How do you test Group “controlling” for x_1?
- A popular choice is to set x_1 to its sample mean, and compare treatments at that point. SAS calls the estimates (Y-hat values) “Least Squares Means.”
- Or, test equal regressions, in which mean response is the same for all values of the covariate(s).

Test for differences at mean of x_1?



“Centering” the independent variables

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



Properties of Centering

- When independent variables are centered, estimates and tests for *intercepts* are affected.
- Relationships between independent variables and dependent 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 independent 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* independent variables are centered by subtracting off their means, then the intercept equals Y-bar, exactly.

Comments

- Often, X=0 is outside the range of independent variable values, and it is hard to say what the intercept *means* in terms of the data.
- When independent 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 independent variables

- Subtract mean (for entire sample) from each quantitative independent 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, some examples