Stat 608 Chapter 5

STARTED: Friday 2/25/22 (week 6, lecture 14)
around 27 min mark

Multiple Linear Regression

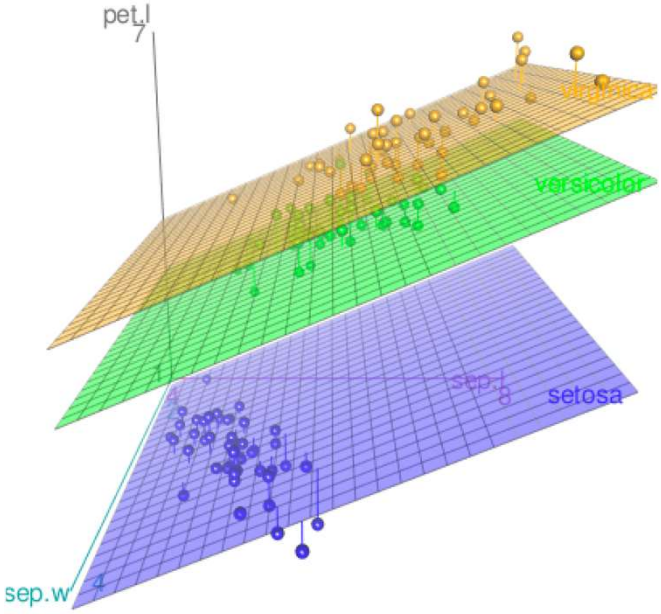
- Chapter 5:
  - Multiple predictor variables
  - ANOVA and ANCOVA
  - Polynomial Regression
  - Assumption that model is valid

- Chapter 6:
  - Leverage points
  - Transformations
  - Relationships between explanatory variables:
    - Multicollinearity
    - Interactions

- Chapter 7:
  - Variable Selection

Types of Multiple Linear Regression

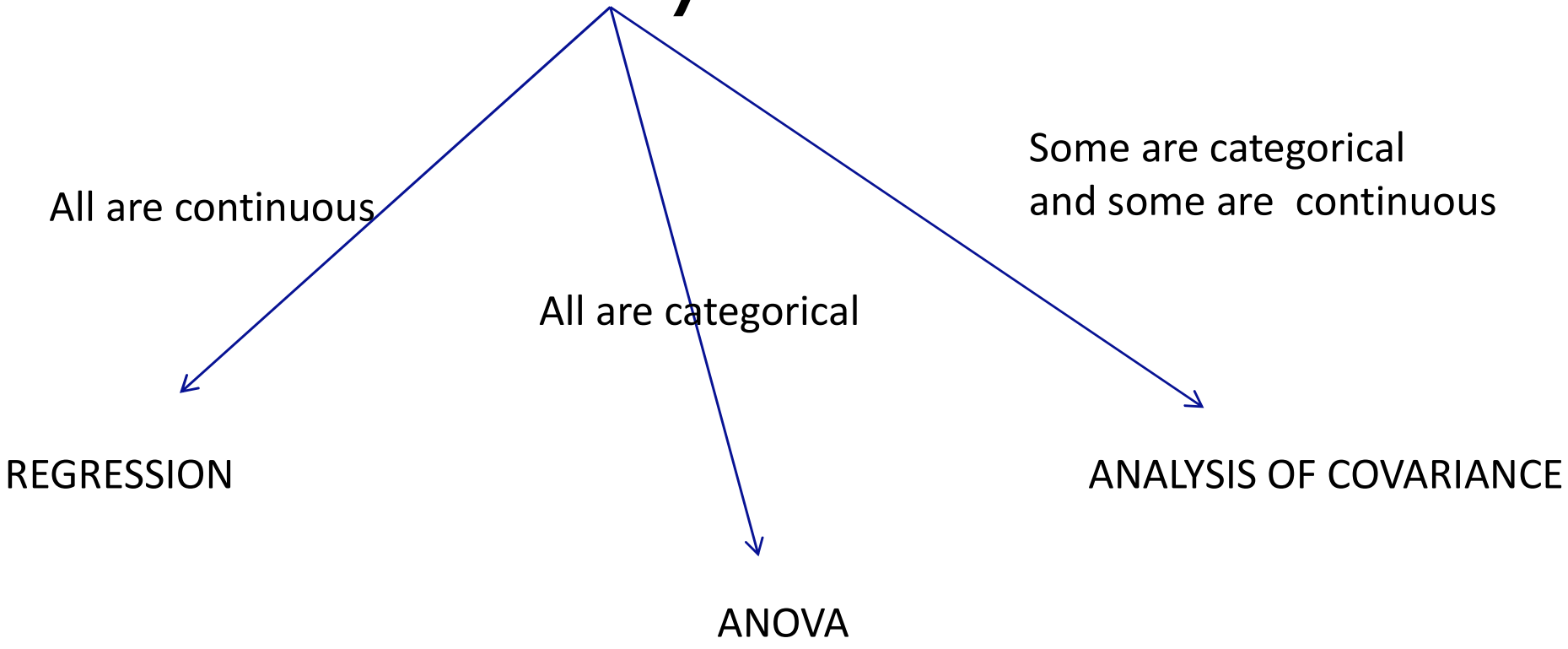
- ANOVA with multiple categories or multiple predictors
- Quantitative and Categorical explanatory variables: ANCOVA (separate lines)
- Polynomial Regression (curves)
- Many explanatory variables (multiple dimensions)
- Combinations of the above!



- (Multivariate Regression actually refers to multiple response variables!)

ANOVA & REGRESSION & ANALYSIS OF COVARIANCE

$$Y = X\beta + e$$



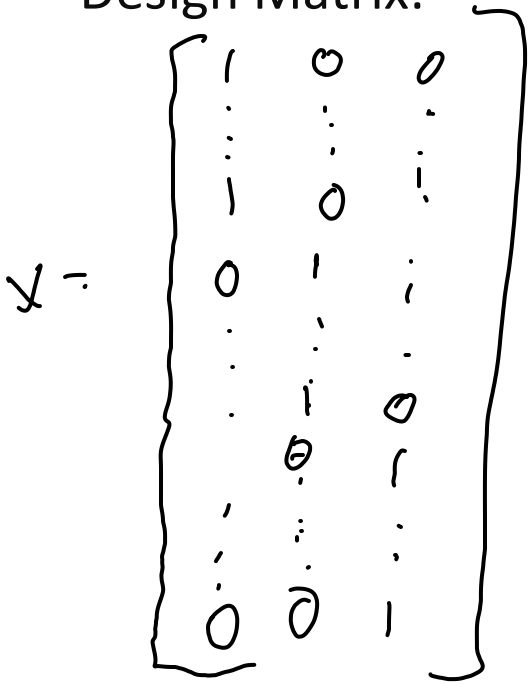
ANOVA (ANalysis Of VAriance)

- One-way ANOVA, without an intercept:

$$y_i = \alpha_j + e_i, \quad j = 1, 2, 3$$

$$y_i = \alpha_1 x_{1i} + \alpha_2 x_{2i} + \alpha_3 x_{3i} + e_i,$$

Design Matrix:



$$\hat{\alpha}_1 = \bar{y}_1$$

$$\hat{\alpha}_2 = \bar{y}_2$$

$$\hat{\alpha}_3 = \bar{y}_3$$

Pro: (X'X) is diagonal: easy to invert!

Con: When we have two-way ANOVA, we have to cut out the last indicator variable…

ANOVA

- One-way ANOVA, with an intercept: x_1, x_2, x_3

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + e_i$$

← must drop 1 during var formodel.

$\alpha_0 = \bar{y}_3$

$\alpha_0 + \alpha_1 = \bar{y}_1$

$\alpha_0 + \alpha_2 = \bar{y}_2$

$\alpha_1 = \mu_1 - \mu_3$

$\alpha_2 = \mu_2 - \mu_3$

Design matrix:

$$X = \begin{bmatrix} 1 & 1 & 0 \\ 1 & \vdots & \vdots \\ & 1 & 0 \\ \vdots & 0 & 1 \\ \vdots & \vdots & \vdots \\ & \vdots & 1 \\ & \vdots & 0 \\ & & \vdots \\ 1 & 0 & 0 \end{bmatrix}$$

Two-Way ANOVA

{ say X_1, X_2 are dummy vars for race (5 race 2)

X_1, X_2: dummy vars for categorical var 1

X_3 : dummy var for categorical var 2

{ say X_3 is gender = { 1 if m, 0 if f

- Model:

$$y_i = \alpha_0 + \alpha_1 x_{1i} + \alpha_2 x_{2i} + \beta x_{3i} + e_i$$

β = mean diff in response comparing gender 1 to gender 2 for the same race

α_1 = mean diff in response comparing race 1 to race 3 for the same gender

α_2 = mean diff in response comparing race 2 to race 3 for the same gender.

α_0 = mean of race 3, gender 2.

Design matrix:

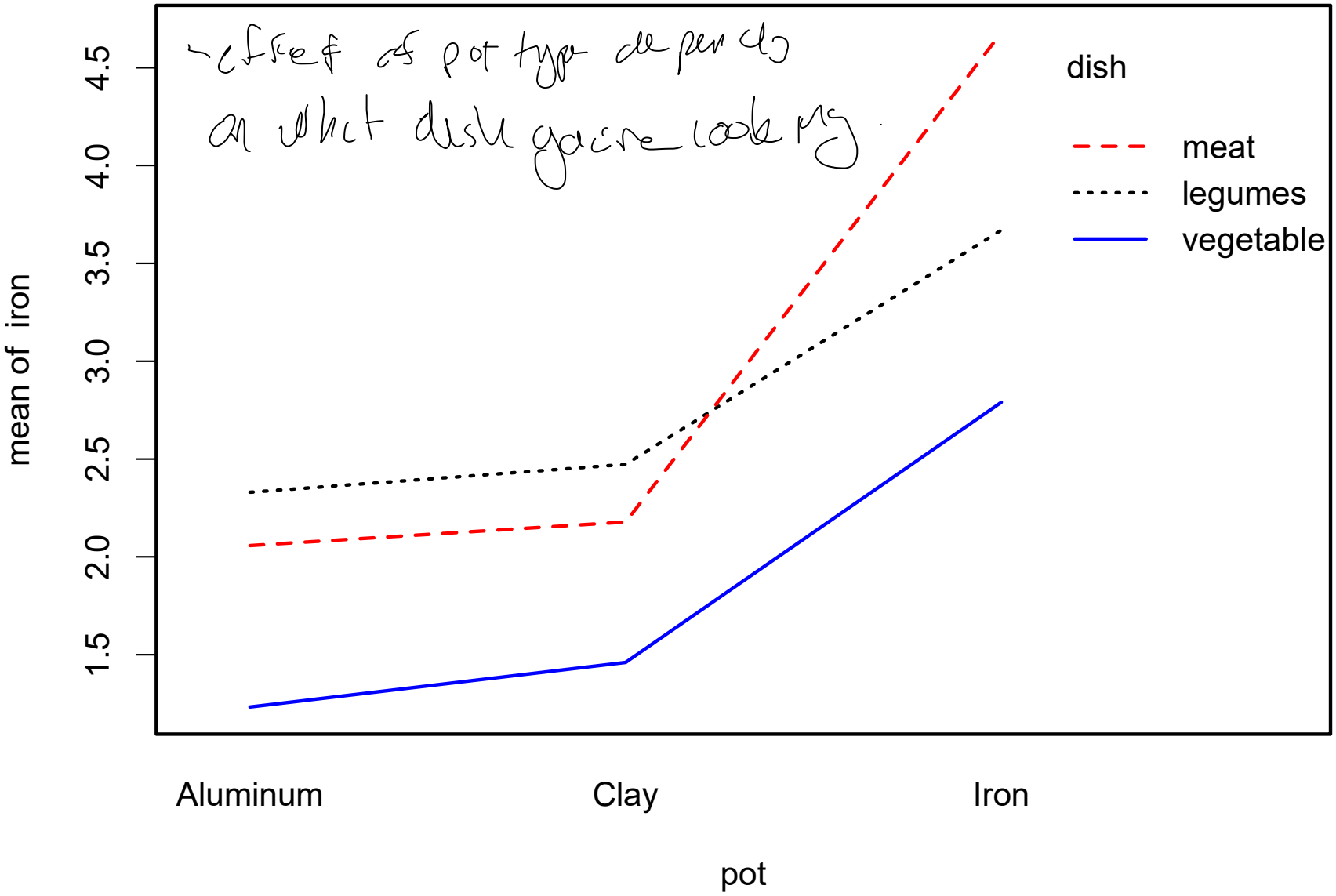
$X =$

$$\begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & \vdots & 1 \\ 1 & 1 & \cdot & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ 1 & \cdot & 1 & 0 \\ 1 & \cdot & 1 & 0 \\ 1 & \cdot & 0 & 1 \\ 1 & & \vdots & 0 \\ 1 & & \cdot & 0 \\ 1 & 0 & 0 & 0 \end{bmatrix}$$

STOP Friday 2/25/22 (Week 4, Lecture 6)

START Monday 2/28/22 (week 7, lecture 17)

Interactions: ANOVA



ANOVA: Interactions

What would the model and design matrix look like in the case of the iron pot example? Interpret the parameter estimates.

$$y_i = \alpha_0 + \alpha_1 meat_i + \alpha_2 legumes_i + \alpha_3 alum_i + \alpha_4 clay_i + \alpha_5 (meat * alum)_i + \alpha_6 (meat * clay)_i + \alpha_7 (legumes * alum)_i + \alpha_8 (legumes * clay)_i + e_i$$

α_0 = mean for veg, iron

$E[y_i | meat, alum] = \alpha_0 + \alpha_1 + \alpha_3 + \alpha_5$ $\Rightarrow \alpha_1 + \alpha_5$: mean diff. comparing meat to veg in alum pot.

$E[y_i | veg, alum] = \alpha_0 + \alpha_3$

$E[y_i | meat, clay] = \alpha_0 + \alpha_1 + \alpha_4 + \alpha_6$ $\Rightarrow \alpha_1 + \alpha_6$: mean diff comparing meat to veg in clay pot.

$E[y_i | veg, clay] = \alpha_0 + \alpha_4$

$E[y_i | meat, iron] = \alpha_0 + \alpha_1$ $\Rightarrow \alpha_1$: mean diff comparing meat to veg in iron pot.

$E[y_i | veg, iron] = \alpha_0$

Analysis of Covariance

- Suppose we have three groups, and want to compare the three means, holding the value of a quantitative variable x constant.
- Example: Compare diets, holding starting BMI constant.
- It's possible to create three separate regression lines, as shown below. We might also create three lines with separate intercepts, but the same slope, or three lines with separate slopes, but the same intercept.

Group 1: $y_{i1} = \beta_{01} + \beta_{11} x_i + e_i$

Group 2: $y_{i2} = \beta_{02} + \beta_{12} x_i + e_i$

Group 3: $y_{i3} = \beta_{03} + \beta_{13} x_i + e_i$

ANCOVA Model

- Where x1 is a quantitative variable and x2 is an indicator (dummy) variable, write down the model for separate slopes, separate intercepts:

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} \cdot x_{2i} + e_i$

$x_2 = 1: (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_{1i} + e_i$

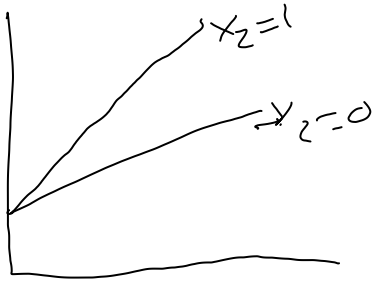
$x_2 = 0: \beta_0 + \beta_1 x_{1i} + e_i$

- Write down the model with separate slopes, but the same intercept:

$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_{1i} x_{2i} + e_i$

$x_2 = 1: \beta_0 + (\beta_1 + \beta_2) x_{1i} + e_i$

$x_2 = 0: \beta_0 + \beta_1 x_i + e_i$

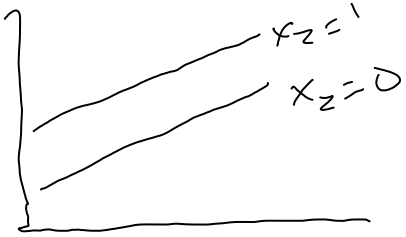


- Write down the model with separate intercepts, but the same slope:

$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + e_i$

$x_2 = 1: (\beta_0 + \beta_2) + \beta_1 x_{1i}$

$x_2 = 0: \beta_0 + \beta_1 x_{1i}$



...

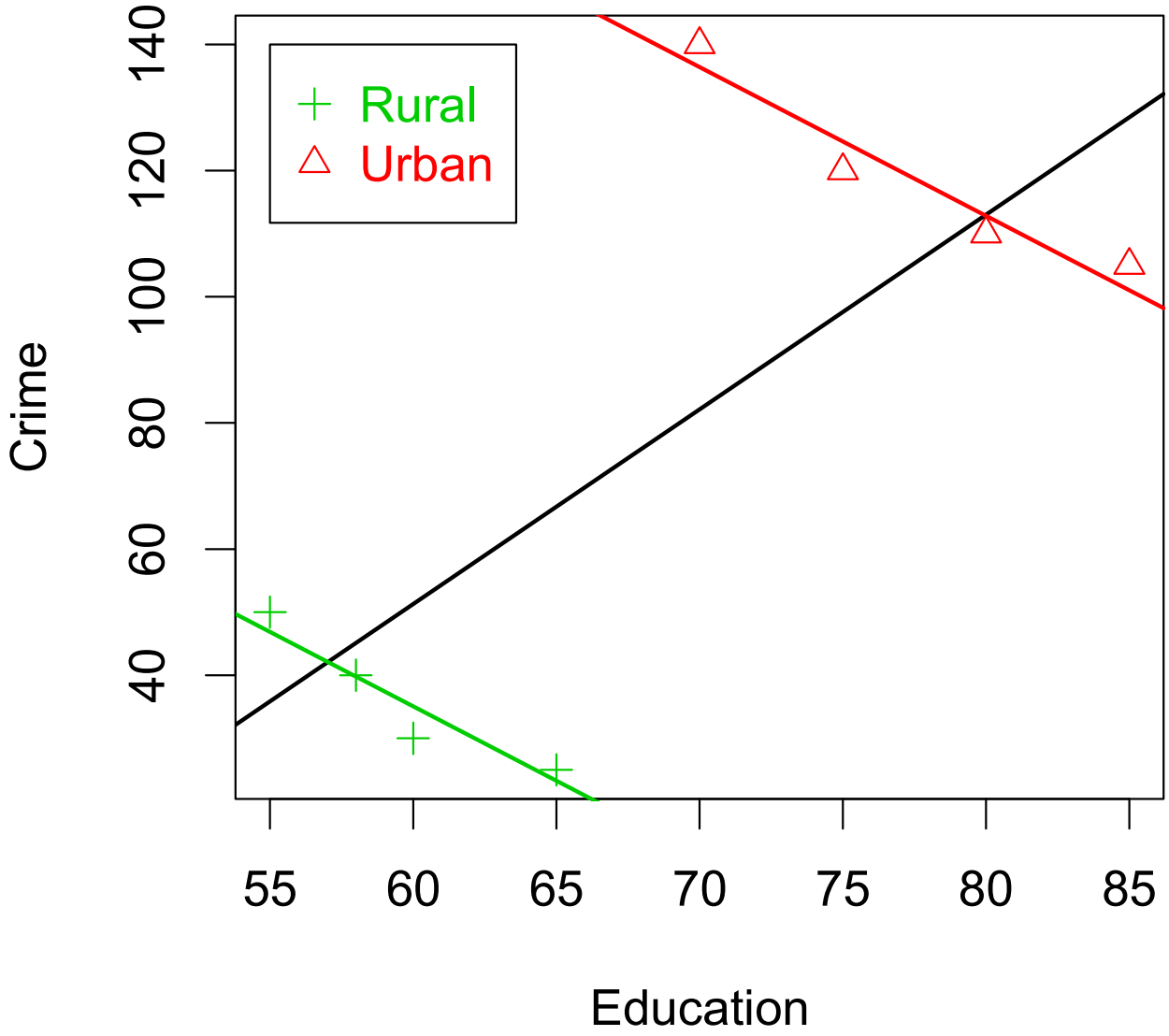
ANCOVA Example

Rats are randomly assigned to be fed 0, 2, 4, and 6 mg of one of two cancerous substances. The response variable y is the number of tumors recorded. What should the model look like? What should the design matrix look like?

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{1i} x_{2i} + \varepsilon$$

$$X = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 2 & 2 \\ 1 & 2 & 2 \\ 1 & 4 & 4 \\ 1 & 4 & 4 \\ 1 & 6 & 6 \\ 1 & 6 & 6 \end{bmatrix}$$

ANCOVA: Same slopes; Simpson's Paradox



Interactions

- An **interaction** between two explanatory variables exists when the effect of one explanatory (X_1) on the response variable (Y) is different for different values of the other explanatory variable (X_2): The second variable changes the *relationship* between the first and the response!

- Ex: Adding sugar (X_1) to coffee makes it much sweeter (Y) when the coffee is stirred (X_2).

- Ex: Injecting one more kg of sand (X_1) into a fracking well has a larger effect on oil production (Y) when the well is shorter (X_2).

- Ex: The amount of iron in food (Y) is higher when cooking in a cast iron pot (X_1). While tomatoes have a tiny amount of iron in them, the acidity in tomatoes means their presence in food (X_2) has a multiplicative effect on cast iron.

- Ex: Third grade math students with ADHD (X_1) have lower math scores (Y) if they studied while listening to music (X_2), while those without ADHD had higher math scores when listening to music.

...

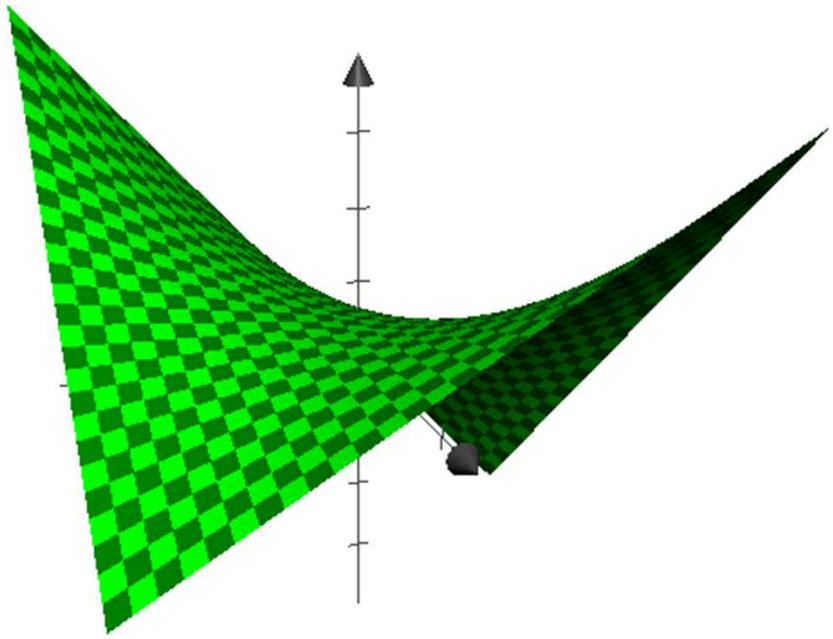
ANCOVA Example

- In the event that we want to fit two separate lines with different slopes, why don't we just break up the data set into subsets according to the categorical variable and fit separate lines?

By considering all the data together in one model we have more statistical power $\rightarrow 1 - P[\text{Type II error}]$

Multiple Regression: Interactions

- Interactions between quantitative variables: fit different kinds of surfaces.



Multiple Regression: Interactions

- If an interaction exists, there are many possible things that could be true:

- The relationship could change direction in the presence of the third variable: the relationship between X_1 and Y is positive before taking into account X_2, and negative afterward. (Simpson's paradox)

- The relationship might not change direction in the presence of a third variable, but merely have a dramatic multiplicative effect: Fast driving (X_1) is much more dangerous (Y) when drunk(X_2).

- Main effects might not be significant, while the interaction is significant.

- Cause and effect could run in many possible directions, but we can only scientifically establish cause and effect through direct experimentation.

...

Interactions Vs. Simpson's Paradox

- Simpson's paradox
  - Can happen merely by adding another variable; no interaction has been added!
  - From one explanatory variable to two, slopes change direction, but lines can *still be parallel*.
  - More education is associated with more crime until we add the urban/rural variable. But education has the SAME effect on both urban and rural areas: decreasing crime.
- Interactions
  - From two explanatory variables to two with a multiplier effect. Lines cross.
  - Cooking in an iron pot has an even LARGER effect in the presence of meat (and tomatoes).

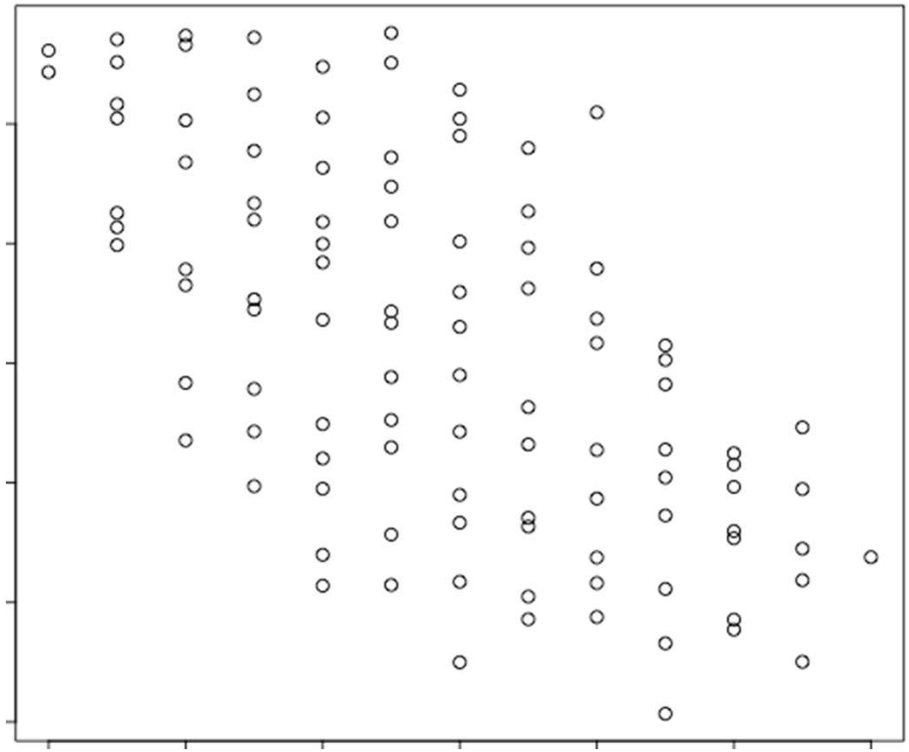
Interactions vs. Independence

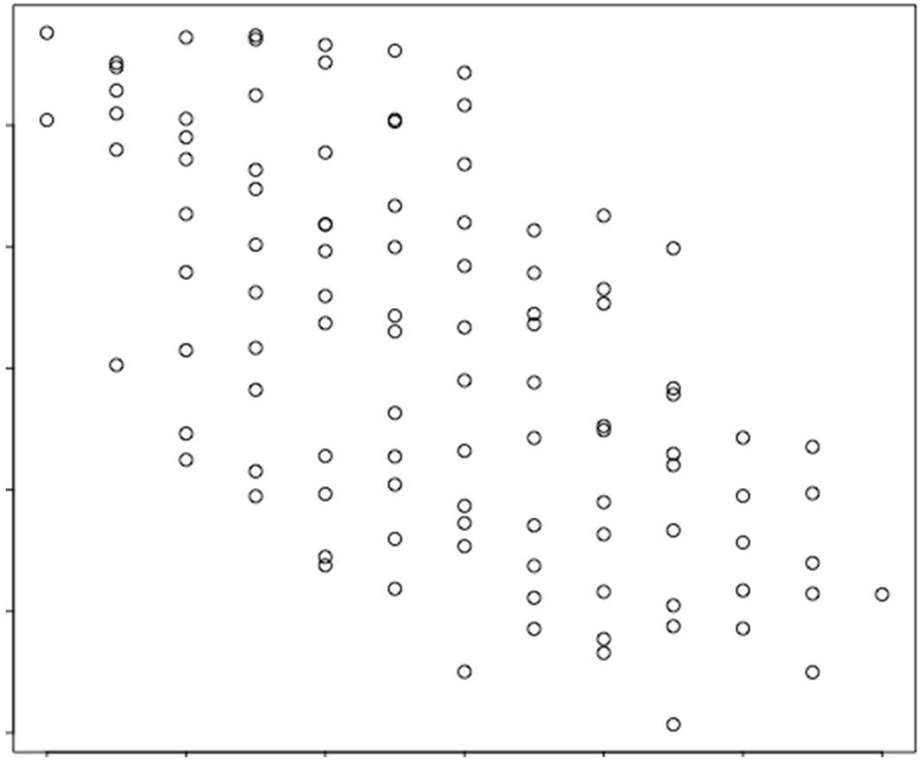
- Variables X_1 and X_2 may be independent, with correlation = 0, yet interact when explaining the response variable. For example:
  - In an experiment, a chemist has designed x_1 = reaction temperatures and x_2 = pressures to be independent: every temperature and pressure combination has been used exactly on two repetitions of the experiment. But the effect of temperature on y = product yield from chemical reaction is different for different pressures.
  - The effect of x_1 = fertilizer on y = crop yield depends on x_2 = soil type.
  - A researcher has randomly sampled houses x_1 = with an ocean view and without that are the same x_2 = age, on average. However, age has a different effect on y = price, depending on whether the house has an ocean view.

STOP today 2/28/22 (week 7. lecture 17)

START Wednesday 3/2/22 (week 7, lecture 18)

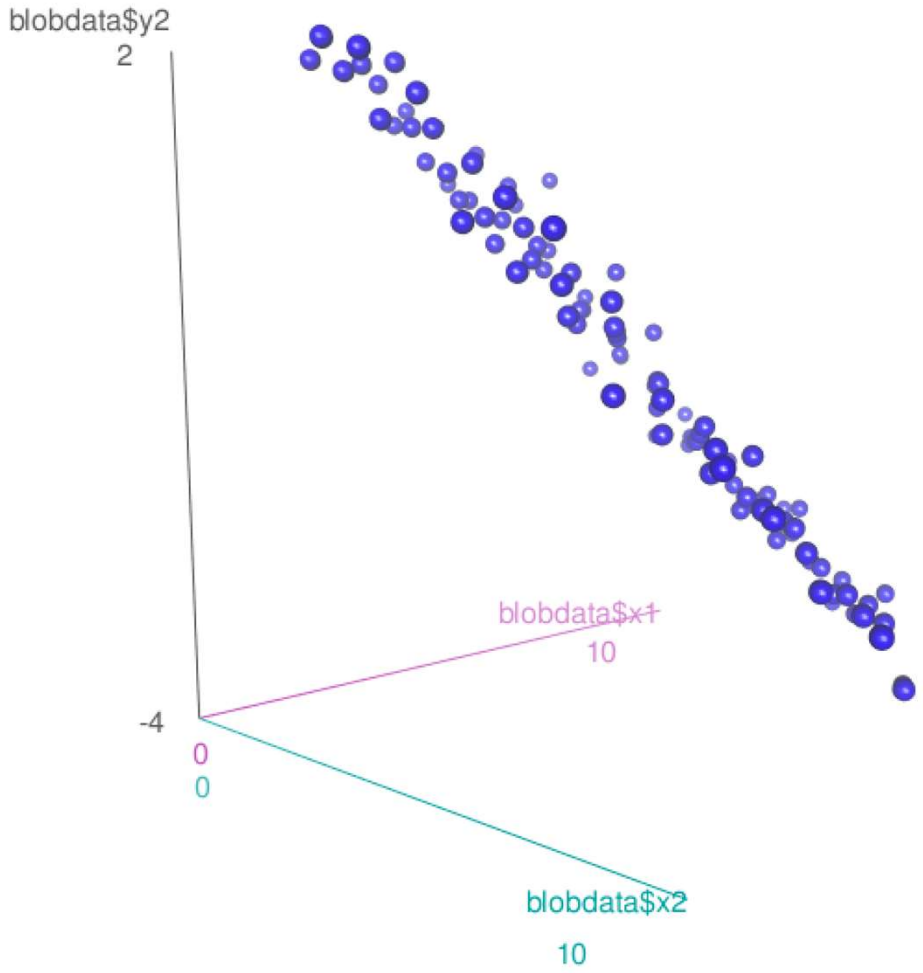
The Fallacy of Bivariate Thinking





The Fallacy of Bivariate Thinking

- Plots of x1 or x2 vs. y may seem to have no relationship with y.
- But the two variables working together may explain much more of the variation in y.
- Ex: Weight, Height, and Body fat percentage.



Rank of X

The rank of our design matrix X should be the number of columns of X. We say our design matrix is not full rank if it isn't.

- If rank (X) < # columns = p + 1, that means there exist a linear combination of the other variables that adds up to one of the variables. Why do we need that extra variable??
- If rank(X) < p + 1, that means rank (X' X) < p+1, so X'X is not invertible.
- If n < p + 1, rank(X) < p + 1 because the rank of X has to be less than or equal to both the number of columns and the number of rows of X. Get a bigger sample size or get rid of some variables.

Rank of X

- R error message:

```
Coefficients: (1 not defined because of
singularities)
```

- SAS error message:

Note: Model is not full rank. Least-squares solutions for the parameters are not unique. Some statistics will be misleading. A reported DF of 0 or B means that the estimate is biased.

Note: The following parameters have been set to 0, since the variables are a linear combination of other variables as shown.

Step 1: Multiple Regression

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0 \text{ vs.}$$

$$H_a : \text{At least one } \beta_i \neq 0$$

$$F = \frac{SSReg/p}{RSS/(n-p-1)}$$

Always conduct this F-test first, before looking at tests for individual variables (next slide).

If this p-value is large, STOP.

Analysis of variance table

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F
Regression	p	SSreg	SSreg/p	$F = \frac{\text{SSreg} / p}{\text{RSS} / (n-p-1)}$
Residual	$n-p-1$	RSS	S^2= RSS/ $(n-p-1)$	
Total	$n-1$	SST = SYY		

Italian Restaurants: Houston

```
my.lm<- lm(Italian$food ~Italian$Service +
Italian$Pct_Liked + Italian$Cost)

anova(my.lm)


Response: Italian$Food
                   Df Sum Sq Mean Sq F value    Pr(>F)
Italian$Service     1 64.619  64.619  36.369 2.605e-07
Italian$Pct_Liked   1 36.510  36.510  20.548 4.133e-05
Italian$Cost        1  1.860   1.860   1.047    0.3116
Residuals          46 81.731   1.777

Residual standard error: 1.333 on 46 degrees of freedom
  (15 observations deleted due to missingness)
Multiple R-squared: 0.5575, Adjusted R-squared: 0.5287
F-statistic: 19.32 on 3 and 46 DF,  p-value: 2.989e-08
```

Step 2: Multiple Regression

$H_0: \beta_i = 0 \quad H_A: \beta_i \neq 0$

our t statistic

$$t_{n-p-1} = \frac{\hat{\beta}_i - \beta_i}{se(\hat{\beta}_i)}$$

$$\hat{\beta}_i \pm t^*_{n-p-1} se(\hat{\beta}_i)$$

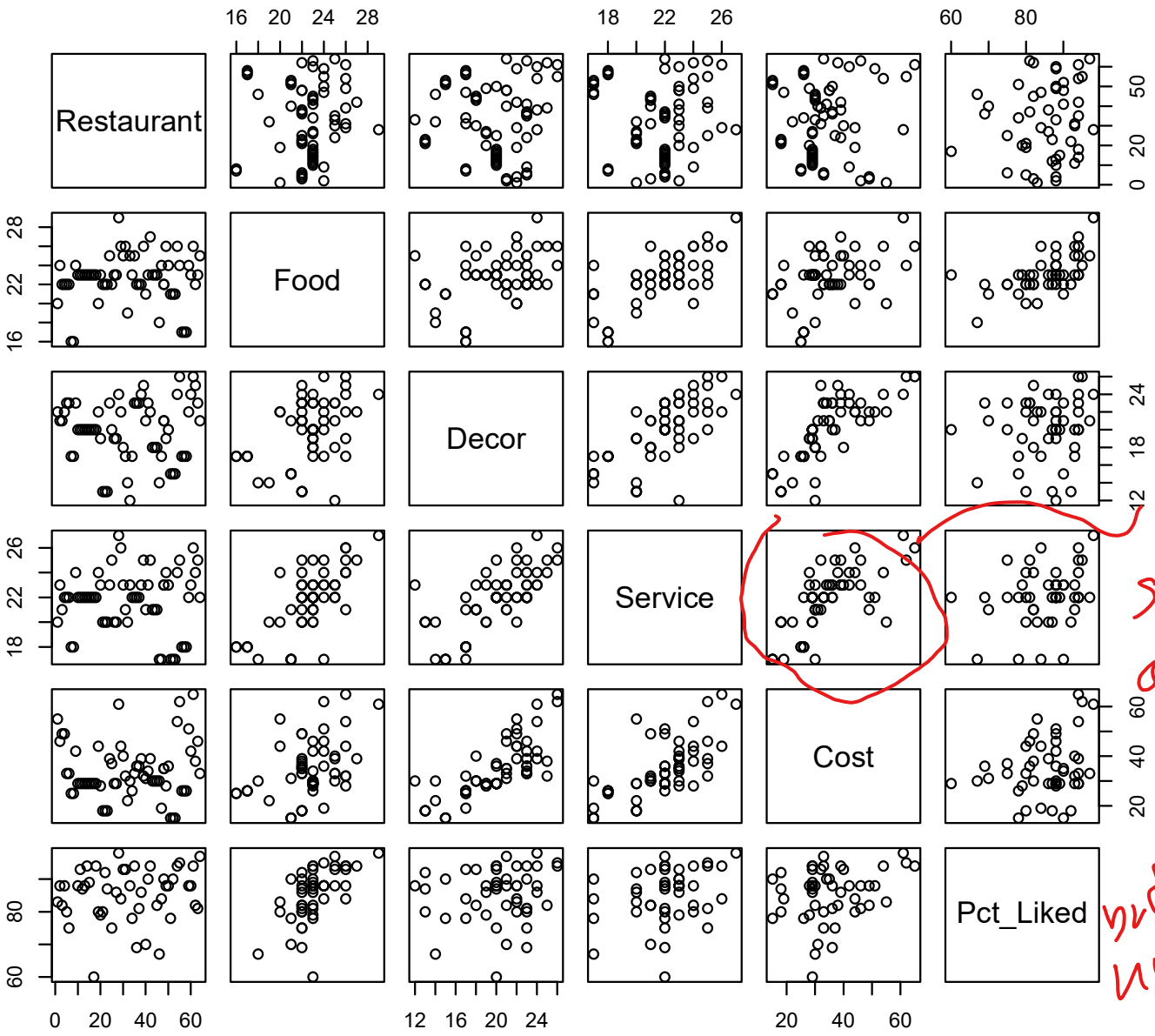
- Note: If we start conducting these tests for many of the variables, performing p separate t-tests, our overall Type I error increases.
- We also run into problems when the predictor variables are highly correlated with each other (see chapter 7).

Step 2: Multiple Regression

- We often write $H_0 : \beta_1 = 0$ without thinking.
- The *implicit* assumption is all the other variables are still in the model.
- If the p-value for one variable is large, the conclusion is that variable has no significant *additional* effect on the response variable (after all the other variables are entered into the model).
- When predictors are correlated, important variables may become insignificant.
- Example: Weight and BMI have large p-values for predicting cholesterol levels.

Italian Restaurants: Houston

plot(Italian)



Italian Restaurants: Houston

```
my.lm<- lm(Italian$food ~Italian$Service +
Italian$Pct_Liked + Italian$Cost)

summary(my.lm)

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)        3.68923    2.66400   1.385  0.17278
Italian$Service    0.47849    0.11744   4.074  0.00018
Italian$Pct_Liked  0.11304    0.02461   4.594 3.38e-05
Italian$Cost      -0.02236    0.02185  -1.023  0.31156
```

Italian Restaurants: Houston

- The coefficient for cost is negative; does that make sense? Interpret the slope for Cost in context.

maybe not the p-value is large so we can't distinguish it from 0.

for every $1 increase in cost, we estimate an average decrease in food quality of 0.02236, holding the other variables constant.

- The p-value for cost is large; does that make sense?

It does if cost doesn't contribute anything to food, once the other variables have been adjusted for.

NOTE: cost & service are highly correlated

Italian Restaurants: Houston

- New model:

```
Coefficients:
                   Estimate Std. Error t value Pr(>|t|)
(Intercept)        10.77121    2.32996   4.623 2.97e-05
Italian$Pct_Liked   0.13047    0.02797   4.665 2.58e-05
Italian$Cost        0.03510    0.01927   1.821   0.0749
```

- The coefficient for %Liked is smaller than the coefficient for Service; does that mean Service is more important when it comes to predicting Food rating?

Not necessarily b/c the variables are measured in different units.

- The p-value for %Liked is smaller than the p-value for Cost; does that mean the association between %Liked and Food rating is stronger than the association between Cost and Food rating?

Not necessarily b/c these p-values are "after adjusting" for the other variables.

· If we looked at Food vs cost & Food vs entry separately we could answer this question.

Italian Restaurants: Houston

- Which variable is most *important*?
  - Sorting by F-value in ANOVA table, t-value (if all variables are quantitative), or p-value will give the same results.
  - If all predictors are independent of each other, sorting by their correlation with the response will also give the same results.

- Ex: Oil Production

Italian Restaurants: Houston

- Calculate an approximate 95% confidence interval for the slope for ~~Service~~ cost.

$0.0351 \pm 2.029 * 0.01927 = [-0.004, 0.074]$

SIOP Wednesday 3/2/22 (week 7, lecture 18)

START Friday 3/4/22 (Week 7 Lecture 19)

Polynomial Regression

- Is the following model linear in the parameters?

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$$

yes. linear in the β coeff.'s

- If a model is linear in the parameters, it means we can write it as:

$$\mathbf{Y} = \mathbf{X\beta} + \mathbf{e}$$

when y is transformed it is no longer a linear model.

ex: $\log(Y) = \beta + \beta_1 x_i + e_i$

$\Rightarrow y = e^{\beta + \beta_1 x_i + e_i}$

Polynomial Regression

- What does the design matrix look like for the following model?

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 x^3 + e$$

$x = 1, 2, 3, 4$

$$X = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 16 \\ 1 & 3 & 9 & 36 \\ 1 & 4 & 16 & 64 \end{bmatrix}$$

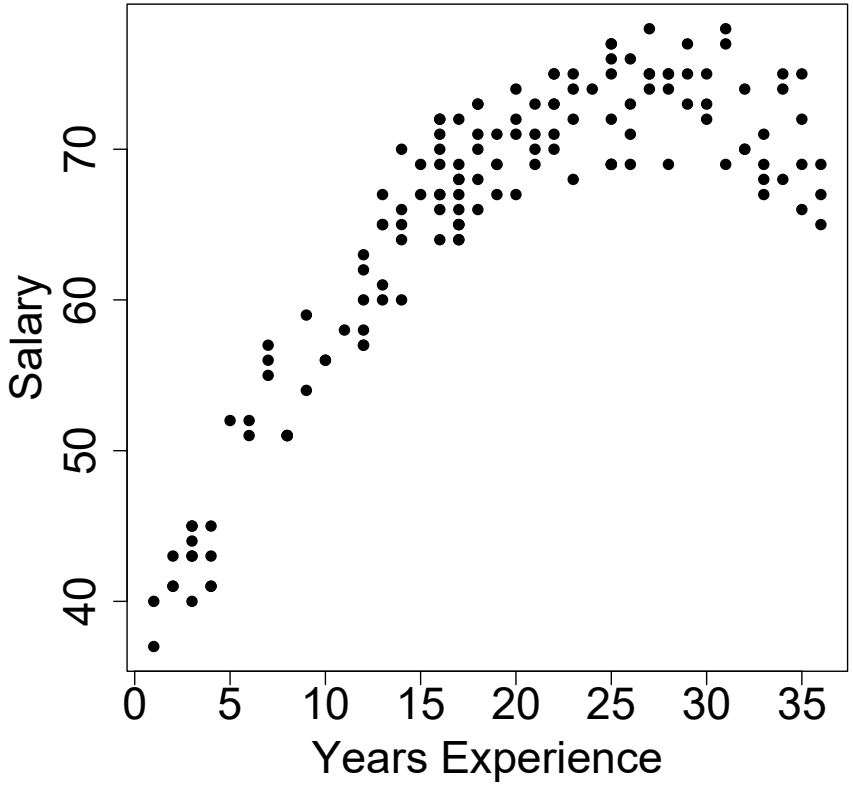
Salary Example

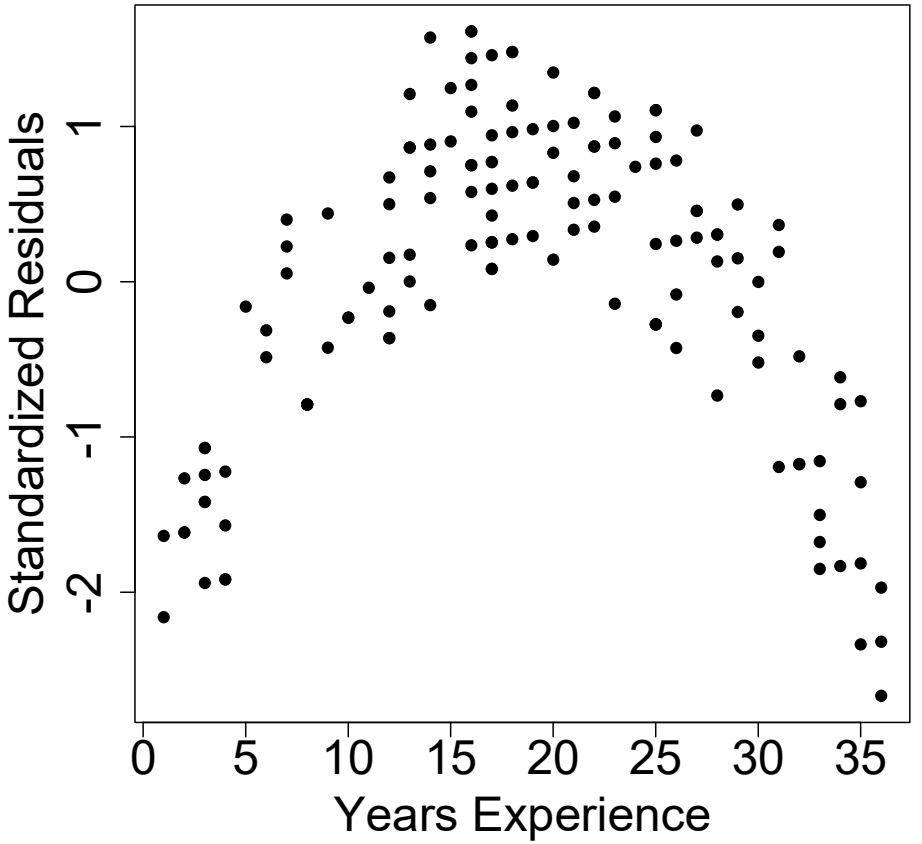
A salary curve relates salary to years of experience. Employees might use it find out where they stand among their peers. Personnel might use it to consider salary adjustments when hiring new professionals.

When we fit the simple linear regression model below, we get the residual plot on the next slide.

$$y_i = \beta_0 + \beta_1 x + e$$

Salary Example: SLR





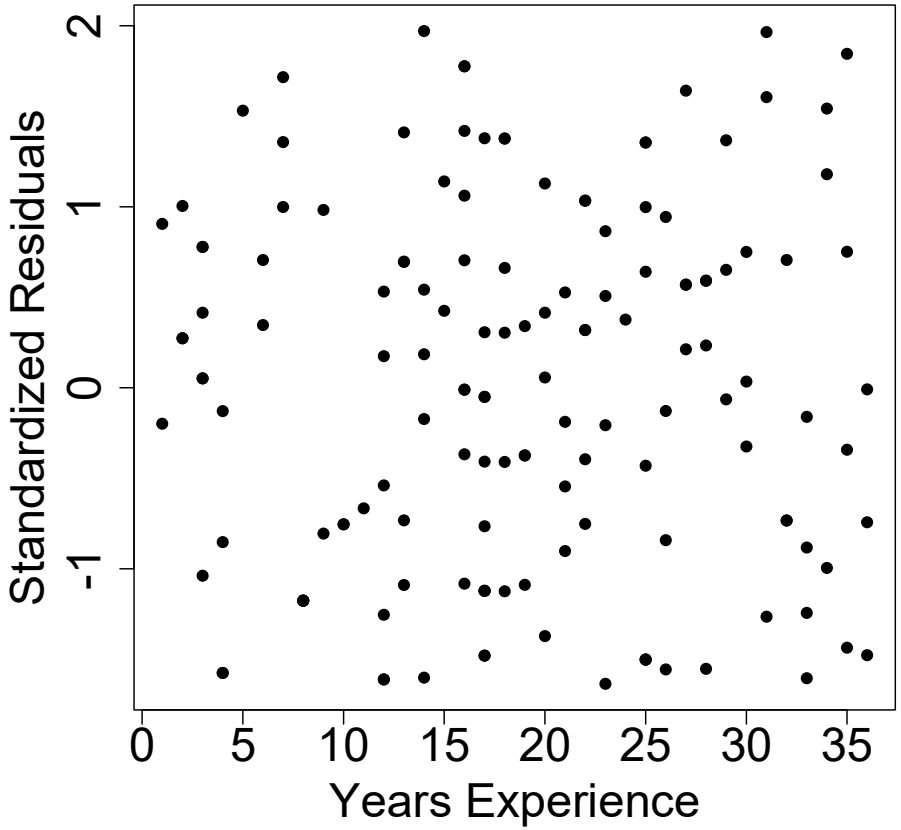
Salary Example: Parabola

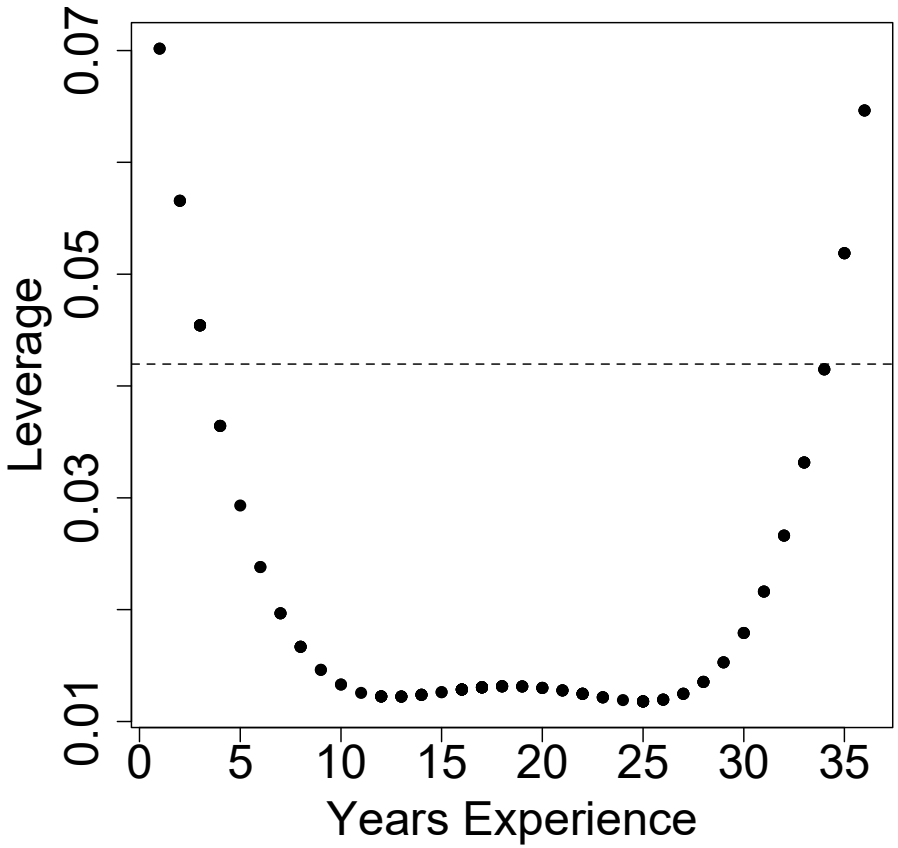
- Clearly there is a non-linear relationship between salary and years of experience.
- Next we fit the model

$$y_i = \beta_0 + \beta_1 x + \beta_2 x^2 + e$$

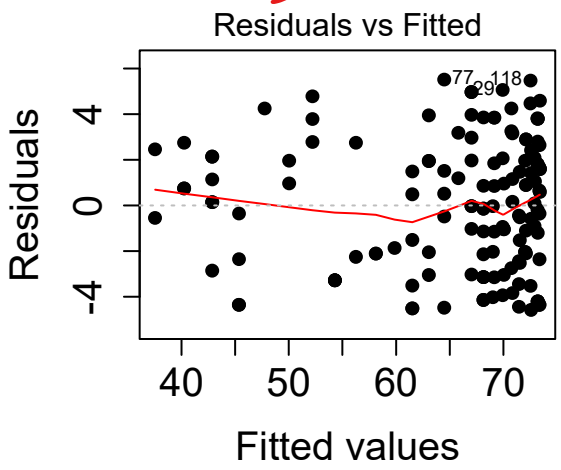
- The cutoff for high leverage is $h_{ii} > 2(p+1) / n$ when there are p predictors plus 1 intercept.

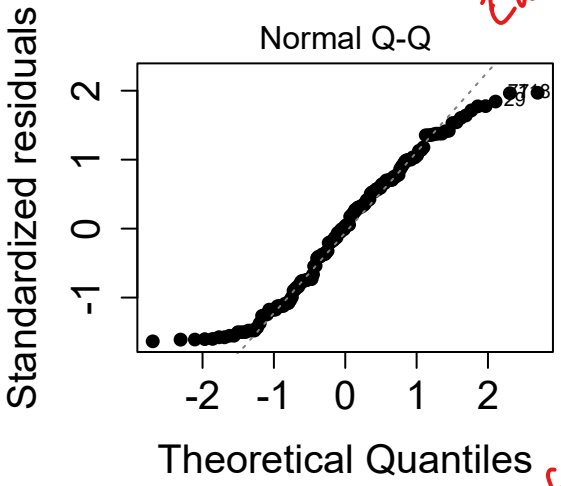
Salary Example: Parabola

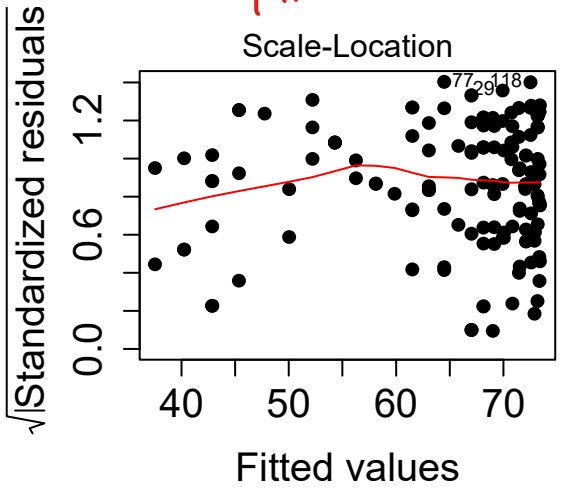


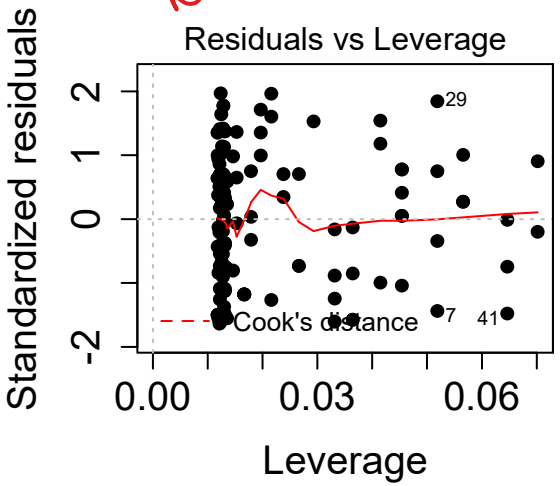


Salary Example: Parabola









Approach: Center explanatory variable

* not included on plots slide

Model Reduction **Method 1 – Partial F-Test**

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + e$$

Suppose we have the model above, and are interested in testing

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0, k < p$$

against the alternative hypothesis

Ha: At least one of the parameters is not 0.

That is, the question is, "Can we drop all of these k variables from our model?" This can be tested using an F-test.

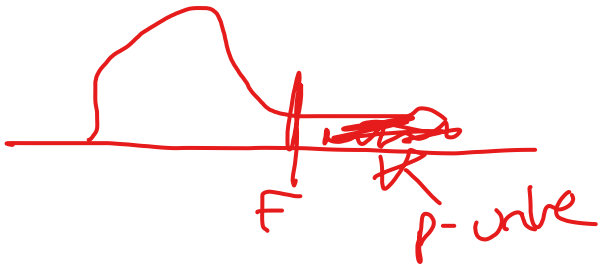
Model Reduction **Method 1 – Partial F-Test**

Let RSS(Full) be the residual sum of squares from the model with all the predictors 1, …, p. Let RSS(Reduced) be the residual sum of squares from the reduced model (with only the remaining predictors that we don't think are 0), and the df's be error degrees of freedom. Then the F-statistic for testing the above hypotheses is given by:

$$F = \frac{(RSS(reduced) - RSS(full))/(df_{reduced} - df_{full})}{RSS(full)/df_{full}}$$

$$= \frac{(RSS(reduced) - RSS(full))/k}{RSS(full)/(n-p-1)}$$

Under H_0 : $F \sim F_{k,n-p-1}$



(This is slide 47 → see note on slide 46)

$$H_0: \begin{bmatrix} \mu_{clay} = \mu_{alum} \\ \mu_{iron} = \mu_{alum} \\ \mu_{clay} = \mu_{iron} \end{bmatrix} \Leftrightarrow \begin{bmatrix} \beta_1 = 0 \\ \beta_2 = 0 \\ \beta_1 - \beta_2 = 0 \end{bmatrix} \qquad A_{3\times 5} = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 \end{bmatrix}, \qquad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix}$$

$H_0: A\beta = \underline{0}$

Model Reduction **Method 2**

$$H_0 : \mathbf{A}_{r\times(p+1)}\boldsymbol{\beta}_{(p+1)\times 1} = h_{r\times 1}$$

H_a: At least one of these not equal.

$$F = \frac{(\mathbf{A}\hat{\boldsymbol{\beta}} - h)'(\mathbf{A}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{A}')^{-1}(\mathbf{A}\hat{\boldsymbol{\beta}} - h)/r}{\hat{\mathbf{e}}'\hat{\mathbf{e}}/(n-p-1)}$$

rank(A) = r

Under H_0: $f \sim F_{r,\,n-p-1}$

r = # of conditions we testing.

Model Reduction

$y_i = \beta_0 + \beta_1 I(clay)_i + \beta_2 I(Iron)_i + \beta_3 I(meat)_i + \beta_4 I(Veg)_i + e_i$

$\mu_{alum,leg} = \beta_0$; $\mu_{clay,leg} = \beta_0 + \beta_1$, $\mu_{iron,leg} = \beta_0 + \beta_2$;

$\mu_{alum,meat} = \beta_0 + \beta_3$, $\mu_{clay,meat} = \beta_0 + \beta_1 + \beta_3$, $\mu_{iron,meat} = \beta_0 + \beta_2 + \beta_3$

$\mu_{alum,veg} = \beta_0 + \beta_0$, $\mu_{clay,veg} = \beta_0 + \beta_1 + \beta_4$, $\mu_{iron,veg} = \beta_0 + \beta_2 + \beta_4$

$\beta_1 - \beta_2 = \mu_{clay,leg} - \mu_{iron,leg} = \mu_{clay,meat} - \mu_{iron,meat} = \mu_{clay,veg} - \mu_{iron,veg}$.

~~if~~ needed space, go up to empty slide 2 slides up.

Caution: Remember multiple comparisons need adjustments!

R^2

- R^2 is often defined as the proportion of the variability in the random variable Y explained by the regression model.

$$SSreg + RSS = SST$$

$$R^2 = \frac{SSreg}{SST} = 1 - \frac{RSS}{SST}$$

R^2: Adding Variables

- Adding irrelevant predictor variables to the regression equation increases R^2.
- **Solution**: R^2 adjusted

$$R^2_{adj} = 1 - \frac{RSS/(n-p-1)}{SST/(n-1)}$$

- The denominator is an unbiased estimate of the variance of Y with all slopes = 0, while the numerator is an unbiased estimate of the variance of the residuals.
- Beware: when used to compare models, R^2 adjusted is biased towards adding too many (irrelevant) predictor variables. (See Chapter 7 for more info.)

The mean function might not be modeled correctly because:

- We didn't add variables we should have.
  - Simpson's paradox: Slope sign changes.
  - Fallacy of univariate thinking: P-values could be large when they should be small.
- We added variables we shouldn't have.
  - Correlated predictors: Slope sign changes.
  - Remember interpretation that we hold other variables constant: P-values could be large when they should be small.
  - Chapter 7: more info.
- We didn't consider interactions.
  - Main effects (before interaction added) could have large p-values while interaction has small p-value.
- We didn't consider polynomial terms.
  - A symmetric parabola may have a 0 slope when fitted with a straight line model.

SHARED R code @ ≈ 37 min mark (Salary & Italian)