Statistical Techniques II

EXST7015

Analysis of Covariance

Simple Linear Regression

- Regression is usually done as a least squares technique applied to QUANTITATIVE VARIABLES.
- ANOVA is the analysis of categorical (class, indicator, group) variables, there are no quantitative "X" variables as in regression, but this is still a least squares technique.

Analysis of Covariance (AnCova)

- It stands to reason that if Regression uses the least squares technique to fit quantitative variables, and ANOVA uses the same approach to fit qualitative variables, that we should be able to put both together into a single analysis.
- We will call this Analysis of Covariance.

AnCova *(continued)*

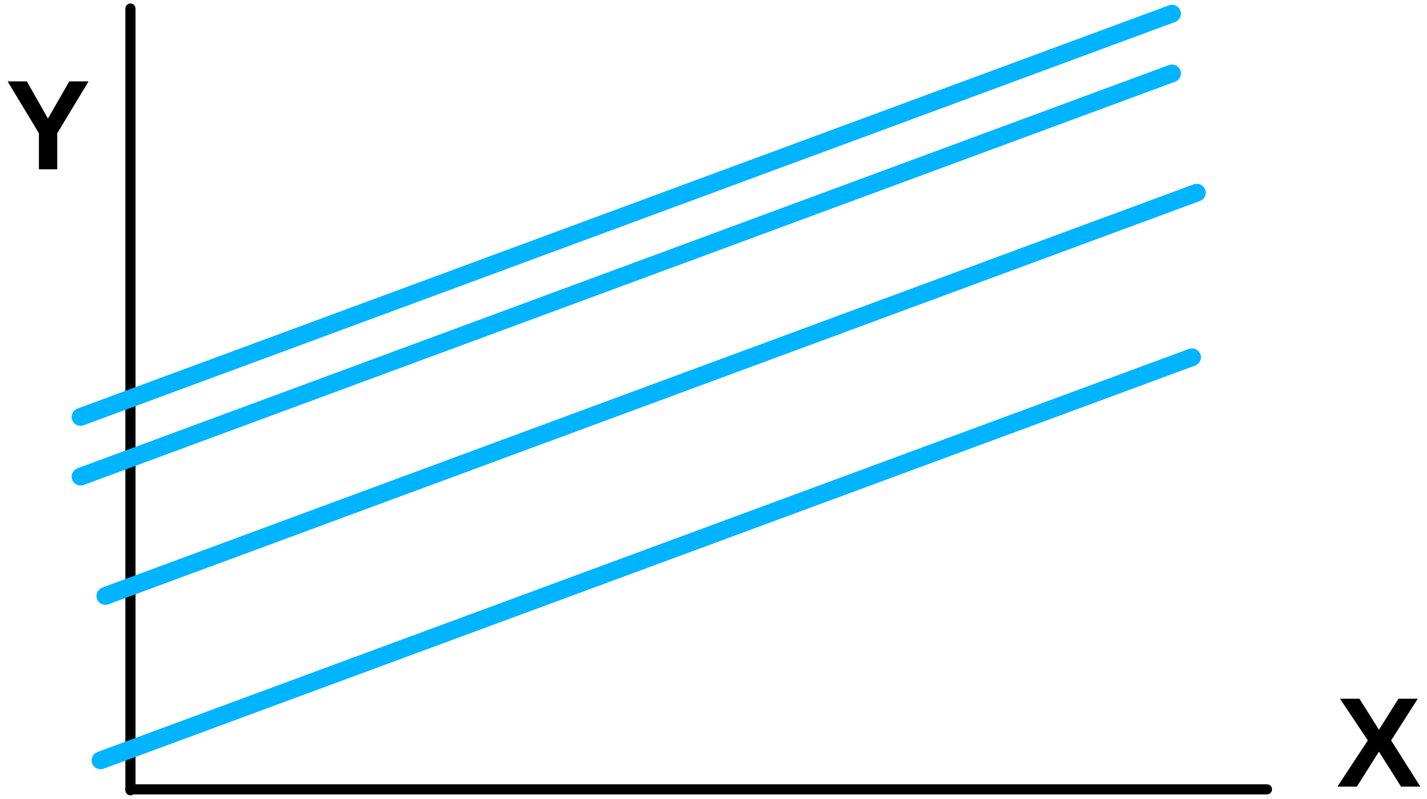
- There are actually two conceptual approaches,
  - Multisource regression - adding class variables to a regression
  - Analysis of Covariance - adding quantitative variables to an ANOVA
- We will talk primarily about Multisource regression.

AnCova *(continued)*

- With multisource regression we start with a regression and ask, would the addition of an indicator or class variable improve the model?
- Adding a class variable to a regression gives each group it's own intercept.
- We may further ask, does each group need it's own slope? This can be fitted with an interaction of the quantitative (X) variable and the group variable.

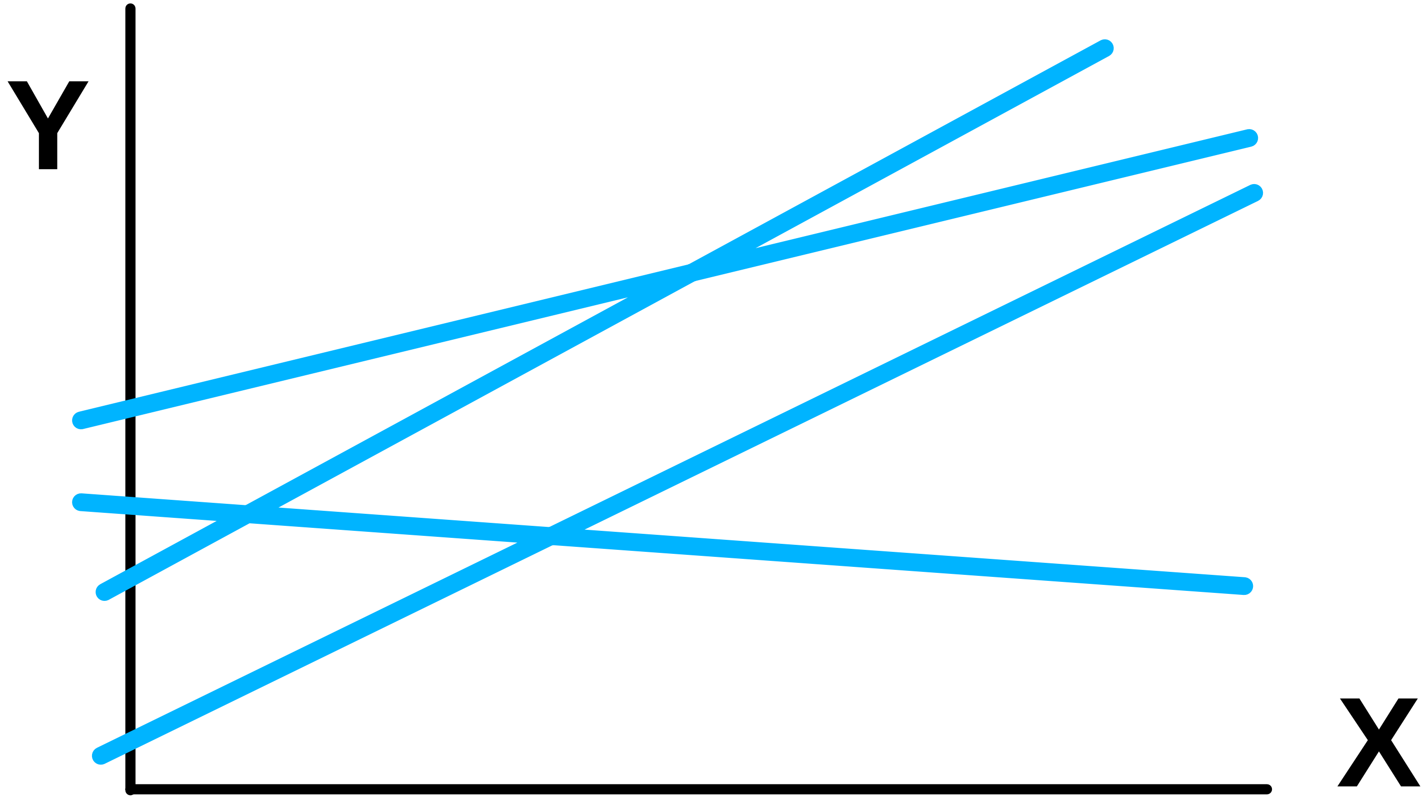
AnCova *(continued)*

- **Adding a class variable fits a separate intercept to each group.**



AnCova *(continued)*

- **Adding an interaction fits a separate slope to each group.**



AnCova *(continued)*

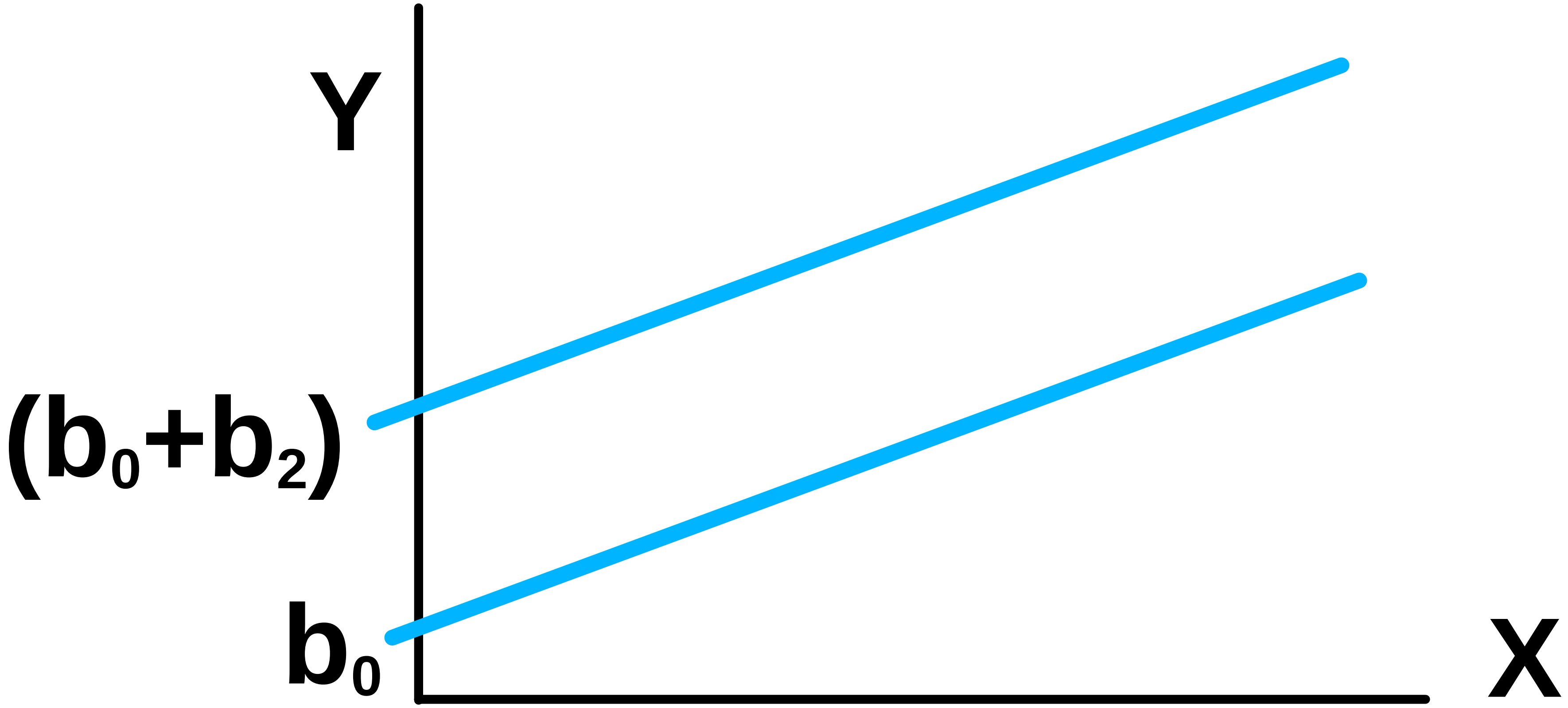
- How do they do that?
- For a simple linear regression we start with,
- $Y_i = b_0 + b_1X_{1i} + e_i$
- Now add an indicator variable. In our example we will add just one, but it could be several. We will call our indicator variable X_{2i}, but it is a variable with values of 0 or 1.

AnCova *(continued)*

- $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + e_i$
- When $X_{2i} = 0$ we get
  - $Y_i = b_0 + b_1X_{1i} + b_20 + e_i$, which reduces to
  - $Y_i = b_0 + b_1X_{1i} + e_i$, a simple linear model
- And when $X_{2i} = 1$ we have
  - $Y_i = b_0 + b_1X_{1i} + b_21 + e_i$, which simplifies to
  - $Y_i = (b_0+b_2) + b_1X_{1i} + e_i$, a simple linear model with intercept (b_0+b_2)

AnCova *(continued)*

- Two lines with intercepts of b_0 and (b_0+b_2). Note that b_2 is a difference or adjustment.



AnCova *(continued)*

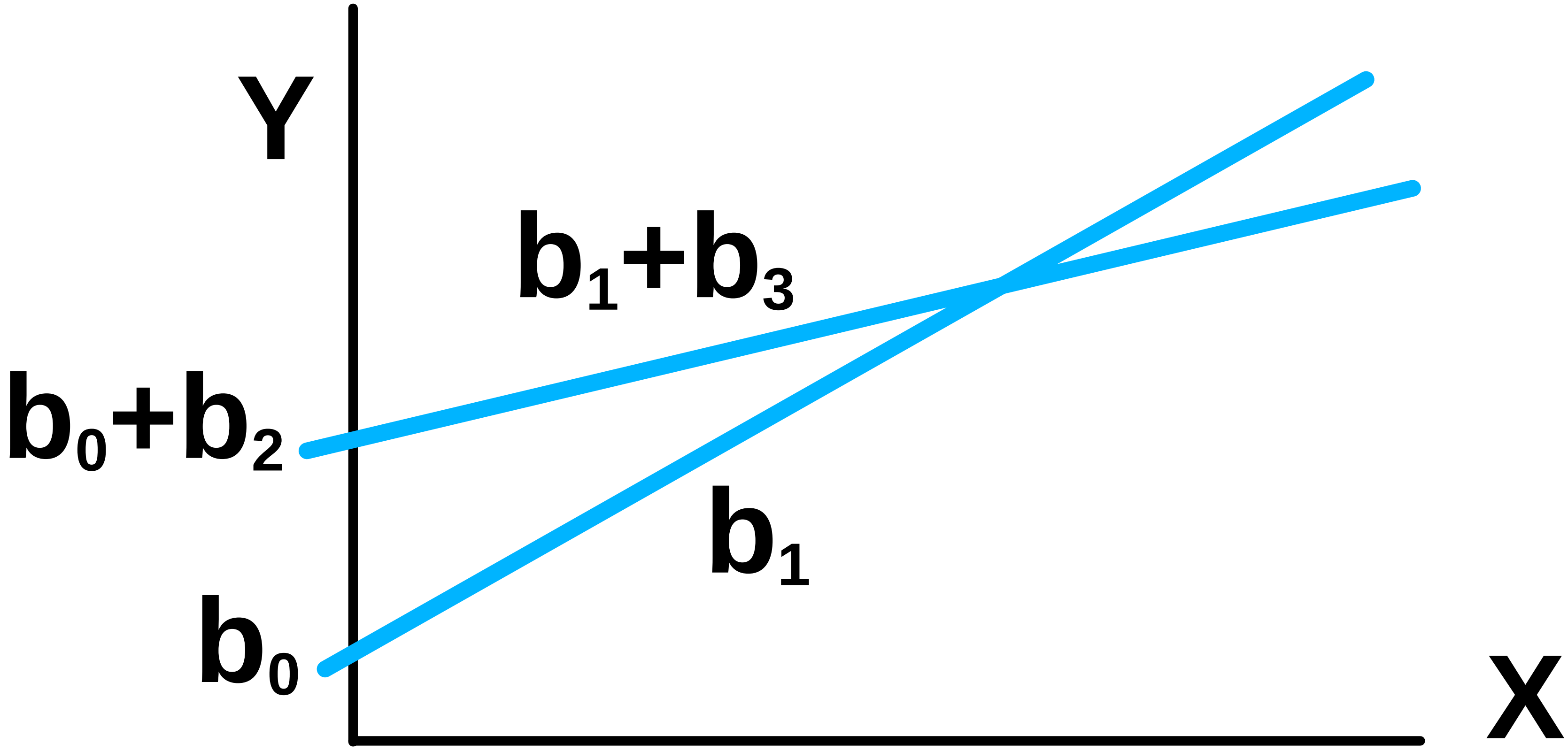
- Adding an interaction between the quantitative variable (X_{1i}) and the indicator variable (X_{2i}) will fit separate slopes.
- With just one indicator variable for two classes the model is
- $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$
-

AnCova *(continued)*

- $Y_i = b_0 + b_1X_{1i} + b_2X_{2i} + b_3X_{1i}X_{2i} + e_i$
- When $X_{2i} = 0$ we get
  - $Y_i = b_0 + b_1X_{1i} + b_20 + b_3X_{1i}0 + e_i$, which reduces to
  - $Y_i = b_0 + b_1X_{1i} + e_i$, a simple linear model
- For $X_{2i} =1$, then
  - $Y_i =b_0 +b_1X_{1i} +b_21 +b_3X_{1i}1 +e_i$, simplifies to
  - $Y_i = (b_0+b_2) + (b_1+b_3)X_{1i} + e_i$, a simple linear model with intercept (b_0+b_2) and slope equal to (b_1+b_3)

AnCova *(continued)*

- Two lines with intercepts of b_0 and (b_0+b_2). Note that b_2 is a difference or adjustment.



AnCova *(continued)*

- **Note that b_0 and b_1 are the intercept and slope for one of the lines (whichever was assigned the 0). The values of b_2 and b_3 are the intercept and slope adjustments (+ or - differences) for the second line.**
- **If these adjustments are not different from zero then the intercept and or slope are not different from each other.**
- **A third or fourth line could be fitted by adding additional indicator variables and interaction with coefficients b_4 and b_5, etc.**

AnCova *(continued)*

- **Conceptual application.**
- **Suppose we have a regression, and we would like to know if perhaps the regression is different for some groups.**
- **We would then start with a simple linear regression and test the hypotheses that the model is improved by adding separate intercepts or slopes.**

AnCova *(continued)*

- The concepts are pretty simple. The additional variables can be tested with extra SS or the General Linear Hypothesis Test (GLHT with full model and reduced model).
  - For a single indicator variable the extra SS are easy enough.
  - For several indicator variables it may be easier to do the GLHT.

AnCova *(continued)*

- **The hypotheses to be tested can be simplified to some extent if we know that for our base model we want a regression. This is multisource regression.**
  - **Start with a simple linear regression.**
  - **Test to see if separate intercepts improve the model (H_o: $\beta_2 = 0$).**
  - **Test to see if separate slopes improve the model (H_o: $\beta_3 = 0$).**

AnCova *(continued)*

- As extra SS we could test extra SS for b_2 and then the extra SS for b_3.
- The only difference is that in practice we usually start with the fullest model (2 slopes and 2 intercepts)
  - and then reduce to 1 slope and 2 intercepts if the extra SS for b_3 is not significant
  - and then to 1 slope and 1 intercept if extra SS for b_2 is not significant.
-

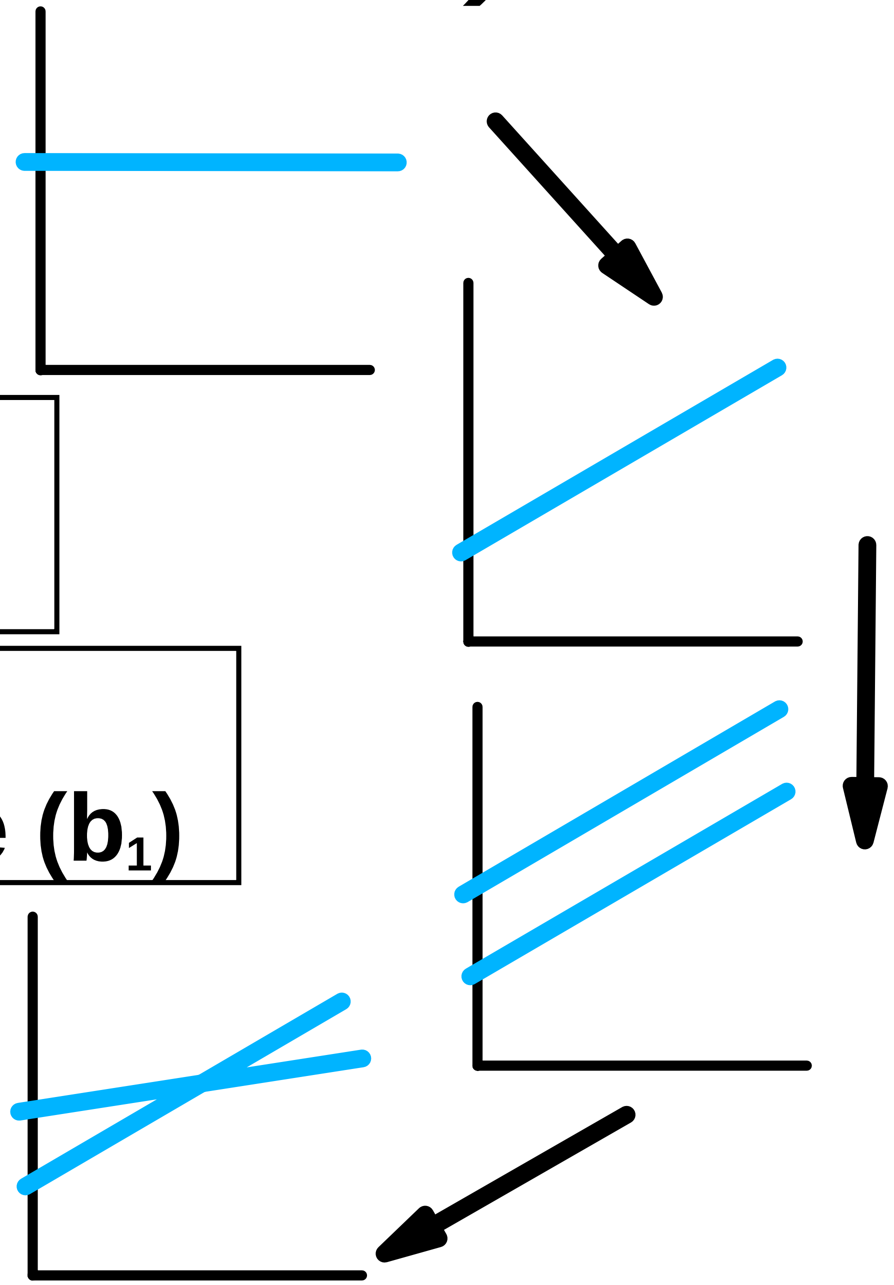
AnCova *(continued)*

- Correction factor,
  1 level = b_0 (fit mean)

add b_1,
1 level (b_0), 1 slope (b_1)

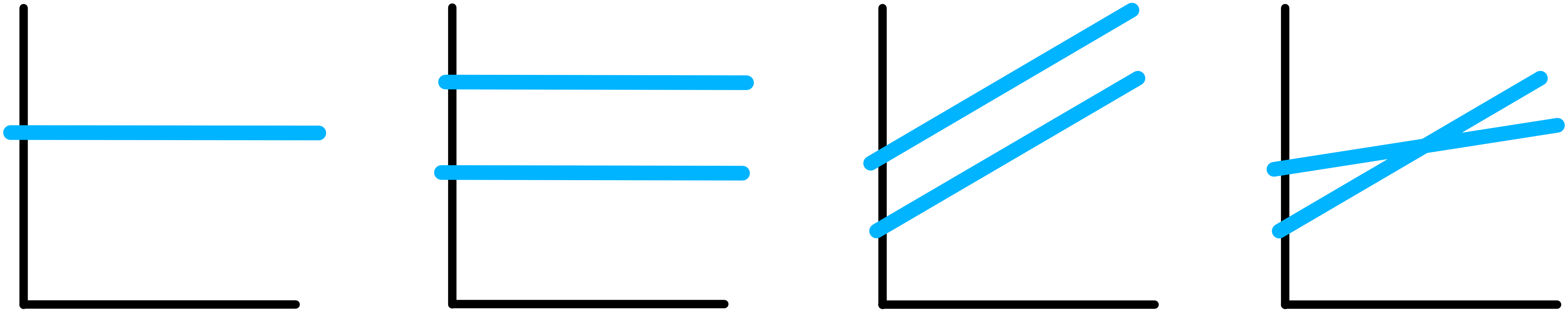
add b_2,
2 levels (b_0, b_2), 1 slope (b_1)

add b3,
2 levels (b_0, b_2) and
2 slopes (b_1 and b3)

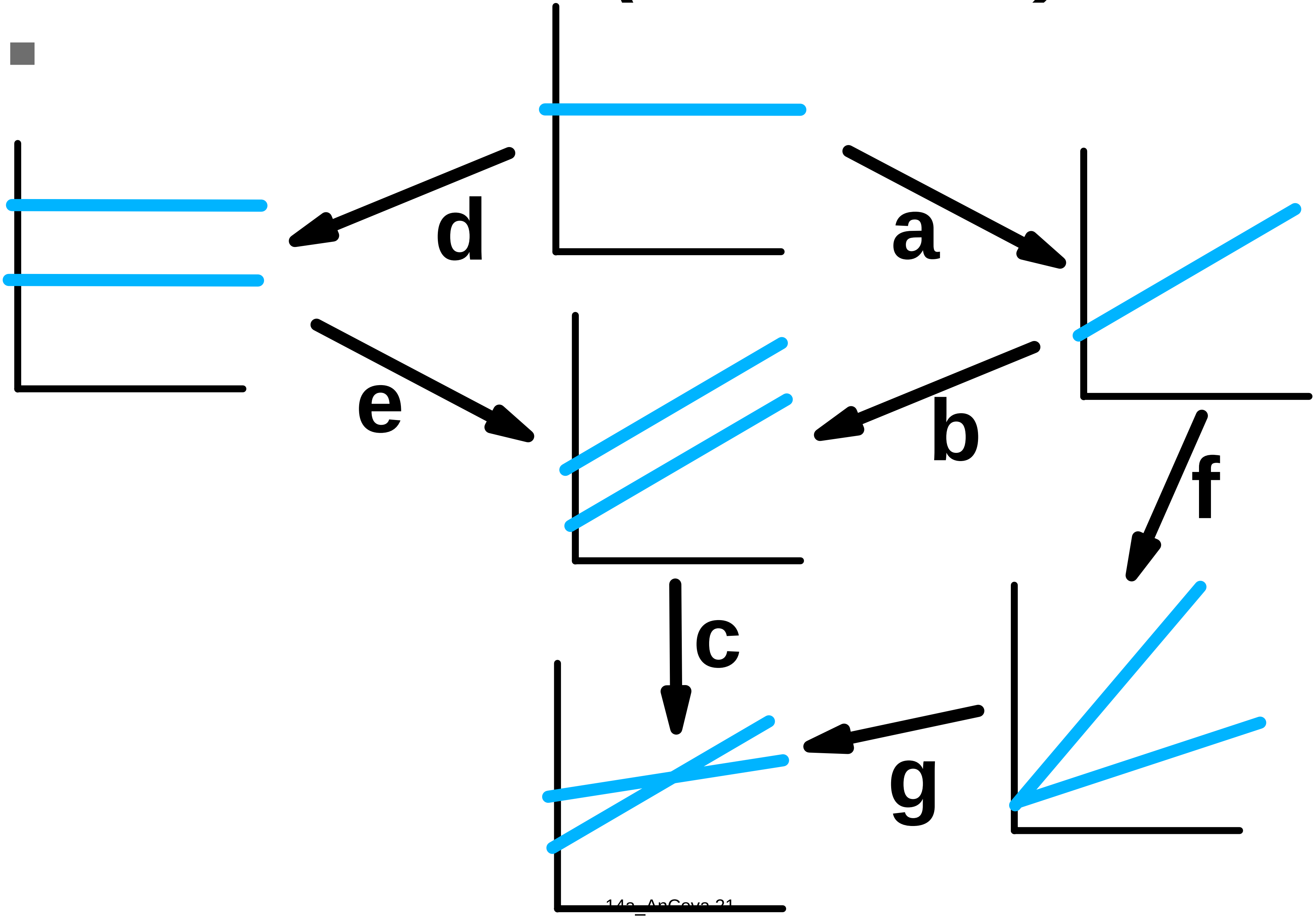


AnCova *(continued)*

- The progression on the previous page corresponds to the likely series in multisource regression,
- however, the categorical variable can be put first and the progression is the usual series for Analysis of Covariance.



AnCova *(continued)*

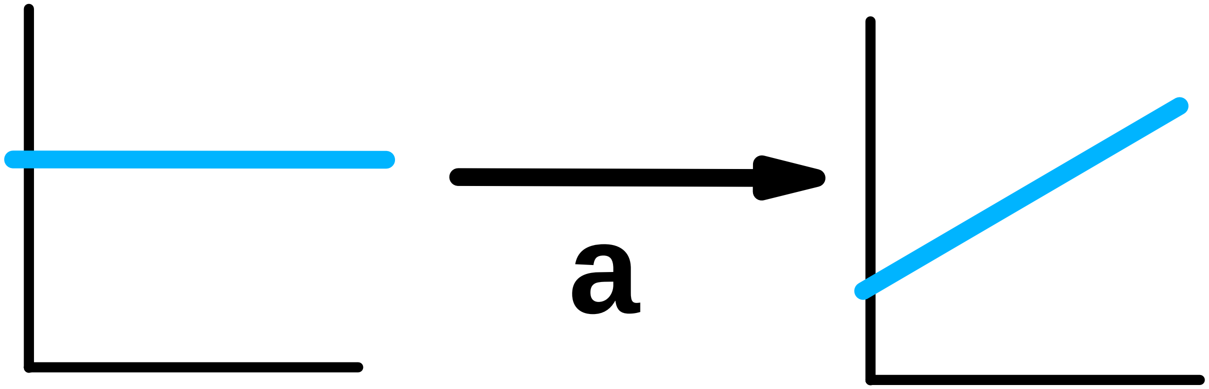


AnCova *(continued)*

- **Examine the possible full and reduced models on the previous page. What is tested in each case, what extra SS is needed for each test and what are the initial and final models for each case.**
- **How would you do these tests in SAS?**

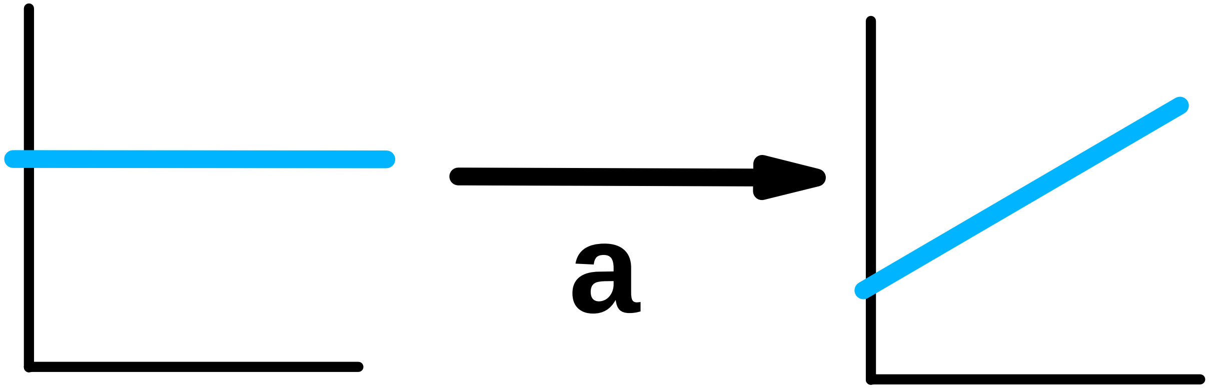
AnCova *(continued)*

- The series starts with "a". This comparison is between a model that is has only an intercept or level adjustment (the correction factor or mean, $Y_i = \bar{Y}$), with a model containing one slope ($Y_i=b_0+b_1X_i$).
-



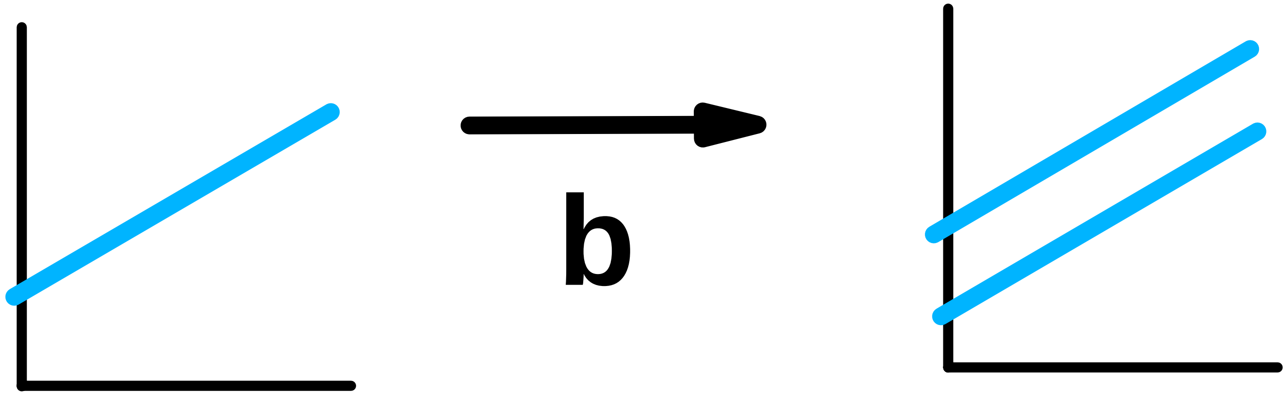
AnCova *(continued)*

- Here we are testing the addition of a slope (H_0:β_1=0).
  - This is simply the test of the model for a simple linear regression.
  - The extra SS are $SSX_1|X_0$, or just SSX_1.



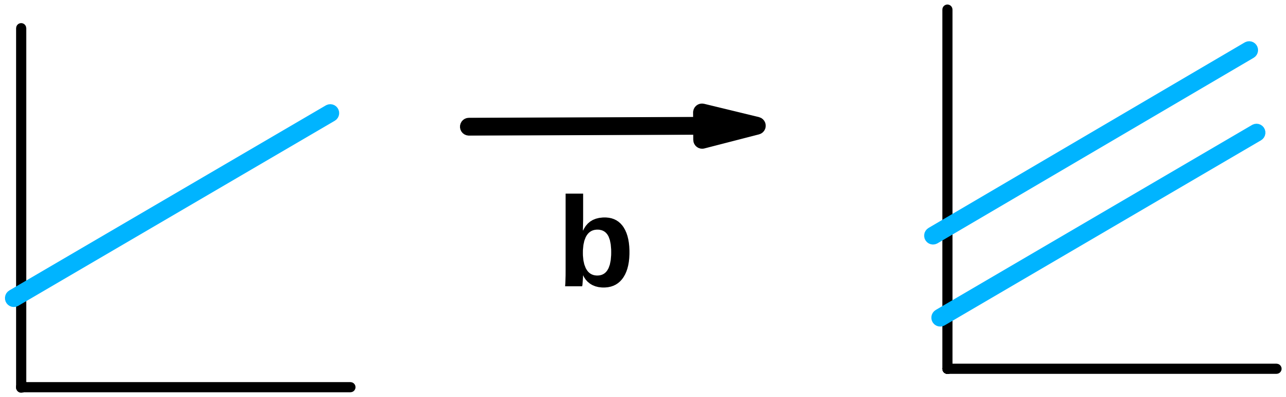
AnCova *(continued)*

- **The second test is "b", which compares a model with one intercept and one slope ($Y_i=b_0+b_1X_i$) with a model containing one slope and two intercepts.**
- **The full model is ($Y_i=b_0+b_1X_{1i}+b_2X_{2i}$).**



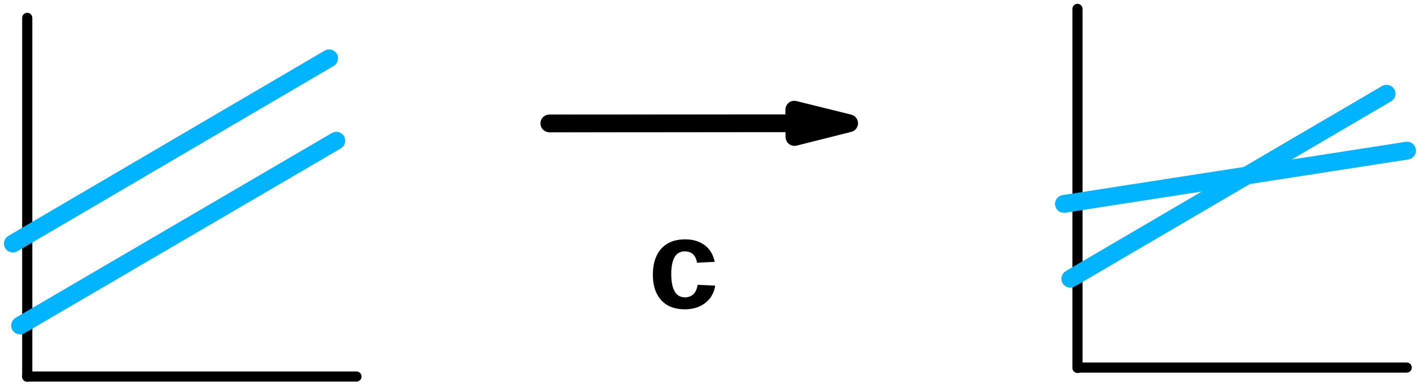
AnCova *(continued)*

- The full model is ($Y_i=b_0+b_1X_{1i}+b_2X_{2i}$).
  - Here we are testing the addition of a level adjustment, or second intercept ($H_0:\beta_2=0$).
  - The extra SS are $SSX_2 \mid X_1$.
  - In plain English this is a test that determines if a model with 2 intercepts is better than a model with 1 intercept.



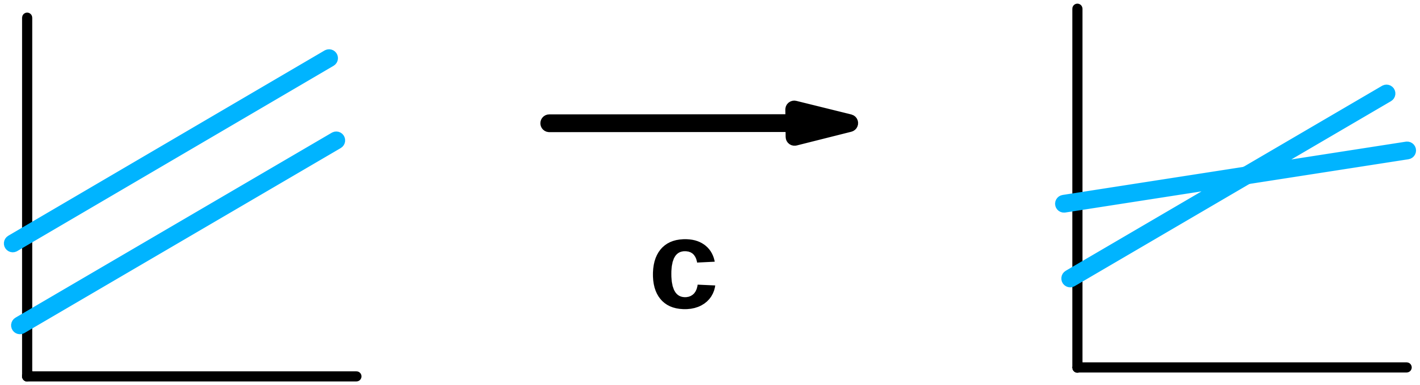
AnCova *(continued)*

- The third test is "c", which compares a model with two intercepts and one slope ($Y_i=b_0+b_1X_{1i}+b_2X_{2i}$) with a model containing two slopes and two intercepts.
- The full model is $Y_i=b_0+b_1X_{1i}+b_2X_{2i}+b_3X_{1i}X_{2i}$



AnCova *(continued)*

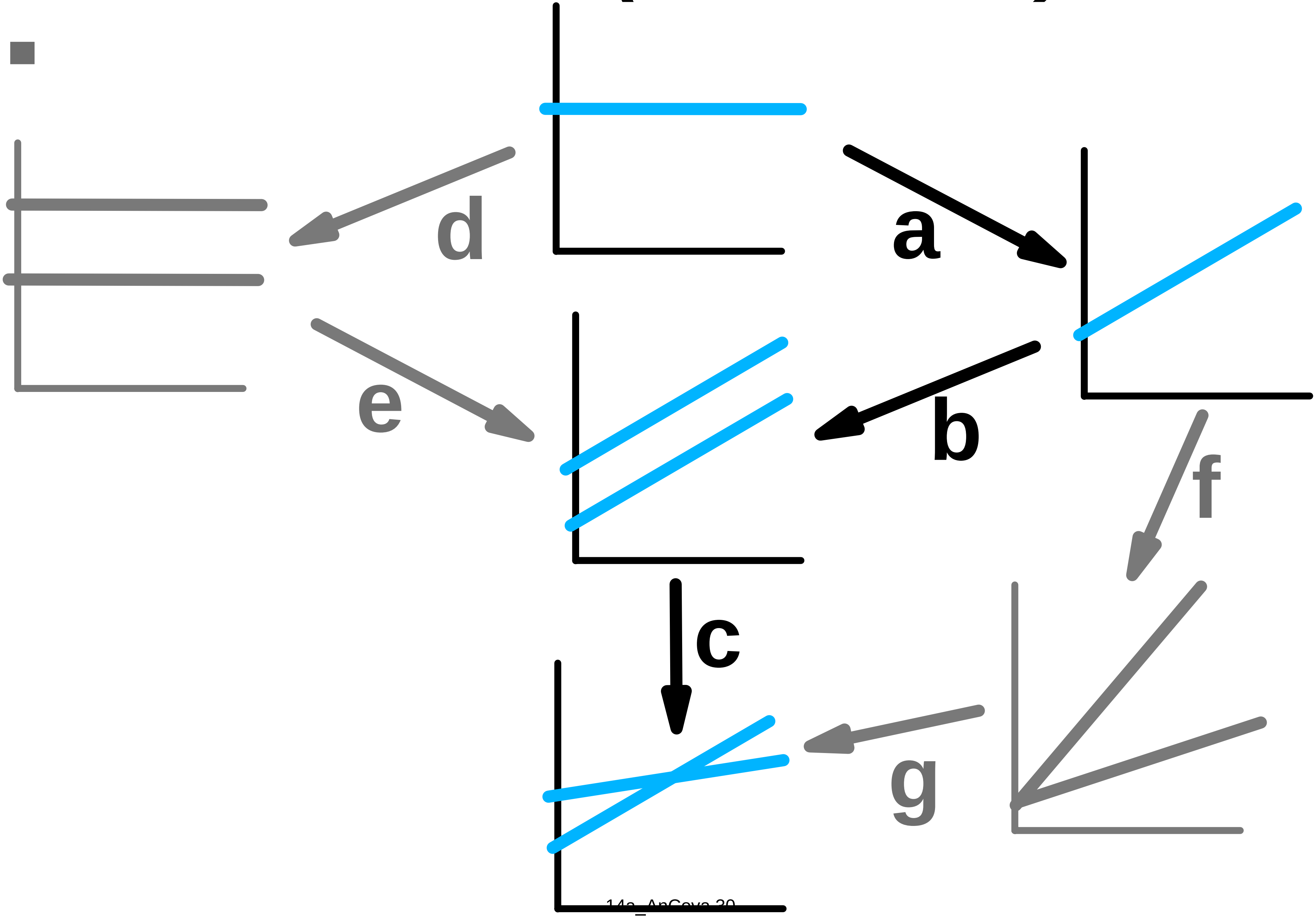
- The full model is $Y_i = b_0 + b_1 X_{1i} + b_2 X_{2i} + b_3 X_{1i} X_{2i}$
  - Here we are testing the addition of a slope adjustment, or second slope ($H_0: \beta_3 = 0$).
  - The extra SS are $SSX_1X_2 \mid X_1 \, X_2$.
  - This test determines if a model with 2 intercepts and 2 slopes is better than a model with 2 intercepts and one slope.



AnCova *(continued)*

- **The series a-b-c is the usual "Multisource regression" series.**
- **Note that the extra SS needed are:**
  - **SSX_1; SSX_2 | X_1; SSX_1X_2 | X_1, X_2.**
- **These are the Type I SS for the SAS model Y=X1 X2 X1*X2;.**
- **Interpretation usually proceeds from the bottom up; do we need 2 slopes and 2 intercepts, if not do we need 2 intercepts, if not is the SLR significant?**

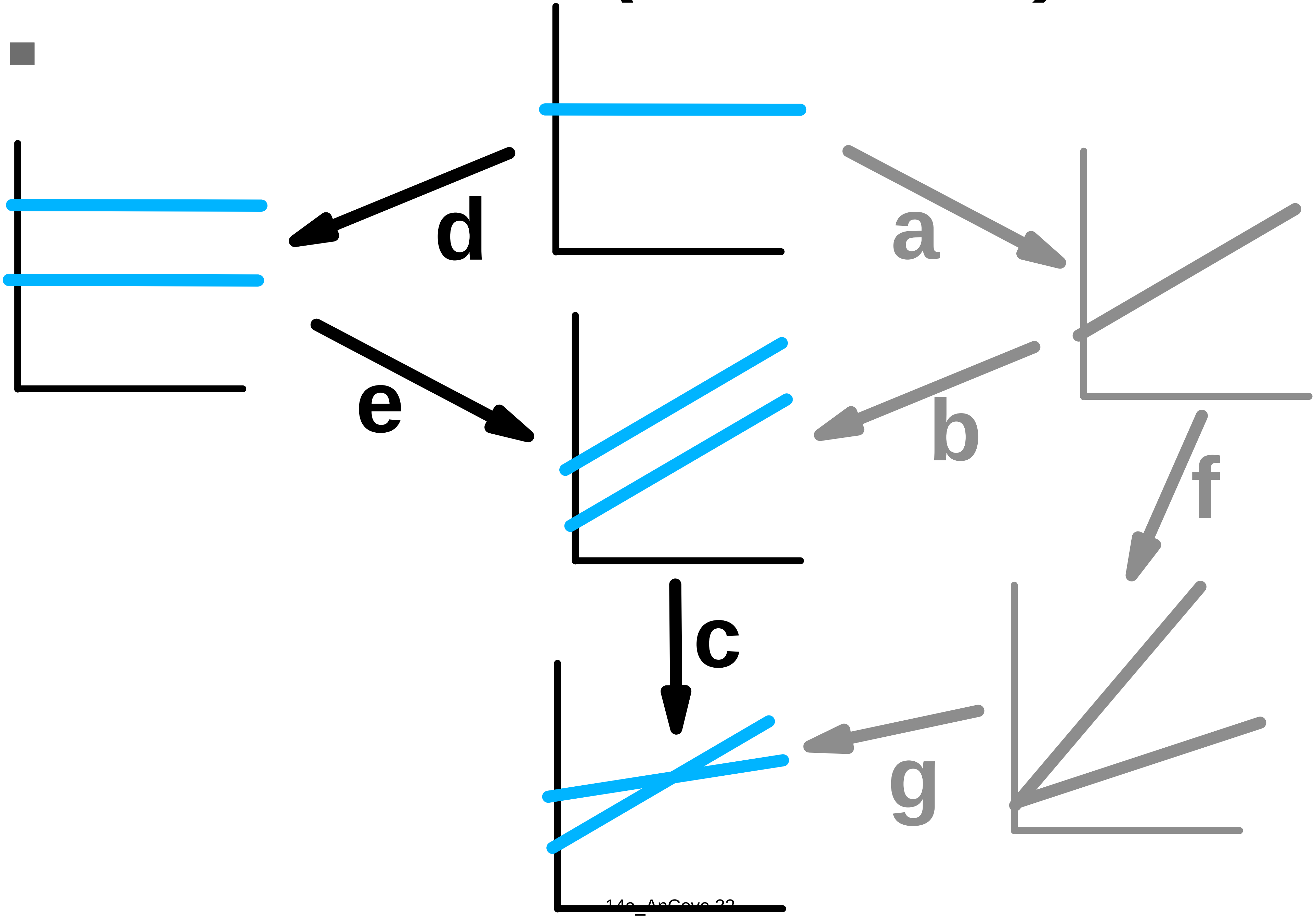
AnCova (*continued*)



AnCova *(continued)*

- **Other series of tests are possible, particularly if you do not feel you should "end up" with a regression.**
- **For designed experiments and ANOVA situations, we are often testing the addition of a slope to a model with categorical variables. The model then uses the series d-e-c.**
-

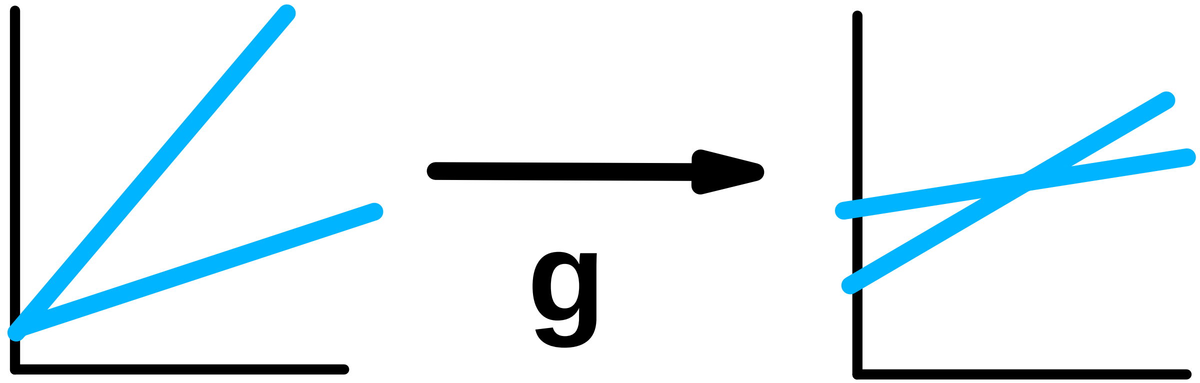
AnCova *(continued)*



AnCova *(continued)*

- One other test of occasional interest is the test denoted by the comparison "g". If you fit a model with two slopes and two intercepts, and cannot reduce to a model with two intercepts and one slope, you may wonder if a model with two slopes and one intercept is appropriate.
- Full model: $Y_i=b_0+b_1X_{1i}+b_2X_{2i}+b_3X_{1i}X_{2i}$
- Reduced model: $Y_i=b_0+b_1X_{1i}+b_3X_{1i}X_{2i}$

AnCova *(continued)*

- The extra SS for this test is SSX2 | X1, X1*X2. This is actually a test of "intercepts fitted last" and would be available in proc glm as the Type III SS for X1*X2 with the model Y=X1 X2 X1*X2;.
- 

AnCova *(continued)*

- A few notes on ANCOVA.
- The slopes and intercepts fitted in the full model (2 slopes and 2 intercepts) are exactly the same slopes and intercepts that would be fitted if the models were done separately.
- However, the combined model has the advantage of having a more powerful pooled error term.

AnCova *(continued)*

- **A few notes on ANCOVA (continued)**
- **In SAS we can**
  - **fit our own indicator variable in proc reg,**
  - **or use the CLASSES statement in PROC MIXED or GLM and let SAS create the indicator variables for us.**
- **We will use MIXED in our example.**

AnCova *(continued)*

- **A few notes on ANCOVA (continued)**
- **When we include the CLASSES statement in SAS, the program assumes that we are doing an ANOVA, and by default does not print the regression coefficients. If we want these we must add the option "/solution" to the model statement. This is true for both PROC MIXED and PROC GLM.**

AnCova Example

- This example is the Atmospheric CO_2 concentration at Mauna Loa Observatory, Hawaii, from 1958 through 2001.
- Contributors are: C. D. Keeling, T. P. Whorf (and the carbon dioxide research group), Scripps Inst. of Oceanography, Univ. of California, La Jolla, California
- The data set has monthly values of CO_2 measured as parts per million of volume (ppmv).

AnCova Example (*continued*)

- **Examine the scatter plot.**

```
            |                                                                                          C
            |                                                                                      B C D
        370 +                                                                                    A C D C
            |                                                                                    C C C B
            |                                                                                  B C B B
            |                                                                                C D D B
            |                                                                              C C C A
            |                                                                            C C C A
        360 +                                                                      A B C B D C B
            |                                                                      B B B D B
            |                                                                    C B C C C
            |                                                                A C D C B B
            |                                                                B E C B C B
            |                                                              C D B B B
        350 +                                                            C B C B
            |                                                          C B E B
            |                                                        C C D B
            |                                                      C D D C
            |                                                  B D D C B
            |                                                B B B C B
        340 +                                              B C D D B
            |                                            C C D B B
            |                                          C B D C B
            |                                        C B D A
            |                                    B C C D C B
            |                                  D C C C C
        330 +                                B C D D B
            |                            C C E D A B A
            |                          D C E C A B
            |                      B C B C B B
            |                    C D E D C B
            |            A C B E D C B B
        320 +        B C D B B C C C B
            |    A C C C C D C C B
            |    B C D D C C B A
            |    C D A B A
            |    B B B
            ---+---------+---------+---------+---------+---------+---------+---------+---------+---------+
               0         5        10        15        20        25        30        35        40        45
                                                          time
NOTE: 7 obs had missing values.
```

AnCova Example *(continued)*

- The scatter plot shows a clear increasing trend. This will be fitted in proc mixed because I want to use type 1 SS more than type 2 or 3.

```
proc mixed data=maunaloa; classes month;
  title2 'Basic Analysis of Covariance using PROC MIXED';
  model co2level = time month time*month / htype=1 3 DDFM=Satterthwaite;
run;
```

- We want to know if the apparent trend is significant and we want to know if the same trend is apparent for each month.

AnCova Example *(continued)*

- **Refer to handout.**
- **We will start with the full model results, and see if we need separate slopes (test MSX1*X2 | X1 X2). Note that in this case we have 12 months, so that actually we need 11 variables to represent the 12 months, though I have only included "X2" to represent them above. Also note that since there are 11 d.f. we should take care to note that we test MS, not SS.**

AnCova Example *(continued)*

- So the first test is of MSX1*X2 | X1 X2. If not significant we do not need separate slopes. We will then want to know if we need separate intercepts (test MSX2 | X1). If not we reduce to a simple linear regression (MSX1).
- These are the TYPE I SS tests in SAS.

```
proc mixed data=maunaloa; classes month;
  title2 'Basic Analysis of Covariance using PROC MIXED';
  model co2level = time month time*month / htype=1 3 DDFM=Satterthwaite;
run;
```

AnCova Example *(continued)*

- relevant PROC MIXED output

```
Type 1 Tests of Fixed Effects
               Num     Den
Effect          DF      DF    F Value    Pr > F
time             1     497    42276.8    <.0001
month           11     497      45.54    <.0001
time*month      11     497       0.26    0.9927
```

- Note 11 d.f. for months and month interactions.
- Also note that the month interaction is not significant.

AnCova Example *(continued)*

- **So we do not need a separate line for each month.**
- **Do we need to consider months at all?**
- **We could fit a reduced model leaving months off, but since we used type I SS the month term is not adjusted for the interaction term and leaving off the interaction term would only make a small difference in the error term.**

AnCova Example *(continued)*

- Therefore, we can interpret the month term without refitting the model.

```
Type 1 Tests of Fixed Effects
               Num     Den
Effect          DF      DF    F Value    Pr > F
time             1     497    42276.8    <.0001
month           11     497      45.54    <.0001
time*month      11     497       0.26    0.9927
```

- We see that months are significant, so in some manner the months have significantly different CO_2 levels.

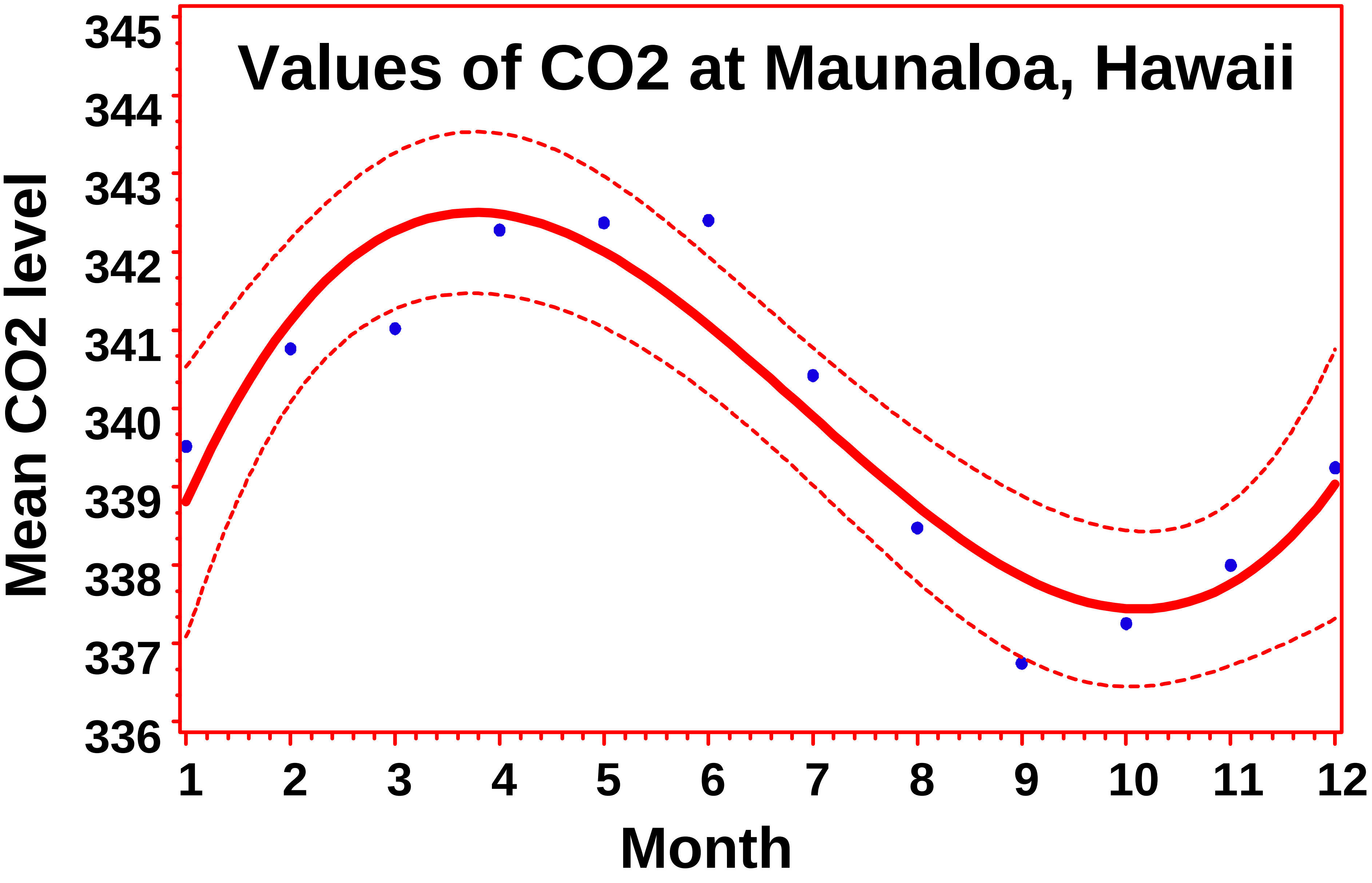
AnCova Example *(continued)*

- In order to examine the month differences I calculated a mean CO2 level for each month and plotted these means (next page).
- I also fitted a cubic model to the means.

AnCova Example *(continued)*

```
                             Plot of mean*mo.  Legend: A = 1 obs, B = 2 obs, etc.
mean |
 343 +
     |
     |
     |
     |                             A        A        A
     |
 342 +
     |
     |
     |
     |
     |
 341 +                    A
     |           A
     |
     |                                                        A
     |
     |
 340 +
     |
     |
     |  A
     |
     |                                                                                                     A
 339 +
     |
     |
     |                                                                 A
     |
     |
 338 +                                                                                            A
     |
     |
     |
     |                                                                                   A
     |
 337 +
     |
     |                                                                          A
     |
     |
     |
 336 +
     |
     ---+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--
        1        2        3        4        5        6        7        8        9       10       11       12
                                                         mo
```

AnCova Example (*continued*)



AnCova Example *(continued)*

- The observed pattern was a reasonable and logical result.
  - Higher at the end of the winter
  - Decreases rapidly over the summer months
  - Starts to increase again at the end of summer
- The pattern also seem to be quite well fitted by a cubic pattern.

AnCova Example *(continued)*

- This suggests that we could simplify our original model.
- 1) We know that we do not need a month by year interaction.
- 2) Results suggest that we can use a cubic model to fit months (only 3 d.f.) instead of the 11 used as a class variable.
- - we will fit a quartic polynomial in order to test for a slightly larger model than the cubic.

AnCova Example *(continued)*

- **I will also fit a quadratic to years to test for significant curvature.**
-
- **The model was refitted.**
- **Type I SS were used because**
- **1) they are appropriate for ANCOVA as we have seen, testing for years first and months later.**
- **2) They are appropriate for polynomials.**

AnCova Example *(continued)*

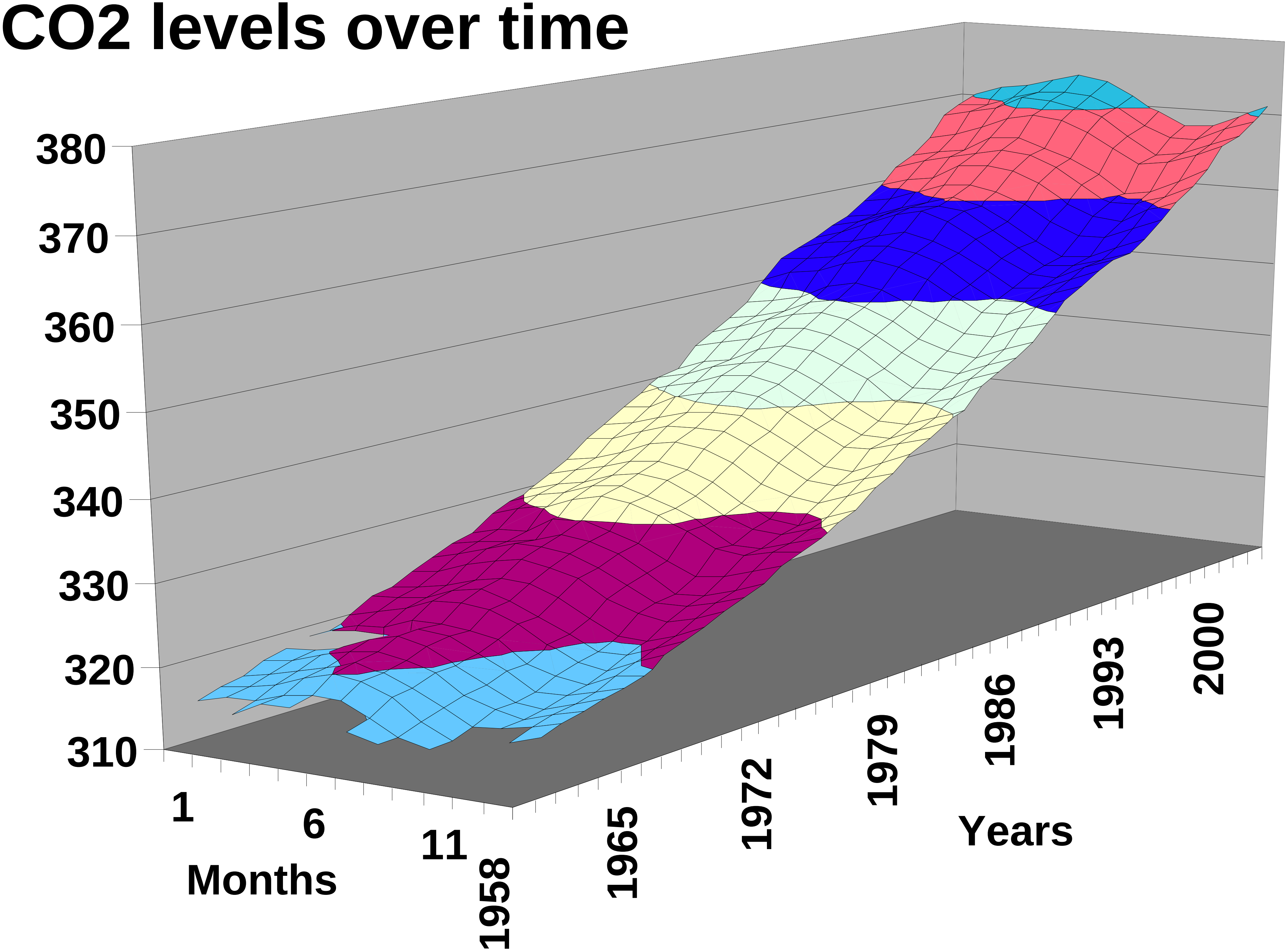
- The refitted model gave these results.

```
         Type 1 Tests of Fixed Effects
                Num     Den
Effect           DF      DF    F Value    Pr > F
time              1     514     197913    <.0001
time*time         1     514    1963.89    <.0001
mo                1     514     615.64    <.0001
mo*mo             1     514     269.43    <.0001
mo*mo*mo          1     514    1076.47    <.0001
mo*mo*mo*mo       1     514     245.26    <.0001
```

AnCova Example *(continued)*

- **The final model fits a quadratic to years and a quartic to months. This forms a response surface (next page).**
-

CO2 levels over time



ANCOVA observations

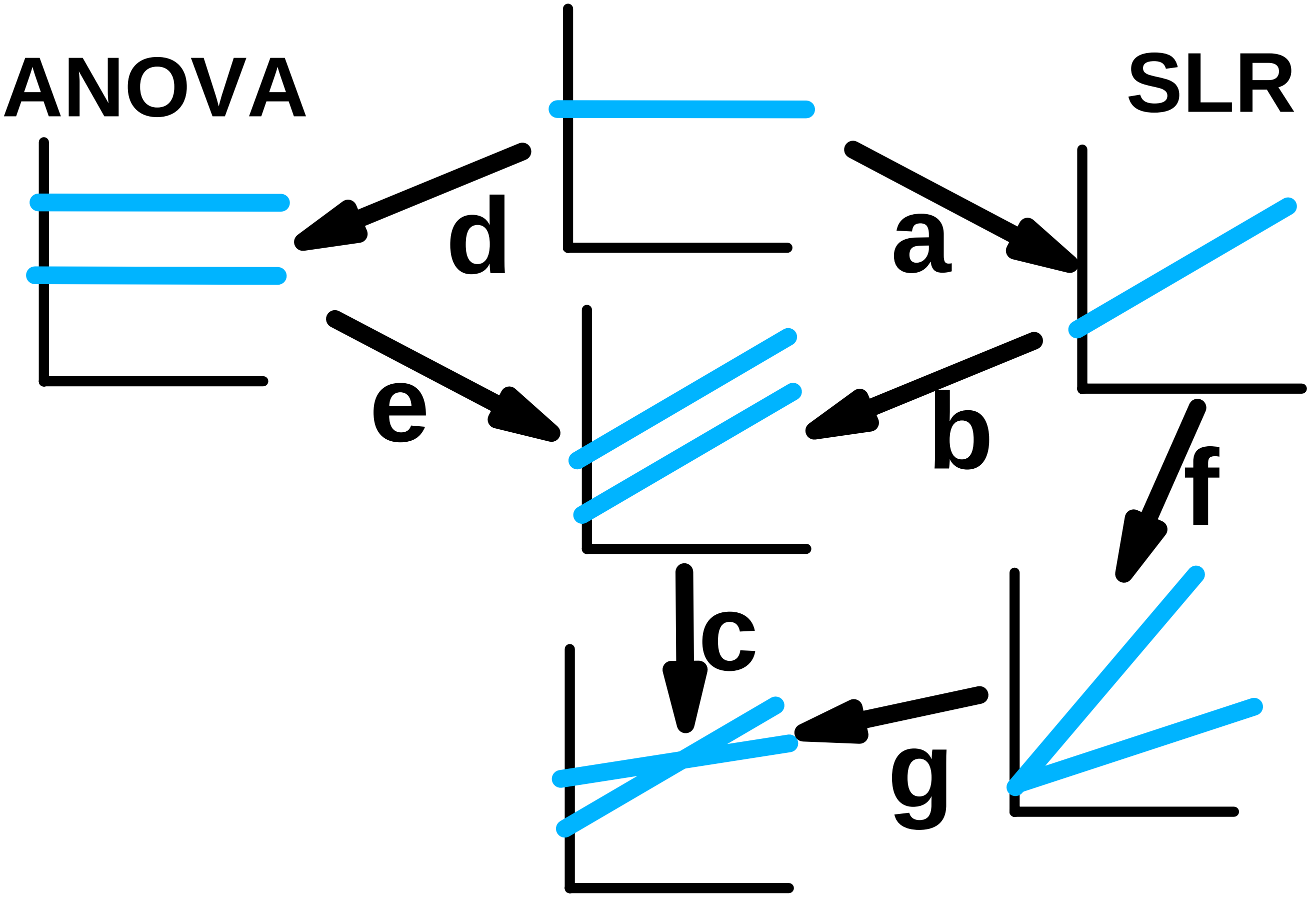
- A few final notes.
- The solution provide with ANCOVA is the correct model for prediction, even though it is fully adjusted and our tests of hypothesis are not.
-

Summary

- Analysis of Covariance is the combination of quantitative variables and categorical variables.
- Multisource regression is the expression of an ANCOVA that reduces to a SLR.
- Analyses that culminate in ANOVAs will be discussed at the end of the course. This type is where the term "Analysis of Covariance" was developed.

Summary *(continued)*

- You should be able to examine the graph below and determine what is tested and what Extra SS are tested.



AnCova *(continued)*

- The model used for testing the differences between the slopes and intercepts is correct for testing, but will yield differences and standard errors of differences, not the actual values.
- These can be obtained in SAS, but this model tests intercepts jointly against zero and slopes jointly against zero.
- This is not likely to be the test we want.