

Applied Regression Analysis

Week 6

1. Homework week 5: highlights
2. Dummy variables
3. Statistical interaction
4. Comparing two straight line regression equations
 - Method 1: fitting separate models
 - Method 2: fitting a single model
5. Homework
6. Homework week 6: highlights

Stanley Lemeshow, Professor of Biostatistics
College of Public Health, The Ohio State University



THE OHIO STATE UNIVERSITY

A dummy variable is any variable in a regression equation that takes on a finite number of values

- used to indicate categories of a nominal scaled variable**

The term “Dummy” simply means that the actual values used (e.g., 0,1 or +1, 0, -1) do not describe a meaningful measurement level of the variable

- instead they only “indicate” the categories of interest**

These variables allow us to broaden the scope of regression analysis to include ANOVA, ANCOVA, Discriminant Analysis, and these variables will be widely used in logistic regression analysis.

Examples:

2 categories
1 dummy variable

$$\left\{ \begin{array}{l} x_1 = \begin{cases} 1 & \text{if patient received Drug A} \\ 0 & \text{otherwise} \end{cases} \\ x_2 = \begin{cases} +1 & \text{if female} \\ -1 & \text{if male} \end{cases} \end{array} \right.$$

3 categories
2 dummy variables

$$\left\{ \begin{array}{cc} x_3 & x_4 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{array} \right. \left\{ \begin{array}{l} \text{if treated at hospital A} \\ \text{if treated at hospital B} \\ \text{if treated at hospital C} \end{array} \right.$$

Another way to code a 3 category variable is:

Instead of $\begin{pmatrix} 0 & 0 \end{pmatrix}$ \longrightarrow

$$\begin{array}{cc} x_3 & x_4 \\ -1 & -1 \\ 1 & 0 \\ 0 & 1 \end{array} \left\{ \begin{array}{l} \text{if treated at hospital A} \\ \text{if treated at hospital B} \\ \text{if treated at hospital C} \end{array} \right.$$

Also, you can mix up the placement of the +1's, 0's and -1's.

Important note:

Different dummy variable definitions will lead to coefficients that have different meaning. However, the test procedures involve the same null hypothesis and test statistics.

General Rule

If the nominal independent variable has k categories, then one must define exactly $k-1$ dummy variables to index these categories, provided that the regression model contains a constant term (i.e., β_0).

If the regression model does not contain a constant term, then k dummy variables are needed to index the k categories of interest.

If k dummy variables are used to describe a nominal variable with k categories in a model containing a constant term, then all the coefficients in the model cannot be uniquely estimated.

Let x_1 and x_2 be two independent variables

Let y represent the dependent variable

Q: How do x_1 and x_2 “interact” to affect y ?

There is “no statistical interaction” between x_1 and x_2 if the relationship between x_1 and y is the “same” regardless of the value of x_2 and the relationship between x_2 and y is the “same” regardless of the value of x_1 .





Example

let Y = height a cake rises

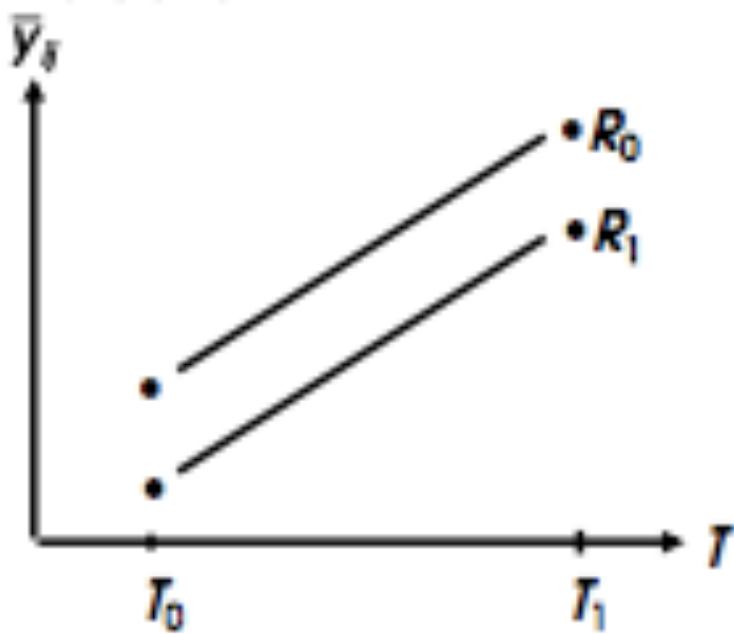
T = temperature \leftarrow two levels: T_0 and T_1

R = recipe \leftarrow two levels: R_0 and R_1

Of interest is to determine how the two independent variables, T and R , jointly affect the height the cake rises.

		Temperature	
		T_0	T_1
Recipe	R_0	 \bar{y}_{11}	 \bar{y}_{12}
	R_1	 \bar{y}_{21}	 \bar{y}_{22}

No interaction:



Here we have no interaction

i.e., The rate of change in Y as a function of temperature is the same regardless of which recipe is used.

i.e., The relationship between Y and T does not, in any way, depend on R .

note: We are not saying that Y and R are unrelated. We are saying that the relationship between Y and T is independent of the relationship between Y and R .

- When this is the case, we say that there is no $T \times R$

Interaction effect.

- This means that we can investigate the effects of T and R on Y independently of one another.

This relationship can be quantified with

$$\mu_{Y|T,R} = \beta_0 + \beta_1 T + \beta_2 R$$

$$\text{where } T = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix} \quad R = \begin{Bmatrix} 0 \\ 1 \end{Bmatrix}$$

Here, the change in Y for a one-unit change in $T = \beta_1$,
Regardless of the level of R .

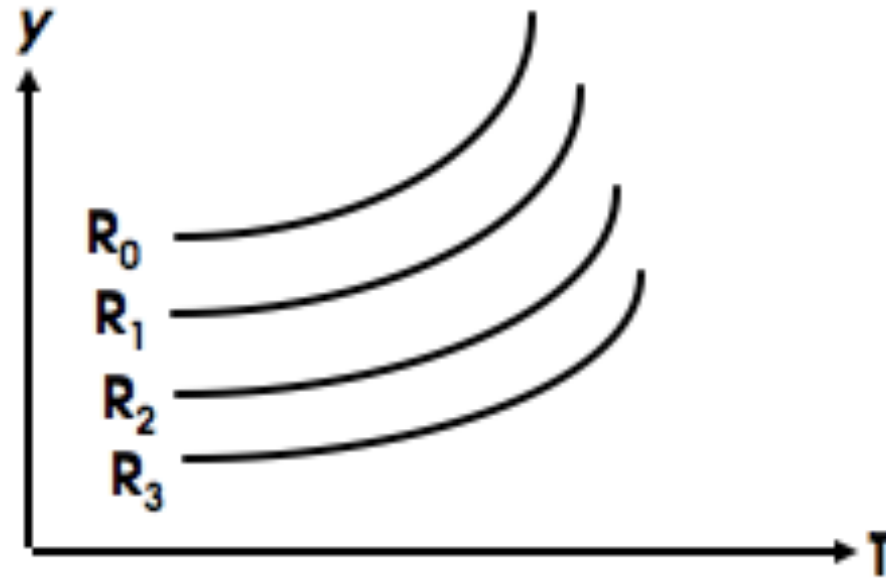
- Changing the level of R has the effect of shifting the straight line either up or down without affecting the value of the slope, β_1 .

$$\text{i.e., } \mu_{Y|T,R_0} = (\beta_0 + \beta_2 R_0) + \beta_1 T = \beta'_0 + \beta_1 T$$

$$\mu_{Y|T,R_1} = (\beta_0 + \beta_2 R_1) + \beta_1 T = \beta''_0 + \beta_1 T$$

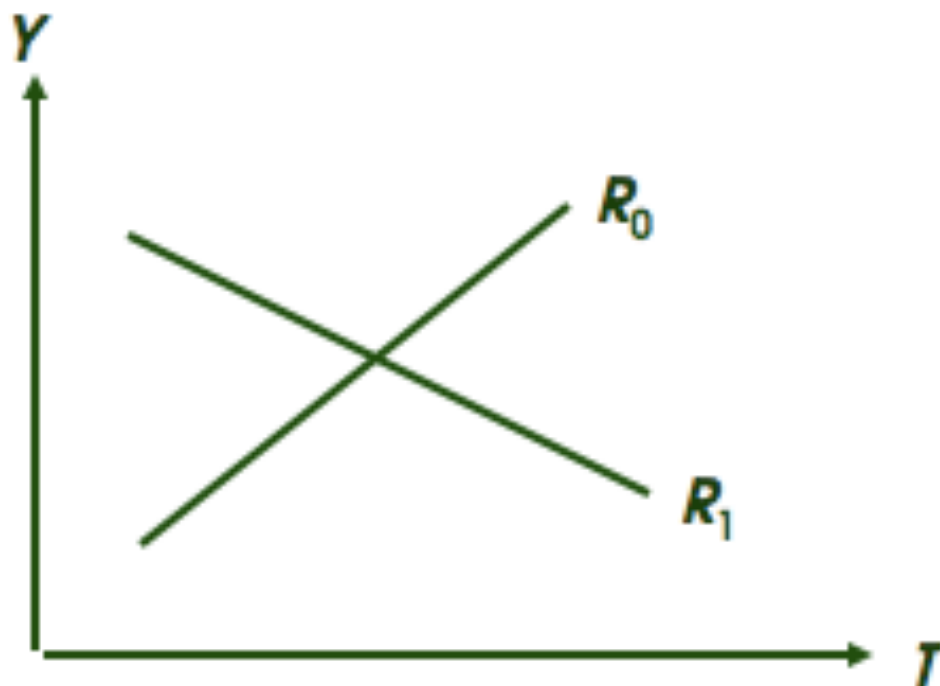
Hence, no interaction is synonymous with “parallelism”

Another example of no interaction is the following:



These response curves (that are non-linear) all have the same general shape, differing from each other only by an additive constant independent of T .

Now, consider the following alternative graph:



Here the relationship between Y and T depends upon R .

- Y decreases with Temperature with Recipe 1
- Y increases with Temperature with Recipe 0.
- Hence there is an $R \times T$ interaction.

We can model this by: $\mu_{Y|R,T} = \beta_0 + \beta_1 T + \beta_2 R + \beta_{12} TR$

$$\mu_{Y|T,R} = \beta_0 + \beta_1 T + \beta_2 R + \beta_{12} TR$$

Here, the change in the mean value of Y for a 1-unit change in T is equal to

$$\begin{aligned} \mu_{Y|T+1,R} - \mu_{Y|T,R} &= \beta_0 + \beta_1 (T+1) + \beta_2 R + \beta_{12} (T+1)R \\ &\quad - (\beta_0 + \beta_1 T + \beta_2 R + \beta_{12} TR) \\ &= \beta_1 + \beta_{12} R \end{aligned}$$

which clearly depends upon the value of R .

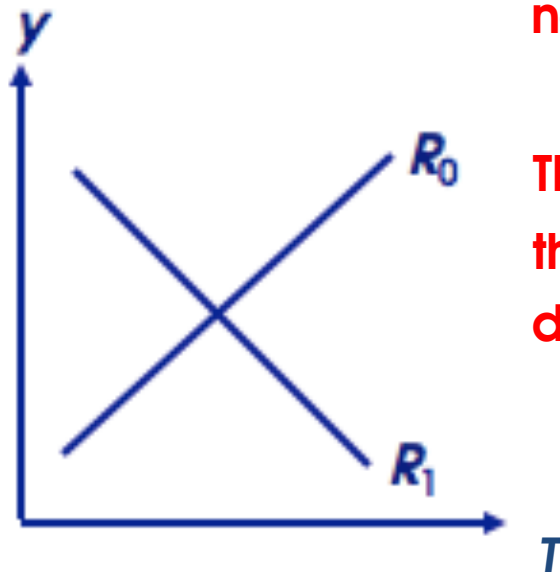
The introduction of a product term such as $\beta_{12} TR$ into the model is one way to account for the fact that two such factors as T and R do not operate independently of one another.

In our example, when $R = R_0$, the model can be written as:

$$\mu_{Y|T,R_0} = (\beta_0 + \beta_2 R_0) + (\beta_1 + \beta_{12} R_0)T$$

and, when $R = R_1$,

$$\mu_{Y|T,R_1} = (\beta_0 + \beta_2 R_1) + (\beta_1 + \beta_{12} R_1)T$$

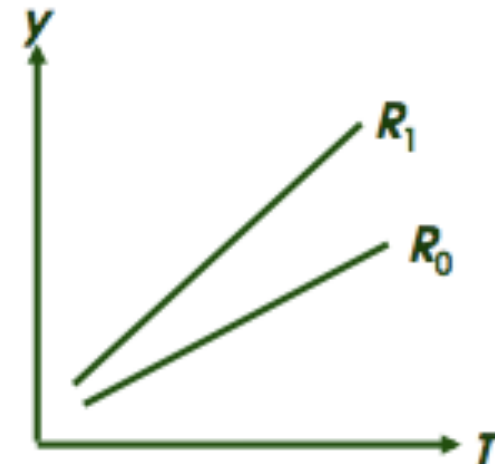


note: the linear effect of T at R_0 is positive $= \beta_1 + \beta_{12} R_0$

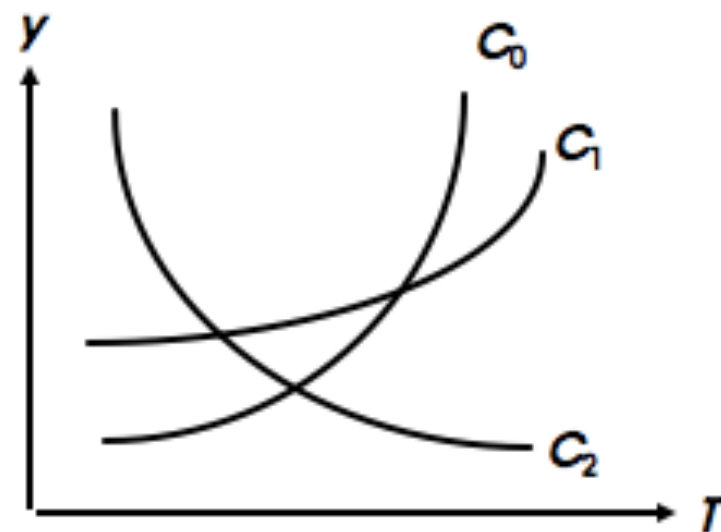
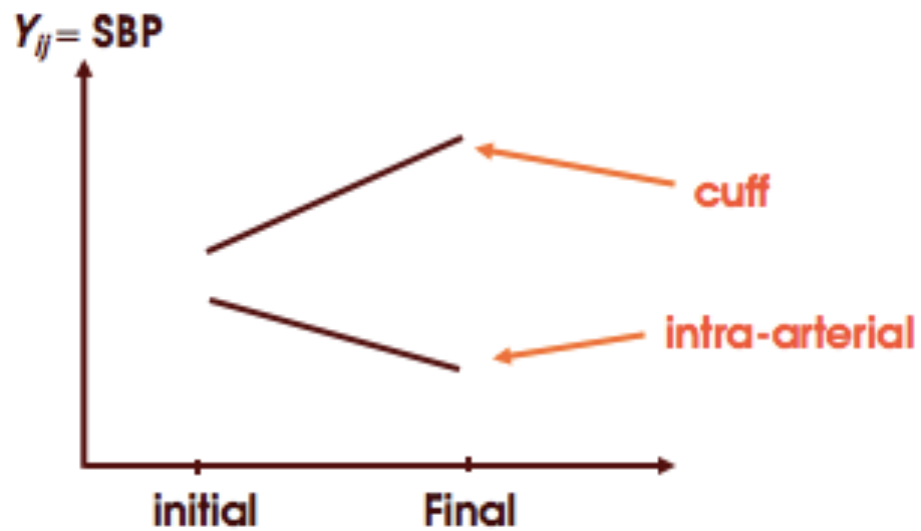
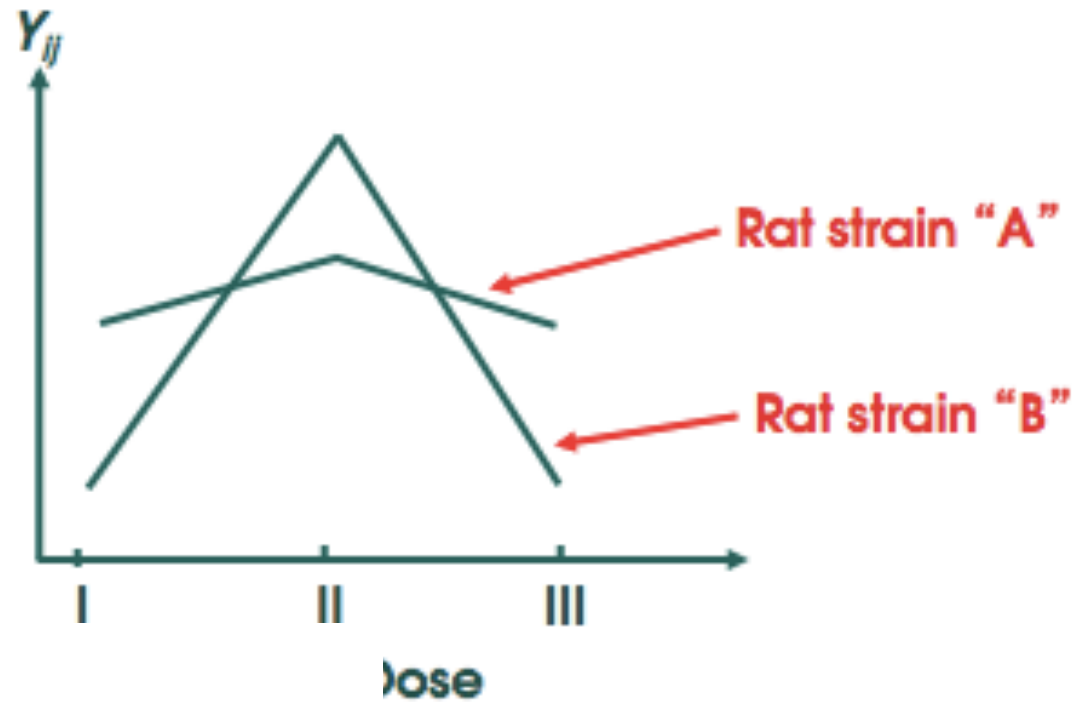
the linear effect of T at R_1 is negative $= \beta_1 + \beta_{12} R_1$

This suggests that the interaction β_{12} is negative since the slope of the linear relationship between Y and T decreases as R changes from R_0 to R_1 .

β_{12} positive could occur in a situation as follows:



Other examples of interaction:



Suppose we want to compare males and females w.r.t. their separate straight line regressions of SBP(y) on AGE (x).

Let $n_m = \#(x, y)$ pairs for the males

$n_f = \#(x, y)$ pairs for the females

Two Methods:

Method 1: fit two separate regression equations:

$$\left. \begin{aligned} y_m &= \beta_{0m} + \beta_{1m}x + \varepsilon \\ y_f &= \beta_{0f} + \beta_{1f}x + \varepsilon \end{aligned} \right\} \text{Treats male and female data separately}$$

Then use statistical methods to compare β_{1m} and β_{1f}
or to compare β_{0m} and β_{0f}

Method 2: Define

$$z = \begin{cases} 0 & \text{if male} \\ 1 & \text{if female} \end{cases}$$

for males: $(x_{1m}, y_{1m}, 0), (x_{2m}, y_{2m}, 0), \dots, (x_{n_m m}, y_{n_m m}, 0)$

for females: $(x_{1f}, y_{1f}, 1), (x_{2f}, y_{2f}, 1), \dots, (x_{n_f f}, y_{n_f f}, 1)$

Then fit a single model:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 xz$$

Note:

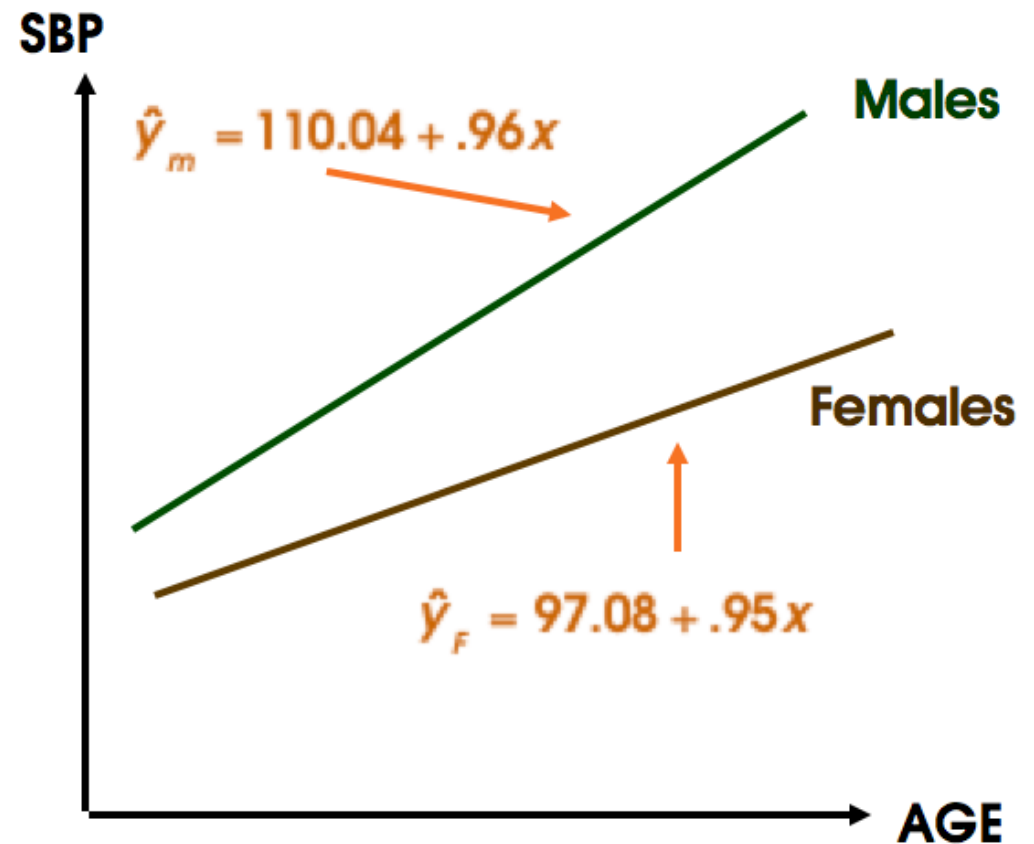
when $z = 0$, $y_m = (\beta_0) + \beta_1 x + \varepsilon = \beta_{0m} + \beta_{1m} x + \varepsilon$

$$z = 1, \quad y_f = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x + \varepsilon = \beta_{0f} + \beta_{1f} x + \varepsilon$$

Hence, this model incorporates two separate regressions within a single model and allows for different slopes (β_1 for males and $\beta_1 + \beta_3$ for females) as well as different intercepts.

EXAMPLE:

	<u>MALES</u>	<u>FEMALES</u>
n	40	29
$\hat{\beta}_0$	110.04	97.08
$\hat{\beta}_1$	0.96	0.95
\bar{x}	46.93	45.07
\bar{y}	155.15	139.86
s_x^2	221.15	242.14
$s_{y x}^2$	71.90	91.46



Here, the separate regressions are $\hat{y}_m = 110.04 + .96x$

$$\hat{y}_f = 97.08 + .95x$$

and the combined model is $\hat{y} = 110.04 + .96x - 12.96z - .012xz$

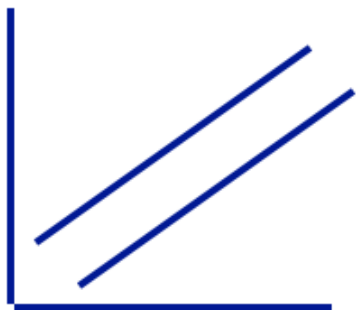
note:

when $z = 0$ (males), $\hat{y} = 110.04 + .96x$

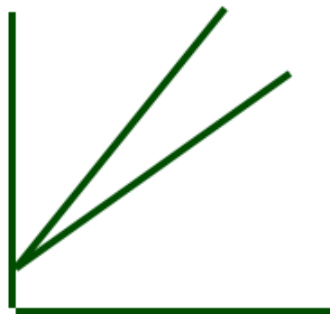
and when $z = 1$ (females), $\hat{y} = 110.04 + .96x$

$$- 12.96(1) - .012x = 97.08 + .95x$$

The appearance of the plots can be:



Parallel



Equal Intercepts



Different slopes
and intercepts



Coincident

Now, we are interested in testing

H_0 : The two regression lines are parallel.

or $H_0 : \beta_3 = 0$ (since then $\beta_{1f} = \beta_1 + \beta_3 = \beta_1 + 0 = \beta_1 = \beta_{1m}$)

This can be tested with the usual Partial F -test (or equivalent t -test) for the significance of the addition of the variable xz to the model already containing x and z .

The ANOVA table is:

Source	df	SS	MS	F
Regression (x)	1	14951.25	14951.25	121.27
Residual	67	8260.51	123.29	
Regression (x,z)	2	18009.78	9004.89	114.25
Residual	66	5201.99	78.82	
Regression (x,z,xz)	3	18010.33	6003.44	75.02
Residual	65	5201.44	80.02	

To test $H_0 : \beta_3 = 0$ we compute

$$F(xz|x,z) = \frac{SS_{\text{reg}}(x,z,xz) - SS_{\text{reg}}(x,z)}{MS_{\text{resid}}(x,z,xz)} = \frac{18010.33 - 18009.78}{80.02} = .007$$

and testing this against an $F(1,65)$ leads to the conclusion that there is no evidence that the two lines are not parallel.

To see whether the two lines coincide we test

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

To do this we use

$$F(xz, z | x) = \frac{\left[SS_{\text{reg}}(x, z, xz) - SS_{\text{reg}}(x) \right] / 2}{MS_{\text{resid}}(x, z, xz)}$$
$$= \frac{[18010.33 - 1495.25] / 2}{80.02} = 103.19$$

and, since $F_{.999}(2, 65) = 7.72$, we reject H_0 with $p < .001$.

Hence there is strong evidence that the two lines are not coincident.

- at this point the complete model could be reduced to the form: $y = \beta_0 + \beta_1 x + \beta_2 z + \varepsilon$