STAT 462

Applied Regression Analysis

7.6 - Interactions Between Quantitative Predictors

Interaction terms between quantitative predictors allow the relationship between the response and one predictor to vary with the values of another predictor. Interestingly, this provides another way to introduce curvature into a multiple linear regression model. For example, consider a model with two quantitative predictors, which we can visualize in a three-dimensional scatterplot with the response values placed vertically as usual and the predictors placed along the two horizontal axes. A multiple linear regression model with just these two predictors results in a flat fitted regression plane (like a flat piece of paper). If, however, we include an interaction between the predictors in our model, then the fitted regression plane looks like a piece of paper that has one edge sloped at one angle and the opposite edge sloped at a different angle, thus creating a three-dimensional curved plane.

Typically, regression models that include interactions between quantitative predictors adhere to the **hierarchy principle**, which says that if your model includes an interaction term, X_1X_2, and X_1X_2 is shown to be a statistically significant predictor of *Y*, then your model should also include the "main effects," X_1 and X_2, whether or not the coefficients for these main effects are significant. Depending on the subject area, there may be circumstances where a main effect could be excluded, but this tends to be the exception.

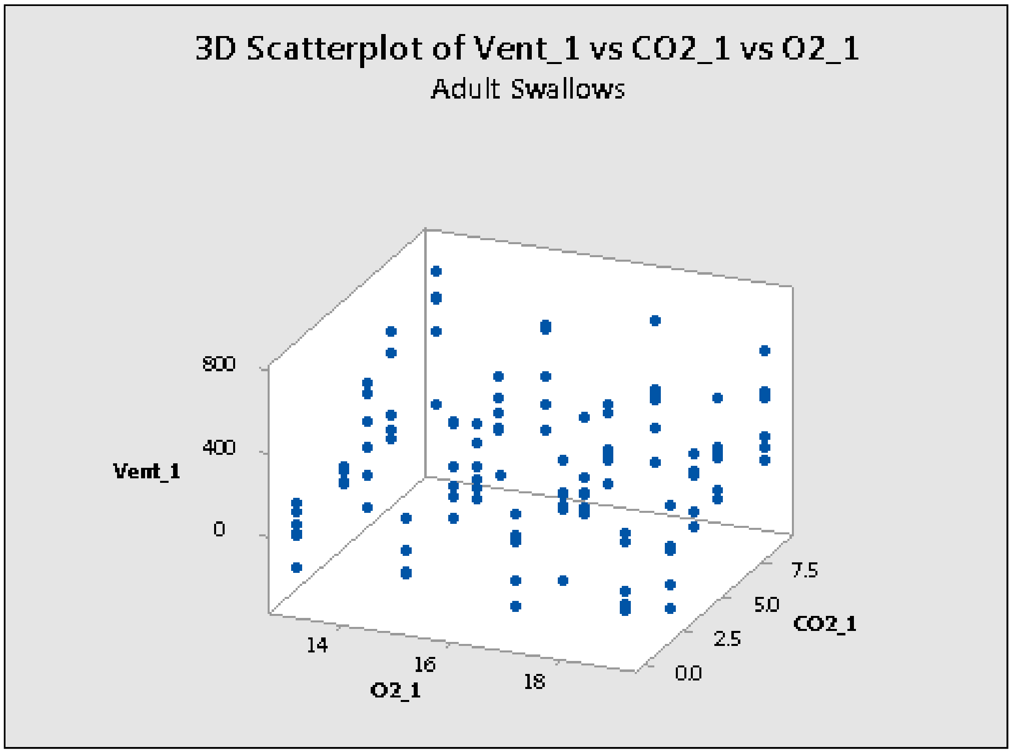
We can use interaction terms in any multiple linear regression model. Here we consider an example with two quantitative predictors and one indicator variable for a categorical predictor. In Lesson 5 we looked at some data resulting from a study in which the researchers (Colby, *et al*, 1987) wanted to determine if nestling bank swallows alter the way they breathe in order to survive the poor air quality conditions of their underground burrows. In reality, the researchers studied not only the breathing behavior of nestling bank swallows, but that of adult bank swallows as well.

To refresh your memory, the researchers conducted the following randomized experiment on 120 nestling bank swallows. In an underground burrow, they varied the percentage of oxygen at four different levels (13%, 15%, 17%, and 19%) and the percentage of carbon dioxide at five different levels (0%, 3%, 4.5%, 6%, and 9%). Under each of the resulting 5×4 = 20 experimental conditions, the researchers observed the total volume of air breathed per minute for each of 6 nestling bank swallows. They replicated the same randomized experiment on 120 adult bank swallows. In this way, they obtained the following data (allswallows.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/allswallows.txt)) on *n* = 240 swallows:

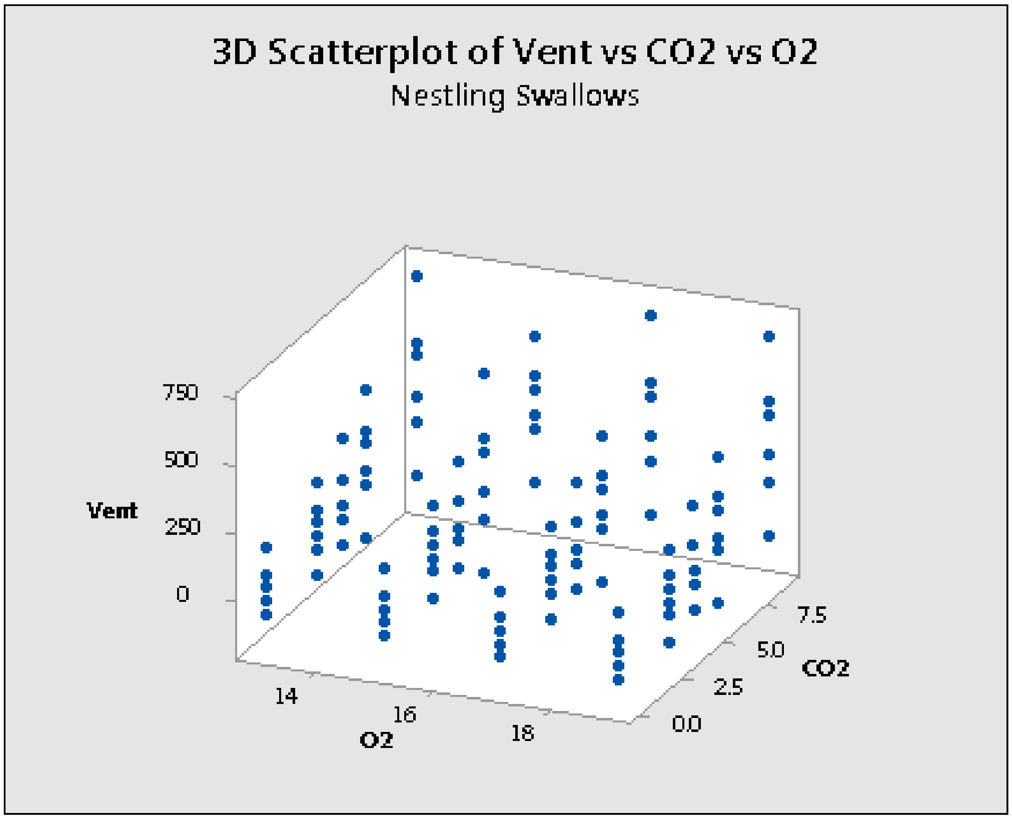
- Response (y): percentage increase in "minute ventilation", (**Vent**), *i.e.*, total volume of air breathed per minute.
- Potential predictor (x_1): percentage of oxygen (**O2**) in the air the swallows breathe
- Potential predictor (x_2): percentage of carbon dioxide (**CO2**) in the air the swallows breathe
- Potential qualitative predictor (x_3): (**Type**) 1 if bird is an adult, 0 if bird is a nestling

Loading [MathJax]/extensions/MathZoom.js

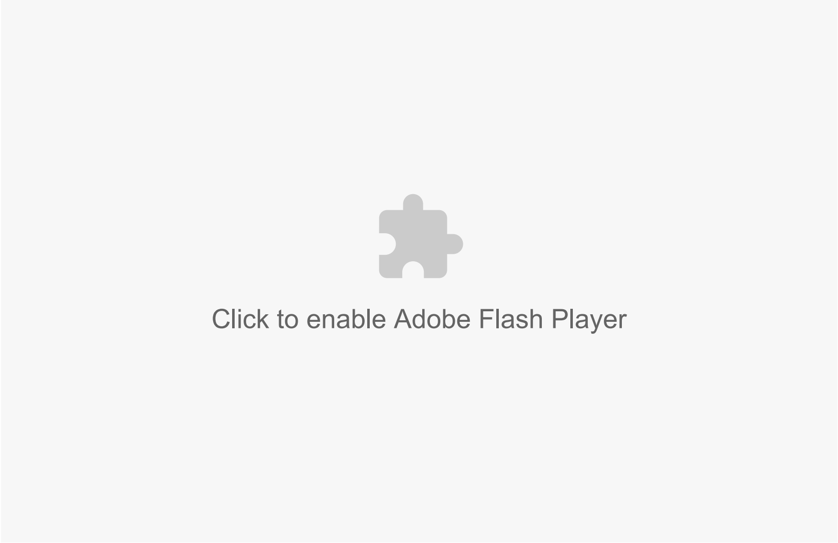
ata for the adult swallows:



and a plot of the resulting data for the nestling bank swallows:



As mentioned previously, the "best fitting" function through each of the above plots will be some sort of surface like a sheet of paper. If you click on the **Draw Plane** button, you will see one possible estimate of the surface for the nestlings:



Click to enable Adobe Flash Player

What we don't know is if the best fitting function —that is, the sheet of paper —through the data will be curved or not. Including interaction terms in the regression model allows the function to have some curvature, while leaving interaction terms out of the regression model forces the function to be flat.

Let's consider the research question "is there any evidence that the adults differ from the nestlings in terms of their minute ventilation as a function of oxygen and carbon dioxide?"

We could start by formulating the following multiple regression model with two quantitative predictors and one qualitative predictor:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

where:

- y_i is the percentage of minute ventilation for swallow i
- x_{i1} is the percentage of oxygen for swallow i
- x_{i2} is the percentage of carbon dioxide for swallow i
- x_{i3} is the type of bird (0, if nestling and 1, if adult) for swallow i

and the independent error terms ε_i follow a normal distribution with mean 0 and equal variance σ^2.

We now know, however, that there is a risk in omitting an important interaction term. Therefore, let's instead formulate the following multiple regression model with three interaction terms:

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \beta_{23} x_{i2} x_{i3}) + \epsilon_i$$

where:

- y_i is the percentage of minute ventilation for swallow i
- x_{i1} is the percentage of oxygen for swallow i
- x_{i2} is the percentage of carbon dioxide for swallow i
- x_{i3} is the type of bird (0, if nestling and 1, if adult) for swallow i
- $x_{i1}x_{i2}$, $x_{i1}x_{i3}$, and $x_{i2}x_{i3}$ are interaction terms

By setting the predictor x_3 to equal 0 and 1 and doing a little bit of algebra we see that our formulated model yields two response functions —one for each type of bird:

Type of bird	Formulated regression function
If a nestling, then x_{i3} = 0 and ...	$\mu_Y = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}$
If an adult, then x_{i3} = 1 and ...	$\mu_Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13}) x_{i1} + (\beta_2 + \beta_{23}) x_{i2} + \beta_{12} x_{i1} x_{i2}$

The $\beta_{12} x_{i1} x_{i2}$ interaction term appearing in both functions allows the two functions to have the same curvature. The additional β_{13} parameter appearing before the x_{i1} predictor in the regression function for the adults allows the adult function to be shifted from the nestling function in the x_{i1} direction by β_{13} units. And, the additional β_{23} parameter appearing before the x_{i2} predictor in the regression function for the adults allows the adult function to be shifted from the nestling function in the x_{i2} direction by β_{23} units.

The results for this model are:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -18.399    160.007  -0.115   0.9086
O2             1.189      9.854   0.121   0.9041
CO2           54.281     25.987   2.089   0.0378 *
Type         111.658    157.742   0.708   0.4797
TypeO2        -7.008      9.560  -0.733   0.4642
TypeCO2        2.311      7.126   0.324   0.7460
CO2O2         -1.449      1.593  -0.909   0.3642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 165.6 on 233 degrees of freedom
```

```
Multiple R-squared:  0.272,     Adjusted R-squared:  0.2533
F-statistic: 14.51 on 6 and 233 DF,  p-value: 4.642e-14
```

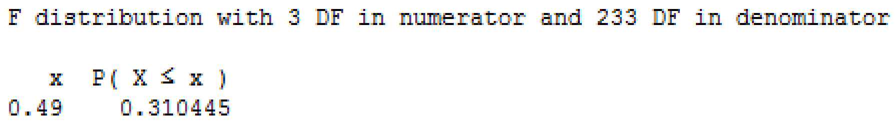
Note that the *P*-values for each of the interaction parameters, β_{12}, β_{13}, and β_{23} are quite large, suggesting there is little evidence for two-way interactions between type of bird, oxygen level, and carbon dioxide level.

Again, however, we should minimize the number of hypothesis tests we perform—and thereby reduce the chance of committing Type I and Type II errors—by instead conducting a general linear *F*-test for testing H_0: $\beta_{12} = \beta_{13} = \beta_{23} = 0$ simultaneously. The residual error sum of squares for this (full) model is 6,388,603 with 233 degrees of freedom. The residual error sum of squares for a (reduced) model that excludes the interaction terms is 6,428,886 with 236 degrees of freedom.

The general linear *F*-statistic is therefore:

$$F^* = \frac{(6428886 - 6388603)/3}{6388603/233} = 0.49$$

And the following output:

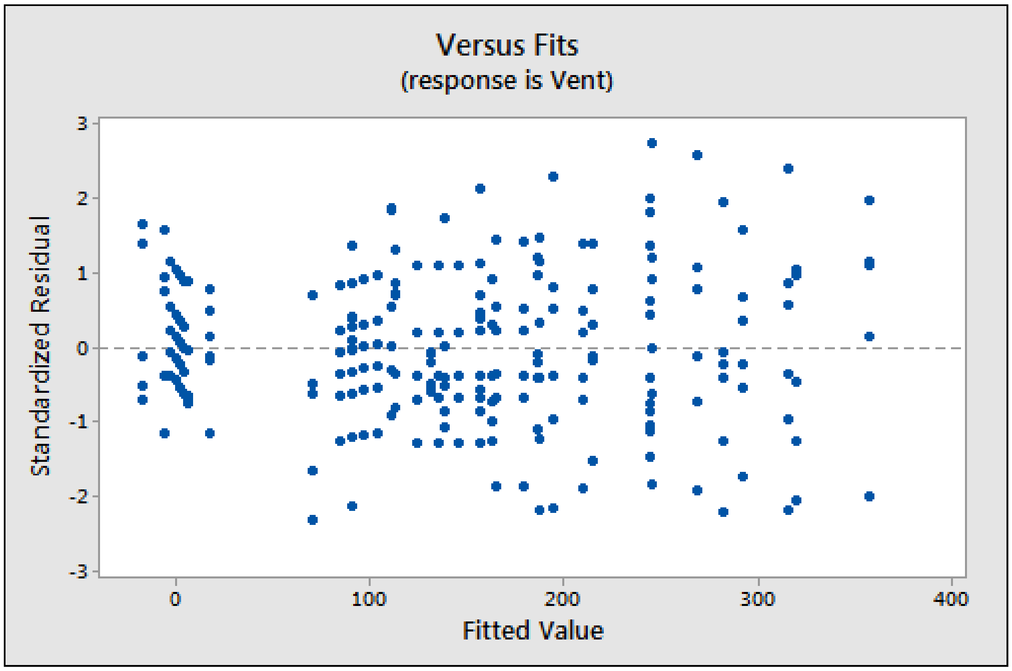


```
F distribution with 3 DF in numerator and 233 DF in denominator

   x  P( X ≤ x )
0.49    0.310445
```

tells us that the probability of observing an *F*-statistic less than 0.49, with 3 numerator and 233 denominator degrees of freedom, is 0.31. Therefore, the probability of observing an *F*-statistic greater than 0.49, with 3 numerator and 233 denominator degrees of freedom, is 1-0.31 or 0.69. That is, the *P*-value is 0.69. There is insufficient evidence at the 0.05 level to conclude that at least one of the interaction parameters is not 0.

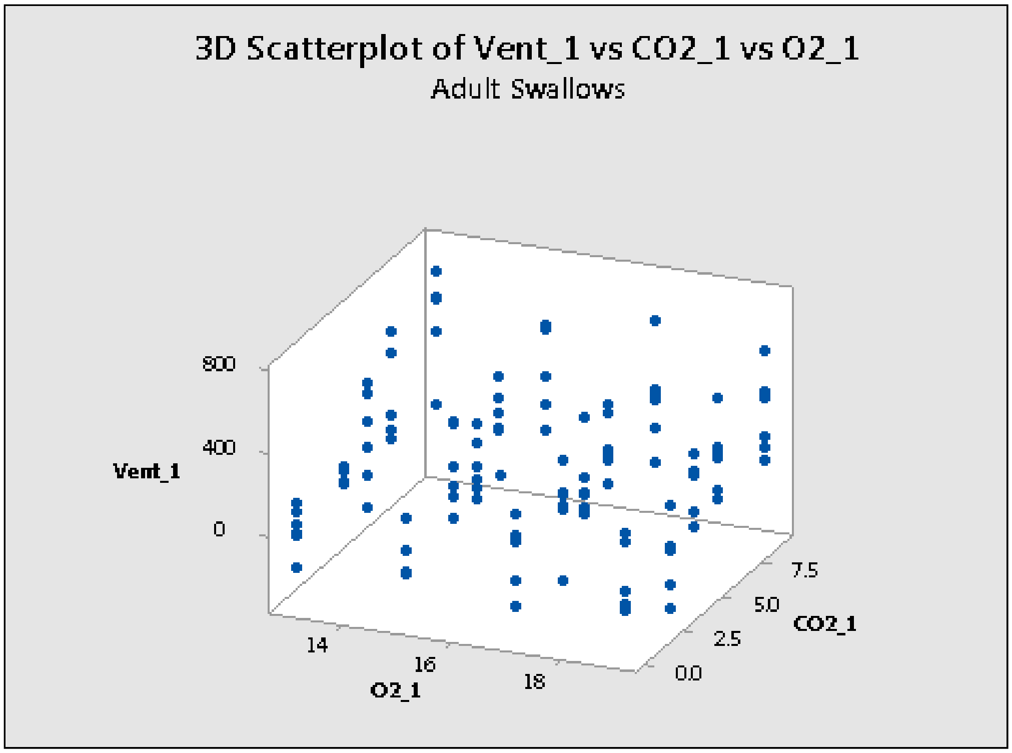
The residual versus fits plot:



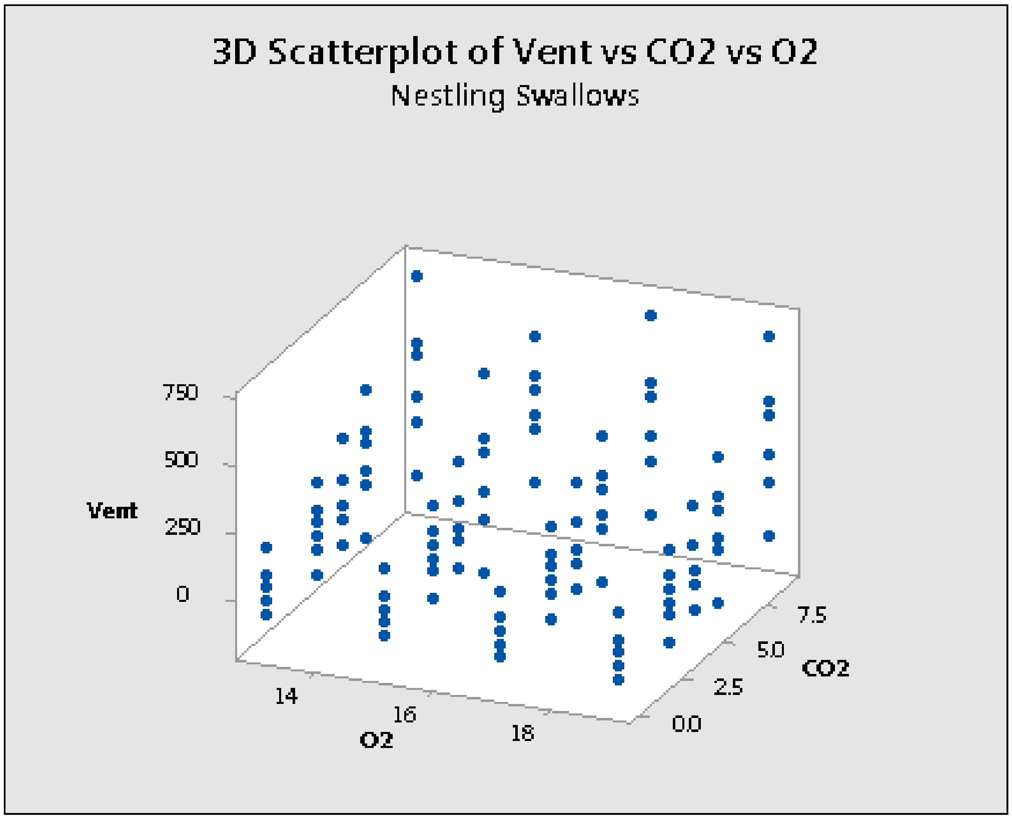
also suggests that there is something not quite right about the fit of the model containing interaction terms.

Loading [MathJax]/extensions/MathZoom.js

Incidentally, if we go back and re-examine the two scatter plots of the data —one for the adults:



and one for the nestlings:



we see that it is believable that there are no interaction terms. If you tried to "draw" the "best fitting" function through each scatter plot, the two functions would probably look like two parallel planes.

So, let's go back to formulating the model with no interactions terms:

Loading [MathJax]/extensions/MathZoom.js

$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3}) + \epsilon_i$$

where:

- y_i is the percentage of minute ventilation for swallow i
- x_{i1} is the percentage of oxygen for swallow i
- x_{i2} is the percentage of carbon dioxide for swallow i
- x_{i3} is the type of bird (0, if nestling and 1, if adult) for swallow i

and the independent error terms ε_i follow a normal distribution with mean 0 and equal variance σ^2.

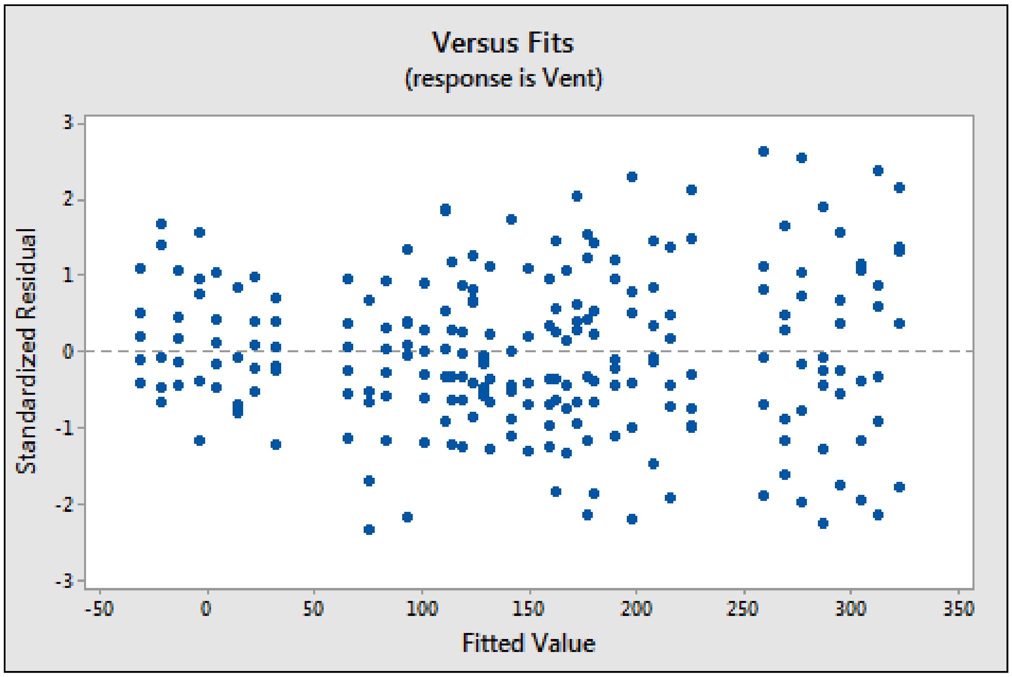
Using software to estimate the regression function, we obtain:

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  136.767     79.334   1.724    0.086 .
O2            -8.834      4.765  -1.854    0.065 .
CO2           32.258      3.551   9.084   <2e-16 ***
Type           9.925     21.308   0.466    0.642
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 165 on 236 degrees of freedom
Multiple R-squared:  0.2675,    Adjusted R-squared:  0.2581
F-statistic: 28.72 on 3 and 236 DF,  p-value: 7.219e-16
```

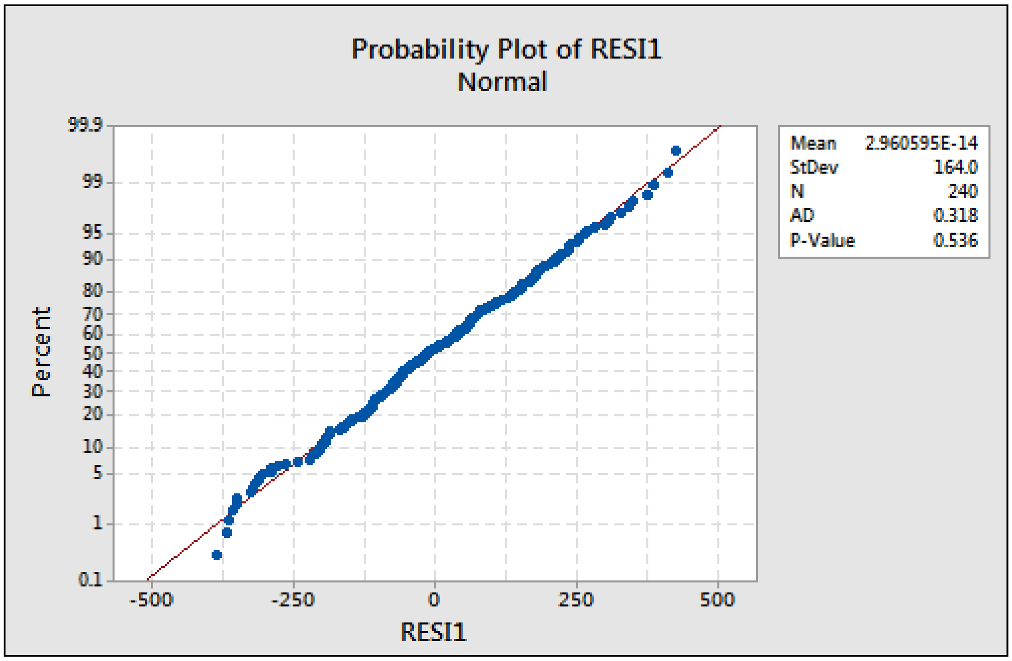
Let's finally answer our primary research question: "is there any evidence that the adult swallows differ from the nestling swallows in terms of their minute ventilation as a function of oxygen and carbon dioxide?" To answer the question, we need only test the null hypothesis $H_0 : \beta_3 = 0$. The software output shows that the P-value is 0.642. We fail to reject the null hypothesis at any reasonable significance level. There is insufficient evidence to conclude that adult swallows differ from nestling swallows with respect to their minute ventilation.

Incidentally, before using the model to answer the research question, we should have assessed the model assumptions. All is fine, however. The residuals versus fits plot for the model with no interaction terms:



shows a marked improvement over the residuals versus fits plot for the model with the interaction terms. Perhaps there is a little bit of fanning? A little bit, but perhaps not enough to worry about.

And, the normal probability plot:



suggests there is no reason to worry about non-normal error terms.

‹ 7.5 - Further Transformation Advice and Box-Cox (/stat462/node/156) up (/stat462/node/85) 7.7 - Polynomial Regression › (/stat462/node/158)

Loading [MathJax]/extensions/MathZoom.js

STAT 462

Applied Regression Analysis

8.5 - Additive Effects

Example

Earlier in this lesson, we investigated a set of data in which the researchers (Daniel, 1999) were interested in determining whether a baby's birth weight was related to his or mother's smoking habits during pregnancy. The researchers collected the following data (birthsmokers.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/birthsmokers.txt)) on a random sample of n = 32 births:

- Response (y): birth **weight** in grams of baby
- Potential predictor (x_1): **smoking** status of mother (yes or no)
- Potential predictor (x_2): length of **gestation** in weeks

For these data, we formulated the following first-order model:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

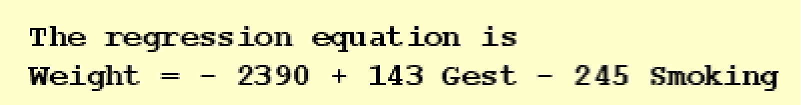
where:

- y_i is the birth weight of baby i in grams
- x_{i1} is the length of gestation of baby i in weeks
- x_{i2} = 1, if baby i's mother smoked and x_{i2} = 0, if not

and the independent error terms ε_i follow a normal distribution with mean 0 and equal variance σ^2.

Do you think the two predictors — the length of gestation and the smoking behavior of the mother — interact? That is, do you think the effect of the gestation length on mean birth weight depends on whether or not the mother is a smoker? Or, equivalently, do you think the effect of smoking on mean birth weight depends on the length of gestation?

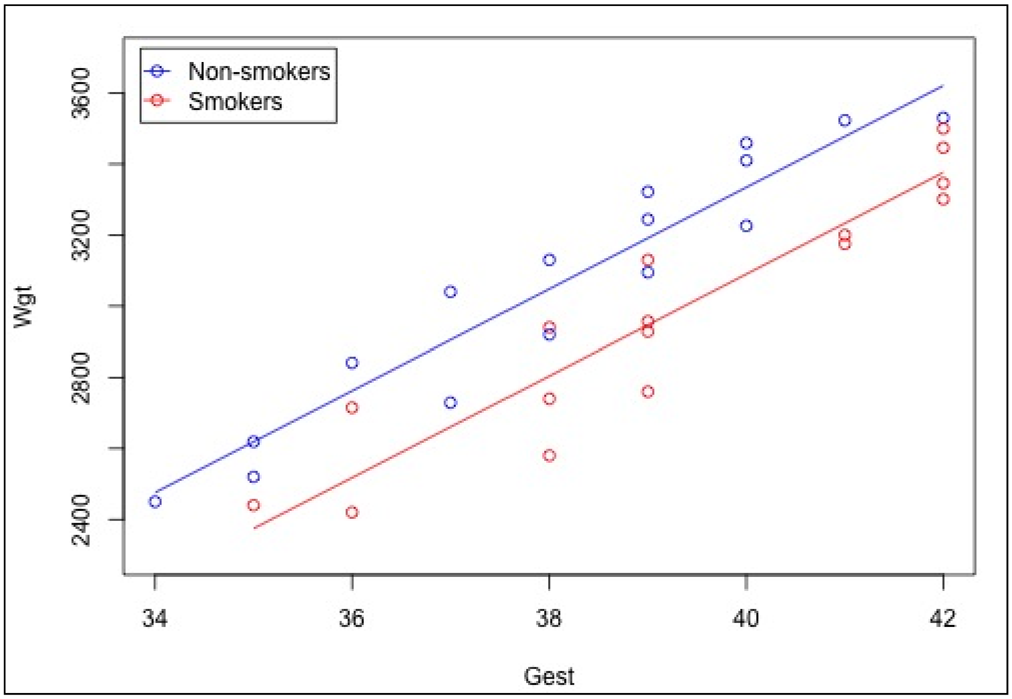
We can take a look at the estimated regression equation to arrive at reasonable answers to these questions. Upon analyzing the sample of n = 32 births, the regression equation is:



```
The regression equation is
Weight = - 2390 + 143 Gest - 245 Smoking
```

Loading [MathJax]/extensions/MathZoom.js

And, a plot of the estimated regression equation looks like:



The blue circles and line represent the data and estimated function for non-smoking mothers (x_2=0), while the red circles and line represent the data and estimated function for smoking mothers (x_2=1). Remember that the two lines in this plot are exactly parallel.

Now, in light of the plot, let's investigate those questions again:

- **Does the effect of the gestation length on mean birth weight depend on whether or not the mother is a smoker?** The answer is no! Regardless of whether or not the mother is a smoker, for each additional one-week of gestation, the mean birth weight is predicted to increase by 143 grams. This lack of interaction between the two predictors is exhibted by the parallelness of the two lines.
- **Does the effect of smoking on mean birth weight depend on the length of gestation?** The answer is no! For a fixed length of gestation, the mean birth weight of babies born to smoking mothers is predicted to be 245 grams lower than the mean birth weight of babies born to non-smoking mothers. Again, this lack of interaction between the two predictors is exhibted by the parallelness of the two lines.

When two predictors do not interact, we say that each predictor has an "**additive effect**" on the response. More formally, a regression model contains additive effects if the response function can be written as a sum of functions of the predictor variables:

$$\mu_y = f_1(x_1) + f_2(x_2) + \ldots + f_{p-1}(x_{p-1})$$

For example, our regression model for the birth weights of babies contains additive effects, because the response function can be written as a sum of functions of the predictor variables:

$$\mu_y = (\beta_0) + (\beta_1 x_{i1}) + (\beta_2 x_{i2})$$

‹ 8.4 - Coding Qualitative Variables (/stat462/node/163) up (/stat462/node/86) 8.6 - Interaction Effects › (/stat462/node/165)

Loading [MathJax]/extensions/MathZoom.js

STAT 462

Applied Regression Analysis

8.6 - Interaction Effects

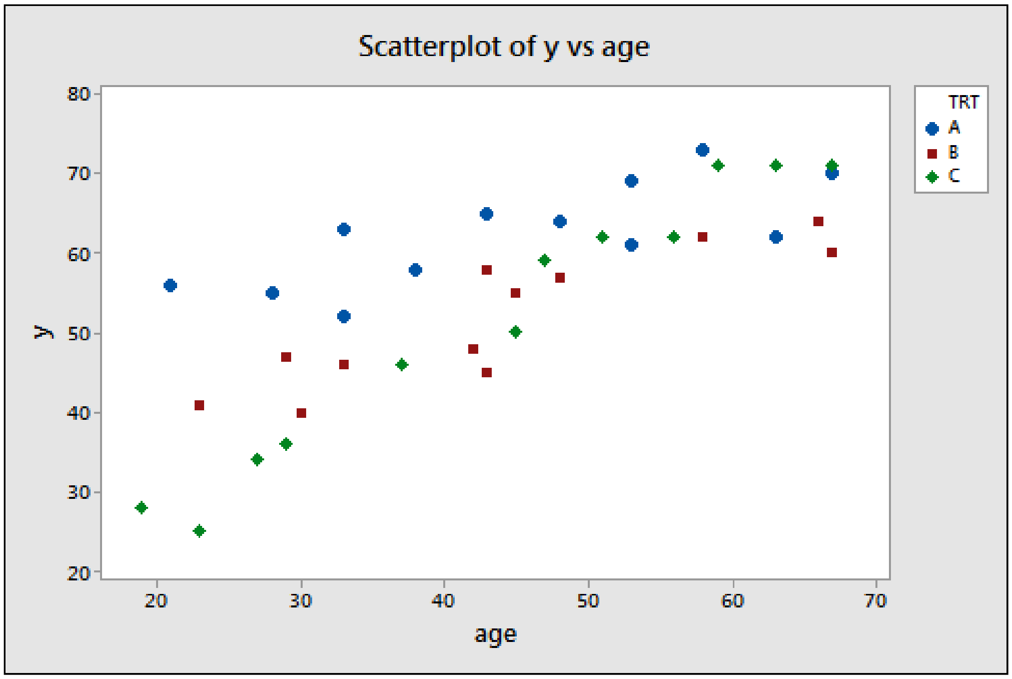
Now that we've clarified what additive effects are, let's take a look at an example where including "**interaction terms**" is appropriate.

Example

Some researchers (Daniel, 1999) were interested in comparing the effectiveness of three treatments for severe depression. For the sake of simplicity, we denote the three treatments A, B, and C. The researchers collected the following data (depression.txt (https://onlinecourses.science.psu.edu/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/depression.txt)) on a random sample of n = 36 severely depressed individuals:

- y_i = measure of the effectiveness of the treatment for individual i
- x_{i1} = age (in years) of individual i
- x_{i2} = 1 if individual i received treatment A and 0, if not
- x_{i3} = 1 if individual i received treatment B and 0, if not

A scatter plot of the data with treatment effectiveness on the y-axis and age on the x-axis looks like:



The blue circles represent the data for individuals receiving treatment A, the red squares represent the data for individuals receiving treatment B, and the green diamonds represent the data for individuals receiving treatment C.

In the previous example, the two estimated regression functions had the same slopes —that is, they were parallel. If you tried to draw three best fitting lines through the data of this example, do you think the slopes of your lines would be the same? Probably not! In this case, we need to include what are called "**interaction terms**" in our formulated regression model.

A (second-order) multiple regression model with interaction terms is:

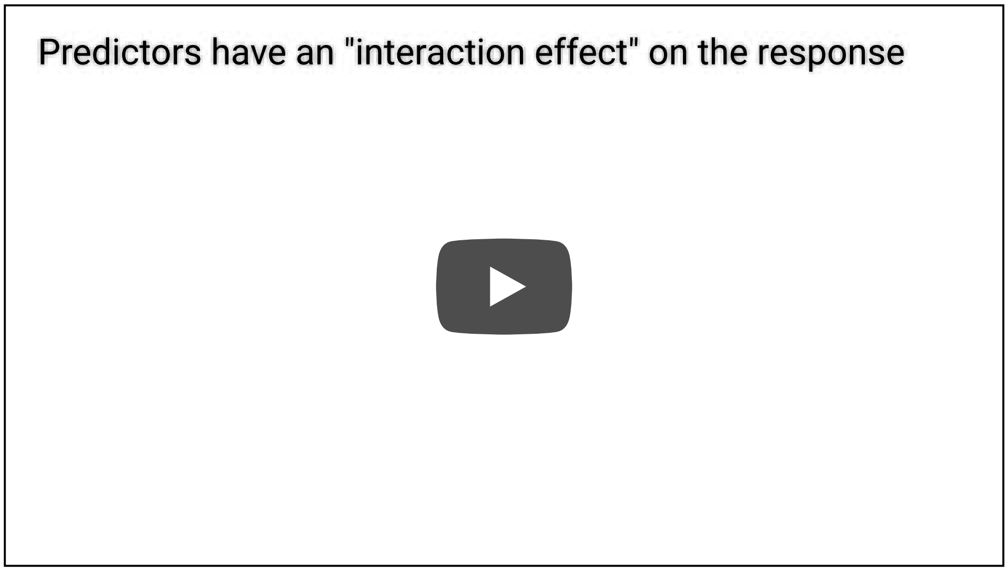
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \epsilon_i$$

where:

- y_i = measure of the effectiveness of the treatment for individual i
- x_{i1} = age (in years) of individual i
- x_{i2} = 1 if individual i received treatment A and 0, if not
- x_{i3} = 1 if individual i received treatment B and 0, if not

and the independent error terms ε_i follow a normal distribution with mean 0 and equal variance σ^2. Perhaps not surprisingly, the terms $x_{i1}x_{i2}$ and $x_{i1}x_{i3}$ are the interaction terms in the model.

Let's investigate our formulated model to discover in what way the predictors have an "**interaction effect**" on the response. We start by determining the formulated regression function for each of the three treatments. In short — after a little bit of algebra (see below) —we learn that the model defines three different regression functions —one for each of the three treatments:



Predictors have an "interaction effect" on the response

Treatment	Formulated regression function
If patient receives A, then (x_{i2} = 1, x_{i3} = 0) and ...	$\mu_Y = (\beta_0 + \beta_2) + (\beta_1 + \beta_{12})x_{i1}$
If patient receives B, then (x_{i2} = 0, x_{i3} = 1) and ...	$\mu_Y = (\beta_0 + \beta_3) + (\beta_1 + \beta_{13})x_{i1}$
If patient receives C, then (x_{i2} = 0, x_{i3} = 0) and ...	$\mu_Y = \beta_0 + \beta_1 x_{i1}$

So, in what way does including the interaction terms, $x_{i1}x_{i2}$ and $x_{i1}x_{i3}$, in the model imply that the predictors have an "**interaction effect**" on the mean response? Note that the slopes of the three regression functions differ —the slope of the first line is $\beta_1 + \beta_{12}$, the slope of the second line is $\beta_1 + \beta_{13}$, and the slope of the third line is β_1. What does this mean in a practical sense? It means that...

- the effect of the individual's age (x_1) on the treatment's mean effectiveness (μ_Y) depends on the treatment (x_2 and x_3), and ...
- the effect of treatment (x_2 and x_3) on the treatment's mean effectiveness (μ_Y) depends on the individual's age (x_1).

In general, then, what does it mean for two predictors "**to interact**"?

- Two predictors interact if the effect on the response variable of one predictor **depends on the value of the other**.
- A slope parameter can no longer be interpreted as the change in the mean response for each unit increase in the predictor, while the other predictors are held constant.

And, what are "**interaction effects**"?

A regression model contains **interaction effects** if the response function is not additive and cannot be written as a sum of functions of the predictor variables. That is, a regression model contains interaction effects if:

$$\mu_Y \neq f_1(x_1) + f_1(x_1) + \cdots + f_{p-1}(x_{p-1})$$

For our example concerning treatment for depression, the mean response:

$$\mu_Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3$$

can not be separated into distinct functions of each of the individual predictors. That is, there is no way of "breaking apart" $\beta_{12}x_1x_2$ and $\beta_{13}x_1x_3$ into distinct pieces. Therefore, we say that x_1 and x_2 interact, and x_1 and x_3 interact.

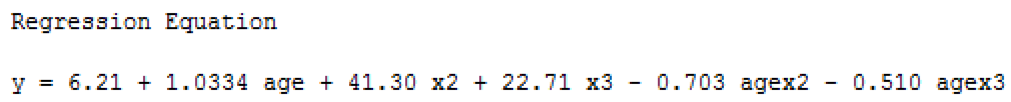
In returning to our example, let's recall that the appropriate steps in any regression analysis are:

- Model building
    - Model formulation
    - Model estimation
    - Model evaluation
- Model use

So far, within the model building step, all we've done is **formulate** the regression model as:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_{12} x_{i1} x_{i2} + \beta_{13} x_{i1} x_{i3} + \epsilon_i$$

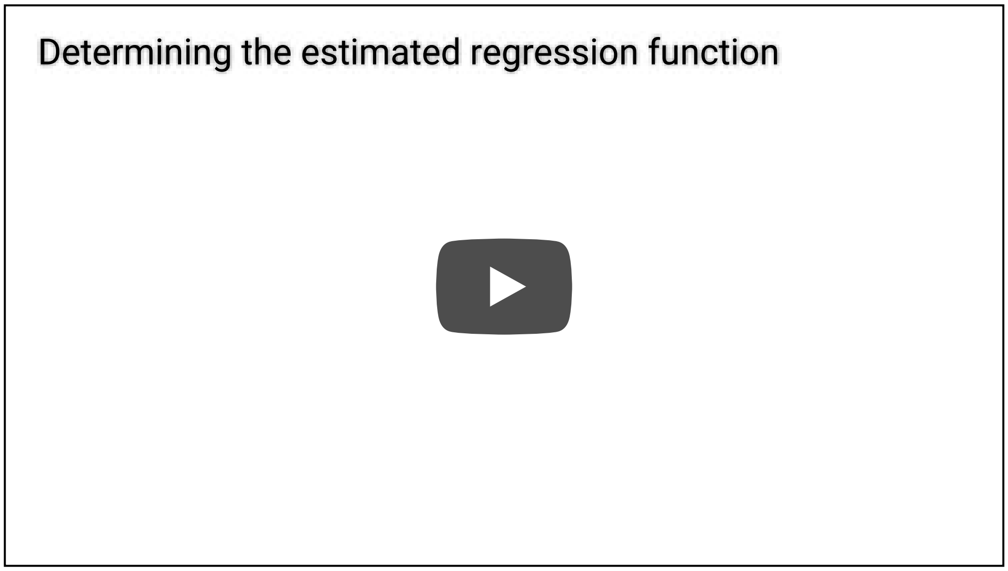
After **estimating** the model, the regression equation is:



```
Regression Equation

y = 6.21 + 1.0334 age + 41.30 x2 + 22.71 x3 - 0.703 agex2 - 0.510 agex3
```

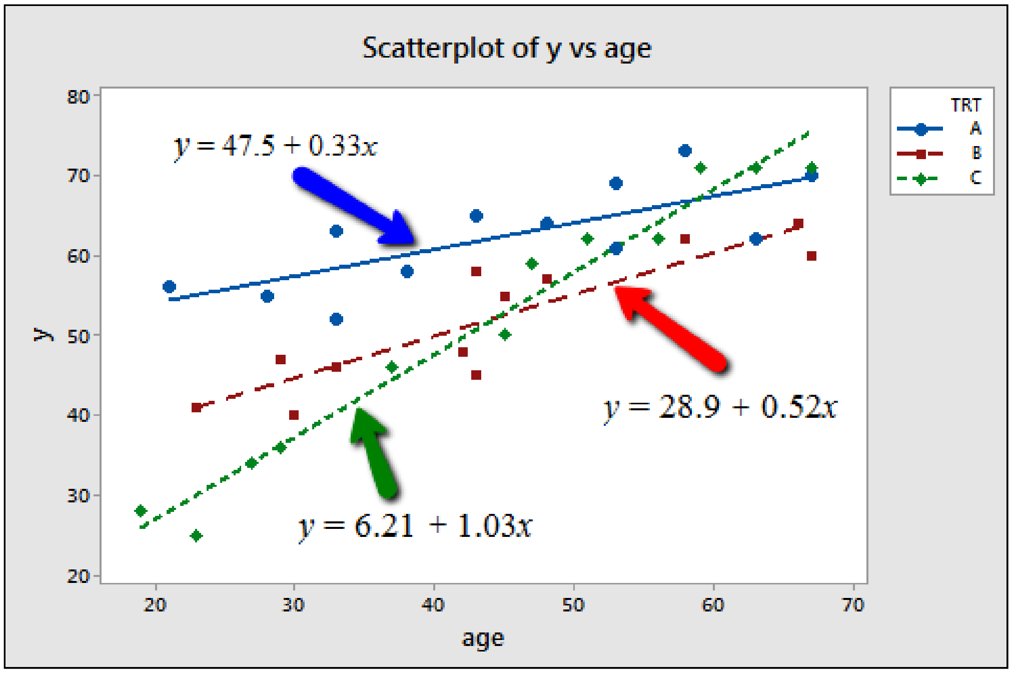
Now, if we plug the possible values for x_2 and x_3 into the estimated regression function, we obtain the three "best fitting" lines —one for each treatment (A, B and C) —through the data. Here's the algebra for determining the estimated regression function for patients receiving treatment A.



Doing similar algebra for patients receiving treatments B and C, we obtain:

Treatment	Estimated regression function
If patient receives A, then ($x_2 = 1, x_3 = 0$) and ...	$\hat{y} = 47.5 + 0.33x_1$
If patient receives B, then ($x_2 = 0, x_3 = 1$) and ...	$\hat{y} = 28.9 + 0.52x_1$
If patient receives C, then ($x_2 = 0, x_3 = 0$) and ...	$\hat{y} = 6.21 + 1.03x_1$

And, plotting the three "best fitting" lines, we obtain:



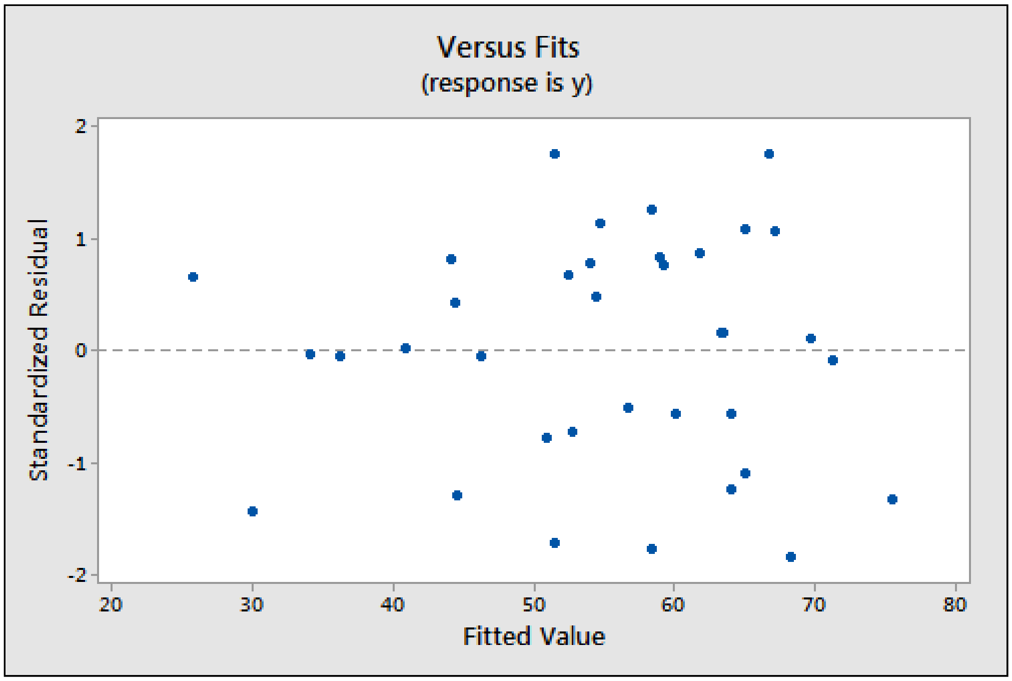
What do the estimated slopes tell us?

- For patients *in this study* receiving treatment A, the effectiveness of the treatment is predicted to increase 0.33 units for every additional year in age.
- For patients *in this study* receiving treatment B, the effectiveness of the treatment is predicted to increase 0.52 units for every additional year in age.
- For patients *in this study* receiving treatment C, the effectiveness of the treatment is predicted to increase 1.03 units for every additional year in age.

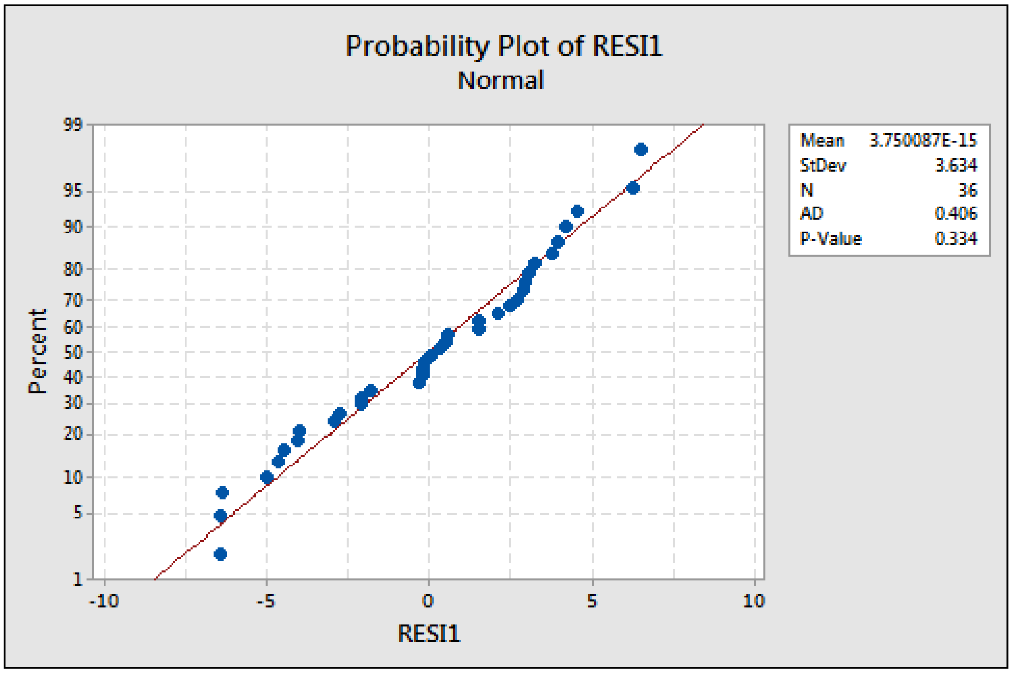
In short, the effect of age on the predicted treatment effectiveness depends on the treatment given. That is, age appears to **interact** with treatment in its impact on treatment effectiveness. The interaction is exhibited graphically by the "nonparallelness" (is that a word?) of the lines.

Of course, our primary goal is not to draw conclusions about this particular sample of depressed individuals, but rather about the entire population of depressed individuals. That is, we want to use our estimated model to draw conclusions about the larger population of depressed individuals. Before we do so, however, we first should **evaluate** the model.

The residuals versus fits plot:



exhibits all of the "good" behavior, suggesting that the model fits well, there are no obvious outliers, and the error variances are indeed constant. And, the normal probability plot:



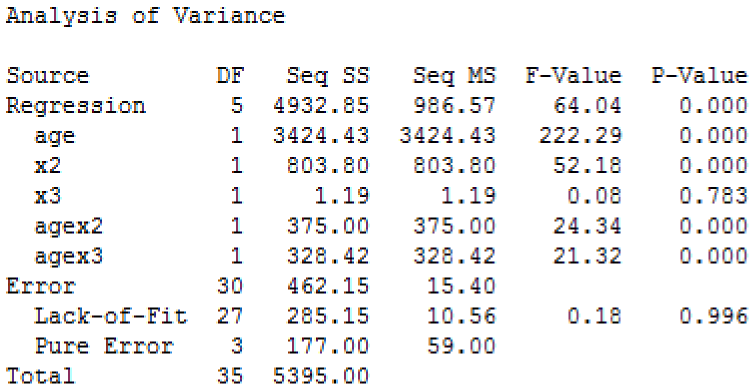
exhibits linear trend and a large *P*-value, suggesting that the error terms are indeed normally distributed.

Having successfully **built** —formulated, estimated, and evaluated —a model, we now can **use** the model to answer our research questions. Let's consider two different questions that we might want answered.

First research question. For every age, is there a difference in the mean effectiveness for the three treatments? As is usually the case, our formulated regression model helps determine how to answer the research question. Our formulated regression model suggests that answering the question involves testing whether the population regression functions are identical.

That is, we need to test the null hypothesis $H_0 : \beta_2 = \beta_3 = \beta_{12} = \beta_{13} = 0$ against the alternative H_A : at least one of these slope parameters is not 0.

We know how to do that! The relevant software output:



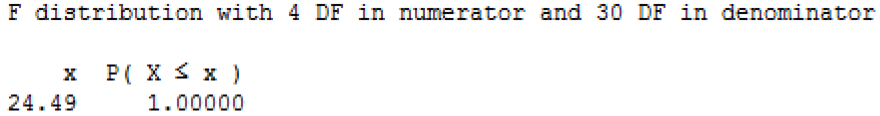
```
Analysis of Variance

Source          DF   Seq SS   Seq MS  F-Value  P-Value
Regression       5  4932.85   986.57    64.04    0.000
  age            1  3424.43  3424.43   222.29    0.000
  x2             1   803.80   803.80    52.18    0.000
  x3             1     1.19     1.19     0.08    0.783
  agex2          1   375.00   375.00    24.34    0.000
  agex3          1   328.42   328.42    21.32    0.000
Error           30   462.15    15.40
  Lack-of-Fit   27   285.15    10.56     0.18    0.996
  Pure Error     3   177.00    59.00
Total           35  5395.00
```

tells us that the appropriate partial *F*-statistic for testing the above hypothesis is:

$$F = \frac{(803.8 + 1.19 + 375 + 328.42)/4}{15.4} = 24.49.$$

To find the *P*-value:



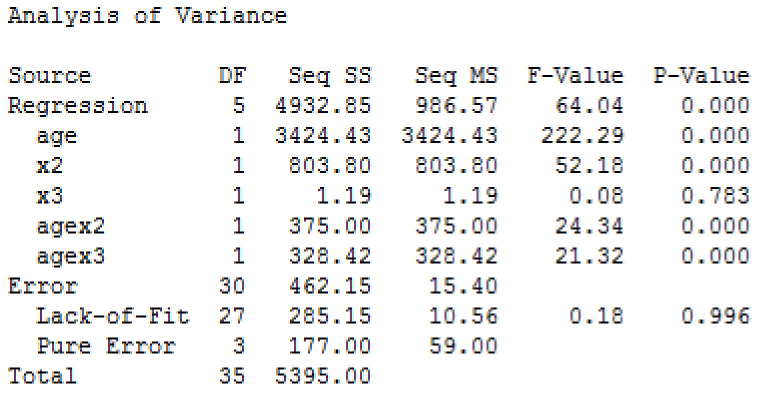
```
F distribution with 4 DF in numerator and 30 DF in denominator

    x  P( X ≤ x )
24.49     1.00000
```

Thus the probability of observing an *F*-statistic —with 4 numerator and 30 denominator degrees of freedom —less than our observed test statistic 24.49 is > 0.999. Therefore, our *P*-value is < 0.001. We can reject our null hypothesis. There is sufficient evidence at the α = 0.05 level to conclude that there is a significant difference in the mean effectiveness for the three treatments.

Second research question. Does the effect of age on the treatment's effectiveness depend on treatment? Our formulated regression model suggests that answering the question involves testing whether the two interaction parameters β_{12} and β_{13} are significant. That is, we need to test the null hypothesis $H_0 : \beta_{12} = \beta_{13} = 0$ against the alternative H_A : at least one of the interaction parameters is not 0.

The relevant software output:



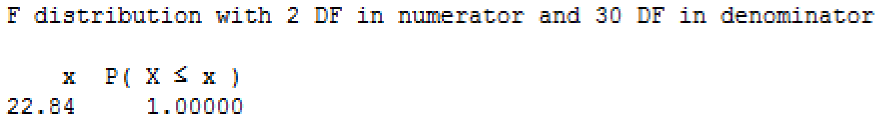
```
Analysis of Variance

Source         DF   Seq SS   Seq MS  F-Value  P-Value
Regression      5  4932.85   986.57    64.04    0.000
  age           1  3424.43  3424.43   222.29    0.000
  x2            1   803.80   803.80    52.18    0.000
  x3            1     1.19     1.19     0.08    0.783
  agex2         1   375.00   375.00    24.34    0.000
  agex3         1   328.42   328.42    21.32    0.000
Error          30   462.15    15.40
  Lack-of-Fit  27   285.15    10.56     0.18    0.996
  Pure Error    3   177.00    59.00
Total          35  5395.00
```

tells us that the appropriate partial *F*-statistic for testing the above hypothesis is:

$$F = \frac{(375 + 328.42)/2}{15.4} = 22.84.$$

To find the *P*-value:



```
F distribution with 2 DF in numerator and 30 DF in denominator

    x  P( X ≤ x )
22.84     1.00000
```

Thus the probability of observing an *F*-statistic — with 2 numerator and 30 denominator degrees of freedom — less than our observed test statistic 22.84 is > 0.999. Therefore, our *P*-value is < 0.001. We can reject our null hypothesis. There is sufficient evidence at the α = 0.05 level to conclude that the effect of age on the treatment's effectiveness depends on the treatment.

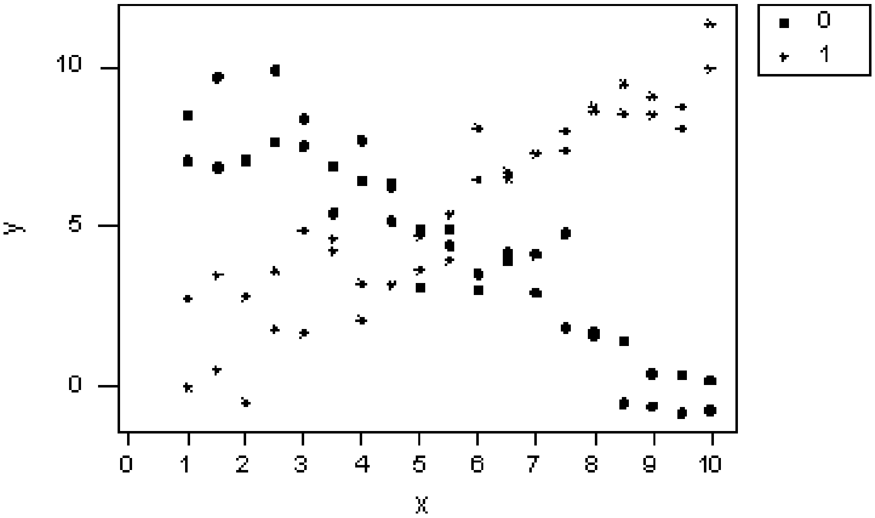
‹ 8.5 - Additive Effects (/stat462/node/164) up (/stat462/node/86) 8.7 - Leaving an Important Interaction Out of a Model › (/stat462/node/166)

STAT 462

Applied Regression Analysis

8.7 - Leaving an Important Interaction Out of a Model

Before we take a look at another example of a regression model containing interaction terms, let's take a little detour to explore the impact of leaving a necessary interaction term out of the model. To do so, let's consider some contrived data for which there is a response *y* and two predictors —one quantitative predictor *x* and one qualitative predictor that takes on values 0 or 1. Looking at a plot of the data:



consider two questions:

- Does the plot suggest *x* is related to *y*? Sure! For the 0 group, as *x* increases, *y* decreases, while for the 1 group, as *x* increases, *y* also increases.
- Does the plot suggest there is a treatment effect? Yes! If you look at any one particular *x* value, say 1 for example, the mean response *y* is about 8 for the 0 group and about 2 for the 1 group. In this sense, there is a treatment effect.

As we now know, the answer to the first question suggests that the effect of *x* on *y* depend on the group. That is, the group and *x* appear to interact. Therefore, our regression model should contain an interaction term between the two predictors. But, let's see what happens if we ignore our intuition and don't add the interaction term! That is, let's

l as:

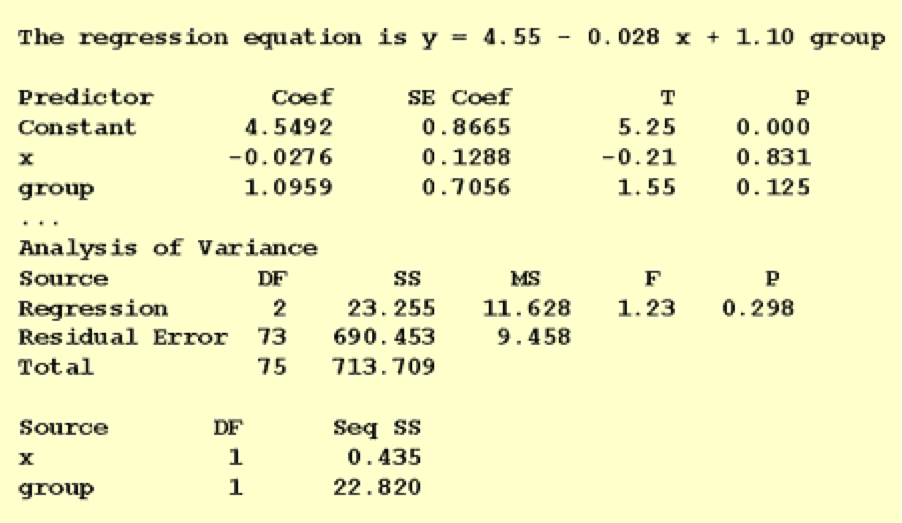
$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}) + \epsilon_i$$

where:

- y_i is the response
- x_{i1} is the quantitative predictor you want to "adjust for "
- x_{i2} is the qualitative group predictor, where 0 denotes the first group and 1 denotes the second group

and the independent error terms ε_i follow a normal distribution with mean 0 and equal variance σ^2.

Now, let's see what conclusions we draw when we fit our contrived data to our formulated model **with no interaction term**:



```
The regression equation is y = 4.55 - 0.028 x + 1.10 group

Predictor         Coef     SE Coef         T        P
Constant        4.5492      0.8665      5.25    0.000
x              -0.0276      0.1288     -0.21    0.831
group           1.0959      0.7056      1.55    0.125
...
Analysis of Variance
Source            DF        SS       MS        F        P
Regression         2    23.255   11.628     1.23    0.298
Residual Error    73   690.453    9.458
Total             75   713.709

Source        DF       Seq SS
x              1        0.435
group          1       22.820
```

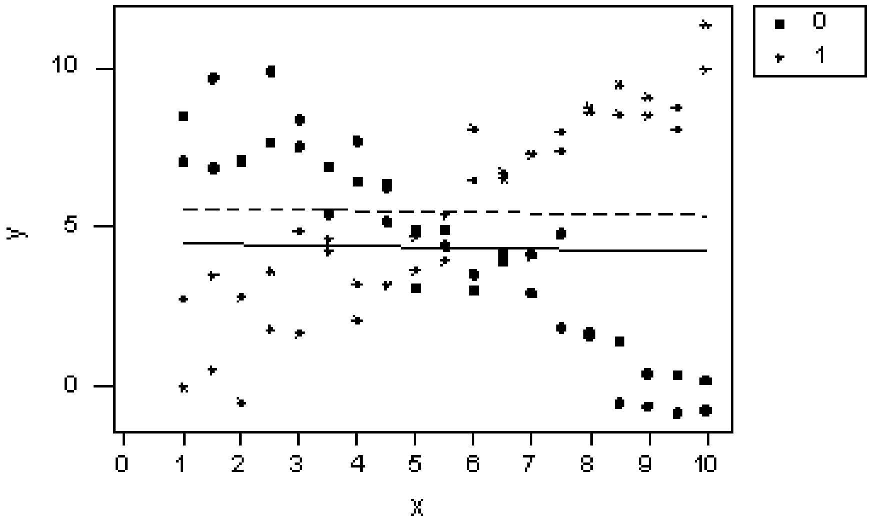
Consider our two research questions:

- Is x related to y? The P-value for testing H_0: $\beta_1 = 0$ is 0.831. There is insufficient evidence at the 0.05 level to conclude that x is related to y. What?! This conclusion contradicts what we'd expect from the plot.
- Is there a treatment effect? The P-value for testing H_0: $\beta_2 = 0$ is 0.125. There is insufficient evidence at the 0.05 level to conclude that there is a treatment effect. Again, this conclusion contradicts what we'd expect from the plot.

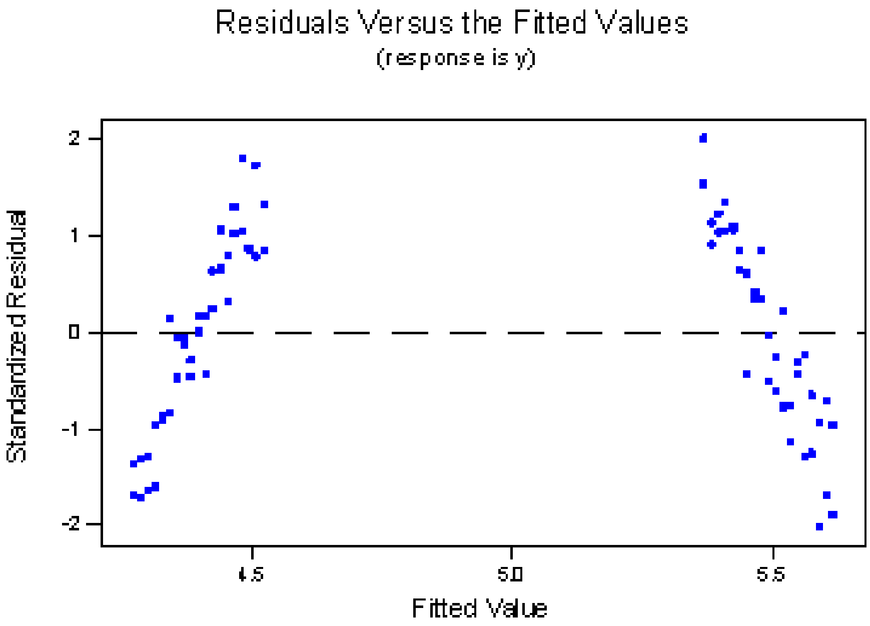
A side note. By conducting the above two tests independently, we increase our chance of making at least one Type I error. Since we are interested in answering both research questions, we could minimize our chance of making a Type I error by conducting the partial F-test for testing, H_0: $\beta_1 = \beta_2 = 0$, that is, that both parameters are simultaneously zero.

Now, let's try to understand why our conclusions don't agree with our intuition based on the plot. If we plug the values 0 and 1 into the *group* variable of the estimated regression equation we obtain two *parallel* lines —one for each group.

A plot of the resulting estimated regression functions:



suggest that the lines don't fit the data very well. By leaving the interaction term out of the model, we have forced the "best fitting lines" to be parallel, when they clearly shouldn't be. The residuals versus fits plot:



provides further evidence that our formulated model does not fit the data well. We now know that the resulting cost is conclusions that just don't make sense.

Let's analyze the data again, but this time with a more appropriately formulated model. Consider the regression model with the interaction term:

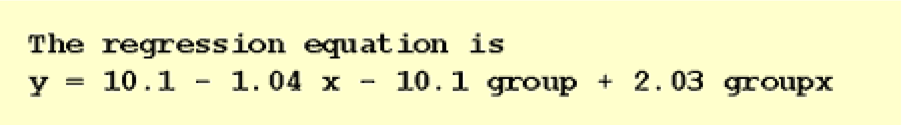
$$y_i = (\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_{12} x_{i1} x_{i2}) + \epsilon_i$$

where:

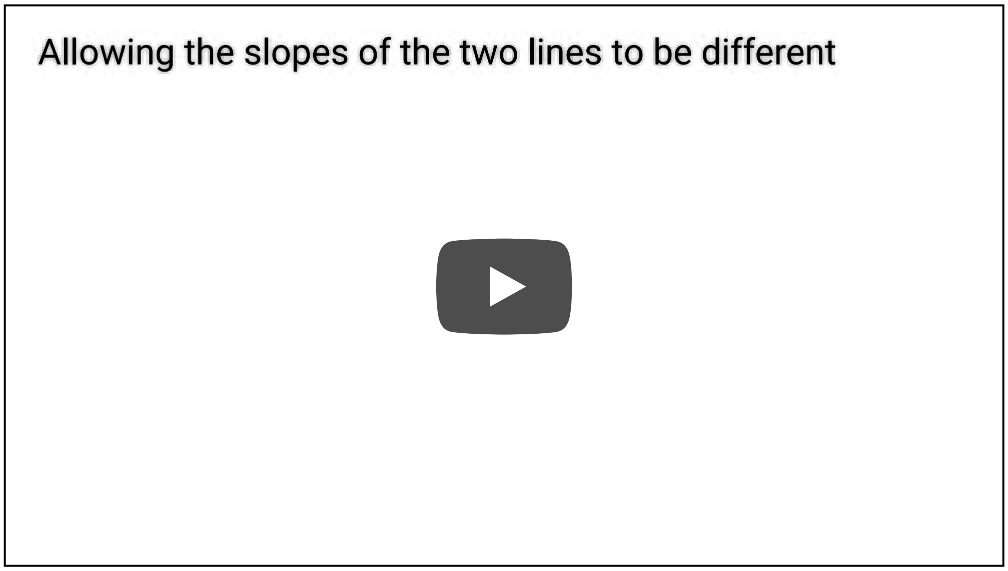
- y_i is the response
- x_{i1} is the quantitative predictor you want to "adjust for "
- x_{i2} is the qualitative group predictor, where 0 denotes the first group and 1 denotes the second group
- $x_{i1}x_{i2}$ is the "missing" interaction term

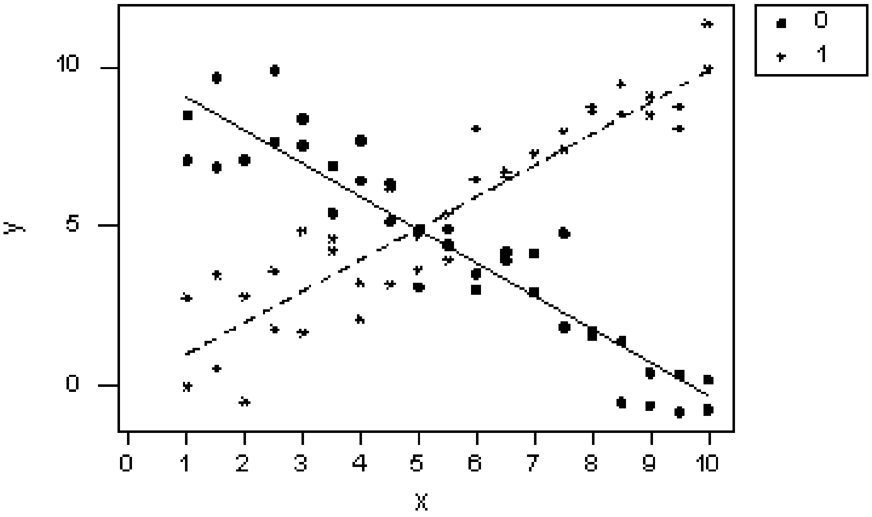
and the independent error terms ε_i follow a normal distribution with mean 0 and equal variance σ^2.

Upon fitting the data to the model with an interaction term, the estimated regression equation is:

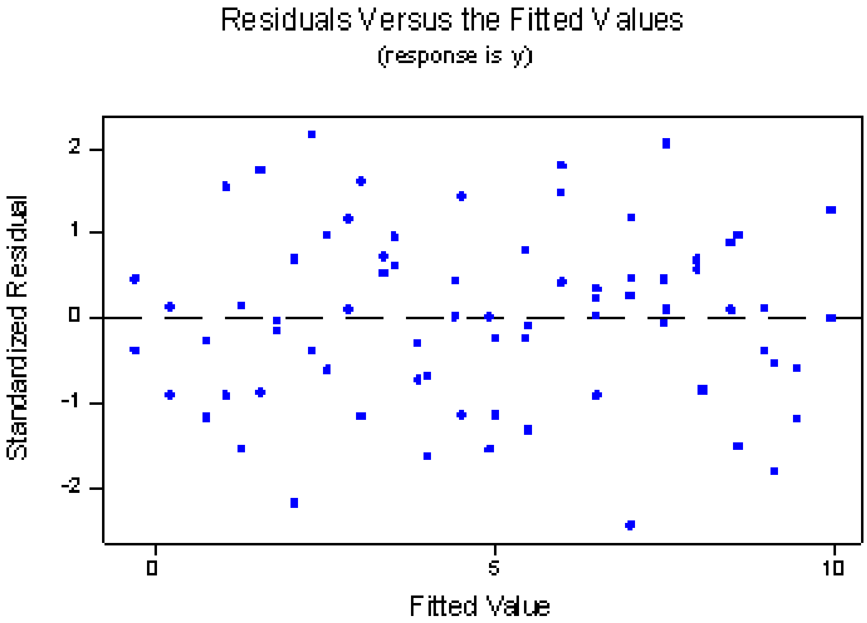

```
The regression equation is
y = 10.1 - 1.04 x - 10.1 group + 2.03 groupx
```

If we now plug the values 0 and 1 into the *group* variable of the estimated regression equation we obtain two *intersecting* lines —one for each group:





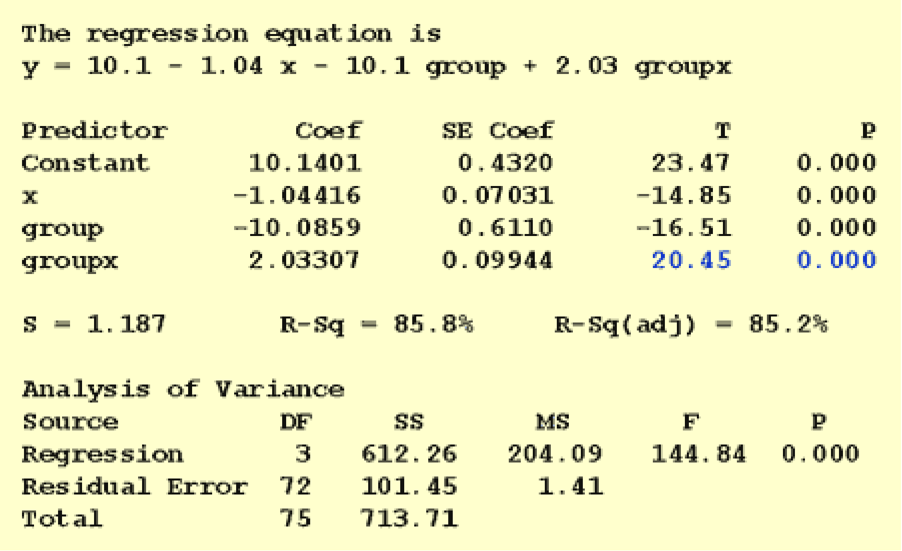
Wow —what a difference! Our formulated regression model now allows the slopes of the two lines to be different. As a result, the lines do a much better job of summarizing the trend in the data. The residuals versus fits plot is about as good as it gets!



The plot provides further evidence that the model with the interaction term does a good job of describing the data.

Okay, so the model with the interaction term does a better job of describing the data than the model with no interaction term. Does it also provide answers to our research questions that make sense?

Let's first consider the question "does the effect of *x* on response *y* depend on the group?" That is, is there an interaction between *x* and group? The software output:



```
The regression equation is
y = 10.1 - 1.04 x - 10.1 group + 2.03 groupx

Predictor        Coef     SE Coef        T       P
Constant      10.1401      0.4320    23.47   0.000
x            -1.04416     0.07031   -14.85   0.000
group        -10.0859      0.6110   -16.51   0.000
groupx        2.03307     0.09944    20.45   0.000

S = 1.187      R-Sq = 85.8%     R-Sq(adj) = 85.2%

Analysis of Variance
Source          DF      SS      MS       F      P
Regression       3  612.26  204.09  144.84  0.000
Residual Error  72  101.45    1.41
Total           75  713.71
```

tells us that the *P*-value for testing $H_0 : \beta_{12} = 0$ is < 0.001. There is strong evidence at the 0.05 level to reject the null hypothesis and conclude that there is indeed an interaction between *x* and group. Aha —our formulated model and resulting analysis yielded a conclusion that makes sense!

Now, what about the research questions "is *x* related to *y*?" and "is there a treatment effect?" Because there is an interaction between *x* and group, it really doesn't make sense to talk about the effect of *x* on *y* without taking into account group. And, it doesn't really make sense to talk about differences in the two groups without taking into account *x*. Thus, neither of these two research questions make sense in the presence of the interaction. This is why you'll often hear statisticians say **"never interpret a main effect in the presence of an interaction."**

Loading [MathJax]/extensions/MathZoom.js

In short, the moral of the story of this little detour that we took is that if we leave an important interaction term out of our model, our analysis can lead us to make erroneous conclusions.

‹ 8.6 - Interaction Effects (/stat462/node/165) up (/stat462/node/86) 8.8 - Further Categorical Predictor Examples › (/stat462/node/168)

8.8 - Further Categorical Predictor Examples

Example 1: Muscle Mass Data

Suppose that we describe y = muscle mass as a function of x_1 = age and x_2 = gender for people in the 40 to 60 year-old age group. We could code the gender variable as $x_2 = 1$ if the subject is female and $x_2 = 0$ if the subject is male.

Consider the multiple regression equation

$$\mathrm{E}(Y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 .$$

The usual slope interpretation will work for β_2, the coefficient that multiplies the gender indicator. Increasing gender by one unit simply moves us from male to female. Thus β_2 = the difference between average muscle mass for females and males of the same age.

Example 2: Real Estate Air Conditioning

Consider the real estate dataset: realestate.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/realestate.txt) . Let us define

- Y = sale price of home
- X_1 = square footage of home
- X_2 = whether home has air conditioning or not.

To put the air conditioning variable into a model create a variable coded as either 1 or 0 to represent the presence or absence of air conditioning, respectively. With a 1, 0 coding for air conditioning and the model:

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

the beta coefficient that multiplies the air conditioning variable will estimate the difference in average sale prices of homes that have air conditioning versus homes that do not, given that the homes have the same square foot area.

Suppose we think that the effect of air conditioning (yes or no) depends upon the size of the home. In other words, suppose that there is interaction between the x-variables. To put an interaction into a model, we multiply the variables involved. The model here is

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \beta_3 x_{i,1}x_{i,2} + \varepsilon_i$$

The data are from $n = 521$ homes. Statistical software output follows. Notice that there is a statistically significant result for the interaction term.

```
              Estimate Std. Error t value Pr(>|t|)
(Intercept)    -3.218     30.085  -0.107 0.914871
SqFeet        104.902     15.748   6.661 6.96e-11 ***
Air           -78.868     32.663  -2.415 0.016100 *
SqFeet.Air     55.888     16.580   3.371 0.000805 ***
```

Software note: We would calculate a new variable by multiplying the square feet size and air conditioning variables. That variable would then be used as a predictor variable, along with the original *x*-variables.
The regression equation is:

$$\text{Average SalePrice} = -3.218 + 104.902 \times \text{SqrFeet} - 78.868 \times \text{Air} + 55.888 \times \text{SqrFeet} \times \text{Air}.$$

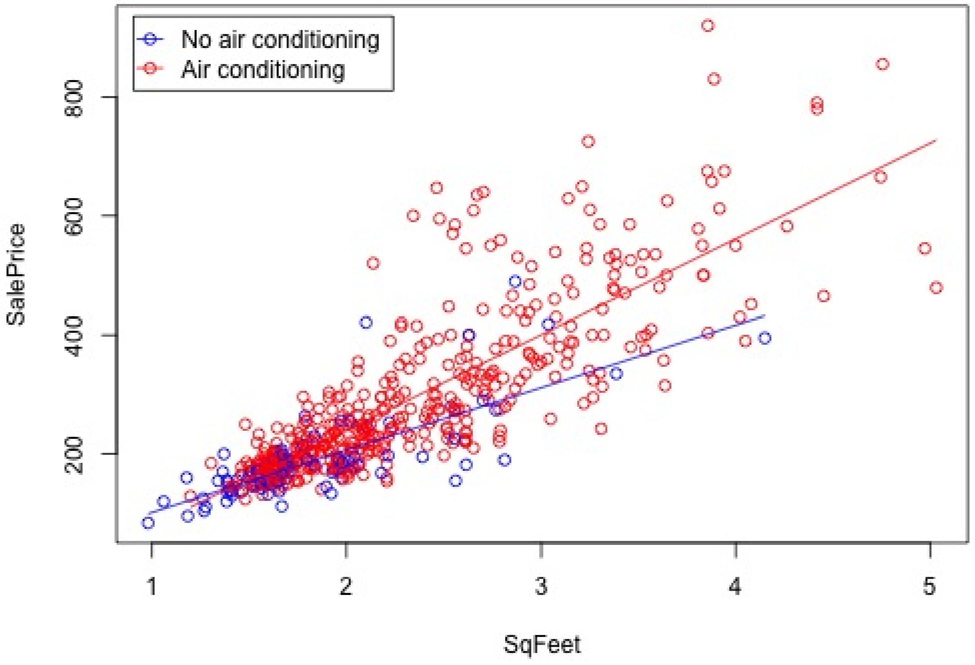
Suppose that a home has air conditioning. That means the variable Air = 1, so we'll substitute Air = 1 in both places that Air occurs in the estimated model. This gives us

$$\begin{aligned}\text{Average SalePrice} &= -3.218 + 104.902 \times \text{SqrFeet} - 78.868(1) + 55.888 \times \text{SqrFeet} \times 1\\ &= -82.086 + 160.790 \times \text{SqrFeet}.\end{aligned}$$

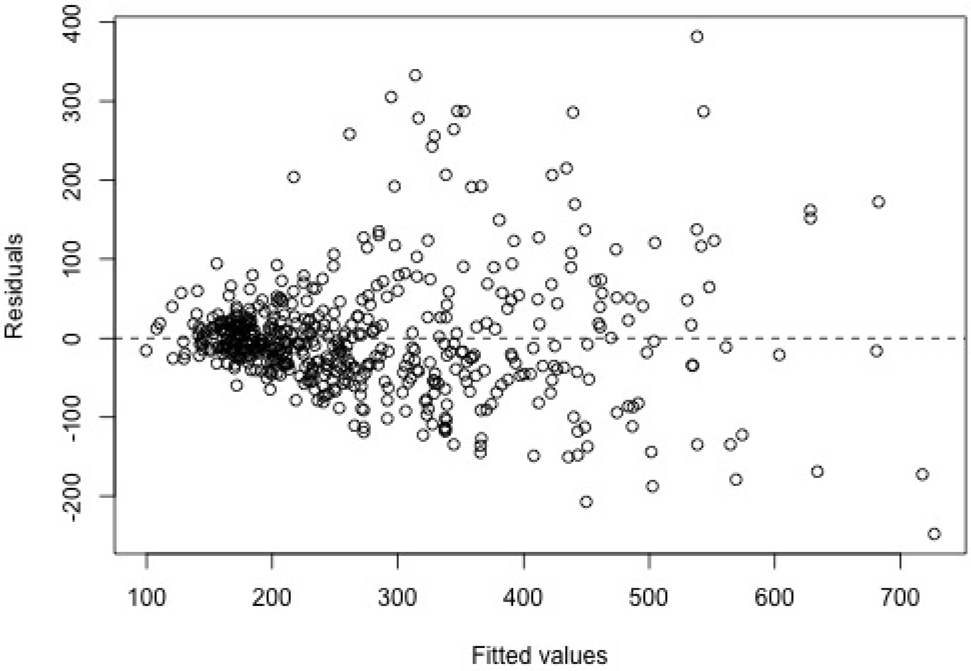
Suppose that a home does not have air conditioning. That means the variable Air = 0, so we'll substitute Air = 0 in both places that Air occurs in the estimated model. This gives us

$$\begin{aligned}\text{Average SalePrice} &= -3.218 + 104.902 \times \text{SqrFeet} - 78.868(0) + 55.888 \times \text{SqrFeet} \times 0\\ &= -3.218 + 104.902 \times \text{SqrFeet}.\end{aligned}$$

The figure below is a graph of the relationship between sale price and square foot area for homes with air conditioning and homes without air conditioning. The equations of the two lines are the equations that we just derived above. The difference between the two lines increases as the square foot area increases. This means that air conditioning versus no air conditioning difference in average sale price increases as the size of the home increases.



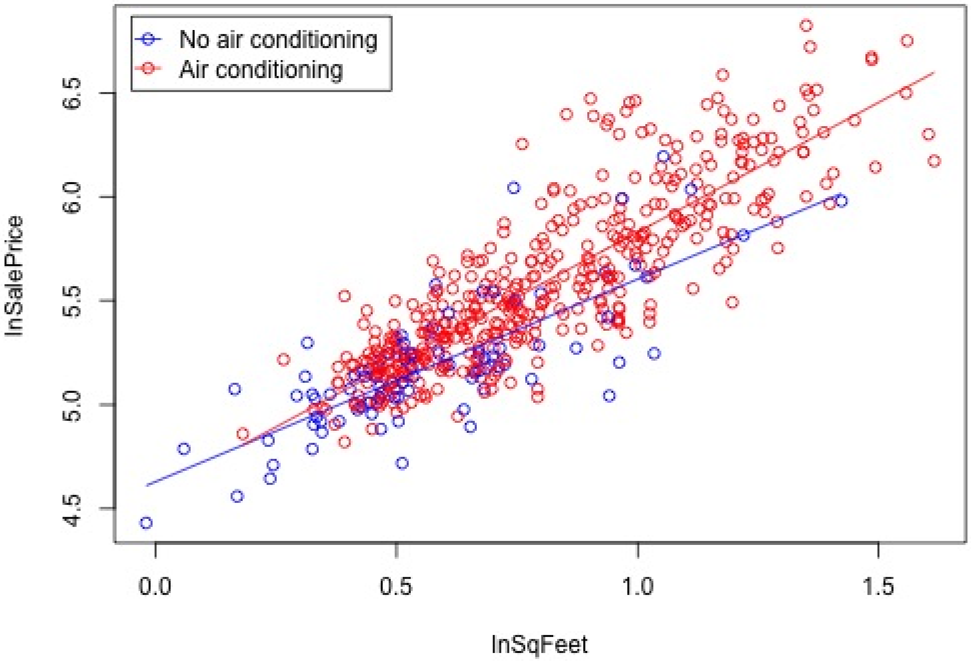
There is an increasing variance problem apparent in the above plot, which is even more obvious from the megaphone pattern in the following residual plot:



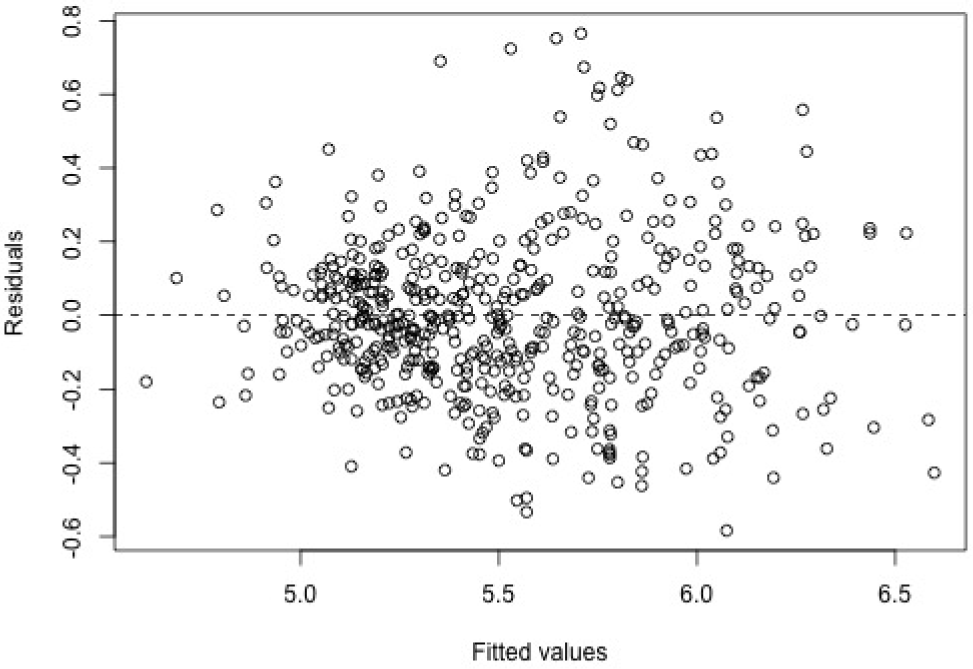
To remedy this, we'll try using log transformations for sale price and square footage (which are quite highly skewed). Now, Y = log(sale price), X_1 = log(home's square foot area), and X_2 = 1 if air conditioning present and 0 if not. After fitting the model:

$$y_i = \beta_0 + \beta_1\, x_{i,1} + \beta_2\, x_{i,2} + \beta_3\, x_{i,1}x_{i,2} + \varepsilon_i$$

the plot showing the regression lines is as follows:



and the residual plot, which shows a vast improvement, is as follows:



Example 3: Hospital Infection Risk Data

Consider the hospital infection risk data: infectionrisk.txt (/stat462/sites/onlinecourses.science.psu.edu.stat462/files/data/infectionrisk.txt) . For this example, the data are limited to observations with average length of stay ≤ 14 days. The overall sample size is $n = 111$. The variables we will analyze are the following:

Y = infection risk in hospital
X_1 = average length of patient's stay (in days)
X_2 = a measure of frequency of giving X-rays
X_3 = indication in which of 4 U.S. regions the hospital is located (north-east, north-central, south, west).

The focus of the analysis will be on regional differences. Region is a categorical variable so we must use indicator variables to incorporate region information into the model. There are four regions. The full set of indicator variables for the four regions is as follows:

I_1 = 1 if hospital is in region 1 (north-east) and 0 if not
I_2 = 1 if hospital is in region 2 (north-central) and 0 if not
I_3 = 1 if hospital is in region 3 (south) and 0 if not
I_4 = 1 if hospital is in region 4 (west), 0 otherwise.

To avoid a linear dependency in the **X** matrix, we will leave out one of these indicators when we forming the model. Using all but the first indicator to describe regional differences (so that "north-east" is the reference region), a possible multiple regression model for E(Y), the mean infection risk, is:

$$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 I_2 + \beta_4 I_3 + \beta_5 I_4.$$

To understand the meanings of the beta coefficients, consider each region separately:

- For hospitals in region 1 (north-east), $I_2 = 0$, $I_3 = 0$, and $I_4 = 0$, so

$$\begin{aligned}\mathrm{E}(Y) &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3(0) + \beta_4(0) + \beta_5(0)\\ &= \beta_0 + \beta_1X_1 + \beta_2X_2.\end{aligned}$$

- For hospitals in region 2 (north-central), $I_2 = 1$, $I_3 = 0$, and $I_4 = 0$, so

$$\begin{aligned}\mathrm{E}(Y) &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3(1) + \beta_4(0) + \beta_5(0)\\ &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3.\end{aligned}$$

- For hospitals in region 3 (south), $I_2 = 0$, $I_3 = 1$, and $I_4 = 0$, so

$$\begin{aligned}\mathrm{E}(Y) &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3(0) + \beta_4(1) + \beta_5(0)\\ &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_4.\end{aligned}$$

- For hospitals in region 4 (west), $I_2 = 0$, $I_3 = 0$, and $I_4 = 1$, so

$$\begin{aligned}\mathrm{E}(Y) &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3(0) + \beta_4(0) + \beta_5(1)\\ &= \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_5.\end{aligned}$$

A comparison of the four equations just given provides these interpretations of the coefficients that multiply indicator variables:

- β_3 = difference in mean infection risk for region 2 (north-central) versus region 1 (north-east), assuming the same values for stay (X_1) and X-rays (X_2).
- β_4 = difference in mean infection risk for region 3 (south) versus region 1 (north-east), assuming the same values for stay (X_1) and X-rays (X_2).
- β_5 = difference in mean infection risk for region 4 (west) versus region 1 (north-east), assuming the same values for stay (X_1) and X-rays (X_2).

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -2.134259   0.877347  -2.433  0.01668 *
Stay         0.505394   0.081455   6.205 1.11e-08 ***
Xray         0.017587   0.005649   3.113  0.00238 **
i2           0.171284   0.281475   0.609  0.54416
i3           0.095461   0.288852   0.330  0.74169
i4           1.057835   0.378077   2.798  0.00612 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.036 on 105 degrees of freedom
Multiple R-squared:  0.4198,    Adjusted R-squared:  0.3922
F-statistic: 15.19 on 5 and 105 DF,  p-value: 3.243e-11
```

Some interpretations of results for individual variables are:

- We have statistical significance for the sample coefficient that multiplies I_4 (p-value = 0.006). This is the sample coefficient that estimates the coefficient β_5, so we have evidence of a difference in the infection risks for hospitals in region 4 (west) and hospitals in region 1 (north-east). The positive coefficient indicates that the infection risk is higher in the west.

- The non-significance for the coefficients multiplying I_2 and I_3 indicates no observed difference between mean infection risks in region 2 (north-central) versus region 1 (north-east) nor between region 3 (south) versus region 1 (north-east).

Next, the finding of a difference between mean infection risks in the north-east and west seems to be strong, but for the sake of example, we'll now consider an overall test of regional differences. There is, in fact, an argument for doing so beyond "for the sake of example." To assess regional differences, we considered three significance tests (for the three indicator variables). When we carry out multiple inferences, the overall error rate is increased so we may be concerned about a "fluke" result for one of the comparisons. If there are no regional differences, we would not have any indicator variables for regions in the model.

- The null hypothesis that makes this happen is $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$.
- The reduced model is simply $\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$. This model has SSE = 123.56 with error df = 108.
- The full model is $\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 I_2 + \beta_4 I_3 + \beta_5 I_4$, the model that we have already estimated. This model has SSE = 112.71 with error df = 105.

The test statistic for $H_0 : \beta_3 = \beta_4 = \beta_5 = 0$ is the general linear F-statistic calculated as

$$F = \frac{\frac{\text{SSE(reduced) - SSE(full)}}{\text{error df for reduced - error df for full}}}{\text{MSE(full)}} = \frac{\frac{123.56 - 112.71}{108 - 105}}{\frac{112.71}{105}} = 3.369.$$

The degrees of freedom for this F-statistic are 3 and 105. We find that the probability of getting an F statistic as extreme or more extreme than 3.369 under an $F_{3,105}$ distribution is 0.021 (i.e., the p-value). We reject the null hypothesis and conclude that at least one of β_3, β_4, and β_5 is not 0. Our previous look at the tests for individual coefficients showed us that it is β_5 (measuring the difference between west and north-east) that we conclude is different from 0.

Finally, the results seem to indicate that the west is the only regional difference we see that has a higher infection risk than the other three regions. (If the north-central and south regions don't differ from the north-east, it is reasonable to think that they don't differ from each other as well.) We can test this by considering a reduced model in which the only region indicator is $I_4 = 1$ if west, and 0 otherwise. The model is

$$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_5 I_4.$$

The null hypothesis leading to this reduced model is $H_0 : \beta_3 = \beta_4 = 0$. This model has SSE = 113.11 with error df = 107.

The full model is still

$$\mathrm{E}(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 I_1 + \beta_4 I_2 + \beta_5 I_3,$$

which has SSE = 112.71 with error df = 105. Finally,

$$F = \frac{\frac{\text{SSE(reduced) - SSE(full)}}{\text{error df for reduced - error df for full}}}{\text{MSE(full)}} = \frac{\frac{113.11 - 112.71}{107 - 105}}{\frac{112.71}{105}} = 0.186.$$

The degrees of freedom for this F-statistic are 2 and 105. We find that the probability of getting an F-statistic as extreme or more extreme than 0.186 under an $F_{2,105}$ distribution is 0.831 (i.e., the p-value). Thus, we cannot reject the null hypothesis and conclude that the west differing from the other three regions seems to be reasonable.

‹ 8.7 - Leaving an Important Interaction Out of a Model (/stat462/node/166) up (/stat462/node/86)