In multiple regression, in addition to interaction, there is one more thing to worry about: **collinearity**.

Let's return to the first (important) step: Look at data! Because there are multiple predictors, there is a matrix of scatterplots:

![Collinearity Diagram](image)

- A linear correlation between \( x_1 \) and \( x_2 \) is called **collinearity**.
- It's a "bad" disease (see below, for why) one cure is to simply exclude one of the predictors \( x_1, x_2 \).

A consequence of collinearity is that it renders the \( \beta \)'s uninterpretable (as the avg. rate of change of \( y \) ---):

- Ordinarily, in \( y = \alpha + \beta_1 x_1 + \beta_2 x_2 \) \( \beta_1 = \text{avg. rate of change in } y \text{ for 1 unit change in } x_1 \) but if \( x_1 \) and \( x_2 \) are correlated, then one cannot hold one of them fixed.

In fact, in an example, \( \text{age} = \alpha + \beta_1(\text{health}) + \beta_2(\text{income}) \) I once got a value of \( \beta_1 \) that was negative, in spite of the positive association displayed in the scatterplot of age vs. health. The culprit was collinearity.
Geometrically, the reason why the \( \beta \)'s become uncertain and uninterpretable is that we are then trying to fit a plane through a cigar-shaped cloud in 3D, as opposed to a planar cloud.

That is ambiguous! There are lots of planes one can fit through a cigar-shaped cloud in 3D. Of course, those different fits differ in their \( \hat{\beta}, \hat{\beta}_1, \hat{\beta}_2 \). That's why they become meaningless. You can also see that the predictions, \( \hat{y} \), are affected by collinearity; however, note that the effect is mostly in their uncertainty. (Move, in Ch. II)

Another bad consequence of coll. is that it effectively reduces the amount of information in the data, which, in turn, leads to more uncertain estimates of the \( \beta \)'s and predictions. We'll see that in Ch. II.

Another bad consequence of coll. is that it can lead to overfitting. This is because the various predictors come with many parameters to be estimated from data, but the various predictors are essentially carrying the same information, i.e. there is effectively more params than data, hence overfitting can happen.
Finally, if there is no collinearity at all, then multiple regression
\[ Y = \alpha + \beta_1 X_1 + \beta_2 X_2 + \ldots \]
will give the same \( \hat{\beta} \) as a bunch of simple regressions:
\[ Y = \alpha + \beta_1 X_1, \quad Y = \alpha + \beta_2 X_2, \quad \ldots \]

So, some collinearity is a good thing; too much is not.

In summary, even though both interaction and collinearity make the \( \beta \)'s uninterpretable, they are very different concepts.

collinearity ≠ interaction.

Q1: Suppose data on \( X_1, X_2, Y \) follow a saddle pattern. Then we have collinearity.

A) Yes  B) No  C) cannot tell.
For different levels of collinearity, the problem of uncertain β's and predictions can be qualitatively different. For very little collinearity, there is a reasonably unique plain one can fit the black dots. For mild collinearity (red), there is no unique surface to fit the "cigar." For extreme collinearity (blue), the "fit" is a "vertical" surface. Think about what this does to the predictions.
Visual assessment of fit quality

Plot \( \hat{y} \) vs. \( y \):
- Random, symmetric about diagonal (Good)
- Need a quadratic model (Bad)

Residual Plot: (Mona Lisa)
- Errors (last column of "table" before)
- Random, symmetric about error = 0 line (Good Model/fit)
- Need a quadratic model (Bad Model/fit)

Things to do:
1) Transform data, or
2) Fit different polynomial
This is FYI, but it will come-up in tomorrow’s lab.

We know that one can overfit data on $x,y$, if one uses a high-order polynomial in poly. regression. Recall that the main reason this can happen is because such a regression model will have a lot of parameters.

In multiple regression there is yet another way that overfitting can happen even w/o including high-order terms in the model.

Consider 3 cases on $y$ and $x_1$:
- A model like $y = \alpha + \beta x_1$ (a line)

  cannot over-fit that data (a plain)

But a model like $y = \alpha + \beta_1 x_1 + \beta_2 x_2$ overfits completely. The reason is because in that case the 3 cases are in 3D (not 2D), and there is always one plane that goes thru 3 points exactly.

Note that the additional variable $x_2$ can even be completely unrelated to $y$! It can even be just random values!

In other words, by arbitrarily making the space big, we opened up the possibility of overfitting.

So one can overfit even a multiple regression model without any non-linear (eg. quadratic, cubic, ...) terms.

You may think this is happening only because I have 3 cases here. But even with more cases, one can still overfit by simply including more (even random) predictors in the model, if there are many more params in regression than cases.

This overfitting problem is not specific to regression. **All** models can overfit when they are too large. CS students: WATCH OUT!
For each of the data sets a) hw_3_dat1.txt and b) hw_3_dat2.txt, find the "best" (OLS) fit, and report R-squared and the standard deviation of the errors. Do not use some ad hoc criterion (like maximum R2) to determine what is the "best" model. Instead, use your knowledge of regression to find the best model, and explain in words why you think you have the best model. Specifically, make sure you address 1) collinearity, 2) interaction, and 3) nonlinearity.

a) Read the data file transform_data.txt from the course website into R, and make a scatterplot of y versus x. Clearly, the relationship is nonlinear and monotonic. I can tell you that a good transformation that linearizes the relationship is to take the sqrt of both x and y. Make a scatterplot of the transformed data.

b) Perform regression on the transformed data, and overlay the regression line on the scatterplot of the transformed data in part a).

c) Fit a regression model of the form y = alpha + beta_1 sqrt(x) + beta_2 x to the original (untransformed data).

d) In a clicker question I claimed that these two models are essentially equivalent. To check that, let's see if they make similar predictions. Make a scatterplot to compare their predictions. Just keep in mind that the second model predicts y, but the first model predicts sqrt(y).

Write code to read in hw_3_mult_simple_dat.txt from the course website. The first two columns are x1 and x2, and the third column is the response y.

a) Write code estimate the regression coefficients in a multiple regression model involving both x1 and x2 as predictors. Do not include any interactions or higher-order terms.
b) Write code to estimate the regression coefficients in a simple regression model involving only x1 as predictor. Repeat for a model involving only x2 as predictor.
c) There is something curious about all the estimates of the beta parameters. What is the curiosity?
d) Provide a brief explanation based on everything discussed in the lecture, and write code to support your explanation.

a) Read the data file bias_0_data.txt into R (it's on the course website), perform regression to predict y from x, make the scatterplot of the predictions versus the observed y values, and overlay a diagonal line (y-intercept=0, slope=1) on it. BUT, because we want to diagonal line to actually appear as diagonal, make sure the range of x and y values shown in the scatterplot is the same; in fact, set that range to (-6,6) for both x and y values. If you don't know how, check the prelabs, looking for xlim and ylim.

b) Now, read in the data file bias_1_data.txt, perform regression, and overlay on the previous plot (in part a) the scatterplot of predictions versus the observed y values. Make these points red. If done correctly, you will see that the predictions are now all positively biased (i.e., consistently shifted up).

c) The scatterplot in part a looks good in that it does not suggests any problems with the model. However, as discussed in class, the scatterplot in part b suggests a positive bias (the predictions are consistently higher than the observed values). Why? Hint: There is something about the data that is causing this bias. What is it?

Write code to read in hw_3_mult_simple_dat.txt from the course website. The first two columns are x1 and x2, and the third column is the response y.

a) Write code estimate the regression coefficients in a multiple regression model involving both x1 and x2 as predictors. Do not include any interactions or higher-order terms.
b) Write code to estimate the regression coefficients in a simple regression model involving only x1 as predictor. Repeat for a model involving only x2 as predictor.
c) There is something curious about all the estimates of the beta parameters. What is the curiosity?
d) Provide a brief explanation based on everything discussed in the lecture, and write code to support your explanation.

a) Read the data file transform_data.txt from the course website into R, and make a scatterplot of y versus x. Clearly, the relationship is nonlinear and monotonic. I can tell you that a good transformation that linearizes the relationship is to take the sqrt of both and x and y. Make a scatterplot of the transformed data.

b) Perform regression on the transformed data, and overlay the regression line on the scatterplot of the transformed data in part a).

c) Fit a regression model of the form y = alpha + beta_1 sqrt(x) + beta_2 x to the original (untransformed data).

d) In a clicker question I claimed that these two models are essentially equivalent. To check that, let's see if they make similar predictions. Make a scatterplot to compare their predictions. Just keep in mind that the second model predicts y, but the first model predicts sqrt(y).