Writing the model as

$$Y = \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon$$

tells you how a unit increase of X_1 interacts w/ a given value of X_2

- You could probably use $\partial/\partial x_1$ as well.

- <u>Hierarchical principle</u>: If you include <u>interaction</u>, include <u>main effect</u> terms too.

2/8/23 <u>Day 7</u>

- We can have interactions of qualitative variables w/ qual or quant variables.

ex) Credit data set. Want to predict balance using income (quantitative) and student (qualitative) variables. With no intxn terms we get a model of the form

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \beta_2 \times \text{student}_i$$

$$= \beta_1 \times \text{income}_i + \begin{cases} \beta_0 + \beta_2 & i \text{ is student} \\ \beta_0 & i \text{ is not student} \end{cases}$$

- Fit two parallel lines to the data for two sets (students & non-students). This is a model limitation.

If we include the intxn term, we get the model

$$\text{balance}_i \approx \beta_0 + \beta_1 \times \text{income}_i + \beta_2 \times \text{student}_i + \beta_3 \times \text{income}_i \times \text{student}_i$$

$$= \begin{cases} (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \times \text{income}_i & i \text{ student} \\ \beta_0 + \beta_1 \times \text{income} & \text{else} \end{cases}$$

Now the intercept <u>and</u> slope change w/ student status

<u>Non-linear relationships</u>

3.3.3 Potential Problems

1. Non-linearity of the response-predictor relationship
2. Correlation of error terms
3. Non-constant variance of error terms
4. Outliers
5. High-leverage points
6. Collinearity

1. Non-linearity of the data

- Harms inference & prediction.

You can plot the residuals vs. the predicted values (ε_i vs. $\hat{y}_i$). If the resulting plot shows no significant trend away from a flatline at 0 you're good. If so, you may need a non-linear model.

2. Correlation of Error Terms

- Means the values of e_i, e_j are independent $\forall i \neq j$.
- If they are correlated, then standard errors of the parameters are much higher, meaning true confidence ivls and true p-values will be higher than those produced by software.
- Correlated residuals often occur in time-series data

- For time series data you can plot the residuals vs. time and check for tracking (similar values between adjacent residuals). ACF may work too.

- Other data can have correlated errors too:

ex) Predict height from weight: errors could be correlated if individuals come from the same family, diet, or environment.

- To solve, need good experimental design.

3. Non-constant variance of the error terms

- Heteroskedasticity violates an assumption in the ~~calculation~~ estimation of the coefficients.
- You can find this if the residual plot has a funnel shape.
- One solution is to apply a concave function to the response because larger responses are shrunk down
  - $\sqrt{\cdot}$, $\log(\cdot)$ do the trick
- If you expect to know the standard error, you can also do weighted least squares

4. Outliers

- An outlier is a data point for which y_i is far from the model value.
  - Inflate RSE and R^2
- Residual plots can help identify standard error, but it can be hard to tell.
- Studentized residuals — $\varepsilon_i / SE(\varepsilon_i)$ — give a standardized error so that if the value is greater than 3 in absolute value, it's a likely error.
- Be careful when removing outliers! A couple is okay, esp. if you find that they're due to collection errors but many outliers can point to (a) missing predictor(s) or some other issue.

5. High Leverage Points

- High leverage points have an unusual predictor value.
- These tend to have a high impact on the ~~line~~ regression line
- A ~~predictor~~ predictor point may have fairly normal values on each individual ~~~~ value, but be far from the center. (Think distance from center of the elipse).

Need each point's leverage statistic. For simple linear regression: leverage is given by (Meaning one variable.)

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^2}{\sum_{i=1}^{n} (x_i - \bar{x})^2}$$

For multiple linear regression it's probably something like

$$h_i = \frac{1}{n} + \frac{(x_i - \bar{x})^{\otimes 2}}{\sum_{i=1}^{n} (x_i - \bar{x})^{\otimes 2}}$$

- $h_i \in [1/n, 1]$, average leverage is always $(p+1)/n$. So if some point has ~~high~~ leverage significantly higher than $(p+1)/n$, we should be suspicious of high leverage
- ☆ Outliers with high leverage can really screw up a model. (Plot standardized residuals vs. leverage to see).

6. Collinearity