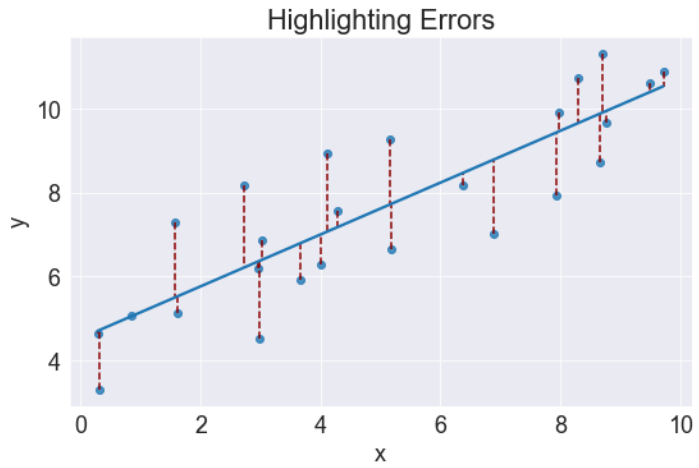
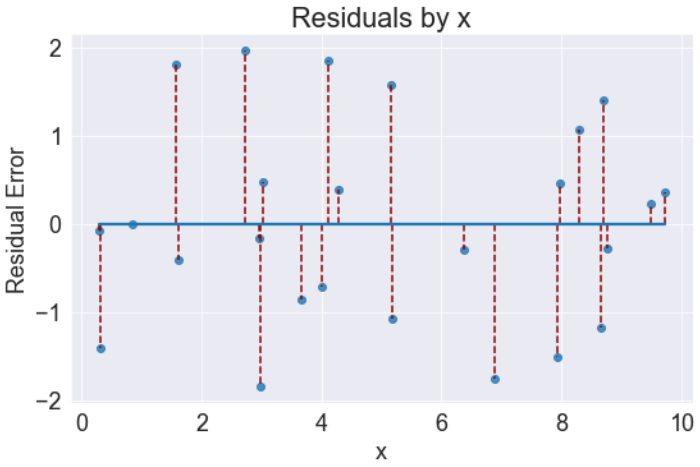
Model Adequacy

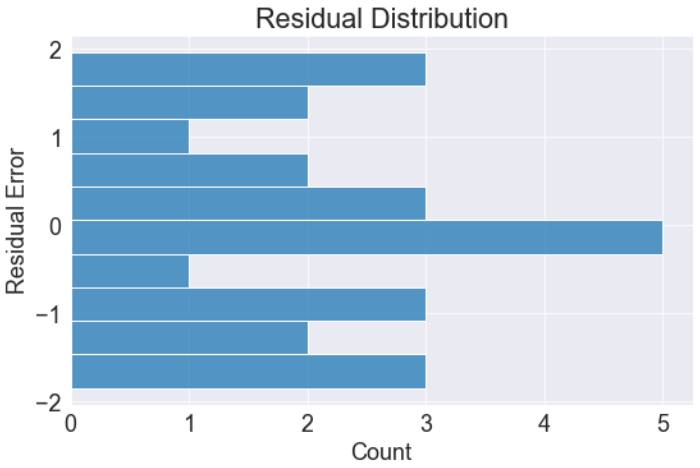
DASC 512

Assumptions (Last Week)

1. The mean value of ϵ is zero for all values of x
2. The variance of ϵ (some σ^2), is constant for all values of x
3. ϵ is normally distributed
4. Each ϵ_i is iid (independent and identically distributed)





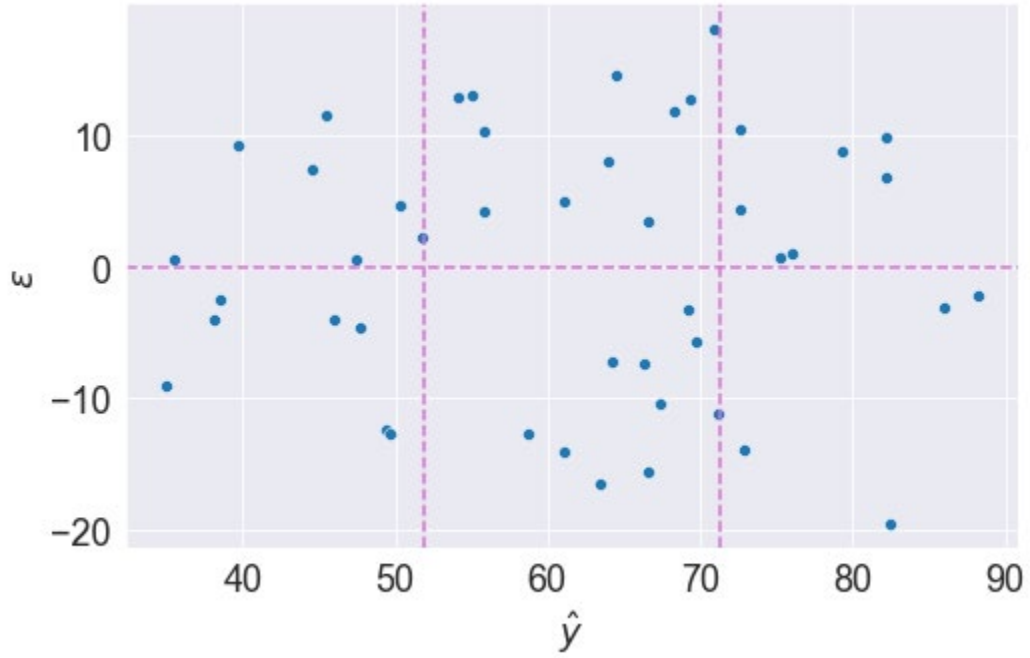


1. The mean value of ϵ is zero

As with simple regression, this is still built into the method

BUT – the mean should be zero for all predicted values

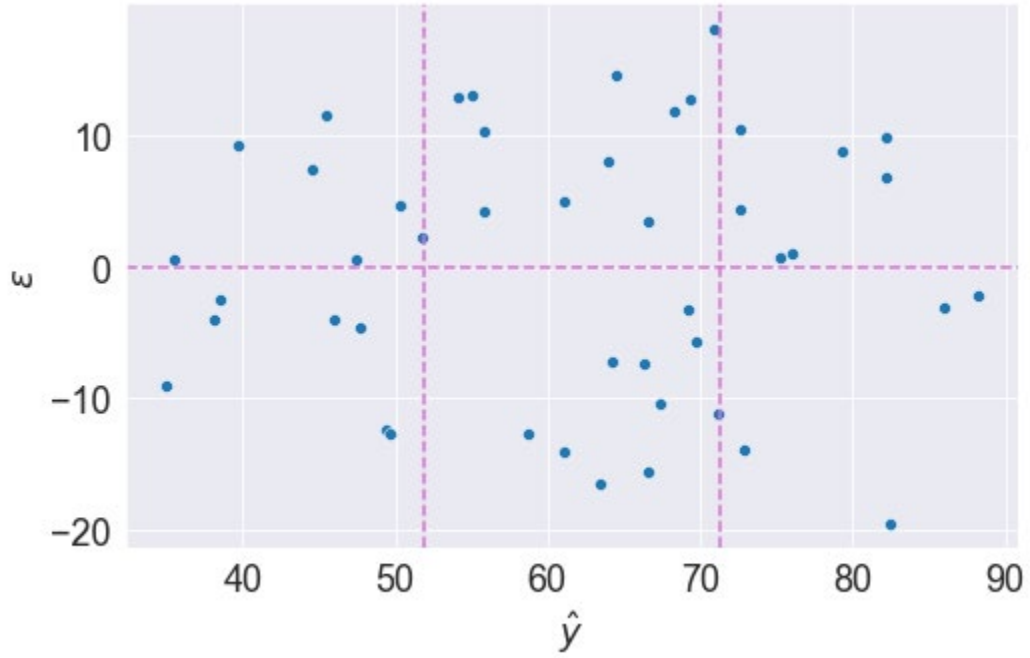
We can identify problems by looking at plots of residual vs predicted value



2. The variance of ϵ (some σ^2), is constant for all values of x

In simple regression we looked at plots of x vs residuals

In multiple regression, we'll look at plots of residual vs. predicted value



Rectifying non-constant mean/variance

Unequal mean and variance can typically be fixed by transforming y

- Common transformations include $\ln y$ and y^{λ}
- Try a transform, refit the model, and check plots again
  - This modified model will lose ease of inference

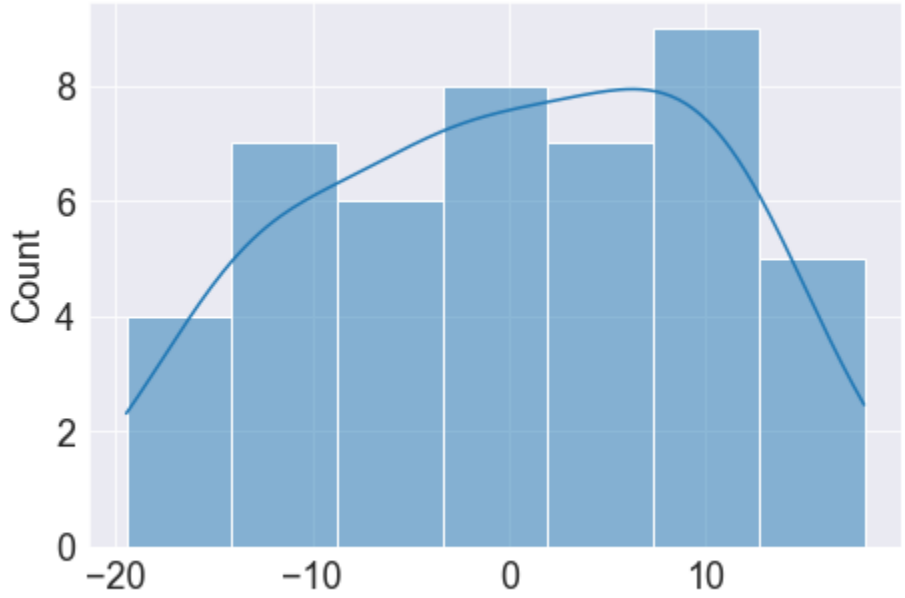
A useful method is the Box-Cox Transformation

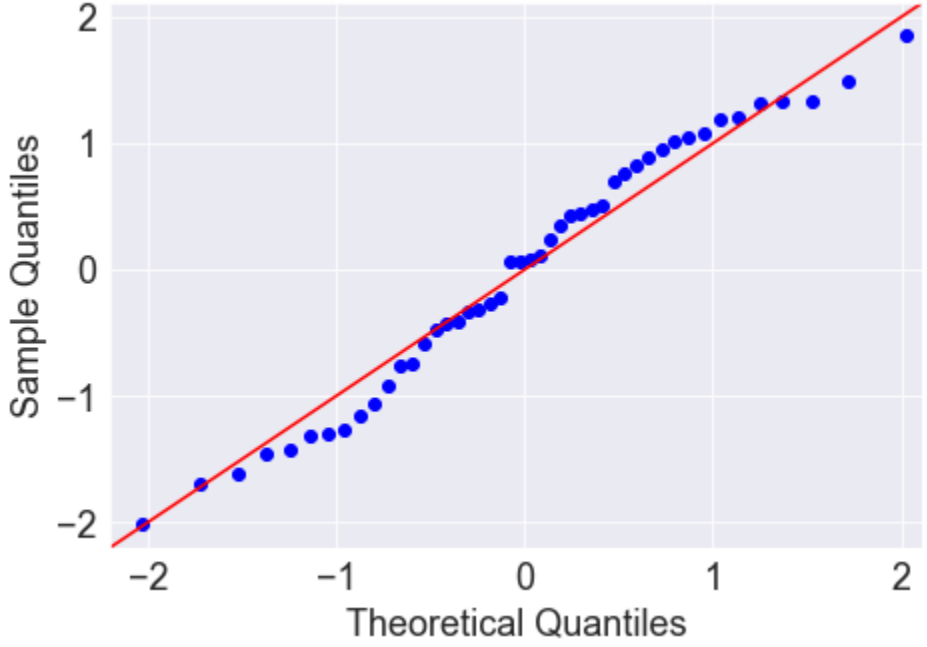
- This finds the optimal value of λ for a y^{λ} transformation to achieve normality
- It also implements $\ln y$ when $\lambda = 0$
- yt, lamb, ci = stats.boxcox(y, alpha=0.05)

3. ϵ is normally distributed

Python automatically conducts normality tests on the residuals (see bottom of summary table)

It is still good practice to check visually with histograms and QQ-plots.





Rectifying non-normality

Slight departures from normality will generally not be cause for concern

- Regression is robust to non-normality, especially for large samples
- Outliers will often result in significant normality tests
- This is part of why visual analysis is so important

For major departures, a transformation may again be required

- If the cause is non-normality of one of the independent variables, consider transforming one or more of them instead

4. Errors are independent

Just like before, we can try to identify this by examining ordered plots of the residuals

- We talked about time plots, but could also be spatial, etc.

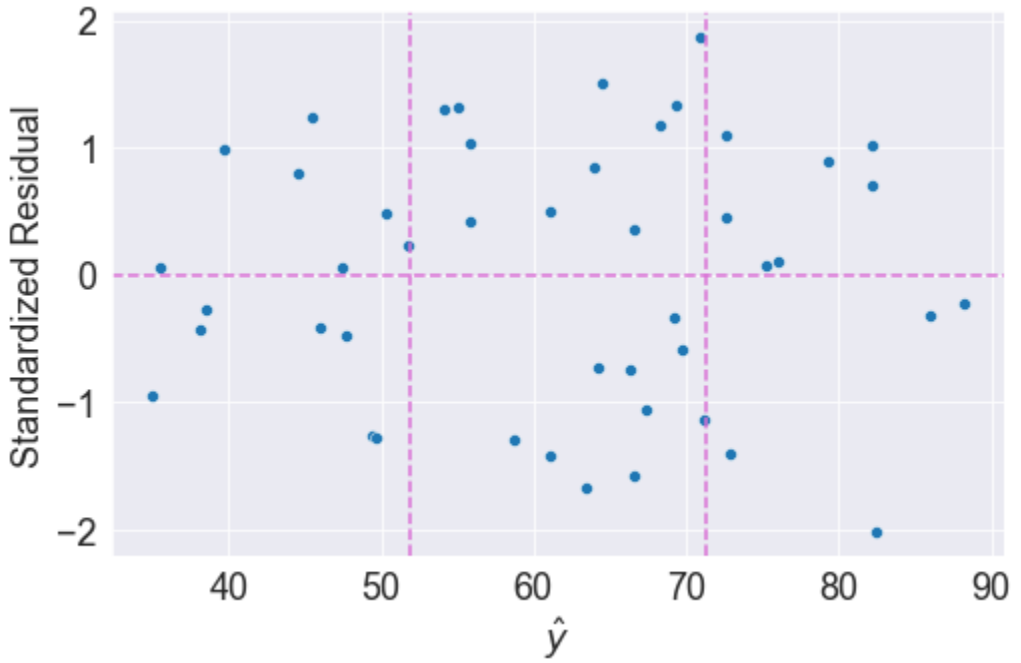
Rectifying independence problems will require time-series modeling, which is beyond the scope of this course

Other considerations

Outliers

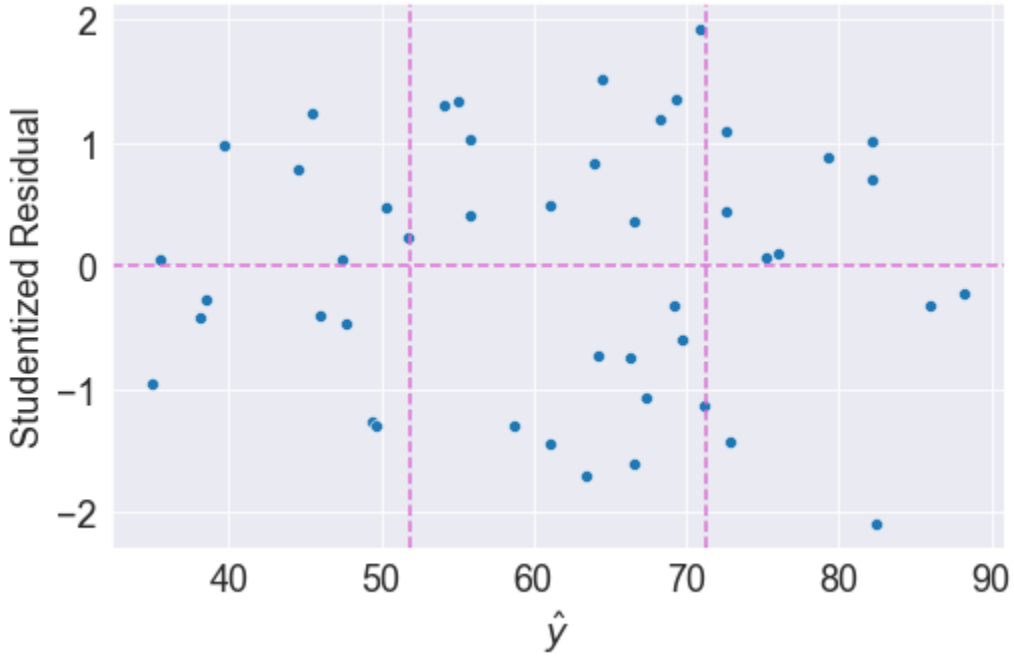
Standardized Residuals: $\frac{\epsilon}{\sqrt{MSE}}$

influence = model.get_influence()
influence.resid_studentized_internal



Outliers are not necessarily a bad thing depending on their location. They can have undue effect on the model, or influence

Influence



Influence is a measure of how much the linear model is affected by the presence of a single data point

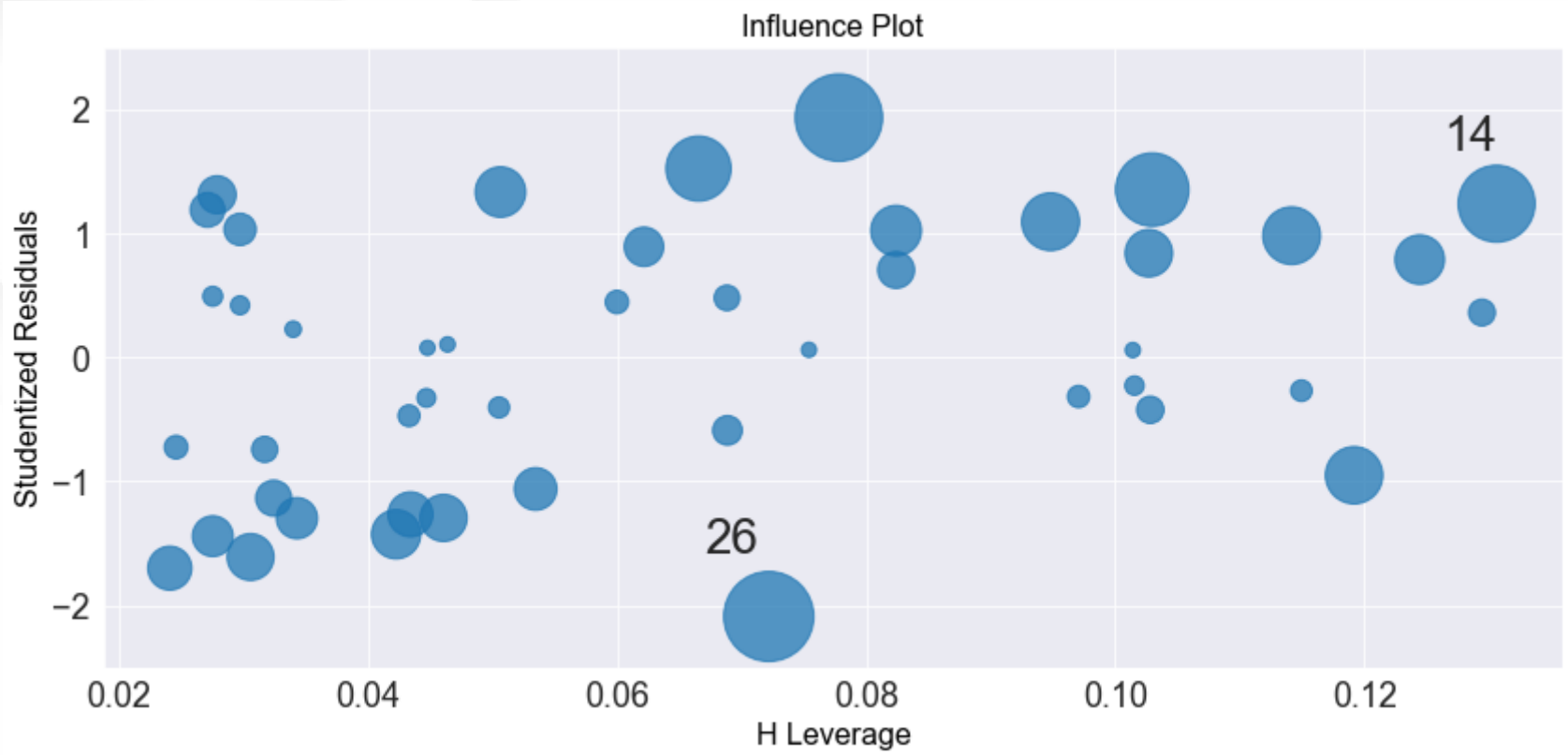
One way to measure influence is with Studentized residuals

Studentized Residuals: Standardized residuals when the model is fit without that point

influence.resid_studentized_external

Influence

Another method is to look at Influence Plots
smg.influence_plot(model)



Overfitting and Underfitting

Overfitting: Captures some of the random noise as part of the model. Typically caused by over-parameterization in regression.

Underfitting: Captures not enough of the deterministic part of the true relationship. Typically cause by under-parameterization in regression.

Both lead to poor predictions on new data sets.

Again, it is best practice to use training, validation, and test sets when you have enough data. The amount of split is up to you as the modeler.

We'll do this in later examples.

Multicollinearity

Multicollinearity occurs when the independent variables are correlated

- May cause t-tests to be non-significant
- May inflate standard errors for $\hat{\beta}$ estimates
- May cause parameters to be opposite-sign from expected

A good diagnostic is the Variance Inflation Factor

- VIF=5 is considered moderate
- VIF=10 is considered something to worry about

Multicollinearity

Multicollinearity is not a concern for

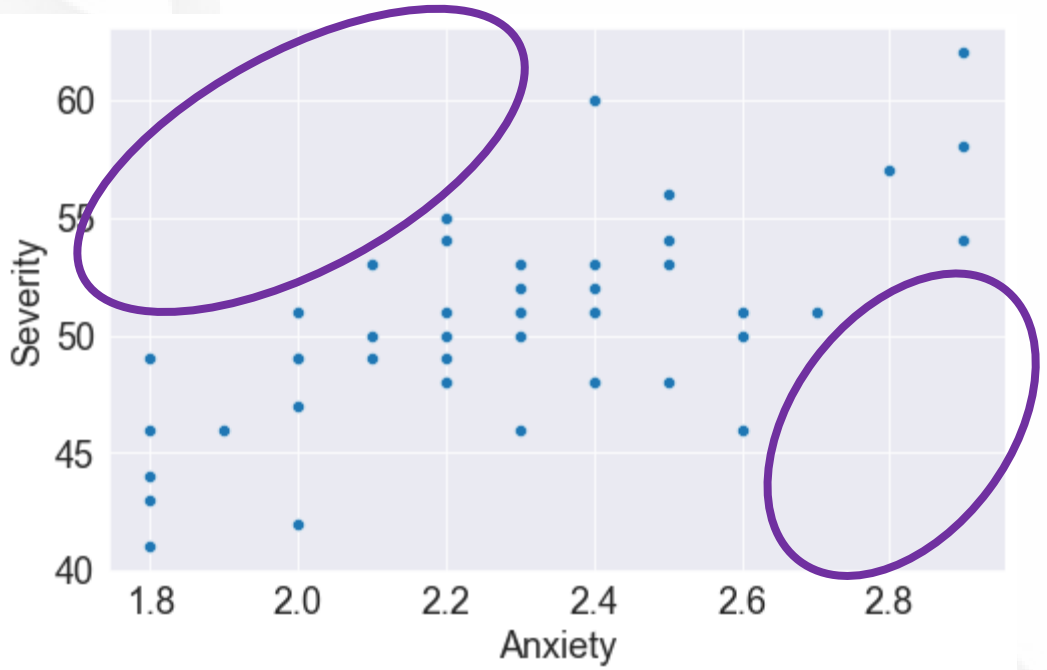
- The intercept
- Higher-order terms being collinear with main effects

If there are issues, you can

- Center/standardize the offending variables
- Remove one or more variables
- Don’t make inferences on $\hat{\beta}$

Extrapolation

- Extrapolation is more complex in the multivariate model
- It is still just as important as in the univariate model!



Next week on DASC 512…

Lots and lots of practice…