Modeling

Suppose we had a linear regression model and we just added spurious variables that are essentially just noise

- If we compute the least squares regression model using the spurious models versus without, we will find a better fit
- One interpretation of this is that we are essentially projecting onto a higher subspace by increasing the number of features
  - Remember that an intuitive interpretation of linear regression is projecting a n dimensional Y onto the k dimensional column space of X
  - The codimension of the column space of X is therefore $n - k$
    - Codimension of a subspace = dimension of main subspace - dimension of subspace
  - Increasing the dimension of X therefore decreases the codimension and leaves less variables in the residual, which exists in a subspace that has dimension equal to the codimension
  - Therefore, the residual length will naturally decrease
- In fact, if we add enough spurious variables to make the dimension of X equal to n, then we will get a perfect fit
- This is a phenomenon known as **overfitting**

A lot of times in the real world, we have a ton of variables, many of which are junk and we need to remove some of them

- This is known as **model selection**
- One way to do this is to do a test for $\beta_j^* \neq 0$ for each of the variables (Wald's test)
  - However we are doing multiple hypothesis tests and therefore need some kind of correction
    - Bonferonni correction is too conservative and BH requires the test statistics to be independent, which is not true here
  - This also only really works in the asymptotics and we are dealing with an effective sample size of $n - k$
    - In some cases, we might have $n < k$
- In reality, what we really want to do is test if $\beta_j^* = 0$ (kind of like the inverse of Wald's test)
  - Our null hypothesis is that it is $\neq 0$
  - But our current hypothesis testing framework doesn't allow us to do this

R Squared

We want to assign a score to each subset S of the variables that we choose

- We can't just choose the residual length because that would just be all of the variables
- We want a dimensionless value between 0 and 1

R Squared

The R^2 or coefficient of determination measures the the fit of the model to the data:

$$R^2(S) = 1 - \frac{\left|\left|Y - X\hat{\beta}(S)\right|\right|^2}{\left|\left|Y - \overline{Y}_n\right|\right|^2}$$

- The denominator is the sum of squared residuals for when we use the best constant function to fit
  - We use the constant $\mathbb{E}[Y]$
- If at least one of the variables in X for every instance is 1, then we can guarantee that this is at least 0
  - That is, we need to use linear regression with a constant term
  - This is because then it will do at least as well as the denominator
- The larger the set S is, the larger this will be, so we will need a way to avoid overfitting

Forward / Greedy Model Selection

Initialize S to the empty set. Then, add the feature that causes the greatest increase in the R^2. Stop once you start to plateau or when you exceed some preset R^2 value

There are other ways to choose scores as well:

- There are some that penalize a larger $|S|$ and then we just want to maximize it