Emily Chen | Scott | STA 371H
3/24/14 | 11AM - MW

In this class period:

- First we will talk about the Hypothesis testing this semester compared to what we learned last semester.
- Then we will finish up the google flu trends forecasting problem from last class.
- Then, we will talk about model building in context of answering focused questions.

The new idea introduced this period is: model building in the context of **answering focused questions** and **forecasting.** (when we say model building→ we mean choosing predictors to include in the model)

There is a spectrum of kinds of problems we will encounter (we use statistical methods to do both of these processes, but use different techniques for each):

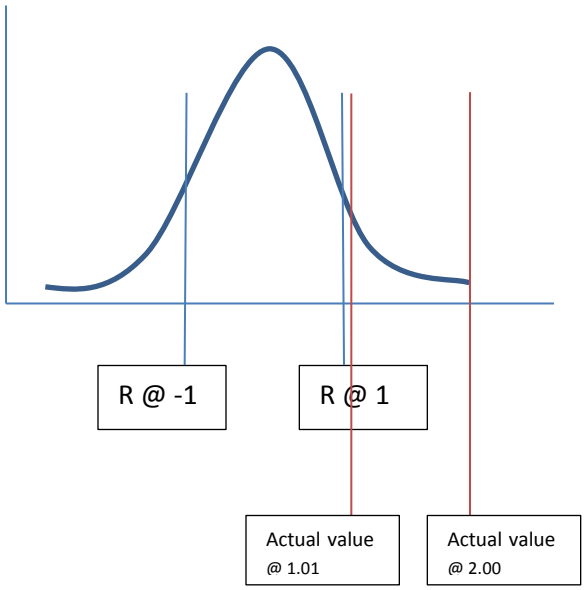
Forecasting Problems	**Very Narrow Focused Questions**
We just want to digest data and forecast predictions	We just want to answer 1 focused question
Ex. Google Flu Trends	Ex. Seatbelt Data Problem

We will use seatbelts .R and seatbelts.csv files in class to see the effect of seatbelts on traffic accident fatalities.

We will first finish off discussion from last class period about **Hypothesis testing**

- Let's talk about **p-values**
  - We remember p-values from STA 309H hypothesis testing
  - But so far, in this class we've been using Neyman Pearson hypothesis testing for weeks and we haven't used p-values yet
- Example: In the cps85 data set, we wanted to see if there was a wage premium for men over women
  - We found the average wage for each gender : 'mean(wage~sex, data=cps85)'
  - We then found this same data in baseline offset form: lm(wage~sex, data=cps85)
    1. The wage premium is around $2.12
  - The null hypothesis is that any difference in wage premium is simply due to chance of our sample
  - Let's walk through **Neyman Pearson Hypothesis testing!**
    1. Our null was there was no wage premium for men over women
    2. The test statistic is decided to be the offset for men's salaries over women's
    3. We then did a perm test to shuffle the sex to simulate the sampling distribution for the test statistic that we see , assuming the null hypothesis is true
    4. We choose a sensible rejection region
    5. We find the size of the rejection region (**alpha**)
    6. We find where our actual difference was and plot it on our histogram. Was this actual value in the rejection region (yes!). →So we can reject it.
       - If our observed wage premium was just 1 cent away from not being rejected (but still in the rejection region), it would still be rejected. Is

this fair that something so close to not rejecting the null would be rejected with the same certainty (look like the same result) as something 1 dollar away from not being rejected? This doesn't seem fair that they would be rejected with the same amount of certainty even though one choice is obviously less worthy of rejection than the other.

- 

  - To account for the difference in the amount of certainty in our rejection of these values, we use **p-values**

- So what was a p-value again?
  - DEFINITION: **P-value** is the probability of observing a **test statistic** as extreme as or more extreme than you "t", given that H_o is true.
    - If we changed "t" to a pre-specified critical value, this definition would define the **alpha level**
  - **Steps for using p-value**
    - We define a null hypothesis
    - We define out test stat as our wage premium
    - We do a perm test to shuffle sec to simulate distribution for our t-statistic, assuming the null hypothesis is true
    - We plot the observed t-statistic on the histogram and find the area under the curve that is at or more-extreme than the "t"
  - Dr. Scott is not a fan of p-value because it is incredibly difficult to understand→ so just know the idea of it for statistical literacy
  - **On an intuitive level:** it is a numerical summary of the evidence of the strength of your t-statistic against the null hypothesis
  - **Just make sure, that you don't confused p-values with alpha values!**

- In the textbook, Dr. Scott makes the difference between alpha and p-value very clear

Now, let's go to the Google Flu Trends Data (**forecasting**):

- Load data from googleflu.R and google.csv into R
- Interpreting the data:
  - The number under each search term is higher if they were googling it more that week
  - The number under "cdcflu" is the activity of the flu that week
- We will use regression modelling to plot cdcflu by week and find the right model to fit the data points
  - When we do, we see that the data seems the be seasonal
- We will use regression modelling with cdcflu as our output and search terms as inputs to predict cdcflu data
  - First, let's find how well each search term predicts cdcflu
    - "lmfull = lm(cdcflu ~ . - week, data=googleflu)"
  - We don't care about the individual coefficients- we just care about the overall predictive value of the model with all of our predictors that we chose to include in the model being used at once
  - But it doesn't make sense to have all of your predictors in the model
    - So we can use **Occam's Razor** → you want your model to fit well, but not "too well" (when we say "too well", we mean overfitting)
      - We want a model that will generate good predictions and fit the data set well- but we don't want the model to be so complex that it won't generalize well for future cases
    - **Occam's Razor idea**→ function of model output that **balances fit and simplicity**
      - Examples are in the notes such as:
        - R squared adjusted
        - BIC
        - AIC
          - $AIC = 2p + \frac{UV}{\hat{\theta}_e^2}$

            $UV = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

            $\hat{\theta}_e^2$ = Residual value
          - = complexity penalty + Poor fit penalty
          - Note: p=# parameters to be estimated
          - Never going to be asked for this formula since we will just use the computer to calculate it- but it is good to know what is
      - Think of it like buying data with a budget

- You want to include data that predicts the y-value well
- We want to spend our data budget wisely to include a few predictors that fit the data well
- The AIC (or BIC, adj R squared) tells you when these predictors are worth splurging your budget on

- How did that work with prediction intervals?
  - $y^* = \hat{y}^* + e^*$

Future systemic model
Ex. $\hat{\beta}_o + \hat{\beta}_1(x^*)$

Of future random part (residuals)

- When we put more data into the model, we get less uncertainty in residuals and more uncertainty in systemic portion since it increases the # of parameters
- If data decreases uncertainty in residuals, more than it increases uncertainty in systemic portion, it's worth it to buy the data

- In the google flu trends data
  - There is no way to try all of the models and find each of their AICs to decide which ones to use (you want a lower AIC) → there's just too many models
  - So we use **(backwards) stepwise selection**
    - Ex. y vs. x_1, x_2, x_3, x_4
    - You start big and consider all possible one variable deletions (or additions, if you are using bidirectional stepwise selection) until the AIC cannot be lowered anymore
    - Y vs. x_1, x_2, x_3, x_4

      $x_1, x_2, x_3,$
      $x_1, x_2, \_, x_4$
      $x_1, \_, x_3, x_4$
      $\_, x_2, x_3, x_4$

      (let's say the yellow model is the one with the lowest AIC)

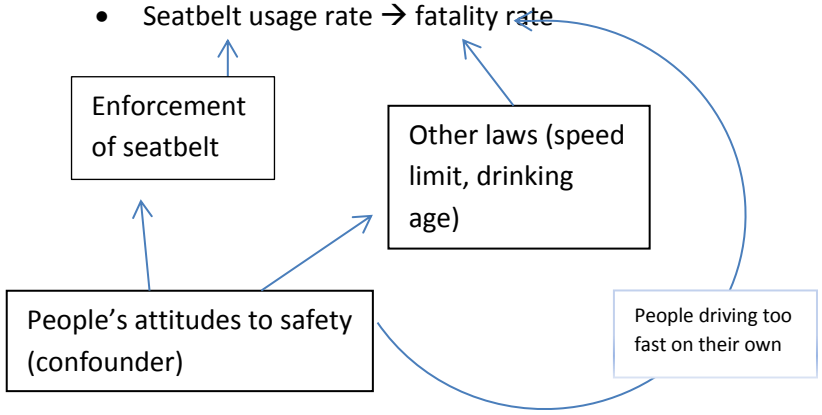
      So then,
      $x_1, x_2,$
      $x_1, \_, x_4$
      $\_, x_2, x_4$

  - Do "lmstep = step(lmfull, direction='both')" to do bidirectional stepwise selection in R → this code does the stepwise process for you
  - Stepwise processes are a good way to start your model. But you move on from here and will do further work on your model after this step

- To check our mode, we can use **cross validation**
  - Starts on line 34 of googleflu.R
  - Label half data as "past" and half as "future"
  - Then fit model on past
  - Then cross validate 1000 times (using forloop or do*(1000)) and average your squared errors over these many samples

Now we can talk about the other side of the spectrum (**very specific question**)

- First, import seatbelts data→ remember to make the header row
- Question: what is the effect of seatbelt usage with the number of traffic fatalities?
- We could just look at the data, but there's confounders
  - Fatalities are the # fatalities on the
  - Seatbelt is the rate of seatbelt usage in the states included in the poll
  - Spead65 and speed 70 indicated whether or not the state had a speed limit of 65 or 70
  - Drinkage is whether or not they had a minimum drinking age of 21
  - Alcohol is whether or not there are open container laws in that state
  - Income is the per capita income in adjusted inflated US dollars
  - Enforce is whether or not they enforce seatbelt usage
    - Secondary is whether police can add on seatbelt charge to another initial offense
    - Primary is when the police can pull you over for not wearing a seatbelt
  - **Some steps to make a great model:**
  - Pictures of the system
    - Seatbelt usage rate → fatality rate



Techniques to get tid of confounders: intercept the back-door paths

  - Pictures of the data