Homework 2

For this and all homeworks, you must explain your approach and how you got to the answer.

1. Validate the estimate for effective degrees of freedom in Wilks Equation 5.12 via simulation.
   (a) Produce sets of random AR(1) time series of length 100 with autocorrelation coefficients -0.9, -0.8, ..., 0, 0.1, 0.2, ..., 0.9.

   (b) Calculate the empirical variance of the means of each set of time series with a given autocorrelation coefficient.

   (c) Use the empirical variance to estimate the effective sample size for sampling variance of the mean.

   (d) Plot your estimate vs. the estimate in Wilks Equation 5.12. How well do they agree?

2. Experiment with using the block bootstrap to assess uncertainty in the trends in correlated data.
   (a) Start with a single correlation coefficient for an AR(1) process, and produce a "dataset" of length 100.

   (b) Calculate the empirical trend in your data using ordinary least squares. Also calculate the expected variance in your estimator (i.e. $Var(\hat{\beta})$) using standard methods that assume independence of the data points.

   (c) Choose an appropriate block size based on equation 5.36 in Wilks. Justify your block size in terms of the autocorrelation properties of the data. Perform a block bootstrap on your data many times (e.g. 10000), recalculate the OLS trend for each resampled time series, and use the distribution of trends to estimate the variance in the OLS estimator for your dataset. How does the block bootstrap estimate compare with the standard estimate from (b)?

   (d) Since we know the generating process for our data, we can again estimate the uncertainty in our slope by generating many more random datasets from the same process. (Note that for an AR(1) process, we could also estimate $Var(\hat{\beta})$ analytically.) How does this estimate of variance compare to our bootstrap estimate?

3. You've finished your degree, and decided to become an energy trader. Your boss has informed you that a good metric for energy demand in the residential sector is so-called heating degree days (HDDs) and cooling degrees days (CDDs) defined as

HDD $= max(T_h - \frac{T_{max}+T_{min}}{2}, 0)$,

where T_h is a baseline temperature above which heating is generally not required, and

CDD $= max(\frac{T_{max}+T_{min}}{2} - T_c, 0)$, where T_c is a baseline temperature below which cooling is generally not required.

We'll set T_h =15°C, and T_c = 20°C.

The concepts of HDDs and CDDs extend to months or seasons. For example, to calculate the CDDs for JJA of some year, simply calculate the CDD for each day in the time interval, and then sum or average, whichever may be more useful. The concepts of HDD and CDD are useful in the energy industry for quantifying power demands, and the related concepts of 'growing' and 'killing' degree days are used in agriculture.

(a) Calculate and plot summer-season (JJA) CDD and winter-season (DJF) HDD for Los Angeles. The year associated with DJF (which will inherently span two calendar years) should be the one containing January and February.

(b) Decide and state what to do about missing data in calculating seasonal CDD and HDD.

(c) Present a few useful exploratory plots of CDD and HDD, and describe the distributions.

(d) Use whatever testing procedure you think best to test H_0: The mean DJF HDD are equal before and after 1980, versus H_a: they are not, and likewise for JJA CDD.

(e) Fit a regression model to the HDD, using an intercept and centered time as predictors, and likewise for CDD. That is, fit a univariate regression model with an intercept, using as a predictor the year, after removing the mean of the vector of years. Centering the predictor aids the interpretation of results, as the y-intercept then corresponds to the expected value in the middle of the time interval. Are there significant trends? Discuss, and compare results with the previous part.

4. Now that we've calculated our HDDs and CDDs, we want to see if we can predict seasonal energy demand in Los Angeles. In particular, your mission is to build separate regression models to predict winter season (DJF) HDD and summer season (JJA) CDD, from climate indices over previous months or seasons. Note that if your model is really good, you should be able to monetize results by trading energy futures. To be clear: use monthly or seasonal averages of the climate indices as you see fit, but make sure they **precede** in time the season being predicted. So to predict JJA

CDDs, the latest predictors would be climate indices for May. Fit models on data prior to 2005, and present predictions for the years 2006-2018.

(a) The website KNMI Climate Explorer (link) hosts time series of monthly climate indices. Select your favorite 5-10 indices (should be close to spanning our dataset time period of 1945-present), and download them. If these indices look like alphabet soup to you, it's worth chatting with your colleagues about what they mean. You should include a few different ENSO indices, the PDO, and whatever else you like. Present a small number of summary plots characterizing your predictors. You need not use the same climate indices as predictors for CDD and HDD, and use multi-month average or monthly values as you see fit. Choose some number of predictors from these indices, remembering the constraint that the predictors have to precede (in time) the predictand. For example, 'Nino 3.4 averaged over March-April-May' could be a potential predictor for summer CDD.

(b) Examine the covariance matrix of your predictors. Is there collinearity? If so, change your predictors to reduce collinearity, and explain.

(c) Read Section 7.4 in Wilks. Build a multivariate linear regression model for summer CDDs and winter HDDs with your chosen predictors (and potentially a time trend term) using stepwise procedures. Describe your final regression model, and discuss uncertainty in your parameters.

(d) Now add one or more additional predictors that are a function of the Los Angeles temperature data for preceding month(s) or season(s). How does the final model change? Are the Los Angeles predictors included in your final model? Are some of the climate indices that were included before now not included?

(e) Predict CDD and HDD for the withheld years, using whatever model you think is most justifiable in each case. To show your results, plot the observed time series and your predicted time series (with uncertainty) on the same plot. You can use the standard uncertainty estimates

(f) As an alternative way to quantify uncertainty in the regression model, perform a residual bootstrap (see below). Plot bootstrap distributions of a few key regression parameters, and compare the variability with that from the 95% intervals that your coding language of choice will provide in standard packages.

(g) End with a short write up of results, explaining what you found and if your model should be used by the energy trading desk.

Residual bootstrap:

1. Calculate the expected values of the response, the $\hat{y}_i$, and the residuals, $\hat{\epsilon}_i = y_i - \hat{y}_i$.

2. Form a bootstrap sample of the response, $(y_1^*, \ldots, y_n^*)$ as $y_i^* = \hat{y}_i + \hat{\epsilon}_i^*$, where the $\hat{\epsilon}_i^*$ are selected, with replacement, from the set of residuals.

3. Re-fit the regression model using the $(y_1^*, \ldots, y_n^*)$, and record the regression parameters.

4. Repeat 2-3 to build up bootstrapped distributions of the regression parameters.