Stat 27850/30850: Problem set 3

1. Selective inference for linear regression.
Suppose that we use the following procedure for model selection, in the setting where $n \geq p$ and the design matrix X is non degenerate (i.e. has rank p). First we fit the least squares coefficients, $\widehat{\beta} \in \mathbb{R}^p$. We next select the model S that consists of the K largest-magnitude coefficients, where K is a fixed number chosen in advance. Suppose that the outcome we observe for this procedure is:

    - Feature X_{j_1} is chosen first (i.e. $|\widehat{\beta}_{j_1}|$ is the largest entry of $\widehat{\beta}$), with sign $s_1 \in \{+1, -1\}$ (i.e. this is the sign of $\widehat{\beta}_{j_1}$)
    - Feature X_{j_2} is chosen second, with sign s_2
    - ...
    - Feature X_{j_K} is chosen last, with sign s_K

    Let $A \subset \mathbb{R}^n$ be the set of all vectors y that would yield this exact same outcome. Write down the set of linear inequalities in y that define the set A, i.e. A is the set of all vectors y that satisfy the entire list of inequalities. For simplicity you can assume there are no ties.

2. Post-selection confidence intervals. Suppose that you observe a single data point

$$X \sim N(\mu, \sigma^2)$$

CI for μ

    where μ is an unknown mean parameter while σ^2 is a known variance. If we believe that μ is a large positive mean, then we consider it to be interesting and will study it further. For example μ might be the increase in survival time when taking a new drug; the data X would be the estimated change in survival time based on a large randomized trial. if it appears that μ is large and positive then we will invest in further clinical trials of the drug.

    To make this decision, we set a threshold $\tau > 0$. If the observed data passes the threshold τ, that is, $X > \tau$, then we will decide to study the effect further.

    In this question, we will work on the problem of building a confidence interval for μ when the effect has been selected for further study. In particular, the ordinary confidence interval $X \pm z_* \sigma$ will not suffice, because it does not take into account the fact that the data has already passed the threshold τ. However, we'll deal with one-sided rather than two-sided inference here to make calculations a bit easier in the post-selection setting.

    As usual we'll use Φ and Φ^{-1} to denote the standard normal CDF and its inverse.

    (a) First let's ignore the selection process and just build a one-sided confidence interval. After observing data $X = x$ we will calculate a value $\mu_0(x) = x - (\text{some margin of error})$ and will claim, with $1 - \alpha$ confidence, that $\mu \geq \mu_0(x)$. Write an expression for $\mu_0(x)$ in terms of x so that this statement is true, that is,

$$\mathbb{P}\{\mu \geq \mu_0(X)\} = 1 - \alpha$$

    where this probability is taken with respect to $X \sim N(\mu, \sigma^2)$. Note that the event $\{\mu \geq \mu_0(X)\}$ is in fact random, even though μ is fixed, because $\mu_0(X)$ is a function of the random variable X.

    (b) Next let's turn to the post-selection version of this problem. Suppose that the true parameter is equal to μ. Calculate a value $x(\mu)$ such that

$X > \tau$

$$\mathbb{P}\{X \leq x(\mu) | X \text{ passes the threshold for further study}\} = 1 - \alpha.$$

    Your equation for $x(\mu)$ will use the function Φ and/or Φ^{-1}.

    (c) Now we'll invert the process. You can assume that $\mu \mapsto x(\mu)$ is a strictly increasing function of μ. Let $x \mapsto \mu(x)$ be the inverse of this function, that is, $\mu(x)$ is the value that satisfies, for any specific value x_1, $x(\mu(x_1)) = x_1$. Then we have

$\mu(x(\mu_1)) = \mu_1$

$$\mu \geq \mu(x) \;\Leftrightarrow\; x \leq x(\mu).$$

$x(\mu) \geq x$

(Note that we do not have a closed form expression for $\mu(x)$, however.)

Now let the true parameter μ be fixed. Explain why it's true that

$$\mathbb{P}\{\mu \geq \mu(X) | X \text{ passes the threshold for further study}\} = 1 - \alpha,$$

where the probability is taken with respect to the draw $X \sim N(\mu, \sigma^2)$. Note that the event $\{\mu \geq \mu(X)\}$ is in fact random, even though μ is fixed, because $\mu(X)$ is a function of the random variable X.

3. In this next problem, test your work above empirically. Fix $\tau = 2, \sigma^2 = 1, \alpha = 0.1$.

   (a) First let's plot the confidence interval as a function of x. In the same figure, plot $\mu_0(x)$ and $\mu(x)$ over a range of x values (but only $x \geq \tau$ since otherwise we would not be interested in that sample). Discuss what you find in your plot.

   There is one caveat: you'll notice that for values of x that are close to the threshold τ (above τ but not by much), R will be unable to find $\mu(x)$. That's because $\mu(x)$ is very far out in the tails of the normal and R will round probabilities to zero. To get around this, here's what I suggest:

   - First set $x_{\text{lower}} = x(-5)$ (here I'm plugging in $\mu = -5$ as a low value; if we go much lower, R will start rounding probabilities to zero in the tails).
   - Then for any x, if $x \leq x_{\text{lower}}$ just set $\mu(x) = -\infty$ (since we know in any case that the right answer would satisfy $\mu(x) \leq -5$ which is very low). If $x > x_{\text{lower}}$ then solve for $\mu(x)$ as above.

   .

   (b) Next we will let μ vary in $\{0, 0.25, 0.5, \ldots, 5\}$ and test the coverage rates. For each value of μ that we're testing, run the following simulation. Generate $X \sim N(\mu, \sigma^2)$; if $X \geq \tau$ then keep this sample, otherwise discard it. Run this until you have 10000 samples, $X_1, \ldots, X_{10000}$. Now for each $i = 1, \ldots, 10000$, construct your (one-sided) confidence intervals: first without accounting for selection, i.e. your claim is that $\mu \geq \mu_0(X_i)$, and then with the correct conditioning to take selection into account, i.e. your claim is that $\mu \geq \mu(X_i)$. Note that to calculate the value $\mu(x)$ you will need to use a numerical solver; use `uniroot` in R. The functions Φ and Φ^{-1} are called `pnorm` and `qnorm` in R.

   Plot the coverage as a function of μ, i.e. for each value of μ that you try, what proportion of the time (out of the 10000 trials) is the statement actually true, both for the "naive" version and for the post-selection version. Then summarize your findings.