STAT30850 Homework 3

Sarah Adilijiang

Problem 1 - Selective Inference

The least squares estimator of coefficients β is: $\hat{\beta} = (X^T X)^{-1} X y$

So the least squares estimator of the jth coefficient is: $\hat{\beta}_j = e_j^T (X^T X)^{-1} X y$, where $e_j = (0, \cdots, 0, 1, 0, \cdots, 0)^T$ only has 1 at the jth element.

Let's define $v_j = e_j^T (X^T X)^{-1} X$, thus we have: $\hat{\beta}_j = v_j y$

Now we have selected the K largest-magnitude coefficients,

$$|\hat{\beta}_{j_1}| \geq |\hat{\beta}_{j_2}| \geq \cdots \geq |\hat{\beta}_{j_K}| \ \geq |\hat{\beta}_h| \qquad \forall\, h \in \{1, \cdots, p\} \text{ and } h \notin \{j_1, \cdots, j_K\}$$

i.e.:

$$s_1 v_{j_1} y \geq s_2 v_{j_2} y \geq \cdots \geq s_K v_{j_K} y \ \geq s_h v_h y \qquad \forall\, h \in \{1, \cdots, p\} \text{ and } h \notin \{j_1, \cdots, j_K\}$$

Therefore, the set of all vectors y that satisfy the entire list of inequalities is:

$$A = \{y : s_1 v_{j_1} y \geq s_2 v_{j_2} y \geq \cdots \geq s_K v_{j_K} y \ \geq s_h v_h y \quad \forall\, h \in \{1, \cdots, p\} \text{ and } h \notin \{j_1, \cdots, j_K\}\}$$

Problem 2 - Post-selective Confidence Intervals

(a) Ignore the selection process

$$X \sim N(\mu, \sigma^2) \quad \Rightarrow \quad \frac{X - \mu}{\sigma} \sim N(0, 1)$$

$$\begin{aligned}
\Rightarrow \quad & \mathbb{P}\left(\frac{X - \mu}{\sigma} \leq \Phi^{-1}(1 - \alpha) \right) = 1 - \alpha \\
\Rightarrow \quad & \mathbb{P}\left(\mu \geq X - \sigma\ \Phi^{-1}(1 - \alpha) \right) = 1 - \alpha \\
\Rightarrow \quad & \mu_0(X) = X - \sigma\ \Phi^{-1}(1 - \alpha)
\end{aligned}$$

(b) Post-selective version

$$1 - \alpha = \mathbb{P}\{X \leq x(\mu) | X > \tau\} = \frac{\mathbb{P}\{\tau < X \leq x(\mu)\}}{\mathbb{P}\{X > \tau\}} = \frac{\mathbb{P}\{\frac{\tau - \mu}{\sigma} < \frac{X - \mu}{\sigma} \leq \frac{x(\mu) - \mu}{\sigma}\}}{\mathbb{P}\{\frac{X - \mu}{\sigma} > \frac{\tau - \mu}{\sigma}\}} = \frac{\Phi(\frac{x(\mu) - \mu}{\sigma}) - \Phi(\frac{\tau - \mu}{\sigma})}{1 - \Phi(\frac{\tau - \mu}{\sigma})}$$

$$\Rightarrow \qquad x(\mu) = \mu + \sigma\ \Phi^{-1} \left(1 - \alpha + \alpha\ \Phi(\frac{\tau - \mu}{\sigma}) \right)$$

(c) Invert the process

Since $\mu \mapsto x(\mu)$ is a strictly inceasing function of μ, and $x \mapsto \mu(x)$ is the inverse of this function, therefore, $\mu(x)$ is also a strictly inceasing function of x. Thus we have that:

$$\mu \geq \mu(X) \quad \Longleftrightarrow \quad x(\mu) \geq x(\mu(X)) = X, \text{ i.e. } X \leq x(\mu)$$

$$\Rightarrow \quad \mathbb{P}\{\mu \geq \mu(X), X > \tau\} = \mathbb{P}\{X \leq x(\mu), X > \tau\}$$

$$\Rightarrow \quad \frac{\mathbb{P}\{\mu \geq \mu(X), X > \tau\}}{\mathbb{P}\{X > \tau\}} = \frac{\mathbb{P}\{X \leq x(\mu), X > \tau\}}{\mathbb{P}\{X > \tau\}}$$

$$\Rightarrow \quad \mathbb{P}\{\mu \geq \mu(X) | X > \tau\} = \mathbb{P}\{X \leq x(\mu) | X > \tau\} = 1 - \alpha$$

Problem 3

(a)

From problem 2, we get that:

$$\mu_0(X) = X - \sigma\ \Phi^{-1}(1-\alpha)$$

And since:

$$\mathbb{P}\{\mu \geq \mu(X) | X > \tau\} = \mathbb{P}\{X \leq x(\mu) | X > \tau\} = 1 - \alpha$$

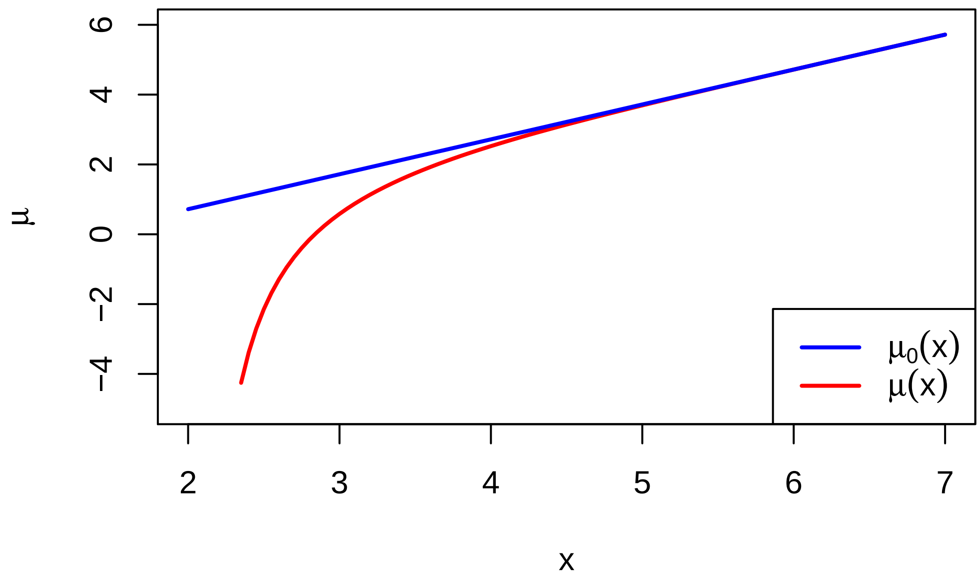
Thus by setting:

$$X = x(\mu), \quad \text{i.e.} \quad x(\mu) - X = \mu + \sigma\ \Phi^{-1}\left(1 - \alpha + \alpha\ \Phi(\frac{\tau - \mu}{\sigma})\right) - X = 0$$

We can solve for the lower boundary, $\mu(X)$, of the condidence interval $\mu \geq \mu(X)$.

```
# functions
mu0 = function(x,sigma,alpha){ x - sigma*qnorm(1-alpha) }

x_mu = function(mu,sigma,alpha,tau){
    mu + sigma*qnorm(1-alpha+alpha*pnorm((tau-mu)/sigma)) }

f_mu = function(mu,sigma,alpha,tau,x){
    mu + sigma*qnorm(1-alpha+alpha*pnorm((tau-mu)/sigma)) - x }

muu = function(xx,sigma,alpha,tau){
    x_lower = x_mu(-5,sigma,alpha,tau)
    mu = rep(-Inf,length(xx))
    for (i in 1:length(xx)){
        if(xx[i] > x_lower){mu[i] = uniroot(f=f_mu, interval=c(-5,20),
                                            x=xx[i], sigma=sigma, alpha=alpha, tau=tau)}
    }
    return(unlist(mu))
}

# calculations
sigma = 1
alpha = 0.1
tau = 2
xx = seq(tau,7,0.05)  # x >= tau = 2, the samples that we are interested in
mu0_xx = mu0(xx,sigma,alpha)
muu_xx = muu(xx,sigma,alpha,tau)
```

```
# plot
plot(xx,muu_xx,ylim=c(-5,6),xlab="x",ylab=expression(mu),type="l",col=2,lwd=2)
lines(xx,mu0_xx,type="l",col=4,lwd=2)
legend("bottomright",legend=c(expression(mu[0](x)),expression(mu(x))),
       lty=1, lwd=2, col=c(4,2))
```



Comment:

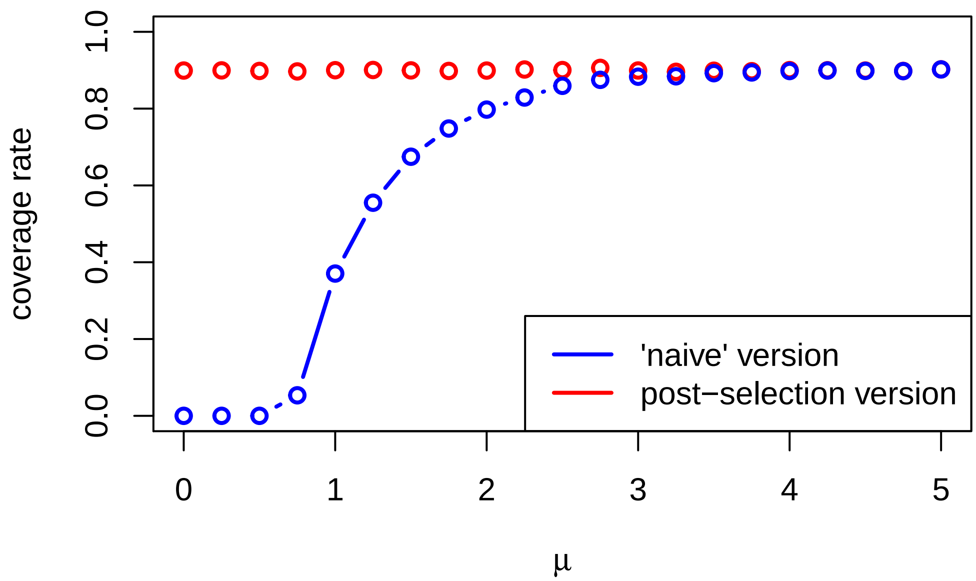
We can see that as the value of x increases from 2 (the value of τ) to around 5, $\mu(x)$ increases quickly and gets closer and closer to $\mu_0(x)$, but is always lower than $\mu_0(x)$. When x becomes larger than 5, $\mu(x)$ is almost the same as $\mu_0(x)$.

(b)

```
mu = seq(0,5,0.25)
cov_rate_mu0 = cov_rate_muu = rep(NA,length(mu))
for (i in 1:length(mu)) {
    # generate 10000 samples with x >= tau = 2
    xx = NULL
    while(length(xx)<10000) {
        x = rnorm(1,mu[i],sigma)
        if(x>=tau) {xx=c(xx,x)}
    }

    # calculate the coverage rate
    mu0_xx = mu0(xx,sigma,alpha)
    muu_xx = muu(xx,sigma,alpha,tau)
    cov_rate_mu0[i] = mean(mu[i] >= mu0_xx)
    cov_rate_muu[i] = mean(mu[i] >= muu_xx)
}
```

```
# plot
plot(mu,cov_rate_muu,ylim=c(0,1),type="b",col=2,lwd=2,
     xlab=expression(mu),ylab="coverage rate")
lines(mu,cov_rate_mu0,type="b",col=4,lwd=2)
legend("bottomright",legend=c("'naive' version","post-selection version"),
```

```
        lty=1, lwd=2, col=c(4,2))
```



Comment:

For the "naive" version, as the value of true μ increases from 0 to around 4, the coverage rate first stays at zero when $\mu \in [0, 0.5]$, and starts to increase fast when $\mu \in [0.5, 2]$, then increases slowly and gets closer to the coverage rate of the post-selective version when $\mu \in [2, 4]$. Finally, when $\mu \geq 4$, the coverage rates of the two versions become almost the same. The coverage rate of the naive version is lower than that of the post-selection version when the true $\mu < 4$. This is because after selecting the samples where $X \geq \tau = 2$, when the true μ is small, especially when $\mu < 2$, the lower boundary of the ordinary confidence interval of μ tends to be larger than the true μ, i.e. $\mu_0(x) > \mu$.

For the post-selection version, as the value of true μ increases, however, the coverage rate keeps almost at the same level, which is around 90%. This is because we are calculating the confidence interval with correct conditioning to take selection into account.