# Final Exam Question 4

## STAT 560 Statistical Theory I

Clayton Allard

December 22, 2022

(a) We iteratively add one variable at a time to see which variable makes the biggest improvement based on the t-test. After we add the variable, we run the whole model to see if every variable still remains significant at the $\alpha$ level from the t-test. If one of the variable's P-values fails to be within the threshold, then we take the previous step's model as our choice. If we get all the way through with all variables being significant, then the full model is the best model.

**Algorithm 1:** Forward Stepwise Algorithm

```
Data: For n > p we have x ∈ ℝ^{n×p}, y ∈ ℝ^n, α ∈ [0,1]
I ← 0 ∈ ℝ^p ;                        /* I_i = 1 if x_i is in the model, 0 otherwise.  */
M ← {1} ;                                                   /* Intercept model.  */
while I ≠ 1 ∈ ℝ^p do
    P ← 1 ∈ ℝ^p ;                                      /* Initialize p-values.  */
    for i ∈ {1, 2, ..., p} do
        if I_i = 1 then
            /* x_i is already in the model.                                      */
            continue
        end
        N ← M append x_i ;                               /* Add variable to model.  */
        P_i ← P-value(N(x_i)) ;                        /* P-value of t-test for x_i.  */
    end
    m ← arg min_i P ;                                            /* Best variable.  */
    p* ← P_m ;                                        /* p* is the lowest P-value.  */
    if p* < α then
        M ← M append x_m;                           /* Add variable to real model.  */
        I_m ← 1;                                        /* Include x_m in the model.  */
        V ← P-value(M);            /* P-values where V_i = 1 if x_i not in model.  */
        /* Element wise product ⇒ Max of variables in the model.                 */
        if max{I ⊙ V} ≥ α then
            /* We want to ensure that all variables maintain a low P-value.   */
            M ← M remove x_m;                     /* Then we go with previous model.  */
            return M
        end
    end
    else
        return M
    end
end
```

(b) To do the backward stepwise algorithm, start with the full model. If every variable's t-test P-value is less than $\alpha$, then that is our model of choice. Otherwise, remove the variable with the highest P-value. Then we run the model of the remaining predictors. If all the variables come out with a P-value less than $\alpha$ from the t-test, then that is the model we choose. Otherwise, continue this process of removing the variable with the highest P-value. This process may continue until we are all the way down to the null model.