Question 1

(5pt) 9-10주차 강의 프레젠테이션 20page에 있는 $SSR + SSE = SST$ 를 증명하시오.

$$SSR + SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

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

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

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

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

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

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

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

$(y_i - \hat{y}_i)$ → 0

☺ $\sum_{i=1}^{n} (y_i - \beta_0 + \beta_1 x_i \cdots) = 0$

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

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

$SSR + SSE = SST$

Question 3

다음은 "gala" 데이터에서의 "Species"를 종속변수, 다른 4개의 변수를 독립변수로 사용한 선형회귀분석을 r에서 실행한 결과를 보여준다.

((a), 5pt, 설명) 예측된 선형회귀분석 모형의 식을 적고, "p-value:2.45e-07" 와 "F-statistic: 19.23 on 4 and 25 DF" 이 설명하고 있는 가설검정과 그 결과를 설명하시오.

선형모형 식: $Species = 13.7567 + 0.2136 \times Elevation - 0.2492 \times Scruz - 0.0666 \times Adjacent + 0.2382 \times Nearest$.

F-statistic은 $p=4$, $n-p-1=25$ 인 자유도에 대해 계산된 값이며, $H_0: \beta_{Elevation} = \beta_{Scruz} = \beta_{Adjacent} = \beta_{Nearest} = 0$ versus $H_1: \beta_j \neq 0$ for some j 에 대한 가설검정이다. 그리고 해당 가설에 대한 p-value가 $2.45e^{-07}$로 0.05보다 낮게 나왔으니 H_0을 reject하여, 변수의 계수 중에 '0'이 아닌 것이 있다.

((b), 5pt, 설명) "Scruz" 와 "Nearest" 두개의 독립변수들을 모두 위 선형모형에서 제외시켜 새로운 선형회귀분석을 하고, 가설검증을 통해 새로운 선형모형과 기존의 선형모형을 비교하시오.

$H_0: \beta_{Scruz} = \beta_{Nearest} = 0$ vs $H_1: \beta_j \neq 0$ for some j in (scruz, Nearest)

해당 가설에서 ANOVA test를 하였을 때, p value가 0.4297로 0.05보다 훨씬 크게 나왔다. 따라서 H_0을 기각할 수 없다. 이는 β_{Scruz}와 $\beta_{Nearest}$가 '0' 임을 의미하므로 기존의 선형모형에서 해당 변수 외에 다른 변수(Elevation, Adjacent)에서 의미있는 변수가 있었다는 것을 알 수 있다.

((c), 5pt, 설명) 위 첨부된 그림의 선형회귀분석 결과에서 우리는 $\hat{\beta}_{Elevation} \approx -4.5\hat{\beta}_{Adjacent}$ 임을 발견할 수 있다. 실제 이 관계식의 통계적 유의성을 유의수준 $\alpha = 0.1$을 사용하여 가설 검정하시오.

$H_0: \beta_{Elevation} = -4.5\beta_{Adjacent}$ vs $H_1: \beta_{Elevation} \neq -4.5\beta_{Adjacent}$

ANOVA의 p-value가 0.6689로 매우 높게 나왔기 때문에 귀무가설을 기각할 수 없다.

Question 4

이 문제는 "faraway" 패키지에 있는 "seatpos" 데이터를 이용한다.

((b), 5pt, 설명) 위의 "seatpos" 데이터에 적용된 Lasso와 Ridge 선형 모형 중 선호하는 모형을 선택하고 그 합리적인 이유를 설명하시오. (이 문제는 Lasso, Ridge 방법 모두 답이 가능. 이 중 하나를 선택하고 그 합리적인 이유를 설명하면 충분)

그래프를 살펴보면 Ridge의 경우 coefficient가 '0'으로 수렴해가는 모습을 보인다. 반면 Lasso의 경우, 진작 의미 없는 변수의 coefficient가 '0'이 되었다. 따라서 Lasso 모형이 Ridge 모형보다 더 가벼운 모형으로 좋은 성능을 낼 수 있을 것이라 생각하여 Lasso를 선택 한다.

Question 5

Lasso는 고차원데이터의 회귀분석에서 가장 많이 쓰이고 있는 방법 중 하나로 중요하지 않는 회귀계수 (coefficient) 를 0으로 만들어 주는 장점이 있다. 하지만, Lasso term (i.e., $\sum_j |\beta_j|$) 에 의해 발생하는 Bias가 존재한다. 이를 보안하기 위해 나온 방법 중 하나가 Adaptive Lasso 로 그 정의는 다음과 같다.

$$\text{minimize}_\beta \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{p} \frac{|\beta_j|}{\max(|\tilde{\beta}_j|, \epsilon)}.$$

여기서 $\tilde{\beta} = (\tilde{\beta}_1, \cdots, \tilde{\beta}_p)'$ 는 Lasso (with $\tilde{\lambda}$) 를 통해 얻어진 회귀계수이다. $\epsilon > 0$은 분모가 0 이 되는 것을 방지하기 위해 필요한 충분히 작은 상수이다.
위의 Adaptive Lasso를 "glmnet" 함수를 사용하여 구현시키는 것이 이 문제의 목표이다.

((a), 5pt, 설명) Adaptive Lasso 를 일반적인 Lasso 문제 형태의 꼴로 바꿔서 표현하시오. 구체적으로는

$$\text{minimize}_\beta \frac{1}{2}\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p} \tilde{x}_{ij}\beta_j)^2 + \sum_{j=1}^{p} |\beta_j|$$

형태로 $\tilde{x}$는 새로운 design 행렬로 $X, \lambda, \tilde{\beta}, \epsilon$ 에 의존한다.

Hint: Use "Change of variables": $\beta_j \leftarrow \lambda\beta_j / \max(|\tilde{\beta}_j|, \epsilon)$ for each j

$$\tilde{\beta}_j = \frac{\beta_j}{\max(|\tilde{\beta}_j|, \epsilon)}$$

$$= \min_{\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\frac{\lambda\beta_j}{\max(|\tilde{\beta}_j|,\epsilon)}\right)^2 + \lambda\sum_{j=1}^{p}\frac{|\beta_j|}{\max(|\tilde{\beta}_j|,\epsilon)}$$

$$= \min_{\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\frac{x_{ij}}{\max(|\tilde{\beta}_j|,\epsilon)}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j$$

$$= \min_{\beta} \frac{1}{2}\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\tilde{x}_{ij}\beta_j\right)^2 + \lambda\sum_{j=1}^{p}\beta_j$$