Problem 1

This problem is about OLS estimation in regression. You can assume that

$$
\begin{aligned}
\boldsymbol{X} &:= [\mathbf{1}_n \mid \boldsymbol{x}_{\cdot 1} \mid \ldots \mid \boldsymbol{x}_{\cdot p}] \quad \text{with column indices } 0, 1, \ldots, p \text{ and row indices } 1, 2, \ldots, n \\
\boldsymbol{H} &:= \boldsymbol{X} \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \\
\boldsymbol{Y} &= \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\mathcal{E}} \\
\boldsymbol{B} &:= \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \boldsymbol{Y} \\
\hat{\boldsymbol{Y}} &= \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} \\
\boldsymbol{E} &:= \boldsymbol{Y} - \hat{\boldsymbol{Y}} = (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}
\end{aligned}
$$

where the entries of $\boldsymbol{X}$ are assumed fixed and known and the entries of $\boldsymbol{\beta}$ are the unknown parameter).

(a) [easy] When we "do inference" for the linear model, what is the parameter vector?

β

(b) [easy] When we "do inference" for the linear model, what are considered the fixed and known quantities?

X, H

(c) [easy] When we "do inference" for the linear model, what are considered the random quantities? And what is the notation for their corresponding realizations?

Y, y and β, B

(d) [easy] What is the "core assumption" in which the classic linear model inference follows?

mean centered, homoscedasicity, and normal dist

(e) [easy] From the core assumption, derive the distribution of $\boldsymbol{B}$.

$\vec{B} = (X^TX)^{-1}X^T\vec{Y} = (X^TX)^{-1}X^T(X\vec{\beta}+\vec{\varepsilon}) = (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon$

$\sim N_{p+1}(\vec{\beta}, \sigma^2(X^TX)^{-1}X^T((X^TX)^{-1}X^T)^T)$

$\sim N_{p+1}(\vec{\beta}, \sigma^2(X^TX)^{-1})$

(f) [easy] From this result, derive the distribution of B_j.

$B_j \sim N(\beta_j, \sigma^2(X^TX)^{-1}_{j,j})$

(g) [easy] From this result, derive the distribution of B_j standardized.

$$\frac{B_j - \beta_j}{\sqrt{\sigma^2(X^TX)^{-1}_{j,j}}} \sim N(0,1)$$

(h) [easy] from the core assumption, derive the distribution of $\hat{\boldsymbol{Y}}$.

$\vec{\hat{Y}} = X\vec{\beta} = H\vec{Y} = H(X\vec{\beta} + \vec{\varepsilon}) = X\vec{\beta} + H\vec{\varepsilon} \sim N_n(X\vec{\beta}, H\sigma^2 I_n H)$

$\sim N_n(X\vec{\beta}, \sigma^2 H)$

(i) [easy] From this result, derive the distribution of $\hat{Y}_i$.

$\hat{Y}_i \sim N(\vec{x}_i \vec{\beta}, \sigma^2 H_{i,i})$

(j) [easy] From this result, derive the distribution of $\hat{Y}_i$ standardized.

$$\frac{\hat{y}_i - \vec{x}_i \vec{\beta}}{\sqrt{\sigma^2 H_{i,i}}}$$

(k) [easy] from the core assumption, derive the distribution of $\boldsymbol{E}$.

$\vec{E} = \vec{Y} - \hat{Y} X\vec{\beta} = (I - H)\vec{Y} = (I - H)(X\beta + \vec{\varepsilon}) = \cancel{(I - H)X\beta} + (I - H)\vec{\varepsilon}$

$= (I - H)\vec{\varepsilon} \sim N_n(\vec{0}_n, (I - H)\sigma^2 I_n (I - H)^T)$

$= N_n(\vec{0}_n, \sigma^2 (I - H))$

(l) [easy] From this result, derive the distribution of E_i.

$\hat{\varepsilon}_i \sim N(0, \sigma^2(1 - H_{i,i}))$

(m) [easy] From this result, derive the distribution of E_i standardized.

$$\frac{e_i}{\sqrt{\sigma^2 (I - H_{i,i})}}$$

(n) [easy] From the core assumption, show that $\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} \sim \chi^2_n$.

$z \cdot z \sim \chi^2_n$

$\Rightarrow \left(\frac{1}{\sigma}\varepsilon\right)\left(\frac{1}{\sigma}\varepsilon\right)$

$= \frac{1}{\sigma^2}\varepsilon^T\varepsilon \sim \chi^2_n$

(o) [easy] Let $\boldsymbol{B}_1 = \boldsymbol{H}$ and let $\boldsymbol{B}_2 = \boldsymbol{I}_n - \boldsymbol{H}$. Justify the use of Cochran's theorem and then find the distributions of $\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \boldsymbol{B}_1 \boldsymbol{\varepsilon}$ and $\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \boldsymbol{B}_2 \boldsymbol{\varepsilon}$.

$\frac{1}{\sigma^2}\varepsilon^T\varepsilon = \frac{1}{\sigma^2}\varepsilon^T H \varepsilon + \frac{1}{\sigma^2}\varepsilon^T(I-H)\varepsilon$

χ^2_{p+1} $\qquad$ $\chi^2_{n-(p+1)}$

\# $\operatorname{rank}(H) = p+1$
$\operatorname{rank}(I-H) = n-(p+1)$

(p) [easy] Show that $\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \boldsymbol{B}_1 \boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\|\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{\beta})\|^2$.

$\frac{1}{\sigma^2}\varepsilon^T H \varepsilon = \frac{1}{\sigma^2}\varepsilon^T H H \varepsilon = \frac{1}{\sigma^2}\|H\varepsilon\|^2 = \frac{1}{\sigma^2}\|H(Y - X\beta)\|^2$

$= \frac{1}{\sigma^2}\|HY - HX\beta\| = \frac{1}{\sigma^2}\|X(B-\beta)\|^2$

(q) [harder] Why is the term $\|\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{\beta})\|^2$ used to measure the model's "estimation error"?

$\|X(B-\beta)\|^2 = \|XB - X\beta\|^2$

Definition of misspecification

(r) [easy] Show that $\frac{1}{\sigma^2}\boldsymbol{\varepsilon}^\top \boldsymbol{B}_2 \boldsymbol{\varepsilon} = \frac{1}{\sigma^2}\|\boldsymbol{E}\|^2$.

$\frac{1}{\sigma^2}\varepsilon^T(I_n-H)\varepsilon = \frac{1}{\sigma^2}\varepsilon^T(I_n-H)^T(I_n-H)\varepsilon = \frac{1}{\sigma^2}\|(I_n-H)\varepsilon\|^2$

$= \frac{1}{\sigma^2}\|(I_n-H)(Y-X\beta)\|^2 =$ ~~$\frac{1}{\sigma^2}\|XB - HXB\|^2$~~

$= \frac{1}{\sigma^2}\|Y - X\beta - HY + HX\beta\|^2 = \frac{1}{\sigma^2}\|Y - X\beta - \hat{Y} + X\beta\|^2$

$= \frac{1}{\sigma^2}\|Y - \hat{Y}\|^2$

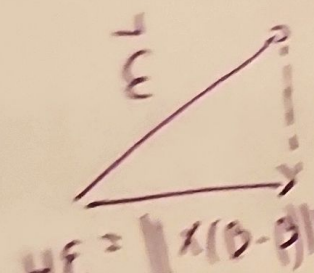
$= \frac{1}{\sigma^2}\|\varepsilon\|^2$

(a) [1 marks] In what scenarios is $\boldsymbol{\varepsilon}^T \boldsymbol{B}_1 \boldsymbol{\varepsilon} > \boldsymbol{\varepsilon}^T \boldsymbol{B}_2 \boldsymbol{\varepsilon}$?

$\varepsilon^T (H \varepsilon^* > \varepsilon^T (I-H) \varepsilon$

when p+1 > n

(b) [1 marks] Draw an illustration of $\boldsymbol{\varepsilon}$ being orthogonally projected onto colsp $[\boldsymbol{X}]$ via projection matrix $\boldsymbol{H}$. Use the previous answers to denote the quantities of the projection and the error of the projection.



$H\varepsilon = H(Y = X\beta)$

$= X(\beta-\hat{\beta})$

(c) [1 marks] A good linear model has a large or a small projection of the error? Discuss.

small because

~~Once the data the error and proj is small~~

~~A explains part of the error~~

If it were large it would mean a lot of the error is found in the colsp of the feature

(d) [1 marks] Find $\mathbb{E}\left[\frac{1}{n}\|\hat{\boldsymbol{\varepsilon}}\|^2\right]$.

$\mathbb{E}\left[\frac{1}{n}\|\hat{\varepsilon}\|^2\right] = n-(p+1)$

(e) [1 marks] Show that $\frac{\|\hat{\boldsymbol{\varepsilon}}\|^2}{n-(p+1)}$ is an unbiased estimate of σ^2.

$\mathbb{E}\left[\frac{\|\hat{\varepsilon}\|^2}{n-(p+1)}\right] = \sigma^2$

(x) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(B_j-\beta_j)}{\|\boldsymbol{E}\|\sqrt{(\boldsymbol{X}^T\boldsymbol{X})^{-1}_{j,j}}} \sim T_{n-(p+1)}$.

$$\frac{B_j - \beta_j}{\sqrt{(X^TX)^{-1}_{j,j}}} \sim N(0,1)$$

$$\sqrt{\frac{\frac{1}{\sigma^2}\|E\|^2}{n-(p+1)}} \sim \chi^2_{n-(p+1)}$$

$$\sim T_{n-(p+1)}$$

$$= \frac{\sqrt{n-(p+1)}\,(B_j-\beta_j)}{\|E\|\sqrt{(X^TX)^{-1}_{j,j}}} \sim T_{n-(p+1)}$$

(y) [easy] Let $H_0 : \beta_j = 0$. Find the test statistic using the fact from the previous question.

Let s_e denote $RMSE := \sqrt{MSE} := \sqrt{SSE/(n-(p+1)} = \sqrt{\|e\|^2/(n-(p+1)}$.

$$\frac{b_j - \beta_j}{s_e\sqrt{(X^TX)^{-1}_{j,j}}}$$

(z) [easy] Consider a new parameter of interest $\mu_\star := \mathbb{E}[Y_\star] = \boldsymbol{x}_\star\boldsymbol{\beta}$, this is the expected response for a unit with measurements given in row vector $\boldsymbol{x}_\star$ whose first entry is 1. Prove that $\dfrac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{\boldsymbol{x}_\star(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,1)$.

$$\hat{Y}_\star = \vec{x}_\star\vec{\beta} \Rightarrow \hat{Y}_\star \sim N(\underbrace{\vec{x}_\star\vec{\beta}}_{\mu_\star}, \sigma^2\vec{x}_\star(X^TX)^{-1}\vec{x}_\star^T)$$

$$\Rightarrow \frac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{x_\star(X^TX)^{-1}x_\star^T}} \sim N(0,1)$$

(aa) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(\hat{Y}_\star - \mu_\star)}{\|\boldsymbol{E}\|\sqrt{\boldsymbol{x}_\star (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_\star^\top}} \sim T_{n-(p+1)}$.

$$\frac{\frac{Y_0 - \mu_0}{\sigma\sqrt{\vec{x}_0(X^TX)^{-1}\vec{x}_0^T}} \sim N(0,1)}{\sqrt{\frac{\frac{1}{\sigma^2}\|E\|^2}{n-(p+1)} \sim \chi^2_{n-(p+1)}}} \sim T_{n-(p+1)}$$

$$= \frac{n-(p+1)\left(\hat{Y}_0 - \mu_0\right)}{\|E\|\sqrt{\vec{x}_0(X^TX)^{-1}\vec{x}_0^T}}$$

(bb) [easy] Let $H_0 : \mu_\star = 17$. Find the test statistic using the fact from the previous question. Let s_e denote the $RMSE$.

$$\frac{\hat{Y}_0 - 17}{s_e\sqrt{\vec{x}_0(X^TX)^{-1}\vec{x}_0^T}}$$

(cc) [easy] Consider a new parameter of interest $y_\star = \boldsymbol{x}_\star\boldsymbol{\beta} + \epsilon_\star$, this is the response for a unit with measurements given in row vector $\boldsymbol{x}_\star$ whose first entry is 1. Prove that $\dfrac{\hat{Y}_\star - y_\star}{\sigma\sqrt{1+\boldsymbol{x}_\star(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,1)$.

$$Y_0 - \hat{Y}_0 = N(0, \sigma^2 + \sigma^2\vec{x}_0(X^TX)^{-1}\vec{x}_0^T) = N(0, \sigma^2(1+\vec{x}_0(X^TX)^{-1}\vec{x}_0^T))$$

$$\frac{Y_0 - \hat{Y}_0}{\sqrt{\sigma^2(1+\vec{x}_0(X^TX)^{-1}\vec{x}_0^T)}} \sim N(0,1)$$

(dd) [easy] Prove that $\dfrac{\sqrt{n-(p+1)}(\hat{Y}_\star - y_\star)}{\|\boldsymbol{E}\|\sqrt{1+\boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim T_{n-(p+1)}.$

$Y_* - \hat{Y}_*$

$\sqrt{\sigma^2(1+\vec{x}_*(X^TX)^{-1}\vec{x}_*^T)}$ } $N(0,1)$

$\sqrt{\dfrac{\|E\|}{\sigma^2}}$ } χ^2_{n-r} $= \dfrac{\sqrt{n-(p+1)}(\hat{Y}_* - \hat{y}_*)}{\|E\|\sqrt{1+x_*(X^TX)^{-1}x_*^T}} \sim T_{n-(p+1)}$

(ee) [easy] Let $H_0 : y_\star = 37$. Find the test statistic using the fact from the previous question. Let s_e denote the $RMSE$.

$$\frac{37 - \hat{y}_*}{s_e\sqrt{1-\vec{x}_*(X^TX)^{-1}\vec{x}_*^T}}$$

(ff) [difficult] Let $S \subseteq \{1, 2, \ldots, p\}$, let $k := |S|$ and let $A = \{0\} \cup S^C$, its complement with zero for the index of the intercept. For convenience, assume you rearrange the columns of the design matrix so that $\boldsymbol{X} = [\boldsymbol{X}_A \mid \boldsymbol{X}_S]$ and the first column is $\mathbf{1}_n$. Let $\boldsymbol{H}_A := \boldsymbol{X}_A(\boldsymbol{X}_A^\top\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^\top$. It is obvious that $\boldsymbol{H} - \boldsymbol{H}_A$ is symmetric as both $\boldsymbol{H}$ and $\boldsymbol{H}_A$ are symmetric. To prove that $\boldsymbol{H} - \boldsymbol{H}_A$ is an orthogonal projection matrix, prove that it is idempotent. Hint: use the Gram-Schmidt decomposition for both matrices and use block matrix format for $\boldsymbol{H}$.

$X_A = Q_A R_A$, $X = QR$

$(H-H_A)(H-H_A)^T = (QQ^T - Q_AQ_A^T)(QQ^T - Q_AQ_A^T)^* = [Q_A|Q_{-A}]\begin{bmatrix}Q_A^T \\ Q_{-A}^T\end{bmatrix} - Q_AQ_A^T)($

$= (\cancel{Q_AQ_A^T} + Q_{-A}Q_{-A}^T) - \cancel{Q_AQ_A^T}(\cancel{Q_AQ_A^T} + Q_AQ_{-A}^T - \cancel{Q_AQ_A^T})$

$= (Q_{-A}Q_{-A}^T)\cancel{(Q_{-A}Q_{-A})^*} = (Q_{-A}Q_{-A}^T)$

$= (H - H_A)$

(gg) [easy] Let $\hat{\boldsymbol{Y}}_A := \boldsymbol{H}_A \boldsymbol{Y}$, the orthogonal projection onto colsp $[\boldsymbol{X}_A]$. Prove that

$$\frac{(n-(p+1))\left\|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A\right\|^2}{k\|\boldsymbol{E}\|^2} \sim F_{k,n-(p+1)}.$$

$\frac{1}{\sigma^2}\vec{\varepsilon}^T(H-H_0)\vec{\varepsilon} \sim \chi^2_k$

$= \frac{1}{\sigma^2}\|\vec{Y}-\vec{\hat{Y}}_A + (I-H_A)X_S\beta_S\|^2$

$\Rightarrow \frac{\frac{1}{\sigma^2}\|\vec{\hat{Y}}-\vec{\hat{Y}}_A\|^2}{k} = \frac{(n-(p+1))\|\vec{Y}-\vec{\hat{Y}}_A\|^2}{k\|\vec{E}\|^2} \sim F_{k,n-(p+1)}$

under H_0 $\quad \frac{\|\vec{E}\|^2}{n-(p+1)}$

(hh) [difficult] Let $\hat{\boldsymbol{E}}_A := (\boldsymbol{I}_n - \boldsymbol{H}_A)\boldsymbol{Y}$, the orthogonal projection onto the colsp $[\boldsymbol{X}_{A^\perp}]$. Prove that $\left\|\hat{\boldsymbol{E}}_A\right\|^2 - \left\|\hat{\boldsymbol{E}}\right\|^2 = \left\|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A\right\|^2$.

$\|(I_n-H_A)Y\|^2 - \|Y-\hat{Y}\|^2 = \|Y-\hat{Y}_A\|^2 - \|Y-\hat{Y}\|^2$

$= (Y-\hat{Y}_A)^T(Y-\hat{Y}_A) - (Y-\hat{Y})^T(Y-\hat{Y})$

$= Y^TY - Y^T\hat{Y}_A - \hat{Y}_A^TY + \hat{Y}_A^T\hat{Y}_A - Y^TY - \hat{Y}^TY + \hat{Y}^TY - \hat{Y}^T\hat{Y}$

$= (\hat{Y}-\hat{Y}_A)^T(\hat{Y}-\hat{Y}_A) = \|\hat{Y}-\hat{Y}_A\|^2$

(ii) [easy] Combining the two previous problems, write the test statistic for $H_0 : \boldsymbol{\beta}_S = \boldsymbol{0}_k$ where β_S denotes the subvector of $\boldsymbol{\beta}$ with indices S. Use the notation $\Delta SSE := SSE_A - SSE$ and MSE.

$F = \frac{\frac{SSE_A - SSE}{k}}{MSE}$

(jj) [difficult] Prove that the square root of the test statistic in (ii) is the same as t-test statistic from (y) when $k = 1$.

(kk) [harder] The point of this exericse is to demonstrate that the estimator used for the omnibus / global / overall F-test is nothing but a special case of the main result from (gg). Let $S = \{1, 2, \ldots, p\}$ and thus $k = p$ and $A = \{0\}$. Using the result from (gg), show that $\dfrac{(n-(p+1))\left|\left|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}_n\right|\right|^2}{p\,||\boldsymbol{E}||^2} \sim F_{p,n-(p+1)}$.

(ll) [easy] Prove that the omnibus / global / overall F-test statistic is $\hat{\hat{F}} = MSR/MSE$ by using the result from (kk).

(mm) [difficult] [MA] Prove that the distribution that realizes the R^2 metric (the proportion of response variance explained by the model) is distributed as Beta $\left(\frac{p}{2}, \frac{n-(p+1)}{2}\right)$. This amounts to proving a fact found at the bottom of the F distribution's Wikipedia page

.

(nn) [easy] Prove that the maximum likelihood estimate for $\boldsymbol{\beta}$ is $\boldsymbol{b}$, the OLS estimator.

$\vec{Y} \sim \mathcal{N}_n(X\vec{\beta}, \sigma^2 I_n)$

$$L(\vec{\beta}, \sigma^2; \vec{y}, X) = \frac{1}{(2\pi)^{n/2}\sqrt{\det(\sigma^2 I_n)}} e^{-\frac{1}{2}(\vec{y}-X\vec{\beta})^T(\sigma^2 I_n)^{-1}(\vec{y}-X\vec{\beta})}$$

$$\ell(\vec{\beta}, \sigma^2; \vec{y}, X) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\|\vec{y}-X\vec{\beta}\|^2$$

$$\frac{d}{d\vec{\beta}}[\ell] = -\frac{1}{2\sigma^2}\frac{d}{d\vec{\beta}}\left[\|\vec{y}-X\vec{\beta}\|^2\right] \overset{set}{=} 0$$

$$\Rightarrow \hat{\beta}^{MLE} = (X^TX)^{-1}X^T\vec{y} = \vec{b}$$

(oo) [harder] Prove that the maximum likelihood estimate for σ^2 is SSE/n.

$$\frac{d}{d\sigma^2}[\ell] = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2}\|\vec{y}-X\vec{\beta}\|^2 \overset{set}{=} 0 \Rightarrow -n + \frac{1}{\sigma^2}\|\vec{y}-X\vec{\beta}\|^2 = 0$$

$$\hat{\sigma}^{2,MLE} = \frac{1}{n}\|\vec{y}-X\vec{\beta}\| = \frac{1}{n}\|\vec{y}-X(X^TX)^{-1}X^T\vec{y}\|$$

$$= \frac{1}{n}\|(I-H)\vec{y}\|^2$$

$$= \frac{1}{n}\|\vec{e}\|^2 = \frac{1}{n}\sum e_i^2 = \frac{1}{n}SSE$$

(pp) [harder] Find the bias of the maximum likelihood estimator for σ^2 using your answers from (w) and (oo).

$$\text{Bias}(\hat{\sigma}^{2\,MLE}) = E\left[\frac{1}{n}SSE\right] - E\left[\frac{SSE}{n-(p+1)}\right]$$

$$= E\left[\|\varepsilon\|^2\left(\frac{1}{n} - \frac{1}{n-(p+1)}\right)\right]$$

$$\frac{1}{n} - \frac{1}{n-(p+1)}\,E\left(\|\varepsilon\|^2\right)$$

$$= \left(\frac{n-(p+1)}{n} - 1\right)\sigma^2 = \left(\frac{-(p+1)}{n}\right)\sigma^2$$

Problem 2

This problem is about two types of Bayesian estimation of the slope parameters in linear regression which lead to the ridge and lasso estimates.

(a) [easy] Write the prior assumption about $\boldsymbol{\beta}$ which yields the ridge estimates.

$$f(\vec{\beta}) = N_{p+1}(\vec{0}_{p+1}, \tau^2 I_{p+1}) \qquad f(\sigma^2) \propto \frac{1}{\sigma^2}$$

(b) [easy] Using the prior and core assumption (which implies a likelihood function for $\boldsymbol{B}$), derive the ridge estimates.

$$f(\vec{\beta}, \sigma^2 | X, \vec{y}) \propto (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{1}{2\sigma^2}\|\vec{y}-X\vec{\beta}\|^2} e^{-\frac{1}{2\tau^2}\|\vec{\beta}\|^2}$$

$$\vec{\beta}^{MAP} = \underset{\vec{\beta}}{\text{argmax}}\left\{\left(-\frac{n}{2}-1\right)\ln(\sigma^2) - \frac{1}{2\sigma^2}\|\vec{y}-X\vec{\beta}\|^2 - \frac{1}{2\tau^2}\|\vec{\beta}\|^2\right\}$$

$$= \text{argmax}\left\{-\frac{1}{2\sigma^2}\left(\|\vec{y}-X\vec{\beta}\|^2 + \frac{\sigma^2}{\tau^2}\|\vec{\beta}\|^2\right)\right\}$$

$$= \text{argmin}\left\{\|\vec{y}-X\vec{\beta}\|^2 + \frac{\sigma^2}{\tau^2}\|\vec{\beta}\|^2\right\}$$

$$= \frac{d}{d\vec{\beta}}\left[(\vec{y}-X\vec{\beta})^T(\vec{y}-X\vec{\beta}) + \lambda\vec{\beta}^T\vec{\beta}\right]$$

$$= \frac{d}{d\vec{\beta}}\left[\vec{y}^T\vec{y} - 2\vec{\beta}^T X^T\vec{y} + \vec{\beta}^T X^T X\vec{\beta} + \vec{\beta}^T(\lambda I_{p+1})\vec{\beta}\right]$$

$$= \frac{d}{d\vec{\beta}}\left[-2\vec{\beta}^T X^T\vec{y} + \vec{\beta}^T(X^TX + \lambda I_{p+1})\vec{\beta}\right]$$

$$-2X^T\vec{y} + 2(X^TX + \lambda I_{p+1})\vec{\beta} \overset{SET}{=} 0$$

$$X^T\vec{y} = (X^TX + \lambda I_{p+1})\vec{\beta} \Rightarrow \vec{b}_{ridge} = (X^TX + \lambda I_{p+1})^{-1}X^T\vec{y}$$

(h) [easy] Describe why Lasso estimation has the added bonus of being able to perform variable selection and ridge does not.

Because Lasso has argmin at $\beta = 0$ due to the Laplace prior

Problem 3

This problem is about the specific robust regression methods we studied.

(a) [easy] If we only know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent, what tried and true method can we employ to get asymptotically valid inference for $\boldsymbol{\beta}$?

Bootstrap

(b) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are iid with expectation zero and variance σ^2 for all values of $\boldsymbol{x}$ (i.e. the errors are "homoskedastic") but the errors are not necessarily normal, what is the asymptotic distribution of $\boldsymbol{B}$?

$\vec{B} \sim N(\vec{\beta}, \sigma^2 (X^T X)^{-1})$

(c) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n \overset{ind}{\sim} \mathcal{N}(0, \sigma_i^2)$ which means the errors are "heteroskedastic", what is the asymptotic distribution of $\boldsymbol{B}$ using the Huber-White estimator?

$\vec{B} \sim N(\vec{\beta}, (X^T X)^{-1} X^T \hat{D} X (X^T X)^{-1})$

(d) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are independent with expectation zero and variance σ_i^2 which means the errors are "heteroskedastic", what is the asymptotic distribution of $\boldsymbol{B}$ using the Huber-White estimator?

$\vec{B} \sim N(\vec{\beta}, (X^T X)^{-1} X^T \hat{D} X (X^T X)^{-1})$

(e) [easy] Is the F-tests we derived under the core assumption valid in any of the four above scenarios? Yes/no

(c) [easy] Write the prior assumption about $\boldsymbol{\beta}$ which yields the lasso estimates.

$\beta_0 \dots \beta_p \overset{iid}{\sim} \text{Laplace}(0, \tau^2) = \frac{1}{2\tau} \cdot e^{-\frac{|\beta|}{\tau}}$

(d) [easy] Using the prior and core assumption (which implies a likelihood function for $\boldsymbol{B}$), derive the lasso estimates to the point where you need to use a computer to run the optimization.

$f(\vec{\beta}, \sigma^2 | X, \vec{y}) \propto (\sigma^2)^{-\frac{n}{2}-1} e^{-\frac{1}{2\sigma^2}\|\vec{y} - X\vec{\beta}\|^2} e^{-\frac{1}{\tau^2}\sum_{j=0}^{p}|\beta_j|}$

$\hat{\beta}^{MAP} = \text{argmax}\left\{\left(-\frac{n}{2}-1\right)\ln\sigma^2 - \frac{1}{2\sigma^2}\|\vec{y} - X\vec{\beta}\|^2 - \frac{1}{\tau^2}\sum_{j=0}^{p}|\beta_j|\right\}$

$= \text{argmax}\left\{-\frac{1}{2\sigma^2}\left(\|\vec{y} - X\vec{\beta}\|^2 + \frac{2\sigma^2}{\tau^2}\sum_{j=0}^{p}|\beta_j|\right)\right\}$

$= \text{argmin}\left\{\|\vec{y} - X\vec{\beta}\|^2 + \frac{2\sigma^2}{\tau^2}\sum|\beta_j|\right\}$

$b_{lasso} = \text{argmin}\{SSE + \lambda\sum|\beta_j|\}$

(e) [easy] Both ridge and lasso shrink the estimate of $\boldsymbol{\beta}$ towards what vector value? $\vec{0}$

(f) [easy] Describe what the prestep called "variable selection" is within the modeling enterprise.

filtering out variables that have little impact on the model

(g) [easy] Describe what the prestep called "variable selection" is within the modeling enterprise.

Problem 4

This problem is about inference for the generalized linear model (glm).

(a) [harder] Let $Y_i \overset{ind}{\sim} \text{Bernoulli}(\theta_i)$ for $i = 1, \ldots, n$ where $\theta_i = \phi(\boldsymbol{x}_i\boldsymbol{\beta})$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. For the link function, use the complementary log-log (i.e. the standard Gumbel CDF). Write out the full likelihood below. No need to simplify or take the log.

$$L(\vec{\theta}^{MLE}; \vec{x}) = \prod \left(\frac{1}{1+e^{-\vec{x}_i \vec{b}}}\right)^{Y_i} \left(\frac{1}{1+e^{\vec{x}_i \vec{b}}}\right)^{1-Y_i}$$

(b) [harder] Given the assumptions in (a), write the likelihood ratio estimate for the omnibus test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$.

$$\hat{LR} = \frac{\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\vec{x}_i \vec{b}}}\right)^{Y_i} \left(\frac{1}{1+e^{\vec{x}_i \vec{b}}}\right)^{1-Y_i}}{\prod_{i=1}^{n} \left(\frac{1}{1+e^{-\tilde{b}_2}}\right)^{Y_i} \left(\frac{1}{1+e^{\tilde{b}_2}}\right)^{1-Y_i}}$$

$$\tilde{b}_2 = \text{argmax} \left(\prod \left(\frac{1}{1+e^{-w}}\right)^{y_i} \left(\frac{1}{1+e^{w}}\right)^{1-y_i} \right)$$

(c) [harder] Let $Y_i \overset{ind}{\sim} \text{Poisson}(\theta_i)$ for $i = 1, \ldots, n$ where $\theta_i = e^{\boldsymbol{x}_i\boldsymbol{\beta}}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. Write out the likelihood ratio when testing $H_0 : \beta_2 = \beta_3 = 0$.

$$\hat{LR} = \frac{\prod_{i=1}^{n} \frac{e^{(\vec{x}_i \vec{b})Y_i} e^{e^{\vec{x}_i \vec{b}}}}{y_i!}}{\frac{e^{\vec{b}(Y_i)} e^{e^{\vec{b}}}}{y_i!} + \prod_{3}^{n} \frac{e^{(\vec{x}_i \vec{b})Y_i} e^{e^{\vec{x}_i \vec{b}}}}{y_i}}$$

(d) [harder] Let $Y_i \overset{ind}{\sim} \text{Weibull}(k, \theta_i)$ for $i = 1, \ldots, n$ where $\theta_i = e^{\boldsymbol{x}_i\boldsymbol{\beta}}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. This uses the alternate parameterization so that $\mathbb{E}[Y_i] = \theta_i\Gamma(1 + 1/k)$. There is a censoring vector $\boldsymbol{c}$ which is 1 when censored on the right (meaning the real y_i is $\geq$ to the observed y_i) and 0 when not censored. Write out the likelihood ratio when testing $H_0 : \beta_2 = \beta_3 = 0$.

$$\tilde{LR} = \prod_{\{i:c_i=0\}} \frac{k}{e^{\vec{x}_i\vec{b}}}\left(\frac{y_i}{e^{\vec{x}_i\vec{b}}}\right)^{k-1} e^{-\left(\frac{y_i}{e^{\vec{x}_i\vec{b}}}\right)^k} \prod_{i:c_i=1} e^{-\left(\frac{y_i}{e^{\vec{x}_i\vec{b}}}\right)^k}$$

$$\prod_{c_i=0} \frac{k}{e^{\vec{x}_i\vec{b}}}\left(\frac{y_i}{e^{\vec{x}_i\vec{b}}}\right)^{k-1} e^{-\left(\frac{y_i}{e^{\vec{x}_i\vec{b}}}\right)} \prod_{i:c_i=1} e^{-\left(\frac{y_i}{x_i\vec{b}_0}\right)}$$

where $\vec{b} = \text{argmax}\left\{\prod_{c_i=0} \frac{k}{e^{\vec{x}_i\vec{b}}}\left(\frac{y_i}{e^{\vec{x}_i\vec{b}}}\right)^{k-1} e^{-\left(\frac{y_i}{e^{\vec{x}_i\vec{b}}}\right)} \prod_{i:c_i=1} e^{\left(\frac{y_i}{e^{\vec{x}_i\vec{b}}}\right)^k}\right.$ under H_0

(e) [difficult] [MA] Let $Y_i \overset{ind}{\sim} \mathcal{N}(\theta_i, \sigma^2)$ for $i = 1, \ldots, n$ where $\theta_i = \boldsymbol{x}_i\boldsymbol{\beta}$ and $\boldsymbol{x}_i \in \mathbb{R}^{p+1}$ whose first entry is always 1. So far, this is the vanilla linear model. However, consider now a wrinkle: there is a censoring vector $\boldsymbol{c}$ which is 1 when censored on the right (meaning the real y_i is $\geq$ to the observed y_i) and 0 when not censored. This is called the Tobit model. Write the likelihood ratio estimate for the omnibus test of $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$.