# MATH 343 / 643 Homework #2

Natasha Watson

Friday 11th April, 2025

## 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?

$$\vec{\boldsymbol{\beta}}$$

The purpose of OLS estimation and inference is to estimate and draw conclusions about the components of $\boldsymbol{\beta}$. IT is the p+1 dimensional vector of true regression coefficients of the model.

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

The design matrix $\boldsymbol{X}$ of predictor values, including an intercept column of 1's. The sample size $n$ and the number of features $p$.

(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?

The error term $\epsilon$ is random, we consider $\epsilon_i$ to be a random variable.

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

The core assumption is the combination of 3 main assumptions about the $\epsilon_i$'s, and becomes our base case:

(1) Let $\epsilon_i = u_{1i} + u_{2i} + u_{3i} + \ldots$, i.e. the sum of many unknown values is likely independent $\Rightarrow \epsilon_i$ is realized from an independent normal distribution

(2) The mean of the Random Variable $\epsilon_i$ was drawn from is zero (mean centered)

(3) The variance is equal $\forall i, \sigma^2$ (Homoscedacity Assumption)

The amalgamation of these three assumptions leads us to the core assumption:

$$\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n \overset{iid}{\sim} \mathcal{N}\left(0,\, \sigma^2\right)$$

Where $\epsilon_1, \epsilon_2, \ldots \epsilon_n$ are its realizations $\leftrightarrow \vec{\mathcal{E}} \sim \mathcal{N}_n\left(\vec{0}_n,\, \sigma^2 I_n\right)$

We are able to say via the core assumption that (1) the relationship between predictors and the response is linear in parameters, the error is normally distributed, there is a constant variance, and the errors are uncorrelated (ind)

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

We recall from Math 340 that if $\vec{X} \sim \mathcal{N}_n\left(\mu,\, \Sigma\right)$ and if $A \in \mathbb{R}^{k \times n}, b \in \mathbb{R}^k$,
then $\vec{Y} = \vec{b} + A\vec{X} \sim \mathcal{N}_k\left(A\vec{\mu} + \vec{b},\, A\Sigma A\top\right)$

So if $\vec{Y} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\mathcal{E}} \sim \mathcal{N}_n\left(\boldsymbol{X}\vec{\boldsymbol{\beta}},\, \sigma^2 I_n\right)$, then we can think of $\vec{B}$ as being a random vector that realizes $\vec{b}$. If we recall that OLS estimator for $\vec{B} = (\boldsymbol{X}\top\boldsymbol{X})^{-1}\boldsymbol{X}\top\vec{Y}$ we can derive its distribution like so:

$$\vec{B} = (\boldsymbol{X}\top\boldsymbol{X})^{-1}\boldsymbol{X}\top\vec{Y}$$

Recall $\vec{Y} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\mathcal{E}}$

$$\vec{B} = (\boldsymbol{X}\top\boldsymbol{X})^{-1}\boldsymbol{X}\top\boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\mathcal{E}}$$

Distribute

$$\Rightarrow (\boldsymbol{X}\top\boldsymbol{X})^{-1}\boldsymbol{X}\top\boldsymbol{X}\vec{\boldsymbol{\beta}} + (\boldsymbol{X}\top\boldsymbol{X})^{-1}\boldsymbol{X}\top\vec{\mathcal{E}}$$

Notice that $(\boldsymbol{X}\top\boldsymbol{X})^{-1}\boldsymbol{X}\top\boldsymbol{X} = \boldsymbol{I}$ and so this term cancels out. We are then left with:

$$\Rightarrow \vec{\boldsymbol{\beta}} + (\boldsymbol{X}\top\boldsymbol{X})^{-1}\boldsymbol{X}\top\vec{\mathcal{E}}$$

And

$$\vec{B} \sim \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}},\, \sigma^2(\boldsymbol{X}\top\boldsymbol{X})^{-1}\right)$$

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

Recall that the elements of a multivariate normal random variable are distributed as univariate normals. Therefore,

$$B_j \sim \mathcal{N}\left(\boldsymbol{\beta}_j,\, \sigma^2(\boldsymbol{X}\top\boldsymbol{X})^{-1}_{j,j}\right)$$

Where $(\boldsymbol{X}\top\boldsymbol{X})^{-1}_{j,j}$ is the j,j-th index of the covariance matrix, because the diagonal entries give the variances of each component

This distribution is the foundation for confidence intervals for $\boldsymbol{\beta}_j$ and let's sue perform t-tests once we estimate $\sigma^2$.

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

$B_j$ standardized:

$$\frac{B_j - \boldsymbol{\beta}_j}{\sqrt{\sigma^2(\boldsymbol{X}\top\boldsymbol{X})^{-1}_{j,j}}} \sim \mathcal{N}\,(0,\, 1)$$

Recall, this is when variance is known! If not then this becomes a t-dsitrbution with n-(p+1) degrees of freedom and we use $\hat{\sigma}^2$

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

Recall from 342 $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$ where $\boldsymbol{X}(\boldsymbol{X}\top\boldsymbol{X})^{-1}\boldsymbol{X}\top$ is the hat matrix. If we express $\hat{\boldsymbol{Y}}$ in terms of $\epsilon$, and we know $\vec{Y} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\mathcal{E}}$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\vec{B} = \boldsymbol{H}\boldsymbol{Y} = \boldsymbol{H}(\boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\mathcal{E}})$$

Distribute

$$\boldsymbol{H}\boldsymbol{X}\vec{\boldsymbol{\beta}} + \boldsymbol{H}\vec{\mathcal{E}}$$

Recall $\boldsymbol{H}$ is in the colsp $[\boldsymbol{X}]$ and so $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$:

$$\hat{\boldsymbol{Y}} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \boldsymbol{H}\vec{\mathcal{E}} \sim ?$$

Recall from math 340:

$$\vec{U} \sim \mathcal{N}_n\left(\vec{0}_n,\, \mathcal{E}\right) \quad \mu \in \mathbb{R}^m; \boldsymbol{A} \in \mathbb{R}^{m\times n} \Rightarrow \vec{\mu} + \boldsymbol{A}\vec{U} \sim \mathcal{N}_m\,(\vec{\mu},\, \boldsymbol{A}\mathcal{E}A\top)$$

Applying this shift and scale method we see,

$$\hat{\boldsymbol{Y}} \sim \mathcal{N}_n\left(\boldsymbol{X}\vec{\boldsymbol{\beta}},\, \boldsymbol{H}(\sigma^2 I_n)\boldsymbol{H}\top\right)$$

If we pull out the scalar $\sigma^2$ in the variance term, then we are left with

$$\sigma^2(\boldsymbol{H}(I_n)\boldsymbol{H}\top) \Rightarrow \sigma^2(\boldsymbol{H}\boldsymbol{H}\top)$$

By idempotentcy and symmetry of the hat matrix, this reduces to:

$$\sigma^2 \boldsymbol{H}$$

And so the final distribution of $\hat{\boldsymbol{Y}}$ is:

$$\hat{\boldsymbol{Y}} \sim \mathcal{N}_n \left( \boldsymbol{X}\vec{\boldsymbol{\beta}},\, \boldsymbol{H}\sigma^2 \right)$$

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

Recall that the elements of a multivariate normal random variable are distributed as univariate normals. Therefore,'

$$\hat{Y}_i \sim \mathcal{N} \left( \vec{x}_{i\cdot}\vec{\boldsymbol{\beta}},\, \sigma^2 \boldsymbol{H}_{i,i} \right)$$

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

Subtract the mean and divide by the standard deviation:

$$\frac{\hat{\boldsymbol{Y}}_i - \vec{x}_{i\cdot}\vec{\boldsymbol{\beta}}}{\sqrt{\sigma^2 \boldsymbol{H}_{i,i}}} \sim \mathcal{N}\left(0,\, 1\right)$$

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

Recall we assumed $\mathcal{E} \sim \mathcal{N}_n \left( \vec{0}_n,\, \sigma^2 \boldsymbol{I}_n \right)$. Rewriting the residuals, we get:

$$\vec{E} = \vec{Y} = \vec{\hat{\boldsymbol{Y}}} = \vec{Y} - \boldsymbol{X}\vec{B}$$

Recall

$$\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}$$

Substitute in OLS estimator $\boldsymbol{B} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$:

$$\boldsymbol{Y} - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$$

Factor out $\boldsymbol{Y}$:

$$= \left( \boldsymbol{I}_n - \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \right) \boldsymbol{Y}$$

Recognize the hat matrix $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top$:

$$= (\boldsymbol{I}_n - \boldsymbol{H})\boldsymbol{Y}$$

Use this:

$$(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\mathcal{E}} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\vec{\boldsymbol{\beta}} + (\boldsymbol{I} - \boldsymbol{H})\mathcal{E} \sim \mathcal{N}_n\left(\vec{0}_n,\ (\boldsymbol{I} - \boldsymbol{H})\sigma^2\boldsymbol{I}_n(\boldsymbol{I} - \boldsymbol{H})\top\right)$$

$$\Rightarrow \vec{E} \sim \mathcal{N}_n\left(\vec{0}_n,\ \sigma^2(\boldsymbol{I} - \boldsymbol{H})\right)$$

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

Recall that the elements of a multivariate normal random variable are distributed as univariate normals. Therefore,

$$E_i \sim \mathcal{N}\left(0,\ \sigma^2(1 - \boldsymbol{H}_{i,i})\right)$$

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

Subtract the mean and divide by the SD:

$$\frac{E_i}{\sqrt{\sigma^2(1 - \boldsymbol{H}_{i,i})}} \sim \mathcal{N}\left(0,\ 1\right)$$

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

Recall $\vec{Z} \sim \mathcal{N}_n\left(\vec{0_n},\ \sigma^2\boldsymbol{I}_n\right) \Rightarrow \vec{Z}\top\vec{Z} \sim \chi^2_n$

This is equivalent to saying $\vec{Z}\top\vec{Z} = \frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{\mathcal{E}} \sim \chi^2_n$, where $\vec{Z} = \frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}$

(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 $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{B}_1\boldsymbol{\mathcal{E}}$ and $\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{B}_2\boldsymbol{\mathcal{E}}$.

We know that $\boldsymbol{B}_1 + \boldsymbol{B}_2 = \boldsymbol{I}_n$ and because $\boldsymbol{H}$ is the orthogonal projection matrix onto the colsp $[\boldsymbol{X}]$, then the $\mathbf{rank}[B_1] = p + 1$ and $\mathbf{rank}[B_2] = n - (p + 1)$. We can use Cochran's theorem because $\mathbf{rank}[B_1] + \mathbf{rank}[B_2] = n$

Therefore,

$$\vec{Z}\top B_1\vec{Z} \sim \chi^2_{p+1}$$

and

$$\vec{Z}\top B_2\vec{Z} \sim \chi^2_{n-(p+1)}$$

Or,

$$\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{H}\boldsymbol{\mathcal{E}} \sim \chi^2_{p+1}$$

and

$$\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mathcal{E}} \sim \chi^2_{n-(p+1)}$$

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

Recall, $\boldsymbol{B} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{Y}$

$$= (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top(\boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\mathcal{E}}) = \boldsymbol{I}\vec{\boldsymbol{\beta}} + (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \vec{\mathcal{E}}$$

$$\Rightarrow \boldsymbol{B} - \boldsymbol{\beta} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \vec{\mathcal{E}}$$

Multiply both sides by $\boldsymbol{X}$

$$\Rightarrow \boldsymbol{X}(\boldsymbol{B}-\boldsymbol{\beta}) = (\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}\vec{\mathcal{E}}$$

Notice $(\boldsymbol{X}^\top \boldsymbol{X})^{-1}\boldsymbol{X}^\top \boldsymbol{X}$ is the hat matrix, which is in the colsp $[\boldsymbol{X}]$. Recall $\boldsymbol{B}_1 = \boldsymbol{H}$. Take the norm and square both sides:

$$\left|\left|\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{\beta})\right|\right|^2 = \left|\left|\boldsymbol{B}_1\vec{\mathcal{E}}\right|\right|^2$$

By idempotency and symmetry:

$$\boldsymbol{\mathcal{E}}\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}}$$

Divide both ides by $\sigma^2$:

$$\boxed{\frac{1}{\sigma^2}\left|\left|\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{\beta})\right|\right|^2 = \frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}}}$$

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

Because when we fit a linear model we are trying to estimate the true mean response vector $\boldsymbol{X}\boldsymbol{\beta}$ by using $\boldsymbol{X}\boldsymbol{B}$ and so the difference between the two $\boldsymbol{X}(\boldsymbol{B}-\boldsymbol{\beta})$ is the error in our estimate of the true mean response.

We take the squared norm as we would in the SSE as it makes mathematical sense (leads to closed form solution of best $\boldsymbol{\beta}$ )

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

We saw $\frac{1}{\sigma^2}\vec{\mathcal{E}}\top(\boldsymbol{I}-\boldsymbol{H})\vec{\mathcal{E}} \sim \chi^2_{n-p-1}$. By idempotentcy we can express this as:

$$\frac{1}{\sigma^2}\vec{\mathcal{E}}\top(\boldsymbol{I}-\boldsymbol{H})(\boldsymbol{I}-\boldsymbol{H})\vec{\mathcal{E}}$$

By symmetry:

$$\frac{1}{\sigma^2}\vec{\mathcal{E}}\top(\boldsymbol{I}-\boldsymbol{H})\top(\boldsymbol{I}-\boldsymbol{H})\vec{\mathcal{E}}$$

Which can be expressed as the norm squared

$$\Rightarrow \left|\left|\frac{1}{\sigma^2}(\boldsymbol{I}-\boldsymbol{H})\vec{\mathcal{E}}\right|\right|^2$$

Recall $\vec{Y} = \boldsymbol{X}\vec{\boldsymbol{\beta}} + \vec{\mathcal{E}}$ and therefore $\vec{\mathcal{E}} = \vec{Y} - \boldsymbol{X}\vec{\boldsymbol{\beta}}$. Substitute this information:

$$\Rightarrow \frac{1}{\sigma^2} \left|\left|(\boldsymbol{I} - \boldsymbol{H})(\vec{Y} - \boldsymbol{X}\vec{\boldsymbol{\beta}})\right|\right|^2$$

Distribute:

$$\frac{1}{\sigma^2} \left|\left|(\boldsymbol{I} - \boldsymbol{H})\vec{Y} - (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{X}\vec{\boldsymbol{\beta}})\right|\right|^2$$

Recall $\boldsymbol{H}$ keeps anything in the colsp $[\boldsymbol{X}]$ while $\boldsymbol{I} - \boldsymbol{H}$ kills anything in the colsp $[\boldsymbol{X}]$, therefore, $\boldsymbol{X}\vec{\boldsymbol{\beta}}$ vanishes when multiplied by $(\boldsymbol{I} - \boldsymbol{H})$. Also, because $(\boldsymbol{I} - \boldsymbol{H})\vec{Y}$ is the error, it goes to 0. Therefore we are left with:

$$\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_2 \boldsymbol{\mathcal{E}} = \frac{1}{\sigma^2} ||\boldsymbol{E}||^2$$

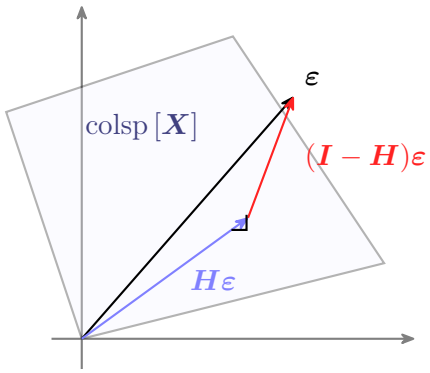
(s) [harder] In what scenarios is $\boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_1 \boldsymbol{\mathcal{E}} > \boldsymbol{\mathcal{E}}^\top \boldsymbol{B}_2 \boldsymbol{\mathcal{E}}$?

Recall, we said $\boldsymbol{B}_1 = \boldsymbol{H}$, the projection onto the colsp $[\boldsymbol{X}]$ and $\boldsymbol{B}_2 = \boldsymbol{I} = \boldsymbol{H}$, the projection of the orthogonal complement of the hat matrix (residual space).

If $\vec{\mathcal{E}}$ lies closer to the colsp $[\boldsymbol{X}]$, then the projection onto that space will be larger and the equality will hold. If $\vec{\mathcal{E}}$ is orthogonal to the colsp $[\boldsymbol{X}]$ then the opposite holds because $\boldsymbol{H}\vec{\mathcal{E}}$ will equal 0.

If the former is the case (equality is true), then that means most of the error lies in the model space and a lot of the error is aligned with the predictors i.e. model is mistaking noise for signal and we might have some overfitting (bad model). If the latter is true, then most of the error is orthogonal to the predictors and so the model is not capturing the error (better model)

(t) [harder] Draw an illustration of $\boldsymbol{\mathcal{E}}$ 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.



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

This is kind of a continuation of part (s). A "good" model has a small projection of the error because if the projection of the error lies in the space which the model is trying to explain, then that means the model is fitting the noise (overfitting). Even if the model has low residuals and appears to fit the data well, this is an indicator that it is actually just capturing random variation as if it is meaningful which will lead to poor generalization to future/new data. i.e. big error projection = bad!

(v) [easy] Find $\mathbb{E}\left[\frac{1}{\sigma^2}\left|\left|\boldsymbol{E}\right|\right|^2\right]$.

Recall we said in 340 that if $X \sim \chi^2_k$ then $\mathbb{E}[X] = k$.
Because $\frac{1}{\sigma^2}\left|\left|\boldsymbol{E}\right|\right|^2 \sim \chi^2_{n-(p+1)}$, $\mathbb{E}\left[\frac{1}{\sigma^2}\left|\left|\boldsymbol{E}\right|\right|^2\right] = n - p - 1$

(w) [easy] Show that $\frac{\left|\left|\boldsymbol{E}\right|\right|^2}{n-(p+1)}$ is an unbiased estimate of $\sigma^2$.

In 342 we saw that the SSE

$$= \vec{e}\top\vec{e} = \left|\left|\vec{e}\right|\right|^2$$

. We proved that the MSE is an unbiased estimate for $\sigma^2$, and the

$$\frac{\left|\left|\boldsymbol{E}\right|\right|^2}{n-(p+1)} = \frac{SSE}{n-(p+1)} = \text{MSE}$$

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

We found via Cochran's theorem that:

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

And we know from our previous work that:

$$\frac{B_j - \boldsymbol{\beta}_j}{\sqrt{\sigma^2(\boldsymbol{X}top\boldsymbol{X})^{-1}_{j,j}}} \sim \mathcal{N}(0,\, 1)$$

We know via 340 that dividing a normal by a chi-squared gives us a student-t distribution, with the same degrees of freedom as the chi-squared we divide by. Thus,

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

(y) [easy] Let $H_0 : \beta_j = 0$. Find the test statistic using the fact from the previous question. Let $s_e$ denote $RMSE := \sqrt{MSE} := \sqrt{SSE/(n-(p+1)} = \sqrt{||e||^2/(n-(p+1)}$.

$$\hat{\hat{t}} = \frac{b_j}{s_e\sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{j,j}}}$$

(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 $\dfrac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{\boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,\,1)$.

$$\Rightarrow \hat{\boldsymbol{Y}} = \vec{X}_*\boldsymbol{B} \sim ?$$

Recall that we previously showed:

$$\boldsymbol{B} \sim \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}},\,\sigma^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$$

And so a linear transformation of this will also be distributed normally:

$$\hat{\boldsymbol{Y}} = \vec{x}_*\boldsymbol{B} \sim \mathcal{N}\left(\vec{x}_*\vec{\boldsymbol{\beta}},\,\sigma^2\vec{x_*}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\vec{x_*}\top\right)$$

Subtracting the mean and dividing by the standard deviation, we get:

$$\frac{\hat{Y}_\star - \mu_\star}{\sigma\sqrt{\boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,\,1)$$

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

We know

$$\frac{\hat{\boldsymbol{Y}}_* - \boldsymbol{x}_*\boldsymbol{B}}{\sigma\sqrt{\boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,\,1)$$

Where $\boldsymbol{x}_*\boldsymbol{B} = \mu_*$. And,

$$\sqrt{\frac{\frac{1}{\sigma^2}\left|\left|\vec{E}\right|\right|^2}{n-p-1}} \sim \chi^2_{n-p-1}$$

And that both of these are independent. Dividing a normal by a chi-squared gives us a student-t and so,

$$\frac{\sqrt{n-(p+1)}(\hat{Y}_\star - \mu_\star)}{||\boldsymbol{E}||\sqrt{\boldsymbol{x}_\star\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{x}_\star^\top}} \sim T_{n-(p+1)}$$

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

$$\hat{\hat{t}} = \frac{\hat{\boldsymbol{y}}_* - \mu_{*_0}}{s_e\sqrt{\boldsymbol{x}_* \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{x}_*^\top}}$$

(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 $\dfrac{\hat{Y}_\star - y_\star}{\sigma\sqrt{1 + \boldsymbol{x}_\star \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,\, 1)$.

$$\boldsymbol{Y}_* - \hat{\boldsymbol{Y}}_* = \boldsymbol{Y}_* - \vec{x}_*\boldsymbol{B} = (\vec{x}_*\boldsymbol{B} + \vec{\mathcal{E}}_*) - (\vec{x}_*\boldsymbol{B})$$
$$\Rightarrow \vec{x}_*\boldsymbol{B} + \vec{\mathcal{E}}_* - \vec{x}_*\boldsymbol{B}$$

Where $\vec{x}_*\boldsymbol{B} + \vec{\mathcal{E}}_* \sim \mathcal{N}(\vec{x}_*\boldsymbol{B},\, \sigma^2)$

So,

$$\boldsymbol{Y}_* \sim \mathcal{N}\left(\vec{x}_*\boldsymbol{B} - \vec{x}_*\boldsymbol{B},\, \sigma^2 + \sigma^2\vec{x}_*(\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1})\vec{x}_*^\top\right)$$

We can subtract the mean and divide by the standard deviation to get the standard normal:

$$\frac{\hat{Y}_\star - y_\star}{\sigma\sqrt{1 + \boldsymbol{x}_\star \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \boldsymbol{x}_\star^\top}} \sim \mathcal{N}(0,\, 1)$$

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

We again have a standard normal divided by a chi squared with n-(p+1) degrees of freedom.

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

$$\hat{\hat{t}} = \frac{37 - \hat{\boldsymbol{y}}_*}{s_e\sqrt{1 + \vec{x}_* \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1} \vec{x}_*^\top}}$$

(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 $\boldsymbol{H}$.

We WTS $\boldsymbol{H} - \boldsymbol{H}_A$ is idempotent via the beginning steps of Gram-Schmidt decomposition.

Recall that $\boldsymbol{X}_A$ is the matrix of the intercept column and the predictors not in the set $S$, while $\boldsymbol{X}_S$ contains the predictors int he set $S$. Recall $\boldsymbol{H} = \boldsymbol{X}\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T$ is the full projection matrix. We have $\boldsymbol{H}_A := \boldsymbol{X}_A(\boldsymbol{X}_A^\top\boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^\top$.

Let us define $P_S := \boldsymbol{H} - \boldsymbol{H}_A$, the projection onto the space in colsp $[\boldsymbol{X}]$, but $\perp$ to colsp $[\boldsymbol{X}_A]$ (It is the extra contribution of $\boldsymbol{X}_S$ after adjsuting for $\boldsymbol{X}_A$

Then we can say

$$P_S = (\boldsymbol{I} - H_A)\boldsymbol{X}_S[X_S^\top(\boldsymbol{I} - \boldsymbol{H}_A)\boldsymbol{X}_S]^{-1}\boldsymbol{X}_S^\top(\boldsymbol{I} - \boldsymbol{H}_A)$$

Here, $(\boldsymbol{I} - H_A)$ is the residual marker for for the subspace colsp $[\boldsymbol{X}_A]$. We are saying take $\boldsymbol{X}_S$ and remove any part that is in the span $\boldsymbol{X}_A$ i.e. perform G-S on $\boldsymbol{X}_S$ relative to $\boldsymbol{X}_A$

Let us define $\boldsymbol{X}_S^\perp := (\boldsymbol{I} - H_A)\boldsymbol{X}_S$

$$\Rightarrow P_S = \boldsymbol{X}_S^\perp[(X_S^\perp)^\top\boldsymbol{X}_S^\perp]^{-1}(\boldsymbol{X}_S^\perp)^\top$$

Square it:

$$P_S^2 = \boldsymbol{X}_S^\perp[(X_S^\perp)^\top\boldsymbol{X}_S^\perp]^{-1}(\boldsymbol{X}_S^\perp)^\top[(X_S^\perp)^\top\boldsymbol{X}_S^\perp]^{-1}(\boldsymbol{X}_S^\perp)^\top$$

Notice $(\boldsymbol{X}_S^\perp)^\top\boldsymbol{X}_S^\perp[(X_S^\perp)^\top\boldsymbol{X}_S^\perp]^{-1} = \boldsymbol{I}$ We now have:

$$P_S^2 = \boldsymbol{X}_S^\perp[(X_S^\perp)^\top\boldsymbol{X}_S^\perp]^{-1}(\boldsymbol{X}_S^\perp)^\top$$

Therefore,

$$(\boldsymbol{H} - \boldsymbol{H}_A)^2 = \boldsymbol{H} - \boldsymbol{H}_A \qquad \square$$

(gg) [easy] Let $\hat{\boldsymbol{Y}}_A := \boldsymbol{H}_A\boldsymbol{Y}$, the orthogonal projection onto colsp $[\boldsymbol{X}_A]$. Prove that $\dfrac{(n-(p+1))\left|\left|\hat{\boldsymbol{Y}} - \hat{\boldsymbol{Y}}_A\right|\right|^2}{k\,||\boldsymbol{E}||^2} \sim F_{k,n-(p+1)}$.

We know the rank of $\boldsymbol{H}_A = p + 1 - k$ and rank of $\boldsymbol{H} - \boldsymbol{H}_A$ is k, so because the ranks total to p+1 we can use cochran's thm;

$$\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{\mathcal{E}} = \frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{H}_A\boldsymbol{\mathcal{E}} + \frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top(\boldsymbol{H} - \boldsymbol{H}_A)\boldsymbol{\mathcal{E}} + \frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\mathcal{E}}$$

Where

$$\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top\boldsymbol{H}_A\boldsymbol{\mathcal{E}} \sim \chi^2_{p+1-k}$$

independent of

$$\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top(\boldsymbol{H} - \boldsymbol{H}_A)\boldsymbol{\mathcal{E}} \sim \chi^2_k$$

and independent of

$$\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\mathcal{E}} \sim \chi^2_{n-(p+1)}$$

By idempotency and symmetry

$$\Rightarrow \frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top(\boldsymbol{H}-\boldsymbol{H}_A)\boldsymbol{\mathcal{E}} = \frac{1}{\sigma^2}\left|\left|\boldsymbol{H}-\boldsymbol{H}_A\boldsymbol{\mathcal{E}}\right|\right|^2$$

Sub for $\boldsymbol{\mathcal{E}}$

$$\frac{1}{\sigma^2}\left|\left|(\boldsymbol{H}-\boldsymbol{H}_A)(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})\right|\right|^2$$

Distribute

$$\frac{1}{\sigma^2}\left|\left|\boldsymbol{H}\boldsymbol{Y}-\boldsymbol{H}_A\boldsymbol{Y}-\boldsymbol{H}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{H}_A\boldsymbol{X}\boldsymbol{\beta}\right|\right|^2$$

We know $\boldsymbol{H}\boldsymbol{Y}=\hat{\boldsymbol{Y}}$ and $\boldsymbol{H}_A\boldsymbol{Y}=\hat{\boldsymbol{Y}}_A$

$$\Rightarrow \frac{1}{\sigma^2}\left|\left|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A-\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{H}_A\boldsymbol{X}\boldsymbol{\beta}\right|\right|^2$$

Augment $\boldsymbol{X}=[\boldsymbol{X}_A \mid \boldsymbol{X}_S]$ and $\boldsymbol{\beta}=[\boldsymbol{\beta}_A \mid \boldsymbol{\beta}_S]^\top$. Now perform block multiplication

$$\Rightarrow \frac{1}{\sigma^2}\left|\left|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A-\boldsymbol{X}_A\boldsymbol{\beta}_A-\boldsymbol{X}_S\boldsymbol{\beta}_S+\boldsymbol{X}_A\boldsymbol{\beta}_A+\boldsymbol{H}_A\boldsymbol{X}_S\boldsymbol{\beta}_S\right|\right|^2$$

$$\Rightarrow \frac{1}{\sigma^2}\left|\left|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A-(\boldsymbol{I}-\boldsymbol{H}_A)\boldsymbol{X}_S\boldsymbol{\beta}_S\right|\right|^2$$

Further,

$$\frac{\frac{\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top(\boldsymbol{H}-\boldsymbol{H}_A)\boldsymbol{\mathcal{E}}}{k}}{\frac{\frac{1}{\sigma^2}\boldsymbol{\mathcal{E}}^\top(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{\mathcal{E}}}{n-p-1}}=\frac{(n-(p+1))\left|\left|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A\right|\right|^2}{k\left|\left|\boldsymbol{E}\right|\right|^2}\sim F_{k,n-(p+1)}$$

Recall: this test tells us if at least one of all the predictors has astatistically significant effect.

(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 $\left|\left|\hat{\boldsymbol{E}}_A\right|\right|^2-\left|\left|\hat{\boldsymbol{E}}\right|\right|^2=\left|\left|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A\right|\right|^2$.

$$\left|\left|\hat{E}_A\right|\right|^2=\left|\left|\hat{E}+(\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A)\right|\right|^2=\hat{E}^\top\hat{E}+2[\hat{E}^\top(\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A)]+(\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A)^\top(\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A)$$

$$\left|\left|\hat{E}_A\right|\right|^2-\left|\left|\hat{E}\right|\right|^2=\left|\left|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A\right|\right|^2+2[\hat{E}^\top(\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A)]$$

Therefore, the equality holds if the dot product of $\hat{E}$ and $(\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A)$ is zero, which it is! This is because $(\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A)\in$ colsp $[\boldsymbol{X}]$ and $\hat{E}\in$ colsp $\left[\boldsymbol{X}^\perp\right]$, making the two vectors orthogonal to each other and so thier dot product is indeed zero. therefore,

$$\left|\left|\hat{\boldsymbol{E}}_A\right|\right|^2-\left|\left|\hat{\boldsymbol{E}}\right|\right|^2=\left|\left|\hat{\boldsymbol{Y}}-\hat{\boldsymbol{Y}}_A\right|\right|^2 \qquad \square$$

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

The test statistic for the partial F-Test is:

$$\frac{\frac{\Delta SSE}{k}}{\frac{SSE}{n-(p+1)}}$$

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

$$\frac{\frac{\Delta SSE}{k}}{\frac{SSE}{n-(p+1)}} = \frac{SSE_A - SSE}{MSE}$$

Recall the T-test

$$\hat{\hat{t}} = \frac{b_j}{s_e\sqrt{\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{j,j}}} \sim T_{n-(p+1)} \Rightarrow \frac{b_j^2}{s_e^2\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}_{j,j}} \sim F_{1,n-(p+1)}$$

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

$$\begin{aligned} F &= \frac{\Delta\text{SSE}}{MSE} \\ &= \frac{\text{SSE}_A - \text{SSE}}{MSE} \\ &= \frac{B_j^2}{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{jj} \cdot MSE} \end{aligned}$$

$$\begin{aligned} \sqrt{F} &= \sqrt{\frac{B_j^2}{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{jj} \cdot MSE}} \\ &= \frac{|B_j|}{\sqrt{(\boldsymbol{X}^\top\boldsymbol{X})^{-1}_{jj}} \cdot \sqrt{MSE}} \\ &= \left|\frac{B_j}{SE(B_j)}\right| \\ &= |t| \end{aligned}$$

(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 $\dfrac{(n-(p+1))\left|\left|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}_n\right|\right|^2}{p\,||\boldsymbol{E}||^2} \sim F_{p,n-(p+1)}$.

Let $\boldsymbol{X}_A = \mathbf{1}$. Find the fitted values under the reduced model:

$$\boldsymbol{H}_A = \boldsymbol{X}_A(\boldsymbol{X}_A^\top \boldsymbol{X}_A)^{-1}\boldsymbol{X}_A^\top$$

Substitute $\boldsymbol{X}_A = \mathbf{1}_n$:

$$\boldsymbol{H}_A = \mathbf{1}_n \left(\mathbf{1}_n^\top \mathbf{1}_n\right)^{-1} \mathbf{1}_n^\top$$

Since:

$$\mathbf{1}_n^\top \mathbf{1}_n = \sum_{i=1}^{n} 1 = n, \quad \text{so} \quad \left(\mathbf{1}_n^\top \mathbf{1}_n\right)^{-1} = \frac{1}{n}$$

Thus:

$$\boldsymbol{H}_A = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top$$

Apply $\boldsymbol{H}_A$ to $\boldsymbol{Y}$:

$$\hat{\boldsymbol{Y}}_A = \boldsymbol{H}_A\boldsymbol{Y} = \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\top\boldsymbol{Y}$$

$$\mathbf{1}_n^\top\boldsymbol{Y} = \sum_{i=1}^{n} y_i = n\bar{y}$$

$$\hat{\boldsymbol{Y}}_A = \frac{1}{n}\mathbf{1}_n(n\bar{y}) = \bar{y}\cdot\mathbf{1}_n$$

Therefore, the fitted values under the reduced mdoel are $\hat{\boldsymbol{Y}}_A = \boldsymbol{H}_A\boldsymbol{Y} = \bar{y}\mathbf{1}$

Plug in:

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

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

We know:

$$\frac{(n-(p+1))\|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}_n\|^2}{p\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2} \sim F_{p,\,n-(p+1)}$$

We can see that $\|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}_n\|^2$ is the sum of square residials due to regression:

$$SSR := \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}_n\|^2$$

Therefore,

$$\frac{\|\hat{\boldsymbol{Y}} - \bar{y}\mathbf{1}_n\|^2}{p} = \frac{SSR}{p} := MSR$$

And we saw earlier that the denominator is the MSE:

$$\frac{\|\boldsymbol{Y} - \hat{\boldsymbol{Y}}\|^2}{n-(p+1)} = \frac{SSE}{n-(p+1)} := MSE$$

So we have,

$$\hat{\hat{F}} = \frac{MSR}{MSE} \sim F_{p,\, n-(p+1)}$$

(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.

From 342, we know that $\vec{b} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$ which came from $\arg\min\left\{||\boldsymbol{y} - \boldsymbol{X}\vec{w}||^2\right\}$. In 343, $\vec{\hat{\theta}} = \vec{b}$ drawn from $\hat{\theta} = \boldsymbol{B}$. Recall 341's use of the method of moments and the maximum liklehood estimate to find estimators. Let us attempt this here; derive $\vec{\hat{\theta}}^{\mathbf{MLE}}$ under core assumption:

$$\boldsymbol{\mathcal{E}} \sim \mathcal{N}_n\left(\vec{0}_n,\ \sigma^2\boldsymbol{I}_n\right) \Rightarrow \boldsymbol{Y} \sim \mathcal{N}_n\left(\boldsymbol{X}\boldsymbol{\beta},\ \sigma^2\boldsymbol{I}_n\right)$$

Use pdf of this and find the likelihood:

$$\Rightarrow \frac{1}{(2\pi)^{n/2}(\sigma^2)^{n/2}} e^{-\frac{1}{2}[(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^\top(\sigma^2\boldsymbol{I}_n)^{-1}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})]}$$

Notice, $-\frac{1}{2}[(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^\top(\sigma^2\boldsymbol{I}_n)^{-1}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})] = -\frac{1}{\sigma^2}\,||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2$

$$\Rightarrow \ell(\boldsymbol{\beta}, \sigma^2; \boldsymbol{y}; \boldsymbol{X}) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln\left(\sigma^2\right)\left[-\frac{1}{\sigma^2}\,||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}||^2\right]$$

Differentiate with respect ot $\boldsymbol{\beta}$

$$\Rightarrow -\frac{1}{2\sigma^2}\frac{\partial}{\partial\boldsymbol{\beta}}\left[\sum_{i=1}^{n}(y_i - \boldsymbol{X}_i\boldsymbol{\beta})^2\right] \overset{\text{set}}{=} 0$$

$$\Rightarrow \vec{\hat{\beta}} = \vec{b} \text{ from OLS} \qquad \square$$

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

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathcal{N}_n(0, \sigma^2 \boldsymbol{I}_n) \Rightarrow \boldsymbol{Y} \sim \mathcal{N}_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$$

$$\ell(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{Y}) = \frac{1}{(2\pi\sigma^2)^{n/2}} e^{(-\frac{1}{2\sigma^2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|^2)}$$

Let $\boldsymbol{b} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{Y}$ be the OLS estimator of $\boldsymbol{\beta}$, and define:

$$SSE := \|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{b}\|^2$$

$$\frac{\partial}{\partial \sigma^2} \ln(\ell) = -\frac{n}{2\sigma^2} + \frac{1}{2(\sigma^2)^2} SSE$$

$$-\frac{n}{2\sigma^2} + \frac{SSE}{2(\sigma^2)^2} = 0 \Rightarrow -n\sigma^2 + SSE \overset{\text{set}}{=} 0$$

$$\Rightarrow \hat{\sigma}^2_{\text{MLE}} = \frac{SSE}{n}$$

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

Recall the bias of an estimator $\hat{\theta}$ for a parameter $\theta$ is:

$$\text{Bias}[\theta] = \mathbb{E}\left[\hat{\theta}\right] - \theta$$

In this case it is:

$$\text{Bias}\left[\sigma^2_{\text{MLE}}\right] = \mathbb{E}\left[\sigma^2_{\text{MLE}}\right] - \sigma^2$$

Recall from the previous problem,

$$\hat{\sigma}^2_{\text{MLE}} = \frac{SSE}{n}$$

Therefor,

$$\mathbb{E}\left[\sigma^2_{\text{MLE}}\right] = \mathbb{E}\left[\frac{SSE}{n}\right] = \frac{1}{n}\mathbb{E}[SSE]$$

But we know

$$\frac{1}{n}\mathbb{E}[SSE] = \frac{n - (p+1)}{n}\sigma^2$$

So,

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

$$\Rightarrow \frac{n\sigma^2 - p\sigma^2 - \sigma^2}{n} - \frac{n\sigma^2}{n}$$

$$\Rightarrow \text{Bias}\left[\sigma^2_{\text{MLE}}\right] = -\frac{(p+1)\sigma^2}{n}$$

The MLE does not account for the fact that we estimated p+1 parameters from the data. Therefore, it systematically underestimates the variance and the unbiased estimator corrects this by dividing n - (p+1) instead of just n.

## 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.

$$f(\vec{\boldsymbol{\beta}}) = \mathcal{N}_{p+1}\left(\vec{0}_{p+1},\, \tau^2 \boldsymbol{I}_{p+1}\right)$$
$$f(\sigma^2) = \frac{1}{\sigma^2}$$

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

$$f(\vec{\boldsymbol{\beta}}; \sigma^2 \mid \vec{y}, X) \propto f(\vec{y} \mid \vec{\boldsymbol{\beta}}, \sigma^2, X) f(\vec{\boldsymbol{\beta}}) f(\sigma^2)$$
$$\propto (\sigma^2)^{-\frac{n}{2}-1} e^{-1/2\sigma^2 \left|\left|\vec{y}-\boldsymbol{X}\vec{\boldsymbol{\beta}}\right|\right|^2} e^{-1/2\tau^2 \left|\left|\vec{\boldsymbol{\beta}}\right|\right|^2}$$

This is not a known distribution. But if we want just point estimation, we don't need the distribution, we can simply just use the MAP:

$$\hat{\vec{\boldsymbol{\beta}}}^{\mathbf{MAP}} = \arg\max \left\{ (-n/2-1) \ln\left(\sigma^2\right) - \frac{1}{2\sigma^2} \left|\left|\vec{y} - \boldsymbol{X}\vec{\boldsymbol{\beta}}\right|\right|^2 - \frac{1}{2\tau^2} \left|\left|\vec{\boldsymbol{\beta}}\right|\right|^2 \right\}$$

End up with:

$$\Rightarrow \arg\min \left\{ SSE + \lambda \left|\left|\vec{\boldsymbol{\beta}}\right|\right|^2 \right\}$$

Where $\lambda = \frac{\sigma^2}{\tau^2}$

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

We use a Laplace prior:

$$f(\vec{\boldsymbol{\beta}}) = \frac{1}{2\tau^2} e^{\frac{|\beta|}{\tau^2}}$$

And jeffries for variance:

$$f(\sigma^2) = \frac{1}{\sigma^2}$$

(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.

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

Again, we do not know this distribution, so we will find the MAP estimate for point estimation:

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

Which becomes:

$$\arg\min \left\{ SSE + \lambda \sum_{j=0}^{p} |\beta_j \right\}$$

Where $\lambda = \frac{2\sigma^2}{\tau^2}$. There is no closed form solution andw e need an optimizer.

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

$$\vec{0}_{p+1}$$

In both cases, ridge and lasso, the prior mean is zero so the MAP estimate is pulled toward zero. When $\lambda = 0$ it becomes the OLS estimate and when $\lambda$ goes to infinity the $\arg\min \{||\,||^2\}\vec{\boldsymbol{\beta}} \to \vec{0}_{p+1}$ in Ridge regression.

Recall, the shrinkage estimator is used to pull the estimated parameters toward a target value to reduce variance at the risk of introducing bias, which can be useful if (a) p is very large or (b) data is noisy. We did this initally because we want to regress p = 20,000 genes but we only have n = 300 people.

Ridge regression shrinks all coefficients towards zero, but keeps them nonzero - does not perform variable selection i.e. keeps all predictors.

Lasso regression shrinks some coefficients exactly to zero, it does both shrinkage and variable selection which can be helpful when you believe only a small number of predictors matter.

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

Variable selection is the process of deciding which predictors to include in the model before fitting the model. This is important because if there is a large number of predictors, they may not all be important and the inclusion of unnecessary variables can create overfitting in your model.

(g) [easy] Describe why Lasso estimation has the added bonus of being able to perform variable selection and ridge does

Lasso has an $\ell_1$ penalty that is not differentiable at zero, so this encourages exactly zero coefficients and will force some of the weights for certain features to zero. Ridge has an $\ell_2$ penalty which is smooth and differentiable everywhere so it will shirnk all the coefficients continuously towards zero, but they will never equal zero unless $\lambda = \infty$ (which is impossible).

## 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 $\boldsymbol{\beta}$?

Perform bootstrap! Sample n rows from $\mathbb{D}$ with replacement to get the bootstrap sample (a bunch of tuples), and the find their OLD estimate.Then, for the j-th estimate, plot the bootstrap distribution.

Ask, does $0 \in \left[Q[b_{b_j}, \frac{\alpha}{2}], Q[b_{b_j}, 1 - \frac{\alpha}{2}]\right]$

(b) [easy] If we know that the errors $\mathcal{E}_1, \ldots, \mathcal{E}_n$ are iid with expectation zero and variance $\sigma^2$ 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 $\boldsymbol{B}$?

$$\boldsymbol{B} \dot{\sim} \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}}, \sigma^2 \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right)$$

(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?

$$\boldsymbol{B} \dot{\sim} \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}}, \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T \hat{\hat{D}} X \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1}\right)$$

Where residuals are in place of variance and

$$\hat{\hat{D}} = \begin{bmatrix} e_1^2 & & & \\ & e_2^2 & & \\ & & \ddots & \\ & & & e_n^2 \end{bmatrix}$$

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

same as previous, except even more approximate

$$\boldsymbol{B} \dot{\sim} \mathcal{N}_{p+1}\left(\vec{\boldsymbol{\beta}},\ \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\hat{\hat{D}}X\left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\right)$$

Where residuals are in place of variance and

$$\hat{\hat{D}} = \begin{bmatrix} e_1^2 & & & \\ & e_2^2 & & \\ & & \ddots & \\ & & & e_n^2 \end{bmatrix}$$

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

No, because the F-tests derived from the core assumption rely on homoscedacity, normality of error, and IID

## 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.

$$\Pi_{i=1}^n \left(e^{-e^{-x_i\beta}}\right)^{Y_i}\left(1 - e^{-e^{-x_i\beta}}\right)^{1-Y_i}$$

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

$$\hat{\hat{\mathbf{LR}}} = \frac{\Pi_{i=1}^n \left(e^{-e^{-x_i\beta}}\right)^{Y_i}\left(1 - e^{-e^{-x_i\beta}}\right)^{1-Y_i}}{\Pi_{i=1}^n \left(e^{-e^{-x_i\tilde{b_0}}}\right)^{Y_i}\left(1 - e^{-e^{-x_i\tilde{b_0}}}\right)^{1-Y_i}}$$

(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 $H_0 : \beta_2 = \beta_3 = 0$.

$$\hat{\hat{\mathbf{LR}}} = \frac{\Pi_{i=1}^n e^{-e^{-x_i\beta}}e^{x_i\beta Y_i}}{\Pi_{i=1}^n e^{-e^{-x_i\tilde{\tilde{b}}}}e^{x_i\tilde{\tilde{b}}Y_i}}$$

Where $\tilde{\tilde{b}} = \arg\max\left\{\sum_{i=1}^n -e^{\vec{x}\vec{w}} + \vec{x}\vec{w}y_i | w_2 = w_3 = 0\right\}$

(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 $H_0 : \beta_2 = \beta_3 = 0$.

$$\Pi_{i:c_i=1} \frac{s}{e^{\vec{x_i}\vec{w}}} \left(\frac{y_i}{e^{\vec{x_i}\vec{w}}}\right)^{s-1} e^{-\left(\frac{y_i}{e^{\vec{x_i}\vec{w}}}\right)^s} \Pi_{i:c_i=0} e^{-\left(\frac{y_i}{e^{\vec{x_i}\vec{w}}}\right)^s}$$

(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 $H_0 : \beta_1 = \beta_2 = \ldots = \beta_p = 0$.