24500 HW7

Bin Yu

Feb 27, 2025

Question 1

(a)

Recall the fact that

$$\frac{\|y - X\hat{\beta}\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Thus,

$$\hat{\sigma}^2 = \frac{\|y - X\hat{\beta}\|^2}{n-p} = \sigma^2 \, \frac{\chi^2_{n-p}}{n-p}.$$

$$E\big(\hat{\sigma}^2\big) \;=\; E\Big(\sigma^2 \cdot \frac{\chi^2_{n-p}}{n-p}\Big) \;=\; \sigma^2 \, \frac{E(\chi^2_{n-p})}{n-p}.$$

Since for χ^2_{n-p} we have

$$E\big(\chi^2_{n-p}\big) = n - p,$$

it follows that

$$E\big(\hat{\sigma}^2\big) \;=\; \sigma^2 \, \frac{n-p}{n-p} \;=\; \sigma^2.$$

Hence, $\hat{\sigma}^2$ is an unbiased estimator of σ^2.

$$\mathrm{Var}\big(\hat{\sigma}^2\big) \;=\; \mathrm{Var}\Big(\sigma^2 \cdot \tfrac{\chi^2_{n-p}}{n-p}\Big) \;=\; \sigma^4 \, \mathrm{Var}\Big(\tfrac{\chi^2_{n-p}}{n-p}\Big).$$

We know

$$\mathrm{Var}\big(\chi^2_{n-p}\big) \;=\; 2\,(n-p).$$

Hence,

$$\mathrm{Var}\Big(\tfrac{\chi^2_{n-p}}{n-p}\Big) = \frac{1}{(n-p)^2}\,\mathrm{Var}(\chi^2_{n-p}) = \frac{2(n-p)}{(n-p)^2} = \frac{2}{n-p}.$$

Therefore,

$$\mathrm{Var}\big(\hat{\sigma}^2\big) = \sigma^4 \, \frac{2}{n-p} = \frac{2\,\sigma^4}{n-p}.$$

A chi-square random variable χ^2_{n-p} can be viewed as the sum of squares of $n-p$ i.i.d. standard normal variables $Z_i \sim N(0,1)$:

$$\chi^2_{n-p} = \sum_{i=1}^{n-p} Z_i^2.$$

For large $n-p$, the distribution of this sum can be approximated by a normal distribution via the Central Limit Theorem, we have:

$$Y = \chi^2_{n-p}, \quad \mu = n-p, \quad \text{and} \quad \sigma_Y^2 = 2(n-p),$$

then

$$\frac{Y-\mu}{\sqrt{2(n-p)}} = \frac{\chi^2_{n-p}-(n-p)}{\sqrt{2(n-p)}} \quad \xrightarrow{d} \quad N(0,1) \quad \text{as } (n-p)\to\infty.$$

Since

$$\hat{\sigma}^2 - \sigma^2 = \sigma^2\left(\tfrac{\chi^2_{n-p}}{n-p} - 1\right),$$

we define

$$Z \;=\; \frac{\chi^2_{n-p}-(n-p)}{\sqrt{2(n-p)}} \;\xrightarrow{d}\; N(0,1).$$

Note that

$$\frac{\chi^2_{n-p}}{n-p} - 1 = \frac{\chi^2_{n-p}-(n-p)}{n-p} = \frac{Z}{\sqrt{n-p}}\sqrt{2}.$$

Thus,

$$\hat{\sigma}^2 - \sigma^2 = \sigma^2\left(\frac{\chi^2_{n-p}}{n-p} - 1\right) = \sigma^2\,\frac{Z}{\sqrt{n-p}}\sqrt{2}.$$

Multiplying both sides by $\sqrt{n-p}$ yields

$$\sqrt{n-p}\,(\hat{\sigma}^2-\sigma^2) = \sigma^2\cdot Z\,\sqrt{2}.$$

As Z converges in distribution to $N(0,1)$, the right-hand side converges to a normal random variable with mean 0 and variance $2\,\sigma^4$. Hence,

$$\sqrt{n-p}\,(\hat{\sigma}^2-\sigma^2) \quad \xrightarrow{d} \quad N(0,\ 2\,\sigma^4).$$

the asymptotic distribution is

$$\sqrt{n-p}\,(\hat{\sigma}^2-\sigma^2) \quad \xrightarrow{d} \quad N\big(0,\ 2\,\sigma^4\big)$$

(b)

We use the Delta method. From part (a), if

$$\sqrt{n-p}\,(\hat{\sigma}^2-\sigma^2) \xrightarrow{d} N(0,\ 2\sigma^4),$$

then for a differentiable g,

$$\sqrt{n-p}\,\big(g(\hat{\sigma}^2)-g(\sigma^2)\big) \xrightarrow{d} N\Big(0,\ \big[g'(\sigma^2)\big]^2\cdot 2\,\sigma^4\Big).$$

We want the limiting variance to be 1, so we need

$$\big[g'(\sigma^2)\big]^2\,2\,\sigma^4 = 1$$

$$g'(\sigma^2) = \frac{1}{\sqrt{2}\,\sigma^2}.$$

Integrating,

$$g(x) = \int \frac{1}{\sqrt{2}\,x}\,dx = \frac{1}{\sqrt{2}}\log(x) + C.$$

Ignore the constant C for inference purposes, so a choice is

$$g(x) = \frac{1}{\sqrt{2}}\,\log(x).$$

Then

$$\sqrt{n-p}\,\big(g(\hat{\sigma}^2)-g(\sigma^2)\big) \xrightarrow{d} N(0,1).$$

(c)

We have found that a convenient transform for $\hat{\sigma}^2$ is

$$g(x) \;=\; \frac{1}{\sqrt{2}}\log(x).$$

Under suitable regularity conditions (and for large $n-p$), one obtains

$$\sqrt{n-p}\,\big(g(\hat{\sigma}^2)-g(\sigma^2)\big) \;=\; \sqrt{n-p}\,\Big(\tfrac{1}{\sqrt{2}}\log(\hat{\sigma}^2) \;-\; \tfrac{1}{\sqrt{2}}\log(\sigma^2)\Big) \;\xrightarrow{d}\; N(0,1).$$

$$\frac{\sqrt{n-p}}{\sqrt{2}}\,\big(\log(\hat{\sigma}^2)-\log(\sigma^2)\big) \;\xrightarrow{d}\; N(0,1).$$

To construct $(1-\alpha)$ confidence interval, begin from the asymptotic result:

$$P\Big(-z_{1-\alpha/2} \;\le\; \frac{\sqrt{n-p}}{\sqrt{2}}\,\big[\log(\hat{\sigma}^2) \;-\; \log(\sigma^2)\big] \;\le\; z_{1-\alpha/2}\Big) \;=\; 1-\alpha.$$

Inside the probability:

$$-z_{1-\alpha/2} \;\le\; \frac{\sqrt{n-p}}{\sqrt{2}}\,\big[\log(\hat{\sigma}^2)-\log(\sigma^2)\big] \;\le\; z_{1-\alpha/2}$$

is equivalent to

$$\log(\hat{\sigma}^2) \;-\; z_{1-\alpha/2}\sqrt{\frac{2}{n-p}} \;\le\; \log(\sigma^2) \;\le\; \log(\hat{\sigma}^2) \;+\; z_{1-\alpha/2}\sqrt{\frac{2}{n-p}}.$$

Exponentiating each part, we obtain

$$\hat{\sigma}^2\,\exp\Big(-z_{1-\alpha/2}\sqrt{\tfrac{2}{n-p}}\Big) \;\le\; \sigma^2 \;\le\; \hat{\sigma}^2\,\exp\Big(z_{1-\alpha/2}\sqrt{\tfrac{2}{n-p}}\Big).$$

Hence, we have the corresponding $(1-\alpha)$ confidence interval for σ^2:

$$\Big[\,\hat{\sigma}^2\,\exp\Big(-z_{1-\alpha/2}\sqrt{\tfrac{2}{n-p}}\Big),\quad \hat{\sigma}^2\,\exp\Big(z_{1-\alpha/2}\sqrt{\tfrac{2}{n-p}}\Big)\Big]$$

Question 2

(a)

We want to solve

$$\hat{\beta}_0 \;=\; \arg\min_{\beta_0}\sum_{i=1}^{n}\big(y_i-\beta_0\big)^2.$$

Define the objective function:

$$Q(\beta_0) \;=\; \sum_{i=1}^{n}(y_i-\beta_0)^2.$$

To find the minimizer, we take the derivative of $Q(\beta_0)$ with respect to β_0 and set it to zero:

$$\frac{\partial}{\partial\beta_0}\sum_{i=1}^{n}(y_i-\beta_0)^2 \;=\; \sum_{i=1}^{n}\frac{\partial}{\partial\beta_0}\big(y_i-\beta_0\big)^2 \;=\; \sum_{i=1}^{n}\big(-2\big)\big(y_i-\beta_0\big).$$

Hence,

$$0 \;=\; \sum_{i=1}^{n}\big(y_i-\beta_0\big) \;=\; \sum_{i=1}^{n}y_i \;-\; n\,\beta_0.$$

$$\hat{\beta}_0 \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i \;=\; \bar{y},$$

the sample mean of the observations.

Because each y_i is distributed as $N(\beta_0, \sigma^2)$ and they are i.i.d., we know

$$\bar{y} \;=\; \hat{\beta}_0 \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i$$

is also normally distributed with

$$E(\hat{\beta}_0) = \beta_0, \quad \mathrm{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n}.$$

Hence,

$$\hat{\beta}_0 \;\sim\; N\Big(\beta_0,\; \frac{\sigma^2}{n}\Big).$$

$$E\big[(\hat{\beta}_0 - \beta_0)^2\big] \;=\; \mathrm{Var}(\hat{\beta}_0) + E^2[(\hat{\beta}_0 - \beta_0)] \;=\; \frac{\sigma^2}{n} + 0 = \frac{\sigma^2}{n}$$

Thus,

$$E\big[(\hat{\beta}_0 - \beta_0)^2\big] \;=\; \frac{\sigma^2}{n}.$$

(b)

To find the minimizer, we take the derivative of $Q(\beta_0)$ with respect to β_0 and set it to zero:

$$\frac{\partial}{\partial \beta_0} Q(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} 2\left(y_i - \beta_0 - \beta_1 x_i\right)(-1) \;=\; -2\sum_{i=1}^{n}\big(y_i - \beta_0 - \beta_1 x_i\big).$$

Setting this to zero:

$$0 \;=\; \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_i) \;=\; \sum_{i=1}^{n} y_i \;-\; n\,\beta_0 \;-\; \beta_1 \sum_{i=1}^{n} x_i.$$

Hence,

$$\sum_{i=1}^{n} y_i \;=\; n\,\beta_0 \;+\; \beta_1 \sum_{i=1}^{n} x_i. \tag{1}$$

For β_1:

$$\frac{\partial}{\partial \beta_1} Q(\beta_0, \beta_1) \;=\; \sum_{i=1}^{n} 2\left(y_i - \beta_0 - \beta_1 x_i\right)(-x_i) \;=\; -2\sum_{i=1}^{n} x_i\big(y_i - \beta_0 - \beta_1 x_i\big).$$

Setting this to zero:

$$0 \;=\; \sum_{i=1}^{n} x_i\left(y_i - \beta_0 - \beta_1 x_i\right) \;=\; \sum_{i=1}^{n} x_i y_i \;-\; \beta_0 \sum_{i=1}^{n} x_i \;-\; \beta_1 \sum_{i=1}^{n} x_i^2.$$

Thus,

$$\sum_{i=1}^{n} x_i y_i \;=\; \beta_0 \sum_{i=1}^{n} x_i \;+\; \beta_1 \sum_{i=1}^{n} x_i^2. \tag{2}$$

Solving equations (1) and (2) gives:

$$\hat{\beta}_1 \;=\; \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 \;=\; \bar{y} \;-\; \hat{\beta}_1\,\bar{x},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$.

In the standard setting of simple linear regression we learned from class:

$$y_i \;=\; \beta_0 \;+\; \beta_1 x_i \;+\; \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0,\sigma^2), \quad i = 1,\ldots,n,$$

with $x_1,\ldots,x_n$ treated as fixed, the well-known OLS result states that:

$$\mathrm{Var}(\hat{\beta}_0) \;=\; \sigma^2\left[\frac{1}{n} \;+\; \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right].$$

Even if the true slope $\beta_1 = 0$, the same formula for $\mathrm{Var}(\hat{\beta}_0)$ remains valid.

Therefore,

$$\mathrm{Var}(\hat{\beta}_0) \;=\; \sigma^2\left[\frac{1}{n} \;+\; \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right].$$

Now calculate $E[\hat{\beta}_1]$. By linearity of expectation,

$$\hat{\beta}_1 \;=\; \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

We examine its expectation. Note that

$$E[\hat{\beta}_1] \;=\; \frac{\sum_{i=1}^{n}(x_i-\bar{x})\,E[\,y_i-\bar{y}\,]}{\sum_{i=1}^{n}(x_i-\bar{x})^2}.$$

Since under the linear model, $E[\,y_i\,] = \beta_0$ and

$$\bar{y} \;=\; \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$E[\bar{y}] \;=\; \beta_0,$$

we have

$$E[\,y_i - \bar{y}\,] = \beta_0 - E[\bar{y}] = \beta_0 - \beta_0 = 0.$$

In this case, $E[\hat{\beta}_1] = 0$, and so $\hat{\beta}_1$ is unbiased for the true slope (which is zero).

Next, consider $\hat{\beta}_0$. Recall that

$$\hat{\beta}_0 \;=\; \bar{y} \;-\; \hat{\beta}_1\,\bar{x}.$$

Taking expectation,

$$E[\hat{\beta}_0] \;=\; E[\bar{y}] \;-\; E[\hat{\beta}_1]\,\bar{x} \;=\; \beta_0 - 0 = \beta_0$$

Hence, $\hat{\beta}_0$ is also an unbiased estimator of β_0.

Finally, the mean squared error (MSE) of $\hat{\beta}_0$ is simply its variance (since it is unbiased):

$$E\big[(\hat{\beta}_0-\beta_0)^2\big] \;=\; \mathrm{Var}(\hat{\beta}_0) + E^2[(\hat{\beta}_0-\beta_0)] = \mathrm{Var}(\hat{\beta}_0) = \mathrm{Var}(\hat{\beta}_0) \;=\; \sigma^2\left[\frac{1}{n} \;+\; \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i-\bar{x})^2}\right].$$

(c)

In part (a), we estimate β_0 by minimizing

$$\sum_{i=1}^{n}(y_i-\beta_0)^2,$$

which leads to the estimator

$$\hat{\beta}_0^{(a)} = \bar{y}.$$

Because $y_i \sim N(\beta_0,\ \sigma^2)$ are i.i.d., $\bar{y}$ is unbiased for β_0 with

$$\mathrm{Var}\big(\hat{\beta}_0^{(a)}\big) \ = \ \frac{\sigma^2}{n}.$$

Hence,

$$E\big[(\hat{\beta}_0^{(a)} - \beta_0)^2\big] \ = \ \frac{\sigma^2}{n}.$$

In part (b), we include a slope parameter β_1 in the model

$$y_i \ = \ \beta_0 + \beta_1\, x_i,$$

and estimate both β_0 and β_1 by ordinary least squares (OLS), even though the true slope is 0. The resulting estimator for the intercept is

$$\hat{\beta}_0^{(b)} \ = \ \bar{y} \ - \ \hat{\beta}_1\, \bar{x}.$$

From standard regression theory, its variance is given by

$$\mathrm{Var}\big(\hat{\beta}_0^{(b)}\big) \ = \ \sigma^2 \left[\frac{1}{n} \ + \ \frac{\bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right].$$

Thus, in this scenario,

$$E\big[(\hat{\beta}_0^{(b)} - \beta_0)^2\big] \ = \ \frac{\sigma^2}{n} \ + \ \frac{\sigma^2\, \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.$$

Therefore, we observe that:

- **Both are correct in expectation.** In each method, $\hat{\beta}_0$ is an unbiased estimator of β_0, so both approaches give valid estimates on average.
- **Different variance (accuracy).** Approach (a) yields $\mathrm{Var}(\hat{\beta}_0^{(a)}) = \frac{\sigma^2}{n}$. Approach (b) inflates the variance by an extra term
$$\frac{\sigma^2\, \bar{x}^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
because it also estimates a slope parameter that, in truth, does not exist ($\beta_1 = 0$).
- **implication.** Therefore, fitting a more complicated model (b) when the simpler model is actually correct will not improve (and can worsen) the precision of estimating β_0. The extra parameter β_1 adds uncertainty to the intercept estimate.

Hence, while both approaches are *correct* in the sense of unbiasedness, method (a) has smaller variance for estimating β_0 in this particular data-generating scenario.

Question 3

(a)

We estimate β_0 by solving

$$\hat{\beta}_0 \ = \ \arg\min_{\beta_0} \sum_{i=1}^{n} (\, y_i - \beta_0\,)^2.$$

From earlier derivations (as in Question 2(a)), we know:

$$\hat{\beta}_0 \;=\; \bar{y} \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(\beta_0 + \beta_1 x_i + \varepsilon_i\big) \;=\; \beta_0 \;+\; \beta_1\,\bar{x} \;+\; \bar{\varepsilon},$$

where $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ and $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n} \varepsilon_i$.

Since we have:

$$E[\,y_i\,] \;=\; \beta_0 + \beta_1\,x_i.$$

Taking $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$, we observe

$$E\big(\hat{\beta}_0\big) \;=\; \frac{1}{n}\sum_{i=1}^{n} E[\,y_i\,] \;=\; \frac{1}{n}\sum_{i=1}^{n}\big(\beta_0 + \beta_1 x_i\big) \;=\; \beta_0 + \beta_1\,\bar{x},$$

which means the estimator is biased for β_0 unless $\bar{x} = 0$.

For the variance, since y_i are i.i.d. with variance σ^2,

$$\mathrm{Var}\big(\hat{\beta}_0\big) \;=\; \mathrm{Var}\Big(\tfrac{1}{n}\sum_{i=1}^{n} y_i\Big) \;=\; \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{Var}(y_i) \;=\; \frac{n\,\sigma^2}{n^2} \;=\; \frac{\sigma^2}{n}.$$

Because

$$\mathrm{Bias}(\hat{\beta}_0) \;=\; E\big[\hat{\beta}_0\big] - \beta_0 \;=\; \beta_1\,\bar{x},$$

the MSE is

$$E\big[(\hat{\beta}_0 - \beta_0)^2\big] \;=\; \mathrm{Var}\big(\hat{\beta}_0\big) \;+\; \big[\mathrm{Bias}(\hat{\beta}_0)\big]^2 \;=\; \frac{\sigma^2}{n} \;+\; \beta_1^2\,\bar{x}^2.$$

(b)

Now we include the slope term and solve:

$$(\hat{\beta}_0,\ \hat{\beta}_1) \;=\; \arg\min_{\beta_0,\,\beta_1}\sum_{i=1}^{n}\big(y_i - \beta_0 - \beta_1 x_i\big)^2.$$

From OLS theory (see Question 2(b)), we get

$$\hat{\beta}_1 \;=\; \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \hat{\beta}_0 \;=\; \bar{y} \;-\; \hat{\beta}_1\,\bar{x}.$$

From the standard OLS estimates we have known that:

$$E[\hat{\beta}_0] \;=\; \beta_0,$$

and the variance is

$$\mathrm{Var}\big(\hat{\beta}_0\big) \;=\; \sigma^2\left[\frac{1}{n} \;+\; \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].$$

Therefore,

$$E\big[(\hat{\beta}_0 - \beta_0)^2\big] \;=\; \sigma^2\left[\frac{1}{n} \;+\; \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].$$

which is the same as the result in Question 2(b).

(c)

Suppose the true data-generating process is

$$y_i \;=\; \beta_0 + \beta_1 x_i + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad \beta_1 \neq 0.$$

However, if we fit the simpler model from part (a), we are assuming:

$$y_i \;\approx\; \beta_0 \quad \text{(no slope term)},$$

which ignores the covariates $\{x_i\}$. Our results show:

If $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \neq 0$, then the simpler model's estimator

$$\hat{\beta}_0^{(\text{mean})} \;=\; \bar{y}$$

will be a biased estimator of β_0. Indeed,

$$\text{Bias}(\hat{\beta}_0^{(\text{mean})}) \;=\; \beta_1\,\bar{x}.$$

This arises because the slope term $\beta_1 x_i$ is ignored, shifting it away from the true β_0.

By contrast, part (b) incorporates the slope parameter β_1, matching the true data distribution. The estimator $\hat{\beta}_0$ in the full regression model is unbiased for β_0, and its MSE depends only on its variance.

Also, the mean squared error (MSE) of the simpler estimator is

$$\text{MSE}_{\text{simple}} = \frac{\sigma^2}{n} + \beta_1^2\,\bar{x}^2.$$

and the full model's MSE is

$$\text{MSE}_{\text{full}} = \sigma^2\left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right].$$

To compare the two, consider the difference:

$$\begin{aligned}
\text{MSE}_{\text{simple}} - \text{MSE}_{\text{full}} &= \left[\frac{\sigma^2}{n} + \beta_1^2\,\bar{x}^2\right] - \left[\frac{\sigma^2}{n} + \frac{\sigma^2\,\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right] \\
&= \beta_1^2\,\bar{x}^2 - \frac{\sigma^2\,\bar{x}^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \\
&= \bar{x}^2\left(\beta_1^2 - \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}\right).
\end{aligned}$$

Thus, the simpler model yields a higher MSE than the full model (i.e., $\text{MSE}_{\text{simple}} > \text{MSE}_{\text{full}}$) if and only if

$$\beta_1^2 > \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Equivalently, if

$$|\beta_1| > \frac{\sigma}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

Therefore, when the magnitude of the slope β_1 exceeds this threshold, the bias incurred by omitting the slope leads to a larger MSE for the simpler model compared to the full model.

Question 4

(a)

Let us denote by S^c the complement of S in $\{0, 1, 2, \ldots, p-1\}$. Hence, β_{S^c} is the vector of parameters β_j with $j \in S^c$, and X_{S^c} is the corresponding submatrix of X consisting of those columns indexed by S^c.

Under the null hypothesis $H_0 : \beta_S = 0$, the regression model only involves the covariates in S^c, i.e.,

$$y \sim N(X_{S^c}\beta_{S^c}, \sigma^2 I_n)$$

Imposing the restriction $\beta_S = 0$ yields the least squares estimator (LSE) for the remaining parameters:

$$\hat{\beta}_{H_0} \;=\; \left(X_{S^c}^\top X_{S^c}\right)^{-1} X_{S^c}^\top\, y.$$

where,

$$X_{S^c} \;=\; \begin{pmatrix} | & & | \\ x_{:,j_1} & \cdots & x_{:,j_{p-s}} \\ | & & | \end{pmatrix},$$

with $j_1, \ldots, j_{p-s} \in S^c$.

We can compute the mean and variance of $\hat{\beta}_{H_0}$ by noting that $\hat{\beta}_{H_0}$ is a linear function of y. Specifically,

$$\hat{\beta}_{H_0} \;=\; (X_{S^c}^\top X_{S^c})^{-1} X_{S^c} y,$$

$$E[\hat{\beta}_{H_0}] = E[(X_{S^c}^\top X_{S^c})^{-1} X_{S^c}^\top\, y] = (X_{S^c}^\top X_{S^c})^{-1} X_{S^c}^\top\, E[y] = (X_{S^c}^\top X_{S^c})^{-1} X_{S^c}^\top\, X_{S^c}\, \beta_{S^c} = \beta_{S^c}.$$

Hence, $\hat{\beta}_{H_0}$ is unbiased for β_{S^c} under H_0.

Since y has covariance matrix $\sigma^2 I_n$ under the model, let $A = (X_{S^c}^\top X_{S^c})^{-1} X_{S^c}^\top$, then $\hat{\beta}_{H_0} = Ay$

$$\mathrm{Var}(\hat{\beta}_{H_0}) \;=\; A\mathrm{Var}(y)\, A^\top \;=\; A\, \sigma^2 I_n\, A^\top \;=\; \sigma^2\, A\, A^\top.$$

$$A\, A^\top = (X_{S^c}^\top X_{S^c})^{-1}\, X_{S^c}^\top\; X_{S^c}\, \left[(X_{S^c}^\top X_{S^c})^{-1}\right]^\top = (X_{S^c}^\top X_{S^c})^{-1},$$

using the fact that $(X_{S^c}^\top X_{S^c})$ is symmetric and invertible. Hence

$$\mathrm{Var}(\hat{\beta}_{H_0}) = \sigma^2 \left(X_{S^c}^\top X_{S^c}\right)^{-1}.$$

Because $\hat{\beta}_{H_0}$ is a linear transformation of the Gaussian vector y, $\hat{\beta}_{H_0}$ itself is normally distributed. Putting it all together,

$$\hat{\beta}_{H_0} \;\sim\; N\Big(\, \beta_{S^c},\; \sigma^2\, (X_{S^c}^\top X_{S^c})^{-1}\Big).$$

(b)

We can write

$$y - y_{H_0} \;=\; y \;-\; \hat{y} \;+\; \hat{y} - y_{H_0}.$$

$$y - y_{H_0} \;=\; (y - \hat{y}) \;+\; (\hat{y} - y_{H_0}).$$

$$\|\, y - y_{H_0}\|^2 \;=\; \left\|\, y - \hat{y} + \hat{y} - y_{H_0}\right\|^2 \;=\; \|\, y - \hat{y}\|^2 \;+\; \|\hat{y} - y_{H_0}\|^2 + 2(y - \hat{y})^\top(\hat{y} - y_{H_0})$$

For the cross product term $(y - \hat{y})^\top(\hat{y} - y_{H_0})$: section*Algebraic Proof of $(y - \hat{y})^\top(\hat{y} - y_{H_0}) = 0$ Using Projection Matrices

Let

$$H \;=\; X\,(X^\top X)^{-1} X^\top.$$

We claim H is the projection matrix onto the column space of X. To verify:

- H is symmetric:

$$H^\top = \left[X\,(X^\top X)^{-1}X^\top\right]^\top = X\left[(X^\top X)^{-1}\right]^\top X^\top = X\,(X^\top X)^{-1}X^\top = H,$$

since $(X^\top X)^{-1}$ is symmetric.

- H is idempotent:

$$H^2 = \left[X\,(X^\top X)^{-1}X^\top\right]\left[X\,(X^\top X)^{-1}X^\top\right] = X\,(X^\top X)^{-1}X^\top X(X^\top X)^{-1}X^\top = X\,(X^\top X)^{-1}X^\top = H.$$

Hence H is projection matrix that projects onto $\mathcal{C}(X)$, the column space of X.

Let S^c be the complement of S in $\{0, 1, \ldots, p-1\}$, and let $X_{S^c} \in R^{n\times(p-s)}$ be the submatrix of X containing the columns indexed by S^c. Define

$$H_c \;=\; X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top.$$

Similarly, one checks that:

$$H_c^\top = H_c, \quad H_c^2 = H_c,$$

so H_c is the projection matrix onto $\mathcal{C}(X_{S^c})$. Because X_{S^c} consists of a subset of the columns of X, we have

$$\mathcal{C}(X_{S^c}) \;\subseteq\; \mathcal{C}(X).$$

In other words, projecting first onto $\mathcal{C}(X_{S^c})$ is a stricter (or smaller) projection than projecting onto $\mathcal{C}(X)$.

Because $\mathcal{C}(X_{S^c}) \subseteq \mathcal{C}(X)$, we have the nesting of subspaces, and we have:

$$H\,H_c = H_c, \qquad H_c\,H = H_c.$$

Here is the verification:

$$\begin{aligned} H\,H_c &= \left[X\,(X^\top X)^{-1}X^\top\right]\left[X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top\right] \\ &= X\,(X^\top X)^{-1}X^\top X_{S^c}(X_{S^c}^\top X_{S^c})^{-1}X_{S^c}^\top. \end{aligned}$$

Because X_{S^c} consists of a subset of the columns of X, each column of X_{S^c} lies in the column space of X. Thus there exists some matrix B such that

$$X_{S^c} \;=\; X\,B.$$

Consequently,

$$X^\top X_{S^c} \;=\; X^\top(X\,B) \;=\; (X^\top X)\,B.$$

$$\begin{aligned} H\,H_c &= \left[X\,(X^\top X)^{-1}X^\top\right]\left[X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top\right] \\ &= X\,(X^\top X)^{-1}\;\underbrace{X^\top X_{S^c}}_{\text{cross-term}}\;\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top \\ &= X\,(X^\top X)^{-1}\left[(X^\top X)\,B\right]\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top \quad (\text{since } X^\top X_{S^c} = (X^\top X)\,B). \end{aligned}$$

Thus

$$H\,H_c \;=\; X\left[(X^\top X)^{-1}(X^\top X)\,B\right]\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top \;=\; X\,B\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top.$$

and $X_{S^c} = X\,B$, therefore,

$$H\,H_c = X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top = H_c$$

Similarly,

$$H_c\,H \;=\; \left[X_{S^c}(X_{S^c}^\top X_{S^c})^{-1}X_{S^c}^\top\right]\left[X(X^\top X)^{-1}X^\top\right].$$

Again substitute $X_{S^c} = X\,B$:

$$X_{S^c}^\top X = (X\,B)^\top X = B^\top \underbrace{X^\top X}_{\text{used in the same manner}},$$

and proceed as above. The same pattern emerges, and one obtains $H_c\, H = H_c$.

In the full model, the fitted values are $\hat{y} \;=\; H\, y$, and hence the residuals are $y - \hat{y} = (I - H)\, y$.

In the restricted model H_0, the fitted values are $y_{H_0} \;=\; H_c\, y$. Hence the difference between the two fitted values is $\hat{y} - y_{H_0} \;=\; (H - H_c)\, y$.

$$\begin{aligned}(y - \hat{y})^\top (\hat{y} - y_{H_0}) &= \Big[(I - H)\, y\Big]^\top \Big[(H - H_c)\, y\Big] \\ &= \; y^\top (I - H)^\top (H - H_c)\, y.\end{aligned}$$

Because H is symmetric, $(I - H)^\top = I - H$. So:

$$= \; y^\top \, (I - H)\, (H - H_c)\, y.$$

$$(I - H)\, (H - H_c) \;=\; (I - H)\, H \;-\; (I - H)\, H_c.$$

for each term:

$$(I - H)\, H = H - H^2.$$

But $H^2 = H$ (idempotent), so

$$H - H^2 = 0.$$

Hence $(I - H)H = 0$. and,

$$(I - H)\, H_c = H_c - H\, H_c.$$

But from above, $H\, H_c = H_c$. Therefore

$$H_c - H\, H_c = 0.$$

Hence $(I - H)H_c = 0$.

Putting these together,

$$(I - H)\, (H - H_c) = (I - H)H \;-\; (I - H)H_c = 0 \;-\; 0 = 0.$$

Thus

$$y^\top (I - H)(H - H_c)\, y \;=\; y^\top \, 0\, y \;=\; 0.$$

Hence

$$(y - \hat{y})^\top (\hat{y} - y_{H_0}) = 0,$$

We obtain:

$$\| \, y - y_{H_0} \|^2 \;=\; \left\| \, y - \hat{y} + \hat{y} - y_{H_0} \right\|^2 \;=\; \| \, y - \hat{y} \|^2 \;+\; \| \hat{y} - y_{H_0} \|^2 + 0 = \| \, y - \hat{y} \|^2 \;+\; \| \hat{y} - y_{H_0} \|^2$$

Thus we have shown the desired identity:

$$\| \, y - y_{H_0} \|^2 \;=\; \| \, y - \hat{y} \|^2 \;+\; \| \hat{y} - y_{H_0} \|^2,$$

(c)

$\|y - \hat{y}\|^2$ and $\|\hat{y} - y_{H_0}\|^2$ can each be written as a quadratic form in y. Specifically,

$$\|y - \hat{y}\|^2 \;=\; \|(I - H)\, y\|^2, \qquad \|\hat{y} - y_{H_0}\|^2 \;=\; \|(H - H_c)\, y\|^2,$$

where

$$H \;=\; X\, (X^\top X)^{-1} X^\top, \quad H_c \;=\; X_{S^c} \left(X_{S^c}^\top X_{S^c}\right)^{-1} X_{S^c}^\top.$$

Since both $(I - H)y$ and $(H - H_c)y$ are linear transformations of the Gaussian vector y. Hence each is a Gaussian vector in R^n. To show the corresponding sums of squares are independent, we only need to show these two Gaussian

vectors are independent as random vectors. For jointly Gaussian vectors, uncorrelatedness implies independence. Hence it suffices to show their covariance is zero.

From part(b), we already know $(I_n - H)(H - H_c) = 0$, and $H - H_c$ is also a projection matrix, indicating that $(H - H_c)^T = (H - H_c)$)

$$\mathrm{Cov}\big((I_n - H)y,\ (H - H_c)y\big) \ = \ (I_n - H)\,\mathrm{Var}[y]\,(H - H_c)^T \ = \ \sigma^2\,(I_n - H)\,(H - H_c) \ = \ 0_n.$$

Therefore, zero covariance implies $\|y - \hat{y}\|^2$ and $\|\hat{y} - y_{H_0}\|^2$ are independent.

(d)

Under H_0, the true model becomes

$$y = X_{S^c}\,\beta_{S^c} + \sigma\,Z,$$

where $Z \sim N(0,\ I_n)$. The restricted fit is $\hat{y}_{H_0} = H_c\,y$. Therefore,

$$y - \hat{y}_{H_0} \ = \ (I_n - H_c)\,y \ = \ (I_n - H_c)\,\big(X_{S^c}\,\beta_{S^c} + \sigma Z\big).$$

$$= \ \big[X_{S^c} - X_{S^c}(X_{S^c}^\top X_{S^c})^{-1} X_{S^c}^\top X_{S^c}\big]\beta_{S^c} \ + \ \sigma\,(I_n - H_c)\,Z$$

$$= \big[X_{S^c} - X_{S^c}\big]\beta_{S^c} \ + \ \sigma\,(I_n - H_c)\,Z$$

$$= \ \sigma\,(I_n - H_c)\,Z.$$

Thus,

$$\|\,y - \hat{y}_{H_0}\|^2 \ = \ \sigma^2\,\|(I_n - H_c)\,Z\|^2.$$

Since $Z \sim N(0, I_n)$, the vector $(I_n - H_c)\,Z$ is a Gaussian in R^n whose dimension equals the rank of $(I_n - H_c)$. By lemma from class,

$$\|(I_n - H_c)\,Z\|^2 \ \sim \ \chi^2_{\mathrm{rank}(I_n - H_c)}.$$

Note:

$$\mathrm{rank}(I_n - H_c) \ = \ \mathrm{Tr}(I_n - H_c) \ = \ \mathrm{Tr}(I_n) \ - \ \mathrm{Tr}(H_c) \ = \ n \ - \ \mathrm{Tr}\big(H_c\big).$$

The $\mathrm{Tr}(H_c)$ equals the dimension of the column space of X_{S^c}, which is $p - s$ if X is full rank ($p \le n$) and $|S| = s$. Hence

$$\mathrm{Tr}(H_c) = p - s \quad \Longrightarrow \quad \mathrm{rank}(I_n - H_c) = n - (p - s) = n - p + s.$$

Therefore,

$$\frac{\|\,y - \hat{y}_{H_0}\|^2}{\sigma^2} \ = \ \|(I_n - H_c)\,Z\|^2 \ \sim \ \chi^2_{n-p+s}.$$

For distribution of $\|\hat{y} - \hat{y}_{H_0}\|^2/\sigma^2$:

The difference between the full fit $\hat{y} = H\,y$ and the restricted fit $\hat{y}_{H_0} = H_c\,y$ is

$$\hat{y} - \hat{y}_{H_0} \ = \ (H - H_c)\,y.$$

Under H_0, write again $y = X_{S^c}\,\beta_{S^c} + \sigma\,Z$. Then

$$(H - H_c)\,y \ = \ (H - H_c)\,\big(X_{S^c}\,\beta_{S^c} + \sigma\,Z\big).$$

Therefore,

$$\hat{y} - \hat{y}_{H_0} = \ (H - H_c)\,y \ = \ (H - H_c)\,\big(X_{S^c}\,\beta_{S^c} + \sigma Z\big)$$

$$= \ \Big[X\,(X^\top X)^{-1}X^\top - X_{S^c}\,\big(X_{S^c}^\top X_{S^c}\big)^{-1}\,X_{S^c}^\top\Big]\big(X_{S^c}\,\beta_{S^c} + \sigma Z\big)$$

Since in (c) we already have:

$$X_{S^c} \ = \ X\,B.$$

Consequently,

$$X^\top X_{S^c} \;=\; X^\top (X\,B) \;=\; (X^\top X)\,B.$$

Substitute:

$$\Big[X\,(X^\top X)^{-1}X^\top - X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top\Big]\big(X_{S^c}\,\beta_{S^c}+\sigma Z\big)$$

$$= \Big[X\,(X^\top X)^{-1}X^\top XB - X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top X_{S^c}\Big]\beta_{S^c}+\sigma\Big[X\,(X^\top X)^{-1}X^\top \;-\; \big(X_{S^c}-X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top\big)\Big]Z$$

$$= \Big[XB - X_{S^c}\Big]\beta_{S^c}+\sigma\Big[X\,(X^\top X)^{-1}X^\top \;-\; \big(X_{S^c}-X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top\big)\Big]Z$$

$$= \big(X_{S^c}-X_{S^c}\big)\,\beta_{S^c} \;+\; \sigma\Big[X\,(X^\top X)^{-1}X^\top \;-\; \big(X_{S^c}-X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top\big)\Big]\,Z$$

$$= \;0\;\cdot\beta_{S^c} \;+\; \sigma\,(H-H_c)\,Z = \;\sigma\,(H-H_c)\,Z.$$

We get

$$\hat{y}-\hat{y}_{H_0} = \sigma\,(H-H_c)\,Z.$$

$$\|\hat{y}-\hat{y}_{H_0}\|^2 \;=\; \sigma^2\,\|(H-H_c)\,Z\|^2.$$

By the same lemma,

$$\frac{\|\hat{y}-y_{H_0}\|^2}{\sigma^2} \;=\; \|(H-H_c)\,Z\|^2 \;\sim\; \chi^2_{\mathrm{rank}(H-H_c)}.$$

Finally, we compute

$$\mathrm{rank}(H-H_c) \;=\; \mathrm{Tr}(H-H_c) \;=\; \mathrm{Tr}(H)-\mathrm{Tr}(H_c).$$

Since H is the full projection of rank p (assuming X is $n\times p$ with full column rank),

$$\mathrm{Tr}(H)=p, \quad \mathrm{Tr}(H_c)=p-s.$$

Hence

$$\mathrm{rank}(H-H_c)=p-(p-s)=s.$$

Therefore,

$$\frac{\|\hat{y}-\hat{y}_{H_0}\|^2}{\sigma^2} \;\sim\; \chi^2_s.$$

Therefore, under the null hypothesis $H_0:\beta_S=0$, we have:

$$\frac{\|\,y-\hat{y}_{H_0}\|^2}{\sigma^2} \;\sim\; \chi^2_{n-p+s}, \qquad \frac{\|\hat{y}-\hat{y}_{H_0}\|^2}{\sigma^2} \;\sim\; \chi^2_s.$$

(e)

Given the model

$$y \;\sim\; N\big(X\beta,\;\sigma^2 I_n\big),$$

we want to test

$$H_0:\beta_S=0 \quad \text{vs.} \quad H_1:\beta_S\neq 0,$$

where $S\subseteq\{0,1,\ldots,p-1\}$ is a subset of indices, of size $s=|S|$, and X_S (respectively X_{S^c}) is the submatrix of X corresponding to those columns (respectively, the complementary columns).

- **Restricted fit:** $\hat{y}_{H_0}=H_c\,y$, where

$$H_c \;=\; X_{S^c}\left(X_{S^c}^\top X_{S^c}\right)^{-1}X_{S^c}^\top.$$

  The residual sum of squares (RSS) under H_0 is

$$\|\,y-\hat{y}_{H_0}\|^2.$$

- **Full fit:** $\hat{y} = H\,y$, where

$$H \;=\; X\,(X^\top X)^{-1}\,X^\top.$$

The residual sum of squares (RSS) under the full model is

$$\|\,y - \hat{y}\|^2.$$

from part(d) we have

$$\|\,y - \hat{y}_{H_0}\|^2 \;\sim\; \sigma^2\,\chi^2_{n-p+s} \quad \text{and} \quad \|\hat{y} - \hat{y}_{H_0}\|^2 \;\sim\; \sigma^2\,\chi^2_s \quad \text{(independently)},$$

under H_0. Also

$$\|\,y - \hat{y}\|^2 = \|\,y - \hat{y}_{H_0}\|^2 - \|\hat{y} - \hat{y}_{H_0}\|^2.$$

Hence $\|\,y - \hat{y}\|^2/\sigma^2 \sim \chi^2_{n-p}$

The "extra sum of squares" contributed by the columns in S is

$$\|\hat{y} - \hat{y}_{H_0}\|^2 \;=\; \|\,y - \hat{y}_{H_0}\|^2 \;-\; \|\,y - \hat{y}\|^2.$$

From the previous results:

$$F \;=\; \frac{\big[\|\hat{y} - \hat{y}_{H_0}\|^2/\sigma^2\big]\Big/\,s}{\big[\|y - \hat{y}\|^2/\sigma^2\big]\Big/\,(n-p)}.$$

$$F = \frac{\|\hat{y} - \hat{y}_{H_0}\|^2/s}{\|y - \hat{y}\|^2/(n-p)} \;\sim\; F_{s,\,n-p} \quad \text{(under } H_0\text{)}.$$

Thus the F-test statistic is

$$F \;=\; \frac{\big[\|\hat{y} - \hat{y}_{H_0}\|^2/s\big]}{\big[\|y - \hat{y}\|^2/(n-p)\big]},$$

and we reject H_0 (i.e. conclude that at least one of $\beta_j, j \in S$ is nonzero) if:

$$\frac{\|\hat{y} - \hat{y}_{H_0}\|^2/s}{\|y - \hat{y}\|^2/(n-p)} > F_{s,\,n-p,1-\alpha}$$

(f)

In this scenario, we are testing:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

Hence the submatrix X_S is all columns of X except the intercept column X_0. Equivalently, X_{S^c} consists only of the intercept column. so

$$X_{S^c} = 1_n,$$

where 1_n denotes the n-vector of all ones.

Under H_0, the restricted least squares fit is given by

$$\hat{y}_{H_0} = H_c\,y, \quad \text{with} \quad H_c = X_{S^c}\,(X_{S^c}^\top X_{S^c})^{-1} X_{S^c}^\top.$$

Since

$$X_{S^c} = 1_n \quad \text{and} \quad X_{S^c}^\top X_{S^c} = 1_n^\top 1_n = n,$$

it follows that

$$H_c = 1_n\,\frac{1}{n}\,1_n^\top = \frac{1}{n}\,1_n 1_n^\top.$$

Thus, the restricted fitted values are

$$\hat{y}_{H_0} = H_c\,y = \frac{1}{n}\,1_n 1_n^\top y,$$

which is the n-vector with every entry equal to the sample mean $\bar{y}$.

From our earlier results (see part (e)), we have shown that

$$\hat{y} - \hat{y}_{H_0} = \sigma\,(H - H_c)\,Z,$$

where

$$H = X\,(X^\top X)^{-1}X^\top \quad \text{and} \quad Z \sim N(0, I_n).$$

Thus,

$$\|\hat{y} - \hat{y}_{H_0}\|^2 = \sigma^2\,\|(H - H_c)Z\|^2.$$

By a standard lemma,

$$\frac{\|\hat{y} - \hat{y}_{H_0}\|^2}{\sigma^2} \sim \chi^2_{\mathrm{rank}(H-H_c)}.$$

Now, the rank of $(H - H_c)$ is given by

$$\mathrm{rank}(H - H_c) = \mathrm{Tr}(H - H_c) = \mathrm{Tr}(H) - \mathrm{Tr}(H_c).$$

Since X is $n \times p$ with full column rank, the projection matrix H has

$$\mathrm{Tr}(H) = p.$$

Moreover, as we have just shown,

$$H_c = \frac{1}{n}\,1_n 1_n^\top,$$

so

$$\mathrm{Tr}(H_c) = \frac{1}{n}\,\mathrm{Tr}(1_n 1_n^\top) = \frac{1}{n}\,(1_n^\top 1_n) = \frac{n}{n} = 1.$$

Thus,

$$\mathrm{rank}(H - H_c) = p - 1.$$

It follows that

$$\frac{\|\hat{y} - \hat{y}_{H_0}\|^2}{\sigma^2} \sim \chi^2_{p-1}.$$

The results in (e) shows that in the full model the residual sum of squares (RSS) is

$$\|y - \hat{y}\|^2,$$

,under the full model, has $n - p$ degrees of freedom:

$$\frac{\|y - \hat{y}\|^2}{\sigma^2} \sim \chi^2_{n-p}.$$

Thus, the standard F-statistic for testing

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_{p-1} = 0$$

is given by

$$F = \frac{\big[\|\hat{y} - \hat{y}_{H_0}\|^2/(p-1)\big]}{\big[\|y - \hat{y}\|^2/(n-p)\big]} \;\sim\; F_{p-1,\,n-p}.$$

Noting that $\hat{y}_{H_0} = \bar{y}\,1_n$ the above F-statistic becomes

$$F = \frac{\big[\|\hat{y} - \bar{y}\,1_n\|^2/(p-1)\big]}{\big[\|y - \hat{y}\|^2/(n-p)\big]} \sim\; F_{p-1,\,n-p},$$

which is exactly the standard result taught in class. Therefore, we reject H_0 at significance level α if

$$F > F_{p-1,n-p,1-\alpha}$$