HW3: Simultaneous Inference and Matrix Representation (ALRM Ch. 4, 5)

MATH/STAT 571A

DUE: 10/06/2023 11:59pm

Homework Guidelines

Please submit your answers on Gradescope as a PDF with pages matched to question answers.

One way to prepare your solutions to this homework is with R Markdown, which provides a way to include mathematical notation, text, code, and figures in a single document. A template `.Rmd` file is available through D2L.

Make sure all solutions are clearly labeled, and please utilize the question pairing tool on Gradescope. You are encouraged to work together, but your solutions, code, plots, and wording should always be your own. Come and see me and/or our TA during office hours or schedule an appointment when you get stuck and can't get unstuck.

I. Mathematical Foundations [25 pts]

(1) [3 pts] (ALRM 4.21) When the predictor variable is coded so that $\bar{X} = 0$ and the normal error regression model (2.1) applies, are b_0 and b_1 independent? Are the joint confidence intervals for β_0 and β_1 then independent? Explain your reasoning.

We can show that cov{b} = 0: $cov\{b_0, b_1\} = -\bar{X}s^2 b_1 = 0$. This means that in the joint confidence interval, b0 will not tend to vary higher or lower than b1. However, being linearly independent (i.e. covariance is 0) does not necessarily imply independence, so b_0 and b_1 are still related since they are regressed using the same dataset.

(2) [5 pts] Suppose you have a sample of n responses, $Y_1, \ldots, Y_n$, and associated predictor values, $X_1, \ldots, X_n$, and you are interested in predicting the expected outcome, $\mathrm{E}[Y_h]$, at level X_h of the predictor. You are also able to make one more observation of the response, Y_{n+1}, at a level, X_{n+1}, of your choosing. What value of X_{n+1} minimizes the variance of the sampling distribution of $\hat{Y}_h$, $\mathrm{Var}(\hat{Y}_h)$?

A more extreme value of X_{n+1} (i.e. far from $\overline{X}$) since it will minimize the variance of the coefficient estimators and $\widehat{Y}_h$.

(3) [3 pts] Use matrix notation and properties of the projection matrix $\mathbf{H} = \mathbf{X}\left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'$ to show property (1.20) of the least-squares fitted regression line: $\sum_{i=1}^{n}\hat{Y}_i e_i = 0$.

Want to show: (HY)'e = 0 We start with the LHS:

$$
\begin{aligned}
(HY)'e &= (HY)'(I - H)Y \\
&= Y'H'(I - H)Y \\
&= Y'H'Y - Y'H'HY \quad \text{*H' = H (symmetric)} \\
&= Y'HY - Y'HY \quad \text{*H'H = H (idempotent)} \\
&= 0
\end{aligned}
$$

(4) [3 pts] (ALRM 5.8) Let $\mathbf{B}$ be defined as follows:

$$\mathbf{B} = \begin{bmatrix} 1 & 5 & 0 \\ 1 & 0 & 5 \\ 1 & 0 & 5 \end{bmatrix}$$

a. Are the column vectors of $\mathbf{B}$ linearly dependent? Yes.

$$\begin{bmatrix} 5 \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ 5 \\ 5 \end{bmatrix} = 5 \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} \text{ linear combination}$$

b. What is the rank of $\mathbf{B}$? Rank 2. $\begin{bmatrix} 5 \\ 0 \\ 0 \end{bmatrix} and \begin{bmatrix} 0 \\ 5 \\ 5 \end{bmatrix}$

c. What must be the determinant of $\mathbf{B}$? The determinant is 0, since the square matrix is not full-rank.
Problem 5.4: Flavor deterioration. The results shown below were obtained in a small-scale experiment to study the relation between °F of storage temperature (X) and number of weeks before flavor deterioration of a food product begins to occur (Y).

i:	1	2	3	4	5
X_i	8	4	0	-4	-8
Y_i	7.8	9.0	10.2	11.0	11.7

(5) [2 pts] (ALRM 5.12) Refer to ALRM **Flavor deterioration Problem** . Find $(\mathbf{X}'\mathbf{X})^{-1}$.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 0.2 & 0 \\ 0 & 0.00625 \end{bmatrix}$$

```
flav_df <- data.frame(cbind(seq(8,-8, by = -4),c(7.8, 9.0, 10.2, 11.0, 11.7)))
rownames(flav_df) <- seq(1,5)
colnames(flav_df) <- c("X","Y")
flav_df
```

```
##    X    Y
## 1  8  7.8
## 2  4  9.0
## 3  0 10.2
## 4 -4 11.0
## 5 -8 11.7
```

```
X <- matrix(cbind(rep(1,5),flav_df$X), nrow = 5, ncol = 2, byrow = FALSE)

solve(t(X) %*% X)
```

```
##       [,1]    [,2]
## [1,]  0.2 0.00000
## [2,]  0.0 0.00625
```

(6) [3 pts] (ALRM 5.23 c.) Refer to ALRM **Flavor deterioration Problems 5.4** and **5.12**.

c. Find the hat matrix **H**.

Verify that the fitted values obtained using the `lm()` function match those you get from using matrix methods. Include your R code.

```
Y <- flav_df$Y
H <- X %*% solve(t(X) %*% X) %*% t(X) %*% Y

fit_flav <- lm(Y~X, flav_df)
fitted <- fit_flav$fitted.values
comparison <- data.frame(cbind(H, matrix(fitted)))
colnames(comparison) <- c("matrix method", "lm fitted")
comparison
```

```
##    matrix method lm fitted
## 1           7.98      7.98
## 2           8.96      8.96
## 3           9.94      9.94
## 4          10.92     10.92
## 5          11.90     11.90
```

(7) [3 pts] (ALRM 5.29) Consider the least squares estimator **b** given in (5.60). Using matrix methods, show that **b** is an unbiased estimator for β.

$$\begin{aligned} E\{b\} &= E\{(X'X)^{-1}X'Y\} \\ &= E\{(X'X)^{-1}X'(X\beta + \epsilon\} \\ &= E\{(X'X)^{-1}X'X\beta\} + (X'X)^{-1}X'E\{\epsilon\} \text{ *goes to I} \\ &= E\{\beta\} \\ &= \beta \text{ is unbiased} \end{aligned}$$

(8) [3 pts] (ALRM 5.31) Derive an expression for the variance-covariance matrix of the fitted values $\hat{Y}_i, i = 1, \ldots, n$ in terms of the hat matrix.

$$\begin{aligned} \sigma^2\{\hat{Y}_i\} &= \sigma^2\{HY\} \\ &= H'\sigma^2\{Y\}H \\ &= H'\sigma^2 IH \\ &= \sigma^2 H'H \\ &= \sigma^2 H \end{aligned}$$

II. Average park size and household income in Tucson [15 pts]

The data in `Tucson_parks.csv` (on D2L) are median acreages (area in acres) of parks and median household incomes aggregated by zip code for the City of Tucson, AZ. Data were compiled from two sources provided by governments in AZ: State **census data**, and Tucson **park properties**.

We will treat the *log* of median acreage as the response variable and seek to understand its relationship with median household income.

(9) [2 pts] Fit a simple linear regression model with log-median acreage as the response and median household income as the predictor using least-squares. Verify you get a point estimate for the intercept of -7.638095e-01. Give a point estimate for β_1.

The SLR point estimate for β_1 is 4.317871e-05.

```
parks <- read.csv("./Datasets/Tucson_parks.csv")
names(parks)
```

```
## [1] "X"                  "ZIP"
## [3] "MEDIAN_ACREAGE"     "MEDIAN_HOUSEHOLD_INCOME"
```

```
fit_parks <- lm(log(MEDIAN_ACREAGE) ~ MEDIAN_HOUSEHOLD_INCOME, data = parks)
coef(fit_parks)
```

```
##             (Intercept) MEDIAN_HOUSEHOLD_INCOME
##           -7.638095e-01            4.317871e-05
```

(10) [2 pts] Obtain Bonferroni joint confidence intervals for β_0 and β_1 using a 95 percent family confidence coefficient. Does either interval overlap 0?

Yes, the confidence interval for the intercept overlaps 0.

```
alpha = .05
confint(object = fit_parks, level = 1-(alpha/2))
```

```
##                                  1.25 %      98.75 %
## (Intercept)             -2.433712e+00 0.9060928219
## MEDIAN_HOUSEHOLD_INCOME  1.514572e-05 0.0000712117
```

(11) [2 pts] Will b_0 and b_1 tend to err in the same direction or in opposite directions in this application? Explain.

They tend to err in opposite directions since they have a negative covariance. Higher estimates of b0 tend to associate with lower estimates of b1, and lower estimates of b0 associate with higher estimates of b1.

```
x <- parks$MEDIAN_HOUSEHOLD_INCOME
x_bar <- mean(x)
mse <- (summary(fit_parks)$sigma)^2
s2_b1 <- mse/sum((x - x_bar)^2)

b_cov <- -1 * x_bar * s2_b1
b_cov
```

```
## [1] -7.324992e-06
```

(12) [3 pts] Use the Working-Hotelling procedure to obtain 90% confidence intervals for the expected log-median acreage of parks in zip codes where the median household incomes are $50,000 and $70,000.

$E\{\widehat{Y}_h = \$50{,}000\}$ 90% CI: (0.8597448, 1.930507)

$E\{\widehat{Y}_h = \$70{,}000\}$ 90% CI: (1.6358492, 2.881551)

```
pred_WH <- function(object, newdata, level = 0.95){
  fit <- predict(object, newdata) ## Yhat
  MSE <- summary(object)$sigma^2
  n <- nrow(object$model) ## extract df from model object
  W <- sqrt(2 * qf(level, 2, n - 2))
  X_obs <- object$model[, attr(object$terms, "term.labels")] ## extract X
  X <- newdata[, attr(object$terms, "term.labels")]
  ME <- W * sqrt(MSE * (1 / n + (X - mean(X_obs))^2 /
                          sum((X_obs - mean(X_obs))^2)))
  upr <- fit + ME
  lwr <- fit - ME
  return(cbind(fit, lwr, upr))
}

wh_ci <- pred_WH(fit_parks,
                 newdata = data.frame(MEDIAN_HOUSEHOLD_INCOME = c(50000,70000)),
                 level = 0.9)
wh_ci
```

```
##          fit       lwr      upr
## 1 1.395126 0.8597448 1.930507
## 2 2.258700 1.6358492 2.881551
```

(13) [3 pts] Use the Bonferroni procedure to obtain 90% confidence intervals for the expected log-median acreage of parks in zip codes where the median incomes are $50,000 and $70,000. Which intervals are narrower?

$E\{\widehat{Y}_h = \$50,000\}$ 90% CI: (0.8229155, 1.967337)

$E\{\widehat{Y}_h = \$70,000\}$ 90% CI: (1.5930028, 2.924398)

The Bonferroni intervals are narrower since they do not span the entire regression line.

```
alpha = .1
bon_ci <- predict(fit_parks,
                  newdata = data.frame(MEDIAN_HOUSEHOLD_INCOME = c(50000,70000)),
                  level = 1-(alpha/2), interval = "confidence")

# Comparison
ci_50k <- rbind(wh_ci[1,2:3], bon_ci[1,2:3])
ci_70k <- rbind(wh_ci[2,2:3], bon_ci[2,2:3])
comparison_df <- as.data.frame(cbind(ci_50k,ci_70k))
names(comparison_df) <- c("50k_lwr", "50k_upr", "70k_lwr", "70k_upr")
rownames(comparison_df) <- c("WH", "Bonferroni")
```

(14) [3 pts] Obtain joint confidence intervals for the expected log-median acreage of parks in zip codes where the median household incomes are $50,000, $60,000, and $70,000 using both Working-Hotelling and Bonferroni procedures. Which intervals are narrower?

The W-H intervals are smaller than Bonferroni intervals for three estimators of mean log_MEDIAN_HOUSEHOLD_INCOME. The Bonferroni method is much more conservative, so corrections become less narrow faster as the number of joint estimations gets larger.

```
wh_ci2 <- pred_WH(fit_parks,
                  newdata = data.frame(MEDIAN_HOUSEHOLD_INCOME = c(50000,60000, 70000)),
                  level = 0.9)
bon_ci2 <- predict(fit_parks,
                   newdata = data.frame(MEDIAN_HOUSEHOLD_INCOME = c(50000, 60000, 70000)),
                   level = 1-(alpha/3), interval = "confidence")

ci2_span_50k <- rbind(diff(wh_ci[1,2:3]), diff(bon_ci[1,2:3]))
ci2_span_60k <- rbind(diff(wh_ci[2,2:3]), diff(bon_ci[2,2:3]))
ci2_span_70k <- rbind(diff(wh_ci2[3,2:3]), diff(bon_ci2[3,2:3]))
comparison2_df <- as.data.frame(cbind(ci2_span_50k, ci2_span_60k, ci2_span_70k))
names(comparison2_df) <- c("50k_ci_size", "60k_ci_size", "70k_ci_size")
rownames(comparison2_df) <- c("WH", "Bonferroni")
comparison2_df
```

```
##            50k_ci_size 60k_ci_size 70k_ci_size
## WH           1.0707625    1.245702    1.245702
## Bonferroni   0.9822895    1.142774    1.254222
```