STAT 425 — Sections 2UG, 2GR, 3UG, 3GR — Spring 2023

Homework 1

Due: February 6, 11:59 PM (US Central)

Please submit your assignment electronically as a PDF document, using the appropriate interface on Canvas. Remember to include relevant computer output.

1. For a constant matrix $\boldsymbol{A}$ and a random vector $\boldsymbol{Z}$,

$$\mathrm{E}(\boldsymbol{AZ}) \;=\; \boldsymbol{A}\,\mathrm{E}(\boldsymbol{Z}) \qquad \mathrm{Var}(\boldsymbol{AZ}) \;=\; \boldsymbol{A}\,\mathrm{Var}(\boldsymbol{Z})\boldsymbol{A}^T$$

   (assuming expectations and variances all exist).

   Consider the linear model $\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$ under the Gauss-Markov conditions. Assume the columns of $\boldsymbol{X}$ are linearly independent. For each of the following random vectors, determine the mean vector and the variance-covariance matrix (in terms of $\boldsymbol{X}$, $\boldsymbol{\beta}$, and σ^2). Simplify as much as possible.

   (a) [2 pts] $\boldsymbol{\varepsilon}$

   (b) [2 pts] $\boldsymbol{Y}$

   (c) [2 pts] $\hat{\boldsymbol{\beta}}$

   (d) [2 pts] $\hat{\boldsymbol{Y}}$ (the random vector for which the realization is the computed vector $\hat{\boldsymbol{y}}$ of fitted values)

   (e) [2 pts] $\hat{\boldsymbol{\varepsilon}} \;=\; \boldsymbol{Y} - \hat{\boldsymbol{Y}}$ [Hint: $\boldsymbol{Y} - \hat{\boldsymbol{Y}} = \big(\boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\big)\boldsymbol{Y}$]

2. The data set `teengamb` concerns a study of teenage gambling in Britain. Fit a regression model with expenditure on gambling as the response, and the sex, status, income, and verbal scores as predictors. (Try `help(teengamb)` for information about the variables.)

   (a) [2 pts] Present a summary of your fitted model. (Use the R `summary` function.)

   (b) [2 pts] Give all of the least squares estimates $\hat{\beta}$.

   (c) [2 pts] What is the *name* for the proportion of variation in the response explained by the predictors? What is its *value*, for the model you fit?

   (d) [2 pts] Which observation (case number) has the largest (positive) residual? Also, what is its fitted value?

   (e) [2 pts] When all other predictors are held constant, what would be the estimated difference in expected expenditure on gambling for a male compared to a female?

   (f) [2 pts] Which independent variables are statistically significant at the 5% (0.05) level?

   (g) [2 pts] For each regression coefficient, compute a 95% confidence interval.

   (h) [2 pts] Predict the amount that a male with average status, income, and verbal score (averaged over all cases) would gamble. Also, give a 95% prediction interval.

   (i) [2 pts] Fit a model with only `income` as a predictor, and use an F-test to compare it with the full model.

3. Using the `seatpos` data set, fit a regression model with `hipcenter` as the response, and `Age`, `Weight`, and `Ht` as predictors.

   (a) [2 pts] Present a summary of your fitted model.

   (b) [2 pts] Test the (null) hypothesis that $\beta_{\texttt{Age}} = 0$.

   (c) [2 pts] Test the (null) hypothesis that $\beta_{\texttt{Age}} = \beta_{\texttt{Weight}} = \beta_{\texttt{Ht}} = 0$.

   (d) [2 pts] Add `HtShoes` as another predictor, and present a summary of your fitted model.

   (e) [2 pts] Use an F-test to test whether $\beta_{\texttt{HtShoes}} = 0$.

   (f) [2 pts] Compare your results with the results of a t-test for $\beta_{\texttt{HtShoes}} = 0$. How similar are the p-values?

4. [GRADUATE SECTION ONLY] Consider the usual simple linear regression model (with intercept) under the usual assumptions about the distribution of the errors. The model is satisfied by pairs $(x_1, Y_1), \ldots (x_n, Y_n)$, where the x values are fixed constants, and the Ys are random variables.

   (a) [3 pts] *Derive* an expression for $\mathrm{Var}\big(\overline{Y}\big)$. (You may use the fact that the Ys are uncorrelated.)

   (b) [3 pts] *Derive* an expression for $\mathrm{Cov}\big(x_i Y_i,\, \overline{Y}\big)$ (the covariance between $x_i Y_i$ and $\overline{Y}$).

   (c) [4 pts] Assume that the least squares *estimators* $\hat{\beta}_0$ and $\hat{\beta}_1$ exist. (They are random because they depend on the random Ys.)
   Derive an expression for $\mathrm{Cov}\big(\hat{\beta}_1, \overline{Y}\big)$.

   Note: In all parts, your derived expression must depend only on the x values and the model parameters.

Some comments:

- Unless otherwise stated, all data sets can be found in either the `faraway` package or the `alr4` package in R.
- Unless otherwise stated, use a 5% level ($\alpha = 0.05$) in all tests.