Assignment 2: STAT20230 Modern Regression Analysis

Instructions for submission: For questions that require handwritten work, please make sure that scans are readable and preferably upload scans as a single file. In questions where software is required, R scripts (.R files) will not be accepted as submission. You should prepare a report containing the code / code output used. This may be a word document, or a RMarkdown/Quarto compiled pdf.

1. Statistical properties of predictions (handwritten work):

   (a) Calculate the expected value of a predicted value $\hat{Y}^*_\epsilon$, where $Y^*_\epsilon = \beta_0 + \beta_1 X^* + \epsilon$ and X^* is the new value of X.

   (b) Calculate the variance of $\hat{Y}^*_\epsilon$.

   (c) ~~Show that (b) is equivalent to the variance of the prediction error, $\epsilon^*_i = Y_i - \hat{Y}^*_{i,\epsilon}$.~~

   (d) Show that the expected value of the mean response prediction $\hat{Y}^*_i = E[\hat{Y}^*_{i,\epsilon}]$ is equivalent to (a) but $Var(\hat{Y}^*_i)$ is smaller than (b).

2. Properties of statistical estimators (handwritten work): In the first part of the course, we defined the following estimator of σ^2:

$$\widehat{\sigma}^2 = \frac{SS_e}{n-p-1},$$

   where p is the number of covariates in the model and $SS_e = \sum_{i=1}^n (Y_i - \widehat{Y}_i)^2$.

   (a) Write down the likelihood function of $\boldsymbol{Y} = (Y_1, \dots, Y_n)$ given $\boldsymbol{X} \equiv \begin{pmatrix} 1 & X_{1,1} & \dots & X_{1,p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_{n,1} & \dots & X_{n,p} \end{pmatrix}$ using the matrix notation. Refer to the multivariate normal distribution form for this.

   (b) Find the maximum likelihood estimator of σ^2. For this you may write the log-likelihood function $\sum_{i=1}^n \log f(Y_1|X_{i1}, \dots, X_{ip}, \beta_0, \beta_1, \dots, \beta_p, \sigma^2)$.

   (c) Use the result that $E[SS_e] = (n-p-1)\sigma^2$ to show that $\hat{\sigma}^2_{mle}$ is not an unbiased estimator of σ^2.

   (d) Show that $E[\hat{\sigma}^2_{mle}]$ is asymptotically σ^2, that is, tends to σ^2 as the sample size increases.

3. Model selection (computational): In this question, we will use the `Crime.csv` data set that contains the following information about crimes in states of the U.S.: VR = violent crime rate, M = percent of population living in metropolitan areas, P = percent of families headed by a single parent and MR = mortality rate. Our goal is to explain VR using the best possible subset of the potential covariates M, P, MR. We can fit $2^3 = 8$ different combinations of (M, P, MR).

   (a) Use `R` to fit all possible models and compute AIC, BIC and R^2_{adj} for each model. Report a table with your results.

   (b) Indicate the best model overall according to each of AIC, BIC and R^2_{adj}.

   (c) Implement a forward stepwise regression that uses BIC. You will start by fitting the null model (the model with no covariates) and computing its BIC. Then, consider all possible one covariate models and compute their BICs. Iterate until there is no improvement in your criteria.

   (d) Does the forward stepwise method find the best possible subset? Compare the solutions to item (a) and (c). Explain why solutions from stepwise regression might differ from the all possible regressions method in (a).

   (e) Explain how you could implement a backward selection algorithm using the F test as a decision rule. You can assume that the confidence level is $\alpha = 0.05$.

4. Multiple regression: (computational) Consider the National Football League data of Assignment 1, file `football.csv`.

   (a) Fit a linear regression model relating the number of games won (y) to the team's passing yardage (x_2), the percentage of rushing plays (x_7), and the opponents' yards rushing (x_8).

   (b) Compute the sums of squares:

   - $SS_T = \sum_{i=1}^{28}(Y_i - \bar{Y})^2$
   - $SS_e = \sum_{i=1}^{28}(Y_i - \widehat{Y}_i)^2$
   - $SS_R = \sum_{i=1}^{28}(\widehat{Y}_i - \bar{Y})^2$
   - What is the relationship between SS_T, SS_e, SS_R?
   - The degrees of freedom of SS_e are $(n - p - 1)$ and the degrees of freedom of SS_R are p. Degrees of freedom obey the same relationship as the sums of squares (previous item). Use this to obtain the degrees of freedom of SS_T.
   - **Mean squares** are the sums of squares divided by their respective degrees of freedom. Compute MST, MSE, MSR which are the Total, Error and Regression Mean Squares.

   (c) Calculate the t statistics for testing the hypothesis $H_0 : \beta_j = 0$ versus $H_0 : \beta_j \neq 0$ for $j = 1, 2, 3$ where $\beta_1, \beta_2, \beta_3$ are the coefficients of x_2, x_7 and x_8. Show how these are obtained from $\hat{\beta}_j$ and include the computation of $Var(\hat{\beta}_j)$. Results obtained using `lm()` will not be accepted.

   (d) Calculate R^2 and R^2_{adj} using (b). Results obtained using `lm()` will not be accepted.

   (e) Use item (b) to test the significance of the regression (F test). Outline the hypothesis of this test, compute the test statistic and conclude using the critical regions approach.

(f) Show numerically that R^2 is equal to the square of the correlation coefficient between Y_i and $\hat{Y}_i$.

(g) Find a 95% CI on the mean number of games won by a team when $x_2 = 2300, x_7 = 56, x_8 = 2100$.

(h) Fit the model using x_7 and x_8 only and compute its error sums of squares.

(i) ~~Use (h) to~~ Perform an F test for $H_0 : \beta_2 = \beta_3 = 0$ versus H_1 : at least one of β_2 or β_3 is different than zero.

(j) Recompute R^2 and R^2_{adj} for the new model. How do these quantities compute to those in (d)?

(k) Recompute the 95% confidence interval on the mean number of games won by a team with the new model, using $x_7 = 56, x_8 = 2100$. Compare the length of the interval to g).

(l) Comment on how removing x_2 from the model changed the model adequacy and its predictions.

5. Types of variables (computational): Consider the `bikesharing` data that records the number of bikes rented (`cnt`) on a given day.

(a) Explain how to code the month of the year `mnth` using indicator variables.

(b) Using R, fit a linear regression model to `cnt` using `hum`: the normalised measure of humidity, `windspeed`: the normalised wind speed on the day, `temp`: the normalised temperature, and `mnth`. Make sure to use `mnth` as a categorical variable using January as the reference category. Include the summary of the fitted model.

(c) Is there an indication that month of the year is an important variable?

(d) What months of the year have a different average number of bike rentals in comparison to January, given `hum, temp` and `windspeed`? Use $\alpha = 0.05$.

(e) Given `hum` $= 0.4$, `temp` $= 0.3$, `windspeed` $= 0.65$, what is the average number of rental in each month? Report a table with your results.