MAT 3375 – Regression Analysis – Questions

1. (a) Let $U_i \sim \chi^2(r_i)$ be independent random variables with $r_1 = 5$, $r_2 = 10$. Set

$$X = \frac{U_1/r_1}{U_2/r_2}.$$

Using R, find s and t such that

$$P(X \leq s) = 0.95 \quad \text{and} \quad P(X \leq t) = 0.99.$$

(b) Let $Z \sim N(0,1)$ and $U \sim \chi^2(10)$ be two independent random variables. Let

$$V = \frac{Z}{\sqrt{U/10}}.$$

Using R, find w such that $P(V \leq w) = 0.95$.

2. Let $f : \mathbb{R}^n \to \mathbb{R}$, $\mathbf{v} \in \mathbb{R}^n$, and $a \in \mathbb{R}$. Define $f(\mathbf{Y}) = \mathbf{Y}^\top \mathbf{v} + a$. Find the gradient of f with respect to $\mathbf{Y}$. Write a function in R that computes $f(\mathbf{Y})$ given $\mathbf{v}, a$. Evaluate the function at $\mathbf{Y} = (1, 0, -1)$, for $\mathbf{v} = (1, 2, -3)$ and $a = -2$.

Note: in the course, we will write vectors either as columns format or as rows, in a more or less arbitrary way. It is up to you to determine which one makes the dimensions compatible.

3. Let $A = \begin{pmatrix} 1 & 1 & 0 \\ 0 & 1 & -1 \end{pmatrix}$, $\boldsymbol{\mu} = (1, 0, 1)$, $\boldsymbol{\Sigma} = \begin{pmatrix} 2 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$, and $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

Let $\mathbf{W} = A\mathbf{Y}$. What distribution does the random vector $\mathbf{W}$ follow? Draw a sample of size 100 for this random vector with R and plot them in a graph. **Note:** you may use the function `mvrnorm()` from the `MASS` package to help along (but you do not have to).

4. Let $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, 9\mathbf{I}_4)$ and set $\overline{Y} = \frac{1}{4}(Y_1 + Y_2 + Y_3 + Y_4)$. Using R, draw 1000 observations from:

(a) $Y_1^2 + Y_2^2 + Y_3^2 + Y_4^2$

(b) $4\overline{Y}^2$

(c) $(Y_1 - \overline{Y})^2 + (Y_2 - \overline{Y})^2 + (Y_3 - \overline{Y})^2 + (Y_4 - \overline{Y})^2$

In each case, plot a histogram of the observations.

5. Consider the function $f : \mathbb{R}^3 \to \mathbb{R}$ defined by

$$f(\mathbf{Y}) = Y_1^2 + \tfrac{1}{2}Y_2^2 + \tfrac{1}{2}Y_3^2 - Y_1Y_2 + Y_1 + 2Y_2 - 3Y_3 - 2.$$

Using R, find the critical point(s) of f. If it is unique, does it give rise to a global maximum of f? A global minimum? A saddle point?

6. (a) Identify the response variable Y and the predictor variable X in each of the examples shown on slides 4 and 5 of the course notes (Chapter 2). Is there a linear relationship between X and Y. Draw the approximate line of linear fit (and give its equation).

Hint: use screenshots and software (Paint, PowerPoint, GIMP, etc.) to overlay the line.

(b) Consider the 4 examples shown on page 9 of the course notes (chapter 2). Is the variance of the error terms constant? Are the error terms independent of each other?

7. Consider the dataset `Autos.xlsx` found on Brightspace. The predictor variable is `VKM.q` (X, the average daily distance driven, in km); the response variable is `CC.q` (Y, the average daily fuel consumption, in L). Use `R` to:

   (a) display the scatterplot of Y versus X;

   (b) determine the number of observations n in the dataset;

   (c) compute the quantities $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum X_iY_i$, $\sum Y_i^2$;

   (d) find the normal equations of the line of best fit;

   (e) find the coefficients of the line of best fit (without using `lm()`), and

   (f) overlay the line of best fit onto the scatterplot.

8. (continuation of the previous question) Use the `R` function `lm()` to obtain the coefficients of the line of best fit and the residuals. Show (by calculating the required quantities directly) that the first 5 properties of residuals (p. 25 in the course notes of Chapter 2) are satisfied.

9. (continuation of the previous question) Using `R`, compute the Pearson and Spearman correlation coefficients between the predictor and the response. Is there a strong or weak linear association between these two variables? Use the correlation values and diagrams to justify your answer.

10. (continuation of the previous question) Using `R`, find the decomposition into sums of squares for the regression.

11. (continuation of the previous question) Using `R`, randomly draw n pairs of observations from the data set. Determine the least squares line of best fit L_n and calculate its coefficient of determination R_n^2. Repeat for $n = 10, 50, 100, 500$ and for all observations. Is there anything interesting to report? If so, how is it explained?

12. Using `R`, plot the residuals corresponding to the ls line of best fit when using all observations in the set. Visually, do the SLR assumptions on the error terms appear to be satisfied? Give a visual approximation of σ^2. Then compute the estimator $\widehat{\sigma}^2$. Compare.

13. Using `R`, compute directly the 95% and the 99% confidence interval of the slope of the regression line.

14. Before even doing the calculations with `R`, do you think we should be able to determine whether the confidence interval for the intercept of the regression line is smaller or larger than the corresponding interval for the slope? If so, why would this be the case? Determine directly the 95% and the 99% confidence interval of the intercept.

15. (continuation of the previous question) Using the fit from the previous questions:

    (a) Test for $H_0 : \beta_0 = 0$ vs. $H_1 : \beta_0 > 0$.

    (b) Test for $H_0 : \beta_1 = 10$ vs. $H_1 : \beta_1 \neq 10$.

    (c) Test for $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$.

    Justify and explain your answers.

16. (continuation of the previous question)

    (a) Using the formulas learned in class, calculate the covariance $\sigma\{b_0, b_1\}$.

(b) Randomly select a sample of 50 pairs of observations from `Autos.xlsx` (with or without remplacement, as desired). Compute the regression parameters $(b_0^{(1)}, b_1^{(1)})$ corresponding to the sample. Repeat the procedure 300 times, to produce 300 pairs $(b_0^{(j)}, b_1^{(j)})$. Display all pairs in a scatter plot.

(c) Comment on the results. Are they consistent with what you obtained in (a)?

17. Determine the 95% confidence interval of the expected response $\mathrm{E}\{Y\}$ when the predictor is $X = X^*$. What is the specific interval when $X^* = 27$? Calculate the mean of the responses $\{Y^*\}$ when $X^* = 27$ in the data. Does this mean fall within the confidence interval? Repeat the exercise for $X^* = 5$. Test $H_0 : \mathrm{E}\{Y^* \mid X^* = 5\} = 0$ vs. $H_1 : \mathrm{E}\{Y^* \mid X^* = 5\} > 0$ at confidence level $\alpha = 0.05$.

18. Determine the 95% prediction interval for a new response Y_p^* when the predictor is $X = X^*$. What is the specific interval when $X^* = 27$? What proportion of the responses Y_p^* fall within the prediction when $X^* = 27$? Repeat the exercise for $X^* = 5$. Are the results compatible with the notion of prediction interval? Is the observation (5.25) probable (at $\alpha = 0.05$)?

19. (continuation of the previous question)

    (a) Perform a 95% joint estimate of the parameters β_0 and β_1 Compare with the results of question 16.

    (b) Find the joint 95% Working-Hostelling confidence band for the mean response $E\{Y\}$ when $X = X^*$. Superimpose the line of best fit and the band on the scatterplot of the observations.

    (c) Find a joint 95% confidence band for the prediction of $g = 20$ new responses Y_k^* at $X = X_k^*$, $k = 1, \ldots, 20$. Superimpose the line of best fit and the band on the scatterplot of the observations.

20. (continuation of the previous question) Perform an analysis of variance to determine if the regression is significant or not.

21. (continuation of the previous question) Express the SLR $Y_i = beta_0 + beta_1 X_i + varepsilon_i$ using matrix notation. With `R`, determine the OLS solution directly (without using `lm()` or the sums $\sum X_i$, $\sum Y_i$, $\sum X_i^2$, $\sum X_i Y_i$, $\sum Y_i^2$).

22. Consider the dataset `Autos.xlsx` found on Brightspace. This time around, we are only interested in the VPAS vehicles. The predictor variables are `VKM.q` (X_1, the average daily distance driven, in km) and `Age` (X_2, the age of the vehicle, in years); the response variable is `CC.q` (Y, the average daily fuel consumption, in L). Use `R` to:

    (a) determine the design matrix $\mathbf{X}$ of the SLR model;

    (b) compute the fitted values of the response $\mathbf{Y}$ if $\boldsymbol{\beta} = (1, 5, 1)$;

    (c) compute the residual sum of squares if $\boldsymbol{\beta} = (1, 5, 1)$.

23. (continuation of the previous question) Determine directly the least squares estimator $\mathbf{b}$ of the SLR problem, using matrix manipulations in `R`. Find the estimated regression function of the response Y. Compute the residual sum of squares in the case $\boldsymbol{\beta} = \mathbf{b}$. Is this value consistent with the result obtained in part (c) of the previous question?

24. (continuation of the previous question) Using only matrix manipulations in `R`, determine the vector of residuals in the SLR problem, as well as SST, SSE, and SSR. Verify that $\mathrm{SST} = \mathrm{SSR} + \mathrm{SSE}$. What is the mean square error of the SLR model?

25. (continuation of the previous question) Assuming the SLR model is valid, test whether the regression is significant using the global F test – use `R` as you see fit (but use it!).

26. (continuation of the previous question) Find the estimated variance-covariance matrix $s^2\{\mathbf{b}\}$ for the OLS estimator $\mathbf{b}$. At a confidence level of 95%, test for

    (a) $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$;

    (b) $H_0 : \beta_2 = 0$ vs. $H_1 : \beta_2 < 0$.

27. (continuation of the previous question) We want to predict the mean response $E\{Y^*\}$ when $\mathbf{X}^* = (20, 5)$. What is the fitted value $\hat{Y}^*$ in this case? Compute a 95% C.I. for the sought quantity.

28. (continuation of the previous question) We want to predict the new response Y_p^* when $\mathbf{X}^* = (20, 5)$. Compute a 95% P.I. for Y_p^*.

29. (continuation of the previous question)

    (a) Give joint 95% C.I. for the regression parameters β_0, β_1, and β_2.

    (b) Give joint 95% C.I. for the expected mean value $E\{Y_\ell^*\}$ using the Working-Hotelling procedure for $\mathbf{X}_1^* = (50, 10), \mathbf{X}_2^* = (20, 5), \mathbf{X}_3^* = (200, 8)$.

30. (continuation of the previous question) Is the multiple linear regression model preferable to the two simple linear regression models for the same subset of `Autos.xlsx` (using X_1 or X_2, but not both)? Support your answer.

31. (continuation of the previous question) Compute the multiple coefficient of determination and the adjusted multiple coefficient of determination directly (without using `lm()`). What do these values tell you about the quality of the fit?

32. (continuation of the previous question) Is the linearity assumption reasonable? Justify your answer.

33. (continuation of the previous question) Is the assumption of constant variance reasonable? Justify your answer.

34. (continuation of the previous question) Is the assumption of independence of the error terms reasonable? Justify your answer.

35. (continuation of the previous question) Is the assumption of normality of the error terms reasonable? Justify your answer.

36. (continuation of the previous question) Overall, do you believe that the multiple linear regression model is appropriate? Justify your answer.

37. (continuation of the previous question) Use appropriate corrective measures to improve the multiple regression results.

38. (continuation of the previous question) Are the predictors in the data set multicollinear? Justify your answer.

39. (continuation of the previous question) For this question, we drop the variable `Age` from the dataset. Fit the response to a cubic regression centered on the predictor $x_1 = X_1 - \overline{X}_1$, by adding one variable at a time, to obtain $E\{Y \mid x_1\} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_1^3$. Using $\alpha = 0.05$, test for $H_0 : \beta_2 = \beta_3 = 0$ vs. $H_1 : \beta_2 \neq 0$ or $\beta_3 \neq 0$.

40. (continuation of the previous question) For this question, we re-introduce the variable `Age` to the data. Build a polynomial model of degree 2 in X_1 and X_2 that includes an interaction term (the full model) and a model that is only of degree 1 in X_1 and X_2, but still contains an interaction term (the reduced model). Determine the coefficients in both cases. Which of the two models is better?

41. Consider the dataset `Autos.xlsx` found on Brightspace. The predictor variable is `Type` (X, vehicle type); the response is `CC.q` (Y, average daily fuel consumption, in L). Using a dummy variable encoding, find the regression model of Y as a function of X. Is this a good model? Justify your answer.

42. Use the data set provided in the example for Section 4.5.

    (a) Find the solution of the WLS problem with $w_i = x_i^2$, $i = 1, \ldots, n$. Plot the results.

    (b) Find the solution of the WLS problem with the procedure described on p.37. Plot the results.

    (c) Which of the two options gives the best fit? Justify your answer.

43. Consider the dataset `Autos.xlsx` found on Brightspace. The predictor variables are `VKM.q` (X_1, average daily distance, in km), `Age` (X_2, vehicle age in years), and `Rural` (X_3, 0 for urban vehicle, 1 for rural vehicle); the response is still `CC.q` (Y, average daily fuel consumption, in L). Use the best subset approach with Mallow's C_p criterion to select the best model.

44. Repeat the previous question, but with the adjusted coefficient of determination R_a^2.

45. Repeat the previous question, but with the backward stepwise selection method and with Mallow's C_p criterion.

46. Repeat the previous question, but with the backward stepwise selection method and with the adjusted coefficient of determination R_a^2.

47. Repeat the previous question, but with the forward stepwise selection method and with Mallow's C_p criterion.

48. Repeat the previous question, but with the forward stepwise selection method and with the adjusted coefficient of determination R_a^2.

49. Consider the dataset `Autos.xlsx` found on Brightspace. The predictor variables are `VKM.q` (X_1, average daily distance, in km) and `Age` (X_2, vehicle age in years), and `Rural` (X_3, 0 for urban vehicle, 1 for rural vehicle; the response is still `CC.q` (Y, average daily fuel consumption, in L). Find the $X-$outliers in the dataset.

50. (continuation of the previous question) Consider the MLR model $\hat{y} = b_0 + b_1 X_1 + b_2 X_2$. Find the $Y-$outliers in the dataset.