STATISTICS 3340, Winter, 2019, midterm practice questions

1. A regression analysis is to be carried out for the model $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$.
   There are eight observations, with $\boldsymbol{y}' = (8, 3, 2, 1, 2, 1, 0, -1)$ and

$$\mathbf{X} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & -1 \\ 1 & 1 & -1 \\ 1 & -1 & 1 \\ 1 & -1 & 1 \\ 1 & -1 & -1 \\ 1 & -1 & -1 \end{bmatrix}$$

   Find the least squares estimator of $\boldsymbol{\beta}$. (Hint: the columns of X are orthogonal to one another.)

2. Which of the following are assumptions upon which linear regression analysis is based? (Circle True if this is an assumption of linear regression. Circle False if this is not an assumption of linear regression.)

   (a) The expected (or mean) value of the dependent variable (Y) is a linear function of the independent variables ($X_1, X_2, \ldots, X_k$). True False

   (b) All observations of the dependent variable (ie all of Y_i) have equal variance. True False

   (c) In a regression problem, if the 95% confidence interval for β_j contains 0, then the 99% confidence interval will also contain 0.
   True False Cannot tell.

   (d) The predictors ($X_1, X_2, \ldots, X_k$) are independent random variables. True False

3. Suppose that X represents the temperature of a randomly chosen October day in Halifax, and that X is a random variable with mean 12 and variance 9. Y represents the temperature on the same day in Dartmouth, and is assumed to be a random variable, with the same mean 12 and variance 9. It is assumed that the covariance between X and Y is 8.

   (a) Find the mean of $(X + Y)/2$.

   (b) Find the variance of $(X - Y)$.

   (c) Find the covariance between $X - Y$ and $(X + Y)/2$.

4. For the simple linear regression model $y = \beta_0 + \beta_1 x + \epsilon$, the following results were obtained using 20 observations.

   $\hat{\boldsymbol{\beta}} = (1.5,\ 2)'$, $SSE = 72$,

   $$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} 20 & 10 \\ 10 & 50 \end{pmatrix}$$

   and

   $$(\boldsymbol{X}'\boldsymbol{X})^{-1} = \frac{1}{900}\begin{pmatrix} 50 & -10 \\ -10 & 20 \end{pmatrix}$$

   The following table provides some R output which includes the appropriate constant(s) needed to construct confidence intervals.

pt(.975,18)	pt(.975,19)	pt(.975,20)	pt(.95,18)	pt(.95,19)	pt(.95,20)
0.8226528	0.8229795	0.823274	0.8287589	0.8290944	0.8293968

qt(.975,18)	qt(.975,19)	qt(.975,20)	qt(.95,18)	qt(.95,19)	qt(.95,20)
2.100922	2.093024	2.085963	1.734064	1.729133	1.724718

   (a) Find a 90% confidence interval for β_0.

   (b) Find a 90% confidence interval for $\beta_0 - 2\beta_1$.

5. In a study of 1006 men in four occupations, a multiple regression was carried out to show how lung function was related to age, smoking and occupation. The four occupations represented in the study were physician, firefighter, farm worker, and chemical worker. The variables in the regression were:

Y	air capacity (ml) that the worker can expire in one second
X_1	age in years
X_2	number of cigarettes smoked per day
X_3	1 if chemical worker, 0 otherwise
X_4	1 if farm worker, 0 otherwise
X_5	1 if firefighter, 0 otherwise

   The model $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + e$ was fit to the data with the following result:

   $$\hat{Y} = 4500 - 39X_1 - 9X_2 - 350X_3 - 380X_4 - 180X_5$$

   It was found that $MSE = 9$ and

   $$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 9 & 8 & 7 & 6 & 7 & 7 \\ 8 & 3 & 5 & 6 & 6 & 1 \\ 1 & 1 & 7 & 4 & 5 & 6 \\ 1 & 6 & 9 & 9 & 8 & 3 \\ 7 & 5 & 5 & 5 & 8 & 8 \\ 7 & 3 & 4 & 1 & 5 & 4 \end{bmatrix}$$

   (a) What is the mean air capacity of a 20 year old farmworker who smokes 10 cigarettes per day?

   (b) What is the estimated difference in air capacity between two farmers of the same age, one of whom smokes 10 cigarettes per day, and the other of whom does not smoke?

   (c) Construct a 95% confidence interval for β_5. [Note: In R you can get the upper 2.5'th percentile of the t distribution with m degrees of freedom, as qt(.975,m). You can also find this value in tables, although tables will NOT be provided with the midterm.]

   (d) Construct a 95% confidence interval for the mean of Y, for a 30 year old chemical worker who does not smoke.

(5) (e) Construct a 95% confidence interval for $\beta_0 + \beta_5$.

6. (a) Prove that $\mathbf{HX} = \mathbf{X}$.
   (b) For a regression which includes an intercept term, prove that $\mathbf{H1} = \mathbf{1}$. (Hint: what is the first column of the $\mathbf{X}$ matrix?)
   (c) Prove that H and $I - H$ are orthogonal matrices.
   (d) Prove that the predicted values $H\boldsymbol{y}$ and the residuals $(I - H)\boldsymbol{y}$ are uncorrelated.

7. An extension of the usual multiple regression model is the 'mixed effects' model

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{Zu} + \boldsymbol{\epsilon}$$

   The mixed effects model assumes that

   - $\boldsymbol{\epsilon}$ is a random vector with mean vector $\mathbf{0}$ and covariance matrix $\sigma^2\boldsymbol{I}$
   - $\boldsymbol{u}$ is a random vector with mean vector zero and covariance matrix $\tau^2\boldsymbol{I}$
   - $\boldsymbol{u}$ and $\boldsymbol{\epsilon}$ are independent of one another
   - $\boldsymbol{X}$, $\boldsymbol{Z}$ are matrices of known constants, and $\boldsymbol{\beta}$ is a vector of constants.

   (a) Find the expected value of $\boldsymbol{y}$.
   (b) Find the covariance matrix of $\boldsymbol{y}$.

8. A linear regression was carried out using three data points (1,4), (2,6) and (3,y_3). The least squares line was $\hat{y} = 1 + 4x$. What was the value of y_3?