STAT 4010/5010, Statistical Methods and Applications II, Exam #1 Review

The following problems should help you review for the first exam (see Canvas for exam information). Note that problems three and four require you to write R code, but on the exam, you will **not** *be required to write R code. However, I may give you R code, and ask you to interpret, the code and/or output.*

1. Please answer 'True in all cases' or 'False for at least one case'. Briefly justify your answers for full credit.

   (a) Residuals are defined as $\widehat{\varepsilon}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 x_{i,1} - ... - \widehat{\beta}_p x_{i,p}$.

   Solution: True by definition.

   (b) The Gauss-Markov Theorem states that the ordinary least squares estimator has the lowest variance of all estimates for $\boldsymbol{\beta}$.

   Solution: False...among all **unbiased** estimators for $\boldsymbol{\beta}$.

   (c) Suppose that SLR has been performed and we hope to learn something about the mean of the response Y at x_a and X_b, where x_a and x_b were not in the original data set. Suppose that x_a is farther from $\bar{x}$ than x_b. The confidence interval for x_a will be narrower than the confidence interval for x_b.

   Solution: False. The $(1-\alpha) \times 100\%$ CI for the mean response, $y^\star$, at predictor value $x^\star$, is given by

   $$\widehat{y^\star} \pm t_{\alpha/2}(n-2)\sqrt{\widehat{\sigma}^2\left(\frac{1}{n} + \frac{(x^\star - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}\right)}.$$

   This will be **wider** when the predictor value gets further away from $\bar{x}$.

   (d) In MLR, $\widehat{\beta}_i$ and $\widehat{\beta}_j$ are correlated for $i \neq j$.

   Solution: True. $Var(\widehat{\boldsymbol{\beta}}) = \sigma^2 \left(X^T X\right)^{-1}$. In Simple linear regression, $Var(\widehat{\boldsymbol{\beta}})$ is a 2×2 matrix, and the off diagonal, which respresents the covariance between the slope and intercept, is $Cov(\widehat{\beta}_0, \widehat{\beta}_1) = -\dfrac{\bar{x}\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}$ (why??). This term is not zero, in general, and thus, since the correlation can only be zero if $Cov(\widehat{\beta}_0, \widehat{\beta}_1) = 0$, in general, $\widehat{\beta}_i$ and $\widehat{\beta}_j$ are correlated for $i \neq j$.

   (e) The problem of fitting the function $y = \theta \sin(x)$, where θ is a unknown parameter to be estimated from the data $(x_1, y_1), ..., (x_n, y_n)$, is a *linear* problem (i.e., you can utilize linear regression techniques such as ordinary least

squares).

Solution: True. Let $v = \sin(x)$, and the problem becomes linear.

(f) The ordinary least squares residuals are homoscedastic. (Challenge Question!)

Solution: False. First, note that $\widehat{\boldsymbol{\varepsilon}} = \mathbf{Y} - X\widehat{\boldsymbol{\beta}} = \mathbf{Y} - H\mathbf{Y}$, where $H = X\left(X^TX\right)^{-1}X^T$ is the hat matrix.

$$\begin{aligned} Var(\widehat{\boldsymbol{\varepsilon}}) &= Var(\mathbf{Y} - H\mathbf{Y}) = Var((I_n - H)\,\mathbf{Y}) \\ &= (I_n - H)\,Var(\mathbf{Y})\,(I_n - H)^T \\ &= (I_n - H)\,\sigma^2 I_n\,(I_n - H)^T \\ &= \sigma^2\,(I_n - H)\,(I_n - H)^T \\ &\overset{why?}{=} \sigma^2\,(I_n - H) \end{aligned}$$

A simple simulation can verify that the diagonal entries of $\sigma^2\,(I_n - H)$ are not all equal. So, the ordinary least squares residuals are **heteroscedastic.**

(g) The regression line always goes through the point $(\bar{x}, \bar{y})$.

Solution: True. Note that our fitted regression line is $y = \widehat{\beta}_0 + \widehat{\beta}_1 x$. Plug in $\bar{x}$, and use the fact that $\widehat{\beta}_0 = \bar{y} - \widehat{\beta}_1\bar{x}$:
$\widehat{\beta}_0 + \widehat{\beta}_1\bar{x} = \bar{y} - \widehat{\beta}_1\bar{x} + \widehat{\beta}_1\bar{x} = \bar{y}$.

(h) The overall/full F-test (i.e., not the partial F-test) tests the hypothesis that $E(Y) = \beta_0$.

Solution: True. The reduced model in the full F-test is $Y = \beta_0 + \varepsilon$. $E(Y) = E(\beta_0 + \varepsilon) = \beta_0$.

2. You are examining the relationship between weight and height, with the hypothesis that an individual's height (x) can somewhat accurately predict the individual's weight (y). In your dataset, the height values range from 62 inches to 74 inches (mean = 68 inches, SD = 3.5 inches), and the weight values range from 114 to 200 pounds (mean = 151 pounds, SD = 27 pounds).

   (a) In R, you calculate estimates for β_0 and β_1. The fitted model is:

   $$\hat{y}_i = -222.48 + 5.48x_i,$$

   where x_i is the height of the i^{th} individual. Interpret both parameters in the model in the context of the original problem.

   Solution: -222.48 is the average weight for an individual with a height of 0 inches. (!)
   As height increases by one inch in the dataset, we can expect weight to increase by 5.48 lbs.

   (b) You calculate a second model, using $x_i' = x_i - 62$, and get the following fitted model:

   $$\hat{y}_i = 117.79 + 5.48x_i'.$$

   Interpret the intercept of this model in the context of the original problem.

   Solution: 117.79 is the average weight for an individual with a height of 62 inches (the min in the dataset).

   (c) You calculate a third model, using $x_i^* = x_i - 68$, and get the following fitted model:

   $$\hat{y}_i = 150.57 + 5.48x_i^*.$$

   Interpret the intercept of this model in the context of the original problem.
   Solution: 150.57 is the average weight for an individual of (sample) average height.

   (d) Which parameterization (i.e. model) do you prefer best and why?

   Solution: Averages are nice! Minimums are nice! Negative weights aren't :/

3. (**Solution in Jupyter Notebook**) Load the hotel data into R. Here's a description of the data:

```
Dataset:  hotel.txt

Source: Y. Xin, S. Lu, N. Zhu, W. Wu (2012). "Energy Consumption Quota of
Four and Five Star Luxury Hotels Buildings in Hainan Province, China,"
Energy and Buildings, Vol. 45, pp. 250-256.
```

```
Description: Energy Consumption and attributes for 19 Luxury Hotels.

Variables/Labels
Hotel Id   (hotel)
Energy Consumption in Kilowatt-hours   (enrgcons)
Area in square meters   (area)
Age in years     (age)
Number of guestrooms    (numrooms)
Occupancy Rate in percent   (occrate)
Effective number of guestrooms  (effrooms = numrooms*occrate/100)
```

(a) Perform simple linear regression using energy consumption as the response and area as the predictor. Report a summary of your model.

(b) Is the slope coefficient significant at the $\alpha = 0.05$ level?

(c) Interpret the slope coefficient in terms of the data.

(d) What percentage of the variability in energy consumption is explained by area?

(e) Now perform multiple linear regression with number of guest rooms, area, and age as predictors (same response). Report a summary of your model. Interpret the coefficient associated with area. Is it different from the SLR model?

(f) Plot the residuals against the fitted values (for the MLR model). Does this plot suggest that any regression assumptions are violated?

(g) Interpret the coefficient associated with the number of guest rooms. Is there anything that seems odd about this interpretation?

(h) Perform a formal test to decide whether the MLR model is necessary, or whether the SLR model will do. Interpret the results.

4. (**Solution in Jupyter Notebook**) In this question, we'll consider how R^2 might be misinterpreted.

(a) Randomly choose $n = 100$ x values between zero and one. Then, simulate Y data such that $Y_i = 1 + 2x_i + \varepsilon_i$, such that $\varepsilon_i \sim N(0, 25)$. Fit a linear regression model in R and report R^2.

(b) True or False: A simple linear regression model may be correct (i.e., it represents the way that the data were generated) but $R^2 \approx 0$.

(c) Now, use the x values from part (a), and let $y = x^2$. So, y and x are perfectly related, but not linearly. Fit a simple linear regression model R and report R^2.

(d) True or False: A simple linear regression model may be incorrect (i.e., it does not represent the way that the data were generated) but $R^2 \approx 1$.

5. Explain why the following situations might arise. There may be more than one correct answer. If you don't think that they can arise, explain why. If you think they always arise, prove it!

(a) You are given a simple linear regression model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$. The null $H_0 : \beta_1 = 0$ is not rejected based on the data used to fit this model, but the variables x and Y in the population are correlated.

Solution: Type II error!

(b) $\sum_{i=1}^n Y_i = \sum_{i=1}^n \widehat{Y}_i$ where $\widehat{Y}_i$ are the fitted values.

Solution: This always happens. Prove it!

(c) $\sum_{i=1}^n \widehat{\varepsilon}_i = \sum_{i=1}^n x_i \widehat{\varepsilon}_i$.

Solution: This always happens. Both are zero by construction of least squares.

(d) The p-value for the partial F hypothesis test is greater than 0.05.

Solution: This happens when the reduced model is sufficient (either there is no relationship between the response and predictors, are there's a type II error).

(e) The 95% confidence interval for β_1 in the model $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ covers zero but the p-value from the t-test for β_1 is less than $\alpha = 0.05$.

Solution: Impossible.

(f) You fit a simple linear regression model, and get the following output (HINT: Is there an inconsistency somewhere?).

```
Call:
lm(formula = y ~ x)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9761     0.9256   1.85    0.294
x            11.0597     1.5991   6.916 4.78e-10 ***
```

Solution: Impossible. Calculate the t-values!

Other Fun Problems (Not Required for Exam...Too Hard!)

6. In multiple linear regression, the parameter β_j $(j = 1, ..., p)$ is often called the *partial regression* parameter/coefficient because it represents the contribution of X_j to the response variable Y after it has been adjusted for the other predictor variables. This question is meant to help us understand what "adjusted for" means. Consider the "supervisor" dataset (library(RSADBE); data(SPD) in R).

   The data consist of the following variables Y = Overall rating of job being done by supervisor; X_1 = Handles employee complaints; X_2 = Does not allow special privileges; X_3 = Opportunity to learn new things; X_4 = Raises based on performance; X_5 = Too critical of poor performance; and X_6 = Rate of advancing to better jobs.

   For simplicity, let's consider just the first two predictors.

   (a) Fit the model $Y_i = \beta_0 + \beta_1 X_{i,1} + \beta_2 X_{i,2} + \varepsilon_i$ in R. Interpret the estimated parameter associated with X_2.

   (b) Now fit the simple linear regression model with Y as the response and X_1 and the predictor. Extract the residuals from this model, and call them `ryx1`.

   (c) Now fit the simple linear regression model with X_2 as the response and X_1 and the predictor. Extract the residuals from this model, and call them `rx2x1`.

   (d) Finally, fit the simple linear regression model where `ryx1` is the response and `rx2x1` is the predictor. What is the estimated slope? How is it related to an estimate in part (a)?

   (e) How does this help explain the interpretation of β_2 as the average change in Y for a one unit increase in X_2, *adjusting for* X_1?

7. Prove that $\widehat{\beta} = (X^T X)^{-1} X^T \mathbf{Y}$ is unbiased and that its variance-covariance matrix is $\sigma^2 (X^T X)^{-1}$.