Midterm Exam 1

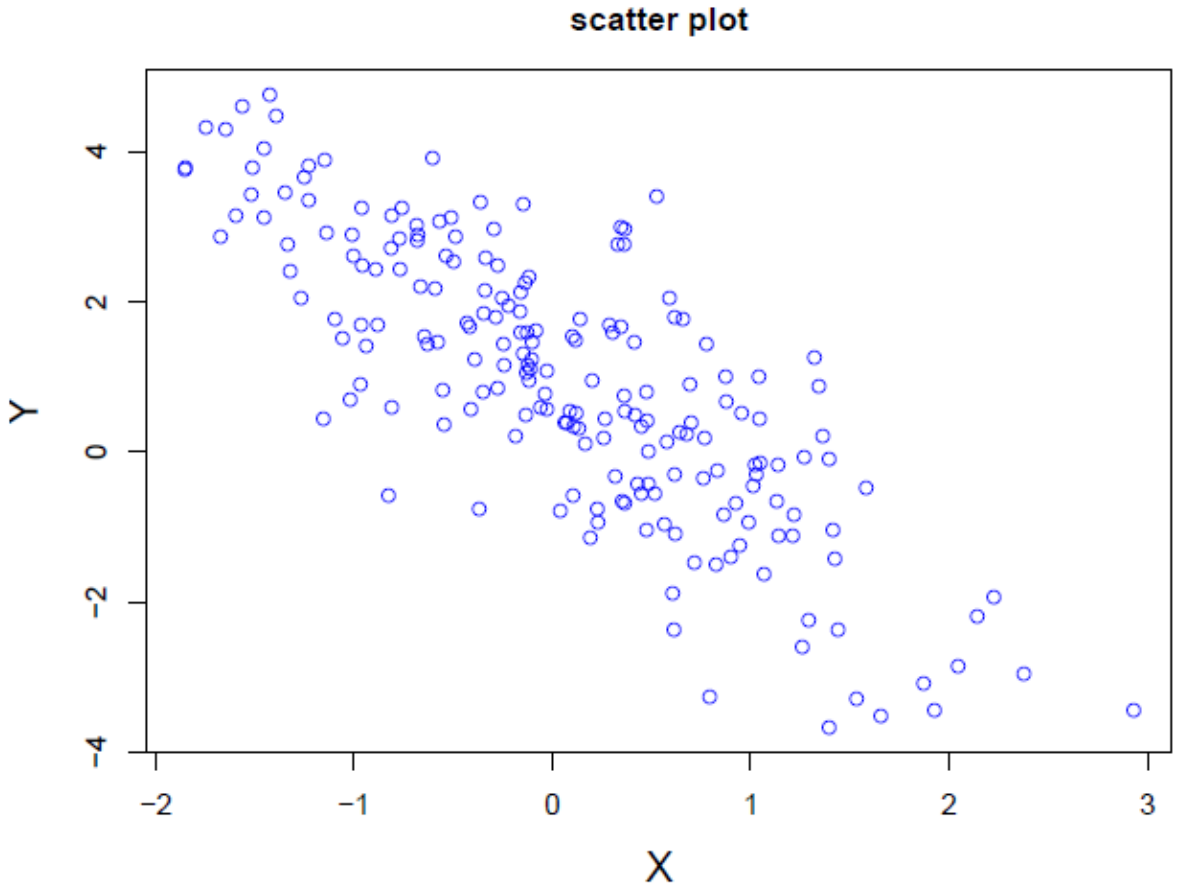
- This is an open book, open notes exam. You may refer to your problem sets and any material distributed in the course. You must work independently, however.
- Unless otherwise specified, there is only one correct answer for a multiple choice question. When you see the words “CIRCLE ALL THAT APPLY” in a multiple choice question, it implies that there are more than one correct answers for that question.

Problems 1-10. (2 points each) True/False Questions: Circle the right answer.

1. In simple linear regression $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, the t-test can be used to test $H_0 : \beta_1 = 0$ versus one-sided alternative $H_1 : \beta_1 > 0$.
   True False
2. In simple linear regression, the confidence interval for the mean value of the response variable given a specific value of the predictor is usually wider than the prediction interval for the individual response given the same value of the predictor.
   True False
3. In multiple linear regression $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \varepsilon_i$, the t-test and partial F-test for $H_0 : \beta_1 = 0$ vs $H_1 : \beta_1 \neq 0$ are equivalent (i.e., yielding the same p-values).
   True ***False***
4. In simple linear regression, the line of best fit based on least square principle maximizes the distance between the observed response variable values and the regression line.
   True ***False***
5. In multiple linear regression $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$, we usually assume that the random error ε_i follows standard normal distribution N(0, 1) in order to perform inference like t or F tests.
   True ***False***
6. In multiple linear regression $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \varepsilon_i$, if we perform an overall F-test and get p-value 0.015, we will conclude that three predictors X_1, X_2 and X_3 all have significant impact on response variable Y under significance level $\alpha = 0.05$.
   True ***False***
7. If **A** is a 4 x 5 matrix (i.e., matrix with 4 rows and 5 columns), and **B** is a 4 x 3 matrix, then **(B’A)’** is a 5 x 3 matrix.
   True ***False***
8. To test the association between a predictor, which is categorical with 2 levels, and a potential confounder, which is categorical with 3 levels, we should use a two-sample t test.
   True ***False***
9. In multiple linear regression models, adjusted R^2 increases whenever new predictors are added to the model.
   True ***False***
10. In simple linear regression $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, (i = 1, 2, …, n) as the value of R^2 becomes smaller, the prediction for Y given $X = x_0$ gets closer to the sample mean of Y.
    True ***False***
11. The proportion of the variation (or variance) in response variable Y that is explained by the least squares regression of Y on X is (CIRCLE ALL THAT APPLY)

a. ***the coefficient of determination.***
b. the Pearson correlation coefficient.
c. the intercept of the least-squares regression line.
d. the slope of the least-squares regression line.
e. ***the square of the Pearson correlation coefficient.***

Problems 12-15. Using data with n = 200 observations, we fit a simple linear regression model with Y as response variable and X as the explanatory variable. We drew the two-way scatter plot for X and Y. Also, we obtain the following SAS output.



A few items have been removed from the SAS output and replaced by clusters of x's.

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F-Value	Prob > F
Model	1	437.82	437.82	xxxx	< 0.0001
Error	xxxx	xxxx	1.19		
Total	199	673.92			

Use the above information answering problems 12 - 15.

12. What is the value of the F statistic in this analysis? What are the degrees of freedom for the numerator and denominator of this F statistic?
    F = MSR / MSE = 437.82 / 1.19 = 367.9160
    The degrees of freedom for the numerator are the number of independent variables. This model is only testing Y and X, so there is only 1 independent variable, and thus, the degree of freedom for the numerator is 1.
    $DF_{denominator}$ = (n - p - 1) = (200 - 1 - 1) = 198
13. What are the null hypothesis and alternative hypothesis for this overall F test showing the above table? What can you conclude based on the testing results?
    $H_0 : \beta_1 = 0$
    $H_a : \beta_1 \neq 0$
    Based on the testing results, the p-value is extremely small at <0.0001. Using a standard significance level of 0.05, we would have evidence to reject the null hypothesis and conclude that the variable β_1 has a statistically significant relationship with Y, explaining the variation in the response variable.
14. What is the proportion of the variability among observed values of Y that is explained by the linear regression model of Y on X?
    R^2 = SSR / SST = 437.82 / 673.92 = 0.6497 ≈ 65%
    a. 35%
    b. <u>65%</u>
    c. 42%
    d. 81%
    e. 59%
15. What is the Pearson correlation coefficient between X and Y?
    r = sqrt(SSR / SST) = sqrt(437.82 / 673.92) = sqrt(0.6497) = 0.8060 ≈ 0.81
    a. 0.59
    b. <u>0.81</u>
    c. -0.59
    d. -0.81
    e. 0.65
16. We fit a simple linear regression model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$, (i = 1, 2, …, n), on n = 15 observations, and obtained least square estimates for β_0, β_1 and σ^2 (which is the variance of error term) as $\hat{\beta}_0 = 3.50$, $\hat{\beta}_1 = 1.50$, $\hat{\sigma}^2 = 1.25$ respectively. If we consider the estimates based on the maximum likelihood principle for the same model and the same data, which of the following statements are correct? (CIRCLE ALL THAT APPLY)
    a. <u>maximum likelihood estimate for β_0 is 3.50</u>
    b. <u>maximum likelihood estimate for β_1 is 1.50</u>
    c. maximum likelihood estimate for σ^2 is 1.08
    d. <u>maximum likelihood estimate for σ^2 is 1.25</u>
    e. maximum likelihood estimate for σ^2 is 1.44

Problems 17-25. In a study for 300 respiratory disease patients, the investigators were interested in assessing the impacts of environmental and genetic factors on patients lung function, which is measured by volume of air expelled in 1 second in liters (FEV). The potential

predictors include age, sex, gene variation and smoking status. The coding sheet for these variables are as follows:

Name	Variable
FEV	Forced expiratory volume in liters in 1 second
Age	Subjects age, in years
Sex	0=Female, 1=Male
Gene variation	0 = no gene variation, 1 = presence of gene variation
Smoking	0=non-smoking, 1=light smoking, 2=heavy smoking

First, we fit main effect model with all the n = 300 observations, using age, sex, gene variation and smoking status as predictors:

$$\begin{aligned} \texttt{FEV}_i \;=\; & \beta_0 + \beta_1\,\texttt{Age}_i + \beta_2\,\texttt{Sex}_i + \beta_3\,\texttt{GeneVariation}_i \\ & + \beta_4\,I(\texttt{Smoking}_i = 1) + \beta_5\,I(\texttt{Smoking}_i = 2) + \varepsilon_i, \quad i = 1, 2, ..., n, \end{aligned} \tag{1}$$

where I(Smoking$_i$ = 1) and I(Smoking$_i$ = 2) are indicators for light and heavy smoking respectively, with non-smoking as the reference level. We obtain the following SAS output for model (1). A few items have been removed and replaced by clusters of x's.

The REG Procedure: Model (1)
Dependent Variable: FEV
Number of Observations Used 300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	xxxx	xxxx	xxxx	97.02	< .0001
Error	xxxx	xxxx	0.37369		
Total	299	291.13370			

sample size n=300, $R^2 = 0.6226$

regression parameter estimates:

$\hat{\beta}_1 = -0.053,\ \hat{\beta}_2 = 0.673,\ \hat{\beta}_3 = -0.366,\ \hat{\beta}_4 = -0.419,\ \hat{\beta}_5 = -0.967$

17. Complete the 5 missing numbers in the Analysis of Variance Table.
    Model DF = number of predictor variables = 5

Error DF = DF_{total} - DF_{model} = 299 - 5 = 294
SSR = R^2 * SST = 0.6226 * 291.13370 = 181.25984
SSE = SST - SSR = 291.13370 - 181.25984 = 109.87386
MSR = SSR / DF_{model} = 181.25984 / 5 = 36.25197

18. Interpret the regression parameter estimates for Age and heavy smoking, respectively.
    $\hat{\beta}_1$ = -0.053
    $\hat{\beta}_5$ = -0.967
    $\hat{\beta}_1$ represents the change in Y for every 1 year increase in patient Age. The value of the parameter estimate means that for every 1 year increase in patient Age, the FEV is expected to reduce by 0.053 liters. $\hat{\beta}_5$ represents the change in Y between the reference level (non-smoking) and heavy smoking status. The parameter estimate means that heavy smoking status results in an expected decrease of FEV by 0.967 liters.

Then, we fit a multiple linear regression model by adding interactions between smoking levels and gene variation into the regression model (1). We call this regression model with interactions "model (2)". We obtain the following SAS output for model (2). A few items have been removed from the SAS output and replaced by clusters of x's.

The REG Procedure: Model (2)
Dependent Variable: FEV
Number of Observations Used 300

Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Prob > F
Model	xxxx	xxxx	xxxx	72.16	< .0001
Error	xxxx	106.64543	xxxx		
Total	299	291.13370			

19. Write out the expression for the regression model (2)? (Don't use matrix form.)
    $FEV_i = \beta_0 + \beta_1 Age_i + \beta_2 Sex_i + \beta_3 GeneVariation_i + \beta_4 I(Smoking_i = 1) + \beta_5 I(Smoking_i = 2) + \beta_6 GeneVariation*I(Smoking_i = 1) + \beta_7 GeneVariation*I(Smoking_i = 2) + \varepsilon_i$, i = 1, 2, …, n
20. Using the information provided for model (1) and model (2), we can perform one hypothesis test to determine whether, collectively, the interaction term(s) contribute to the variability in FEV significantly given that Age, Sex, Gene variation and Smoking levels are retained in the model. What hypothesis test will we use? Write out the null hypothesis H_0 and alternative hypothesis H_1? (note: if there are more than one interaction terms, test them collectively as a group.)
    The hypothesis test we will use is an overall F-test to compare variation from the interaction terms in model 2 to the variation from model 1's interaction terms.
    $H_0 : \beta_6 = \beta_7 = 0$
    H_1 : at least one of $\beta_6, \beta_7 \neq 0$
21. What's the value of the test statistic in problem 20? What are the degrees of freedom of this test statistic? (Don't need to calculate p-value, which is < 0.05.)

SSR = SST - SSE = 291.13370 - 106.64543 = 184.48827
Error DF = DF_{total} - DF_{model} = 299 - 7 = 292
F = [(SSR_2 - SSR_1) / (DF_{model2} - DF_{model1})] / [SSE_2 / DF_{error2}]
F = ([184.48827 - 181.25984] / [7 - 5]) / [106.64543 / 292]
F = (3.22843 / 2) / 0.36522
F = 4.4198
DF_{num} = DF_{model2} - DF_{model1} = 7 - 5 = 2
DF_{den} = DF_{error2} = 292
The test statistic is 4.4198. The DF of the numerator is 2. The DF of the denominator is 292.

22. Fitting the interaction model, we get the parameter estimate $\hat{\beta}_7$ = -0.663 for the interaction between gene variation and heavy smoking, and its standard error as s.e.($\hat{\beta}_7$) = 0.242. We use a t test for $H_0 : \beta_7 = 0$ versus $H_1 : \beta_7 \neq 0$. What's the value of t-test statistics and its degree of freedom? (Don't need to calculate the p-value, which is < 0.01.)
    $t = \hat{\beta}_7 - \mu$ / s.e.($\hat{\beta}_7$) = (-0.663 - 0) / 0.242 = -2.7397. The degrees of freedom are based on the DF_{error}, or 292. So the test statistic is -2.7397 with 292 degrees of freedom.
23. The parameter estimate for the interaction term between gene variation and heavy smoking is $\hat{\beta}_7$ = -0.663, as given in the previous problem. Interpret regression parameter estimate $\hat{\beta}_7$.
    β_7 represents the interaction between gene variation and heavy smoking. This indicates that if the patient has a gene variation and is a heavy smoker compared to the reference level of a non-smoker, there can be an expected decrease in FEV of 0.663 liters.
24. Under the interaction model (2), we want to test whether response variable FEV is significantly associated with Gene variation among heavy smoking people, write out the null and alternative hypotheses.
    $H_0 : \beta_7 = 0$
    $H_1 : \beta_7 \neq 0$
25. Under the interaction model (2), we want to test whether there is significant difference between the expected change of FEV associated with gene variation among light smoking people and the expected change of FEV associated with gene variation among heavy smoking people, and write out the null and alternative hypotheses.
    $H_0 : \beta_6 - \beta_7 = 0$
    $H_1 : \beta_6 - \beta_7 \neq 0$
26. Let $Y = X\beta + \varepsilon$ be a matrix representation of a multiple linear regression model with 10 predictors and sample size n = 100 observations. Regression parameter vector $\beta = (\beta_0, \beta_1, \ldots, \beta_{10})'$, and the hat matrix $H = X(X'X)^{-1}X'$. Which of the following statements are correct? (CIRCLE ALL THAT APPLY)
    a. ***the matrix X has 10 columns***
    b. the matrix Y has 10 rows
    c. ***the hat matrix H is a square matrix***
    d. the hat matrix H is a diagonal matrix
    e. ***the hat matrix H is a symmetric matrix***