Homework 4 / Final
Due Friday, July 15

The first four questions concern a study of the association between smoking and blood pressure. Researchers are interested both in

* whether there is a relationship between the amount of smoking and blood pressure, and

* the role of a specific gene in the relationship.

The researchers believe that positive gene expression (GENE=+) strengthens risk factor effects on blood pressure.

Two hundred participants are randomly selected from a health management organization. Blood pressure, smoking history and gene expression status are recorded:

Outcome variable: Systolic blood pressure (mg/m^2)

Smoking history: Coded as two dummy variables, SMOKF & SMOKC. **SMOKF = 1** if person was a **former smoker** but has not smoked within the last year, and 0 otherwise. **SMOKC = 1** if person is a **current smoker** (or recently quit–within the last year), and 0 otherwise. The **reference group** is persons who "**never smoked**."

Genetic status: Coded as a dummy variable GYES = 1 if GENE=+ and 0 otherwise.

1. Researchers consider whether genetic status **confounds** the relationship between smoking history and blood pressure. Which of the following choices best describes what it means for genetic status to confound the relationship between smoking history and blood pressure? Choose only one answer:

a) The difference in mean blood pressure between smokers and non-smokers is **not** the same for those with GENE=+ as for those with GENE=-.

b) The difference in mean blood pressure between smokers and non-smokers **is** the same for those with GENE=+ as for those with GENE=-.

c) The overall difference in mean blood pressure between smokers and non-smokers is **not** the same as the difference comparing persons with the same genetic status.

d) The overall difference in mean blood pressure between smokers and non-smokers **is** the same as the difference comparing persons with the same genetic status.

e) Smoking status and genetic status are related to one another.

2. The researchers fit a model with both main effects and all possible interactions. Here are the coefficient estimates and their standard errors:

Coefficient	Estimate	S.E.	Parameter
Intercept	125.8	20.9	β_0
GYES	10.6	6.1	β_1
SMOKF	3.2	3.7	β_2
SMOKC	4.3	4.2	β_3
SMOKF*GYES	2.3	5.2	β_4
SMOKC*GYES	18.3	7.5	β_5

Which of the following choices best interprets the current smoking-by-positive gene expression coefficient (SMOKC*GYES, that is, the number 18.3? We estimate that .... : Choose only one answer.

a) Mean blood pressure is 18.3 points higher in current smokers than in never smokers.

b) The difference in mean blood pressure comparing current smokers and never smokers is 18.3 points larger in those with GENE=+ than GENE=-.

c) Mean blood pressure is 18.3 points higher in current smokers than in former smokers.

d) The difference in mean blood pressure comparing current smokers and former smokers is 18.3 points larger in those with GENE=- than GENE=+.

e) The difference in mean blood pressure between current smokers and never smokers comparing persons with the same genetic status is 18.3 points higher than the overall difference.

3. The researchers theorize that **any association between smoking and systolic blood pressure is limited to those with positive gene expression** (no association if negative gene expression). Which of the following choices is a hypothesis that correctly states the researchers' theory in terms of parameters of the above model? Choose only one answer.

a) H_0: $\beta_4 = \beta_5 = 0$

b) H_0: $\beta_2 = \beta_3 = 0$

c) H_0: $\beta_1 = 0$

d) H_0: $\beta_1 = \beta_4 = \beta_5 = 0$

e) H_0: $\beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$

4. Suppose the researchers aim to test $H_0: \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$. Which of the following accomplishes their aim? Choose only one answer.

a) Create 95% confidence interval for β_1 and see whether it includes 0 or not.

b) Fit 4 simple linear regressions, one with each covariate corresponding to β_2, β_3, β_4, β_5, create a 95% confidence interval for each of the 4 parameters, and see whether any of them excludes 0.

c) Compare the sum of squares for error (SSE) for the model described in question 2 and a model that only includes GYES as a covariate using a F-test.

d) Compare the sum of squares for error (SSE) for the model described in question 2 and a model that includes all the five covariates except GYES.

e) Evaluate whether R^2 differs between the model described in question 2 and a model that only includes GYES as a covariate .

The last four questions concern a study of 50 head-injured children. Severity of injury is measured as

$$X_i = \text{lesion (injury) volume in brain in } mm^3$$

The children are asked to perform a series of tasks. The tasks have equal difficulty, and testing continues until the first time a task is failed. As response, researchers record

$$Y_i = \text{number of tasks attempted by child i.}$$

Y_i is distributed as a geometric random variable with mass function, mean and variance

$$\Pr\{Y_i = y\} = (1-p_i)^{y-1}p_i;\ E[Y_i] = 1/p_i;\ \text{Var}[Y_i] = (1-p_i)/p_i^2,$$

$p_i = \Pr\{\text{task failure}|\text{child i}\}$. The goal is to describe p_i as a function of the single covariate X_i using a generalized linear model.

5. Assuming the link function $g(m) = 1/m$ and a linear predictor in X, write the systematic part of the model (the regression equation).

6. Write the likelihood function for θ = (intercept,slope) explicitly in terms of θ.

The researchers fit their model and find

	Coef	SE
Intercept	0.096	0.030
lesion volume	0.027	0.013

7. Which of the following best characterizes the interpretation of β_1 (the lesion volume coefficient)? Choose only one answer:

a) The log odds of task failure.

b) The log odds ratio for association between lesion volume and task failure.

c) The mean number of tasks tried per task failure.

d) The difference in the mean number of tasks tried per task failure comparing children whose lesion volumes differ by 1 mm^3.

e) The difference in the probability of task failure comparing children whose lesion volumes differ by 1 mm^3.

8. Suppose 10 tasks are presented without stopping testing (testing continues through all 10 tries even if there are task failures.) Using the data above provide a 95% confidence interval for the expected number of tasks failed in the 10 task tries.

7. The maximum likelihood estimate is obtained by which method? Choose only one answer.

a) Set the likelihood function equal to 0 and solve for θ.

b) It equals the maximum value of the likelihood function.

c) It equals that value of θ at which the likelihood function achieves its maximum value.

d) It equals that value of θ at which the score function achieves its maximum value.

e) It equals