1. [10 marks] **Circle the letter corresponding to your choice.**

(a) Which statement is **FALSE**?

$\boxed{\text{A:}}$ For Poisson data the likelihood ratio statistic is a continuous random variable.

B: The distribution of the likelihood ratio statistic based on a random sample $Y_1, Y_2, \ldots, Y_n$ is approximately $\chi^2(1)$ for large n.

C: For Exponential data, the likelihood ratio statistic is a continuous random variable.

D: For Binomial(n, θ) data, an approximate 95% confidence interval for θ based on the asymptotic Normal pivotal quantity can contain values outside the interval $[0, 1]$.

E: For Exponential(θ) data, an approximate 95% confidence interval for θ based on a 15% likelihood interval only contains values of θ greater than zero.

(b) Which of the following statements is **TRUE**?

$\boxed{\text{A:}}$ For the $G(\mu, \sigma)$ model with σ known the $p - value$ for the likelihood ratio test of $H_0 : \mu = \mu_0$ is exact.

B: The $p - value$ obtained using the likelihood ratio test statistic is the same $p - value$ obtained using the test based on the asymptotic Normal pivotal quantity.

C: The likelihood ratio test statistic can only be used for Exponential, Binomial and Poisson models.

D: The observed value of the likelihood ratio test statistic is always between 0 and 1.

(c) Let $y_1, y_2, \ldots, y_{40}$ be a random sample from an Exponential(θ) distribution. Suppose $[1, 3.5]$ is a 10% likelihood interval for the unknown parameter θ. If we use the likelihood ratio test statistic to test $H_0 : \theta = 4$, then we can conclude

A: the approximate $p - value$ is larger than 0.1.

$\boxed{\text{B:}}$ the approximate $p - value$ is smaller than 0.03.

C: the approximate $p - value$ is larger than 0.03.

D: nothing about the $p - value$ because it is not related to the likelihood interval.

(d) Suppose that a data set $y_1, y_2, \ldots, y_{25}$ is assumed to be an observed random sample from a $G(\mu, \sigma)$ distribution where μ and σ are unknown. Suppose also that the data set is stored in the variable y and that the command

```
t.test(y,0,conf.level=0.95)
```

has been run in R and the following output obtained:

```
One Sample t-test
data:  y
t = 3.3621, df = 24, p-value = 0.002587
alternative hypothesis:  true mean is not equal to 0
95 percent confidence interval:
0.4553305 1.9030695
sample estimates:
mean of x
1.1792
```

Based on this information the sample standard deviation to 3 decimal places is equal to

A: 3.362

$\boxed{\text{B:}}$ 1.754

C: 3.075

D: Not enough information to determine.

(e) In the simple linear regression model, which of the following random variables does NOT have a Gaussian distribution?

A: $\tilde{\alpha}$

B: $\tilde{\beta}$

C: $\frac{\tilde{\beta}-\beta}{\sigma/\sqrt{S_{XX}}}$

$\boxed{\text{D:}}$ S_e^2

E: $\bar{Y}$

(f) Suppose for data (x_i, y_i), $i = 1, 2, \ldots, n$ we assume the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, n$ independently. Which statement is **FALSE**?

A: $\sum_{i=1}^{n} \left(y_i - \hat{\alpha} - \hat{\beta} x_i\right) = 0$

B: $S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})^2$

C: $S_{XX} = \sum_{i=1}^{n} (x_i - \bar{x})\, x_i$

D: $\hat{\beta} = S_{XY}/S_{XX}$

$\boxed{\text{E:}}$ The least squares estimate of α and β, and the maximum likelihood estimates of α and β both minimize the function $g(\alpha, \beta) = \sum_{i=1}^{n} |y_i - \alpha - \beta x_i|$.

(g) Suppose for data (x_i, y_i), $i = 1, 2, \ldots, n$ we assume the model $Y_i \sim G(\alpha + \beta x_i, \sigma)$, $i = 1, 2, \ldots, n$ independently. Which statement is **FALSE**?

A: The parameter σ represents the variability in the response variate in the study population for each value of the explanatory variate x.

B: The parameter β represents the change in the mean of the response variate in the study population for a one unit increase in the explanatory variate.

$\boxed{\text{C:}}$ The parameter α represents the intercept of the least squares line.

D: The parameter $\mu(x) = \alpha + \beta x$ represents the mean response in the study population for units with explanatory variate equal to x.

(h) Which of the following statements about the simple linear regression model is **TRUE**?

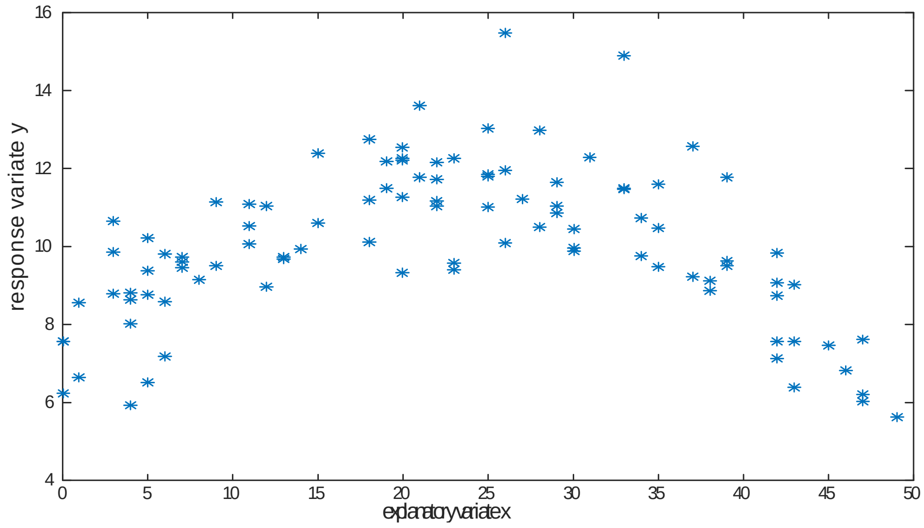
A: If the $p-value$ for testing $H_0 : \beta = \beta_0$ is less than 0.001 then we can conclude that there is a linear relationship between the explanatory variate x and the response variate Y.

$\boxed{\text{B:}}$ The explanatory variates $x_1, x_2, \ldots, x_n$ should be considered as known constants.

C: The relationship between the sample correlation r and the least squares estimate of the slope $\hat{\beta}$ is $r = \hat{\beta}\,(S_{yy}/S_{xx})^{1/2}$.

D: S_e is the maximum likelihood estimator of σ.

(i) The scatter plot for data (x_i, y_i), $i = 1, 2, \ldots, 100$ is



Based on this plot we would conclude that:

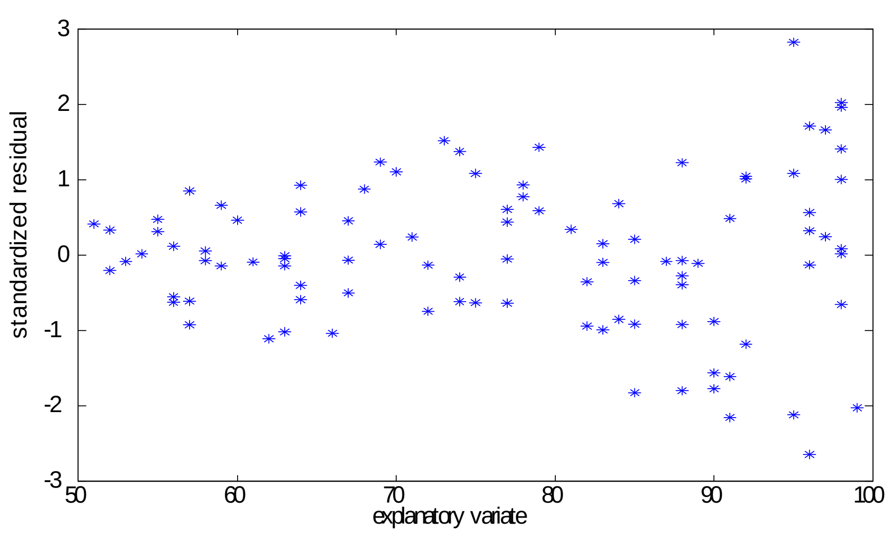
A: the simple linear regression model is an appropriate model for these data.

$\boxed{\textbf{B:}}$ the simple linear regression model is not an appropriate model for these data because the assumption that the mean of the response variate is a linear function of the explanatory variate does not hold.

C: the simple linear regression model is not an appropriate model for these data because the sample size is too small.

D: the simple linear regression model is not an appropriate model for these data because the assumption of constant standard deviation does not hold.

(j) Suppose the simple linear regression model has been fit to the data (x_i, y_i), $i = 1, 2, \ldots, 100$. The standardized residual plot $(x_i, \hat{r}_i^*)$, $i = 1, 2, \ldots, 100$ with $\hat{r}_i^* = \left(y_i - \hat{\alpha} - \hat{\beta}x_i\right)/s_e$ for these data is:



Based on this plot we would conclude that:

$\boxed{\text{A:}}$ the simple linear regression model is not an appropriate model for these data because the assumption of constant standard deviation does not hold.

B: the simple linear regression model is not an appropriate model for these data because the assumption that the mean of the response variate is a linear function of the explanatory variate does not hold.

C: the simple linear regression model is not an appropriate model for these data because the Gaussian distribution assumption for the response variate does not hold

D: the simple linear regression model is an appropriate model for these data.

2. [11 marks] Suppose $y_1, y_2, \ldots, y_n$ is an observed random sample from the distribution with probability function

$$f(y;\theta) = \binom{y+2}{y}(1-\theta)^3\theta^y \quad \text{for } y = 0, 1, \ldots; \quad \theta \in (0,1)$$

where θ is an unknown parameter.

(a) [4] Find the maximum likelihood estimate $\hat{\theta}$ for θ. Show your steps clearly.

The likelihood function, after dropping the constants, is

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n}(1-\theta)^3\theta^{y_i} = (1-\theta)^{3n}\theta^{\sum\limits_{i=1}^{n} y_i} \\ &= (1-\theta)^{3n}\theta^{n\bar{y}} \quad \text{for} \quad \theta \in (0,1) \end{aligned}$$

The log likelihood function is

$$l(\theta) = 3n\log(1-\theta) + (n\bar{y})\log\theta \quad \text{for} \quad \theta \in (0,1)$$

Since

$$\begin{aligned} \frac{dl(\theta)}{d\theta} &= \frac{-3n}{1-\theta} + \frac{n\bar{y}}{\theta} = \frac{n}{\theta(1-\theta)}\left[-3\theta + \bar{y}(1-\theta)\right] \\ &= \frac{n}{\theta(1-\theta)}\left[-\theta(3+\bar{y}) + \bar{y}\right] = 0 \end{aligned}$$

if

$$\theta = \frac{\bar{y}}{3+\bar{y}}$$

therefore the maximum likelihood estimate of θ is

$$\hat{\theta} = \frac{\bar{y}}{3+\bar{y}}$$

(b) [2] If $n = 30$ and the observed sample mean is $\bar{y} = 3.9$, show that $R(0.5) = 0.171$ where $R(\theta)$ is the relative likelihood function.

The relative likelihood function is

$$\begin{aligned} R(\theta) &= \frac{L(\theta)}{L(\hat{\theta})} = \left(\frac{1-\theta}{1-\hat{\theta}}\right)^{3n}\left(\frac{\theta}{\hat{\theta}}\right)^{n\bar{y}} \\ &= \left[\frac{(1-\theta)(3+\bar{y})}{3}\right]^{3n}\left[\frac{\theta(3+\bar{y})}{\bar{y}}\right]^{n\bar{y}} \end{aligned}$$

Since $n = 30$, $\bar{y} = 3.9$,

$$\begin{aligned} R(0.5) &= \left[\frac{(1-0.5)(3+3.9)}{3}\right]^{90}\left[\frac{0.5(3+3.9)}{3.9}\right]^{117} \\ &= 0.171 \end{aligned}$$

(c) [5] Given that $n = 30$ and $R(0.5) = 0.171$, use the likelihood ratio test statistic to test the null hypothesis $H_0 : \theta = 0.5$. Show your work. Write your final numerical answers to 3 decimal places in the space provided.

(i) [1] The observed value of the likelihood ratio test statistic is ______3.532______ .

$$\lambda(0.5) = -2\log(R(0.5)) = -2\log(0.171) = 3.532$$

(ii) [2] The approximate $p - value$, using the Normal table, is ______0.06______.

$$p - value = 2\left[1 - P(Z \leq \sqrt{3.532})\right] = 2(1 - 0.96995) = 0.060 = 0.06$$

where $Z \sim N(0,1)$.

(iii) [2] State your conclusion regarding the hypothesis $H_0 : \theta = 0.5$ in a sentence.

Since $0.05 < p - value < 0.1$, we conclude that there is weak evidence (some evidence) based on the data against the null hypothesis $H_0 : \theta = 0.5$.

Note: The $p - value$ must be referred to in the conclusion.

3. [9 marks] Suppose the data set $x_1, x_2, \ldots, x_{30}$ are stored in the vector x and the data set $y_1, y_2, \ldots, y_{30}$ are stored in the vector y in R. These data are to be analyzed using the simple linear regression model

$$Y_i \sim G(\alpha + \beta x_i, \sigma) \quad i = 1, 2, \ldots, 30 \text{ independently}$$

where α, β, σ are unknown parameters and the x_i's are known constants.

The following code was run in R:

```
RegModel<-lm(y~x)
summary(RegModel)
```

The output obtained was:

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

Residuals:
Min      1Q      Median   3Q     Max
-7.8891 -4.5914 -0.5018 4.7169 11.3140

Coefficients:
              Estimate  Std.  Error  t value     Pr(>|t|)
 (Intercept)   3.3816     1.8718      1.807      0.081582
      x       -0.6947     0.1680     -4.134    0.000293 ***

Signif.  codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.'  0.1 ' ' 1

Residual standard error:  5.707 on 28 degrees of freedom
Multiple R-squared:  0.379,   Adjusted R-squared:  0.3568
F-statistic:  17.09 on 1 and 28 DF, p-value:  0.0002931
```

Answer the following questions based on this information. Use all the decimals given in the output.

(a) [1] The least squares estimate of β is $\underline{\quad -0.6947 \quad}$.

(b) [1] The maximum likelihood estimate of α is $\underline{\quad 3.3816 \quad}$.

(c) [1] The equation of the fitted least squares line is $\underline{\quad y = 3.3816 - 0.6947x \quad}$.

(d) [1] The value of the test statistic for testing $H_0 : \beta = 0$ is equal to $\underline{\quad -4.134 \quad}$.

(e) [1] The $p - value$ for testing $H_0 : \beta = 0$ is equal to $\underline{\quad 0.000293 \text{ or } 0.0002931 \quad}$.

(f) [2] State your conclusion with justification regarding the hypothesis $H_0 : \beta = 0$ in a sentence.

Since $p - value = 0.000293 < 0.001$, we conclude that there is very strong evidence based on the data against the null hypothesis $H_0 : \beta = 0$.

Note: The $p - value$ must be referred to in the conclusion.

(g) [2] The following additional code was run:

```
xbar<-mean(x)
Sxx<-(n-1)*var(x)
se<-summary(RegModel)$sigma
cat("xbar = ", xbar,", Sxx = ", Sxx, ", se = ", se)
```

The output obtained was:

```
xbar = 9.253333 , Sxx = 1153.275 , se = 5.70683
```

Based on this information and the information from the output on the previous page determine a 95% prediction interval for a response at $x = 2$ is (show your work).

$[-10.150, 14.135]$.

From t tables $P(T \leq 2.0484) = 0.975$ where $T \backsim t(28)$.

Predicted value for $x = 2$ is $\hat{\mu}(2) = 3.3816 - 0.6947(2) = 1.9922$.

The 95% prediction interval for a response at $x = 2$ is

$$
\begin{aligned}
& 3.3816 - 0.6947(2) \pm 2.0484(5.70683)\sqrt{1 + \frac{1}{30} + \frac{(2 - 9.253333)^2}{1153.275}} \\
= \; & 1.9922 \pm 12.1426 \\
= \; & [-10.1504, 14.1348]
\end{aligned}
$$