2016 COMPREHENSIVE WRITTEN EXAMINATION

STOR 664 Questions

Problem A (38 points)

The *gala* dataset contains information on species diversity on the Galapagos Islands. The relationship between the number of plant species and several geographic variables is of interest.

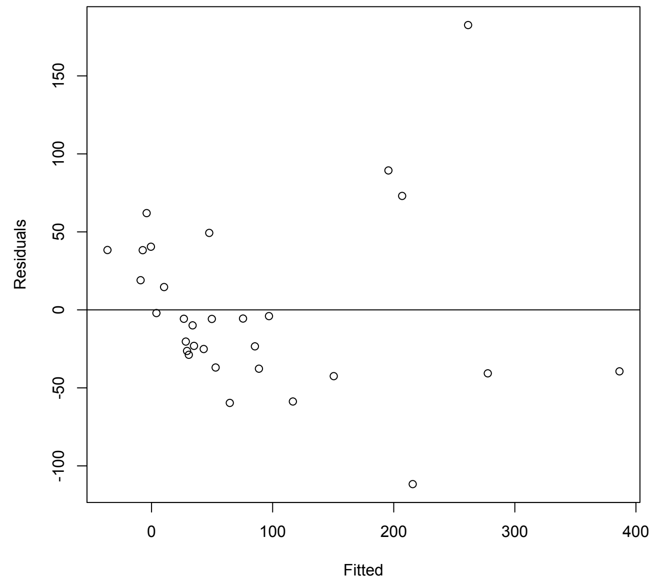
The edited R output and other summaries are given below.

```
Call:
lm(formula = Species ~ Area + Elevation + Scruz + Nearest + Adjacent,
    data = gala)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.068221  19.154198   0.369 0.715351
Area        -0.023938   0.022422  -1.068 0.296318
Elevation    0.319465   0.053663   5.953 3.82e-06 ***
Scruz       -0.240524   0.215402  -1.117 0.275208
Nearest      0.009144   1.054136   0.009 0.993151
Adjacent    -0.074805   0.017700  -4.226 0.000297 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 60.98 on 24 degrees of freedom
Multiple R-squared:  0.7658,  Adjusted R-squared:  ?
F-statistic:  ? on ? and ? DF,  p-value: 6.838e-07

>plot(fitted(lmod1),residuals(lmod1),xlab="Fitted",ylab="Residuals")
>abline(h=0)
```

(a) (8 points) Based on the R output, write out the model fitted with clear notations. Give the two types of assumptions commonly used in linear models and explain their implications.

(b) (12 points) What is the sample size n for this study? Calculate SSE, SSR, SSTO, and the adjusted R-squared for this model.

(c) (10 points) Calculate the missing F-statistic and the corresponding degrees of freedom. Write out the hypothesis test corresponding to F-statistic in the output. What conclusion can you draw from the output about the test?



(d) (8 points) Comment on the plot above on whether the model assumption is reasonable and discuss possible remedy if necessary.

Problem B (62 points)

Consider the general linear model

$$Y = X_1\beta_1 + X_2\beta_2 + \epsilon; \qquad \epsilon \sim N(0, \sigma^2 I_n), \tag{1}$$

where Y and ϵ are both $n \times 1$, β_1 and β_2 are respectively $p_1 \times 1$ and $p_2 \times 1$ vectors of unknown parameters, and $X = (X_1, X_2)$ is specified and of full rank $(p_1 + p_2)$. The variance σ^2 of the observations is unknown. Assume (1) is the true model and answer the following questions:

1. (32 points) Suppose one ignores or is unaware of the X_2 covariates and fits the following model

$$Y = X_1\beta_1 + \epsilon; \qquad \epsilon \sim N(0, \sigma^2 I_n). \tag{2}$$

   (a) (10 points) Calculate the standard least squares estimator $\tilde{\beta}_1$ based on this model, and demonstrate whether $\tilde{\beta}_1$ an unbiased estimator of β_1. If not, what is the bias?

   (b) (7 points) Let $\tilde{Y} = X_1\tilde{\beta}_1 = H_1 Y$. Express H_1 using the given observations. Without the need of justification, discuss properties of the matrix H_1.

   (c) (15 points) Suppose the first column of X_1 is a constant vector of 1, and the other columns include covariates without standardization. Let X_1^* represent the standardized version of X_1, i.e., for each of the columns 2 to p_1 of X_1, by subtracting its column mean and dividing by its column standard deviation to construct X_1^*.

   Let the corresponding least squares solution $\tilde{Y}_1^* = H_1^* Y$ by using X_1^*. Discuss the relationship between H_1 and H_1^*. Use matrix algebra to formally prove your claim.

2. (30 points) Suppose one uses the correct model, i.e., model (1). Denote the corresponding least squares estimate of the parameters (β_1, β_2) as $(\hat{\beta}_1, \hat{\beta}_2)$.

   (a) (15 points) Let $A_1 = I_n - H_1$ with I_n a standard identity matrix and H_1 is defined as in Part 1. Let $R_1 = A_1 Y$ and $Z = A_1 X_2$. Fit a linear model $R_1 = Z\gamma + \epsilon$. Derive the least squares estimate $\hat{\gamma}$ for γ and discuss its relationship with $\hat{\beta}_2$. Prove your claim.

   (b) (15 points) Let $R_2 = Y - X_1\hat{\beta}$. Fit a linear model $R_2 = X_2\alpha + \epsilon$. Derive the least squares estimate $\hat{\alpha}$ for α and discuss its relationship with $\hat{\beta}_2$. Prove your claim.