STAT7630 - HW 2

D. Pluta

2024-02-08

Due: February 22nd, 2024 by 11:59pm

Submission: AU Course Website

Homework 2

STAT 7630, SPRING 2024

1. *Simple Linear Regression* The data set in 'fev.txt' is from an observational study of factors potentially influencing respiratory health as measured by functional effective volume (FEV). Specific interest is in the difference of individuals who smoke versus those who do not.
    a. Provide a histogram of the response variable. Use informative labels for the plot axes. Does this plot suggest the data is appropriate for application of simple linear regression? Explain.
    b. Provide a boxplot of FEV by Smoking status, with informative axis labels. Does this plot tell us anything about whether simple linear regression is an appropriate model?
    c. Fit a linear regression model to determine if smoking is significantly associated with FEV. Include a nicely formatted table of the regression results with columns for the coefficient name, point estimate, standard error, and P-value.
    d. Find the 95% confidence interval for the effect of smoking on FEV.
    e. Give a practical interpretation of the confidence interval you found in (c).
    f. Based on the model estimates, give 95% confidence intervals for the average FEV among the smoking group, and among the nonsmoking group.
    g. Produce the 'residuals by fitted' scatterplot from the model you fit. From this plot, does the data appear to violate any of the four assumptions of linear regression? Explain.
    h. Produce the QQ-plot of the standardized residuals. Do the data appear to satisfy or violate any of the assumptions of linear regression based on this plot? Explain.
2. *Multiple Linear Regression* Continuing to work with the FEV data set, we will now consider the effect of smoking on FEV when adjusting for height.
    a. Generate a scatterplot of height versus FEV. Use informative axis labels.
    b. To see the marginal effect of height, fit the simple linear regression model with FEV as the response, and height as the predictor. Provide a nicely formatted table of the regression results, similar to Problem 1.
    c. For each of the four assumptions of linear regression, assess whether the data appear to satisfy or violate the assumption. Include any plots you used.
    d. Produce a boxplot of height by smoking status. Do you observe anything from this plot that may be relevant to the regression results?
    e. Fit a linear regression model to estimate the effect of smoking on FEV adjusting for height.

Give a formal (i.e., in mathematical notation) specification of the model you fit. Include a table of the regression results.

f. Produce a 95% confidence interval for the effect of smoking on FEV, adjusted for height.

g. Interpret this confidence interval in the application context, and discuss how this result compares to the effect of smoking you found in Problem 1.

h. Compute the 95% CI for the expected difference when comparing the FEV of smokers who are 1 inch taller than nonsmokers.

3. Consider the simple linear regression model, and let $\hat{Y}_h = \hat{\beta}_0 + \hat{\beta}_1 X_h$ be the MLE of the mean response for covariate value X_h.

a. Show $\hat{Y}_h$ is unbiased for $\mathbb{E}[Y|X = X_h]$.

b. Write out the formulas for the variance of $\hat{\beta}_0$ and $\hat{\beta}_1$, and $\mathrm{Cov}(\bar{Y}, \hat{\beta}_1)$. Use these formulas to calculate the variance of $\hat{Y}_h$.

c. What is the distribution of $\hat{Y}_h$?

d. How does the uncertainty about our estimate of $\hat{Y}_h$ change as the difference between X_h and $\bar{X}$ increases?

e. Suppose we estimate σ^2 by $s^2 = SSE/(n-2)$. Derive the distribution for

$$\frac{\hat{Y}_h - \mathbb{E}[Y|X = X_h]}{\sqrt{s^2\left[\frac{1}{n} + \frac{(X_h - \bar{X})^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]}}.$$

You can use the fact that $SSE/\sigma^2 \sim \chi^2_{n-2}$ without proof.

f. Suppose we observe a new observation Y_{new} at covariate value $X = X_{new}$. What is the $(1-\alpha)100\%$ prediction interval for Y_{new}.

g. Give an intuitive explanation for why the prediction interval for Y_{new} is different from the confidence interval for $\hat{Y}_h$.

4. Suppose $Y \sim \mathrm{Exp}(\mu)$, which has pdf $f(y) = \frac{1}{\mu}\exp(-y/\mu)$.

a. Use the following `R` code to generate data from the model $Y_i \sim \mathrm{Exp}(0.05/X_i)$, and provide a scatterplot of Y against X.

```
set.seed(123)
n <- 500
X <- rnorm(n, 3, 1)
Y <- rep(n, X)
```

b. Fit the model $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ using the `lm` function. Provide a plot of the best fit line on the scatterplot of Y vs X. Also provide the residual vs fitted plot. What assumptions of the linear regression model appear to be violated by the data?

c. Derive the variance stabilizing transformation for $Y \sim \mathrm{Exp}(\mu)$.

d. Apply the variance stabilizing transformation to the data Y, and then fit the linear regression model to the transformed data. Provide the new scatterplot and residuals vs fitted plot. Explain

whether the transformed data appear to satisfy the assumptions of linear regression better than the original data?

5. Consider the multiple linear regression model

$$Y_i = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_{p-1,i} X_{i,p-1} + \varepsilon_i,$$

where X_{ij} is the jth covariate for the ith observation. Assume the errors are normally distributed with zero mean and variance σ^2.

   a. What is the interpretation of β_1 in this model?
   b. Write the matrix form of the model. Label the response vector, design matrix, coefficient vector, and error vector, and indicate the dimensions of each in terms of the number of observations n and the number of covariates p.
   c. Write the likelihood $\mathscr{L}$ and log-likelihood ℓ of $\theta = (\beta, \sigma^2)$.
   d. Find the partial derivative of the log-likelihood with respect to β.
   e. Solve $\frac{\partial \ell}{\partial \beta} = 0$ for $\hat $, the MLE of the coefficient vector.
   f. Calculate $\mathrm{Var}(\hat{\beta})$.
   g. Write distribution of $Y|X_h$.
   h. From (g) and the properties of the multivariate normal distribution, derive the distribution of $\hat{Y}_h = \mathbb{E}[Y|X = X_h]$.