STAT 425 Assignment 2

Due Monday, February 22, 11:59pm. Submit through Moodle.

Name: (insert your name here)

Netid: (insert)

Submit your computational work both as an R markdown (*.Rmd) document and as a pdf, along with any files needed to run the code. Embed your answers to each problem in the document below after the question statement. If you have hand-written work, please scan or take pictures of it and include in a pdf file, ideally combined with your pdf output file from R Markdown.

Problem 1

In this problem we have data of the form $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. We construct a new variable z_i from the x_i values by subtracting their sample mean, so $z_i = x_i - \bar{x}$ for $i = 1, 2, \ldots, n$, where $\bar{x}_i = n^{-1} \sum_{i=1}^{n} x_i$. We consider two models:

$$\begin{aligned} &\text{Model 1:} \quad y_i = \beta_0 + \beta_1 x_i + e_i, \quad i = 1, 2, \ldots, n \\ &\text{Model 2:} \quad y_i = \alpha_0 + \alpha_1 z_i + e_i, \quad i = 1, 2, \ldots, n \end{aligned}$$

Here the x_i values are considered a fixed set of numbers, and the errors e_i are considered to be uncorrelated random variables with $E(e_i) = 0$ and $\text{var}(e_i) = \sigma^2$.

a) By subtracting one model from the other and averaging, show that $\alpha_0 = \beta_0 + \beta_1 \bar{x}$.

Answer: (here or indicate where it is in the attached pdf file.)

b) Substituting $\alpha_0 = \beta_0 + \beta_1 \bar{x}$ into Model 2, show that $(\alpha_1 - \beta_1)(x_i - \bar{x}) = 0$, for $i = 1, 2, \ldots, n$. Assume $\sum_{i=1}^{n} (x_i - \bar{x})^2 > 0$, i.e., the x_i values are not all the same. Show why this implies $\alpha_1 = \beta_1$.

Answer:

Now we use matrix notation and rewrite Model 2 as $\mathbf{y} = \mathbf{Z}\alpha + \mathbf{e}$ with:

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2, \\ \vdots \\ y_n \end{pmatrix} \qquad \mathbf{Z} = \begin{pmatrix} 1 & z_1 \\ 1 & z_2 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix} \qquad \alpha = \begin{pmatrix} \alpha_1 \\ \alpha_2 \end{pmatrix} \qquad \mathbf{e} = \begin{pmatrix} e_1 \\ e_2, \\ \vdots \\ e_n \end{pmatrix}$$

c) Show

$$\mathbf{Z}^T\mathbf{Z} = \begin{pmatrix} n & 0 \\ 0 & S_{xx} \end{pmatrix}$$

where $S_{xx} = \sum_{i=1}^{n}(x_i - \bar{x})^2$.

Answer:

d) The LS estimate of the coefficient vector α solves the matrix equation:

$$\mathbf{Z}^T\mathbf{Z}\,\hat{\alpha} = \mathbf{Z}^T\mathbf{y}$$

Solve the equation algebraically to get simplified expressions for $\hat{\alpha}_1$ and $\hat{\alpha}_2$ expressed in terms of the original x_i, y_i and n values.

Answer:

e) Derive a simplified expressions for the entries of $\mathrm{cov}(\hat{\alpha})$, the 2×2 covariance matrix of $\hat{\alpha} = \begin{pmatrix} \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix}$. (Note that this still depends on the unknown σ^2.)

Answer:

Problem 2

This problem considers the `prostate` data in the library `faraway`. See `help(prostate)` for more information about the data set.

a) Fit a linear model with `lpsa` as the response and all the other variables as predictor variables. Display the model summary and identify all variables whose coefficient t values are statistically significant at the $\alpha = 0.05$ level. In each case, state what the null hypothesis is for the t test.

Answer:

b) Fit a reduced model with `lpsa` as the response but removing `lcp`, `gleason`, and `pgg45` as predictors in the model. Display the model summary, and provide both 90% and 95% confidence intervals for the coefficient of `lbph`. Do either or both of them include 0?

Answer:

c) Perform an F test comparing the model from Part b) as the null model, and the model from Part a) as the alternative model. Based on your calculations, does the test reject or accept the null hypothesis at the level $\alpha = 0.05$?

Answer:

d) Here are the data for observation #32:

```
> prostate[32,]
      lcavol lweight age     lbph svi      lcp gleason pgg45    lpsa
32 0.1823216  6.1076  65 1.704748   0 -1.38629       6     0 2.00821
```

Pretending you did not know the value for `lpsa` for this observation, use your model from Part b) to compute a 95% prediction interval for observation 32 based on the predictors in the model. Does the interval include the value observed in the data?

Answer:

e) It is often useful to make a scatter plot of the residuals, $\mathbf{r} = \mathbf{y} - \hat{\mathbf{y}}$, on the vertical axis versus the fitted values, $\hat{\mathbf{y}}$ on the horizontal axis. Make such a scatter plot for the model in Part b), and add a horizontal line at a vertical height 0. Hint: `abline(h=0)`. Is there any trend or pattern, or does the point cloud appear random without any systematic trend or curvature? Explain briefly.

Answer:

Problem 3:

This problem refers to the `punting` data in the `faraway` library. The average distance punted and hang times of 10 punts of a football were measured for 13 volunteers. The left and right leg strength and flexibility were also recorded for each volunteer.

a) Fit a regression model with `Distance` as the response, and `RStr`, `LStr`, `RFlex` and `LFlex` as predictors (left and right leg strength, and left and right leg flexibility). Present a summary of the fitted model. Which if any predictors are significant at the 5% level?

Answer:

b) Use an F-test to determine whether collectively these four predictors have any relationship with the response, i.e., test the (null) hypothesis that $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}} = \beta_{\texttt{RFlex}} = \beta_{\texttt{LFlex}} = 0$. (Here we are referring to the coefficient for predictor X_j in the model as β_{X_j}.) What do you conclude?

Answer:

c) Now we wish to test whether $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}}$ but not necessarily 0. Under the hypothesis that these two coefficients are equal write out the regression model formula and show that it is equivalent to replacing `RStr` and `LStr` in the model by the single variable `Str` = `RStr` + `LStr`.

Answer:

d) Use an F-test to test whether $\beta_{\texttt{RStr}} = \beta_{\texttt{LStr}}$. Note that the reduced model implied by this hypothesis entails replacing `RStr` and `LStr` in the `lm` model formula by `I(RStr+LStr)` (using the syntax of R).

Answer:

e) Make a plot of residuals versus fitted values for the reduced model considered in Part d). Does the plot show any trend or pattern, or does it appear to be random noise? Explain briefly.

Answer: