PSTAT 126 - Assignment 6

Fall 2022

Kayla Katakis

Note: ***Submit both your `Rmd` and generated pdf file to Gradescope.*** *Use the same indentation level as* ***Solution*** *markers to write your solutions. Improper indentation will break your document.*

1. Using the prostate data from the faraway package with lpsa (log prostate specific antigen) as response and lcavol (log cancer volume) as predictor, the fitted model is

$$\text{lpsa } = 1.507 + 0.719 \text{ lcavol}$$

Provide an interpretation of the estimated coefficient for lcavol based on the fact that both variables are log-transformed.

Solution:

2. In a study of faculty salaries in a small college in the Midwest, a linear regression model was fit, giving the fitted mean function

$$E(\text{ Salary } \mid \text{ Sex }) = 24697 - 3340 \text{ Sex},$$

where Sex equals 1 if the faculty member was female and 0 if male. The response Salary is measured in dollars (the data are from the 1970s).

(a) Give a sentence that describes the meaning of the two estimated coefficients.

Solution:

(b) An alternative mean function fit to these data with an additional term, Years, the number of years employed at this college, gives the estimated mean function

$$E(\text{ Salary } \mid \text{ Sex, Years }) = 18065 + 201 \text{ Sex } + 759 \text{ Years }.$$

The important difference between these two mean functions is that the coefficient for Sex has changed signs. Provide an explanation as to how this could happen.

Solution: $\beta = (X'X)^{-1}X'Y$ is dependent on X, therefore its value will change when new variables are introduced.

3. This problem uses the data set cakes from the alr4 package, which contains the results of a baking experiment on n = 14 packaged cake mixes. The variables X1 and X2 data are the predictors representing baking time in minutes and baking temperature in degrees Fahrenheit, respectively. The response Y is a palatability score indicating quality of the cake.

(a) Fit the model

$$E(\text{Y} \mid \text{X1}, \text{X2}) = \beta_0 + \beta_1\text{X1} + \beta_2\text{X2} + \beta_{11}\text{X1}^2 + \beta_{22}\text{X2}^2 + \beta_{12}\text{X1X2}$$

and verify that the p-values for the quadratic terms and the interaction are all less than 0.005.

Solution: All p-values are lessthan 0.005.

```
library(alr4)
```

```
## Loading required package: car

## Loading required package: carData

##
## Attaching package: 'car'

## The following object is masked from 'package:dplyr':
##
##     recode

## The following object is masked from 'package:purrr':
##
##     some

## Loading required package: effects

## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

```
data(cakes)
attach(cakes)
x1_sq = X1*X1
x2_sq = X2*X2
x1_x2 = X1*X2
cake_model <-lm(Y~X1+X2+x1_sq+x2_sq+x1_x2, data = cakes)
summary(cake_model)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + x1_sq + x2_sq + x1_x2, data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4912 -0.3080  0.0200  0.2658  0.5454
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.204e+03  2.416e+02  -9.125 1.67e-05 ***
## X1           2.592e+01  4.659e+00   5.563 0.000533 ***
## X2           9.918e+00  1.167e+00   8.502 2.81e-05 ***
## x1_sq       -1.569e-01  3.945e-02  -3.977 0.004079 **
## x2_sq       -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
## x1_x2       -4.163e-02  1.072e-02  -3.883 0.004654 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4288 on 8 degrees of freedom
## Multiple R-squared:  0.9487, Adjusted R-squared:  0.9167
## F-statistic:  29.6 on 5 and 8 DF,  p-value: 5.864e-05
```

(b) The cake experiment was carried out in two blocks of seven observations each. It is possible that the response might differ by block, due to differences in air temperature or humidity, for example. Add a main effect for the Block variable to model in part a), fit the model, and summarize results.

Solution: The block in which the experiment was carried out is not a significant predictor.

```
cake_model2 <-lm(Y~X1+X2+x1_sq+x2_sq+x1_x2+ block, data = cakes)
summary(cake_model2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + x1_sq + x2_sq + x1_x2 + block, data = cakes)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.4525 -0.3046  0.0200  0.2924  0.4883
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.205e+03  2.542e+02  -8.672 5.43e-05 ***
## X1           2.592e+01  4.903e+00   5.287 0.001140 **
## X2           9.918e+00  1.228e+00   8.080 8.56e-05 ***
## x1_sq       -1.569e-01  4.151e-02  -3.779 0.006898 **
## x2_sq       -1.195e-02  1.660e-03  -7.197 0.000178 ***
## x1_x2       -4.163e-02  1.128e-02  -3.690 0.007754 **
## block1       1.143e-01  2.412e-01   0.474 0.650014
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4512 on 7 degrees of freedom
## Multiple R-squared:  0.9503, Adjusted R-squared:  0.9077
## F-statistic: 22.31 on 6 and 7 DF,  p-value: 0.0003129
```

4. The data BGSall in the alr4 package contains information on n = 136 children in the Berkeley Guidance study, including heights at ages 9 and 18 (HT9 and HT18), and gender (Sex = 0 for male, 1 for female). Consider the regression of HT18 on HT9 and the grouping factor Sex.

   (a) Draw the scatterplot of HT18 versus HT9, using a different symbol for males and females. Comment on the information in the graph about an appropriate mean function for these data.

   Solution:

   (b) Obtain the appropriate test for a parallel regression model.

   Solution:

   (c) Assuming the parallel regression model is adequate, estimate a 95% confidence interval for the difference between males and females. For the parallel regression model, this is the difference in the intercepts of the two groups.

   Solution:

5. The data set infmort from the faraway package contains information on the mortality of infants for 105 nations. The variable mortality gives the number of deaths per 1000 live births, while income is the per capita income in US dollars and region indicates the geographic area of the nation. Consider the model

   $$E(\log(\text{ mortality }) \mid \text{ income,region }) = \beta_0 + \beta_1 \log(\text{ income }) + \beta_2 \text{ region } + \beta_{12} \text{ region } * \log(\text{ income })$$

   (a) State the null and alternative hypotheses for the overall F-test for this model. Perform the test and summarize results.

   Solution:

(b) Explain the practical meaning of the hypothesis $\mathcal{H}_0 : \beta_{12} = \beta_2 = 0$ in the context of the above model.

Solution:

(c) Perform a test for the hypothesis in part b) and summarize your results.

Solution: