Homework 6

Emily Lu

5/26/2019

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

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

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

Given both variables are log-transformed, we would use the formula:

$$100[(1+p)^{\beta_j} - 1].$$

If we increase lcavol by 200%, then the average of lpsa changes by

$$100[(1+2)^{0.719} - 1] = 120.32\%.$$

.

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

$$E(\textbf{Salary}|\textbf{Sex}) = 24697 - 3340\textbf{Sex}$$

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.

$\beta_0 = 24697$ is the expected salary for males and $\beta_1 = -3340$ is the change in the expected salaries between males and females, i.e. females' expected salaries are \$3340 less than males' expected salaries.

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

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

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.

When the number of years employed at this college is factored into the expected salary function, the coefficient for Sex has changed signs. This could happen by taking

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

$$\iff E(\text{Salary}|\text{Sex}) = 18065 + 201\text{Sex} + 759E(\text{Years}|\text{Sex})$$

by replacing the Years with the conditional expectation of Years given the other 3 terms. Solving for the $E(\text{Years}|\text{Sex})$, we get

$$E(\text{Years}|\text{Sex}) = \frac{24697 - 3340\text{Sex}}{759} - \frac{18065 + 201\text{Sex}}{759} \approx 8.7 - 4.7\text{Sex}.$$

So, females work approximately $8.7 - 4.7 = 4$ years on average and males work approximately 8.7 years on average. This is consistent with the mean function of

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

when we plug in Years $= 8.7$ for males and Years $= 4$ for females in the mean function,

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

3. This problem uses the dataset 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

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

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

```
library(alr4)
```

```
## Loading required package: car
```

```
## Loading required package: carData
```

```
## Loading required package: effects
```

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

```
attach(cakes)
fit <- lm(Y ~ X1 + X2 + I(X1^2) + I(X2^2) +I(X1*X2), data = cakes)
summary(fit)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(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 ***
## I(X1^2)     -1.569e-01  3.945e-02  -3.977 0.004079 **
## I(X2^2)     -1.195e-02  1.578e-03  -7.574 6.46e-05 ***
## I(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
```

Then the fitted model is

$$\hat{Y} = -2.204e^{3} + 2.592e\text{X1} + 9.918\text{X2} - 1.569e^{-1}\text{X1}^2 - 1.195e^{-2}\text{X2}^2 - 4.163e^{-3}\text{X1X2}.$$

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.

```
fit2 <- lm(Y~ X1 + X2 + I(X1^2) + I(X2^2) +I(X1*X2) + block, data = cakes)
summary(fit2)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X2 + I(X1^2) + I(X2^2) + I(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 ***
## I(X1^2)     -1.569e-01  4.151e-02  -3.779 0.006898 **
## I(X2^2)     -1.195e-02  1.660e-03  -7.197 0.000178 ***
## I(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
```

The fitted model is

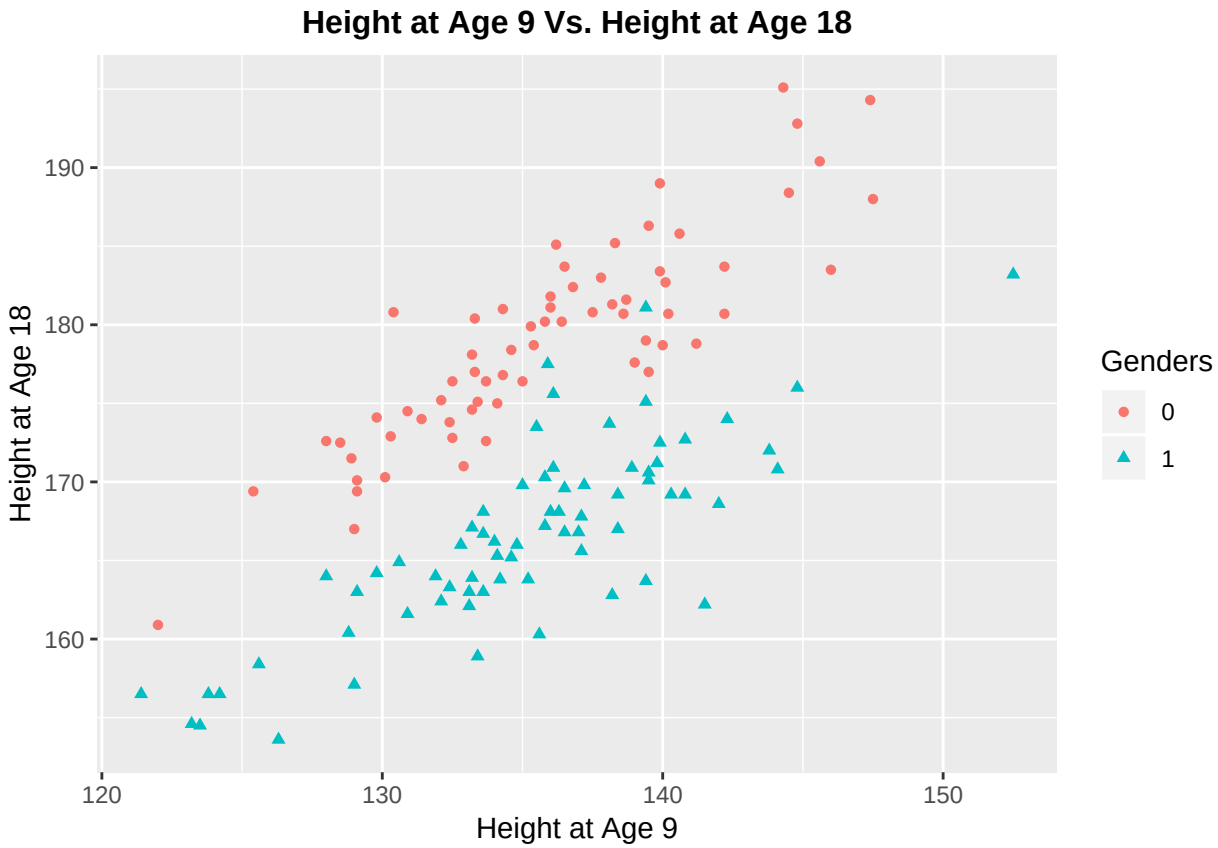
$$\hat{Y} = -2.205e^{3} + 2.592e\text{X1} + 9.918\text{X2} - 1.569e^{-1}\text{X1}^2 - 1.195e^{-2}\text{X2}^2 - 4.163e^{-3}\text{X1X2} + 1.143e^{-1}\text{block}.$$

The result suggests that the block variable is not significant.

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.

```
attach(BGSall)
library(ggplot2)
HT9 <- BGSall$HT9
HT18 <- BGSall$HT18
Sex <- BGSall$Sex
Genders <- factor(Sex)
input = BGSall[,c('HT9', 'HT18', 'Sex')]
ggplot(input, aes(x = HT9, y = HT18)) +
  geom_point(aes(shape = Genders,  color = Genders)) +
  labs(title = "Height at Age 9 Vs. Height at Age 18",
       x = "Height at Age 9", y = "Height at Age 18") +
  theme(plot.title = element_text(hjust = 0.5, size = 12, face = "bold"))
```



Group Female's mean function is $Y_i = \beta_0 + \beta_2 x_{i2}$ and Group Male's mean function is $Y_i = \beta_0 + \beta_1 + \beta_2 x_{i2}$. β_0 is the mean height at age 18 and β_2 is the difference in mean height at age 18 between male & female.

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

```
Sex1 <- ifelse(BGSall$Sex == 0, 'Male', 'Female')
# parallel model, so remove interaction terms below
par.lm <- lm(BGSall$HT18 ~ Sex1 + BGSall$HT9, data = input)
summary(par.lm)
```

```
##
```

```
## Call:
## lm(formula = BGSall$HT18 ~ Sex1 + BGSall$HT9, data = input)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.4694  -2.0952  -0.0136   1.7101  10.4467
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 36.82147    7.29177    5.05 1.43e-06 ***
## Sex1Male    11.69584    0.59036   19.81  < 2e-16 ***
## BGSall$HT9   0.96006    0.05388   17.82  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.432 on 133 degrees of freedom
## Multiple R-squared:  0.8516, Adjusted R-squared:  0.8494
## F-statistic: 381.7 on 2 and 133 DF,  p-value: < 2.2e-16
```

With $H_0 : \beta_2 = 0$ vs. $H_A : \beta_2 \neq 0$, we would reject H_0 since the p-value $= 2.2e^{-16} < \alpha = 0.05$. This result means that sex has a significant effect in determining the height for males and females at age 18.

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.

Given the degree of freedom is 133, then the t-value is 1.98. Thus, the 95% CI for the difference between males and females is (11.69584-1.98x0.59036, 11.69584+1.98x0.59036) $\longleftrightarrow$ (10.53, 12.86). An alternative method is:

```
confint(par.lm, level = 0.95)
```

```
##                   2.5 %    97.5 %
## (Intercept) 22.3986375 51.244301
## Sex1Male    10.5281335 12.863548
## BGSall$HT9   0.8534845  1.066628
```

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

$$E(log(\textbf{mortality})|\textbf{income, region}) = \beta_0 + \beta_1 log(\textbf{income}) + \beta_2 \textbf{region} + \beta_{12} \textbf{region} log(\textbf{income}).$$

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

Null hypothesis is $H_0 : \beta_1 = \beta_2 = \beta_{12} = 0$ (which means the fitted model is not significant) and alternative hypothesis is $H_A : H_0 : \beta_1 \neq \beta_2 \neq \beta_{12} \neq 0$ (which means the model is significant).

```
library(faraway)
```

```
##
## Attaching package: 'faraway'
```

```
## The following objects are masked from 'package:alr4':
##
##     cathedral, pipeline, twins

## The following objects are masked from 'package:car':
##
##     logit, vif
```

```
attach(infmort)
full.lm <- lm(log(mortality) ~ log(income) + region + region*log(income), data = infmort)
summary(full.lm)
```

```
##
## Call:
## lm(formula = log(mortality) ~ log(income) + region + region *
##     log(income), data = infmort)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.46809 -0.26530 -0.02148  0.27478  3.14219
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  4.9385     0.6362   7.763 1.06e-11 ***
## log(income)                 -0.0112     0.1235  -0.091   0.9280
## regionEurope                 2.0882     1.8422   1.134   0.2599
## regionAsia                   1.2634     0.8561   1.476   0.1434
## regionAmericas               1.5661     1.1856   1.321   0.1898
## log(income):regionEurope    -0.5205     0.2516  -2.069   0.0413 *
## log(income):regionAsia      -0.3798     0.1580  -2.404   0.0182 *
## log(income):regionAmericas  -0.3978     0.1979  -2.010   0.0473 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5971 on 93 degrees of freedom
##   (4 observations deleted due to missingness)
## Multiple R-squared:  0.6464, Adjusted R-squared:  0.6198
## F-statistic: 24.29 on 7 and 93 DF,  p-value: < 2.2e-16
```

From the summary of the model, the F-statistic is 24.29 on 7 and 93 DF and the p-value is $2.2e^{-16}$ which is less than 0.05 so we would reject the null hypothesis. This means the fitted regression model is significant in determining the log(mortality).

b) Explain the practical meaning of the hypothesis $H_0 : \beta_{12} = \beta_2 = 0$ in the context of the above model.

In the context of the model above, the hypothesis $H_0 : \beta_{12} = \beta_2 = 0$ means that the region has no impact on the relationship between income and mortality, i.e. log(mortality) is independent of the region and interaction between the region & log(income) for the given income and region.

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

```
# restricted model (or reduced model), includes fewer possible predictors
reduced.lm <- lm(log(mortality) ~ log(income), data = infmort)
anova(reduced.lm, full.lm)
```

```
## Analysis of Variance Table
##
## Model 1: log(mortality) ~ log(income)
## Model 2: log(mortality) ~ log(income) + region + region * log(income)
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     99 46.685
## 2     93 33.152  6    13.533 6.3274 1.31e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since the p-value $= 6.418e^{-5}$ is less than the significant level, 0.05, we would reject the null hypothesis and conclude that $\beta_{12} \neq 0$, $\beta_2 \neq 0$. Thus, the region and the interaction between region & log(income) are significant variables in determining log(mortality) for the given income and region.