HW3 Math158

Joshua Jansen-Montoya

2022-09-21

Problem 3.1

For the prostate data, fit a model with lpsa as the response and the other variables as predictors: 1. Compute 90 and 95% CIs for the parameter associated with age. Using just these intervals, what could we have deduced about the p-value for age in the regression summary? 2. Compute and display a 95% joint confidence region for the parameters associated with age and lbph. Plot the origin on this display. The location of the origin on the display tells us the outcome of a certain hypothesis test. State that test and its outcome. 3. In the text, we made a permutation test corresponding to the F-test for the significance of all the predictors. Execute the permutation test corresponding to the t-test for age in this model. (Hint: {summary(g)$coef[4,3] gets you the t-statistic you need if the model is called g.) 4. Remove all the predictors that are not significant at the 5% level. Test this model against the original model. Which model is preferred?

Answer 3.1

1. We can construct the following linear model and compute the following CIs for the parameter associated with age as follows,

```
library(faraway)
reg <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate)
confint(reg, level = .90)
```

```
##                       5 %         95 %
## (Intercept) -1.485718237  2.824391633
## lcavol       0.440867156  0.733176497
## lweight      0.171846568  0.737088281
## age         -0.038210200 -0.001064151
## lbph         0.009890745  0.204217317
## svi          0.360029029  1.172285623
## lcp         -0.256770899  0.045822373
## gleason     -0.216620186  0.306903382
## pgg45       -0.002824333  0.011874796
```

```
confint(reg, level = .95)
```

```
##                     2.5 %      97.5 %
## (Intercept) -1.906960983 3.245634379
## lcavol       0.412298699 0.761744954
## lweight      0.116603435 0.792331414
## age         -0.041840618 0.002566267
## lbph        -0.009101499 0.223209561
## svi          0.280644232 1.251670420
## lcp         -0.286344443 0.075395916
## gleason     -0.267786053 0.358069248
## pgg45       -0.004260932 0.013311395
```

```
  summary(reg)
```

```
##
## Call:
## lm(formula = lpsa ~ lcavol + lweight + age + lbph + svi + lcp +
##     gleason + pgg45, data = prostate)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.7331 -0.3713 -0.0170  0.4141  1.6381
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.669337   1.296387   0.516  0.60693
## lcavol       0.587022   0.087920   6.677 2.11e-09 ***
## lweight      0.454467   0.170012   2.673  0.00896 **
## age         -0.019637   0.011173  -1.758  0.08229 .
## lbph         0.107054   0.058449   1.832  0.07040 .
## svi          0.766157   0.244309   3.136  0.00233 **
## lcp         -0.105474   0.091013  -1.159  0.24964
## gleason      0.045142   0.157465   0.287  0.77503
## pgg45        0.004525   0.004421   1.024  0.30886
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7084 on 88 degrees of freedom
## Multiple R-squared:  0.6548, Adjusted R-squared:  0.6234
## F-statistic: 20.86 on 8 and 88 DF,  p-value: < 2.2e-16
```

Looking over the results from our tests, we can see that 0 is in the interval for our age variable for only the 95% interval, not for our 90% interval We can see that this agrees with our linear model which puts age significant to $\alpha = 0.1\%$ (as indicated by the "."), and that the p-value is 8.29%. 2. We can plot the data as follows using the following R-code

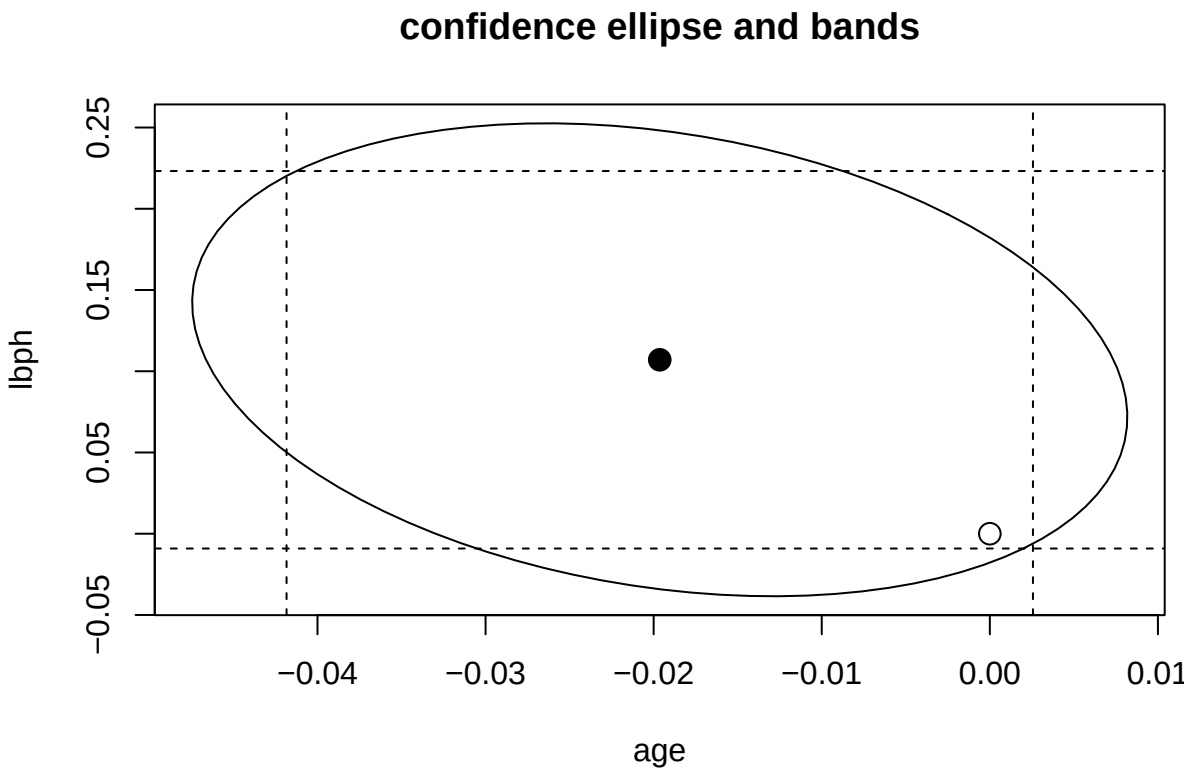
```
  library(faraway)
  require(ellipse)
```

```
## Loading required package: ellipse
```

```
##
## Attaching package: 'ellipse'
```

```
## The following object is masked from 'package:graphics':
##
##     pairs
```

```
  lmod <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate)
  plot(ellipse(lmod,c(4,5)),type="l", main = "confidence ellipse and bands")
  points(coef(lmod)[4], coef(lmod)[5], pch=19, cex = 1.5)
  points(0,0,cex=1.5)
  abline(v = confint(lmod)[4,], lty=2)
  abline(h = confint(lmod)[5,], lty=2)
```



3. We can answer this problem by executing the following R code,

```
library(faraway)
reg <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate)
regs <- summary(reg)
nreps <- 4000
set.seed(123)
tstats <- numeric(nreps)
for(i in 1:nreps){
lmods <- lm(sample(lpsa) ~lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostat
tstats[i] <- summary(lmods)$coef[4,3]
}
summary(reg)$coef[3,]
```

```
##    Estimate  Std. Error     t value    Pr(>|t|)
## 0.454467424 0.170012435 2.673142255 0.008955363
```

```
mean(tstats > abs(regs$coef[4,3]))
```

```
## [1] 0.03825
```

We can see that it is quite rare that our random sample t-stats is greater than our calculate t-stat from the p-value we have obtained, and thus, we can conclude that our model is sufficient. 4. We can see from the model that the one that we can remove lcp, gleason, and pgg45 from our variables to get the following model, from which we can use anova to compare.

```
library(faraway)
reg <- lm(lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason + pgg45, data = prostate)
reg2 <- lm(lpsa ~ lcavol + lweight + age + lbph + svi , data = prostate)
anova(reg2, reg)
```

```
## Analysis of Variance Table
```

```
##
## Model 1: lpsa ~ lcavol + lweight + age + lbph + svi
## Model 2: lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
##     pgg45
##   Res.Df    RSS Df Sum of Sq     F Pr(>F)
## 1     91 45.526
## 2     88 44.163  3    1.3625 0.905 0.4421
```

We can see that compared to the base model, since we have too large of a p-value and thus, we can accept the reduced variable model (where lcp, gleason, and pgg45 are all removed), which is our null hypothesis model.

Problem 3.2

Thirty samples of cheddar cheese were analyzed for their content of acetic acid, hydrogen sulfide and lactic acid. Each sample was tasted and scored by a panel of judges and the average taste score produced.Use the cheddar data to answer the following: 1. Fit a regression model with taste as the response and the three chemical contents as predictors. Identify the predictors that are statistically significant at the 5% level. 2. Acetic and H2S are measured on a log scale. Fit a linear model where all three predictors are measured on their original scale. Identify the predictors that are statistically significant at the 5% level for this model. 3. Can we use an F-test to compare these two models? Explain. Which model provides a better fit to the data? Explain your reasoning. 4. If H2S is increased 0.01 for the model used in (a), what change in the taste would be expected? 5. What is the percentage change in H2S on the original scale corresponding to an additive increase of 0.01 on the (natural) log scale?

Answer 3.2

1. We can fit the following regression model with taste as the response variable and the three chemical contents as predictors as follows,

```
library(faraway)
reg <- lm(taste ~ Acetic + H2S + Lactic, data = cheddar)
summary(reg)
```

```
##
## Call:
## lm(formula = taste ~ Acetic + H2S + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.390  -6.612  -1.009   4.908  25.449
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.8768    19.7354  -1.463  0.15540
## Acetic        0.3277     4.4598   0.073  0.94198
## H2S           3.9118     1.2484   3.133  0.00425 **
## Lactic       19.6705     8.6291   2.280  0.03108 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.13 on 26 degrees of freedom
## Multiple R-squared:  0.6518, Adjusted R-squared:  0.6116
## F-statistic: 16.22 on 3 and 26 DF,  p-value: 3.81e-06
```

From which we can see that only lactic and H2S are significant to a $\alpha = 0.05$ level. 2. We can construct the desired linear regression model as follows by exponentiating the logged vectors as follows,

```
library(faraway)
regExp <- lm(taste ~ exp(Acetic) + exp(H2S) + Lactic, data = cheddar)
summary(regExp)
```

```
##
## Call:
## lm(formula = taste ~ exp(Acetic) + exp(H2S) + Lactic, data = cheddar)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -16.209  -7.266  -1.651   7.385  26.335
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.897e+01  1.127e+01  -1.684   0.1042
## exp(Acetic)  1.891e-02  1.562e-02   1.210   0.2371
## exp(H2S)     7.668e-04  4.188e-04   1.831   0.0786 .
## Lactic       2.501e+01  9.062e+00   2.760   0.0105 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 11.19 on 26 degrees of freedom
## Multiple R-squared:  0.5754, Adjusted R-squared:  0.5264
## F-statistic: 11.75 on 3 and 26 DF,  p-value: 4.746e-05
```

Using the exponentiated vectors, we can see that only Lactic acid is significant for the 5% level of this model. 3. We cannot use a F-statistic test to compare the two models, as we do not have a reduced baseline model that we are comparing to, since each model has the same number of predictor variables, just that two of them are exponentiated. The F-statistic test is to see if we can discard any predictor variables using a baseline model, which does not apply here. 4. If H2S is increased by 0.01, we would expect taste to increase by about 0.0169 points. 5. The percentage change corresponding to 0.01 on the natural log scale would be an increase of 101%.

Problem 3.3

Using the teengamb data, fit a model with gamble as the response and the other variables as predictors. 1. Which variables are statistically significant at the 5% level? 2. What interpretation should be given to the coefficient for sex? 3. Fit a model with just income as a predictor and use an F-test to compare it to the full model.

Answer 3.3

1. We can find which variables are statistically significant at the 5% level using the following linear model.

```
library(faraway)
?teengamb
reg <- lm(gamble ~ sex + status + income + verbal + gamble, data = teengamb)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 5 in
## model.matrix: no columns are assigned
```

```
summary(reg)
```

```
##
```

```
## Call:
## lm(formula = gamble ~ sex + status + income + verbal + gamble,
##     data = teengamb)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -51.082 -11.320  -1.451   9.452  94.252
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  22.55565   17.19680   1.312   0.1968
## sex         -22.11833    8.21111  -2.694   0.0101 *
## status        0.05223    0.28111   0.186   0.8535
## income        4.96198    1.02539   4.839 1.79e-05 ***
## verbal       -2.95949    2.17215  -1.362   0.1803
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 22.69 on 42 degrees of freedom
## Multiple R-squared:  0.5267, Adjusted R-squared:  0.4816
## F-statistic: 11.69 on 4 and 42 DF,  p-value: 1.815e-06
```

From this, we can see that only sex and income are statistically significant at a 5% level. 2. The coefficient for sex should be interpreted as if you are female, then according to the model you will spend 22 pounds less on gambling per year. 3. Finally, we can complete a F-test for this for our two models as follows,

```
library(faraway)
reg <- lm(gamble ~ sex + status + income + verbal + gamble, data = teengamb)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 5 in
## model.matrix: no columns are assigned
```

```
regInc <- lm(gamble ~ income, data = teengamb)
anova(regInc, reg)
```

```
## Analysis of Variance Table
##
## Model 1: gamble ~ income
## Model 2: gamble ~ sex + status + income + verbal + gamble
##   Res.Df   RSS Df Sum of Sq      F  Pr(>F)
## 1     45 28009
## 2     42 21624  3    6384.8 4.1338 0.01177 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We can see from our F-statistic test that we have a p-value of 1.1% which is significant to our 5% level and indicates that we can ignore the null hypothesis in favor of the full model.