R Lab #2 (Ch. 7) – Solutions

1. 7.8.1 Polynomial Regression

Work through section 7.8.1 by fitting a fourth-degree polynomial to the Wage data using the various methods on pages 288 and 289. Then recreate the left-hand plot from Figure 7.1 on page 267 by entering the code in the middle of page 289. Next confirm that the fitted values from the first two models (fit and fit2) are identical by following the directions at the bottom of page 289 (note that the maximum absolute difference may not match the value provided in the lab exactly, but it should still be essentially zero). Next use the anova function to compare five sequential polynomial models as described on page 290, and also confirm the results at the bottom of page 290 and top of page 291. Finally, use the anova function to compare three models with the following predictor terms:

- Model 1: education, age
- Model 2: education, poly(age, 2)
- Model 3: education, poly(age, 3)

What is the most appropriate description of the results for this last ANOVA?

a) The p-value for model 1 vs. model 2 is essentially zero, and the p-value for model 2 vs. model 3 is essentially zero.
b) The p-value for model 1 vs. model 2 is essentially zero, and the p-value for model 2 vs. model 3 is approximately 3%.
c) The p-value for model 1 vs. model 2 is approximately 3%, and the p-value for model 2 vs. model 3 is essentially zero.
d) The p-value for model 1 vs. model 2 is approximately 3%, and the p-value for model 2 vs. model 3 is approximately 3%.

Solution: B

```
fit.1 <- lm(wage ~ education + age, data = Wage)
fit.2 <- lm(wage ~ education + poly(age, 2), data = Wage)
fit.3 <- lm(wage ~ education + poly(age, 3), data = Wage)
anova(fit.1, fit.2, fit.3)
```

2. 7.8.1 Polynomial Regression

Work through section 7.8.1 by fitting a fourth-degree polynomial to the Wage data using the various methods on pages 288 and 289. Then recreate the left-hand plot from Figure 7.1 on page 267 by entering the code in the middle of page 289. Next confirm that the fitted values from the first two models (fit and fit2) are identical by following the directions at the bottom of page 289 (note that the maximum absolute difference may not match the value provided in the lab exactly, but it should still be essentially zero). Next use the anova function to compare five sequential polynomial models as described on page 290, and also confirm the results at the bottom of page

290 and top of page 291. Finally, use the anova function to compare three models with the following predictor terms:

- Model 1: education, age
- Model 2: education, poly(age, 2)
- Model 3: education, poly(age, 3)

What is the value of the t-statistic for the last orthogonal polynomial in Model 3 (i.e. poly(age, 3)3)? Round your answer to 2 decimal places.

Solution: 2.12

coef(summary(fit.3)) #poly(age, 3)3 = 2.119808

3. 7.8.1 Step Functions

Since we have not yet covered logistic regression, do not worry about working through the lower half of page 291 and upper half of page 292. Fit the step function at the bottom of page 292 and confirm the results provided. Then fit another step function, this time using cut(age,c(0,25,40,60,80)), which creates cutpoints at 25, 40, and 60 years of age. Based on the results of this model fit, match the predicted average salaries in thousands of dollars to the appropriate age group.

- a) (0,25] – 76
- b) (25, 40] – 110
- c) (40, 60] – 118
- d) (60, 80] - 113

Solution:

table(cut(age,c(0,25,40,60,80)))
fit <- lm(wage ~ cut(age,c(0,25,40,60,80)), data = Wage)
coef(summary(fit))
predict(fit, data.frame(age = seq(0:80)))

4. 7.8.2 Splines

Follow the directions on pages 293 and 294 to fit a series of splines to the Wage data:

- A cubic regression spline with knots at ages 25, 40, and 60.
- A natural regression spline with 4 degrees of freedom.
- A smoothing spline with smoothing parameter chosen by generalized cross-validation (use cv=FALSE rather than cv=TRUE).
- A local linear regression fit with span 0.5.

Find predicted salaries in thousands of dollars at age 80 for each of the models. For the regression splines and local linear regression, enter predict(fit,data.frame(age=80)), where fit is the name of the fitted model. For the smoothing spline, enter fit$y[fit$x==80], where fit is the name of the fitted model. Match the predictions to the appropriate model.

a) A cubic regression spline with knots at ages 25, 40, and 60. – 77
b) A natural regression spline with 4 degrees of freedom. – 96
c) A smoothing spline with smoothing parameter chosen by generalized cross-validation. – 88
d) A local linear regression fit with span 0.5. – 79

Solution:

```
fit <- lm(wage ~ bs(age, knots = c(25, 40, 60)), data = Wage)
predict(fit, data.frame(age = 80))
fit2 <- lm(wage ~ ns(age, df = 4), data = Wage)
predict(fit2, data.frame(age = 80))
fit3 <- smooth.spline(age, wage, cv = FALSE)
fit3$y[fit3$x==80]
fit4 <- loess(wage ~ age, span = .5, data = Wage)
predict(fit4, data.frame(age = 80))
```

5. 7.8.3 Generalized Additive Models

Fit the three generalized additive models detailed on pages 294-295 and make sure you can replicate the results (you may notice some small discrepancies in the summary command likely due to updates to the algorithms):

- Model 1: smoothing spline of age with 5 degrees of freedom; education
- Model 2: year; smoothing spline of age with 5 degrees of freedom; education
- Model 3: smoothing spline of year with 4 degrees of freedom; smoothing spline of age with 5 degrees of freedom; education

Work through the analyses of these models on pages 295 and 296. Since we haven't covered logistic regression yet, don't worry about fitting the logistic regression GAM detailed at the bottom of page 296 and top of page 297.

Enter coef(gam.m2) to look at the coefficient estimates for Model 2. What is the predicted additional salary (in thousands of dollars) for those with an advanced degree (relative to those with less than a high school education)?

a) 11
b) 24
c) 38
d) 63

Solution: D

```
gam.m2 <- gam(wage ~ year + s(age, 5) + education, data = Wage)
coef(gam.m2)
```

6. 7.8.3 Generalized Additive Models

Fit the three generalized additive models detailed on pages 294-295 and make sure you can replicate the results (you may notice some small discrepancies in the summary command likely due to updates to the algorithms):

- Model 1: smoothing spline of age with 5 degrees of freedom; education
- Model 2: year; smoothing spline of age with 5 degrees of freedom; education
- Model 3: smoothing spline of year with 4 degrees of freedom; smoothing spline of age with 5 degrees of freedom; education

Work through the analyses of these models on pages 295 and 296. Since we haven't covered logistic regression yet, don't worry about fitting the logistic regression GAM detailed at the bottom of page 296 and top of page 297.

Enter predict(gam.m2,data.frame(year=2006,age=50,education="3. Some College")) to predict the salary (in thousands of dollars) in 2006 for a person aged 50 with some college. (Round your answer to the nearest whole number.)

Solution: 115

```
gam.m2 <- gam(wage ~ year + s(age, 5) + education, data = Wage)
predict(gam.m2, data.frame(year = 2006, age = 50, education = "3. Some College"))
```