Anthony Cunningham

STAT 3200

Due 3/20/2017

**Homework 5**

#1. > attach(Duncan)

> fit=lm(prestige~education, data=Duncan)

> summary(fit)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 0.28400 5.09306 0.056 0.956

education 0.90200 0.08455 10.668 1.17e-13 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 16.69 on 43 degrees of freedom

Multiple R-squared: 0.7258, Adjusted R-squared: 0.7194

F-statistic: 113.8 on 1 and 43 DF, p-value: 1.171e-13

\* Yes, education is a significant predictor of job prestige, shown by the p-value in the F-test. The p-value is 1.171 x 10-13, which is very small and easily shows statistical significance.

#2. A. 1. > is.factor(type) \* R distinguishes “type” as a factor, not a numeric

[1] TRUE value.

> is.numeric(type)

[1] FALSE

2. > unique(type) \* There are 3 levels of “type”: prof, bc, and wc.

[1] prof wc bc

Levels: bc prof wc

3. prof = Professional or managerial-type occupation

wc = White-collar type occupation

bc = Blue-collar type occupation

B. > plot(education, prestige, pch=as.numeric(type), col=as.numeric(type))

> legend(8,95, levels(type), pch=1:3, col=1:3)

> fit=lm(prestige~education, data=Duncan)

> abline(fit)



\* Yes, it appears that the linear relationship between prestige and education differ by the type of occupations. It seems like the intercepts of all three groups would be different, with the blue-collar type having the lowest intercept and the white-collar and professional types having similar intercepts. Also, the slope of the white-collar type would appear to differ significantly from the other two types, being more flat, while the blue-collar and professional types appear to have similar slopes.

#3. A. > type.dummy1 = rep(0,nrow(Duncan))

> type.dummy1[type=="prof"]=1

> type.dummy2 = rep(0,nrow(Duncan))

> type.dummy2[type=="wc"]=1

> data.frame(Duncan, type.dummy1, type.dummy2)

type income education prestige type.dummy1 type.dummy2

accountant prof 62 86 82 1 0

pilot prof 72 76 83 1 0

architect prof 75 92 90 1 0

author prof 55 90 76 1 0

chemist prof 64 86 90 1 0

minister prof 21 84 87 1 0

professor prof 64 93 93 1 0

dentist prof 80 100 90 1 0

reporter wc 67 87 52 0 1

engineer prof 72 86 88 1 0

> fit = lm(prestige~education + type.dummy1 + type.dummy2)

> summary(fit)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.4692 5.0039 1.693 0.098132 .

education 0.5642 0.1564 3.608 0.000832 \*\*\*

type.dummy1 26.0881 9.8458 2.650 0.011393 \*

type.dummy2 -6.5000 8.6042 -0.755 0.454302

Residual standard error: 14.01 on 41 degrees of freedom

Multiple R-squared: 0.8159, Adjusted R-squared: 0.8024

F-statistic: 60.56 on 3 and 41 DF, p-value: 4.066e-15

B. > contrasts(type)

prof wc \* The blue-collar type category is the baseline.

bc 0 0

prof 1 0

wc 0 1

> fit.both=lm(prestige~education + type)

> summary(fit.both)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 8.4692 5.0039 1.693 0.098132 .

education 0.5642 0.1564 3.608 0.000832 \*\*\*

typeprof 26.0881 9.8458 2.650 0.011393 \*

typewc -6.5000 8.6042 -0.755 0.454302

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.01 on 41 degrees of freedom

Multiple R-squared: 0.8159, Adjusted R-squared: 0.8024

F-statistic: 60.56 on 3 and 41 DF, p-value: 4.066e-15

\* Yes, the two models are exactly the same, since we manually chose the blue-collar category as our baseline, while R automatically selected the blue-collar category as its baseline since it’s first alphabetically. Nothing else changed, so, therefore, the two models are exactly the same.

C. Yi = B0 + B1xi + B2z1i + B3z2i + ei where Yi is prestige rating, xi is % of males in a given occupation with a high school diploma, z1i is a dummy variable for “prof” type, and z2i is a dummy variable for “wc” type.

Coding:

z1 z2

Category: prof 1 0

wc 0 1

bc 0 0 \*baseline

Fitted Model: Yi-hat = 8.4692 + 0.5642xi + 26.0881z1i – 6.5z2i + ei

D. 0.5642 = change in prestige level after a 1% increase in male high school graduation rate after accounting for effect of type

26.0881 = the difference in the prestige level between professional and blue-collar jobs when male high school graduation rate is 0% in a given occupation

-6.5 = the difference in the prestige level between white-collar and blue-collar jobs when male high school graduation rate is 0% in a given occupation

E. H0: B1=0 Ha: B1 ≠ 0

t = 3.608 p-value = 0.000832

df = 41 \* Education is a significant predictor for prestige after accounting for occupation type.

F. Fitted Models

(prof) Yi-hat = (8.4692 + 26.0881) + 0.5642xi = 34.5573 + 0.5642xi

(wc) Yi-hat = (8.4692 – 6.5) + 0.5642xi = 1.9692 + 0.5642xi

(bc) Yi-hat = 8.4692 + 0.5642xi

\* Only the intercepts differ since we are only looking at the additive model (i.e. without interaction).

G. > plot(education, prestige, pch=as.numeric(type), col=as.numeric(type))

> legend(8,95, levels(type), pch=1:3, col=1:3)

> intercept = fit.both$coefficients[1] + fit.both$coefficients[3]

> slope = fit.both$coefficients[2]

> abline(intercept, slope, col="red")

> abline(fit.both)

> intercept.wc = fit.both$coefficients[1] + fit.both$coefficients[4]

> slope = fit.both$coefficients[2]

> abline(intercept.wc, slope, col="green")



H. > anova(fit, fit.both) H0: B2 = B3 = 0

Analysis of Variance Table Ha: B2 ≠ 0 or B3 ≠ 0 or both ≠ 0

F = 10.033; numdf = 2; dendf = 41

Model 1: prestige ~ education p = 0.0002839

Model 2: prestige ~ education + type

Res.Df RSS Df Sum of Sq F Pr(>F) \* Type is a significant predictor prestige after accounting for Education.

1 43 11980.9

2 41 8044.1 2 3936.8 10.033 0.0002839

#4. A. > fit.interact = lm(prestige ~ education + type + education:type)

> summary(fit.interact)

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -2.8544 9.8767 -0.289 0.7741

education 1.0112 0.3707 2.728 0.0095 \*\*

typeprof 42.6603 19.4757 2.190 0.0345 \*

typewc 16.2190 23.4148 0.693 0.4926

education:typeprof -0.5115 0.4223 -1.211 0.2331

education:typewc -0.6323 0.4979 -1.270 0.2116

Residual standard error: 14.03 on 39 degrees of freedom

Multiple R-squared: 0.8243, Adjusted R-squared: 0.8018

F-statistic: 36.59 on 5 and 39 DF, p-value: 1.015e-13

\* We reject H0 due to sufficiently small p-value. Therefore, at least 1 of the predictors (education, type, or some combination of interaction) is significant in predicting prestige.

B. Model: Yi = B0 + B1xi + Bz1z1i + Bz2z2i + Bint, z1z1ixi + Bint, z2z2ixi + ei

Yi is prestige rating, xi is % of males in a given occupation with a high school diploma, z1i is a dummy variable for “prof” type, and z2i is a dummy variable for “wc” type, and z1ixi and z2ixi are the interaction terms between the two predictors

Coding:

z1 z2

Category: prof 1 0

wc 0 1

bc 0 0 \*baseline

Fitted Model:

Yi-hat = -2.8544 + 1.0112xi + 42.6603z1i + 16.2190z2i – 0.5115z1ixi – 0.6323z2ixi + ei

C. (prof) Yi-hat = (-2.8544 + 42.6603) + (1.0012 – 0.5115)xi = 39.8059 + 0.4897xi

(wc) Yi-hat = (-2.8544 + 16.2190) + (1.0012 – 0.6323)xi = 13.3646 + 0.3689xi

(bc) Yi-hat = -2.8544 + 1.0112xi

\* All of the intercepts differ for the three levels, while the slopes are slightly different.

D. > plot(education, prestige, pch=as.numeric(type), col=as.numeric(type))

> legend(8,95, levels(type), pch=1:3, col=1:3)

> intercept = fit.interact$coefficients[1] + fit.interact$coefficients[3]

> slope = fit.interact$coefficients[2] + fit.interact$coefficients[5]

> intercept.wc = fit.interact$coefficients[1] + fit.interact$coefficients[4]

> slope.wc = fit.interact$coefficients[2] + fit.interact$coefficients[6]

> intercept.bc = fit.interact$coefficients[1]

> slope.bc = fit.interact$coefficients[2]

> abline(intercept.bc, slope.bc, col="black")

> abline(intercept.wc, slope.wc, col="green")

> abline(intercept, slope, col="red")



E. > anova(fit.both, fit.interact)

Analysis of Variance Table

Model 1: prestige ~ education + type

Model 2: prestige ~ education + type + education:type

Res.Df RSS Df Sum of Sq F Pr(>F)

1 41 8044.1

2 39 7676.7 2 367.42 0.9333 0.4019

H0: Bint, z1 = Bint, z2 = 0 Ha: Bint, z1 ≠ 0 or Bint, z2 ≠ 0 or both ≠ 0

F = 0.9333 num df = 2 den df = 39

p-value = 0.4019

\* No, there is not significant difference in the effect of education on prestige among the three types of job occupations. Therefore, the main effects model is sufficient.

#5. A. \* It is sufficient to use the Main Effects Model (without interaction).

B. > par(mfrow=c(2,1))

> plot(fit.dummy$residuals)

> abline(h=0)

> qqnorm(fit.dummy$residuals)

> qqline(fit.dummy$residuals)



\* The constant variance assumption may be violated, as the dispersion of the residuals, especially past x > 25, does not appear to be constant. The QQ Plot shows that the data is skewed to the right, so a transformation of the data may be in order. In this case, we would descend the ladder of powers, perhaps using a log or square-root transformation. Hopefully, this will help make our residuals be more dispersed around 0 on account of the increase in variance as the fitted values increase.