Anthony Cunningham

STAT 3200

Due Fri, March 31

**Homework 6**

#1. A. > is.factor(region) \*Yes, the variable “region” is a factor.

[1] TRUE

> ?Angell \* Regions: *E* = Northeastern U.S. cities

\* four total groups *MW* = Midwestern U.S. cities

*S* = Southeastern U.S. cities

*W* = Western U.S. cities

> table(region)

Region \* This is an unbalanced One-Way ANOVA.

E MW S W

9 14 14 6

> tapply(mobility, region, mean)

E MW S W

15.90000 26.05714 32.45714 37.40000

> boxplot(mobility~region)



\* The sample means appear to differ, especially with regards to the northeast region (E), whose sample mean appears to be significantly lower than the other three groups’. Also, the constant variance assumption may be violated, since the spread of the southeast region (S) appears to be at least twice as big as the spread of both the northeast region and the west region (if one does not include the outlier), violating the rule of thumb when determining if the constant variance assumption is met in One-Way ANOVA.

#2. A. > contrasts(region)

MW S W

E 0 0 0 \* Northeastern US cities are the baseline group.

MW 1 0 0

S 0 1 0

W 0 0 1

B. Population Regression Model: Yi = B0 + BMWD1i + BSD2i + BWD3i + *e*i

\* Yi denotes mobility, B0 denotes mobility of baseline group (group E), BMW denotes mean effect of MW group on mobility, D1i denotes dummy variable corresponding to MW group, and likewise for groups S (BsD2i) and W (BWD3i).

\* We are using the baseline = 0 constraint so as to not over-parametrize the model.

Models of Individual Regions:

E Region (Baseline): Yi = B0 + *e*i

MW Region: Yi = (B0 + BMW) + *e*i

S Region: Yi = (B0 + BS) + *e*i

W Region: Yi = (B0 + BW) + *e*i

Assumptions: 1. 4 independent random samples from each of 4 populations.

2. Each population *i* is normally-distributed about unknown mean µi.

3. All populations have the same variance.

C. > fit = lm(mobility~region)

> summary(fit)

Coefficients:

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

(Intercept) 15.900 2.297 6.921 2.75e-08 \*\*\*

regionMW 10.157 2.945 3.449 0.00136 \*\*

regionS 16.557 2.945 5.623 1.73e-06 \*\*\*

regionW 21.500 3.633 5.919 6.72e-07 \*\*\*

Residual standard error: 6.892 on 39 degrees of freedom

Multiple R-squared: 0.5397, Adjusted R-squared: 0.5043

F-statistic: 15.24 on 3 and 39 DF, p-value: 1.025e-06

1) Overall Fitted Model: Y-hat = 15.9 + 10.157D1 + 16.557D2 + 21.500D3

Separate Fitted Models: E Group (Baseline): Y-hat = 15.9

MW Group: Y-hat = 15.9 + 10.157 = 26.057

S Group: Y-hat = 15.9 + 16.557 = 32.457

W Group: Y-hat = 15.9 + 21.5 = 37.4

\* Yes, these results are consistent with the sample means calculated in 1A.

2) B0 = 15.9 = the expected mobility of observations in E Region

BMW = 10.157 = the expected difference in mobility of observations in MW Region as compared to those in the baseline region (E Region)

BS = 16.557 = the expected difference in mobility of observations in S Region as compared to those in the baseline region

BW = 21.5 = the expected difference in mobility of observations in W Region as compared to those in the baseline region

3) H0: µMW = µE or H0: BMW = 0

HA: µMW ≠ µE HA: BMW ≠

t = 3.449 df = nMW + nE – 2 = 9 + 14 – 2 = 21

p-value = 0.00136 \* Reject H0; µMW is significantly different from µE

4) H0: µE = µMW = µS = µW

HA: at least one µi is different for *i* = E,MW,S,W

or

H0: BMW = BS = BW = 0

HA: at least one Bi for *i* = MW,S,W ≠ 0

F = 15.24 numdf = k = 3 dendf = N – k – 1 = 39

p-value = 1.025 x 10-6

\* Reject H0; at least one region mean is significantly different from the others and region is a significant predictor for mobility

D. > anova(fit)

Analysis of Variance Table

Response: mobility

Df Sum Sq Mean Sq F value Pr(>F)

region 3 2171.9 723.95 15.24 1.025e-06 \*\*\*

Residuals 39 1852.6 47.50

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Source | Sum of Squares | Df | Mean Square | F |
| Regression  Residuals | 2171.9  1852.6 | 3  39 | 723.95  47.50 | 15.25 |
| Total | 4024.5 | 42 |  |  |

\* p-value = 1.025 x 10-6 \* We reject H0. Therefore, at least one region mean is significantly different from the others and region is a significant predictor for mobility. This result is consistent with the F test in 2C.

#3. > pairwise.t.test(mobility, region, p.adjust.method="bonferroni")

E MW S \* baseline region (E region) is statistically significantly different

MW 0.0082 - - from all other regions; also, regions W and MW are significantly

S 1e-05 0.1114 - different

W 4e-06 0.0101 0.8979

\* p-value for regions MW vs. E individual t-test with the Bonferroni correction (p-value = 0.0082) is m-choose-2 = 4-choose-2 = 6 times larger than the individual t-test without correction (p-value = 0.00136)

#4. A. H0: *o*E = *oMW = oS = oW*

HA: at least one *oi* is significantly different from the others

\* α = 0.05

> leveneTest(mobility, region)

Levene's Test for Homogeneity of Variance (center = median)

Df F value Pr(>F)

group 3 2.4694 0.07625 \* p-value = 0.07625

39

\* We fail to reject H0. Therefore, the constant variance assumption is valid at the α = 0.05 confidence level, but just barely. This assumption would have been violated at the α = 0.10 confidence level, according to Levene’s Test.

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

> plot(fit$residuals)

> abline(h=0)

> qqnorm(fit$residuals)

> qqline(fit$residuals)



\* I think both the constant variance and normality of population *i* about its unknown mean *µi* are reasonable assumptions to make just by looking at the residual and qq normal plot of residuals, respectively. The observations on the residual plot don’t appear to follow any discernible pattern, while the observations on the qq normal plot don’t deviate too much from normality.

#5. H0: avg. mean of regions E and S = avg. mean of regions MW and W

= [µ + (µ + α2)]/2 = [(µ + α1) + (µ + α3)]/2

= α2 – α1 – α3 = 0

= (µ, α1, α2, α3)(0, -1, 1, -1) = 0

Ha: (µ, α1, α2, α3)(0, -1, 1, -1) ≠ 0

> linearHypothesis(fit, c(0,-1,1,-1), rhs=0)

Linear hypothesis test

Hypothesis:

- regionMW + regionS - regionW = 0

Model 1: restricted model

Model 2: mobility ~ region

Res.Df RSS Df Sum of Sq F Pr(>F) \* p-value = 0.01667

1 40 2394.7

2 39 1852.6 1 542.06 11.411 0.001667

\* We reject H0. Therefore, we have evidence that the average mean mobility of northeastern and southeastern cities is significantly different than the average mean mobility of midwestern and western cities.

#6. A. > dummy.1=rep(0, nrow(Angell))

> dummy.1[region=="E"]=1

> dummy.1[region=="W"]=-1 Region D1 D2 D3

> dummy.2=rep(0, nrow(Angell)) E 1 0 0

> dummy.2[region=="MW"]=1 MW 0 1 0

> dummy.2[region=="W"]=-1 S 0 0 1

> dummy.3=rep(0, nrow(Angell)) W -1 -1 -1

> dummy.3[region=="S"]=1

> dummy.3[region=="W"]=-1

B. > fit2 = lm(mobility ~ dummy.1 + dummy.2 + dummy.3)

> summary(fit2)

Coefficients:

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

(Intercept) 27.954 1.118 25.014 < 2e-16 \*\*\*

dummy.1 -12.054 1.972 -6.113 3.61e-07 \*\*\*

dummy.2 -1.896 1.716 -1.105 0.2759

dummy.3 4.504 1.716 2.624 0.0123 \*

Residual standard error: 6.892 on 39 degrees of freedom

Multiple R-squared: 0.5397, Adjusted R-squared: 0.5043

F-statistic: 15.24 on 3 and 39 DF, p-value: 1.025e-06

1) (Intercept) = µ-hat = 27.954 is the average mean mobility of all observations, regardless of region

Dummy.1 = α1 -hat = -12.054 is the difference of mean mobility of cities in region E than µ

Dummy.2 = α2 -hat = -1.896 is the difference of mean mobility in region MW than µ

Dummy.3 = α3 –hat = 4.504 is the difference of mean mobility in region S than µ

2) Region E: Y-hat = µ-hat + α1-hat = 27.954 - 12.054 = 15.9

Region MW: Y-hat = µ-hat + α2 –hat = 27.954 - 1.896 = 26.058

Region S: Y-hat = µ-hat + α3 –hat = 27.954 + 4.504 = 32.458

Region W: Y-hat = µ-hat – (α1 –hat + α2 –hat + α3 -hat) = 27.945 – (-12.054 – 1.896 + 4.504) = 27.954 – (-9.446) = 37.4

\* Yes, these estimated group means are the same as those that were calculated earlier.

3) F = 15.24 df = 3(num) & 39(den) p-value = 1.025x10-6

\* We reject H0. Therefore, we have evidence that at least one regions’ estimated mean mobility is significantly different from the grand mean.

\* Yes, the overall F-test just conducted is the same as the one conducted earlier with the baseline=0 constraint.

C. > contrasts(region) = contr.sum(levels(region))

> contrasts(region)

[,1] [,2] [,3]

E 1 0 0

MW0 1 0

S 0 0 1

W -1 -1 -1

> fit3 = lm(mobility~region)

> summary(fit3)

Coefficients:

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

(Intercept) 27.954 1.118 25.014 < 2e-16 \*\*\*

region1 -12.054 1.972 -6.113 3.61e-07 \*\*\*

region2 -1.896 1.716 -1.105 0.2759

region3 4.504 1.716 2.624 0.0123 \*

Residual standard error: 6.892 on 39 degrees of freedom

Multiple R-squared: 0.5397, Adjusted R-squared: 0.5043

F-statistic: 15.24 on 3 and 39 DF, p-value: 1.025e-06

\* Yes, the results are exactly the same as when the dummy variables were created manually. This is because our sum-to-zero group (region W) is exactly the same as the one created by R (since W is last alphabetically).