

# Answers to Application Activities for Finding the Best Fit in Multiple Regression

---

## 1 Howell (2002) data

Use the HowellChp15Data.sav file (imported into R as `howell`). It has 6 variables in it, and 50 rows. By Crawley's rule of thumb, we should not estimate more than  $50/3$  parameters, or about 16. If we start with a full factorial model we will have one 4-way interaction, four 3-way interactions, seven 2-way interactions, and four single terms, for a total of 16 parameters. We should therefore be safe to run a full factorial model at first.

Start with the full factorial model:

```
model1=lm(overall~exam*grade*knowledge*teach,data=howell)
```

```
summary(model1)
```

Result:  $R^2 = 0.79$  but none of the factors are statistical. Residual standard error is 0.34 on 34 df.

Pull out the largest term first:

```
model2=update(model1,~.-exam:grade:knowledge:teach, data=howell)
```

```
summary(model2)
```

```
anova(model1, model2)
```

Result:  $R^2 = 0.79$ , residual error 0.34 on 35 df, but none of the factors are statistical. The ANOVA finds no difference between models ( $p > .05$ ). Pull out the term `exam:knowledge:teach` because it has the largest  $p$ -value of any of the four 3-way interactions.

```
model3=update(model2, ~. - exam: knowledge:teach, data=howell)
```

```
summary(model3)
```

```
anova(model2, model3)
```

Result:  $R^2 = 0.79$ , residual error 0.33 on 36 df, but none of the factors are statistical. The ANOVA finds no difference between models ( $p > .05$ ). Pull out the term `grade:knowledge:teach` because it has the largest  $p$ -value of any of the remaining three 3-way interactions.

```
model4=update(model3, ~. - grade:knowledge:teach, data=howell)
```

```
summary(model4)
```

```
anova(model3, model4)
```

Result:  $R^2 = 0.78$ , residual error 0.33 on 37 df, but none of the factors are statistical. The ANOVA finds no difference between models ( $p > .05$ ). Pull out the term `exam:grade:teach` because it has the largest  $p$ -value of any of the remaining two 3-way interactions.

```
model5=update(model4, ~. - exam:grade:teach, data=howell)
```

```
summary(model5)
```

```
anova(model4, model5)
```

Result:  $R^2 = 0.78$ , residual error 0.33 on 38 df, but none of the factors are statistical. The

ANOVA finds no difference between models ( $p > .05$ ). Pull out the term

`exam:grade:knowledge`.

```
model6=update(model5,~.- exam:grade:knowledge, data=howell)
```

```
summary(model6)
```

```
anova(model5, model6)
```

Result:  $R^2 = 0.77$ , residual error 0.33 on 39 df, but none of the factors are statistical (getting

closer though! One  $p$ -value is under  $p = .10$ ). The ANOVA finds no difference between models

( $p > .05$ ). Pull out the term `exam:grade` because it has the largest  $p$ -value of any of the six 2-way interactions.

```
model7=update(model6,~.- exam:grade, data=howell)
```

```
summary(model7)
```

```
anova(model6, model7)
```

Result:  $R^2 = 0.77$ , residual error 0.33 on 40 df, but none of the factors are statistical (almost!).

The ANOVA finds no difference between models ( $p > .05$ ). Pull out the term `exam:knowledge`

because it has the largest  $p$ -value of any of the remaining five 2-way interactions.

```
model8=update(model7,~.- exam:knowledge, data=howell)
```

```
summary(model8)
```

```
anova(model7, model8)
```

Result:  $R^2 = 0.77$ , residual error 0.32 on 41 df, but none of the factors are statistical (almost!).

The ANOVA finds no difference between models ( $p > .05$ ). Pull out the term **grade:teach** because it has the largest  $p$ -value of any of the remaining four 2-way interactions.

```
model9=update(model8, ~.- grade:teach, data=howell)
```

```
summary(model9)
```

```
anova(model8, model9)
```

Result:  $R^2 = 0.76$ , residual error 0.32 on 42 df, but none of the factors are statistical. The

ANOVA finds no difference between models ( $p > .05$ ). Pull out the term **grade:knowledge** because it has the largest  $p$ -value of any of the remaining three 2-way interactions.

```
model10=update(model9, ~.- grade:knowledge, data=howell)
```

```
summary(model10)
```

```
anova(model9, model10)
```

Result:  $R^2 = 0.75$ , residual error 0.32 on 43 df, but none of the factors are statistical. The

ANOVA finds no difference between models ( $p > .05$ ). Pull out the term **exam:teach** because it has the largest  $p$ -value of any of the remaining two 2-way interactions.

```
model11=update(model10,~.- exam:teach, data=howell)
```

```
summary(model11)
```

```
anova(model10, model11)
```

Result:  $R^2 = 0.75$ , residual error 0.32 on 44 df, but none of the factors are statistical. The ANOVA finds no difference between models ( $p > .05$ ). Pull out the term **knowledge:teach** because it has the largest  $p$ -value of any of the remaining one 2-way interactions.

```
model12=update(model11,~.- knowledge:teach, data=howell)
```

```
summary(model12)
```

```
anova(model11, model12)
```

Result:  $R^2 = 0.75$ , residual error 0.32 on 45 df, and now two of the factors (**knowledge** and **teach**) are statistical. The ANOVA finds no difference between models ( $p > .05$ ). Pull out the term **exam** because it has the largest  $p$ -value of any of the remaining single terms.

```
model13=update(model12,~.- exam, data=howell)
```

```
summary(model13)
```

```
anova(model12, model13)
```

Result:  $R^2 = 0.75$ , residual error 0.32 on 46 df, and still two of the factors (**knowledge** and **teach**) are statistical. The ANOVA finds no difference between models ( $p > .05$ ). Pull out the term **grade** because it is the only remaining term that is not statistical.

```
model14=update(model13,~.- grade, data=howell)
```

```
summary(model14)
```

```
anova(model13, model14)
```

Result:  $R^2 = 0.74$ , residual error 0.32 on 47 df, and all of the factors in the regression equation (**knowledge** and **teach**) are statistical. The ANOVA finds no difference between models ( $p > .05$ ). This is our minimal adequate model. If we try to take out any of the remaining predictors, our Anova will find a difference between models. Let's try it:

```
model15=update(model14,~.- knowledge, data=howell)
```

```
summary(model15)
```

```
anova(model14, model15)
```

Indeed, the ANOVA shows a  $p$ -value of 0.0001721, meaning there *is* a difference between models, and model 15 has a smaller  $R^2$  (0.64) and a higher residual error (0.37 on 48 df).

Therefore, Model 14 is the model to stick with. The regression equation is exactly the same one that we found in Exercise #4 for Section 7.4.9 of the book, except now we know for sure this is the best regression equation instead of just hoping it is.

Overall course rating =  $-1.30 + .54$  (instructor's knowledge of subject matter) +  $.71$ (teaching skills of the instructor)

To get the relative importance of the terms:

```
library(relaimpo)
```

```
calc.relimp(model14)
```

The importance of Teaching skills of the instructor is highest at 46%, followed by instructor's knowledge at 28%.

Examine regression assumptions:

```
plot(model14,cex=1.5)
```

```
vif(model14)
```

Overall, this model seems to satisfy regression assumptions quite well.

## 2 Dewaele and Pavlenko (2001–3)

Use the data file BEQ.Swear.sav and import it into R as `beq.swear`.

It has 21 variables in it, and 1028 rows. There would be no problem modeling as large a model as we wanted with this large an N! But we are simply going to model a full-factorial model with one response variable and three explanatory variables.

```
model1=lm(swear2~I2freq*weight2*I2speak,data=beq.swear)
```

I will demonstrate the use of the `boot.step(AIC)` command here.

```
library(bootStepAIC)
```

```
boot.stepAIC(model1,data=beq.swear)
```

The procedure selects a final model with all three single terms and two interaction terms:

```
swear2 ~ l2freq + weight2 + l2speak + l2freq:l2speak + weight2:l2speak
```

The statistical significance shows that L2 frequency was not statistically significant very often (only 14% of the time), but it will be retained because it is used in one of the interaction terms.

We can put these terms into a model and find the R<sup>2</sup>:

```
model2=lm(swear2 ~ l2freq + weight2 + l2speak + l2freq:l2speak + weight2:l2speak,
```

```
data=beq.swear)
```

```
summary(model2)
```

The  $R^2 = 0.30$ , meaning this model explains 30% of the variance in scores on how frequently a person swears in their L2. We notice that all of the terms in the model are statistical, except for L2 frequency.

To check on the relative importance:

```
calc.relimp(model2)
```



The term that has a largest importance is L2 frequency of use, with 12%; next comes L2 speaking at 9%, then the weight the person gives to swearing in their L2 at 7%, and last the interaction terms for 1% each. Since these two terms account for such a very small amount of the variance but add complexity to the equation, there might be reasons for the researcher to just pull them out.

The regression equation, using the unstandardized regression coefficients found in the `summary()` command, is:

$$\text{Frequency of swearing in the L2} = 0.24 - 0.01 (\text{frequency of using L2}) + 0.55 (\text{weight given to swearing in L2}) + 0.22 (\text{L2 speaking proficiency}) + 0.08 (\text{interaction between frequency of using L2 and L2 speaking proficiency}) - 0.08 (\text{interaction between weight given to swearing in L2 and L2 speaking proficiency})$$

For regression assumptions, use this code:

```
vif(model2)
```

```
plot(model2,cex=1.5)
```

The VIF numbers are way over 5 (all are over 10) which may indicate problems with multicollinearity. Check on the multicollinearity using R Commander to choose STATISTICS > SUMMARIES > CORRELATION MATRIX. Choose `l2freq`, `l2speak` and `weight2`. The correlation

matrix doesn't show any correlations over  $r = .70$ —not even close, so it seems strange that the VIF numbers are so high.

The Residuals vs. Fitted plot and Scale-Location plot show the data is constrained in a certain way on both sides of the graph, indicating heteroscedasticity. For the Normal Q-Q plot there is a lot of data but it seems to mostly follow the straight line. Lastly the Residuals vs. Leverage plot does not seem to show outliers.

### 3 Larson-Hall (2008)

Use the data file LarsonHall2008.sav, imported into R as `lh2008`. It has 13 variables in it, and 200 rows. By Crawley's rule of thumb, we should not estimate more than  $200/3$  parameters, or about 60, so we are fine. If we start with a full factorial model we will have one 3-way interaction, three 2-way interactions and three single terms. If we add in quadratic (squared terms) for each of the three single terms the total will be 10 parameters.

```
model1=lm(gjtscore~aptscore*rlwscore*totalhrs+l(aptscore^2)+l(rlwscore^2)+l(totalhrs^2),data=lh2008)
```

I will use the `boot.step(AIC)` command here.

```
boot.stepAIC(model1,data= lh2008)
```

The procedure selects a final model with all three single terms, two quadratic terms and two interaction terms:

```
gjtscor ~ aptscore + rlwscore + totalhrs + (aptscore)2 + (rlwscore)2 + aptscore:rlwscore +  
rlwscore:totalhrs
```

The statistical significance shows that rlwscore was not statistically significant consistently (only 33% of the time), but it will be retained because it is used in both of the interaction terms.

We can put these terms into a model and find the R<sup>2</sup>:

```
model2=lm(gjtscor ~ aptscore + rlwscore + totalhrs +  
l(aptscore^2)+l(rlwscore^2)+aptscore:rlwscore + rlwscore:totalhrs,  
data=lh2008)  
summary(model2)
```

The R<sup>2</sup> = 0.25, meaning this model explains 25% of the variance in scores on the grammaticality test. We notice that not all of the terms in the model are statistical; **rlwscore** is not, nor is the interaction between aptitude and **rlwscore**.

To check on the relative importance:

```
calc.relimp(model2)
```

The term that has a largest importance is the squared term of RLW score, with 8%; next comes the RLW score by itself at 7%. The rest of the terms add very little, so it looks like our best predictors for how a person will score on the grammar test are their scores on the pronunciation test. Total hours a person spends studying contributes only 3% to the variance, and aptitude score contributes only 1.5%. The squared term of aptitude adds 2%. Last, the interaction between aptitude and RLW score adds 1% and the interaction between RLW score and total hours of study adds 2%.

The regression equation, using the unstandardized regression coefficients found in the `summary()` command, is:

$$\begin{aligned} \text{Scores on the GJT} = & 0.01 - 3.0 (\text{aptitude score}) - 0.19 (\text{RLW score}) + 0.01 (\text{total hours of study}) + \\ & 0.07 (\text{aptitude score squared}) + 0.01 (\text{RLW score squared}) - 0.02 (\text{aptitude score: RLW score}) - \\ & 0.001 (\text{RLW score:total hours}) \end{aligned}$$

For regression assumptions, use this code:

```
vif(model2)
```

```
plot(model2,cex=1.5)
```

The VIF numbers are way over 5 (all are over 20) which may indicate problems with multicollinearity, but a check on using R Commander to choose STATISTICS > SUMMARIES > CORRELATION MATRIX previous with this data did not find any high correlations at all.

The Residuals vs. Fitted plot and Scale-Location plot do not show any patterns of data being constrained which might indicate heteroscedasticity. For the Normal Q-Q plot the data seems to mostly follow the straight line. Lastly the Residuals vs. Leverage plot does not seem to show outliers.