Paul Johnson April 25, 2013

# Data Management

```
library(foreign)
library(rockchalk)
i <- 38
dat <- read.dta(paste("../student-test2/student-",i,".dta", sep = ""))
```

The variables pprof and pnet are scored as numeric, but really they are factors. So convert them to prevent future mis-understandings.

```
dat$pprof <- factor(dat$pprof, labels = c("NO", "YES"))
dat$pnet <- factor(dat$pnet, labels = c("NO","YES"))
```

```
datsum <- summarize(dat)
```

Table would need some hand customization

```
library(xtable)
print(xtable(datsum$numeric, caption = "Best Automatic Summary Table for Numerics", label =
    "table1"), "latex")
```

| | act | harv | ibs | sal1 | sal2 | sal3 | sat |
|---|---|---|---|---|---|---|---|
| 0% | 4.13 | 1223.00 | 74.07 | 4765.00 | 5666.00 | 148800.00 | 1202.00 |
| 25% | 18.40 | 1519.00 | 92.80 | 16820.00 | 19420.00 | 161500.00 | 1501.00 |
| 50% | 22.34 | 1624.00 | 99.66 | 20220.00 | 23630.00 | 165500.00 | 1604.00 |
| 75% | 25.35 | 1730.00 | 106.70 | 24150.00 | 27500.00 | 169400.00 | 1706.00 |
| 100% | 35.75 | 2092.00 | 129.30 | 34690.00 | 39220.00 | 183600.00 | 2075.00 |
| mean | 22.08 | 1623.00 | 99.71 | 20430.00 | 23400.00 | 165400.00 | 1603.00 |
| sd | 5.05 | 156.80 | 9.62 | 5318.00 | 5753.00 | 5755.00 | 153.60 |
| var | 25.52 | 24590.00 | 92.45 | 28280000.00 | 33090000.00 | 33120000.00 | 23600.00 |
| NA's | 20.00 | 50.00 | 0.00 | 12.00 | 0.00 | 0.00 | 31.00 |
| N | 551.00 | 551.00 | 551.00 | 551.00 | 551.00 | 551.00 | 551.00 |

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

```
print(datsum$factors)
```

```
           gender                        major                       pnet
F             :282.0000    N             :196.0000    NO            :383.0000
M             :269.0000    H             :191.0000    YES           :168.0000
NA's          :  0.0000    S             :164.0000    NA's          :  0.0000
entropy       :  0.9996    NA's          :  0.0000    entropy       :  0.8872
normedEntropy:   0.9996    entropy       :  1.5807    normedEntropy:   0.8872
N             :551.0000    normedEntropy:   0.9973    N             :551.0000
                           N             :551.0000
           pprof
NO            :385.0000
YES           :166.0000
NA's          :  0.0000
entropy       :  0.8828
normedEntropy:   0.8828
N             :551.0000
```

# Aptitude Test Variables

There's severe multicollinearity between the variables harv, sat, and act. It seems clear we can't estimate both sat and harv, and several students noticed that since harv is a summary of the other tests, then there's some reason to suppose sat is a better variable. (I know for a fact that harv = sat + act).

Please find Table 2. I left the Iowa Basic Skills variable in my best model, mainly because I wanted to estimate that coefficient, even though the F test below indicates one can exclude harv and ibs from the "full" model without losing any sleep.

```
m1s <- lm(sal1 ~ sat, data = dat)
m1a <- lm(sal1 ~ act, data = dat)
m1i <- lm(sal1 ~ ibs, data = dat)
m1h <- lm(sal1 ~ harv, data = dat)
m1all <- lm(sal1 ~ sat + act + ibs + harv, data = dat)
m1best <- lm(sal1 ~ sat + act + ibs, data = dat)
```

```
mcDiagnose(m1all)
```

```
The following auxiliary models are being estimated and returned in a list:
sat ~ act + ibs + harv
<environment: 0x1d0e930>
act ~ sat + ibs + harv
<environment: 0x1d0e930>
ibs ~ sat + act + harv
<environment: 0x1d0e930>
harv ~ sat + act + ibs
<environment: 0x1d0e930>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998201 0.8634427 0.1836673 0.9998236
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
5558.431349    7.322935    1.224991 5669.773483
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.29 0.37 1.00
act  0.29 1.00 0.31 0.32
ibs  0.37 0.31 1.00 0.38
harv 1.00 0.32 0.38 1.00
```

```
niceLabels <- c(act = "ACT", sat = "SAT", harv = "Harvard SS", ibs = "Iowa BS",  majorS = "
    Major: Soc.", majorN = "Major: Nat.", majorH = "Major: Hum.", pnetYES = "Parent Network
    : Yes",  pprofYES="Prof. Parents: Yes", genderM = "Gender: Male", "log(harv)"= "ln(
    Harvard SS)",
    "I(harv * harv)"= "Harvard SS$^2$", major2H = "Major 2: Hum.", major2N = "Major 2: Nat.
      ")
outreg(list(m1s, m1a, m1i, m1h, m1all, m1best), tight = TRUE, modelLabels = c("SAT","ACT","
    IBS","Harvard SS", "All", "Best"), varLabels = niceLabels, title = paste("Regression
    with sal1: Student-",i, sep=""), label = "tab:tab2")
```

Could conduct an F test of the hypothesis that $b_{ibs} = b_{harv} = 0$. But which model should I be testing? Test the one with all the variables, to see if *harv* and *ibs* should both be set to 0. To do that, I need to take the data frame used to fit m1all and use it to fit the restricted model. Otherwise, the F test fails.

```
m1alldf <- model.frame(m1all)
m1restricted <- lm(sal1 ~ sat + act, data = m1alldf)
anova(m1restricted, m1all)
```

```
Analysis of Variance Table

Model 1: sal1 ~ sat + act
Model 2: sal1 ~ sat + act + ibs + harv
```

Table 2: Regression with sal1: Student-38

| | SAT<br>Estimate<br>(S.E.) | ACT<br>Estimate<br>(S.E.) | IBS<br>Estimate<br>(S.E.) | Harvard SS<br>Estimate<br>(S.E.) | All<br>Estimate<br>(S.E.) | Best<br>Estimate<br>(S.E.) |
|---|---|---|---|---|---|---|
| (Intercept) | 1366.273 | 14012.493* | 15390.728* | 2147.496 | 3020.383 | 2014.322 |
| | (2321.029) | (1027.3) | (2385.287) | (2367.858) | (2971.642) | (2859.456) |
| SAT | 11.802* | . | . | . | 34.365 | 10.636* |
| | (1.44) | | | | (114.016) | (1.609) |
| ACT | . | 290.75* | . | . | 242.844 | 237.508* |
| | | (45.271) | | | (124.53) | (47.48) |
| Iowa BS | . | . | 50.529* | . | -46.868 | -40.445 |
| | | | (23.802) | | (27.233) | (26.249) |
| Harvard SS | . | . | . | 11.176* | -23.749 | . |
| | | | | (1.451) | (114.016) | |
| N | 508 | 519 | 539 | 491 | 446 | 488 |
| RMSE | 4977.665 | 5180.147 | 5300.667 | 5026.981 | 4947.153 | 4913.319 |
| $R^2$ | 0.117 | 0.074 | 0.008 | 0.108 | 0.151 | 0.164 |
| adj $R^2$ | 0.115 | 0.072 | 0.006 | 0.106 | 0.143 | 0.158 |

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    443 1.0866e+10
2    441 1.0793e+10  2  73036750 1.4921  0.226
```

Noticing this sample size problem, I wondered if I should re-do Table 2 so that all are fitted on the exact same data. Since I exclude harv, should those cases that are missing on harv "come back to life" when I exclude harv from the model? I think so. Still, there is something unappetizing about this. Fitting harv causes a loss of cases, no matter how we look at it. So for the best model and the ones for sat and ibs, I use the sample from the best model, but when harv enters the picture, we lose some cases.

```
m1best <- lm(sal1 ~ sat + act + ibs, data = dat)
dat2 <- model.frame(m1best)
m1s <- lm(sal1 ~ sat, data = dat2)
m1a <- lm(sal1 ~ act, data = dat2)
m1i <- lm(sal1 ~ ibs, data = dat2)
m1h <- lm(sal1 ~ harv, data = dat[row.names(dat2), ])
m1all <- lm(sal1 ~ sat + act + ibs + harv, data = dat[row.names(dat2), ])
```

```
outreg(list(m1s, m1a, m1i, m1h, m1all, m1best), tight = TRUE, modelLabels = c("SAT","ACT","
    IBS","Harvard SS", "All", "Best"), varLabels = niceLabels)
```

| | SAT<br>Estimate<br>(S.E.) | ACT<br>Estimate<br>(S.E.) | IBS<br>Estimate<br>(S.E.) | Harvard SS<br>Estimate<br>(S.E.) | All<br>Estimate<br>(S.E.) | Best<br>Estimate<br>(S.E.) |
|---|---|---|---|---|---|---|
| (Intercept) | 800.397<br>(2408.302) | 13385.305*<br>(1040.358) | 13998.427*<br>(2584.216) | 1128.619<br>(2513.809) | 3020.383<br>(2971.642) | 2014.322<br>(2859.456) |
| SAT | 12.147*<br>(1.493) | . | . | . | 34.365<br>(114.016) | 10.636*<br>(1.609) |
| ACT | . | 313.525*<br>(45.91) | . | . | 242.844<br>(124.53) | 237.508*<br>(47.48) |
| Iowa BS | . | . | 63.304*<br>(25.8) | . | -46.868<br>(27.233) | -40.445<br>(26.249) |
| Harvard SS | . | . | . | 11.726*<br>(1.54) | -23.749<br>(114.016) | . |
| N | 488 | 488 | 488 | 446 | 446 | 488 |
| RMSE | 5029.337 | 5120.991 | 5328.172 | 5031.909 | 4947.153 | 4913.319 |
| $R^2$ | 0.12 | 0.088 | 0.012 | 0.115 | 0.151 | 0.164 |
| adj $R^2$ | 0.118 | 0.086 | 0.01 | 0.113 | 0.143 | 0.158 |

$*p \leq 0.05$

Deciding what's "important"? We have lots of ways. If I've settled on a "best" model, it seems like I should be confined to the variables in that model. And the diagnostics should not depend on harv. Here are the partial and semi-partial correlations.

```
getPartialCor(m1best)
```

```
            sal1
sal1 -1.00000000
sat   0.28782340
act   0.22171546
ibs  -0.06986502
```

```
getDeltaRsquare(m1best)
```

```
The deltaR-square values: the change in the R-square
      observed when a single term is removed.
Same as the square of the 'semi-partial correlation coefficient'
    deltaRsquare
sat  0.075555160
act  0.043245352
ibs  0.004102991
```

I admit, it is tough to conceptualize the scales of the different variables. I suppose I could scale the sat, act, and ibs scores so that they are all on the same 0-100 scale. Then I'll re-run the model. (This is called "percent of maximum" scoring (POMS)). Since we KNOW from previous work that re-scaling a variable has absolutely no substantive impact on the fit, and it is just for convenience of interpretation, this is an innocuous change.

```
dat2$satpoms <- 100*(dat2$sat - min(dat2$sat))/(max(dat2$sat) - min(dat2$sat))
dat2$actpoms <- 100*(dat2$act - min(dat2$act))/(max(dat2$act) - min(dat2$act))
dat2$ibspoms <- 100*(dat2$ibs - min(dat2$ibs))/(max(dat2$ibs) - min(dat2$ibs))
summarize(dat2[, c("satpoms","actpoms","ibspoms")])
```

```
$numerics
     actpoms ibspoms satpoms
0%      0.00    0.00    0.00
25%    45.64   36.01   33.68
50%    57.51   48.47   45.64
75%    66.78   61.75   57.46
100%  100.00  100.00  100.00
```

```
mean    56.80   48.37   45.56
sd      15.99   17.64   17.73
var    255.50  311.30  314.30
NA's     0.00    0.00    0.00
N      488.00  488.00  488.00

$factors
NULL
```

```
m1poms <- lm(sal1 ~ satpoms + actpoms + ibspoms, data = dat2)
summary(m1poms)
```

```
Call:
lm(formula = sal1 ~ satpoms + actpoms + ibspoms, data = dat2)

Residuals:
     Min        1Q    Median        3Q       Max
-14817.3   -3404.2     -51.5    3423.2   13166.0

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12909.77     938.48  13.756  < 2e-16 ***
satpoms        91.59      13.85   6.612 1.01e-10 ***
actpoms        75.10      15.01   5.002 7.94e-07 ***
ibspoms       -21.45      13.92  -1.541    0.124
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4913 on 484 degrees of freedom
Multiple R^2: 0.1635,   Adjusted R^2: 0.1583
F-statistic: 31.54 on 3 and 484 DF,  p-value: < 2.2e-16
```

Oh, one more thing. Recall my point that partial and semi-partial correlations are completely worthless when 1) there is multicollinearity and 2) we are uncertain which variables should be in consideration. Notice how crazy your conclusions would be if you based them on the "full" model.

```
options(scipen = 10)
getPartialCor(m1all)
```

```
            sal1
sal1 -1.000000000
sat   0.014351086
act   0.092463157
ibs  -0.081677677
harv -0.009918179
```

```
getDeltaRsquare(m1all)
```

```
The deltaR-square values: the change in the R-square
      observed when a single term is removed.
Same as the square of the 'semi-partial correlation coefficient'
      deltaRsquare
sat  0.00017493272
act  0.00732282083
ibs  0.00570329432
harv 0.00008354464
```

```
options(scipen = 5)
```

# Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-38

| | Test Scores Only | All Predictors |
|---|---|---|
| | Estimate | Estimate |
| | (S.E.) | (S.E.) |
| (Intercept) | 6066.045 | 2040.439 |
| | (3094.77) | (2835.008) |
| SAT | 10.115* | 10.96* |
| | (1.744) | (1.575) |
| ACT | 234.898* | 222.418* |
| | (51.237) | (46.369) |
| Iowa BS | -41.79 | -39.294 |
| | (28.559) | (25.832) |
| Major: Soc. | . | 1712.961* |
| | | (548.145) |
| Major: Nat. | . | 5363.589* |
| | | (524.128) |
| Prof. Parents: Yes | . | 1443.799* |
| | | (473.646) |
| Parent Network: Yes | . | 849.349 |
| | | (478.811) |
| Gender: Male | . | -904.242* |
| | | (439.109) |
| N | 500 | 500 |
| RMSE | 5405.945 | 4872.851 |
| $R^2$ | 0.129 | 0.3 |
| adj $R^2$ | 0.124 | 0.288 |

$*p \leq 0.05$

```
m2small <- lm(sal2 ~ sat + act + ibs, data = dat)
m2all <- lm(sal2 ~ sat + act + ibs + major + pprof + pnet + gender, data = dat)
outreg(list(m2small, m2all), tight = TRUE, title = paste("Regression with sal2: Student-",i
    , sep=""),modelLabels = c("Test Scores Only","All Predictors"), varLabels = niceLabels,
     label = "table3")
```

Fancy T test. Lets use the big model to find out if $b_{pnetYES} = b_{pprofYES}$.

```
m2allc <- coef(m2all)
m2allv <- vcov(m2all)
numer <- m2allc["pprofYES"] - m2allc["pnetYES"]
names(numer) <- "difference"
denom <- sqrt(m2allv["pprofYES","pprofYES"] + m2allv["pnetYES","pnetYES"] - 2 * m2allv["
    pprofYES","pnetYES"])
print(paste("Fancy T: ", "Numerator = ", numer, "Denominator = ", denom))
```

```
[1] "Fancy T:  Numerator =  594.450079550957 Denominator =  679.977471756474"
```

```
tval <- numer/denom
print("T ratio is")
```

```
[1] "T ratio is"
```

```
tval
```

```
difference
 0.8742203
```

```
print("The two-tailed test would have p value")
```

```
[1] "The two-tailed test would have p value"
```

```
2 * pt(abs(tval), df = m2all$df, lower.tail = FALSE)
```

```
difference
 0.3824257
```

Could I make a function that "just" gets that right and would I be damaging students by ruining their educational experience? This would be very easy if the output had the variable names "pprof" and "pnet", but because I've made them factors, they are now pprofYES and pnetYES, and thus either my function has to be clever or the user's have to be clever in naming their request.

```
fancyT <- function(model, parm1, parm2){
      mc <- coef(model)
      mv <- vcov(model)
      numer <- mc[parm1] - mc[parm2]
      denom <- sqrt(mv[parm1, parm1]
          + mv[parm2, parm2] - 2 * mv[parm1, parm2])
      tval <- numer/denom
      tdf <- model$df
      tvalp <- 2 * pt(abs(tval), df = tdf, lower.tail = FALSE)
    res <- c(numer, denom, tval, tdf, tvalp)
    names(res) <- c("parm1 - parm2", "SE(parm1 - parm2)", "T", "df", "p-value")
    res
 }
fancyT(m2all, parm1 = "pprofYES", parm2 = "pnetYES")
```

```
    parm1 - parm2 SE(parm1 - parm2)                 T                df           p-value
      594.4500796         679.9774718         0.8742203       491.0000000         0.3824257
```

```
m2all <- lm(sal2 ~ sat + act + ibs + major + pprof + pnet + gender, data = dat)
m2alldf <- model.frame(m2all)
m2small <- lm(sal2 ~ sat + act + ibs, data = m2alldf)
anova(m2small, m2all)
```

```
Analysis of Variance Table

Model 1: sal2 ~ sat + act + ibs
Model 2: sal2 ~ sat + act + ibs + major + pprof + pnet + gender
  Res.Df         RSS Df  Sum of Sq      F    Pr(>F)
1    496 14495224766
2    491 11658635301  5 2836589465 23.892 < 2.2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Nonlinear

```
nm1 <- lm(sal3 ~ harv + gender + major + pprof + pnet, data = dat)
nm2 <- lm(sal3 ~ log(harv) + gender + major + pprof + pnet, data = dat)
nm3 <- lm(sal3 ~ harv + I(harv*harv) + gender + major + pprof + pnet, data = dat)
library(rockchalk)
nd <- rockchalk::newdata(nm1, predVals = list(harv = plotSeq(dat$harv, 20)))
nd$m1fit <- predict(nm1, newdata = nd)
nd$m2fit <- predict(nm2, newdata = nd)
nd$m3fit <- predict(nm3, newdata = nd)
```

For the regression table, please see Table 4

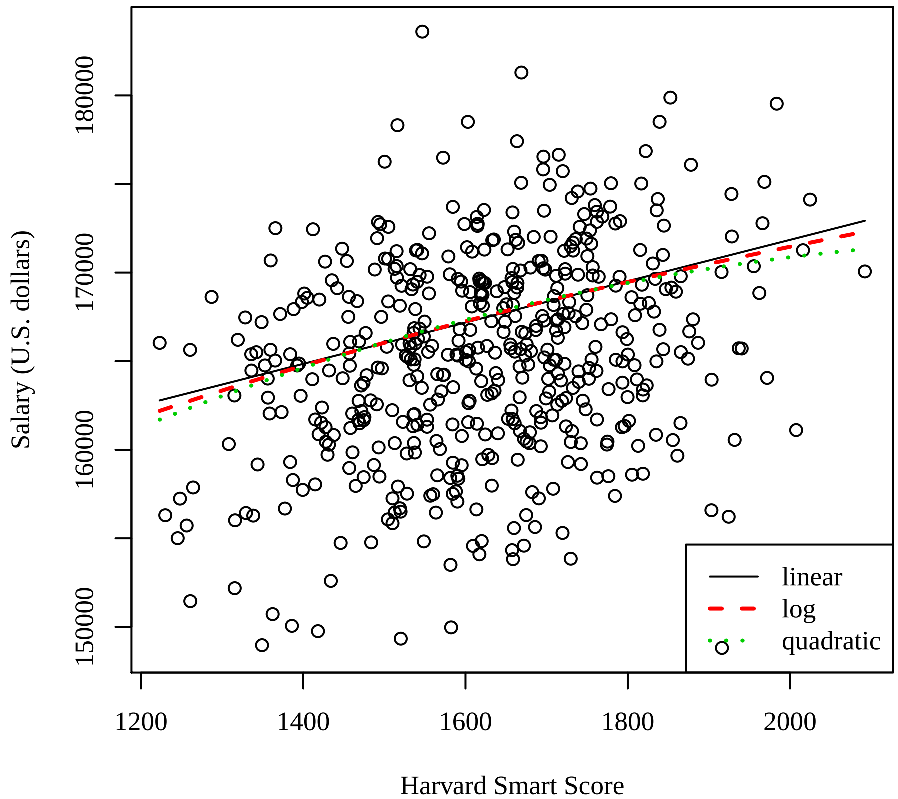
Table 4: Regression with sal3: Student-38

| | Linear Estimate (S.E.) | Log Estimate (S.E.) | Quadratic Estimate (S.E.) |
|---|---|---|---|
| (Intercept) | 143050.495* | 22849.066 | 122517.688* |
| | (2419.091) | (17186.422) | (18266.056) |
| Harvard SS | 11.667* | . | 37.262 |
| | (1.452) | | (22.616) |
| Gender: Male | 723.002 | 742.39 | 763.849 |
| | (455.399) | (454.991) | (456.69) |
| Major: Soc. | 2567.389* | 2554.307* | 2540.202* |
| | (564.689) | (564.124) | (565.035) |
| Major: Nat. | 5463.52* | 5446.568* | 5425.697* |
| | (544.6) | (543.896) | (545.462) |
| Prof. Parents: Yes | 931.7 | 932.398 | 931.763 |
| | (494.177) | (493.694) | (494.034) |
| Parent Network: Yes | 19.526 | 14.582 | 8.236 |
| | (497.388) | (496.919) | (497.344) |
| ln(Harvard SS) | . | 18834.842* | . |
| | | (2323.842) | |
| Harvard SS$^2$ | . | . | -0.008 |
| | | | (0.007) |
| N | 501 | 501 | 501 |
| RMSE | 5068.824 | 5063.878 | 5067.357 |
| $R^2$ | 0.246 | 0.248 | 0.248 |
| adj $R^2$ | 0.237 | 0.239 | 0.237 |

$*p \leq 0.05$

```
outreg(list(nm1, nm2, nm3), tight = TRUE, title = paste("Regression with sal3: Student-",i,
    sep=""), modelLabels = c("Linear", "Log", "Quadratic"), varLabels = niceLabels, label
    = "table4")
```

```
plot(sal3 ~ harv, data = dat, xlab = "Harvard Smart Score", ylab = "Salary (U.S. dollars)")
lines(m1fit ~ harv, data = nd, lty = 1, col = 1)
lines(m2fit ~ harv, data = nd, lty = 2, col = 2, lwd = 2)
lines(m3fit ~ harv, data = nd, lty = 3, col = 3, lwd = 2)
legend("bottomright", legend = c("linear", "log", "quadratic"), lty = c(1,2,3), col = c
    (1,2,3), lwd = c(1,2,2))
```




```
cm1 <- lm(sal2 ~ major, data = dat)
dat$major2  <- relevel(dat$major, ref = "S")
cm2 <- lm(sal2 ~ major2, data = dat)
cm3 <- lm(sal2 ~ sat + act + ibs + major + pprof + pnet + gender, data = dat)
cm4 <- lm(sal2 ~ sat + act + ibs + major2 + pprof + pnet + gender, data = dat)
```

```
outreg(list(cm1, cm2, cm3, cm4), tight = TRUE, title = paste("Categorical Regressions:
    Student-", i, sep=""), modelLabels = c("major", "major2", "major full", "major2 full"),
     varLabels = niceLabels)
```

```
predictOMatic(cm1)
```

```
$major
              fit  major
N (40%) 26068.23     N
H (30%) 21125.72     H
S (30%) 22863.84     S

attr(,"flnames")
[1] "major"
```

```
predictOMatic(cm2)
```

```
$major2
              fit  major2
N (40%) 26068.23      N
H (30%) 21125.72      H
S (30%) 22863.84      S

attr(,"flnames")
[1] "major2"
```

Table 5: Categorical Regressions: Student-38

| | major<br>Estimate<br>(S.E.) | major2<br>Estimate<br>(S.E.) | major full<br>Estimate<br>(S.E.) | major2 full<br>Estimate<br>(S.E.) |
|---|---|---|---|---|
| (Intercept) | 21125.725* | 22863.837* | 2040.439 | 3753.399 |
| | (388.165) | (418.9) | (2835.008) | (2834.624) |
| Major: Soc. | 1738.112* | . | 1712.961* | . |
| | (571.095) | | (548.145) | |
| Major: Nat. | 4942.504* | . | 5363.589* | . |
| | (545.436) | | (524.128) | |
| Major 2: Hum. | . | -1738.112* | . | -1712.961* |
| | | (571.095) | | (548.145) |
| Major 2: Nat. | . | 3204.392* | . | 3650.629* |
| | | (567.72) | | (542.752) |
| SAT | . | . | 10.96* | 10.96* |
| | | | (1.575) | (1.575) |
| ACT | . | . | 222.418* | 222.418* |
| | | | (46.369) | (46.369) |
| Iowa BS | . | . | -39.294 | -39.294 |
| | | | (25.832) | (25.832) |
| Prof. Parents: Yes | . | . | 1443.799* | 1443.799* |
| | | | (473.646) | (473.646) |
| Parent Network: Yes | . | . | 849.349 | 849.349 |
| | | | (478.811) | (478.811) |
| Gender: Male | . | . | -904.242* | -904.242* |
| | | | (439.109) | (439.109) |
| N | 551 | 551 | 500 | 500 |
| RMSE | 5364.543 | 5364.543 | 4872.851 | 4872.851 |
| $R^2$ | 0.134 | 0.134 | 0.3 | 0.3 |
| adj $R^2$ | 0.13 | 0.13 | 0.288 | 0.288 |

$*p \leq 0.05$