Paul Johnson April 25, 2013

# Data Management

```
library(foreign)
library(rockchalk)
i <- 20
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% | 4.77 | 1229.00 | 71.18 | 4113.00 | 4664.00 | 144900.00 | 1207.00 |
| 25% | 18.52 | 1511.00 | 93.25 | 16120.00 | 18680.00 | 161700.00 | 1492.00 |
| 50% | 22.08 | 1609.00 | 100.50 | 19740.00 | 23020.00 | 165300.00 | 1586.00 |
| 75% | 25.27 | 1720.00 | 107.50 | 24090.00 | 27080.00 | 169100.00 | 1699.00 |
| 100% | 36.54 | 2184.00 | 131.40 | 34580.00 | 39110.00 | 181700.00 | 2159.00 |
| mean | 21.92 | 1618.00 | 100.20 | 19930.00 | 22870.00 | 165300.00 | 1597.00 |
| sd | 5.00 | 157.20 | 10.26 | 5588.00 | 5975.00 | 5797.00 | 156.80 |
| var | 24.97 | 24710.00 | 105.40 | 31220000.00 | 35700000.00 | 33610000.00 | 24600.00 |
| NA's | 16.00 | 50.00 | 0.00 | 11.00 | 0.00 | 0.00 | 25.00 |
| N | 542.00 | 542.00 | 542.00 | 542.00 | 542.00 | 542.00 | 542.00 |

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
           gender                        major                      pnet
F             :298.0000   H             :187.0000   NO            :380.0000
M             :244.0000   N             :185.0000   YES           :162.0000
NA's          :  0.0000   S             :170.0000   NA's          :  0.0000
entropy       :  0.9928   NA's          :  0.0000   entropy       :  0.8799
normedEntropy:  0.9928    entropy       :  1.5837   normedEntropy:  0.8799
N             :542.0000   normedEntropy:  0.9992    N             :542.0000
                          N             :542.0000
           pprof
NO            :366.0000
YES           :176.0000
NA's          :  0.0000
entropy       :  0.9094
normedEntropy:  0.9094
N             :542.0000
```

# 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: 0x1ce5738>
act ~ sat + ibs + harv
<environment: 0x1ce5738>
ibs ~ sat + act + harv
<environment: 0x1ce5738>
harv ~ sat + act + ibs
<environment: 0x1ce5738>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998296 0.8568236 0.1924534 0.9998338
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
5867.259314    6.984391    1.238319 6016.579221
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.38 0.38 1.00
act  0.38 1.00 0.35 0.41
ibs  0.38 0.35 1.00 0.38
harv 1.00 0.41 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-20

| | 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) | -1951.97 | 13489.204* | 10878.447* | -2060.672 | 800.208 | -901.542 |
| | (2350.501) | (1067.725) | (2341.964) | (2411.094) | (2975.682) | (2797.942) |
| SAT | 13.709* | . | . | . | 34.469 | 12.114* |
| | (1.466) | | | | (121.191) | (1.703) |
| ACT | . | 291.976* | . | . | 196.976 | 158.604* |
| | | (47.582) | | | (130.667) | (52.87) |
| Iowa BS | . | . | 90.364* | . | -37.137 | -20.048 |
| | | | (23.266) | | (26.473) | (25.433) |
| Harvard SS | . | . | . | 13.682* | -22.487 | . |
| | | | | (1.484) | (121.234) | |
| N | 507 | 516 | 531 | 481 | 444 | 492 |
| RMSE | 5172.892 | 5418.798 | 5514.779 | 5112.422 | 5167.09 | 5170.355 |
| $R^2$ | 0.148 | 0.068 | 0.028 | 0.151 | 0.154 | 0.158 |
| adj $R^2$ | 0.146 | 0.066 | 0.026 | 0.149 | 0.147 | 0.153 |

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    441 1.1775e+10
2    439 1.1721e+10  2  54284211 1.0166 0.3627
```

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 Estimate (S.E.) | ACT Estimate (S.E.) | IBS Estimate (S.E.) | Harvard SS Estimate (S.E.) | All Estimate (S.E.) | Best Estimate (S.E.) |
|---|---|---|---|---|---|---|
| (Intercept) | -1763.945 | 13391.948* | 11846.84* | -1262.525 | 800.208 | -901.542 |
| | (2403.628) | (1111.625) | (2452.907) | (2554.512) | (2975.682) | (2797.942) |
| SAT | 13.574* | . | . | . | 34.469 | 12.114* |
| | (1.501) | | | | (121.191) | (1.703) |
| ACT | . | 296.085* | . | . | 196.976 | 158.604* |
| | | (49.613) | | | (130.667) | (52.87) |
| Iowa BS | . | . | 80.256* | . | -37.137 | -20.048 |
| | | | (24.429) | | (26.473) | (25.433) |
| Harvard SS | . | . | . | 13.156* | -22.487 | . |
| | | | | (1.574) | (121.234) | |
| N | 492 | 492 | 492 | 444 | 444 | 492 |
| RMSE | 5207.204 | 5430.861 | 5563.831 | 5203.859 | 5167.09 | 5170.355 |
| $R^2$ | 0.143 | 0.068 | 0.022 | 0.136 | 0.154 | 0.158 |
| adj $R^2$ | 0.141 | 0.066 | 0.02 | 0.135 | 0.147 | 0.153 |

$*p \leq 0.05$

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   0.30651688
act   0.13456387
ibs  -0.03566036
```

```
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  0.087259383
act  0.015518410
ibs  0.001071466
```

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%    43.19   35.51   29.77
50%    54.27   48.03   39.61
75%    63.89   59.78   51.36
100%  100.00  100.00  100.00
```

```
mean    53.78    47.66    40.57
sd      15.55    17.07    16.44
var    241.80   291.20   270.10
NA's     0.00     0.00     0.00
N      492.00   492.00   492.00

$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
-16490.9  -3241.3     41.2   3775.3  12831.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13048.58     942.46  13.845  < 2e-16 ***
satpoms       115.36      16.22   7.114 4.05e-12 ***
actpoms        50.39      16.80   3.000  0.00284 **
ibspoms       -12.08      15.32  -0.788  0.43092
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5170 on 488 degrees of freedom
Multiple R^2: 0.1585,   Adjusted R^2: 0.1533
F-statistic: 30.64 on 3 and 488 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   0.013573267
act   0.071762096
ibs  -0.066802855
harv -0.008852484
```

```
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  0.00015581325
act  0.00437711673
ibs  0.00379042639
harv 0.00006627045
```

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

# Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-20

| | Test Scores Only | All Predictors |
|---|---|---|
| | Estimate | Estimate |
| | (S.E.) | (S.E.) |
| (Intercept) | 2465.215 | -905.474 |
| | (2995.491) | (2838.411) |
| SAT | 13.086* | 12.368* |
| | (1.818) | (1.693) |
| ACT | 126.237* | 153.445* |
| | (56.685) | (52.862) |
| Iowa BS | -32.921 | -23.151 |
| | (27.259) | (25.502) |
| Major: Soc. | . | 1765.572* |
| | | (574.458) |
| Major: Nat. | . | 4891.704* |
| | | (562.013) |
| Prof. Parents: Yes | . | 638.475 |
| | | (498.353) |
| Parent Network: Yes | . | 867.397 |
| | | (508.261) |
| Gender: Male | . | 602.498 |
| | | (466.497) |
| N | 501 | 501 |
| RMSE | 5574.823 | 5177.39 |
| $R^2$ | 0.139 | 0.265 |
| adj $R^2$ | 0.134 | 0.253 |

$*p \leq 0.05$

```
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 =  -228.922243832021 Denominator =  688.33149202552"
```

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

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

```
tval
```

```
difference
-0.3325756
```

```
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
 0.7395963
```

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
     -228.9222438         688.3314920        -0.3325756       492.0000000         0.7395963
```

```
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    497 15446088691
2    492 13188240058  5 2257848633 16.846 2.215e-15 ***
---
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

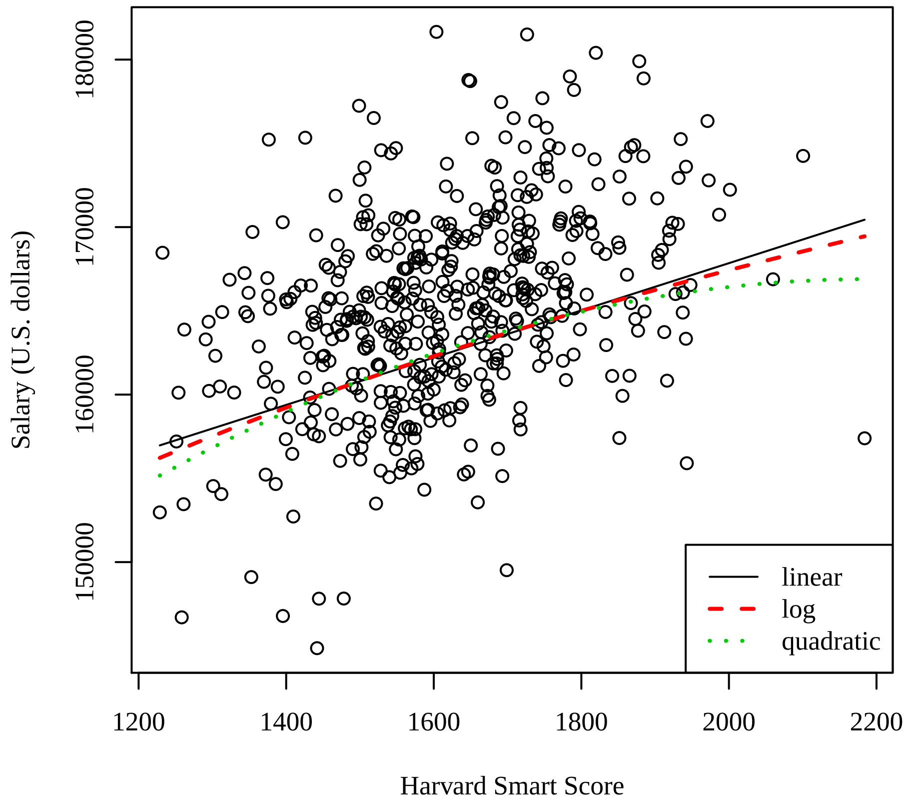
Table 4: Regression with sal3: Student-20

| | Linear Estimate (S.E.) | Log Estimate (S.E.) | Quadratic Estimate (S.E.) |
|---|---|---|---|
| (Intercept) | 139638.495* | -7432.923 | 106223.761* |
| | (2345.413) | (16894.32) | (16838.475) |
| Harvard SS | 14.104* | . | 55.332* |
| | (1.419) | | (20.623) |
| Gender: Male | 500.668 | 517.54 | 544.92 |
| | (448.219) | (447.363) | (447.377) |
| Major: Soc. | 1588.495* | 1591.071* | 1593.667* |
| | (552.22) | (551.095) | (550.517) |
| Major: Nat. | 5036.183* | 5045.43* | 5063.034* |
| | (536.938) | (535.811) | (535.444) |
| Prof. Parents: Yes | 836.762 | 841.779 | 857.524 |
| | (474.541) | (473.552) | (473.186) |
| Parent Network: Yes | 168.893 | 176.111 | 186.169 |
| | (489.328) | (488.322) | (487.89) |
| ln(Harvard SS) | . | 23005.32* | . |
| | | (2286.728) | |
| Harvard SS$^2$ | . | . | -0.013* |
| | | | (0.006) |
| N | 492 | 492 | 492 |
| RMSE | 4930.047 | 4919.975 | 4914.792 |
| $R^2$ | 0.296 | 0.299 | 0.301 |
| adj $R^2$ | 0.287 | 0.29 | 0.291 |

$*p \leq 0.05$

```
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
H (30%) 20628.97     H
N (30%) 25610.98     N
S (30%) 22363.60     S

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

```
predictOMatic(cm2)
```

```
$major2
              fit  major2
H (30%) 20628.97      H
N (30%) 25610.98      N
S (30%) 22363.60      S

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

Table 5: Categorical Regressions: Student-20

| | 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) | 20628.966* | 22363.596* | -905.474 | 860.097 |
| | (409.954) | (429.963) | (2838.411) | (2818.788) |
| Major: Soc. | 1734.63* | . | 1765.572* | . |
| | (594.08) | | (574.458) | |
| Major: Nat. | 4982.017* | . | 4891.704* | . |
| | (581.327) | | (562.013) | |
| Major 2: Hum. | . | -1734.63* | . | -1765.572* |
| | | (594.08) | | (574.458) |
| Major 2: Nat. | . | 3247.387* | . | 3126.133* |
| | | (595.607) | | (576.025) |
| SAT | . | . | 12.368* | 12.368* |
| | | | (1.693) | (1.693) |
| ACT | . | . | 153.445* | 153.445* |
| | | | (52.862) | (52.862) |
| Iowa BS | . | . | -23.151 | -23.151 |
| | | | (25.502) | (25.502) |
| Prof. Parents: Yes | . | . | 638.475 | 638.475 |
| | | | (498.353) | (498.353) |
| Parent Network: Yes | . | . | 867.397 | 867.397 |
| | | | (508.261) | (508.261) |
| Gender: Male | . | . | 602.498 | 602.498 |
| | | | (466.497) | (466.497) |
| N | 542 | 542 | 501 | 501 |
| RMSE | 5606.037 | 5606.037 | 5177.39 | 5177.39 |
| $R^2$ | 0.123 | 0.123 | 0.265 | 0.265 |
| adj $R^2$ | 0.12 | 0.12 | 0.253 | 0.253 |

$*p \leq 0.05$