Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 3
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%	2.74	1132.00	55.81	3490.00	5608.00	147700.00	1114.00
25%	18.40	1523.00	93.80	16850.00	19620.00	161500.00	1489.00
50%	21.60	1623.00	100.60	20230.00	23310.00	165200.00	1594.00
75%	25.06	1725.00	107.80	23940.00	27550.00	169200.00	1698.00
100%	38.66	2235.00	133.20	36230.00	39240.00	180900.00	2203.00
mean	21.71	1626.00	100.60	20460.00	23300.00	165100.00	1598.00
sd	5.24	160.90	10.67	5510.00	5910.00	5684.00	159.60
var	27.41	25890.00	113.90	30360000.00	34920000.00	32310000.00	25480.00
NA's	17.00	64.00	0.00	9.00	0.00	0.00	29.00
N	555.00	555.00	555.00	555.00	555.00	555.00	555.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet                   pprof
M             :283.0000   H             :200.0000   NO            :384.000   NO            :385
    .0000
F             :272.0000   S             :179.0000   YES           :171.000   YES           :170
    .0000
NA's          :  0.0000   N             :176.0000   NA's          :  0.000   NA's          :  0
    .0000
entropy       :  0.9997   NA's          :  0.0000   entropy       :  0.891   entropy       :  0
    .8889
normedEntropy:  0.9997   entropy       :  1.5826   normedEntropy:  0.891   normedEntropy:  0
    .8889
N             :555.0000   normedEntropy:  0.9985   N             :555.000   N             :555
    .0000
                          N             :555.0000
```

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: 0x1f7ffa8>
act ~ sat + ibs + harv
<environment: 0x1f7ffa8>
ibs ~ sat + act + harv
<environment: 0x1f7ffa8>
harv ~ sat + act + ibs
<environment: 0x1f7ffa8>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998605 0.8938904 0.3097037 0.9998645
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
7169.887309    9.424220    1.448653 7377.709484
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.42 0.45 1.00
act  0.42 1.00 0.48 0.44
ibs  0.45 0.48 1.00 0.46
harv 1.00 0.44 0.46 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-3

	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)	-577.868	14100.025*	8901.89*	-213.082	28.736	237.406
	(2258.729)	(964.692)	(2178.554)	(2369.528)	(2809.057)	(2650.858)
SAT	13.091*	.	.	.	0.716	10.702*
	(1.406)				(128.887)	(1.64)
ACT	.	287.427*	.	.	128.928	147.173*
		(43.223)			(139.942)	(50.327)
Iowa BS	.	.	115.003*	.	17.082	-2.901
			(21.558)		(26.734)	(25.097)
Harvard SS	.	.	.	12.787*	8.994	.
				(1.45)	(128.836)	
N	517	529	546	484	443	500
RMSE	5102.056	5226.89	5376.112	5135.465	5059.092	5014.586
R^2	0.144	0.077	0.05	0.139	0.143	0.152
adj R^2	0.142	0.076	0.048	0.137	0.135	0.147

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    440 1.1221e+10
2    438 1.1210e+10  2  10588492 0.2069 0.8132
```

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)	-16.748	14184.785*	10095.968*	640.059	28.736	237.406
	(2299.604)	(990.714)	(2255.212)	(2462.422)	(2809.057)	(2650.858)
SAT	12.675*	.	.	.	0.716	10.702*
	(1.433)				(128.887)	(1.64)
ACT	.	278.621*	.	.	128.928	147.173*
		(44.453)			(139.942)	(50.327)
Iowa BS	.	.	100.957*	.	17.082	-2.901
			(22.366)		(26.734)	(25.097)
Harvard SS	.	.	.	12.11*	8.994	.
				(1.508)	(128.836)	
N	500	500	500	443	443	500
RMSE	5052.57	5232.288	5326.867	5087.903	5059.092	5014.586
R^2	0.136	0.073	0.039	0.127	0.143	0.152
adj R^2	0.134	0.071	0.037	0.126	0.135	0.147

$*p \leq 0.05$

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 -1.000000000
sat   0.281159354
act   0.130190261
ibs  -0.005191063
```

```
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 7.278363e-02
act 1.461994e-02
ibs 2.285013e-05
```

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%    43.53   48.72   34.50
50%    52.45   57.53   43.99
75%    61.86   66.76   53.17
100%  100.00  100.00  100.00
```

```
mean    52.66    57.41    44.30
sd      14.67    13.77    14.49
var    215.20   189.60   210.00
NA's     0.00     0.00     0.00
N      500.00   500.00   500.00

$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
-13358.4  -3151.2   -318.8   3553.3  14939.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12402.311   1065.842  11.636  < 2e-16 ***
satpoms       116.514     17.857   6.525 1.68e-10 ***
actpoms        52.865     18.077   2.924  0.00361 **
ibspoms        -2.247     19.435  -0.116  0.90801
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5015 on 496 degrees of freedom
Multiple R^2: 0.1521,   Adjusted R^2: 0.1469
F-statistic: 29.65 on 3 and 496 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 -1.0000000000
sat   0.0002653927
act   0.0439784161
ibs   0.0305174602
harv  0.0033356049
```

```
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  0.00000006034596
act  0.00166031393854
ibs  0.00079867794310
harv 0.00000953288368
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-3

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	5202.869	31.438
	(2897.641)	(2717.472)
SAT	10.93*	10.835*
	(1.776)	(1.624)
ACT	142.184*	152.446*
	(54.976)	(50.258)
Iowa BS	-26.26	-5.179
	(27.204)	(25.008)
Major: Soc.	.	1844.269*
		(544.211)
Major: Nat.	.	5198.956*
		(546.412)
Prof. Parents: Yes	.	1483.025*
		(485.523)
Parent Network: Yes	.	1361.856*
		(485.17)
Gender: Male	.	-315.289
		(447.411)
N	509	509
RMSE	5507.608	5025.337
R^2	0.119	0.273
adj R^2	0.113	0.262

$*p \leq 0.05$

```
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 =  121.168825328281 Denominator =  673.774389779744"
```

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

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

```
tval
```

```
difference
 0.1798359
```

```
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
 0.8573543
```

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
121.1688253	673.7743898	0.1798359	500.0000000	0.8573543

```
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    505 15318540272
2    500 12627008158  5 2691532114 21.316 < 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

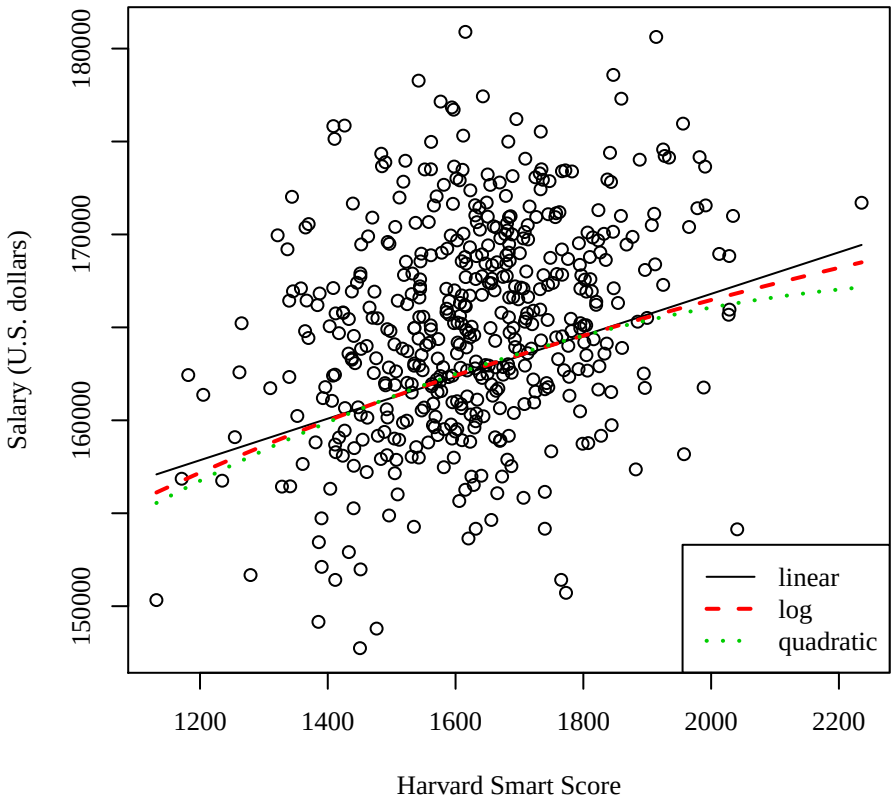
Table 4: Regression with sal3: Student-3

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	144169.89*	27949.586	126207.563*
	(2362.834)	(16776.176)	(15465.669)
Harvard SS	11.187*	.	33.38
	(1.408)		(18.937)
Gender: Male	259.237	260.001	263.848
	(453.829)	(453.307)	(453.667)
Major: Soc.	2565.358*	2583.407*	2601.906*
	(543.704)	(543.056)	(544.379)
Major: Nat.	4970.296*	4964.403*	4959.474*
	(559.583)	(558.899)	(559.439)
Prof. Parents: Yes	1213.136*	1216.639*	1219.032*
	(492.151)	(491.591)	(491.983)
Parent Network: Yes	-210.807	-223.981	-240.294
	(493.221)	(492.594)	(493.665)
ln(Harvard SS)	.	18190.408*	.
		(2267.07)	
Harvard SS^2	.	.	-0.007
			(0.006)
N	491	491	491
RMSE	5002.998	4997.241	5001.029
R^2	0.229	0.231	0.231
adj R^2	0.219	0.221	0.22

$*p \leq 0.05$

```
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 (40%) 21019.93     H
S (30%) 23281.70     S
N (30%) 25923.80     N

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
H (40%) 21019.93      H
S (30%) 23281.70      S
N (30%) 25923.80      N

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

Table 5: Categorical Regressions: Student-3

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	21019.928* (393.512)	23281.698* (415.956)	31.438 (2717.472)	1875.706 (2753.166)
Major: Soc.	2261.77* (572.6)	.	1844.269* (544.211)	.
Major: Nat.	4903.87* (575.17)	.	5198.956* (546.412)	.
Major 2: Hum.	.	-2261.77* (572.6)	.	-1844.269* (544.211)
Major 2: Nat.	.	2642.1* (590.751)	.	3354.687* (561.406)
SAT	.	.	10.835* (1.624)	10.835* (1.624)
ACT	.	.	152.446* (50.258)	152.446* (50.258)
Iowa BS	.	.	-5.179 (25.008)	-5.179 (25.008)
Prof. Parents: Yes	.	.	1483.025* (485.523)	1483.025* (485.523)
Parent Network: Yes	.	.	1361.856* (485.17)	1361.856* (485.17)
Gender: Male	.	.	-315.289 (447.411)	-315.289 (447.411)
N	555	555	509	509
RMSE	5565.105	5565.105	5025.337	5025.337
R^2	0.116	0.116	0.273	0.273
adj R^2	0.113	0.113	0.262	0.262

$*p \leq 0.05$