Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 23
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%	7.51	1223.00	59.11	6283.00	8200.00	148700.00	1208.00
25%	18.70	1503.00	93.43	16490.00	18460.00	161200.00	1491.00
50%	21.99	1606.00	100.70	20090.00	23020.00	164900.00	1587.00
75%	25.45	1719.00	106.80	23510.00	26840.00	168800.00	1698.00
100%	42.75	2100.00	134.20	39450.00	46420.00	184000.00	2076.00
mean	22.04	1614.00	100.30	20030.00	22870.00	164900.00	1597.00
sd	4.81	162.10	10.25	5479.00	5964.00	5492.00	161.10
var	23.13	26290.00	105.10	30010000.00	35560000.00	30160000.00	25940.00
NA's	20.00	53.00	0.00	13.00	0.00	0.00	27.00
N	569.00	569.00	569.00	569.00	569.00	569.00	569.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                  major                   pnet                   pprof
F             :295.000   H             :218.0000   NO            :400.0000   NO            :386
    .0000
M             :274.000   N             :180.0000   YES           :169.0000   YES           :183
    .0000
NA's          :  0.000   S             :171.0000   NA's          :  0.0000   NA's          :  0
    .0000
entropy       :  0.999   NA's          :  0.0000   entropy       :  0.8776   entropy       :  0
    .9061
normedEntropy:  0.999   entropy       :  1.5768   normedEntropy:  0.8776   normedEntropy:  0
    .9061
N             :569.000   normedEntropy:  0.9949   N             :569.0000   N             :569
    .0000
                         N             :569.0000
```

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: 0x23d70d8>
act ~ sat + ibs + harv
<environment: 0x23d70d8>
ibs ~ sat + act + harv
<environment: 0x23d70d8>
harv ~ sat + act + ibs
<environment: 0x23d70d8>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998377 0.8482877 0.2555119 0.9998414
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
6163.040188    6.591423    1.343205 6305.046432
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.40 0.38 1.00
act  0.40 1.00 0.44 0.43
ibs  0.38 0.44 1.00 0.39
harv 1.00 0.43 0.39 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-23

	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)	-1702.31	13701.199*	13966.067*	-2822.326	-835.917	-405.839
	(2184.809)	(1077.58)	(2265.343)	(2262.403)	(2807.718)	(2639.193)
SAT	13.624*	.	.	.	150.077	13.608*
	(1.362)				(115.045)	(1.566)
ACT	.	287.126*	.	.	284.779*	132.821*
		(47.764)			(129.473)	(53.204)
Iowa BS	.	.	60.451*	.	-53.576*	-41.365
			(22.473)		(26.493)	(24.842)
Harvard SS	.	.	.	14.138*	-135.639	.
				(1.396)	(114.851)	
N	529	536	556	505	462	510
RMSE	5038.698	5308.365	5447.986	5077.934	5058.456	4993.669
R^2	0.16	0.063	0.013	0.169	0.194	0.184
adj R^2	0.158	0.062	0.011	0.168	0.187	0.179

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F  Pr(>F)
1    459 1.1821e+10
2    457 1.1694e+10  2 127689520 2.4951 0.08361 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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)	-2608.918	13847.452*	12963.976*	-3512.886	-835.917	-405.839
	(2212.656)	(1102.104)	(2375.826)	(2356.42)	(2807.718)	(2639.193)
SAT	14.223*	.	.	.	150.077	13.608*
	(1.381)				(115.045)	(1.566)
ACT	.	282.213*	.	.	284.779*	132.821*
		(48.869)			(129.473)	(53.204)
Iowa BS	.	.	70.798*	.	-53.576*	-41.365
			(23.569)		(26.493)	(24.842)
Harvard SS	.	.	.	14.604*	-135.639	.
				(1.455)	(114.851)	
N	510	510	510	462	462	510
RMSE	5018.301	5344.719	5469.008	5087.533	5058.456	4993.669
R^2	0.173	0.062	0.017	0.18	0.194	0.184
adj R^2	0.171	0.06	0.016	0.178	0.187	0.179

$*p \leq 0.05$

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   0.36041147
act   0.11030293
ibs  -0.07382157
```

```
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  0.121811750
act  0.010049706
ibs  0.004470979
```

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%    31.73   45.73   32.39
50%    41.20   55.34   43.62
75%    51.04   63.43   55.90
100%  100.00  100.00  100.00
```

```
mean    41.19    54.85    44.48
sd      13.76    13.70    18.57
var    189.20   187.80   344.90
NA's     0.00     0.00     0.00
N      510.00   510.00   510.00

$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
-14560  -3356    -114    3563   15229

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14588.46     968.48  15.063   <2e-16 ***
satpoms       118.02      13.58   8.691   <2e-16 ***
actpoms        46.81      18.75   2.496   0.0129 *
ibspoms       -31.05      18.65  -1.665   0.0965 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4994 on 506 degrees of freedom
Multiple R^2: 0.1841,   Adjusted R^2: 0.1792
F-statistic: 38.05 on 3 and 506 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 -1.00000000
sat   0.06090893
act   0.10234914
ibs  -0.09417553
harv -0.05516081
```

```
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   0.003000008
act   0.008528801
ibs   0.007209270
harv  0.002458844
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-23

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	1271.187	-849.013
	(2876.757)	(2647.65)
SAT	13.831*	13.833*
	(1.702)	(1.551)
ACT	113.767*	123.951*
	(57.73)	(52.73)
Iowa BS	-29.114	-38.296
	(27.051)	(24.917)
Major: Soc.	.	2115.535*
		(536.803)
Major: Nat.	.	5309.938*
		(527.921)
Prof. Parents: Yes	.	786.612
		(468.342)
Parent Network: Yes	.	999.384*
		(489.414)
Gender: Male	.	-111.043
		(438.711)
N	523	523
RMSE	5496.982	5000.392
R^2	0.16	0.312
adj R^2	0.155	0.301

$*p \leq 0.05$

```
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 =  -212.772138095386 Denominator =  692.693721562343"
```

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

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

```
tval
```

```
difference
-0.3071663
```

```
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
 0.7588412
```

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
    -212.7721381        692.6937216        -0.3071663       514.0000000         0.7588412
```

```
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    519 15682523471
2    514 12852012711  5 2830510760 22.64 < 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

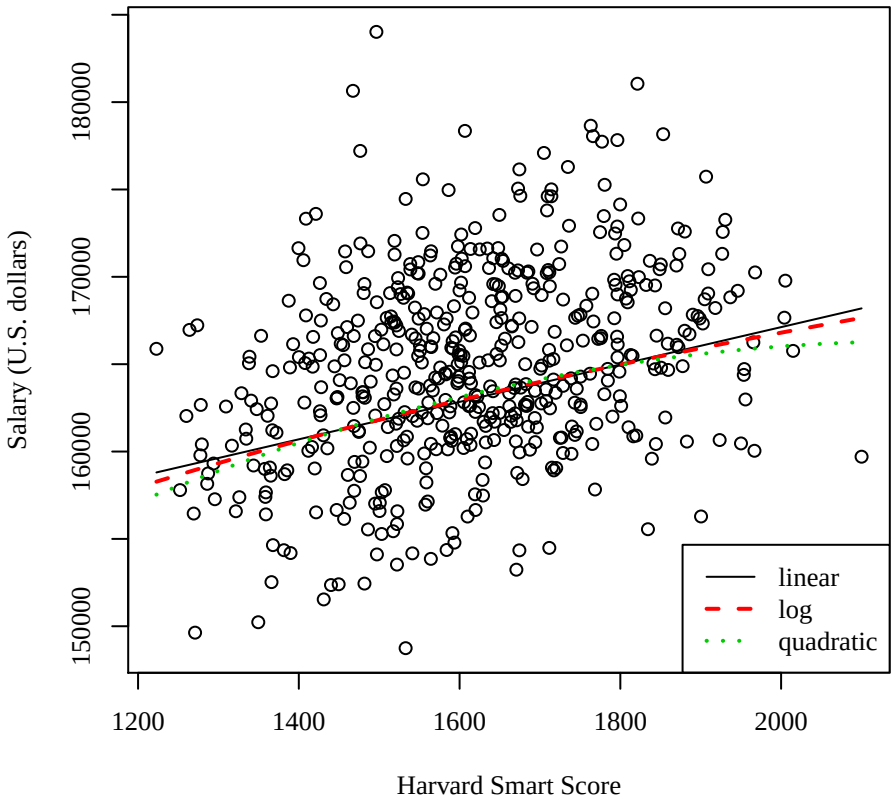
Table 4: Regression with sal3: Student-23

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	145712.82*	35043.852*	120421.51*
	(2186.026)	(15710.12)	(16972.747)
Harvard SS	10.704*	.	42.227*
	(1.327)		(21.02)
Gender: Male	-344.56	-338.02	-327.045
	(431.442)	(430.948)	(431.067)
Major: Soc.	1211.018*	1209.266*	1207.9*
	(525.318)	(524.717)	(524.674)
Major: Nat.	3725.103*	3700.867*	3656.88*
	(515.667)	(515.089)	(517.028)
Prof. Parents: Yes	1941.061*	1931.274*	1912.699*
	(460.457)	(459.921)	(460.276)
Parent Network: Yes	-221.022	-222.552	-231.267
	(466.759)	(466.217)	(466.233)
ln(Harvard SS)	.	17334.731*	.
		(2127.652)	
Harvard SS^2	.	.	-0.01
			(0.006)
N	516	516	516
RMSE	4880.401	4874.798	4874.382
R^2	0.212	0.214	0.216
adj R^2	0.203	0.205	0.205

$*p \leq 0.05$

```
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%) 20397.59     H
N (30%) 25887.02     N
S (30%) 22842.37     S

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
H (40%) 20397.59      H
N (30%) 25887.02      N
S (30%) 22842.37      S

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

Table 5: Categorical Regressions: Student-23

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	20397.59* (373.678)	22842.374* (421.918)	-849.013 (2647.65)	1266.522 (2684.072)
Major: Soc.	2444.784* (563.605)	.	2115.535* (536.803)	.
Major: Nat.	5489.429* (555.652)	.	5309.938* (527.921)	.
Major 2: Hum.	.	-2444.784* (563.605)	.	-2115.535* (536.803)
Major 2: Nat.	.	3044.645* (589.176)	.	3194.403* (559.094)
SAT	.	.	13.833* (1.551)	13.833* (1.551)
ACT	.	.	123.951* (52.73)	123.951* (52.73)
Iowa BS	.	.	-38.296 (24.917)	-38.296 (24.917)
Prof. Parents: Yes	.	.	786.612 (468.342)	786.612 (468.342)
Parent Network: Yes	.	.	999.384* (489.414)	999.384* (489.414)
Gender: Male	.	.	-111.043 (438.711)	-111.043 (438.711)
N	569	569	523	523
RMSE	5517.294	5517.294	5000.392	5000.392
R^2	0.147	0.147	0.312	0.312
adj R^2	0.144	0.144	0.301	0.301

$*p \leq 0.05$