Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 28
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%	7.55	1041.00	70.12	3679.00	6644.00	146400.00	1029.00
25%	18.30	1494.00	92.18	16690.00	19360.00	161100.00	1477.00
50%	22.13	1611.00	99.11	20540.00	23620.00	165100.00	1590.00
75%	25.18	1719.00	105.80	24600.00	27880.00	169300.00	1698.00
100%	35.04	2059.00	131.20	36680.00	39410.00	181600.00	2038.00
mean	21.86	1607.00	99.19	20700.00	23690.00	165200.00	1587.00
sd	5.13	171.60	10.47	5688.00	5986.00	5922.00	169.60
var	26.36	29460.00	109.60	32350000.00	35840000.00	35070000.00	28770.00
NA's	16.00	47.00	0.00	10.00	0.00	0.00	35.00
N	561.00	561.00	561.00	561.00	561.00	561.00	561.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet
F             :285.0000   S             :195.0000   NO            :388.0000
M             :276.0000   N             :192.0000   YES           :173.0000
NA's          :  0.0000   H             :174.0000   NA's          :  0.0000
entropy       :  0.9998   NA's          :  0.0000   entropy       :  0.8913
normedEntropy:  0.9998   entropy       :  1.5832   normedEntropy:  0.8913
N             :561.0000   normedEntropy:  0.9989   N             :561.0000
                          N             :561.0000
          pprof
NO            :389.0000
YES           :172.0000
NA's          :  0.0000
entropy       :  0.8892
normedEntropy:  0.8892
N             :561.0000
```

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: 0x15030a8>
act ~ sat + ibs + harv
<environment: 0x15030a8>
ibs ~ sat + act + harv
<environment: 0x15030a8>
harv ~ sat + act + ibs
<environment: 0x15030a8>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998637 0.8677410 0.2014071 0.9998676
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
7339.007240    7.560920    1.252203 7551.042755
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.48 0.40  1.0
act  0.48 1.00 0.37  0.5
ibs  0.40 0.37 1.00  0.4
harv 1.00 0.50 0.40  1.0
```

```
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-28

	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)	-2638.911	13655.438*	12875.486*	-2460.308	-1404.892	-542.651
	(2158.621)	(1029.179)	(2281.409)	(2182.747)	(2748.355)	(2613.153)
SAT	14.772*	.	.	.	-66.762	13.005*
	(1.353)				(121.783)	(1.588)
ACT	.	322.469*	.	.	102.307	164.312*
		(45.875)			(129.943)	(52.234)
Iowa BS	.	.	78.879*	.	-21.541	-29.049
			(22.868)		(25.605)	(24.088)
Harvard SS	.	.	.	14.377*	79.579	.
				(1.351)	(121.769)	
N	517	535	551	504	461	506
RMSE	5202.526	5461.175	5632.063	5191.781	5207.081	5164.622
R^2	0.188	0.085	0.021	0.184	0.209	0.204
adj R^2	0.186	0.083	0.019	0.183	0.202	0.199

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    458 1.2394e+10
2    456 1.2364e+10  2  30287446 0.5585 0.5724
```

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)	-2519.911	13235.827*	12615.148*	-3034.634	-1404.892	-542.651
	(2169.899)	(1058.451)	(2376.429)	(2281.214)	(2748.355)	(2613.153)
SAT	14.69*	.	.	.	-66.762	13.005*
	(1.36)				(121.783)	(1.588)
ACT	.	346.31*	.	.	102.307	164.312*
		(47.216)			(129.943)	(52.234)
Iowa BS	.	.	82.245*	.	-21.541	-29.049
			(23.772)		(25.605)	(24.088)
Harvard SS	.	.	.	14.787*	79.579	.
				(1.411)	(121.769)	
N	506	506	506	461	461	506
RMSE	5205.893	5491.701	5709.959	5242.126	5207.081	5164.622
R^2	0.188	0.096	0.023	0.193	0.209	0.204
adj R^2	0.186	0.095	0.021	0.191	0.202	0.199

$*p \leq 0.05$

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   0.34336122
act   0.13903520
ibs  -0.05374718
```

```
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  0.106383372
act  0.015689807
ibs  0.002305995
```

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%    39.04   36.45   44.22
50%    53.07   47.68   55.66
75%    64.09   58.91   66.31
100%  100.00  100.00  100.00
```

```
mean    51.88    47.94    55.29
sd      18.83    17.51    16.90
var    354.50   306.40   285.70
NA's     0.00     0.00     0.00
N      506.00   506.00   506.00

$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
-18070.9  -3548.5     91.7   3593.9  13019.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12049.04     896.44  13.441  < 2e-16 ***
satpoms       131.09      16.00   8.191 2.16e-15 ***
actpoms        45.17      14.36   3.146  0.00176 **
ibspoms       -17.74      14.71  -1.206  0.22840
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5165 on 502 degrees of freedom
Multiple R^2: 0.204, Adjusted R^2: 0.1993
F-statistic:  42.9 on 3 and 502 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  -0.02566346
act   0.03684457
ibs  -0.03936573
harv  0.03058975
```

```
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  0.0005212471
act  0.0010751380
ibs  0.0012275447
harv 0.0007407740
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-28

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	3878.432	536.789
	(2765.91)	(2692.761)
SAT	13.278*	12.758*
	(1.683)	(1.597)
ACT	150.066*	163.965*
	(55.522)	(52.408)
Iowa BS	-44.594	-31.294
	(25.673)	(24.295)
Major: Soc.	.	1212.012*
		(568.708)
Major: Nat.	.	4478.034*
		(570.138)
Prof. Parents: Yes	.	670.616
		(501.284)
Parent Network: Yes	.	989.764*
		(493.339)
Gender: Male	.	116.727
		(458.5)
N	515	515
RMSE	5526.745	5176.101
R^2	0.176	0.285
adj R^2	0.172	0.273

$*p \leq 0.05$

```
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 =  -319.147493742078 Denominator =  715.713070723945"
```

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

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

```
tval
```

```
difference
-0.4459154
```

```
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
 0.6558491
```

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
-319.1474937	715.7130707	-0.4459154	506.0000000	0.6558491

```
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    511 15608450818
2    506 13556761576  5 2051689242 15.316 4.95e-14 ***
---
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

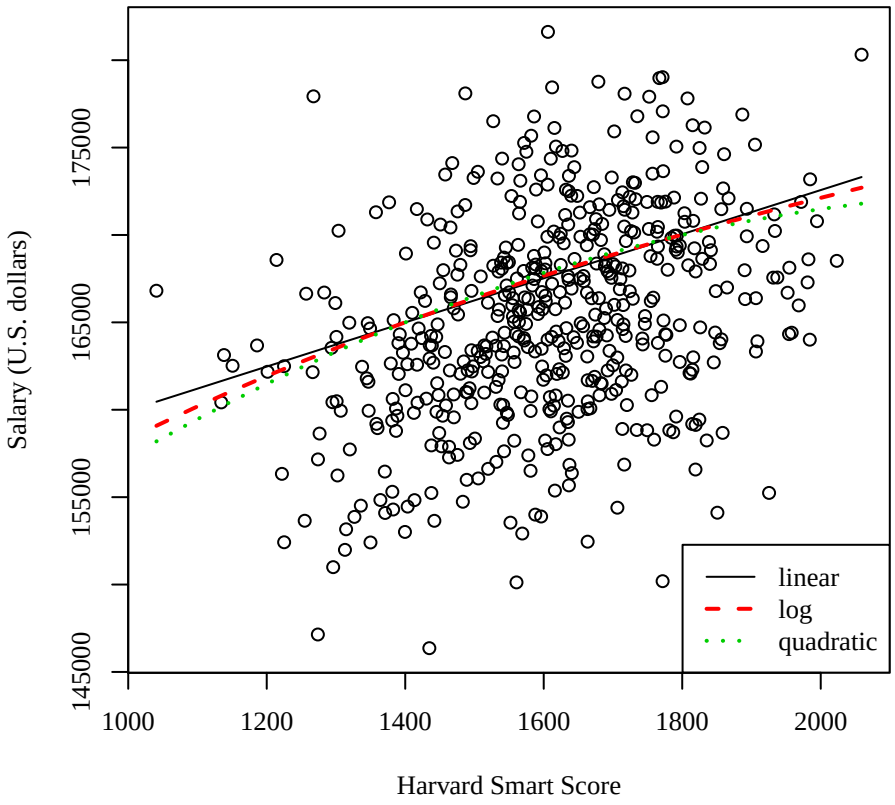
Table 4: Regression with sal3: Student-28

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	141772.027*	14816.173	120489.739*
	(2173.973)	(15073.795)	(13983.124)
Harvard SS	12.624*	.	39.7*
	(1.301)		(17.623)
Gender: Male	-639.377	-627.259	-620.084
	(443.111)	(442.566)	(442.689)
Major: Soc.	1981.654*	1938.684*	1889.176*
	(553.709)	(553.128)	(556.209)
Major: Nat.	5549.696*	5569.946*	5587.086*
	(551.571)	(550.739)	(551.359)
Prof. Parents: Yes	1468.636*	1461.656*	1443.987*
	(479.256)	(478.521)	(478.875)
Parent Network: Yes	1420.647*	1395.63*	1367.757*
	(483.721)	(482.75)	(484.286)
ln(Harvard SS)	.	19963.212*	.
		(2039.107)	
Harvard SS^2	.	.	-0.008
			(0.006)
N	514	514	514
RMSE	4995.693	4988.845	4988.939
R^2	0.302	0.304	0.305
adj R^2	0.294	0.295	0.295

$*p \leq 0.05$

```
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
S (30%) 22853.41     S
N (30%) 26346.26     N
H (30%) 21681.72     H

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
S (30%) 22853.41      S
N (30%) 26346.26      N
H (30%) 21681.72      H

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

Table 5: Categorical Regressions: Student-28

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	21681.72* (429.082)	22853.406* (405.32)	536.789 (2692.761)	1748.8 (2696.879)
Major: Soc.	1171.686* (590.25)	.	1212.012* (568.708)	.
Major: Nat.	4664.539* (592.421)	.	4478.034* (570.138)	.
Major 2: Hum.	.	-1171.686* (590.25)	.	-1212.012* (568.708)
Major 2: Nat.	.	3492.853* (575.443)	.	3266.022* (552.943)
SAT	.	.	12.758* (1.597)	12.758* (1.597)
ACT	.	.	163.965* (52.408)	163.965* (52.408)
Iowa BS	.	.	-31.294 (24.295)	-31.294 (24.295)
Prof. Parents: Yes	.	.	670.616 (501.284)	670.616 (501.284)
Parent Network: Yes	.	.	989.764* (493.339)	989.764* (493.339)
Gender: Male	.	.	116.727 (458.5)	116.727 (458.5)
N	561	561	515	515
RMSE	5659.981	5659.981	5176.101	5176.101
R^2	0.109	0.109	0.285	0.285
adj R^2	0.106	0.106	0.273	0.273

$*p \leq 0.05$