Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 45
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%	7.88	1145.00	70.54	4202.00	6728.00	147500.00	1130.00
25%	18.72	1513.00	93.03	16610.00	19430.00	161300.00	1490.00
50%	22.05	1628.00	98.87	20440.00	23530.00	165300.00	1608.00
75%	25.31	1742.00	107.00	23810.00	27120.00	169600.00	1717.00
100%	36.42	2153.00	128.20	34860.00	38190.00	179800.00	2127.00
mean	21.95	1626.00	99.69	20250.00	23230.00	165400.00	1602.00
sd	4.88	161.30	10.04	5085.00	5650.00	5946.00	159.30
var	23.78	26020.00	100.70	25860000.00	31930000.00	35350000.00	25370.00
NA's	12.00	46.00	0.00	7.00	0.00	0.00	32.00
N	550.00	550.00	550.00	550.00	550.00	550.00	550.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet                   pprof
F             :284.0000   S             :207.000   NO            :385.0000   NO            :377
    .0000
M             :266.0000   N             :177.000   YES           :165.0000   YES           :173
    .0000
NA's          :  0.0000   H             :166.000   NA's          :  0.0000   NA's          :  0
    .0000
entropy       :  0.9992   NA's          :  0.000   entropy       :  0.8813   entropy       :  0
    .8984
normedEntropy:  0.9992   entropy       :  1.579   normedEntropy:  0.8813   normedEntropy:  0
    .8984
N             :550.0000   normedEntropy:  0.996   N             :550.0000   N             :550
    .0000
                          N             :550.000
```

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: 0x3070e88>
act ~ sat + ibs + harv
<environment: 0x3070e88>
ibs ~ sat + act + harv
<environment: 0x3070e88>
harv ~ sat + act + ibs
<environment: 0x3070e88>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998463 0.8652250 0.2474483 0.9998504
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
6507.220048    7.419773    1.328812 6686.367792
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.43 0.45 1.00
act  0.43 1.00 0.38 0.45
ibs  0.45 0.38 1.00 0.46
harv 1.00 0.45 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-45

	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)	4703.434*	12403.866*	13797.445*	3141.178	7696.384*	6935.047*
	(2144.124)	(954.389)	(2157.502)	(2182.524)	(2616.436)	(2473.583)
SAT	9.669*	.	.	.	5.156	7.179*
	(1.331)				(112.245)	(1.536)
ACT	.	356.463*	.	.	268.782*	274.966*
		(42.508)			(121.687)	(48.379)
Iowa BS	.	.	64.794*	.	-51.076*	-43.291
			(21.544)		(25.218)	(23.925)
Harvard SS	.	.	.	10.517*	2.081	.
				(1.335)	(112.249)	
N	512	531	543	499	462	502
RMSE	4803.704	4785.657	5047.543	4816.609	4703.282	4668.182
R^2	0.094	0.117	0.016	0.111	0.14	0.148
adj R^2	0.092	0.116	0.015	0.109	0.133	0.143

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    459 1.0200e+10
2    457 1.0109e+10  2  90816368 2.0527 0.1296
```

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)	4808.577*	12661.687*	14243.643*	4820.276*	7696.384*	6935.047*
	(2165.015)	(974.092)	(2240.187)	(2299.488)	(2616.436)	(2473.583)
SAT	9.575*	.	.	.	5.156	7.179*
	(1.345)				(112.245)	(1.536)
ACT	.	341.054*	.	.	268.782*	274.966*
		(43.348)			(121.687)	(48.379)
Iowa BS	.	.	59.094*	.	-51.076*	-43.291
			(22.335)		(25.218)	(23.925)
Harvard SS	.	.	.	9.424*	2.081	.
				(1.409)	(112.249)	
N	502	502	502	462	462	502
RMSE	4808.623	4760.243	5011.354	4825.958	4703.282	4668.182
R^2	0.092	0.11	0.014	0.089	0.14	0.148
adj R^2	0.09	0.108	0.012	0.087	0.133	0.143

$*p \leq 0.05$

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   0.20505600
act   0.24680931
ibs  -0.08081638
```

```
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  0.037411736
act  0.055287225
ibs  0.005603395
```

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%    37.95   39.27   36.07
50%    49.35   49.52   47.82
75%    60.51   63.47   58.84
100%  100.00  100.00  100.00
```

```
mean    49.23    50.70    47.27
sd      17.19    17.37    16.01
var    295.50   301.70   256.40
NA's     0.00     0.00     0.00
N      502.00   502.00   502.00

$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
-13590.3  -3011.6     -5.8   3311.2  14282.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 14160.63     803.54  17.623  < 2e-16 ***
satpoms        71.59      15.31   4.675 3.78e-06 ***
actpoms        78.48      13.81   5.684 2.25e-08 ***
ibspoms       -24.98      13.81  -1.809    0.071 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4668 on 498 degrees of freedom
Multiple R^2: 0.1477,   Adjusted R^2: 0.1425
F-statistic: 28.76 on 3 and 498 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   0.0021487323
act   0.1027758195
ibs  -0.0943221824
harv  0.0008671952
```

```
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  0.0000039703735
act  0.0091803324140
ibs  0.0077192207693
harv 0.0000006466938
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-45

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	9848.132*	6570.105*
	(2810.418)	(2489.608)
SAT	7.3*	7.051*
	(1.738)	(1.541)
ACT	260.734*	276.571*
	(54.937)	(48.386)
Iowa BS	-41.281	-46.215
	(27.162)	(24.02)
Major: Soc.	.	2563.436*
		(512.629)
Major: Nat.	.	6230.674*
		(531.631)
Prof. Parents: Yes	.	1379.602*
		(445.374)
Parent Network: Yes	.	1057.012*
		(453.201)
Gender: Male	.	154.01
		(418.971)
N	508	508
RMSE	5318.367	4676.458
R^2	0.113	0.321
adj R^2	0.108	0.31

$*p \leq 0.05$

```
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 =  322.590198937927 Denominator =  642.335047292636"
```

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

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

```
tval
```

```
difference
 0.5022148
```

```
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
 0.6157379
```

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
322.5901989	642.3350473	0.5022148	499.0000000	0.6157379

```
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    504 14255656162
2    499 10912758652  5 3342897510 30.572 < 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

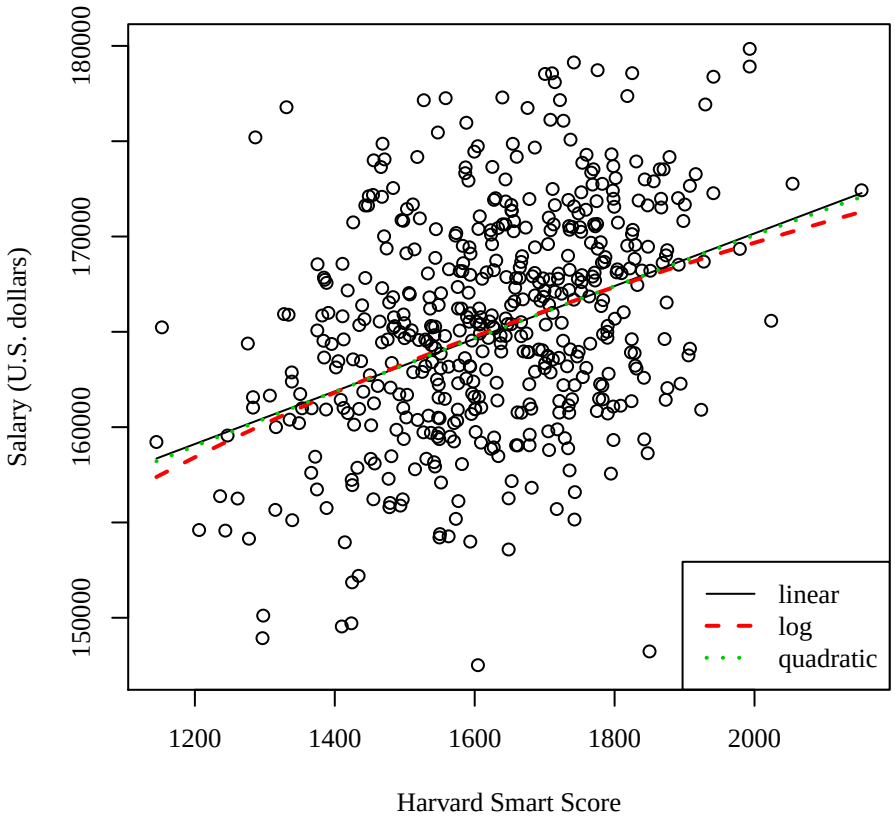
Table 4: Regression with sal3: Student-45

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	140294.097*	-24.648	138365.29*
	(2217.581)	(15791.32)	(16007.226)
Harvard SS	13.812*	.	16.227
	(1.34)		(19.897)
Gender: Male	-223.004	-239.319	-226.376
	(431.054)	(431.143)	(432.371)
Major: Soc.	2246.313*	2242.965*	2245.165*
	(528.305)	(528.363)	(528.914)
Major: Nat.	6079.66*	6076.175*	6078.4*
	(541.704)	(541.732)	(542.341)
Prof. Parents: Yes	605.122	615.639	607.218
	(465.398)	(465.463)	(466.178)
Parent Network: Yes	-306.129	-304.758	-306.095
	(470.6)	(470.629)	(471.067)
ln(Harvard SS)	.	22030.873*	.
		(2138.436)	
Harvard SS^2	.	.	-0.001
			(0.006)
N	504	504	504
RMSE	4813.735	4813.992	4818.514
R^2	0.321	0.321	0.321
adj R^2	0.313	0.313	0.312

$*p \leq 0.05$

```
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 (40%) 22975.44     S
N (30%) 26262.49     N
H (30%) 20317.29     H

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
S (40%) 22975.44      S
N (30%) 26262.49      N
H (30%) 20317.29      H

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

Table 5: Categorical Regressions: Student-45

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	20317.293* (399.313)	22975.437* (357.588)	6570.105* (2489.608)	9133.541* (2522.85)
Major: Soc.	2658.144* (536.023)	.	2563.436* (512.629)	.
Major: Nat.	5945.195* (555.871)	.	6230.674* (531.631)	.
Major 2: Hum.	.	-2658.144* (536.023)	.	-2563.436* (512.629)
Major 2: Nat.	.	3287.051* (526.698)	.	3667.238* (498.217)
SAT	.	.	7.051* (1.541)	7.051* (1.541)
ACT	.	.	276.571* (48.386)	276.571* (48.386)
Iowa BS	.	.	-46.215 (24.02)	-46.215 (24.02)
Prof. Parents: Yes	.	.	1379.602* (445.374)	1379.602* (445.374)
Parent Network: Yes	.	.	1057.012* (453.201)	1057.012* (453.201)
Gender: Male	.	.	154.01 (418.971)	154.01 (418.971)
N	550	550	508	508
RMSE	5144.794	5144.794	4676.458	4676.458
R^2	0.174	0.174	0.321	0.321
adj R^2	0.171	0.171	0.31	0.31

$*p \leq 0.05$