Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 25
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%	7.50	1037.00	71.59	-1012.00	5029.00	147400.00	1023.00
25%	18.36	1527.00	93.42	17160.00	19560.00	161300.00	1506.00
50%	22.14	1632.00	100.30	20820.00	23600.00	165300.00	1608.00
75%	25.58	1730.00	107.50	24260.00	27620.00	169300.00	1707.00
100%	36.26	2158.00	134.60	37460.00	38860.00	182000.00	2130.00
mean	22.04	1628.00	100.30	20590.00	23490.00	165200.00	1608.00
sd	5.25	162.50	10.17	5442.00	5933.00	6000.00	158.80
var	27.59	26410.00	103.40	29620000.00	35200000.00	36000000.00	25230.00
NA's	18.00	50.00	0.00	6.00	0.00	0.00	25.00
N	508.00	508.00	508.00	508.00	508.00	508.00	508.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet
M             :263.0000   H             :179.0000   NO            :362.0000
F             :245.0000   N             :167.0000   YES           :146.0000
NA's          :  0.0000   S             :162.0000   NA's          :  0.0000
entropy       :  0.9991   NA's          :  0.0000   entropy       :  0.8653
normedEntropy:  0.9991   entropy       :  1.5837   normedEntropy:  0.8653
N             :508.0000   normedEntropy:  0.9992   N             :508.0000
                          N             :508.0000
          pprof
NO            :351.0000
YES           :157.0000
NA's          :  0.0000
entropy       :  0.8921
normedEntropy:  0.8921
N             :508.0000
```

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: 0x17b8460>
act ~ sat + ibs + harv
<environment: 0x17b8460>
ibs ~ sat + act + harv
<environment: 0x17b8460>
harv ~ sat + act + ibs
<environment: 0x17b8460>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998357 0.8634810 0.2447183 0.9998397
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
6085.392212    7.324988    1.324009 6238.030063
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.39 0.41 1.00
act  0.39 1.00 0.41 0.41
ibs  0.41 0.41 1.00 0.42
harv 1.00 0.41 0.42 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-25

	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)	-361.199	13391.472*	10193.499*	-374.398	-759.03	-134.153
	(2329.005)	(1019.238)	(2365.127)	(2418.531)	(2928.925)	(2794.552)
SAT	13.01*	.	.	.	-139.493	10.273*
	(1.443)				(118.866)	(1.634)
ACT	.	329.884*	.	.	70.97	201.957*
		(44.99)			(127.356)	(50.722)
Iowa BS	.	.	103.559*	.	-2.654	-2
			(23.442)		(27.563)	(25.858)
Harvard SS	.	.	.	12.873*	149.941	.
				(1.479)	(118.9)	
N	479	484	502	453	416	461
RMSE	4989.402	5164.62	5344.406	5089.546	4933.286	4912.582
R^2	0.146	0.1	0.038	0.144	0.183	0.174
adj R^2	0.144	0.098	0.036	0.142	0.176	0.169

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    413 1.0042e+10
2    411 1.0003e+10  2  38990190 0.801 0.4496
```

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)	-5.896	13457.311*	10170.646*	-677.082	-759.03	-134.153
	(2369.094)	(1044.851)	(2464.346)	(2505.635)	(2928.925)	(2794.552)
SAT	12.831*	.	.	.	-139.493	10.273*
	(1.467)				(118.866)	(1.634)
ACT	.	325.627*	.	.	70.97	201.957*
		(46.25)			(127.356)	(50.722)
Iowa BS	.	.	103.937*	.	-2.654	-2
			(24.391)		(27.563)	(25.858)
Harvard SS	.	.	.	13.058*	149.941	.
				(1.531)	(118.9)	
N	461	461	461	416	416	461
RMSE	4993.671	5124.325	5290.305	5016.602	4933.286	4912.582
R^2	0.143	0.097	0.038	0.149	0.183	0.174
adj R^2	0.141	0.096	0.036	0.147	0.176	0.169

$*p \leq 0.05$

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   0.282082039
act   0.183104769
ibs  -0.003617148
```

```
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 7.139575e-02
act 2.864982e-02
ibs 1.080564e-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%    37.80   34.92   43.69
50%    51.04   46.19   52.92
75%    62.76   57.00   61.97
100%  100.00  100.00  100.00
```

```
mean    50.40    45.90    52.82
sd      17.96    16.04    14.34
var    322.60   257.30   205.50
NA's     0.00     0.00     0.00
N      461.00   461.00   461.00

$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
-15484.3  -3116.4     31.1   3466.6  17415.2

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 11741.800    972.586  12.073  < 2e-16 ***
satpoms       113.742     18.096   6.285 7.64e-10 ***
actpoms        58.083     14.588   3.982 7.96e-05 ***
ibspoms        -1.261     16.304  -0.077    0.938
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4913 on 457 degrees of freedom
Multiple R^2: 0.1741,   Adjusted R^2: 0.1687
F-statistic: 32.12 on 3 and 457 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  -0.057788948
act   0.027476986
ibs  -0.004749445
harv  0.062083656
```

```
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   0.00273605900
act   0.00061694906
ibs   0.00001841954
harv  0.00315947442
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-25

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	2665.581	-166.277
	(3097.483)	(2861.43)
SAT	10.647*	10.29*
	(1.811)	(1.635)
ACT	185.238*	208.869*
	(56.34)	(50.756)
Iowa BS	-3.098	-2.134
	(28.702)	(25.867)
Major: Soc.	.	1868.862*
		(561.913)
Major: Nat.	.	5516.15*
		(554.398)
Prof. Parents: Yes	.	909.09
		(495.332)
Parent Network: Yes	.	1030.403*
		(506.622)
Gender: Male	.	-450.934
		(458.404)
N	465	465
RMSE	5473.223	4923.385
R^2	0.145	0.316
adj R^2	0.14	0.304

$*p \leq 0.05$

```
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 =  -121.31350325342 Denominator =  707.045162399911"
```

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

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

```
tval
```

```
difference
-0.1715782
```

```
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
 0.8638453
```

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
-121.3135033	707.0451624	-0.1715782	456.0000000	0.8638453

```
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    461 13809794462
2    456 11053310555  5 2756483907 22.744 < 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

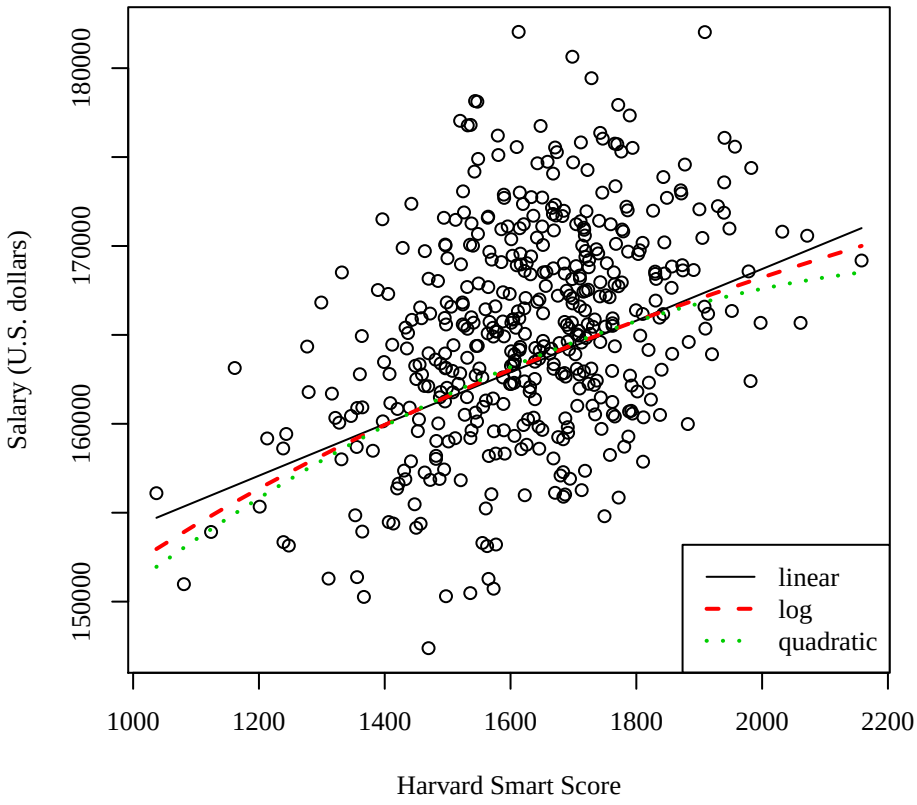
Table 4: Regression with sal3: Student-25

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	139432.305*	-8741.205	115572.858*
	(2473.868)	(17153.123)	(14597.946)
Harvard SS	14.53*	.	44.507*
	(1.466)		(18.136)
Gender: Male	213.473	221.494	236.94
	(478.112)	(477.051)	(477.397)
Major: Soc.	705.981	664.792	627.446
	(579.164)	(577.704)	(579.98)
Major: Nat.	4641.768*	4601.129*	4562.261*
	(577.571)	(576.201)	(578.444)
Prof. Parents: Yes	1316.062*	1297.109*	1281.7*
	(509.994)	(508.865)	(509.429)
Parent Network: Yes	-416.057	-385.336	-355.349
	(527.006)	(525.843)	(527.259)
ln(Harvard SS)	.	23254.26*	.
		(2316.263)	
Harvard SS^2	.	.	-0.009
			(0.006)
N	458	458	458
RMSE	5068.151	5056.56	5058.346
R^2	0.289	0.292	0.293
adj R^2	0.279	0.283	0.282

$*p \leq 0.05$

```
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%) 21175.16     H
N (30%) 26545.56     N
S (30%) 22913.04     S

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
H (40%) 21175.16      H
N (30%) 26545.56      N
S (30%) 22913.04      S

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

Table 5: Categorical Regressions: Student-25

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	21175.159* (411.05)	22913.038* (432.079)	-166.277 (2861.43)	1702.585 (2853.181)
Major: Soc.	1737.879* (596.367)	.	1868.862* (561.913)	.
Major: Nat.	5370.405* (591.662)	.	5516.15* (554.398)	.
Major 2: Hum.	.	-1737.879* (596.367)	.	-1868.862* (561.913)
Major 2: Nat.	.	3632.526* (606.461)	.	3647.288* (566.959)
SAT	.	.	10.29* (1.635)	10.29* (1.635)
ACT	.	.	208.869* (50.756)	208.869* (50.756)
Iowa BS	.	.	-2.134 (25.867)	-2.134 (25.867)
Prof. Parents: Yes	.	.	909.09 (495.332)	909.09 (495.332)
Parent Network: Yes	.	.	1030.403* (506.622)	1030.403* (506.622)
Gender: Male	.	.	-450.934 (458.404)	-450.934 (458.404)
N	508	508	465	465
RMSE	5499.468	5499.468	4923.385	4923.385
R^2	0.144	0.144	0.316	0.316
adj R^2	0.141	0.141	0.304	0.304

$*p \leq 0.05$