Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 7
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%	6.60	1027.00	71.42	6423.00	7472.00	147800.00	1006.00
25%	18.48	1525.00	94.13	16780.00	19740.00	161000.00	1504.00
50%	21.93	1625.00	100.10	20370.00	22930.00	164600.00	1603.00
75%	25.57	1733.00	107.20	23860.00	27200.00	169500.00	1711.00
100%	35.79	2049.00	125.60	38450.00	43610.00	181000.00	2031.00
mean	22.04	1625.00	100.30	20530.00	23440.00	165100.00	1603.00
sd	4.98	159.70	10.22	5463.00	5600.00	5796.00	157.90
var	24.78	25500.00	104.50	29840000.00	31360000.00	33590000.00	24950.00
NA's	16.00	72.00	0.00	11.00	0.00	0.00	32.00
N	564.00	564.00	564.00	564.00	564.00	564.00	564.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
            gender              major                    pnet                    pprof
F              :283     H           :193.0000     NO          :398.0000     NO           :392
      .0000
M              :281     N           :186.0000     YES         :166.0000     YES          :172
      .0000
NA's           :  0     S           :185.0000     NA's        :  0.0000     NA's         :  0
      .0000
entropy        :  1     NA's        :  0.0000     entropy     :  0.8742     entropy      :  0
      .8873
normedEntropy:  1      entropy      :  1.5847     normedEntropy:  0.8742    normedEntropy:  0
      .8873
N              :564     normedEntropy:  0.9998    N           :564.0000     N            :564
      .0000
                        N           :564.0000
```

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: 0x247c0a8>
act ~ sat + ibs + harv
<environment: 0x247c0a8>
ibs ~ sat + act + harv
<environment: 0x247c0a8>
harv ~ sat + act + ibs
<environment: 0x247c0a8>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998536 0.8794579 0.2241021 0.9998579
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
6831.716930    8.295855    1.288829 7038.836596
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.46 0.42 1.00
act  0.46 1.00 0.39 0.49
ibs  0.42 0.39 1.00 0.42
harv 1.00 0.49 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-7

	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)	-868.537	13576.904*	10575.894*	-1663.119	-1298.826	-846.268
	(2273.797)	(1033.961)	(2251.43)	(2364.755)	(2881.853)	(2715.138)
SAT	13.303*	.	.	.	-242.809	11.127*
	(1.412)				(125.302)	(1.651)
ACT	.	315.051*	.	.	-141.412	149.781*
		(45.735)			(139.954)	(52.368)
Iowa BS	.	.	99.153*	.	3.177	1.812
			(22.318)		(26.679)	(24.936)
Harvard SS	.	.	.	13.655*	254.643*	.
				(1.448)	(125.297)	
N	521	537	553	483	442	505
RMSE	5066.464	5248.623	5372.166	5058.287	5054.432	5040.323
R^2	0.146	0.081	0.035	0.156	0.17	0.161
adj R^2	0.144	0.08	0.033	0.154	0.162	0.156

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    439 1.1270e+10
2    437 1.1164e+10  2 105660856 2.0679 0.1277
```

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)	-814.953	13663.383*	10358.568*	-1473.914	-1298.826	-846.268
	(2301.442)	(1064.161)	(2374.515)	(2445.635)	(2881.853)	(2715.138)
SAT	13.273*	.	.	.	-242.809	11.127*
	(1.429)				(125.302)	(1.651)
ACT	.	309.264*	.	.	-141.412	149.781*
		(47.278)			(139.954)	(52.368)
Iowa BS	.	.	100.562*	.	3.177	1.812
			(23.534)		(26.679)	(24.936)
Harvard SS	.	.	.	13.513*	254.643*	.
				(1.501)	(125.297)	
N	505	505	505	442	442	505
RMSE	5074.891	5272.985	5395.631	5079.916	5054.432	5040.323
R^2	0.146	0.078	0.035	0.156	0.17	0.161
adj R^2	0.145	0.077	0.033	0.154	0.162	0.156

$*p \leq 0.05$

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    0.288335971
act    0.126751895
ibs    0.003245868
```

```
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   7.60518e-02
act   1.36949e-02
ibs   8.83654e-06
```

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%     40.49    42.17    48.57
50%     52.07    53.23    58.05
75%     64.78    65.92    68.76
100%   100.00   100.00   100.00
```

```
mean    52.60    53.49    58.17
sd      17.02    18.86    15.43
var    289.70   355.80   238.20
NA's     0.00     0.00     0.00
N      505.00   505.00   505.00

$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
-13697.4   -3633.3     -37.5    3506.5   16640.4

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.147e+04  9.732e+02  11.785  < 2e-16 ***
satpoms     1.140e+02  1.692e+01   6.740 4.37e-11 ***
actpoms     4.372e+01  1.529e+01   2.860  0.00441 **
ibspoms     9.808e-01  1.350e+01   0.073  0.94211
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5040 on 501 degrees of freedom
Multiple R^2: 0.1613,   Adjusted R^2: 0.1563
F-statistic: 32.11 on 3 and 501 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  -0.092301110
act  -0.048278379
ibs   0.005696261
harv  0.096762524
```

```
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   0.00713456560
act   0.00193979870
ibs   0.00002694212
harv  0.00784761514
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-7

	Test Scores Only	All Predictors
	Estimate	Estimate
	(S.E.)	(S.E.)
(Intercept)	3310.526	293.182
	(2793.404)	(2736.388)
SAT	11.138*	11.222*
	(1.676)	(1.599)
ACT	124.909*	137.716*
	(53.253)	(50.83)
Iowa BS	-4.827	1.164
	(25.69)	(24.789)
Major: Soc.	.	774.508
		(543.337)
Major: Nat.	.	3365.763*
		(537.362)
Prof. Parents: Yes	.	1590.476*
		(482.335)
Parent Network: Yes	.	637.259
		(487.288)
Gender: Male	.	-116.226
		(444.238)
N	516	516
RMSE	5244.319	4995.948
R^2	0.138	0.225
adj R^2	0.133	0.213

$*p \leq 0.05$

```
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 =  953.217029996019 Denominator =  689.823675161957"
```

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

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

```
tval
```

```
difference
  1.381827
```

```
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
 0.1676334
```

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
      953.2170300       689.8236752         1.3818271       507.0000000         0.1676334
```

```
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    512 14081474433
2    507 12654465846  5 1427008587 11.435 1.791e-10 ***
---
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

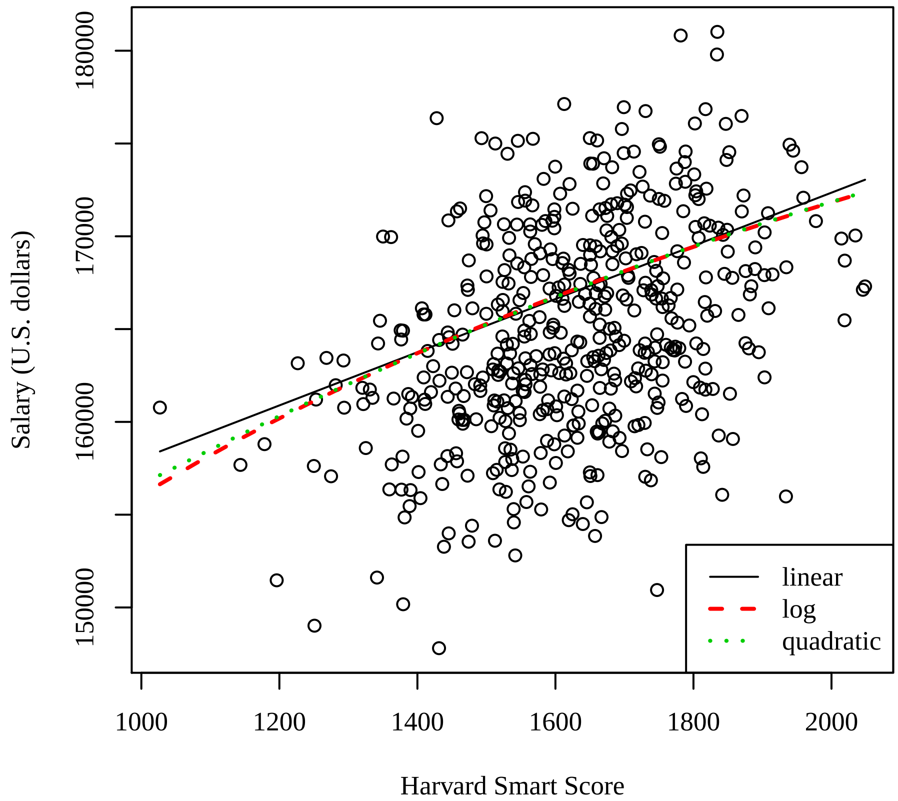
Table 4: Regression with sal3: Student-7

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	139284.775*	-5819.123	128872.713*
	(2329.927)	(16362.607)	(15375.221)
Harvard SS	14.329*	.	27.386
	(1.391)		(19.109)
Gender: Male	314.361	295.566	301.079
	(443.471)	(443.461)	(444.137)
Major: Soc.	1388.411*	1376.453*	1378.811*
	(544.271)	(544.212)	(544.749)
Major: Nat.	4407.843*	4399.005*	4404.574*
	(543.419)	(543.351)	(543.737)
Prof. Parents: Yes	1186.405*	1164.077*	1166.953*
	(480.201)	(480.092)	(481.302)
Parent Network: Yes	-408.554	-396.626	-395.213
	(490.377)	(490.39)	(491.032)
ln(Harvard SS)	.	22793.825*	.
		(2211.635)	
Harvard SS^2	.	.	-0.004
			(0.006)
N	492	492	492
RMSE	4911.603	4911.216	4914.292
R^2	0.27	0.27	0.27
adj R^2	0.26	0.261	0.26

$*p \leq 0.05$

```
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 (30%) 22192.89     H
N (30%) 25383.95     N
S (30%) 22800.66     S

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

```
predictOMatic(cm2)
```

```
$major2
              fit  major2
H (30%) 22192.89      H
N (30%) 25383.95      N
S (30%) 22800.66      S

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

Table 5: Categorical Regressions: Student-7

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	22192.886*	22800.656*	293.182	1067.689
	(391.286)	(399.657)	(2736.388)	(2714.604)
Major: Soc.	607.77	.	774.508	.
	(559.312)		(543.337)	
Major: Nat.	3191.06*	.	3365.763*	.
	(558.544)		(537.362)	
Major 2: Hum.	.	-607.77	.	-774.508
		(559.312)		(543.337)
Major 2: Nat.	.	2583.29*	.	2591.255*
		(564.44)		(541.465)
SAT	.	.	11.222*	11.222*
			(1.599)	(1.599)
ACT	.	.	137.716*	137.716*
			(50.83)	(50.83)
Iowa BS	.	.	1.164	1.164
			(24.789)	(24.789)
Prof. Parents: Yes	.	.	1590.476*	1590.476*
			(482.335)	(482.335)
Parent Network: Yes	.	.	637.259	637.259
			(487.288)	(487.288)
Gender: Male	.	.	-116.226	-116.226
			(444.238)	(444.238)
N	564	564	516	516
RMSE	5435.92	5435.92	4995.948	4995.948
R^2	0.061	0.061	0.225	0.225
adj R^2	0.058	0.058	0.213	0.213

$*p \leq 0.05$