Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 37
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%	7.89	1049.00	71.22	4327.00	6805.00	150200.00	1035.00
25%	18.41	1503.00	92.33	17380.00	20270.00	161800.00	1490.00
50%	21.69	1618.00	99.74	20850.00	23760.00	165500.00	1596.00
75%	25.18	1731.00	106.60	24080.00	27680.00	169800.00	1709.00
100%	36.42	2136.00	125.40	34170.00	38950.00	185300.00	2103.00
mean	21.78	1621.00	99.30	20600.00	23780.00	165800.00	1600.00
sd	4.90	170.60	10.07	5351.00	5645.00	6113.00	169.30
var	24.04	29110.00	101.40	28640000.00	31870000.00	37370000.00	28650.00
NA's	15.00	48.00	0.00	15.00	0.00	0.00	25.00
N	535.00	535.00	535.00	535.00	535.00	535.00	535.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender              major                pnet                pprof
F            :269   N            :197.0000   NO           :371.0000   NO           :388
    .0000
M            :266   S            :186.0000   YES          :164.0000   YES          :147
    .0000
NA's         :  0   H            :152.0000   NA's         :  0.0000   NA's         :  0
    .0000
entropy      :  1   NA's         :  0.0000   entropy      :  0.8891   entropy      :  0
    .8482
normedEntropy:  1   entropy      :  1.5765   normedEntropy:  0.8891   normedEntropy:  0
    .8482
N            :535   normedEntropy:  0.9946   N            :535.0000   N            :535
    .0000
                    N            :535.0000
```

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: 0x259eb48>
act ~ sat + ibs + harv
<environment: 0x259eb48>
ibs ~ sat + act + harv
<environment: 0x259eb48>
harv ~ sat + act + ibs
<environment: 0x259eb48>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998745 0.8663860 0.2657006 0.9998773
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
7967.962160    7.484245    1.361842 8147.984251
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.39 0.46 1.00
act  0.39 1.00 0.40 0.41
ibs  0.46 0.40 1.00 0.46
harv 1.00 0.41 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-37

	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)	3637.145	13928.385*	9410.155*	2675.305	1333.103	1100.922
	(2188.928)	(1047.054)	(2259.229)	(2237.948)	(2799.637)	(2638.457)
SAT	10.573*	.	.	.	-122.08	8.63*
	(1.36)				(128.307)	(1.556)
ACT	.	305.265*	.	.	28.077	179.787*
		(47.031)			(135.658)	(52.332)
Iowa BS	.	.	112.754*	.	5.314	16.943
			(22.644)		(28.443)	(26.505)
Harvard SS	.	.	.	11.036*	131.557	.
				(1.372)	(128.147)	
N	496	506	520	473	440	482
RMSE	5117.508	5162.29	5232.746	5082.805	5101.359	5044.792
R^2	0.109	0.077	0.046	0.121	0.145	0.144
adj R^2	0.107	0.075	0.044	0.119	0.138	0.139

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    437 1.1349e+10
2    435 1.1320e+10  2  28601590 0.5495 0.5776
```

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)	2870.392	13928.173*	9021.274*	1828.931	1333.103	1100.922
	(2218.01)	(1076.549)	(2398.645)	(2331.083)	(2799.637)	(2638.457)
SAT	11.014*	.	.	.	-122.08	8.63*
	(1.377)				(128.307)	(1.556)
ACT	.	303.208*	.	.	28.077	179.787*
		(48.299)			(135.658)	(52.332)
Iowa BS	.	.	115.849*	.	5.314	16.943
			(24.045)		(28.443)	(26.505)
Harvard SS	.	.	.	11.502*	131.557	.
				(1.428)	(128.147)	
N	482	482	482	440	440	482
RMSE	5110.916	5230.583	5314.095	5132.803	5101.359	5044.792
R^2	0.118	0.076	0.046	0.129	0.145	0.144
adj R^2	0.116	0.074	0.044	0.127	0.138	0.139

$*p \leq 0.05$

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   0.24591208
act   0.15523203
ibs   0.02922582
```

```
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 0.0551003510
act 0.0211378140
ibs 0.0007318273
```

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%    36.47   38.63   42.77
50%    48.14   52.15   52.80
75%    60.52   65.29   63.16
100%  100.00  100.00  100.00
```

```
mean    48.53    51.68    53.12
sd      17.31    18.58    15.86
var    299.60   345.30   251.40
NA's     0.00     0.00     0.00
N      482.00   482.00   482.00

$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
-12892.4  -3310.5    212.1   3349.2  15014.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12660.001    922.924  13.717  < 2e-16 ***
satpoms        92.135     16.611   5.547 4.82e-08 ***
actpoms        51.293     14.930   3.436 0.000643 ***
ibspoms         9.188     14.374   0.639 0.522971
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5045 on 478 degrees of freedom
Multiple R^2: 0.1439,   Adjusted R^2: 0.1386
F-statistic: 26.79 on 3 and 478 DF,  p-value: 4.956e-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  -0.045572247
act   0.009922891
ibs   0.008956801
harv  0.049162638
```

```
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   0.00177840076
act   0.00008414825
ibs   0.00006855932
harv  0.00207036622
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-37

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	3472.31	1476.926
	(2788.205)	(2648.445)
SAT	9.249*	8.844*
	(1.621)	(1.527)
ACT	150.497*	171.552*
	(54.49)	(51.36)
Iowa BS	21.442	18.282
	(27.776)	(26.204)
Major: Soc.	.	1013.823
		(578.516)
Major: Nat.	.	4098.088*
		(567.595)
Prof. Parents: Yes	.	534.288
		(509.445)
Parent Network: Yes	.	1241.737*
		(493.675)
Gender: Male	.	201.374
		(453.669)
N	495	495
RMSE	5352.933	5030.042
R^2	0.132	0.241
adj R^2	0.126	0.229

$*p \leq 0.05$

```
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 =  -707.448461537461 Denominator =  723.647202503542"
```

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

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

```
tval
```

```
difference
-0.9776151
```

```
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
 0.3287511
```

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
    -707.4484615        723.6472025        -0.9776151       486.0000000         0.3287511
```

```
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    491 14069058245
2    486 12296443544  5 1772614700 14.012 8.223e-13 ***
---
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

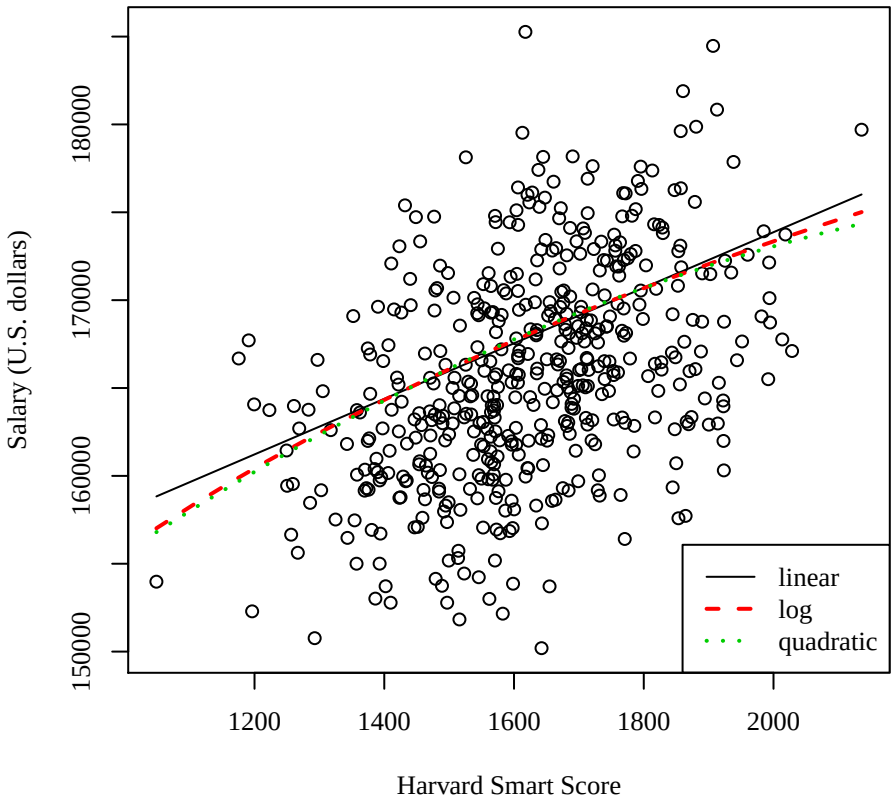
Table 4: Regression with sal3: Student-37

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	137111.829*	-24100.347	118982.224*
	(2216.513)	(15762.752)	(14824.422)
Harvard SS	15.808*	.	38.458*
	(1.342)		(18.362)
Gender: Male	-300.167	-304.141	-299.455
	(457.788)	(457.241)	(457.536)
Major: Soc.	2334.138*	2353.783*	2356.06*
	(582.33)	(581.542)	(582.278)
Major: Nat.	5440.097*	5478.577*	5489.471*
	(572.258)	(571.528)	(573.334)
Prof. Parents: Yes	1419.539*	1418.794*	1415.346*
	(514.57)	(513.948)	(514.298)
Parent Network: Yes	74.452	76.757	80.67
	(503.224)	(502.615)	(502.971)
ln(Harvard SS)	.	25296.332*	.
		(2135.675)	
Harvard SS^2	.	.	-0.007
			(0.006)
N	487	487	487
RMSE	5038.681	5032.584	5035.903
R^2	0.332	0.334	0.335
adj R^2	0.324	0.326	0.325

$*p \leq 0.05$

```
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
N (40%) 26026.79     N
S (30%) 23093.87     S
H (30%) 21711.59     H

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
N (40%) 26026.79      N
S (30%) 23093.87      S
H (30%) 21711.59      H

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

Table 5: Categorical Regressions: Student-37

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	21711.585* (434.792)	23093.869* (393.049)	1476.926 (2648.445)	2490.749 (2683.319)
Major: Soc.	1382.283* (586.116)	.	1013.823 (578.516)	.
Major: Nat.	4315.2* (578.71)	.	4098.088* (567.595)	.
Major 2: Hum.	.	-1382.283* (586.116)	.	-1013.823 (578.516)
Major 2: Nat.	.	2932.917* (548.041)	.	3084.264* (535.814)
SAT	.	.	8.844* (1.527)	8.844* (1.527)
ACT	.	.	171.552* (51.36)	171.552* (51.36)
Iowa BS	.	.	18.282 (26.204)	18.282 (26.204)
Prof. Parents: Yes	.	.	534.288 (509.445)	534.288 (509.445)
Parent Network: Yes	.	.	1241.737* (493.675)	1241.737* (493.675)
Gender: Male	.	.	201.374 (453.669)	201.374 (453.669)
N	535	535	495	495
RMSE	5360.476	5360.476	5030.042	5030.042
R^2	0.102	0.102	0.241	0.241
adj R^2	0.098	0.098	0.229	0.229

$*p \leq 0.05$