Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 48
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%	4.58	1182.00	60.59	6938.00	7100.00	145400.00	1164.00
25%	18.84	1502.00	92.08	17060.00	19640.00	161600.00	1486.00
50%	21.82	1615.00	99.42	20730.00	23250.00	165400.00	1587.00
75%	25.30	1722.00	106.60	23780.00	27240.00	169700.00	1701.00
100%	36.81	2174.00	129.70	38360.00	41130.00	181200.00	2146.00
mean	21.91	1617.00	99.37	20480.00	23360.00	165400.00	1593.00
sd	4.92	166.10	10.31	5151.00	5645.00	5834.00	160.40
var	24.19	27600.00	106.20	26530000.00	31870000.00	34030000.00	25720.00
NA's	10.00	53.00	0.00	10.00	0.00	0.00	37.00
N	529.00	529.00	529.00	529.00	529.00	529.00	529.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender              major                pnet                pprof
M             :266   H             :179.0000   NO            :364.0000   NO            :364
    .0000
F             :263   S             :175.0000   YES           :165.0000   YES           :165
    .0000
NA's          :  0   N             :175.0000   NA's          :  0.0000   NA's          :  0
    .0000
entropy       :  1   NA's          :  0.0000   entropy       :  0.8954   entropy       :  0
    .8954
normedEntropy:  1   entropy       :  1.5849   normedEntropy:  0.8954   normedEntropy:  0
    .8954
N             :529   normedEntropy:  0.9999   N             :529.0000   N             :529
    .0000
                     N             :529.0000
```

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: 0x2647b10>
act ~ sat + ibs + harv
<environment: 0x2647b10>
ibs ~ sat + act + harv
<environment: 0x2647b10>
harv ~ sat + act + ibs
<environment: 0x2647b10>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998610 0.8674211 0.2584727 0.9998647
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
7191.848549    7.542677    1.348568 7392.741106
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.45 0.43 1.00
act  0.45 1.00 0.43 0.47
ibs  0.43 0.43 1.00 0.44
harv 1.00 0.47 0.44 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-48

	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)	213.557	12546.329*	9922.131*	1143.895	1128.705	411.137
	(2153.162)	(981.704)	(2144.021)	(2163.264)	(2673.5)	(2559.434)
SAT	12.642*	.	.	.	80.613	10.207*
	(1.344)				(119.021)	(1.571)
ACT	.	361.153*	.	.	267.632*	196.624*
		(43.726)			(126.601)	(51.17)
Iowa BS	.	.	106.217*	.	-2.563	-6.154
			(21.452)		(25.937)	(24.546)
Harvard SS	.	.	.	11.91*	-71.123	.
				(1.331)	(119.048)	
N	483	509	519	466	425	474
RMSE	4736.027	4844.206	5037.598	4751.66	4710.263	4688.058
R^2	0.155	0.119	0.045	0.147	0.177	0.185
adj R^2	0.154	0.117	0.043	0.145	0.169	0.18

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    422 9326663910
2    420 9318364283  2    8299627 0.187 0.8295
```

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)	-41.271	12909.97*	9968.829*	719.628	1128.705	411.137
	(2181.51)	(1016.093)	(2281.375)	(2276.246)	(2673.5)	(2559.434)
SAT	12.808*	.	.	.	80.613	10.207*
	(1.362)				(119.021)	(1.571)
ACT	.	340.766*	.	.	267.632*	196.624*
		(45.279)			(126.601)	(51.17)
Iowa BS	.	.	104.658*	.	-2.563	-6.154
			(22.842)		(25.937)	(24.546)
Harvard SS	.	.	.	12.122*	-71.123	.
				(1.401)	(119.048)	
N	474	474	474	425	425	474
RMSE	4755.783	4896.596	5070.526	4767.756	4710.263	4688.058
R^2	0.158	0.107	0.043	0.15	0.177	0.185
adj R^2	0.156	0.105	0.041	0.148	0.169	0.18

$*p \leq 0.05$

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   0.28708777
act   0.17452406
ibs  -0.01156341
```

```
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 0.0732017528
act 0.0256023956
ibs 0.0001089848
```

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%    44.13   45.92   32.73
50%    53.44   56.09   43.09
75%    64.30   66.51   54.74
100%  100.00  100.00  100.00
```

```
mean    53.69    56.08    43.72
sd      15.43    14.76    16.35
var    238.00   218.00   267.40
NA's     0.00     0.00     0.00
N      474.00   474.00   474.00

$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
-11925.4  -3306.6    -53.6   3348.7  13721.3

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12822.918    957.640  13.390  < 2e-16 ***
satpoms       100.195     15.421   6.497 2.09e-10 ***
actpoms        63.372     16.492   3.843 0.000138 ***
ibspoms        -4.254     16.969  -0.251 0.802151
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4688 on 470 degrees of freedom
Multiple R^2: 0.185, Adjusted R^2: 0.1798
F-statistic: 35.57 on 3 and 470 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   0.033030951
act   0.102606993
ibs  -0.004821496
harv -0.029139118
```

```
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  0.00089942490
act  0.00876191752
ibs  0.00001914354
harv 0.00069979427
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-48

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	4757.51	-897.904
	(2871.675)	(2627.444)
SAT	9.479*	10.821*
	(1.766)	(1.582)
ACT	211.253*	182.539*
	(57.444)	(51.191)
Iowa BS	-12.501	-5.057
	(27.535)	(24.565)
Major: Soc.	.	2290.232*
		(522.779)
Major: Nat.	.	5472.239*
		(528.872)
Prof. Parents: Yes	.	1291.547*
		(464.007)
Parent Network: Yes	.	1840.045*
		(462.188)
Gender: Male	.	-254.577
		(430.188)
N	483	483
RMSE	5304.858	4709.782
R^2	0.142	0.331
adj R^2	0.136	0.319

$*p \leq 0.05$

```
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 =  -548.497084873856 Denominator =  655.448220081472"
```

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

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

```
tval
```

```
difference
-0.8368275
```

```
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
 0.4031112
```

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
    -548.4970849       655.4482201        -0.8368275       474.0000000         0.4031112
```

```
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    479 13479786355
2    474 10514289983  5 2965496372 26.738 < 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

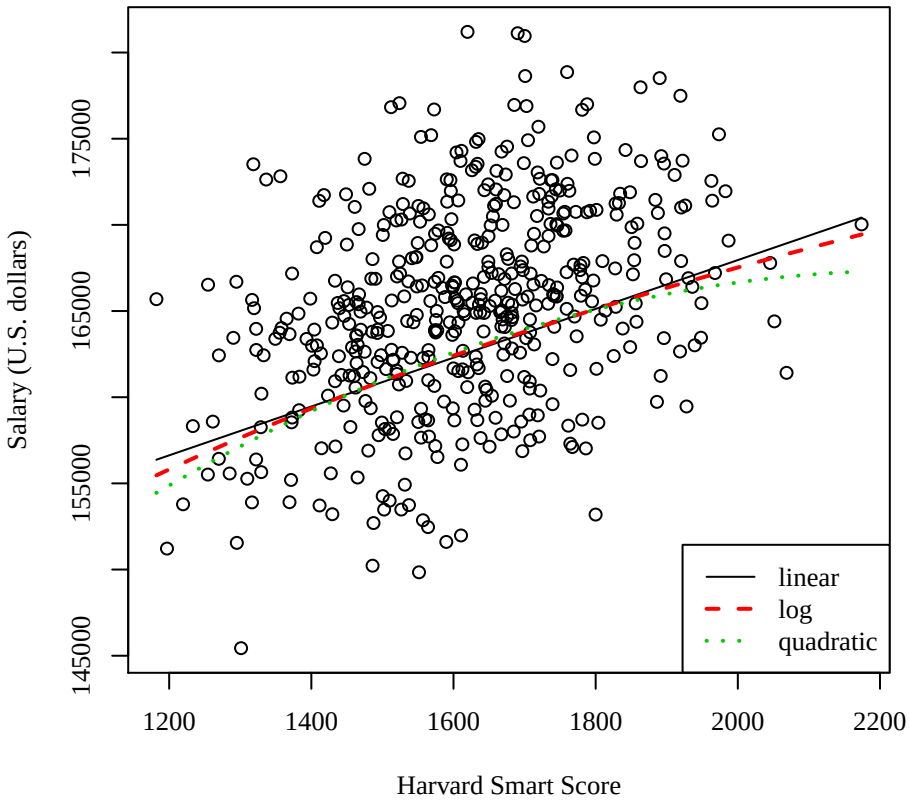
Table 4: Regression with sal3: Student-48

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	140006.663*	-6429.764	110180.656*
	(2371.094)	(16710.203)	(16298.597)
Harvard SS	14.161*	.	51.177*
	(1.41)		(20.063)
Gender: Male	-372.733	-349.752	-308.004
	(467.012)	(466.031)	(467.124)
Major: Soc.	1949.426*	1965.514*	2000.552*
	(574.005)	(572.89)	(573.196)
Major: Nat.	5030.419*	5011.934*	4985.626*
	(566.787)	(565.526)	(565.848)
Prof. Parents: Yes	1099.563*	1118.383*	1143.287*
	(507.649)	(506.729)	(506.895)
Parent Network: Yes	-18.209	9.514	50.297
	(505.243)	(504.311)	(505.303)
ln(Harvard SS)	.	22932.442*	.
		(2257.444)	
Harvard SS^2	.	.	-0.011
			(0.006)
N	476	476	476
RMSE	5076.247	5065.853	5063.197
R^2	0.273	0.276	0.278
adj R^2	0.263	0.266	0.267

$*p \leq 0.05$

```
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%) 20875.47     H
S (30%) 23289.85     S
N (30%) 25983.92     N

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
H (30%) 20875.47      H
S (30%) 23289.85      S
N (30%) 25983.92      N

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

Table 5: Categorical Regressions: Student-48

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	20875.47* (392.624)	23289.849* (397.085)	-897.904 (2627.444)	1392.328 (2621.268)
Major: Soc.	2414.379* (558.418)	.	2290.232* (522.779)	.
Major: Nat.	5108.452* (558.418)	.	5472.239* (528.872)	.
Major 2: Hum.	.	-2414.379* (558.418)	.	-2290.232* (522.779)
Major 2: Nat.	.	2694.073* (561.564)	.	3182.007* (528.805)
SAT	.	.	10.821* (1.582)	10.821* (1.582)
ACT	.	.	182.539* (51.191)	182.539* (51.191)
Iowa BS	.	.	-5.057 (24.565)	-5.057 (24.565)
Prof. Parents: Yes	.	.	1291.547* (464.007)	1291.547* (464.007)
Parent Network: Yes	.	.	1840.045* (462.188)	1840.045* (462.188)
Gender: Male	.	.	-254.577 (430.188)	-254.577 (430.188)
N	529	529	483	483
RMSE	5252.947	5252.947	4709.782	4709.782
R^2	0.137	0.137	0.331	0.331
adj R^2	0.134	0.134	0.319	0.319

$*p \leq 0.05$