Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 47
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%	5.83	1197.00	68.63	5235.00	5936.00	147100.00	1180.00
25%	18.00	1521.00	92.26	16390.00	19380.00	162000.00	1495.00
50%	21.41	1626.00	99.88	20330.00	23080.00	165700.00	1600.00
75%	25.01	1731.00	106.20	23790.00	27350.00	169300.00	1706.00
100%	37.65	2224.00	126.40	39430.00	44350.00	185300.00	2188.00
mean	21.58	1622.00	99.49	20220.00	23180.00	165500.00	1597.00
sd	5.16	154.30	9.67	5441.00	5947.00	5645.00	154.50
var	26.60	23800.00	93.51	29600000.00	35370000.00	31870000.00	23880.00
NA's	18.00	46.00	0.00	14.00	0.00	0.00	24.00
N	530.00	530.00	530.00	530.00	530.00	530.00	530.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
             gender                    major                        pnet                         pprof
F            :270.0000    S            :182.0000    NO           :368.0000    NO
    :363.000
M            :260.0000    H            :177.0000    YES          :162.0000    YES
    :167.000
NA's         :  0.0000    N            :171.0000    NA's         :  0.0000    NA's           :
    0.000
entropy      :  0.9997    NA's         :  0.0000    entropy      :  0.8881    entropy        :
    0.899
normedEntropy:  0.9997    entropy      :  1.5845    normedEntropy:  0.8881    normedEntropy:
    0.899
N            :530.0000    normedEntropy:  0.9997    N            :530.0000    N
    :530.000
                          N            :530.0000
```

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: 0x1483d10>
act ~ sat + ibs + harv
<environment: 0x1483d10>
ibs ~ sat + act + harv
<environment: 0x1483d10>
harv ~ sat + act + ibs
<environment: 0x1483d10>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998305 0.8797761 0.2845976 0.9998357
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
5900.931243    8.317811    1.397815 6085.011357
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.44 0.46 1.00
act  0.44 1.00 0.45 0.47
ibs  0.46 0.45 1.00 0.47
harv 1.00 0.47 0.47 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-47

	SAT	ACT	IBS	Harvard SS	All	Best
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)
(Intercept)	-2145.121	12196.556*	9359.671*	-905.262	975.246	286.385
	(2365.229)	(992.595)	(2447.341)	(2425.627)	(2911.342)	(2799.865)
SAT	14.011*	.	.	.	-235.799*	11.536*
	(1.474)				(117.659)	(1.732)
ACT	.	370.691*	.	.	18.468	256.605*
		(44.63)			(131.066)	(51.483)
Iowa BS	.	.	109.006*	.	-40.538	-39.922
			(24.441)		(28.525)	(27.895)
Harvard SS	.	.	.	13.02*	246.818*	.
				(1.488)	(117.764)	
N	492	499	516	470	432	477
RMSE	5029.055	5139.123	5343.891	4949.183	4839.845	4936.58
R^2	0.156	0.122	0.037	0.141	0.203	0.199
adj R^2	0.154	0.12	0.035	0.139	0.195	0.194

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F  Pr(>F)
1    429 1.0156e+10
2    427 1.0002e+10  2 153602617 3.2787 0.03863 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

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	ACT	IBS	Harvard SS	All	Best
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)
(Intercept)	-2327.242	12180.269*	9540.694*	-1823.431	975.246	286.385
	(2408.6)	(1014.852)	(2568.078)	(2524.119)	(2911.342)	(2799.865)
SAT	14.149*	.	.	.	-235.799*	11.536*
	(1.503)				(117.659)	(1.732)
ACT	.	374.406*	.	.	18.468	256.605*
		(45.871)			(131.066)	(51.483)
Iowa BS	.	.	107.604*	.	-40.538	-39.922
			(25.716)		(28.525)	(27.895)
Harvard SS	.	.	.	13.633*	246.818*	.
				(1.551)	(117.764)	
N	477	477	477	432	432	477
RMSE	5053.982	5155.4	5406.337	4973.156	4839.845	4936.58
R^2	0.157	0.123	0.036	0.152	0.203	0.199
adj R^2	0.155	0.121	0.034	0.15	0.195	0.194

$*p \leq 0.05$

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   0.29289102
act   0.22338629
ibs  -0.06566203
```

```
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  0.075137523
act  0.042056914
ibs  0.003467354
```

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%    41.95   40.75   31.06
50%    53.82   53.73   41.19
75%    66.56   65.13   51.90
100%  100.00  100.00  100.00
```

```
mean    54.30    53.23    41.14
sd      17.83    16.67    15.28
var    317.90   277.80   233.30
NA's     0.00     0.00     0.00
N      477.00   477.00   477.00

$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
-14833.6   -3645.0     -95.4    3484.9   13370.6

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12652.29     882.15  14.343  < 2e-16 ***
satpoms       116.37      17.47   6.662 7.52e-11 ***
actpoms        74.13      14.87   4.984 8.74e-07 ***
ibspoms       -23.08      16.13  -1.431    0.153
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4937 on 473 degrees of freedom
Multiple R^2: 0.1993,  Adjusted R^2: 0.1942
F-statistic: 39.23 on 3 and 473 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  -0.096531641
act   0.006818591
ibs  -0.068611902
harv  0.100908021
```

```
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  0.00749934703
act  0.00003707043
ibs  0.00377108580
harv 0.00820189882
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-47

	Test Scores Only	All Predictors
	Estimate	Estimate
	(S.E.)	(S.E.)
(Intercept)	5647.406	899.824
	(3067.008)	(2833.076)
SAT	11.667*	11.249*
	(1.891)	(1.707)
ACT	261.848*	242.018*
	(56.79)	(51.302)
Iowa BS	-67.046*	-38.405
	(30.653)	(27.751)
Major: Soc.	.	2131.958*
		(546.928)
Major: Nat.	.	5230.281*
		(554.619)
Prof. Parents: Yes	.	1554.216*
		(485.221)
Parent Network: Yes	.	1537.898*
		(485.047)
Gender: Male	.	-835.599
		(448.804)
N	490	490
RMSE	5501.999	4951.956
R^2	0.16	0.326
adj R^2	0.155	0.315
$*p \leq 0.05$		

```
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 =  16.3185694513402 Denominator =  675.386206584203"
```

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

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

```
tval
```

```
difference
0.02416183
```

```
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
 0.9807335
```

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
      16.31856945      675.38620658        0.02416183      481.00000000        0.98073354
```

```
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    486 14712191030
2    481 11795016357  5 2917174673 23.792 < 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

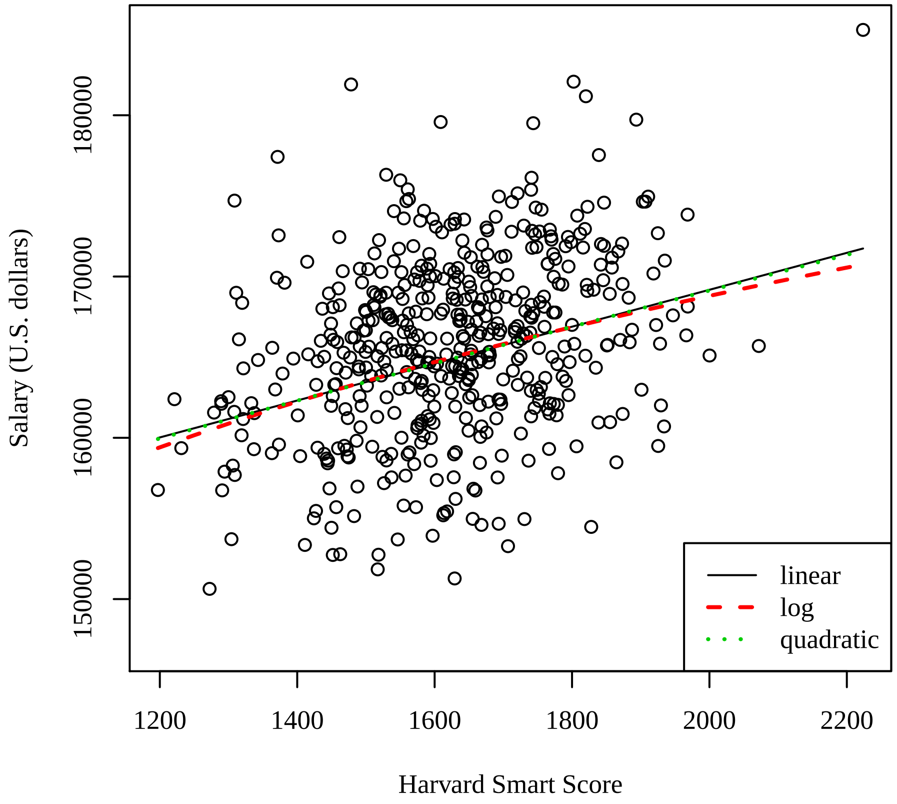
Table 4: Regression with sal3: Student-47

	Linear	Log	Quadratic
	Estimate	Estimate	Estimate
	(S.E.)	(S.E.)	(S.E.)
(Intercept)	144218.294*	26961.673	143157.712*
	(2310.063)	(16505.049)	(16311.147)
Harvard SS	11.445*	.	12.762
	(1.389)		(20.095)
Gender: Male	-145.771	-162.221	-147.264
	(428.101)	(428.109)	(429.15)
Major: Soc.	2070.01*	2075.547*	2070.797*
	(522.001)	(522.036)	(522.684)
Major: Nat.	5950.156*	5955.556*	5950.939*
	(527.009)	(527.026)	(527.695)
Prof. Parents: Yes	722.779	732.057	723.832
	(459.982)	(460.048)	(460.742)
Parent Network: Yes	449.588	431.83	447.403
	(464.883)	(464.899)	(466.557)
ln(Harvard SS)	.	18387.686*	.
		(2232.319)	
Harvard SS^2	.	.	0
			(0.006)
N	484	484	484
RMSE	4703.88	4704.086	4708.797
R^2	0.303	0.303	0.303
adj R^2	0.295	0.294	0.293

$*p \leq 0.05$

```
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 (30%) 22948.85     S
H (30%) 20563.47     H
N (30%) 26146.99     N

attr(,”flnames”)
[1] ”major”
```

```
predictOMatic(cm2)
```

```
$major2
             fit major2
S (30%) 22948.85      S
H (30%) 20563.47      H
N (30%) 26146.99      N

attr(,”flnames”)
[1] ”major2”
```

Table 5: Categorical Regressions: Student-47

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	20563.473*	22948.845*	899.824	3031.782
	(413.936)	(408.21)	(2833.076)	(2826.245)
Major: Soc.	2385.372*	.	2131.958*	.
	(581.359)		(546.928)	
Major: Nat.	5583.512*	.	5230.281*	.
	(590.506)		(554.619)	
Major 2: Hum.	.	-2385.372*	.	-2131.958*
		(581.359)		(546.928)
Major 2: Nat.	.	3198.14*	.	3098.322*
		(586.507)		(547.317)
SAT	.	.	11.249*	11.249*
			(1.707)	(1.707)
ACT	.	.	242.018*	242.018*
			(51.302)	(51.302)
Iowa BS	.	.	-38.405	-38.405
			(27.751)	(27.751)
Prof. Parents: Yes	.	.	1554.216*	1554.216*
			(485.221)	(485.221)
Parent Network: Yes	.	.	1537.898*	1537.898*
			(485.047)	(485.047)
Gender: Male	.	.	-835.599	-835.599
			(448.804)	(448.804)
N	530	530	490	490
RMSE	5507.058	5507.058	4951.956	4951.956
R^2	0.146	0.146	0.326	0.326
adj R^2	0.143	0.143	0.315	0.315

$*p \leq 0.05$