Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 6
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%	6.51	1168.00	72.85	5665.00	8637.00	150300.00	1159.00
25%	19.15	1526.00	93.97	16920.00	19680.00	161900.00	1504.00
50%	22.23	1634.00	99.97	20680.00	23550.00	165600.00	1611.00
75%	25.68	1753.00	107.10	24500.00	27400.00	169200.00	1726.00
100%	34.37	2094.00	136.40	35640.00	38640.00	182600.00	2083.00
mean	22.17	1634.00	100.20	20560.00	23490.00	165600.00	1613.00
sd	4.96	158.70	9.85	5477.00	5694.00	5590.00	157.10
var	24.63	25190.00	97.08	29990000.00	32420000.00	31250000.00	24680.00
NA's	20.00	58.00	0.00	10.00	0.00	0.00	30.00
N	537.00	537.00	537.00	537.00	537.00	537.00	537.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet                    pprof
F             :281.0000   N             :188.0000   NO            :393.0000   NO            :382.000
M             :256.0000   S             :180.0000   YES           :144.0000   YES           :155.000
NA's          :  0.0000   H             :169.0000   NA's          :  0.0000   NA's          :  0.000
entropy       :  0.9984   NA's          :  0.0000   entropy       :  0.8388   entropy       :  0.867
normedEntropy:  0.9984   entropy       :  1.5836   normedEntropy:  0.8388   normedEntropy:  0.867
N             :537.0000   normedEntropy:  0.9991   N             :537.0000   N             :537.000
                          N             :537.0000
```

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: 0x2d9dfe8>
act ~ sat + ibs + harv
<environment: 0x2d9dfe8>
ibs ~ sat + act + harv
<environment: 0x2d9dfe8>
harv ~ sat + act + ibs
<environment: 0x2d9dfe8>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998097 0.8528129 0.2526962 0.9998147
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
5253.541414    6.794073    1.338144 5396.679347
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.41 0.44 1.00
act  0.41 1.00 0.41 0.44
ibs  0.44 0.41 1.00 0.44
harv 1.00 0.44 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-6

	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)	-123.742	13777.483*	10060.439*	-421.211	-212.441	-232.553
	(2369.632)	(1062.055)	(2391.573)	(2383.015)	(2932.481)	(2856.592)
SAT	12.82*	.	.	.	-298.093*	9.759*
	(1.462)				(111.702)	(1.695)
ACT	.	303.613*	.	.	-150.767	195.133*
		(46.797)			(124.691)	(52.676)
Iowa BS	.	.	104.762*	.	1.487	6.877
			(23.749)		(28.073)	(27.047)
Harvard SS	.	.	.	12.877*	308.725*	.
				(1.451)	(111.7)	
N	497	507	527	470	423	477
RMSE	5112.01	5216.367	5382.936	5004.892	4892.892	5025.432
R^2	0.134	0.077	0.036	0.144	0.169	0.15
adj R^2	0.133	0.075	0.034	0.142	0.161	0.145

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F  Pr(>F)
1    420 1.0190e+10
2    418 1.0007e+10  2 182883047 3.8195 0.02271 *
---
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 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)	691.175	13496.322*	9452.738*	-187.454	-212.441	-232.553
	(2449.958)	(1098.581)	(2509.396)	(2528.258)	(2932.481)	(2856.592)
SAT	12.3*	.	.	.	-298.093*	9.759*
	(1.514)				(111.702)	(1.695)
ACT	.	315.763*	.	.	-150.767	195.133*
		(48.34)			(124.691)	(52.676)
Iowa BS	.	.	110.415*	.	1.487	6.877
			(24.959)		(28.073)	(27.047)
Harvard SS	.	.	.	12.721*	308.725*	.
				(1.542)	(111.7)	
N	477	477	477	423	423	477
RMSE	5097.367	5210.871	5331.17	4961.228	4892.892	5025.432
R^2	0.122	0.082	0.04	0.139	0.169	0.15
adj R^2	0.12	0.08	0.038	0.137	0.161	0.145

$*p \leq 0.05$

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   0.25585023
act   0.16791081
ibs   0.01169038
```

```
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 0.0595262829
act 0.0246554617
ibs 0.0001161588
```

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%    45.37   37.29   37.33
50%    56.53   47.45   48.86
75%    68.81   60.32   61.16
100%  100.00  100.00  100.00
```

```
mean    56.26    47.91    48.85
sd      17.73    17.24    16.70
var    314.50   297.10   278.80
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
-13270.9  -3148.6    308.6   3014.5  14787.8

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12850.640    915.139  14.042  < 2e-16 ***
satpoms        90.179     15.667   5.756 1.55e-08 ***
actpoms        54.364     14.675   3.704 0.000237 ***
ibspoms         3.906     15.362   0.254 0.799400
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5025 on 473 degrees of freedom
Multiple R^2: 0.1502,   Adjusted R^2: 0.1448
F-statistic: 27.86 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  -0.129429350
act  -0.059037184
ibs   0.002590367
harv  0.133966936
```

```
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  0.014162815208
act  0.002907466632
ibs  0.000005577924
harv 0.015191740029
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-6

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	2739.384	259.668
	(2970.493)	(2852.192)
SAT	9.638*	9.872*
	(1.762)	(1.673)
ACT	201.987*	180.917*
	(54.666)	(52.21)
Iowa BS	6.919	10.513
	(28.211)	(26.762)
Major: Soc.	.	965.447
		(564.951)
Major: Nat.	.	3944.039*
		(554.679)
Prof. Parents: Yes	.	354.648
		(506.358)
Parent Network: Yes	.	1263.175*
		(510.922)
Gender: Male	.	108.744
		(457.097)
N	487	487
RMSE	5277.834	4997.009
R^2	0.139	0.236
adj R^2	0.134	0.224

$*p \leq 0.05$

```
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 =  -908.526488996943 Denominator =  736.551854770968"
```

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

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

```
tval
```

```
difference
 -1.233486
```

```
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
 0.2180006
```

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
    -908.5264890        736.5518548       -1.2334861       478.0000000        0.2180006
```

```
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    483 13454222726
2    478 11935706007  5 1518516719 12.163 4.139e-11 ***
---
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

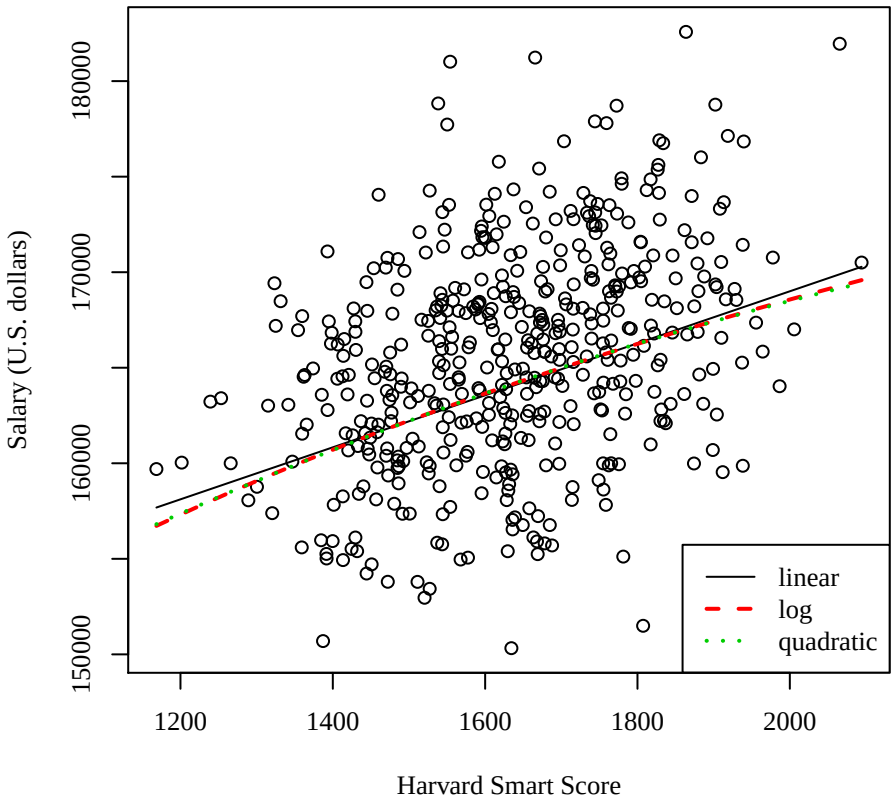
Table 4: Regression with sal3: Student-6

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	141078.467*	343.76	129137.814*
	(2349.306)	(16804.28)	(18416.679)
Harvard SS	13.61*	.	28.348
	(1.406)		(22.589)
Gender: Male	99.926	108.822	108.731
	(446.943)	(446.792)	(447.418)
Major: Soc.	702.387	711.899	710.281
	(543.49)	(543.254)	(543.954)
Major: Nat.	4374.913*	4403.193*	4404.687*
	(547.553)	(547.349)	(549.775)
Prof. Parents: Yes	1812.199*	1819.608*	1819.348*
	(496.517)	(496.379)	(496.939)
Parent Network: Yes	202.048	200.485	199.213
	(504.411)	(504.225)	(504.735)
ln(Harvard SS)	.	22038.807*	.
		(2271.477)	
Harvard SS^2	.	.	-0.005
			(0.007)
N	479	479	479
RMSE	4851.787	4849.945	4854.733
R^2	0.268	0.268	0.268
adj R^2	0.259	0.259	0.258

$*p \leq 0.05$

```
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%) 25541.19     N
S (30%) 23084.63     S
H (30%) 21641.60     H

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
N (40%) 25541.19      N
S (30%) 23084.63      S
H (30%) 21641.60      H

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

Table 5: Categorical Regressions: Student-6

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	21641.602* (420.805)	23084.628* (407.745)	259.668 (2852.192)	1225.115 (2884.574)
Major: Soc.	1443.026* (585.946)	.	965.447 (564.951)	.
Major: Nat.	3899.588* (579.877)	.	3944.039* (554.679)	.
Major 2: Hum.	.	-1443.026* (585.946)	.	-965.447 (564.951)
Major 2: Nat.	.	2456.562* (570.47)	.	2978.591* (551.572)
SAT	.	.	9.872* (1.673)	9.872* (1.673)
ACT	.	.	180.917* (52.21)	180.917* (52.21)
Iowa BS	.	.	10.513 (26.762)	10.513 (26.762)
Prof. Parents: Yes	.	.	354.648 (506.358)	354.648 (506.358)
Parent Network: Yes	.	.	1263.175* (510.922)	1263.175* (510.922)
Gender: Male	.	.	108.744 (457.097)	108.744 (457.097)
N	537	537	487	487
RMSE	5470.467	5470.467	4997.009	4997.009
R^2	0.08	0.08	0.236	0.236
adj R^2	0.077	0.077	0.224	0.224

$*p \leq 0.05$