Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 36
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%	4.52	1079.00	68.61	4776.00	8184.00	142900.00	1064.00
25%	18.63	1528.00	92.53	16860.00	19800.00	162000.00	1500.00
50%	21.76	1629.00	99.52	20750.00	23540.00	166100.00	1602.00
75%	24.82	1728.00	106.40	24140.00	27850.00	169500.00	1699.00
100%	40.68	2050.00	131.10	38300.00	41610.00	184700.00	2020.00
mean	21.81	1622.00	99.76	20620.00	23740.00	165800.00	1596.00
sd	4.98	158.70	10.26	5523.00	5957.00	5912.00	156.70
var	24.82	25190.00	105.30	30510000.00	35490000.00	34950000.00	24550.00
NA's	15.00	57.00	0.00	11.00	0.00	0.00	28.00
N	559.00	559.00	559.00	559.00	559.00	559.00	559.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet
F             :287.0000   N             :206.0000   NO            :393.0000
M             :272.0000   S             :192.0000   YES           :166.0000
NA's          :  0.0000   H             :161.0000   NA's          :  0.0000
entropy       :  0.9995   NA's          :  0.0000   entropy       :  0.8775
normedEntropy:  0.9995   entropy       :  1.5775   normedEntropy:  0.8775
N             :559.0000   normedEntropy:  0.9953   N             :559.0000
                          N             :559.0000
          pprof
NO            :381.0000
YES           :178.0000
NA's          :  0.0000
entropy       :  0.9027
normedEntropy:  0.9027
N             :559.0000
```

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: 0x1bc0580>
act ~ sat + ibs + harv
<environment: 0x1bc0580>
ibs ~ sat + act + harv
<environment: 0x1bc0580>
harv ~ sat + act + ibs
<environment: 0x1bc0580>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998410 0.8639193 0.2357991 0.9998446
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
6289.372022    7.348578    1.308556 6434.301556
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.35 0.38 1.00
act  0.35 1.00 0.41 0.38
ibs  0.38 0.41 1.00 0.39
harv 1.00 0.38 0.39 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-36

	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)	1574.551	11787.951*	7875.587*	991.662	-49.217	-477.49
	(2312.781)	(1001.787)	(2239.361)	(2399.948)	(2774.778)	(2657.551)
SAT	11.904*	.	.	.	-26.194	7.766*
	(1.441)				(117.442)	(1.559)
ACT	.	407.055*	.	.	239.882	289.923*
		(44.761)			(127.702)	(51.169)
Iowa BS	.	.	127.827*	.	17.91	23.605
			(22.339)		(25.13)	(24.218)
Harvard SS	.	.	.	12.171*	34.344	.
				(1.472)	(117.484)	
N	520	534	548	491	457	507
RMSE	5143.21	5133.692	5369.653	5169.794	4961.615	4975.749
R^2	0.116	0.135	0.057	0.123	0.173	0.177
adj R^2	0.115	0.133	0.055	0.121	0.166	0.172

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    454 1.1142e+10
2    452 1.1127e+10  2  14979301 0.3042 0.7378
```

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)	2181.722	11901.764*	8247.506*	1739.191	-49.217	-477.49
	(2341.952)	(1040.691)	(2279.48)	(2459.432)	(2774.778)	(2657.551)
SAT	11.538*	.	.	.	-26.194	7.766*
	(1.458)				(117.442)	(1.559)
ACT	.	399.32*	.	.	239.882	289.923*
		(46.515)			(127.702)	(51.169)
Iowa BS	.	.	123.925*	.	17.91	23.605
			(22.708)		(25.13)	(24.218)
Harvard SS	.	.	.	11.722*	34.344	.
				(1.508)	(117.484)	
N	507	507	507	457	457	507
RMSE	5164.198	5114.39	5320.234	5108.985	4961.615	4975.749
R^2	0.11	0.127	0.056	0.117	0.173	0.177
adj R^2	0.109	0.126	0.054	0.115	0.166	0.172

$*p \leq 0.05$

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   0.21679376
act   0.24494071
ibs   0.04341944
```

```
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   0.04057390
act   0.05250959
ibs   0.00155394
```

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%    33.17   37.86   46.05
50%    42.43   50.06   56.35
75%    51.67   61.38   66.43
100%  100.00  100.00  100.00
```

```
mean    42.80    50.01    55.87
sd      14.83    16.68    16.48
var    220.10   278.20   271.50
NA's     0.00     0.00     0.00
N      507.00   507.00   507.00

$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
-14231.9  -3232.4    125.1   3257.8  14890.8

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11649.15     927.21  12.564  < 2e-16 ***
satpoms        74.19      14.90   4.981 8.73e-07 ***
actpoms        95.53      16.86   5.666 2.46e-08 ***
ibspoms        14.74      15.12   0.975     0.33
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4976 on 503 degrees of freedom
Multiple R^2: 0.1773,   Adjusted R^2: 0.1724
F-statistic: 36.13 on 3 and 503 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 -1.00000000
sat  -0.01049039
act   0.08801187
ibs   0.03350361
harv  0.01374864
```

```
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  0.00009103223
act  0.00645689709
ibs  0.00092946958
harv 0.00015637444
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-36

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	2026.022	-420.715
	(2884.751)	(2687.758)
SAT	8.536*	8.236*
	(1.687)	(1.557)
ACT	274.831*	282.318*
	(55.342)	(51.251)
Iowa BS	20.619	18.439
	(26.263)	(24.311)
Major: Soc.	.	2392.588*
		(557.85)
Major: Nat.	.	5103.866*
		(547.602)
Prof. Parents: Yes	.	1175.817*
		(480.545)
Parent Network: Yes	.	739.003
		(488.283)
Gender: Male	.	-559.005
		(445.721)
N	517	517
RMSE	5444.716	5015.348
R^2	0.155	0.29
adj R^2	0.15	0.279

$*p \leq 0.05$

```
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 =  436.813784292402 Denominator =  705.924372314972"
```

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

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

```
tval
```

```
difference
 0.6187827
```

```
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
 0.5363369
```

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
     436.8137843       705.9243723         0.6187827       508.0000000         0.5363369
```

```
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    513 15207848030
2    508 12778087532  5 2429760498 19.319 < 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

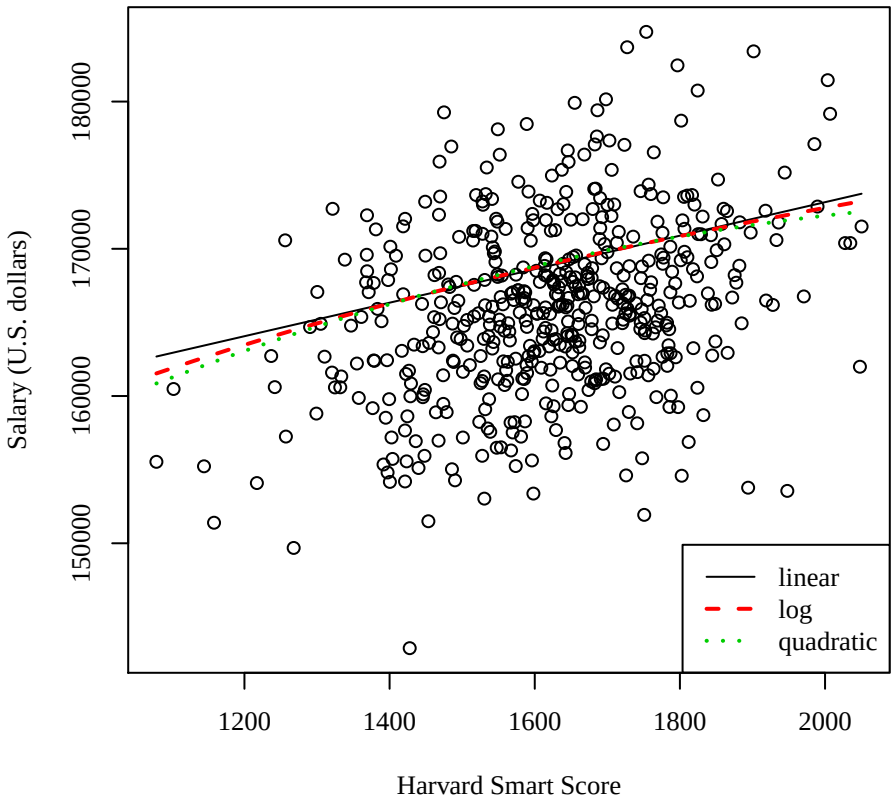
Table 4: Regression with sal3: Student-36

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	144456.535*	28356.738	125773.91*
	(2366.903)	(16789.845)	(15543.683)
Harvard SS	11.379*	.	34.864
	(1.437)		(19.364)
Gender: Male	-997.655*	-993.875*	-996.212*
	(457.09)	(456.509)	(456.871)
Major: Soc.	2447.765*	2447.285*	2453.164*
	(574.444)	(573.687)	(574.184)
Major: Nat.	5951.357*	5959.462*	5972.921*
	(569.705)	(568.911)	(569.706)
Prof. Parents: Yes	1165.72*	1154.09*	1138.95*
	(487.188)	(486.563)	(487.45)
Parent Network: Yes	306.343	292.469	279.681
	(499.017)	(498.446)	(499.257)
ln(Harvard SS)	.	18217.091*	.
		(2274.894)	
Harvard SS^2	.	.	-0.007
			(0.006)
N	502	502	502
RMSE	5090.006	5083.421	5087.545
R^2	0.279	0.281	0.282
adj R^2	0.271	0.273	0.271

$*p \leq 0.05$

```
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%) 26032.49     N
S (30%) 23714.23     S
H (30%) 20839.65     H

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
N (40%) 26032.49      N
S (30%) 23714.23      S
H (30%) 20839.65      H

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

Table 5: Categorical Regressions: Student-36

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	20839.645* (440.433)	23714.232* (403.313)	-420.715 (2687.758)	1971.873 (2690.932)
Major: Soc.	2874.587* (597.196)	.	2392.588* (557.85)	.
Major: Nat.	5192.844* (587.868)	.	5103.866* (547.602)	.
Major 2: Hum.	.	-2874.587* (597.196)	.	-2392.588* (557.85)
Major 2: Nat.	.	2318.257* (560.596)	.	2711.278* (528.756)
SAT	.	.	8.236* (1.557)	8.236* (1.557)
ACT	.	.	282.318* (51.251)	282.318* (51.251)
Iowa BS	.	.	18.439 (24.311)	18.439 (24.311)
Prof. Parents: Yes	.	.	1175.817* (480.545)	1175.817* (480.545)
Parent Network: Yes	.	.	739.003 (488.283)	739.003 (488.283)
Gender: Male	.	.	-559.005 (445.721)	-559.005 (445.721)
N	559	559	517	517
RMSE	5588.471	5588.471	5015.348	5015.348
R^2	0.123	0.123	0.29	0.29
adj R^2	0.12	0.12	0.279	0.279

$*p \leq 0.05$