Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 49
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%	5.06	1189.00	71.74	1649.00	3637.00	148900.00	1172.00
25%	18.81	1501.00	93.14	16660.00	19220.00	162000.00	1480.00
50%	22.22	1611.00	99.53	20330.00	23620.00	165700.00	1587.00
75%	25.45	1727.00	106.20	24000.00	27130.00	169500.00	1703.00
100%	37.57	2013.00	128.40	37440.00	42510.00	182900.00	2093.00
mean	21.97	1612.00	99.64	20410.00	23400.00	165900.00	1592.00
sd	4.91	158.00	9.85	5568.00	5987.00	5475.00	157.40
var	24.07	24970.00	96.99	31000000.00	35840000.00	29970000.00	24770.00
NA's	14.00	57.00	0.00	6.00	0.00	0.00	29.00
N	573.00	573.00	573.00	573.00	573.00	573.00	573.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
             gender                     major                    pnet
M             :291.0000  N             :197.0000  NO            :390.0000
F             :282.0000  H             :189.0000  YES           :183.0000
NA's          :  0.0000  S             :187.0000  NA's          :  0.0000
entropy       :  0.9998  NA's          :  0.0000  entropy       :  0.9037
normedEntropy:   0.9998  entropy       :  1.5846  normedEntropy:   0.9037
N             :573.0000  normedEntropy:   0.9998  N             :573.0000
                         N             :573.0000
             pprof
NO            :408.0000
YES           :165.0000
NA's          :  0.0000
entropy       :  0.8661
normedEntropy:   0.8661
N             :573.0000
```

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: 0x1790200>
act ~ sat + ibs + harv
<environment: 0x1790200>
ibs ~ sat + act + harv
<environment: 0x1790200>
harv ~ sat + act + ibs
<environment: 0x1790200>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998445 0.8731232 0.2745554 0.9998490
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
6431.858894    7.881664    1.378465 6623.275766
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.44 0.44 1.00
act  0.44 1.00 0.44 0.46
ibs  0.44 0.44 1.00 0.45
harv 1.00 0.46 0.45 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-49

	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)	-2395.083	9386.866*	3179.789	-3286.397	-5408.059*	-5460.342*
	(2253.281)	(981.442)	(2268.293)	(2291.875)	(2713.769)	(2586.023)
SAT	14.379*	.	.	.	174.182	8.864*
	(1.41)				(114.158)	(1.574)
ACT	.	501.253*	.	.	532.095*	344.559*
		(43.58)			(127.217)	(50.057)
Iowa BS	.	.	173.069*	.	37.597	42.744
			(22.669)		(26.63)	(25.062)
Harvard SS	.	.	.	14.672*	-165.373	.
				(1.416)	(114.136)	
N	538	553	567	510	472	525
RMSE	5144.615	5034.997	5305.542	5049.954	4847.456	4896.945
R^2	0.163	0.194	0.094	0.175	0.267	0.255
adj R^2	0.161	0.192	0.092	0.173	0.261	0.25

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    469 1.1064e+10
2    467 1.0973e+10  2  90259731 1.9206 0.1477
```

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)	-2745.147	9468.935*	2894.324	-3544.453	-5408.059*	-5460.342*
	(2290.987)	(1013.944)	(2382.288)	(2400.53)	(2713.769)	(2586.023)
SAT	14.591*	.	.	.	174.182	8.864*
	(1.434)				(114.158)	(1.574)
ACT	.	500.673*	.	.	532.095*	344.559*
		(45.084)			(127.217)	(50.057)
Iowa BS	.	.	176.162*	.	37.597	42.744
			(23.781)		(26.63)	(25.062)
Harvard SS	.	.	.	14.851*	-165.373	.
				(1.483)	(114.136)	
N	525	525	525	472	472	525
RMSE	5172.687	5092.957	5386.178	5124.617	4847.456	4896.945
R^2	0.165	0.191	0.095	0.176	0.267	0.255
adj R^2	0.164	0.189	0.093	0.174	0.261	0.25

$*p \leq 0.05$

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    0.23946354
act    0.28872066
ibs    0.07451178
```

```
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  0.045333556
act  0.067772186
ibs  0.004160663
```

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%    42.33    34.85    33.23
50%    52.63    46.98    45.08
75%    62.75    59.40    57.54
100%  100.00   100.00   100.00
```

```
mean    51.93    47.05    45.40
sd      15.18    18.25    17.10
var    230.40   333.10   292.30
NA's     0.00     0.00     0.00
N      525.00   525.00   525.00

$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
-17954.1   -3316.8     142.1    3372.7   13088.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  9838.18     835.59  11.774  < 2e-16 ***
satpoms        81.71      14.51   5.630 2.95e-08 ***
actpoms       112.02      16.27   6.883 1.69e-11 ***
ibspoms        23.17      13.59   1.706   0.0887 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4897 on 521 degrees of freedom
Multiple R^2: 0.2548,  Adjusted R^2: 0.2505
F-statistic: 59.37 on 3 and 521 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    0.07043023
act    0.19002051
ibs    0.06519177
harv  -0.06689734
```

```
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   0.003652332
act   0.027445105
ibs   0.003126998
harv  0.003293502
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-49

	Test Scores Only	All Predictors
	Estimate	Estimate
	(S.E.)	(S.E.)
(Intercept)	-2744.117	-5377.035*
	(2816.641)	(2594.468)
SAT	8.745*	8.843*
	(1.722)	(1.568)
ACT	350.897*	341.982*
	(54.733)	(49.822)
Iowa BS	46.083	43.571
	(27.399)	(25.005)
Major: Soc.	.	2501.366*
		(527.531)
Major: Nat.	.	5065.077*
		(515.572)
Prof. Parents: Yes	.	1209.976*
		(469.236)
Parent Network: Yes	.	1046.193*
		(459.165)
Gender: Male	.	-648.783
		(426.096)
N	531	531
RMSE	5373.66	4890.23
R^2	0.225	0.364
adj R^2	0.22	0.354

$*p \leq 0.05$

```
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 =  163.783340649466 Denominator =  684.132015156818"
```

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

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

```
tval
```

```
difference
 0.2394031
```

```
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
  0.810887
```

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
      163.7833406       684.1320152         0.2394031       522.0000000         0.8108870
```

```
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    527 15217771581
2    522 12483291013  5 2734480568 22.869 < 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

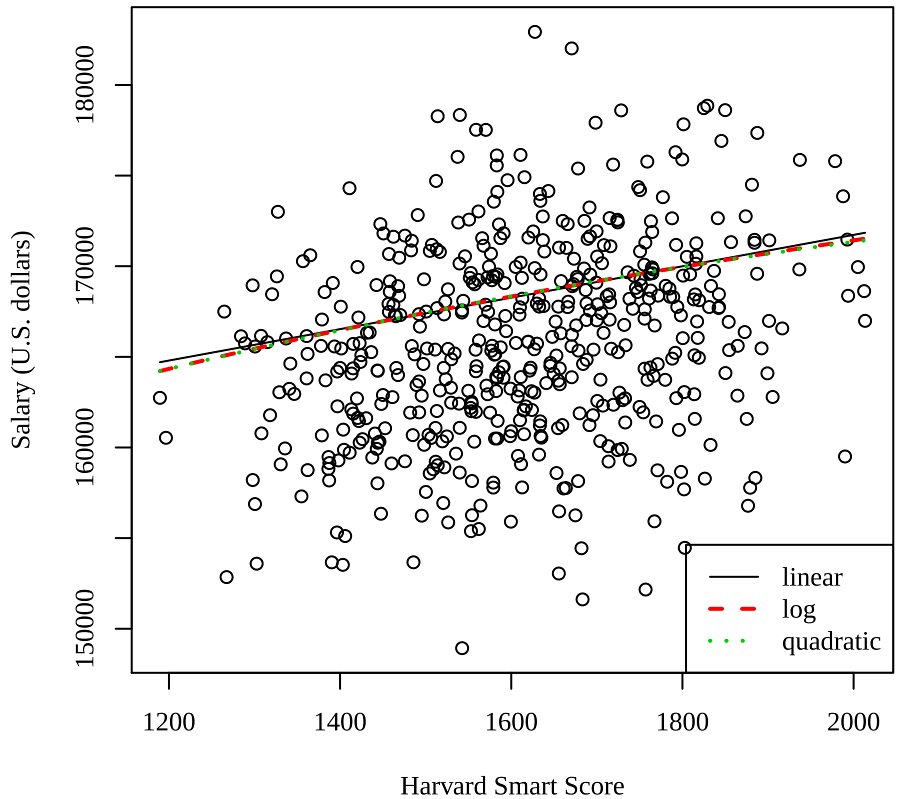
Table 4: Regression with sal3: Student-49

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	149669.073*	61208.083*	141616.199*
	(2236.28)	(15986.963)	(17985.16)
Harvard SS	8.666*	.	18.748
	(1.355)		(22.383)
Gender: Male	-168.084	-173.388	-174.206
	(427.573)	(427.507)	(428.123)
Major: Soc.	1639.216*	1636.913*	1637.821*
	(530.85)	(530.764)	(531.275)
Major: Nat.	4896.364*	4892.887*	4893.704*
	(523.635)	(523.561)	(524.078)
Prof. Parents: Yes	839.95	826.57	823.591
	(474.498)	(474.442)	(476.252)
Parent Network: Yes	-650.08	-649.388	-650.184
	(459.357)	(459.285)	(459.717)
ln(Harvard SS)	.	13879.835*	.
		(2165.334)	
Harvard SS^2	.	.	-0.003
			(0.007)
N	516	516	516
RMSE	4854.815	4854.03	4858.618
R^2	0.215	0.215	0.216
adj R^2	0.206	0.206	0.205

$*p \leq 0.05$

```
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 (30%) 26043.74     N
H (30%) 20813.81     H
S (30%) 23231.70     S

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

```
predictOMatic(cm2)
```

```
$major2
              fit major2
N (30%) 26043.74      N
H (30%) 20813.81      H
S (30%) 23231.70      S

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

Table 5: Categorical Regressions: Student-49

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	20813.81*	23231.702*	-5377.035*	-2875.668
	(407.123)	(409.294)	(2594.468)	(2587.357)
Major: Soc.	2417.891*	.	2501.366*	.
	(577.296)		(527.531)	
Major: Nat.	5229.929*	.	5065.077*	.
	(569.883)		(515.572)	
Major 2: Hum.	.	-2417.891*	.	-2501.366*
		(577.296)		(527.531)
Major 2: Nat.	.	2812.038*	.	2563.711*
		(571.436)		(520.943)
SAT	.	.	8.843*	8.843*
			(1.568)	(1.568)
ACT	.	.	341.982*	341.982*
			(49.822)	(49.822)
Iowa BS	.	.	43.571	43.571
			(25.005)	(25.005)
Prof. Parents: Yes	.	.	1209.976*	1209.976*
			(469.236)	(469.236)
Parent Network: Yes	.	.	1046.193*	1046.193*
			(459.165)	(459.165)
Gender: Male	.	.	-648.783	-648.783
			(426.096)	(426.096)
N	573	573	531	531
RMSE	5597.01	5597.01	4890.23	4890.23
R^2	0.129	0.129	0.364	0.364
adj R^2	0.126	0.126	0.354	0.354

$*p \leq 0.05$