Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 32
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%	7.01	1235.00	70.34	4785.00	8770.00	149600.00	1214.00
25%	18.92	1513.00	94.23	17200.00	19810.00	161600.00	1485.00
50%	22.22	1619.00	100.90	20530.00	23460.00	165300.00	1589.00
75%	25.43	1743.00	107.60	24120.00	27390.00	169600.00	1709.00
100%	36.33	2085.00	129.00	37560.00	42280.00	180000.00	2064.00
mean	21.98	1624.00	100.80	20520.00	23520.00	165600.00	1597.00
sd	4.96	158.90	9.78	5265.00	5742.00	5539.00	155.60
var	24.57	25260.00	95.60	27720000.00	32970000.00	30680000.00	24200.00
NA's	15.00	47.00	0.00	9.00	0.00	0.00	28.00
N	525.00	525.00	525.00	525.00	525.00	525.00	525.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet
M             :284.0000   S             :187.0000   NO            :384.0000
F             :241.0000   N             :177.0000   YES           :141.0000
NA's          :  0.0000   H             :161.0000   NA's          :  0.0000
entropy       :  0.9952   NA's          :  0.0000   entropy       :  0.8394
normedEntropy:  0.9952   entropy       :  1.5822   normedEntropy:  0.8394
N             :525.0000   normedEntropy:  0.9983   N             :525.0000
                          N             :525.0000
          pprof
NO            :346.0000
YES           :179.0000
NA's          :  0.0000
entropy       :  0.9257
normedEntropy:  0.9257
N             :525.0000
```

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: 0x2a08f98>
act ~ sat + ibs + harv
<environment: 0x2a08f98>
ibs ~ sat + act + harv
<environment: 0x2a08f98>
harv ~ sat + act + ibs
<environment: 0x2a08f98>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998240 0.8501209 0.2242051 0.9998280
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
5682.072185    6.672043    1.289001 5813.060764
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.35 0.40 1.00
act  0.35 1.00 0.38 0.37
ibs  0.40 0.38 1.00 0.40
harv 1.00 0.37 0.40 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-32

	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)	1097.812	13323.748*	6171.482*	2076.412	-1423.95	-2711.891
	(2301.596)	(1027.683)	(2336.616)	(2288.902)	(2828.647)	(2756.156)
SAT	12.136*	.	.	.	-110.181	8.579*
	(1.435)				(110.761)	(1.61)
ACT	.	328.192*	.	.	97.428	190.392*
		(45.52)			(118.711)	(49.806)
Iowa BS	.	.	142.337*	.	51.505	52.776*
			(23.065)		(26.371)	(26.089)
Harvard SS	.	.	.	11.358*	117.675	.
				(1.403)	(110.789)	
N	488	501	516	470	431	473
RMSE	4928.196	5014.473	5085.301	4830.226	4709.355	4818.272
R^2	0.128	0.094	0.069	0.123	0.169	0.17
adj R^2	0.126	0.093	0.067	0.121	0.161	0.165

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F  Pr(>F)
1    428 9562682495
2    426 9447838578  2 114843917 2.5891 0.07627 .
---
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)	1213.904	13379.852*	5817.584*	2299.81	-1423.95	-2711.891
	(2346.253)	(1052.569)	(2454.13)	(2433.753)	(2828.647)	(2756.156)
SAT	12.069*	.	.	.	-110.181	8.579*
	(1.461)				(110.761)	(1.61)
ACT	.	324.53*	.	.	97.428	190.392*
		(46.814)			(118.711)	(49.806)
Iowa BS	.	.	145.504*	.	51.505	52.776*
			(24.213)		(26.371)	(26.089)
Harvard SS	.	.	.	11.223*	117.675	.
				(1.493)	(110.789)	
N	473	473	473	431	431	473
RMSE	4932.654	5027.427	5086.284	4838.169	4709.355	4818.272
R^2	0.126	0.093	0.071	0.116	0.169	0.17
adj R^2	0.125	0.091	0.069	0.114	0.161	0.165

$*p \leq 0.05$

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   0.23886511
act   0.17382703
ibs   0.09300576
```

```
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  0.050219113
act  0.025858825
ibs  0.007241724
```

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%    40.55   41.16   32.50
50%    51.57   52.20   44.64
75%    62.72   63.43   58.19
100%  100.00  100.00  100.00
```

```
mean    50.90    52.13    45.13
sd      16.86    16.49    18.28
var    284.20   272.10   334.30
NA's     0.00     0.00     0.00
N      473.00   473.00   473.00

$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
-12635.6  -3047.4    224.6   3184.0  15732.4

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12754.34     876.52  14.551  < 2e-16 ***
satpoms        72.90      13.69   5.327 1.55e-07 ***
actpoms        55.82      14.60   3.823  0.00015 ***
ibspoms        30.94      15.29   2.023  0.04365 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4818 on 469 degrees of freedom
Multiple R^2: 0.1701,   Adjusted R^2: 0.1647
F-statistic: 32.03 on 3 and 469 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  -0.04814062
act   0.03973237
ibs   0.09420646
harv  0.05139350
```

```
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   0.001931056
act   0.001314435
ibs   0.007443836
harv  0.002201552
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-32

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	626.734	-2288.092
	(2994.819)	(2749.535)
SAT	8.73*	8.736*
	(1.759)	(1.598)
ACT	202.396*	190.634*
	(54.218)	(49.131)
Iowa BS	44.261	45.103
	(28.403)	(25.905)
Major: Soc.	.	1099.447*
		(540.305)
Major: Nat.	.	4776.304*
		(545.212)
Prof. Parents: Yes	.	1709.057*
		(465.562)
Parent Network: Yes	.	1709.407*
		(496.033)
Gender: Male	.	78.981
		(444.088)
N	482	482
RMSE	5305.379	4804.463
R^2	0.147	0.308
adj R^2	0.142	0.296

$*p \leq 0.05$

```
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 =  -0.349212295007874 Denominator =  651.808894562018"
```

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

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

```
tval
```

```
   difference
-0.0005357587
```

```
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
 0.9995728
```

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
   -0.3492122950    651.8088945620    -0.0005357587    473.0000000000    0.9995727523
```

```
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    478 13454286180
2    473 10918196826  5 2536089354 21.974 < 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

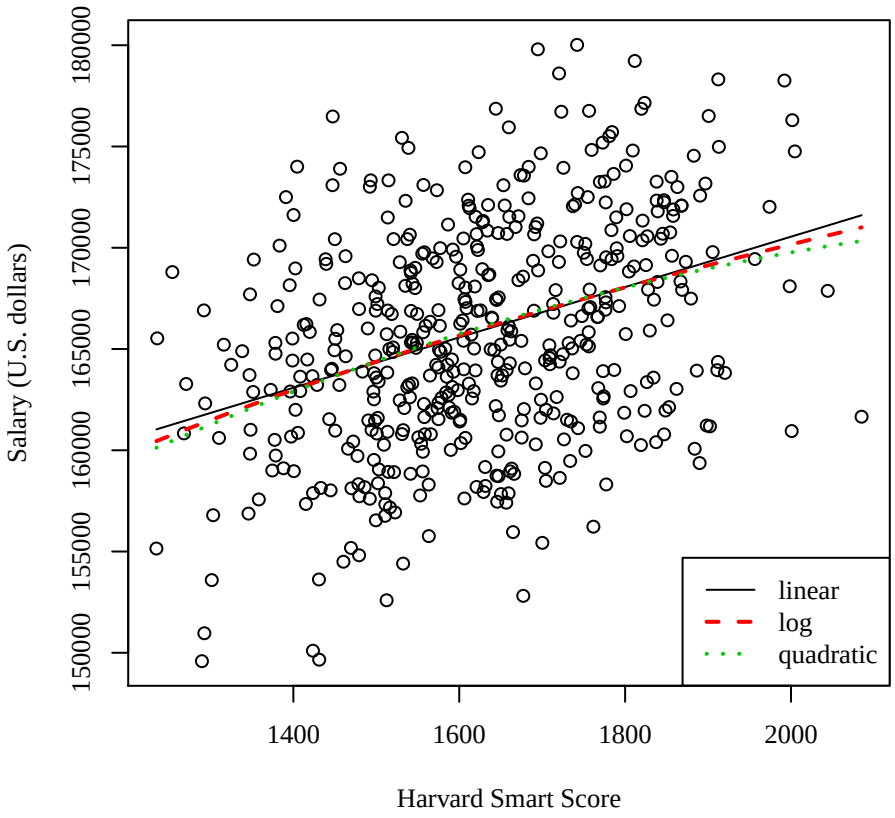
Table 4: Regression with sal3: Student-32

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	141958.453*	13347.98	123571.76*
	(2305.424)	(16438.728)	(18368.714)
Harvard SS	12.436*	.	35.191
	(1.379)		(22.596)
Gender: Male	881.689*	876.168*	870.053*
	(440.819)	(440.496)	(440.961)
Major: Soc.	2834.595*	2857.017*	2876.058*
	(541.056)	(540.678)	(542.604)
Major: Nat.	4830.961*	4855.338*	4878.571*
	(548.748)	(548.298)	(550.763)
Prof. Parents: Yes	426.595	412.718	399.915
	(463.733)	(463.369)	(464.478)
Parent Network: Yes	603.864	621.871	634.449
	(495.787)	(495.511)	(496.704)
ln(Harvard SS)	.	20140.216*	.
		(2222.818)	
Harvard SS^2	.	.	-0.007
			(0.007)
N	478	478	478
RMSE	4781.099	4777.641	4781.008
R^2	0.263	0.264	0.264
adj R^2	0.253	0.254	0.253

$*p \leq 0.05$

```
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 (40%) 22534.81     S
N (30%) 26504.04     N
H (30%) 21369.49     H

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
S (40%) 22534.81      S
N (30%) 26504.04      N
H (30%) 21369.49      H

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

Table 5: Categorical Regressions: Student-32

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	21369.491* (419.285)	22534.806* (389.046)	-2288.092 (2749.535)	-1188.645 (2746.569)
Major: Soc.	1165.314* (571.976)	.	1099.447* (540.305)	.
Major: Nat.	5134.547* (579.403)	.	4776.304* (545.212)	.
Major 2: Hum.	.	-1165.314* (571.976)	.	-1099.447* (540.305)
Major 2: Nat.	.	3969.232* (557.911)	.	3676.857* (530.152)
SAT	.	.	8.736* (1.598)	8.736* (1.598)
ACT	.	.	190.634* (49.131)	190.634* (49.131)
Iowa BS	.	.	45.103 (25.905)	45.103 (25.905)
Prof. Parents: Yes	.	.	1709.057* (465.562)	1709.057* (465.562)
Parent Network: Yes	.	.	1709.407* (496.033)	1709.407* (496.033)
Gender: Male	.	.	78.981 (444.088)	78.981 (444.088)
N	525	525	482	482
RMSE	5320.125	5320.125	4804.463	4804.463
R^2	0.145	0.145	0.308	0.308
adj R^2	0.142	0.142	0.296	0.296

$*p \leq 0.05$