Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 2
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%	6.85	1205.00	72.17	2421.00	5431.00	143400.00	1183.00
25%	19.26	1520.00	93.15	16790.00	19600.00	161300.00	1497.00
50%	22.65	1622.00	99.56	20560.00	23640.00	165100.00	1599.00
75%	25.69	1739.00	106.50	24420.00	27850.00	169300.00	1714.00
100%	37.45	2092.00	133.10	37210.00	42340.00	181600.00	2072.00
mean	22.44	1624.00	99.87	20590.00	23610.00	165200.00	1602.00
sd	4.88	157.20	9.61	5315.00	5795.00	5904.00	152.90
var	23.81	24730.00	92.40	28250000.00	33580000.00	34850000.00	23380.00
NA's	23.00	48.00	0.00	8.00	0.00	0.00	28.00
N	556.00	556.00	556.00	556.00	556.00	556.00	556.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
           gender                   major                    pnet                    pprof
M            :302.0000    N            :197.000    NO           :401.0000    NO           :388
    .0000
F            :254.0000    H            :183.000    YES          :155.0000    YES          :168
    .0000
NA's         :  0.0000    S            :176.000    NA's         :  0.0000    NA's         :  0
    .0000
entropy      :  0.9946    NA's         :  0.000    entropy      :  0.8538    entropy      :  0
    .8839
normedEntropy:  0.9946    entropy      :  1.583    normedEntropy:  0.8538    normedEntropy:  0
    .8839
N            :556.0000    normedEntropy:  0.999    N            :556.0000    N            :556
    .0000
                          N            :556.000
```

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: 0x14d4d78>
act ~ sat + ibs + harv
<environment: 0x14d4d78>
ibs ~ sat + act + harv
<environment: 0x14d4d78>
harv ~ sat + act + ibs
<environment: 0x14d4d78>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998256 0.8443680 0.2284092 0.9998294
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
5733.412202    6.425414    1.296024 5862.322878
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 0.38
ibs  0.40 0.38 1.00 0.41
harv 1.00 0.38 0.41 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-2

	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)	-1092.42	13375.928*	3186.927	-2022.885	-7392.686*	-6123.751*
	(2244.669)	(1033.894)	(2247.002)	(2257.011)	(2777.649)	(2647.4)
SAT	13.492*	.	.	.	-64.856	9.93*
	(1.395)				(109.564)	(1.572)
ACT	.	320.81*	.	.	51.397	129.002*
		(45.037)			(117.615)	(49.296)
Iowa BS	.	.	174.371*	.	88.253*	78.277*
			(22.411)		(26.327)	(25.398)
Harvard SS	.	.	.	13.971*	75.026	.
				(1.383)	(109.591)	
N	520	525	548	500	453	497
RMSE	4860.962	5055.813	5047.503	4861.181	4734.534	4730.343
R^2	0.153	0.088	0.1	0.17	0.203	0.192
adj R^2	0.151	0.087	0.098	0.168	0.196	0.188

$*p \le 0.05$

```
  Res.Df        RSS Df Sum of Sq      F   Pr(>F)
1    450 1.0310e+10
2    448 1.0042e+10  2 267289005 5.9621 0.002783 **
---
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)	-1097.699 (2278.37)	13584.928* (1075.933)	3526.911 (2334.508)	-1752.927 (2386.134)	-7392.686* (2777.649)	-6123.751* (2647.4)
SAT	13.48* (1.416)	.	.	.	-64.856 (109.564)	9.93* (1.572)
ACT	.	306.735* (46.69)	.	.	51.397 (117.615)	129.002* (49.296)
Iowa BS	.	.	170.157* (23.302)	.	88.253* (26.327)	78.277* (25.398)
Harvard SS	.	.	.	13.729* (1.461)	75.026 (109.591)	.
N	497	497	497	453	453	497
RMSE	4829.736	5038.308	4991.39	4833.842	4734.534	4730.343
R^2	0.155	0.08	0.097	0.164	0.203	0.192
adj R^2	0.153	0.078	0.095	0.162	0.196	0.188

$*p \leq 0.05$

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 -1.0000000
sat   0.2736401
act   0.1170487
ibs   0.1374864
```

```
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   0.06536017
act   0.01121700
ibs   0.01555821
```

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%     40.82   34.12   35.32
50%     51.96   44.27   46.82
75%     61.90   56.26   59.79
100%   100.00  100.00  100.00
```

```
mean    51.24    45.24    47.14
sd      15.83    15.79    17.23
var    250.70   249.40   297.00
NA's     0.00     0.00     0.00
N      497.00   497.00   497.00

$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
-12804.4   -3157.9    -158.6    3355.3   12070.7

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12156.14     844.16  14.400  < 2e-16 ***
satpoms        88.25      13.97   6.317 5.97e-10 ***
actpoms        39.47      15.08   2.617  0.00915 **
ibspoms        47.67      15.47   3.082  0.00217 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4730 on 493 degrees of freedom
Multiple R^2: 0.1925,  Adjusted R^2: 0.1876
F-statistic: 39.17 on 3 and 493 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  -0.02795593
act   0.02064160
ibs   0.15642398
harv  0.03232751
```

```
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  0.0006233564
act  0.0003397195
ibs  0.0199900671
harv 0.0008337723
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-2

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	-4105.968	-6327.032*
	(2949.214)	(2691.611)
SAT	10.369*	10.116*
	(1.738)	(1.568)
ACT	130.698*	129.252*
	(54.88)	(49.541)
Iowa BS	81.308*	74.759*
	(28.076)	(25.372)
Major: Soc.	.	2096.556*
		(531.402)
Major: Nat.	.	5424.199*
		(513.153)
Prof. Parents: Yes	.	1007.722*
		(461.643)
Parent Network: Yes	.	879.344
		(475.052)
Gender: Male	.	240.606
		(426.403)
N	505	505
RMSE	5289.758	4749.948
R^2	0.169	0.337
adj R^2	0.164	0.326

$*p \leq 0.05$

```
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 =  128.378282618634 Denominator =  668.841737553164"
```

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

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

```
tval
```

```
difference
 0.1919412
```

```
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
 0.8478668
```

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
      128.3782826       668.8417376         0.1919412       496.0000000         0.8478668
```

```
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    501 14018748883
2    496 11190755743  5 2827993140 25.069 < 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

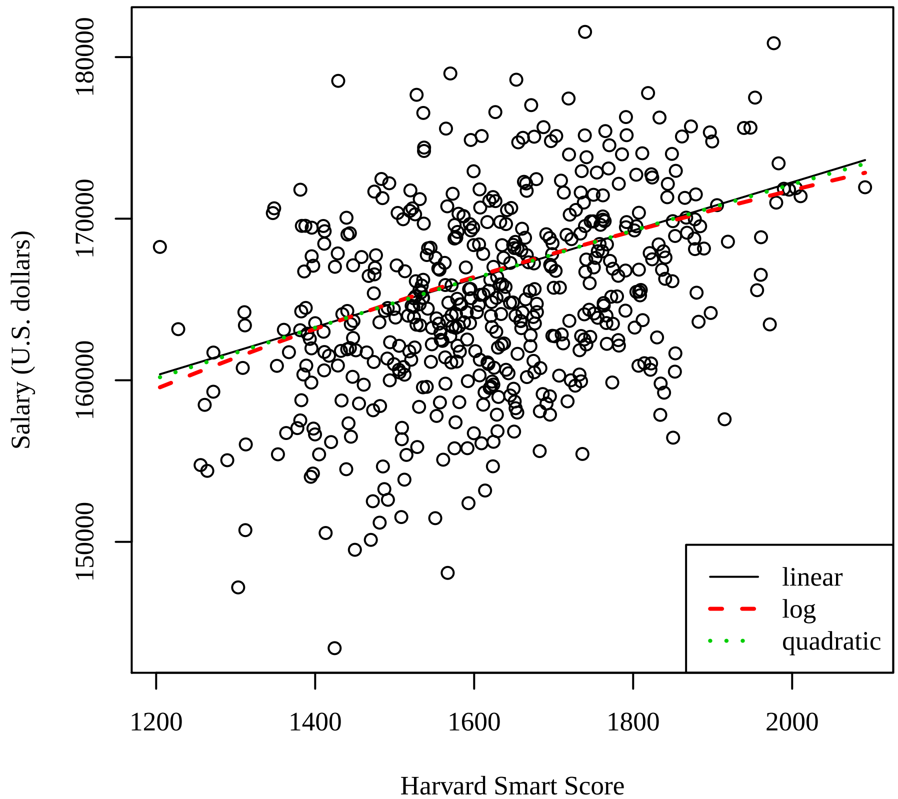
Table 4: Regression with sal3: Student-2

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	137789.01*	-15730.952	134641.608*
	(2326.083)	(16676.52)	(18060.017)
Harvard SS	14.942*	.	18.837
	(1.399)		(22.211)
Gender: Male	-819.797	-818.81	-819.348
	(439.176)	(439.295)	(439.609)
Major: Soc.	3146.001*	3144.324*	3146.255*
	(539.898)	(540.033)	(540.423)
Major: Nat.	5402.677*	5412.402*	5405.601*
	(530.225)	(530.376)	(530.999)
Prof. Parents: Yes	1804.167*	1823.416*	1809.102*
	(477.648)	(477.732)	(478.934)
Parent Network: Yes	308.605	313.681	310.125
	(483.576)	(483.669)	(484.121)
ln(Harvard SS)	.	24062.819*	.
		(2255.635)	
Harvard SS^2	.	.	-0.001
			(0.007)
N	508	508	508
RMSE	4910.905	4912.216	4915.662
R^2	0.327	0.327	0.327
adj R^2	0.319	0.319	0.318

$*p \leq 0.05$

```
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%) 26421.01     N
H (30%) 21245.15     H
S (30%) 22916.32     S

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

```
predictOMatic(cm2)
```

```
$major2
              fit major2
N (40%) 26421.01      N
H (30%) 21245.15      H
S (30%) 22916.32      S

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

Table 5: Categorical Regressions: Student-2

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	21245.148*	22916.319*	-6327.032*	-4230.477
	(397.253)	(405.076)	(2691.611)	(2652.517)
Major: Soc.	1671.171*	.	2096.556*	.
	(567.36)		(531.402)	
Major: Nat.	5175.857*	.	5424.199*	.
	(551.73)		(513.153)	
Major 2: Hum.	.	-1671.171*	.	-2096.556*
		(567.36)		(531.402)
Major 2: Nat.	.	3504.686*	.	3327.643*
		(557.389)		(515.735)
SAT	.	.	10.116*	10.116*
			(1.568)	(1.568)
ACT	.	.	129.252*	129.252*
			(49.541)	(49.541)
Iowa BS	.	.	74.759*	74.759*
			(25.372)	(25.372)
Prof. Parents: Yes	.	.	1007.722*	1007.722*
			(461.643)	(461.643)
Parent Network: Yes	.	.	879.344	879.344
			(475.052)	(475.052)
Gender: Male	.	.	240.606	240.606
			(426.403)	(426.403)
N	556	556	505	505
RMSE	5373.945	5373.945	4749.948	4749.948
R^2	0.143	0.143	0.337	0.337
adj R^2	0.14	0.14	0.326	0.326

$*p \leq 0.05$