Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 50
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%	8.42	1107.00	68.10	524.20	5456.00	143600.00	1089.00
25%	18.95	1504.00	92.94	16340.00	19460.00	161900.00	1489.00
50%	22.08	1616.00	100.40	20460.00	23600.00	165600.00	1600.00
75%	25.34	1722.00	105.80	24420.00	27540.00	169600.00	1701.00
100%	35.08	2106.00	125.00	35200.00	40320.00	183800.00	2087.00
mean	22.07	1610.00	99.72	20330.00	23390.00	165600.00	1594.00
sd	4.65	160.00	9.53	5527.00	5963.00	5808.00	158.90
var	21.64	25610.00	90.83	30540000.00	35560000.00	33730000.00	25240.00
NA's	14.00	54.00	0.00	9.00	0.00	0.00	33.00
N	486.00	486.00	486.00	486.00	486.00	486.00	486.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet
F             :249.0000   N             :176.0000   NO            :339.0000
M             :237.0000   S             :156.0000   YES           :147.0000
NA's          :  0.0000   H             :154.0000   NA's          :  0.0000
entropy       :  0.9996   NA's          :  0.0000   entropy       :  0.8843
normedEntropy:  0.9996   entropy       :  1.5823   normedEntropy:  0.8843
N             :486.0000   normedEntropy:  0.9983   N             :486.0000
                          N             :486.0000
          pprof
NO            :338.0000
YES           :148.0000
NA's          :  0.0000
entropy       :  0.8868
normedEntropy:  0.8868
N             :486.0000
```

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: 0x1bde468>
act ~ sat + ibs + harv
<environment: 0x1bde468>
ibs ~ sat + act + harv
<environment: 0x1bde468>
harv ~ sat + act + ibs
<environment: 0x1bde468>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998488 0.8488331 0.1865056 0.9998523
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
6612.319959    6.615203    1.229265 6769.163002
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.39 0.38 1.00
act  0.39 1.00 0.33 0.42
ibs  0.38 0.33 1.00 0.39
harv 1.00 0.42 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-50

	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)	-223.777	13265.347*	11240.008*	295.221	1151.797	1121.424
	(2439.697)	(1204.447)	(2641.098)	(2529.763)	(3245.734)	(3038.97)
SAT	12.911*	.	.	.	-133.25	10.77*
	(1.522)				(132.975)	(1.734)
ACT	.	319.047*	.	.	48.52	208.34*
		(53.386)			(146.136)	(57.894)
Iowa BS	.	.	91.143*	.	-26.15	-25.265
			(26.357)		(30.813)	(28.507)
Harvard SS	.	.	.	12.403*	144.27	.
				(1.563)	(133.081)	
N	447	463	477	425	387	435
RMSE	5109.912	5351.244	5464.112	5174.443	5190.18	5089.487
R^2	0.139	0.072	0.025	0.13	0.153	0.158
adj R^2	0.137	0.07	0.023	0.127	0.144	0.152

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    384 1.0339e+10
2    382 1.0290e+10  2  48297803 0.8965 0.4089
```

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)	310.476	13006.219*	12783.93*	120.46	1151.797	1121.424
	(2483.26)	(1225.197)	(2789.215)	(2640.071)	(3245.734)	(3038.97)
SAT	12.574*	.	.	.	-133.25	10.77*
	(1.548)				(132.975)	(1.734)
ACT	.	334.948*	.	.	48.52	208.34*
		(54.476)			(146.136)	(57.894)
Iowa BS	.	.	76.201*	.	-26.15	-25.265
			(27.873)		(30.813)	(28.507)
Harvard SS	.	.	.	12.504*	144.27	.
				(1.63)	(133.081)	
N	435	435	435	387	387	435
RMSE	5153.595	5305.349	5484.968	5230.954	5190.18	5089.487
R^2	0.132	0.08	0.017	0.133	0.153	0.158
adj R^2	0.13	0.078	0.015	0.13	0.144	0.152

$*p \leq 0.05$

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   0.28668847
act   0.17079276
ibs  -0.04265101
```

```
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  0.075444012
act  0.025313534
ibs  0.001535346
```

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%    39.29   43.55   40.39
50%    51.43   56.77   51.33
75%    63.41   66.10   61.85
100%  100.00  100.00  100.00
```

```
mean    50.94    55.44    50.79
sd      17.53    16.61    16.01
var    307.50   276.00   256.40
NA's     0.00     0.00     0.00
N      435.00   435.00   435.00

$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
-12597.0  -4111.0     81.9   3802.0  14717.5

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 12884.50    1045.25  12.327  < 2e-16 ***
satpoms       107.46      17.30   6.213 1.23e-09 ***
actpoms        55.54      15.43   3.599 0.000357 ***
ibspoms       -14.37      16.21  -0.886 0.375969
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5089 on 431 degrees of freedom
Multiple R^2: 0.1575,   Adjusted R^2: 0.1517
F-statistic: 26.86 on 3 and 431 DF,  p-value: 6.045e-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  -0.05120320
act   0.01698500
ibs  -0.04338018
harv  0.05538097
```

```
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  0.0022273326
act  0.0002445159
ibs  0.0015975395
harv 0.0026067900
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-50

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	2701.513	896.534
	(3276.963)	(3042.094)
SAT	11.686*	10.573*
	(1.88)	(1.737)
ACT	185.764*	203.341*
	(62.833)	(57.943)
Iowa BS	-20.166	-17.267
	(30.85)	(28.363)
Major: Soc.	.	2907.484*
		(612.195)
Major: Nat.	.	5145.826*
		(591.626)
Prof. Parents: Yes	.	1072.612*
		(522.547)
Parent Network: Yes	.	574.973
		(535.842)
Gender: Male	.	-842.427
		(487.594)
N	441	441
RMSE	5536.233	5079.837
R^2	0.145	0.288
adj R^2	0.139	0.275

$*p \leq 0.05$

```
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 =  497.63946538673 Denominator =  735.905304907821"
```

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

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

```
tval
```

```
difference
 0.6762276
```

```
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
  0.499258
```

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
497.6394654	735.9053049	0.6762276	432.0000000	0.4992580

```
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    437 13393993459
2    432 11147647791  5 2246345668 17.41 1.029e-15 ***
---
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

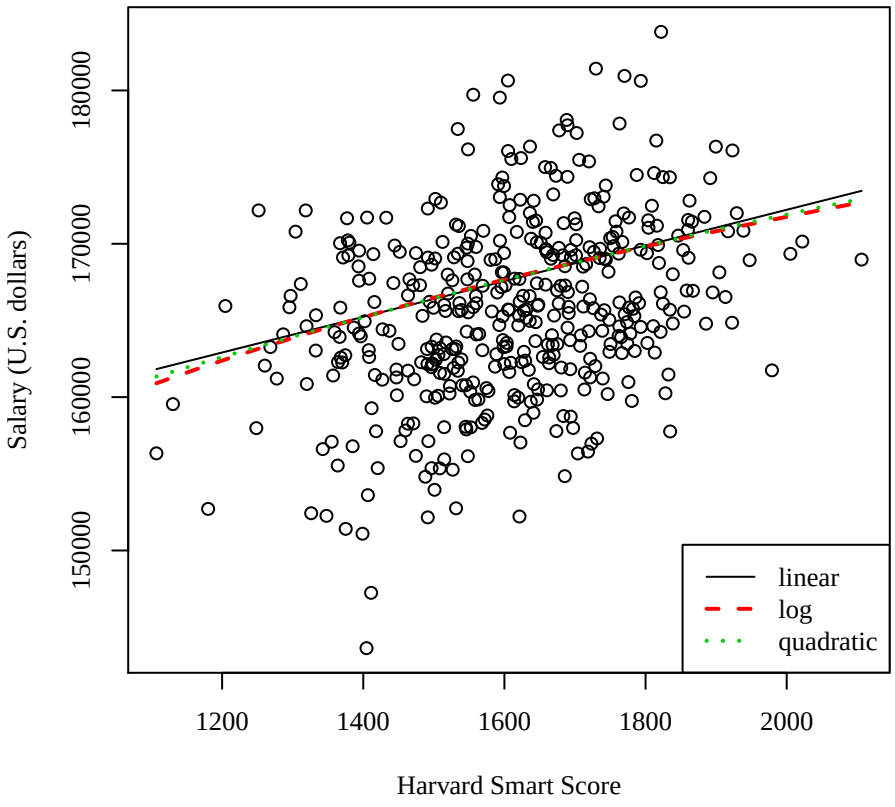
Table 4: Regression with sal3: Student-50

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	143943.543*	27259.039	138243.878*
	(2502.559)	(17897.618)	(17649.65)
Harvard SS	11.635*	.	18.834
	(1.539)		(22.121)
Gender: Male	215.632	206.591	209.608
	(489.571)	(489.612)	(490.434)
Major: Soc.	1694.392*	1706.971*	1703.688*
	(616.79)	(616.722)	(618.096)
Major: Nat.	5004.312*	5016.005*	5012.007*
	(597.646)	(597.561)	(598.74)
Prof. Parents: Yes	1473.184*	1473.178*	1473.215*
	(533.446)	(533.505)	(534.007)
Parent Network: Yes	327.216	330.384	329.048
	(529.872)	(529.909)	(530.46)
ln(Harvard SS)	.	18351.019*	.
		(2429.319)	
Harvard SS^2	.	.	-0.002
			(0.007)
N	432	432	432
RMSE	5068.127	5068.68	5073.463
R^2	0.264	0.264	0.264
adj R^2	0.254	0.253	0.252

$*p \leq 0.05$

```
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%) 25821.49     N
S (30%) 23515.74     S
H (30%) 20491.48     H

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
N (40%) 25821.49      N
S (30%) 23515.74      S
H (30%) 20491.48      H

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

Table 5: Categorical Regressions: Student-50

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	20491.479* (447.705)	23515.744* (444.826)	896.534 (3042.094)	3804.018 (3059.92)
Major: Soc.	3024.265* (631.118)	.	2907.484* (612.195)	.
Major: Nat.	5330.012* (613.046)	.	5145.826* (591.626)	.
Major 2: Hum.	.	-3024.265* (631.118)	.	-2907.484* (612.195)
Major 2: Nat.	.	2305.747* (610.946)	.	2238.342* (586.865)
SAT	.	.	10.573* (1.737)	10.573* (1.737)
ACT	.	.	203.341* (57.943)	203.341* (57.943)
Iowa BS	.	.	-17.267 (28.363)	-17.267 (28.363)
Prof. Parents: Yes	.	.	1072.612* (522.547)	1072.612* (522.547)
Parent Network: Yes	.	.	574.973 (535.842)	574.973 (535.842)
Gender: Male	.	.	-842.427 (487.594)	-842.427 (487.594)
N	486	486	441	441
RMSE	5555.877	5555.877	5079.837	5079.837
R^2	0.135	0.135	0.288	0.288
adj R^2	0.132	0.132	0.275	0.275

$*p \leq 0.05$