Paul Johnson April 25, 2013

Data Management

```
library(foreign)
library(rockchalk)
i <- 17
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%	7.98	1215.00	68.96	1844.00	1792.00	150600.00	1194.00
25%	18.36	1526.00	93.59	16560.00	19120.00	161800.00	1505.00
50%	21.70	1632.00	99.46	20200.00	23110.00	165600.00	1612.00
75%	25.07	1733.00	106.20	24030.00	27450.00	169700.00	1709.00
100%	38.17	2095.00	129.00	40510.00	44170.00	182000.00	2065.00
mean	21.79	1629.00	99.77	20470.00	23340.00	165600.00	1606.00
sd	4.89	152.30	10.10	5626.00	6079.00	5780.00	149.70
var	23.96	23190.00	101.90	31650000.00	36950000.00	33410000.00	22410.00
NA's	20.00	58.00	0.00	7.00	0.00	0.00	39.00
N	586.00	586.00	586.00	586.00	586.00	586.00	586.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

```
          gender                   major                    pnet
M             :299.0000   H             :203.0000   NO            :425.0000
F             :287.0000   S             :195.0000   YES           :161.0000
NA's          :  0.0000   N             :188.0000   NA's          :  0.0000
entropy       :  0.9997   NA's          :  0.0000   entropy       :  0.8482
normedEntropy:  0.9997   entropy       :  1.5843   normedEntropy:  0.8482
N             :586.0000   normedEntropy:  0.9996   N             :586.0000
                          N             :586.0000
          pprof
NO            :417.0000
YES           :169.0000
NA's          :  0.0000
entropy       :  0.8666
normedEntropy:  0.8666
N             :586.0000
```

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: 0x1522708>
act ~ sat + ibs + harv
<environment: 0x1522708>
ibs ~ sat + act + harv
<environment: 0x1522708>
harv ~ sat + act + ibs
<environment: 0x1522708>
Drum roll please!

And your R_j Squareds are (auxiliary Rsq)
      sat       act       ibs      harv
0.9998154 0.8536989 0.2125148 0.9998204
The Corresponding VIF, 1/(1-R_j^2)
        sat         act         ibs        harv
5416.906207    6.835218    1.269865 5568.138421
Bivariate Correlations for design matrix
      sat  act  ibs harv
sat  1.00 0.42 0.37 1.00
act  0.42 1.00 0.40 0.44
ibs  0.37 0.40 1.00 0.38
harv 1.00 0.44 0.38 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-17

	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)	-1748.642	12316.78*	9346.443*	-394.015	-1223.002	-2269.609
	(2432.437)	(1032.367)	(2268.875)	(2486.385)	(3048.388)	(2850.984)
SAT	13.833*	.	.	.	-63.986	10.893*
	(1.509)				(117.672)	(1.738)
ACT	.	377.574*	.	.	146.272	216.007*
		(46.209)			(132.176)	(54.61)
Iowa BS	.	.	111.507*	.	-6.147	6.398
			(22.625)		(27.01)	(25.212)
Harvard SS	.	.	.	12.848*	74.955	.
				(1.52)	(117.617)	
N	540	559	579	522	473	521
RMSE	5253.851	5355.42	5515.578	5295.576	5249.75	5197.498
R^2	0.135	0.107	0.04	0.121	0.161	0.169
adj R^2	0.134	0.105	0.039	0.119	0.154	0.164

$*p \leq 0.05$

```
  Res.Df        RSS Df Sum of Sq      F Pr(>F)
1    470 1.2910e+10
2    468 1.2898e+10  2  12335484 0.2238 0.7996
```

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)	-2127.131	12464.658*	9598.293*	-1803.603	-1223.002	-2269.609
	(2470.148)	(1087.402)	(2403.612)	(2608.251)	(3048.388)	(2850.984)
SAT	14.134*	.	.	.	-63.986	10.893*
	(1.533)				(117.672)	(1.738)
ACT	.	371.265*	.	.	146.272	216.007*
		(48.748)			(132.176)	(54.61)
Iowa BS	.	.	109.739*	.	-6.147	6.398
			(23.964)		(27.01)	(25.212)
Harvard SS	.	.	.	13.766*	74.955	.
				(1.594)	(117.617)	
N	521	521	521	473	473	521
RMSE	5275.646	5397.703	5579.732	5309.935	5249.75	5197.498
R^2	0.141	0.101	0.039	0.137	0.161	0.169
adj R^2	0.139	0.099	0.037	0.135	0.154	0.164

$*p \leq 0.05$

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   0.26567997
act   0.17138796
ibs   0.01115916
```

```
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 0.0630944659
act 0.0251415211
ibs 0.0001034666
```

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%    36.48   40.92   35.59
50%    48.12   50.80   47.74
75%    60.08   61.80   59.15
100%  100.00  100.00  100.00
```

```
mean    48.51    51.37    47.07
sd      17.08    17.02    17.32
var    291.70   289.60   300.10
NA's     0.00     0.00     0.00
N      521.00   521.00   521.00

$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
-17668.0  -3411.5    -87.9   3163.5  20934.0

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 12904.856    870.154  14.831  < 2e-16 ***
satpoms        94.896     15.144   6.266  7.8e-10 ***
actpoms        61.411     15.525   3.955  8.7e-05 ***
ibspoms         3.839     15.127   0.254      0.8
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5197 on 517 degrees of freedom
Multiple R^2: 0.1692,   Adjusted R^2: 0.1644
F-statistic:  35.1 on 3 and 517 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  -0.02512769
act   0.05108807
ibs  -0.01051883
harv  0.02944560
```

```
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   0.00052977056
act   0.00219423120
ibs   0.00009278783
harv  0.00072765570
```

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

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-17

	Test Scores Only Estimate (S.E.)	All Predictors Estimate (S.E.)
(Intercept)	257.419	-2940.523
	(3081.419)	(2892.576)
SAT	12.625*	11.143*
	(1.872)	(1.753)
ACT	201.155*	215.203*
	(58.921)	(54.841)
Iowa BS	-15.017	6.852
	(27.238)	(25.559)
Major: Soc.	.	2206.966*
		(552.838)
Major: Nat.	.	4585.221*
		(562.554)
Prof. Parents: Yes	.	160.728
		(507.436)
Parent Network: Yes	.	1803.058*
		(518.632)
Gender: Male	.	691.894
		(457.056)
N	528	528
RMSE	5627.848	5232.223
R^2	0.157	0.279
adj R^2	0.153	0.268

$*p \leq 0.05$

```
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 =  -1642.32914134853 Denominator =  747.901771077101"
```

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

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

```
tval
```

```
difference
 -2.195916
```

```
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
0.02853974
```

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
  -1642.32914135       747.90177108       -2.19591557       519.00000000       0.02853974
```

```
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    524 16596480198
2    519 14208228439  5 2388251760 17.448 5.528e-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

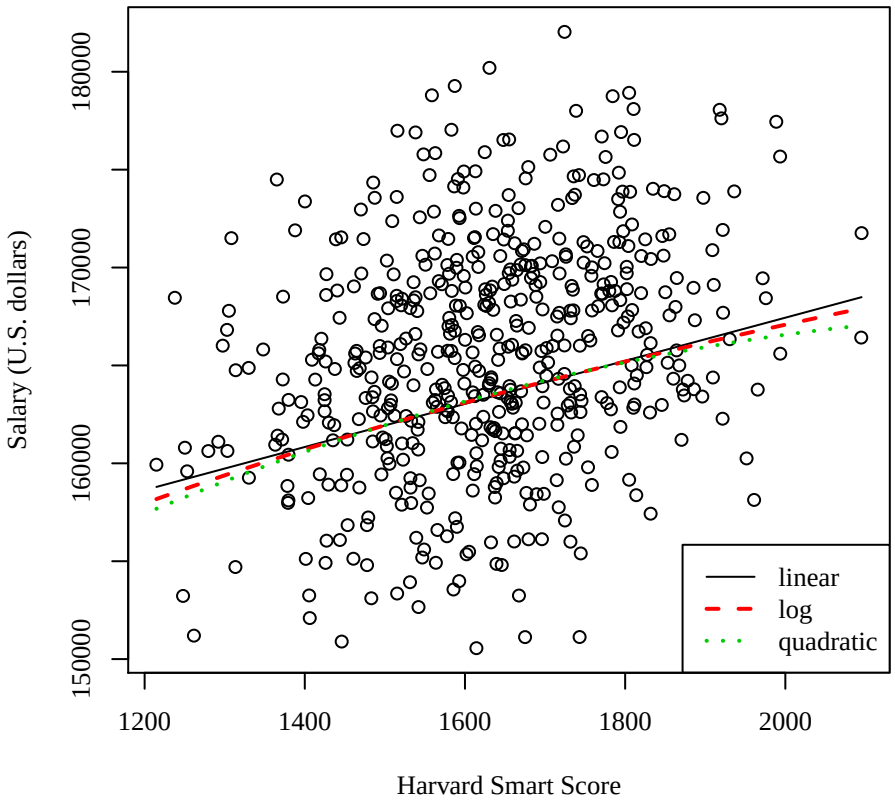
Table 4: Regression with sal3: Student-17

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	145422.799*	31316.088	126237.349*
	(2470.912)	(17927.918)	(18988.227)
Harvard SS	11.008*	.	34.724
	(1.508)		(23.321)
Gender: Male	222.563	234.999	246.874
	(458.022)	(457.641)	(458.626)
Major: Soc.	1592.343*	1599.733*	1609.628*
	(553.694)	(553.216)	(553.933)
Major: Nat.	4922.184*	4936.866*	4953.571*
	(561.062)	(560.624)	(561.886)
Prof. Parents: Yes	720.335	729.645	739.751
	(502.776)	(502.427)	(503.119)
Parent Network: Yes	-496.95	-492.653	-489.233
	(510.984)	(510.574)	(511.021)
ln(Harvard SS)	.	17861.72*	.
		(2428.931)	
Harvard SS^2	.	.	-0.007
			(0.007)
N	528	528	528
RMSE	5244.394	5240.726	5244.2
R^2	0.208	0.209	0.21
adj R^2	0.199	0.2	0.199

$*p \leq 0.05$

```
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
H (30%) 20974.30     H
S (30%) 23269.99     S
N (30%) 25972.97     N

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

```
predictOMatic(cm2)
```

```
$major2
            fit major2
H (30%) 20974.30      H
S (30%) 23269.99      S
N (30%) 25972.97      N

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

Table 5: Categorical Regressions: Student-17

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	20974.299* (402.523)	23269.991* (410.697)	-2940.523 (2892.576)	-733.556 (2920.923)
Major: Soc.	2295.692* (575.062)	.	2206.966* (552.838)	.
Major: Nat.	4998.671* (580.497)	.	4585.221* (562.554)	.
Major 2: Hum.	.	-2295.692* (575.062)	.	-2206.966* (552.838)
Major 2: Nat.	.	2702.979* (586.195)	.	2378.255* (568.098)
SAT	.	.	11.143* (1.753)	11.143* (1.753)
ACT	.	.	215.203* (54.841)	215.203* (54.841)
Iowa BS	.	.	6.852 (25.559)	6.852 (25.559)
Prof. Parents: Yes	.	.	160.728 (507.436)	160.728 (507.436)
Parent Network: Yes	.	.	1803.058* (518.632)	1803.058* (518.632)
Gender: Male	.	.	691.894 (457.056)	691.894 (457.056)
N	586	586	528	528
RMSE	5735.07	5735.07	5232.223	5232.223
R^2	0.113	0.113	0.279	0.279
adj R^2	0.11	0.11	0.268	0.268

$*p \leq 0.05$