

Data Management

```
library(foreign)
library(rockchalk)
i <- 40
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%	7.68	1211.00	70.28	1775.00	3596.00	149800.00	1192.00
25%	18.92	1510.00	93.13	17100.00	19630.00	161900.00	1500.00
50%	22.37	1625.00	99.55	20580.00	23580.00	165700.00	1609.00
75%	25.37	1728.00	106.20	24060.00	27500.00	169500.00	1705.00
100%	36.72	2228.00	135.00	34890.00	40120.00	187600.00	2199.00
mean	22.22	1624.00	99.68	20570.00	23600.00	165800.00	1608.00
sd	4.87	165.90	9.58	5397.00	5833.00	5891.00	162.60
var	23.68	27520.00	91.76	29130000.00	34020000.00	34700000.00	26430.00
NA's	15.00	51.00	0.00	7.00	0.00	0.00	28.00
N	545.00	545.00	545.00	545.00	545.00	545.00	545.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

	gender		major		pnet
M	:292.0000	N	:189.0000	NO	:391.0000
F	:253.0000	S	:183.0000	YES	:154.0000
NA's	: 0.0000	H	:173.0000	NA's	: 0.0000
entropy	: 0.9963	NA's	: 0.0000	entropy	: 0.8589
normedEntropy	: 0.9963	entropy	: 1.5840	normedEntropy	: 0.8589
N	:545.0000	normedEntropy	: 0.9994	N	:545.0000
		N	:545.0000		
	pprof				
NO	:363.0000				
YES	:182.0000				
NA's	: 0.0000				
entropy	: 0.9189				
normedEntropy	: 0.9189				
N	:545.0000				

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 $\text{harv} = \text{sat} + \text{act}$).

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(sall ~ sat, data = dat)
m1a <- lm(sall ~ act, data = dat)
m1i <- lm(sall ~ ibs, data = dat)
m1h <- lm(sall ~ harv, data = dat)
m1all <- lm(sall ~ sat + act + ibs + harv, data = dat)
m1best <- lm(sall ~ sat + act + ibs, data = dat)
```

```
mcDiagnose(m1all)
```

The following auxiliary models are being estimated and returned in a list:

```
sat ~ act + ibs + harv
<environment: 0x1c5c6d8>
act ~ sat + ibs + harv
<environment: 0x1c5c6d8>
ibs ~ sat + act + harv
<environment: 0x1c5c6d8>
harv ~ sat + act + ibs
<environment: 0x1c5c6d8>
Drum roll please!
```

And your R_j Squareds are (auxiliary Rsq)

```
      sat      act      ibs      harv
0.9998521 0.8542723 0.2035374 0.9998553
The Corresponding VIF, 1/(1-Rj2)
      sat      act      ibs      harv
6761.856129  6.862115  1.255552 6912.748087
```

Bivariate Correlations for design matrix

```
      sat  act  ibs  harv
sat  1.00 0.36 0.36 1.00
act  0.36 1.00 0.37 0.38
ibs  0.36 0.37 1.00 0.37
harv 1.00 0.38 0.37 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 SS2", 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 sall: 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(sall ~ sat + act, data = m1alldf)
anova(m1restricted, m1all)
```

Analysis of Variance Table

```
Model 1: sall ~ sat + act
Model 2: sall ~ sat + act + ibs + harv
```

Table 2: Regression with sall: Student-40

	SAT	ACT	IBS	Harvard SS	All	Best
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)
(Intercept)	-253.285 (2191.009)	11757.916* (1015.106)	5990.875* (2363.159)	-1134.648 (2230.531)	-4713.942 (2793.658)	-4765.642 (2640.765)
SAT	12.976* (1.357)	.	.	.	-135.019 (113.079)	9.713* (1.443)
ACT	.	398.508* (44.625)	.	.	73.675 (122.831)	240.769* (48.593)
Iowa BS	.	.	146.267* (23.601)	.	36.064 (26.69)	44.328 (25.009)
Harvard SS	.	.	.	13.352* (1.367)	145.526 (113.239)	.
N	510	523	538	487	449	496
RMSE	4980.962	4975.044	5218.204	5005.029	4804.168	4737.268
R^2	0.153	0.133	0.067	0.164	0.214	0.218
adj R^2	0.151	0.131	0.065	0.163	0.207	0.213

* $p \leq 0.05$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	446	1.0336e+10				
2	444	1.0248e+10	2	88363815	1.9143	0.1487

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(sall ~ sat + act + ibs, data = dat)
dat2 <- model.frame(m1best)
m1s <- lm(sall ~ sat, data = dat2)
m1a <- lm(sall ~ act, data = dat2)
m1i <- lm(sall ~ ibs, data = dat2)
m1h <- lm(sall ~ harv, data = dat[row.names(dat2), ])
m1all <- lm(sall ~ 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	ACT	IBS	Harvard SS	All	Best
	Estimate	Estimate	Estimate	Estimate	Estimate	Estimate
	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)	(S.E.)
(Intercept)	-479.91 (2179.864)	12051.96* (1052.311)	5615.456* (2434.289)	-1273.617 (2276.488)	-4713.942 (2793.658)	-4765.642 (2640.765)
SAT	13.132* (1.349)	.	.	.	-135.019 (113.079)	9.713* (1.443)
ACT	.	385.54* (46.175)	.	.	73.675 (122.831)	240.769* (48.593)
Iowa BS	.	.	150.551* (24.289)	.	36.064 (26.69)	44.328 (25.009)
Harvard SS	.	.	.	13.44* (1.393)	145.526 (113.239)	.
N	496	496	496	449	449	496
RMSE	4895.569	5003.292	5148.238	4913.027	4804.168	4737.268
R^2	0.161	0.124	0.072	0.172	0.214	0.218
adj R^2	0.159	0.122	0.07	0.171	0.207	0.213

* $p \leq 0.05$

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(mlbest)
```

```

      sal1
sal1 -1.00000000
sat  0.29033455
act  0.21800581
ibs  0.07965494

```

```
getDeltaRsquare(mlbest)
```

```

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 0.072026420
act 0.039042141
ibs 0.004996203

```

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%     38.77    35.91    30.63
50%     50.79    45.35    41.38
75%     60.94    55.46    50.81
100%    100.00   100.00   100.00

```

```

mean  50.22  45.59  41.29
sd    16.77  14.73  16.21
var   281.30 216.90 262.80
NA's  0.00   0.00   0.00
N     496.00 496.00 496.00

```

```

$ factors
NULL

```

```

mlpoms <- lm(sall ~ satpoms + actpoms + ibspoms, data = dat2)
summary(mlpoms)

```

```

Call:
lm(formula = sall ~ satpoms + actpoms + ibspoms, data = dat2)

```

```

Residuals:
    Min       1Q   Median       3Q      Max
-14020.9  -3226.0   226.4   3331.6  12655.9

```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 11780.87    842.97  13.975 < 2e-16 ***
satpoms       97.76     14.53   6.730 4.74e-11 ***
actpoms       69.92     14.11   4.955 9.98e-07 ***
ibspoms       28.67     16.18   1.772  0.0769 .

```

```

Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

```

Residual standard error: 4737 on 492 degrees of freedom
Multiple R-squared: 0.2176, Adjusted R-squared: 0.2128
F-statistic: 45.6 on 3 and 492 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(mlall)

```

```

           sall
sall -1.00000000
sat  -0.05657499
act   0.02845405
ibs   0.06399324
harv  0.06087603

```

```

getDeltaRsquare(mlall)

```

```

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  0.0025239357
act  0.0006369083
ibs  0.0032321191
harv 0.0029237625

```

```

options(scipen = 5)

```

Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-40

	Test Scores Only	All Predictors
	Estimate	Estimate
	(S.E.)	(S.E.)
(Intercept)	-697.423 (2909.734)	-5081.244 (2685.737)
SAT	9.503* (1.597)	9.847* (1.442)
ACT	240.994* (53.87)	241.716* (48.632)
Iowa BS	37.203 (27.489)	44.672 (24.865)
Major: Soc.	.	1984.498* (526.485)
Major: Nat.	.	5098.597* (522.13)
Prof. Parents: Yes	.	1780.787* (451.948)
Parent Network: Yes	.	582.014 (473.57)
Gender: Male	.	-236.365 (424.633)
N	503	503
RMSE	5265.227	4743.673
R^2	0.176	0.338
adj R^2	0.171	0.327

* $p \leq 0.05$

```
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 = 1198.77317075194 Denominator = 642.982470237977"
```

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

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

```
tval
```

```
difference
1.864395
```

```
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
0.06285914
```

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
1198.77317075	642.98247024	1.86439479	494.00000000	0.06285914

```
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      499 13833582608
2      494 11116203437  5 2717379170 24.152 < 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

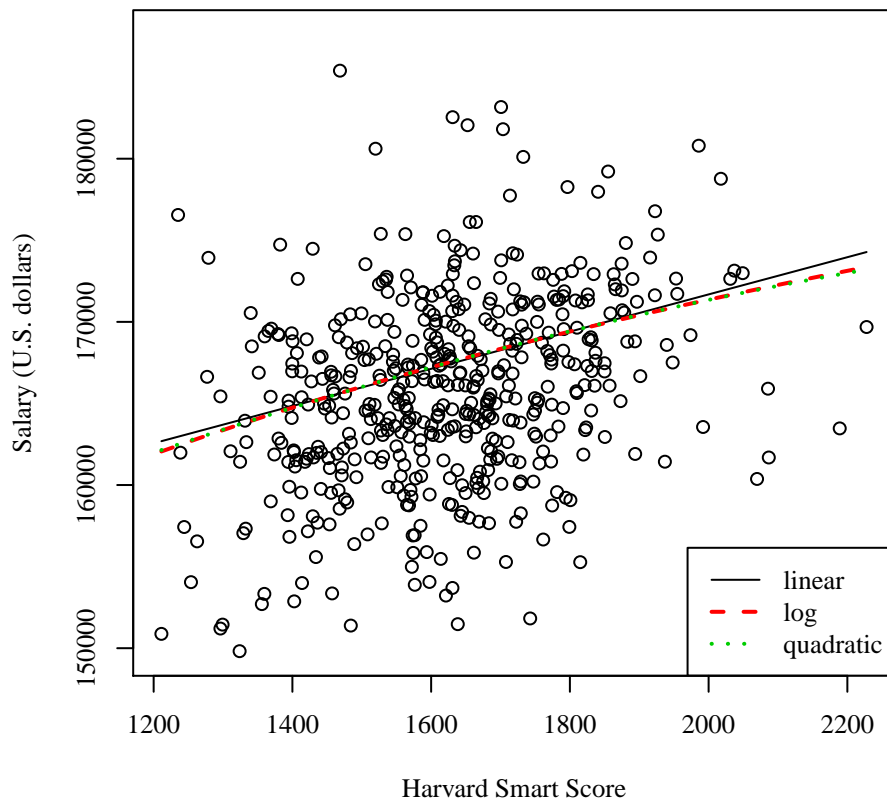
Table 4: Regression with sal3: Student-40

	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	144418.156* (2344.891)	25905.052 (16734.881)	135134.04* (15540.021)
Harvard SS	11.404* (1.394)	.	22.807 (18.919)
Gender: Male	-420.864 (461.442)	-400.012 (461.448)	-398.553 (463.217)
Major: Soc.	2200.762* (571.983)	2196.323* (571.803)	2195.274* (572.429)
Major: Nat.	4877.868* (566.282)	4869.734* (566.05)	4869.636* (566.815)
Prof. Parents: Yes	901.915 (491.196)	892.146 (490.973)	891.561 (491.815)
Parent Network: Yes	1072.681* (507.982)	1082.617* (507.799)	1082.396* (508.567)
ln(Harvard SS)	.	18548.574* (2260.707)	.
Harvard SS ²	.	.	-0.003 (0.006)
N	494	494	494
RMSE	5106.556	5104.855	5109.887
R^2	0.231	0.232	0.232
adj R^2	0.222	0.222	0.221

* $p \leq 0.05$

```
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%) 26175.40 N
S (30%) 23447.84 S
H (30%) 20958.73 H

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

```
predictOMatic(cm2)
```

```
$major2
      fit major2
N (30%) 26175.40 N
S (30%) 23447.84 S
H (30%) 20958.73 H

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

Table 5: Categorical Regressions: Student-40

	major	major2	major full	major2 full
	Estimate	Estimate	Estimate	Estimate
	(S.E.)	(S.E.)	(S.E.)	(S.E.)
(Intercept)	20958.727*	23447.839*	-5081.244	-3096.747
	(413.64)	(402.179)	(2685.737)	(2707.939)
Major: Soc.	2489.112*	.	1984.498*	.
	(576.928)		(526.485)	
Major: Nat.	5216.672*	.	5098.597*	.
	(572.461)		(522.13)	
Major 2: Hum.	.	-2489.112*	.	-1984.498*
		(576.928)		(526.485)
Major 2: Nat.	.	2727.56*	.	3114.1*
		(564.236)		(512.408)
SAT	.	.	9.847*	9.847*
			(1.442)	(1.442)
ACT	.	.	241.716*	241.716*
			(48.632)	(48.632)
Iowa BS	.	.	44.672	44.672
			(24.865)	(24.865)
Prof. Parents: Yes	.	.	1780.787*	1780.787*
			(451.948)	(451.948)
Parent Network: Yes	.	.	582.014	582.014
			(473.57)	(473.57)
Gender: Male	.	.	-236.365	-236.365
			(424.633)	(424.633)
N	545	545	503	503
RMSE	5440.583	5440.583	4743.673	4743.673
R^2	0.133	0.133	0.338	0.338
adj R^2	0.13	0.13	0.327	0.327

* $p \leq 0.05$