

## Data Management

```
library(foreign)
library(rockchalk)
i <- 46
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%	6.78	1105.00	72.13	3412.00	3319.00	150100.00	1087.00
25%	18.14	1512.00	92.80	17320.00	19960.00	161900.00	1494.00
50%	22.20	1632.00	99.68	20710.00	23520.00	165700.00	1607.00
75%	25.32	1729.00	106.80	24240.00	27470.00	169600.00	1702.00
100%	36.90	2114.00	133.80	37210.00	40850.00	181500.00	2089.00
mean	21.86	1620.00	99.94	20830.00	23670.00	165700.00	1597.00
sd	5.13	166.20	10.27	5300.00	5777.00	5796.00	163.00
var	26.37	27620.00	105.40	28090000.00	33380000.00	33600000.00	26570.00
NA's	16.00	64.00	0.00	6.00	0.00	0.00	22.00
N	551.00	551.00	551.00	551.00	551.00	551.00	551.00

Table 1: Best Automatic Summary Table for Numerics

Let students figure way to beautify this:

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

<b>gender</b>		<b>major</b>		<b>pnet</b>	
M	:283.0000	H	:190.0000	NO	:411.0000
F	:268.0000	S	:188.0000	YES	:140.0000
NA's	: 0.0000	N	:173.0000	NA's	: 0.0000
entropy	: 0.9995	NA's	: 0.0000	entropy	: 0.8177
normedEntropy	: 0.9995	entropy	: 1.5837	normedEntropy	: 0.8177
N	:551.0000	normedEntropy	: 0.9992	N	:551.0000
		N	:551.0000		
<b>pprof</b>					
NO	:366.0000				
YES	:185.0000				
NA's	: 0.0000				
entropy	: 0.9207				
normedEntropy	: 0.9207				
N	:551.0000				

## 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: 0x25ccde0>
act ~ sat + ibs + harv
<environment: 0x25ccde0>
ibs ~ sat + act + harv
<environment: 0x25ccde0>
harv ~ sat + act + ibs
<environment: 0x25ccde0>
Drum roll please!
```

And your R<sub>j</sub> Squareds are (auxiliary Rsq)

```
      sat      act      ibs      harv
0.9998428 0.8555866 0.2427534 0.9998461
The Corresponding VIF, 1/(1-Rj2)
      sat      act      ibs      harv
6361.368359  6.924563  1.320574 6495.971146
```

Bivariate Correlations for design matrix

```
      sat  act  ibs  harv
sat  1.00 0.32 0.40 1.00
act  0.32 1.00 0.39 0.35
ibs  0.40 0.39 1.00 0.41
harv 1.00 0.35 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 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-46

	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)	2421.29 (2101.21)	15442.99* (990.469)	11803.934* (2198.637)	3003.027 (2233.748)	521.62 (2688.048)	832.356 (2535.19)
SAT	11.551* (1.309)	.	.	.	20.336 (110.944)	9.961* (1.475)
ACT	.	244.485* (44.126)	.	.	118.503 (118.021)	122.661* (47.028)
Iowa BS	.	.	90.434* (21.91)	.	21.783 (25.87)	14.209 (24.156)
Harvard SS	.	.	.	11.026* (1.372)	-10.427 (110.899)	.
N	524	529	545	481	448	508
RMSE	4868.99	5191.984	5223.525	4981.279	4861.055	4864.49
$R^2$	0.13	0.055	0.03	0.119	0.146	0.147
adj $R^2$	0.128	0.053	0.029	0.117	0.139	0.142

\* $p \leq 0.05$

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	445	1.0485e+10				
2	443	1.0468e+10	2	16766122	0.3548	0.7015

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)	2257.418 (2131.732)	15498.987* (992.323)	10653.576* (2235.368)	2287.335 (2244.685)	521.62 (2688.048)	832.356 (2535.19)
SAT	11.639* (1.33)	.	.	.	20.336 (110.944)	9.961* (1.475)
ACT	.	242.968* (44.117)	.	.	118.503 (118.021)	122.661* (47.028)
Iowa BS	.	.	102.13* (22.339)	.	21.783 (25.87)	14.209 (24.156)
Harvard SS	.	.	.	11.494* (1.381)	-10.427 (110.899)	.
N	508	508	508	448	448	508
RMSE	4897.67	5104.646	5150.117	4877.942	4861.055	4864.49
$R^2$	0.132	0.057	0.04	0.134	0.146	0.147
adj $R^2$	0.13	0.055	0.038	0.133	0.139	0.142

\* $p \leq 0.05$

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  0.28803561
act  0.11540565
ibs  0.02619361

```

```
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 0.0772053886
act 0.0115190999
ibs 0.0005859089

```

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%     37.77    33.10    40.12
50%     51.69    44.14    51.53
75%     61.84    55.12    61.33
100%    100.00   100.00   100.00

```

```

mean   50.20   44.41   50.65
sd     17.06   16.59   16.33
var    291.10  275.20  266.60
NA's   0.00    0.00    0.00
N      508.00  508.00  508.00

```

```

$ 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
-13697.3  -3171.9   -80.6   3315.7  14884.9

```

```

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13520.21    854.25  15.827 < 2e-16 ***
satpoms      99.80     14.78   6.753 4.01e-11 ***
actpoms      36.95     14.16   2.608 0.00937 **
ibspoms       8.77     14.91   0.588 0.55663

```

```

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

```

```

Residual standard error: 4864 on 504 degrees of freedom
Multiple R2: 0.1466, Adjusted R2: 0.1415
F-statistic: 28.86 on 3 and 504 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 -1.000000000
sat   0.008708575
act   0.047651313
ibs   0.039973184
harv -0.004466966

```

```

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 0.00006475343
act 0.00194299921
ibs 0.00136636777
harv 0.00001703608

```

```

options(scipen = 5)

```

## Additional Variables

Please see Table 3 for the regressions.

Table 3: Regression with sal2: Student-46

	Test Scores Only	All Predictors
	Estimate	Estimate
	(S.E.)	(S.E.)
(Intercept)	3407.459 (2785.846)	1035.534 (2572.905)
SAT	9.776* (1.632)	10.159* (1.484)
ACT	156.015* (51.83)	141.316* (47.167)
Iowa BS	12.748 (26.653)	10.376 (24.295)
Major: Soc.	.	1736.645* (525.295)
Major: Nat.	.	5379.954* (531.192)
Prof. Parents: Yes	.	436.459 (459.573)
Parent Network: Yes	.	896.266 (496.376)
Gender: Male	.	-717.527 (432.17)
N	513	513
RMSE	5392.509	4881.671
$R^2$	0.132	0.296
adj $R^2$	0.127	0.284

\* $p \leq 0.05$ 

```
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 = -459.807019146227 Denominator = 682.097554251401"
```

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

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

```
tval
```

```
difference
-0.6741074
```

```
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
0.5005521
```

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
-459.8070191	682.0975543	-0.6741074	504.0000000	0.5005521

```
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     509 14801287103
2     504 12010680234  5 2790606869 23.42 < 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-46

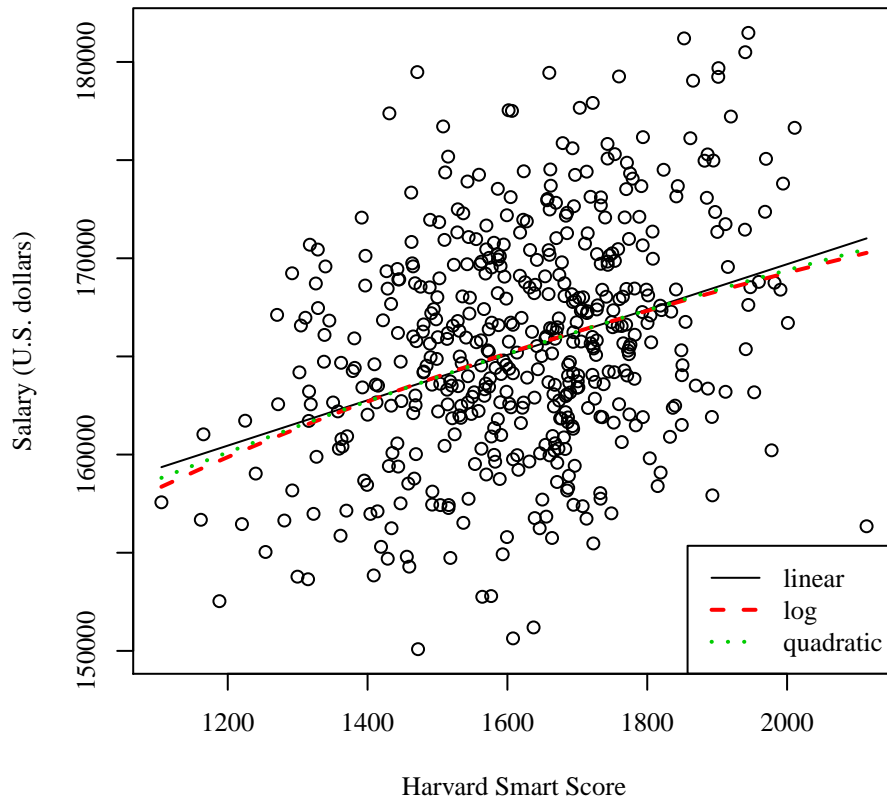
	Linear Estimate (S.E.)	Log Estimate (S.E.)	Quadratic Estimate (S.E.)
(Intercept)	143662.156* (2201.792)	26691.986 (15562.257)	137660.001* (14920.748)
Harvard SS	11.553* (1.327)	.	19.1 (18.603)
Gender: Male	669.056 (440.331)	666.428 (440.222)	667.579 (440.729)
Major: Soc.	2259.652* (527.955)	2231.063* (528.025)	2241.468* (530.302)
Major: Nat.	5778.419* (548.338)	5773.34* (548.216)	5775.138* (548.875)
Prof. Parents: Yes	1037.634* (468.945)	1053.242* (468.947)	1048.742* (470.147)
Parent Network: Yes	168.93 (499.209)	153.533 (499.151)	159.814 (500.146)
ln(Harvard SS)	.	18375.409* (2106.749)	.
Harvard SS <sup>2</sup>	.	.	-0.002 (0.006)
N	487	487	487
RMSE	4839.943	4838.81	4844.156
$R^2$	0.292	0.292	0.292
adj $R^2$	0.283	0.283	0.281

\* $p \leq 0.05$ 

```
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%) 21474.06  H
S (30%) 23104.11  S
N (30%) 26697.29  N

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

```
predictOMatic(cm2)
```

```
$major2
      fit major2
H (30%) 21474.06  H
S (30%) 23104.11  S
N (30%) 26697.29  N

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

Table 5: Categorical Regressions: Student-46

	major Estimate (S.E.)	major2 Estimate (S.E.)	major full Estimate (S.E.)	major2 full Estimate (S.E.)
(Intercept)	21474.057* (389.485)	23104.113* (391.551)	1035.534 (2572.905)	2772.179 (2569.394)
Major: Soc.	1630.057* (552.277)	.	1736.645* (525.295)	.
Major: Nat.	5223.236* (564.184)	.	5379.954* (531.192)	.
Major 2: Hum.	.	-1630.057* (552.277)	.	-1736.645* (525.295)
Major 2: Nat.	.	3593.179* (565.612)	.	3643.309* (534.291)
SAT	.	.	10.159* (1.484)	10.159* (1.484)
ACT	.	.	141.316* (47.167)	141.316* (47.167)
Iowa BS	.	.	10.376 (24.295)	10.376 (24.295)
Prof. Parents: Yes	.	.	436.459 (459.573)	436.459 (459.573)
Parent Network: Yes	.	.	896.266 (496.376)	896.266 (496.376)
Gender: Male	.	.	-717.527 (432.17)	-717.527 (432.17)
N	551	551	513	513
RMSE	5368.676	5368.676	4881.671	4881.671
$R^2$	0.14	0.14	0.296	0.296
adj $R^2$	0.136	0.136	0.284	0.284

\* $p \leq 0.05$