HW 11 Stat 139

Callin Switzer

November 25, 2014

#1

(#10(a-d))

Here are the formulas I used for calculations.

$\hat{\sigma}^2$ = Residual SS / d.f.

$\hat{\sigma}$ is the "Residual standard error" from the lm() function

AdjR^2 = 100*((total mean square) - (residual mean square))/total mean square

Cp = p + (n-p)$\frac{\hat{\sigma^2} - \hat{\sigma^2_{full}}}{\hat{\sigma^2_{full}}}$

BIC = n log(SSRes / n) + p log(n)

Note: R calculates BIC with this formula: $BIC = n + n\dot{l}og(2\pi) + nlog(RSS/n) + log(n)(p + 1)$

	variables	ResidSS	df	sigmaSq	AdjRSq	Cp	BIC
1	None	8100.00	27.00	300.00	0.00	1.00	162.02
2	A	6240.00	26.00	240.00	0.20	-3.20	158.05
3	B	5980.00	26.00	230.00	0.23	-4.07	156.86
4	C	6760.00	26.00	260.00	0.13	-1.47	160.29
5	AB	5500.00	25.00	220.00	0.27	-3.67	157.84
6	AC	5250.00	25.00	210.00	0.30	-4.50	156.54
7	BC	5750.00	25.00	230.00	0.23	-2.83	159.09
8	ABC	5160.00	24.00	215.00	0.28	-2.80	159.39

Table 1: Parts A-D

(#10e, i-iv)

	BestCombo
max R^2	AC
min sigmaSq	AC
min Cp	AC
min BIC	AC

Table 2: 1e, part i-iv

#11 (still part of #1)

The F-statistic will be $\frac{ExtraSumOfSquares/ExtraDegreesOfFreedom}{\hat{\sigma}^2_{full}}$

The Extra sum fo squares is = Residual Sum of squares (reduced) - Residual sum of Squares (full)

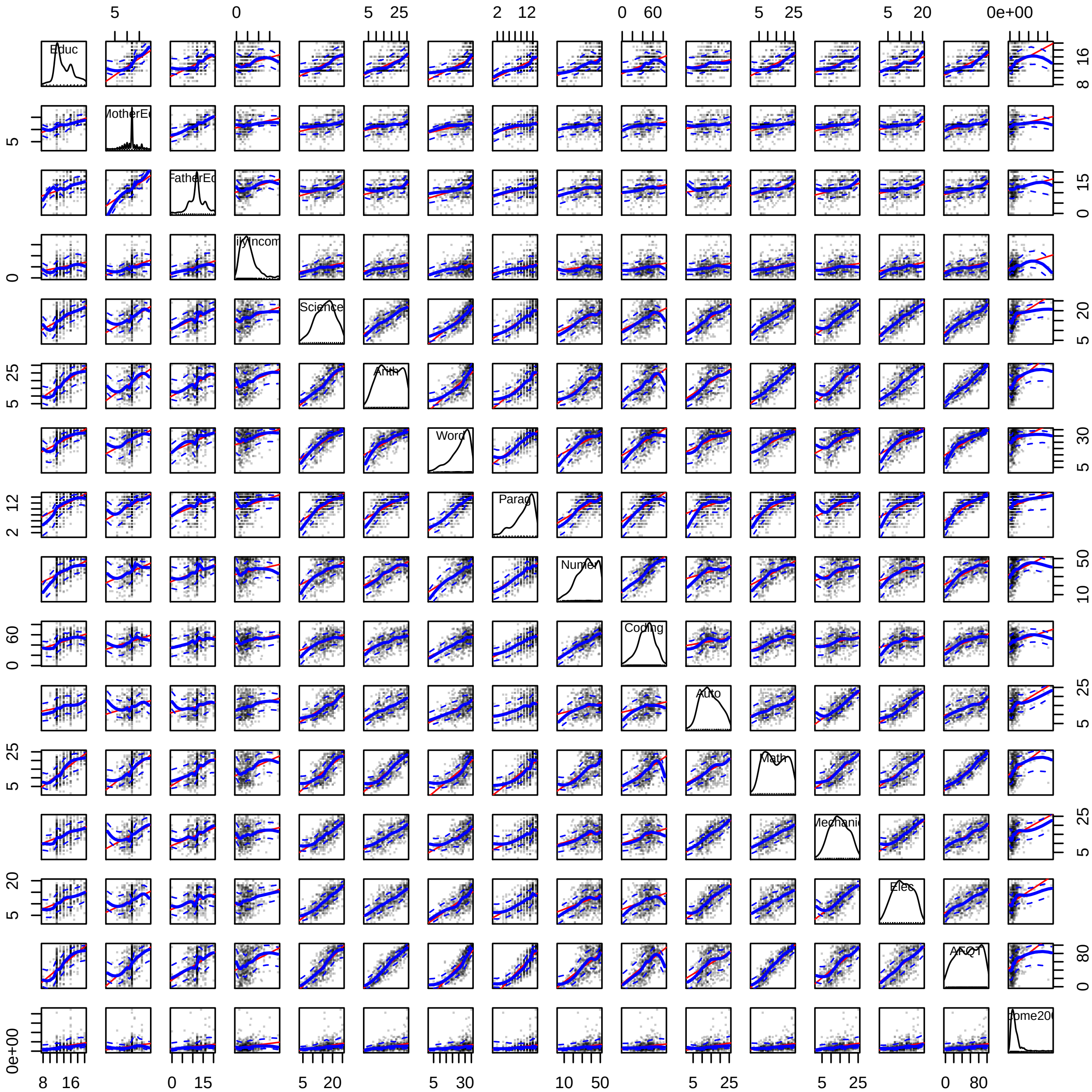
The single-variable model with the smallest residual sum of squares is B.

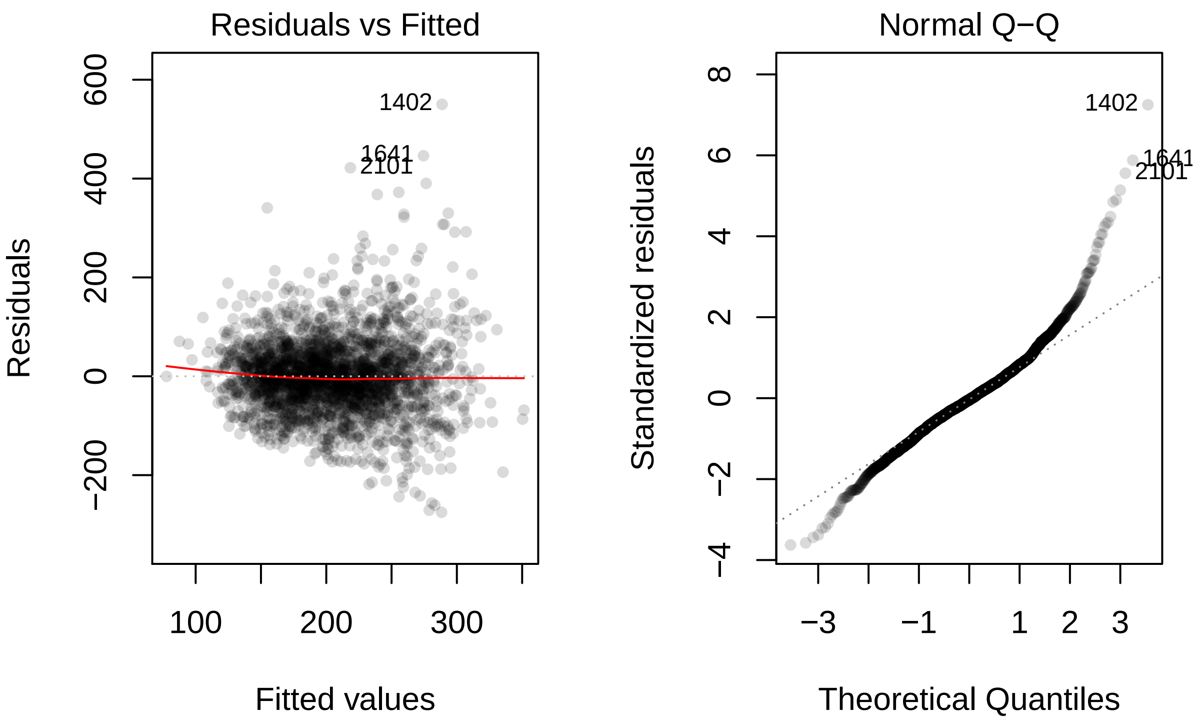
An F-statistic will be ((8100 - 5980)/1)/300 = 7.0666667. We'll compare this to an F-distribution with numerator df =1 , and denominator df = 27. The p-value is 0.0130366.

Now we move on to the second step. AB is the two-variable model that includes B and has the lowest Residual Sum of Squares. When we do an extra sum of squares F-test, the test statistic is ((5980 - 5500)/1)/230 = 2.0869565. We'll use an F-distribution with df1 = 1, and df2 = 26. the p-value is 0.1605063. The F-statistic is also under 4, so we'll just keep the original model, B. Notably, this is not the model with the lowest BIC that we found in the earlier part.

2a

To decide if variables need to be transfomed, I looked at the scatterplot matrix of all the continuous variables. I decided that Income2005 should be log-transformed. Some predictors that may need log transfomed included word, parag, and numer. After considering them for transformation, I decided not to transform them. I decided to sqrt transform Income2005 after looking at the normal Q-Q plot for the model when Income2005 was sqrt transformed (shown below).


Educ
MotherEd
FatherEd
Science
Arith
Word
Parag
Numer
Coding
Auto
Math
Mechanic
Elec
AFQT



#2b

I first made the model with interaction terms, and I used sqrt transformed Income2005 (because that's what I decided on in part 2a). I looked at the residual and normal Q-Q plots, and they looked somewhat troubling – the residuals looked a bit fanned, and the Q-Q plot looked long-tailed (I didn't picture them below).

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	11.5905	31.0427	0.37	0.7089
Educ	13.5203	2.3575	5.74	0.0000
Race2	-19.3562	45.6561	-0.42	0.6716
Race3	26.5312	32.5854	0.81	0.4156
Educ:Race2	0.5063	3.4078	0.15	0.8819
Educ:Race3	-1.3721	2.4580	-0.56	0.5767

Table 3: Model for 2b with interaction terms

The slope of the line when Race = 2 will be 14.0265881, which is the sum of the coefficients, Educ and Educ:Race2 from the table above.

The slope of the line when Race = 3 will be 12.1481679, which is the sum of the coefficients, Educ and Educ:Race3 from the table above.

Below, I redefine the reference level to be Race = 2 and rerun the regression so that R is automatically performing a t-test to compare the two slopes.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.7657	33.4789	-0.23	0.8166
Educ	14.0266	2.4608	5.70	0.0000
Race1	19.3562	45.6561	0.42	0.6716
Race3	45.8873	34.9141	1.31	0.1889
Educ:Race1	-0.5063	3.4078	-0.15	0.8819
Educ:Race3	-1.8784	2.5573	-0.73	0.4627

Table 4: Model for 2b with interaction terms, and new reference level

The table above shows that Race = 3 is not significantly different form Race = 2. The p-value is 0.4626858

#2c

Below is a table that shows the results for the significant terms after backward selection, based on AIC.

```
##                             Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 -10.8204    52.0596 -0.2078   0.8354
## Imagazine                    14.9559    14.9207  1.0024   0.3163
## Inewspaper                    5.9357     7.6868  0.7722   0.4401
## Ilibrary                    -41.4723    24.7382 -1.6764   0.0938
## MotherEd                     11.6975     4.4812  2.6103   0.0091
## FatherEd                     -4.3606     3.0591 -1.4254   0.1542
## FamilyIncome78                0.0024     0.0006  3.8609   0.0001
## Race1                         7.1628    18.1049  0.3956   0.6924
## Race3                         3.0406    13.9588  0.2178   0.8276
## Gendermale                   14.9002    15.7224  0.9477   0.3434
## Educ                          3.7413     4.2767  0.8748   0.3818
## Science                       1.1543     1.8754  0.6155   0.5383
## Arith                        -1.2674     2.3106 -0.5485   0.5834
## Word                         -1.6547     1.6021 -1.0328   0.3018
## Parag                         5.2141     5.0609  1.0303   0.3030
## Numer                        -0.2336     1.2022 -0.1943   0.8459
## Coding                        0.7969     0.4933  1.6153   0.1064
## Auto                          1.2727     1.8114  0.7026   0.4823
## Math                         -3.2072     3.0737 -1.0434   0.2968
## Mechanic                      0.2425     1.8102  0.1340   0.8934
## Elec                          0.5958     3.4434  0.1730   0.8626
## AFQT                          0.8860     0.9593  0.9235   0.3558
## Imagazine:FatherEd           -1.9911     1.0868 -1.8320   0.0671
## Imagazine:Gendermale         30.7395     7.3800  4.1653   0.0000
## Imagazine:Science             2.5073     1.0878  2.3050   0.0212
## Imagazine:Elec               -3.7520     1.3013 -2.8834   0.0040
## Inewspaper:FamilyIncome78    -0.0009     0.0004 -2.2311   0.0258
## Ilibrary:Gendermale         -18.9144     8.4533 -2.2375   0.0253
## Ilibrary:Educ                 2.6862     1.9164  1.4017   0.1611
## Ilibrary:Numer                0.9074     0.5216  1.7396   0.0820
## Ilibrary:Coding              -0.5594     0.3493 -1.6014   0.1094
## Ilibrary:Math                -2.2439     0.9166 -2.4481   0.0144
## Ilibrary:Mechanic             3.3265     0.9341  3.5611   0.0004
```

```
## MotherEd:Educ              -0.6981     0.3522 -1.9824   0.0475
## MotherEd:Numer             -0.1310     0.0753 -1.7384   0.0823
## MotherEd:Math               0.4006     0.1662  2.4103   0.0160
## MotherEd:Elec              -0.3091     0.2109 -1.4657   0.1429
## FatherEd:Educ               0.3544     0.2440  1.4524   0.1465
## FatherEd:Arith             -0.1910     0.1000 -1.9101   0.0562
## FatherEd:Elec               0.4283     0.1685  2.5409   0.0111
## FamilyIncome78:Gendermale   0.0007     0.0002  2.9850   0.0029
## FamilyIncome78:Science     -0.0001     0.0000 -2.6552   0.0080
## FamilyIncome78:Coding       0.0000     0.0000 -1.8473   0.0648
## FamilyIncome78:AFQT         0.0000     0.0000  2.8838   0.0040
## Race1:Gendermale          -12.7095    15.1613 -0.8383   0.4020
## Race3:Gendermale           13.6455    11.0106  1.2393   0.2153
## Race1:Math                  2.6416     2.8351  0.9318   0.3515
## Race3:Math                 -2.5409     2.0325 -1.2501   0.2114
## Race1:AFQT                 -0.7564     0.6262 -1.2080   0.2272
## Race3:AFQT                  0.3059     0.4560  0.6708   0.5024
## Gendermale:Arith            2.4222     0.9270  2.6129   0.0090
## Gendermale:Word            -1.2653     0.7200 -1.7573   0.0790
## Gendermale:Auto             2.5253     1.0098  2.5009   0.0125
## Gendermale:Math            -2.5094     0.9351 -2.6837   0.0073
## Educ:Parag                 -0.6117     0.3542 -1.7270   0.0843
## Educ:Numer                  0.2209     0.0926  2.3870   0.0171
## Educ:Elec                   0.5072     0.2100  2.4150   0.0158
## Science:Coding             -0.1079     0.0366 -2.9493   0.0032
## Science:Math                0.2200     0.0926  2.3761   0.0176
## Arith:Word                  0.2049     0.0993  2.0631   0.0392
## Arith:Auto                 -0.2797     0.1190 -2.3517   0.0188
## Arith:Elec                  0.3050     0.1207  2.5277   0.0115
## Arith:AFQT                 -0.0412     0.0205 -2.0104   0.0445
## Word:Parag                  0.3068     0.1787  1.7171   0.0861
## Word:AFQT                  -0.1164     0.0283 -4.1167   0.0000
## Parag:Mechanic             -0.3246     0.1516 -2.1410   0.0324
## Parag:Elec                 -0.6488     0.2380 -2.7264   0.0064
## Parag:AFQT                  0.1357     0.0492  2.7562   0.0059
## Numer:Math                 -0.0950     0.0402 -2.3627   0.0182
## Coding:Auto                 0.0646     0.0287  2.2504   0.0245
## Coding:Math                 0.0761     0.0272  2.8010   0.0051
## Auto:Elec                  -0.2856     0.1132 -2.5222   0.0117
## Auto:AFQT                   0.0826     0.0299  2.7665   0.0057

## [1] "The AIC from this model is 22156.96"
```

#2d

Here are the forward stepwise results:

```
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 33.8430    28.1260  1.2033   0.2290
## Educ                         4.4871     2.4282  1.8479   0.0647
## Gendermale                  12.8758     8.8363  1.4571   0.1452
## Numer                        2.0340     0.5100  3.9883   0.0001
## FamilyIncome78               0.0010     0.0004  2.7287   0.0064
## Arith                       -0.2130     0.5330 -0.3996   0.6895
## Auto                         1.3796     0.4261  3.2377   0.0012
## Science                     -4.3299     1.7274 -2.5067   0.0122
## Ilibrary                    -8.5391    11.7823 -0.7247   0.4687
## Inewspaper                   5.9959     7.2520  0.8268   0.4084
## Math                         1.4882     0.6054  2.4583   0.0140
## Imagazine                    7.0217    12.6693  0.5542   0.5795
## Gendermale:FamilyIncome78    0.0007     0.0002  2.9773   0.0029
## Gendermale:Arith             1.8861     0.7180  2.6268   0.0087
## FamilyIncome78:Inewspaper   -0.0010     0.0004 -2.5111   0.0121
## Gendermale:Imagazine        23.9329     6.9335  3.4518   0.0006
## Gendermale:Math             -1.5357     0.8170 -1.8797   0.0603
## Science:Ilibrary             0.9845     0.7270  1.3542   0.1758
## Numer:Imagazine             -0.3888     0.3320 -1.1710   0.2417
## Educ:Science                 0.2712     0.1352  2.0053   0.0450
## Numer:Science               -0.0538     0.0328 -1.6400   0.1011
```

```
## [1] "The AIC for the forward-selected model is 22172.97"
```

#2e

Here are the stepwise results:

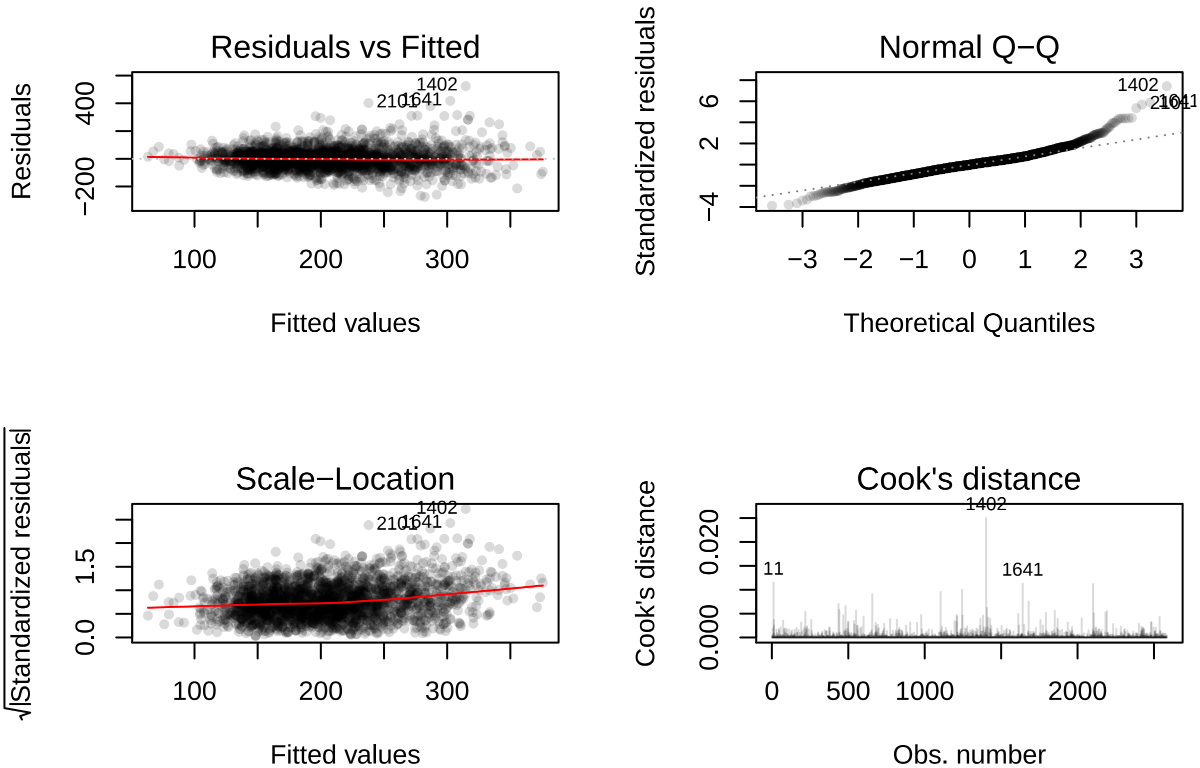
```
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   16.3611    29.8689  0.5478   0.5839
## Imagazine                     16.0691    14.4408  1.1128   0.2659
## Inewspaper                     5.3356     7.6027  0.7018   0.4829
## Ilibrary                     -15.4788    10.3484 -1.4958   0.1348
## MotherEd                       2.4036     2.0683  1.1621   0.2453
## FatherEd                      -0.8830     1.4353 -0.6152   0.5385
## FamilyIncome78                 0.0013     0.0005  2.4141   0.0158
## Gendermale                    20.3139    10.0116  2.0290   0.0426
## Educ                           4.1751     2.0836  2.0038   0.0452
## Science                       -2.7876     0.9286 -3.0018   0.0027
## Arith                         -0.4411     0.5889 -0.7490   0.4539
## Word                           1.1008     0.7307  1.5065   0.1321
## Numer                          1.7605     0.6765  2.6022   0.0093
## Coding                         0.4801     0.2103  2.2833   0.0225
## Auto                           1.6804     0.4870  3.4505   0.0006
## Math                           1.1918     0.6732  1.7702   0.0768
## Mechanic                      -0.8758     1.5373 -0.5697   0.5689
## Elec                          -2.7202     2.3683 -1.1486   0.2508
## AFQT                           0.2758     0.1793  1.5385   0.1241
## Imagazine:Gendermale          31.7736     7.2503  4.3824   0.0000
## FamilyIncome78:Gendermale      0.0006     0.0002  2.5313   0.0114
## Inewspaper:FamilyIncome78     -0.0009     0.0004 -2.1934   0.0284
## Ilibrary:Mechanic              2.0816     0.7434  2.8001   0.0051
## Word:Mechanic                 -0.1579     0.0502 -3.1445   0.0017
## Gendermale:Arith               2.0998     0.7178  2.9254   0.0035
## Educ:Elec                      0.4196     0.1600  2.6229   0.0088
## Imagazine:Elec                -3.4571     1.2267 -2.8182   0.0049
## Gendermale:Math               -1.7287     0.8247 -2.0962   0.0362
## Ilibrary:Gendermale          -14.9685     7.5485 -1.9830   0.0475
## MotherEd:Numer                -0.0929     0.0560 -1.6570   0.0976
## FatherEd:Mechanic              0.2195     0.0942  2.3292   0.0199
## Imagazine:FatherEd            -2.4213     1.0527 -2.3000   0.0215
## Imagazine:Science              1.9892     1.0254  1.9399   0.0525
## FamilyIncome78:Coding          0.0000     0.0000 -1.6764   0.0938
## Imagazine:FamilyIncome78       0.0005     0.0003  1.5309   0.1259
```

```
## [1] "The AIC for the forward-selected model is 22160.84"
```

#2f

I choose the model from the backward variable selection procedure. It has the lowest AIC. Below is a model check. Here are the assumptions of multiple regression: 1. Linearity 2. Constant variance along the line(s) 3. Normality of each subpopulation of responses 4. Independence:Location in relation to the mean cannot be predicted with knowledge of other responses 5. Random sample

The data do not fit the assumptions of normality and equal variance. I can tell because the residual plot shows a bit of fanning, and the scale-location plot shows an increase in residuals. The Normal Q-Q plot has points that don't line up along the line very well. Also, I'm not sure if a random sample was used.



#2g

Below are the 95% confidence and prediction intervals for the sqrt of income. Since I was using sqrt of income2005 as a predictor, I squared the interval to get the correct interval for actual dollars.

```
## [1] "Here is the 95% confidence interval for the mean"

##        fit      lwr      upr
## 1 49052.48 45440.31 52802.77

## [1] "Here is the 95% prediction interval"

##        fit     lwr      upr
## 1 49052.48 6481.13 131370.3
```

#3a

I did an F-test on two models. The full model contained all four components and the AFQT, and the reduced model contained only the AFQT (See handwritten paper for calculations). I also checked my work with the anova() function. Based on my results, I reject the null hypothesis that the reduced model with only AFQT is the best model.

[1] "Sigma from the full model is 81.1261301354893 with df = 2578" [1] "Sigma from the reduced model is 82.2191536738501 with df = 2582" [1] "Here is the anova table to compare with my calculations:"

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2582	17454292.19				
2	2578	16966975.50	4	487316.70	18.51	0.0000

#3b

I used an F-test to compare the two models – the full model contained the four components and the AFQT, and the reduced model contained only the four components. I didn't need to conduct an F-test in this case, because I found in part 3a that the four components were useful predictors (more useful that AFQT alone). Below is the results from my model comparison. I fail to reject the null hypothesis that the reduced model is sufficient.

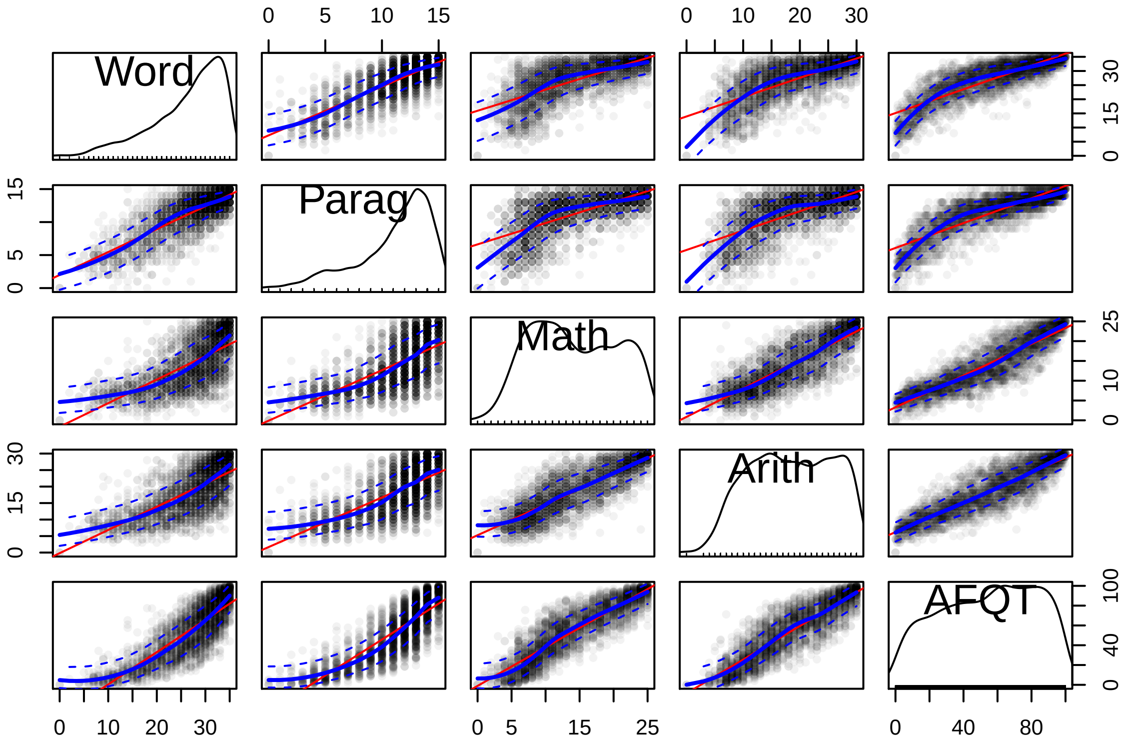
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	2579	16970202.99				
2	2578	16966975.50	1	3227.49	0.49	0.4838

#3c

Below, I've printed tables for the regressions and a scatterplot matrix to show that the components are all correlated with AFQT. The scatterplot matrix shows why the SE and estimated slope of AFQT differ so much when the four components are added: the four components are all correlated, and they're correlated with AFQT quite strongly. When explanatory variables are multicollinear, then the estimates and SEs can differ drastically when you include or exclude those variables.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	131.8163	10.1652	12.97	0.0000
Word	0.0098	0.4278	0.02	0.9817
Parag	-1.6593	0.9900	-1.68	0.0939
Math	2.1178	0.5942	3.56	0.0004
Arith	2.8129	0.4807	5.85	0.0000
AFQT	0.1612	0.2302	0.70	0.4838

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	146.2909	3.5608	41.08	0.0000
AFQT	1.0677	0.0583	18.32	0.0000



#3d

Here is the regression:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-33.9380	0.5562	-61.02	0.0000
Word	0.7641	0.0333	22.91	0.0000
Parag	2.3086	0.0714	32.31	0.0000
Math	1.6357	0.0393	41.61	0.0000
Arith	1.0262	0.0358	28.66	0.0000

The R^2 is 0.938, and the tolerance is 0.062. The vif is 1/tolerance, which is 16.032 for AFQT

3e

$Var(\hat{\beta}_j^{full}) = 0.0530042$

$Var(\hat{\beta}_j^{red}) = 0.0033958$

$MSR_{X_j} = 0.9735887$

$VIF_{X_j} = 16.0323005$

$Var(\hat{\beta}_j^{red}) * MSR_{X_j} * VIF_{X_j} = 0.0530042$ which is the same as $Var(\hat{\beta}_j^{full})$

#3f

Yes, we can approximate the relationship between AFQT and the components.

#3fi

The SSOM has 15 parameters (not including $\hat{\sigma}$)

#3fii

According to the formula in the book, there are 1337 possible heriarchical models if we consider all first and second-order terms and interactions.

If we're not considering any interactions, the total number of models becomes 461

#3fii

[1] "Here is the model with highest adj R^2 value"

	11 (1)
Word	TRUE
Parag	TRUE
Math	TRUE
Arith	TRUE
I.Word.2.	TRUE
I.Parag.2.	TRUE
I.Math.2.	TRUE
I.Arith.2.	TRUE
Word.Parag	FALSE
Word.Math	FALSE
Word.Arith	TRUE
Parag.Math	TRUE
Parag.Arith	FALSE
Math.Arith	TRUE

I used the regsubsets() function to find a list of models. I started by looking at the one with the max adjr2 value. This model was heirarchical, and it's R^2_{adj} is 0.9522792

#3fiv

Here is a summary of the best model.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-7.8349	1.1880	-6.59	0.0000
Word	-0.3309	0.1224	-2.70	0.0069
Parag	-1.6940	0.2456	-6.90	0.0000
Math	1.7670	0.1546	11.43	0.0000
Arith	1.0757	0.1398	7.70	0.0000
I(Word^2)	0.0208	0.0035	5.95	0.0000
I(Math^2)	-0.0089	0.0075	-1.18	0.2386
I(Arith^2)	-0.0091	0.0058	-1.58	0.1140
I(Parag^2)	0.1875	0.0169	11.07	0.0000
Parag:Math	0.0355	0.0163	2.18	0.0294
Parag:Arith	0.0104	0.0153	0.68	0.4966
Math:Arith	-0.0213	0.0096	-2.22	0.0262
Word:Arith	0.0145	0.0063	2.29	0.0221

Here is the fitted equation:

E{AFQT| Work, Parag, Math, Arith} = -7.835 + -0.331 * Word + -1.694 * Parag + 1.767 * Math + 1.076 * Arith + 0.021 * $Word^2$ + -0.009 * $Math^2$ + -0.009 * $Arith^2$ + 0.187 * $Parag^2$ + 0.036 * Parag * Math + 0.01 * Parag * Arith + -0.021 * Math * Arith + 0.014 * Word * Arith

#4a

(ex. 13, from ch.12)

I think that there is a danger of using variable selection techniques.

The R^2 is 0.1495745.

#4b

This model is suggested by forward selection:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-0.2092	0.1025	-2.04	0.0440
X3	-0.2606	0.1263	-2.06	0.0418
X1	-0.2105	0.1111	-1.90	0.0610
X9	-0.1825	0.1035	-1.76	0.0812

#4c

The model with the following X's has the smallest Cp Statistic:

	3 (1)
X1	TRUE
X2	FALSE
X3	TRUE
X4	FALSE
X5	FALSE
X6	FALSE
X7	FALSE
X8	FALSE
X9	TRUE
X10	FALSE

#4d

The model with the following X's has the smallest BIC:

	1 (1)
X1	FALSE
X2	FALSE
X3	TRUE
X4	FALSE
X5	FALSE
X6	FALSE
X7	FALSE
X8	FALSE
X9	FALSE
X10	FALSE

#4e

There is danger in using variable selection techniques – you'll get significant results due to chance. As shown above, we know that the data are independent, yet there are still some variables that seem significant. If we use variable selection techniques, we risk getting meaningless results.

Code

```r
# #1
options(xtable.comment = FALSE)
library(xtable)

variables <- c("None", "A", "B", "C", "AB", "AC", "BC", "ABC")
ResidSS <- c(8100, 6240, 5980, 6760, 5500, 5250, 5750, 5160)
df <- c(27, 26, 26,26,25,25,25,24)

sigmaSq <- ResidSS/df
AdjRSq <- (ResidSS[1]/df[1] - ResidSS/df)/(ResidSS[1]/df[1])

p <- c(1,2,2,2,3,3,3,4) # number of regression coefs
n <- 28
Cp <- p + (n-p)*(sigmaSq - sigmaSq[1])/(sigmaSq[1])
BIC <- n*log(ResidSS / n) + p * log(n)
#n*log(sigmaSq) + p * log(n) # formula from book is slightly different

myDF <- data.frame(variables, ResidSS, df, sigmaSq, AdjRSq, Cp, BIC)
print(xtable(myDF, caption = "Parts A-D"))

# x <- rnorm(28,100)
# y <- x + rnorm(28, 100, 16)
#
# plot(y~x)
#
# foo <- lm(y~x)
# foo1 <- lm(y~1)
# summary(foo)
#
# # how to calculate residual standard error
# sqrt(sum((foo$residuals^2)/(length(foo$residuals)-length(foo$coefficients))))
# summary(foo)
# summary(foo1)
# sum(foo$residuals^2)
# BIC(foo)
# n*log(18.06^2) + 2*log(n)
#
# ## this is the formula for BIC in R
# ## which agrees with this description: http://www.stat.wisc.edu/courses/st333-larget/aic.pdf
# ## BIC = n + n log 2*pi + n log(RSS/n) + (log n)(p + 1)
# n+n*log(2*pi) + n * log(sum(foo$residuals^2)/n) + log(n)*(3)
#
#
# ((8100/27 - 6240/26)/(8100/27))
# sqrt(sum(foo$residuals^2)/98)
#
# # multiple R^2
# (sum(foo1$residuals^2) - sum(foo$residuals^2))/sum(foo1$residuals^2)
#
# # Adj R^2
# (sum(foo1$residuals^2)/99 - sum(foo$residuals^2)/98)/(sum(foo1$residuals^2)/99)
```

```r
#
# # Adjusted R^2 in R
# 1-(1-(sum(foo1$residuals^2) - sum(foo$residuals^2))/sum(foo1$residuals^2)) * (99/ 98)
#
# .55332/16


### (#10e, i-iv)

tabb <- sapply(X=c("sigmaSq", "Cp", "BIC"),
       function(thing) {
            as.character(myDF$variables[ myDF[thing] == min(myDF[thing])])
            })
BestCombo <- c(as.character(myDF$variables[ myDF["AdjRSq"] == max(myDF["AdjRSq"])]), tabb)

names(BestCombo) <- c("max R^2", "min sigmaSq", " min Cp", "min BIC")

print(xtable(data.frame(BestCombo), caption = "1e, part i-iv"))

### #11 (still part of #1)

Fpv <- pf(q = ((8100 - 5980)/1)/300, df1 = 1, df2 = 27, lower.tail = FALSE)

Fpv2 <- pf(q = ((5980 - 5500)/1)/230, df1 = 1, df2 = 26, lower.tail = FALSE)

# 2a

library(car)
finc <- read.csv("data/ex1223.csv")
finc$Race <- as.factor(finc$Race)
#colnames(finc)
# pairs(finc[,c("Educ","MotherEd", "FatherEd", "FamilyIncome78", "Science",
#               "Arith", "Word", "Parag", "Numer", "Coding", "Auto", "Math", "Mechanic",
#               "Elec", "AFQT", "Income2005")])


# I just used a sample, so it wouldn't take so long!
scatterplotMatrix(finc[sample(1:nrow(finc), 500),c("Educ","MotherEd",
                                                   "FatherEd", "FamilyIncome78", "Science",
              "Arith", "Word", "Parag", "Numer", "Coding", "Auto", "Math", "Mechanic",
              "Elec", "AFQT", "Income2005")], col = c("red", "blue", rgb(0,0,0,0.2)), pch = ".")

# looks like we should square parag and sqrt income2005
finc$SIncome2005 <- sqrt(finc$Income2005)
finc$Par2 <- finc$Parag^2

vars <- c("Educ","MotherEd", "FatherEd", "FamilyIncome78", "Science",
              "Arith", "Word", "Parag", "Numer", "Coding", "Auto", "Math", "Mechanic",
              "Elec", "AFQT",  "Income2005")


#paste(vars, collapse = "+")
```

```
modM <- lm(sqrt(Income2005) ~ Educ+MotherEd+FatherEd+FamilyIncome78+
                Educ+Science+Arith+Word+Parag+Numer+Coding+Auto+
                Math+Mechanic+Elec+AFQT+Race, data = finc)
par(mfrow = c(1,2))
plot(modM, which = 1:2, col = rgb(0,0,0, 0.15), pch = 20)
par(mfrow = c(1,1))

# dat <-sqrt(finc$Income2005)
# hist(dat, freq = FALSE, breaks = 100)
# lines(x = seq(min(dat), max(dat), length = 100), y = dnorm(x = seq(min(dat), to = max(dat), length =

# #2b

mod2b <- lm(sqrt(Income2005)~ Educ*Race,  data = finc)
# par(mfrow = c(1,2))
# plot(mod2b, which = 1:2)
# par(mfrow = c(1,1))
print(xtable(summary(mod2b), caption= "Model for 2b with interaction terms"))
#mod2b$coefficients[2] + mod2b$coefficients[5]
#levels(finc$Race)
finc$Race <- relevel(finc$Race, ref = "2")

mod2bb <- lm(sqrt(Income2005)~ Educ*Race,  data = finc)
print(xtable(summary(mod2bb), caption= "Model for 2b with interaction terms, and new reference level"))
foo <- summary(mod2bb)
# foo$coefficients[6, 4]

# #2c

library(MASS)

regDat <- finc[,2:22]
FullMod <- lm(sqrt(Income2005) ~ .^2, regDat)
backMod <- stepAIC(object = FullMod, direction = "backward", trace = FALSE)


#bb <- step(object = FullMod, direction = "backward", trace = T)
#print(xtable(backMod, caption = "Results from backward variable selection, based on AIC"))
print(round(summary(backMod)$coefficients, 4))
foo <- extractAIC(backMod)[[2]]
print(paste("The AIC from this model is", round(foo,2)))

# #2d

IntOnlyMod <- lm(sqrt(Income2005) ~ 1, regDat)

forMod <- step(IntOnlyMod, scope = list(upper = FullMod), direction = "forward", k = 2, trace = FALSE)
print(round(summary(forMod)$coefficients, 4))

print(paste("The AIC for the forward-selected model is", round(extractAIC(forMod)[[2]], 2)))

# #2e
```

```r
MainMod <- lm(sqrt(Income2005)~., regDat)
stepMod <- step(MainMod, scope = list(lower = IntOnlyMod, upper = FullMod), direction = "both", trace =

print(round(summary(stepMod)$coefficients, 4))
print(paste("The AIC for the forward-selected model is", round(extractAIC(stepMod)[[2]], 2)))

# #2f

par(mfrow = c(2,2))
plot(backMod, which = 1:4, col = rgb(0,0,0, 0.15), pch = 20)
par(mfrow = c(1,1))

# #2g

#summary(backMod)


NewDataSet <- data.frame(Imagazine = 1, Inewspaper  = 1, Ilibrary = 1,
                         MotherEd = 12, FatherEd = 12, FamilyIncome78 = median(finc$FamilyIncome78),
                         Race = "3", Gender = "male", Educ = 12, Science = mean(finc$Science),
                         Arith = mean(finc$Arith), Word = mean(finc$Word), Parag = mean(finc$Parag),
                         Numer = mean(finc$Numer), Coding = mean(finc$Coding), Auto = mean(finc$Auto),
                         Math = mean(finc$Auto), Mechanic = mean(finc$Mechanic),
                         Elec = mean(finc$Elec), AFQT = mean(finc$AFQT))


ci <- predict(backMod, new = NewDataSet, interval="confidence", level = .95)
print("Here is the 95% confidence interval for the mean")
ci^2
pii <- predict(backMod, new = NewDataSet, interval="predict", level = 0.95)
print("Here is the 95% prediction interval")
pii^2

# #3a

Mod3aFull <- lm(sqrt(Income2005)~Word+Parag+Math+Arith+AFQT, data = finc)
m3a <- summary(Mod3aFull)
#m3a$sigma^2*2578
print(paste("Sigma from the full model is",m3a$sigma, "with df =", m3a$df[2]))

Mod3aReduced <- lm(sqrt(Income2005)~AFQT, data = finc)
m3aR <- summary(Mod3aReduced)
print(paste("Sigma from the reduced model is",m3aR$sigma, "with df =", m3aR$df[2]))
ESS <- (m3aR$sigma^2*2582) - (m3a$sigma^2*2578)

FStat <- (ESS/4)/(m3a$sigma^2)

#pf(q = 18.511, df1 = 4, df2 = 2578, lower.tail = F)

print("Here is the anova table to compare with my calculations:")
print(xtable(anova(Mod3aReduced, Mod3aFull)))

# #3b
```

```r
Mod3bFull <- lm(sqrt(Income2005)~Word+Parag+Math+Arith+AFQT, data = finc)

Mod3bReduced <- lm(sqrt(Income2005)~Word+Parag+Math+Arith, data = finc)

print(xtable(anova(Mod3bReduced, Mod3bFull)))

# #3c

library(car)
print(xtable(summary(Mod3bFull)), floating = F)
print(xtable(summary(Mod3aReduced)), floating = F)

scatterplotMatrix(finc[c("Word", "Parag", "Math", "Arith", "AFQT")],
                  col = c("red", "blue", rgb(0,0,0,0.05)), pch = 20)

# #3d

tolReg <- lm(AFQT~Word+Parag+Math+Arith, data = finc)
fo <- summary(tolReg)
print(xtable(fo), floating = F)


tolerance <- 1-fo$r.squared
vifF <- 1/tolerance

#vif(Mod3bFull) this is true

# 3e

varFull <- (summary(Mod3bFull)$coefficients[6,2])^2
varRed <- (m3aR$coefficients[2,2])^2

vb <- (varRed) * (vifF) * ((m3a$sigma)^2/(m3aR$sigma)^2)


# #3f


# #3fi

library(leaps)
#AFQT~Word+Parag+Math+Arith, data = finc)
dat3f <- finc[c("AFQT", "Word", "Parag", "Math", "Arith")]
mod3fi <- lm(AFQT~ . ^2 + I(Word^2) + I(Parag^2) + I(Math^2) + I(Arith^2), data = dat3f)
# length(mod3fi$coefficients)

# Calculating total number of models
K = 4 # number of variables
p = 15 # total numer of parameters in max model
numx <- function(p = 1){
     sum(sapply(X = 0:K, FUN = function(x){
     choose(K, x) * choose(choose(x+1, 2), p-1-x)
}))}
```

```
#sum(sapply(1:15, numx))


# #3fii

leaps <- regsubsets(AFQT~ . ^2 + I(Word^2) + I(Parag^2) + I(Math^2) + I(Arith^2), data = dat3f, nbest=5
foo <- summary(leaps, matrix.logical = T, scale = "adjr2")
fdf <- data.frame(foo$outmat)

print("Here is the model with highest adj R^2 value")
print(xtable(t(fdf[foo$adjr2 == max(foo$adjr2), ])))

# #3fiv

bestMod <- lm(AFQT~Word+Parag*Math + Parag*Arith+Math*Arith+Word*Arith + I(Word^2) + I(Math^2) + I(Arit
print(xtable(summary(bestMod)), floating = F)

toe <- summary(bestMod)
cofs <- toe$coefficients[,1]
cofs <- round(cofs, 3)

# #4a

Y <- rnorm(100)
X <- sapply(1:10, function(o) rnorm(100))
df <- as.data.frame(cbind(Y, X))
colnames(df) = c("Y", paste("X", 1:10, sep = ""))

#head(df)

mod4 <- lm(Y~., data = df)
md <- summary(mod4)


# #4b

intMod <- lm(Y~1, data = df)
forMod <- step(object = intMod, scope = list(upper = mod4), direction = "forward",k = 2, trace =F)
print(xtable(forMod), floating = F)

# #4c

leaps4 <- regsubsets(Y~ . , data =df, nbest=5, nvmax = 15, method = "exhaustive")
#plot((leaps4), scale = "Cp")
fo <- summary(leaps4, scale = "Cp", matrix.logical = T)
fodf <- as.data.frame(fo$outmat)
print(xtable(t(fodf[fo$cp == min(fo$cp), ])), floating = F)

# #4d

fo <- summary(leaps4, scale = "bic", matrix.logical = T)
fodf <- as.data.frame(fo$outmat)
print(xtable(t(fodf[fo$bic == min(fo$bic), ])))
```