Multiple Regression

Rob Root

4/30/2018

Introduction to multiple regression

An example and some theory

Consider a math department trying to predict the success of its students in first semester Calculus. The students' admissions profiles might offer a range of data might offer some indication of the success students can expect in Calculus. Consider in particular the possibility of making a prediction using some subset of a student's x_1: grade in high school algebra, x_2 (standardized on a scale from 0 to 30): score on the ACT math test, x_3: score on the ACT natural science test, and x_4: high school class rank (as a percentile). To predict the grade of future students, we might begin by using the linear regression model, assuming that the grade of any particular student i, is a normal random variable with mean determined by the four x-variables aready measured: algebra grade x_{i1}, ACT math score x_{i2}, ACT science score x_{i3}, and class percentile x_{i4}. We can then assume

$$Y_i = Y(x_{i1}, x_{i2}, x_{i3}, x_{i4}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \epsilon_i \text{ where } \epsilon_i \sim N(0, \sigma) \text{ or}$$
$$Y(\vec{x}_i) = \vec{\beta} \cdot \vec{x}_i + \epsilon_i \text{ where } \vec{x}_i = (1, x_{i1}, x_{i2}, x_{i3}, x_{i4}),\ \vec{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4).$$

Here the error variables ϵ_j are iid σ multiples of the standard normal. If we accept this model then we can use data to find estimators of the parameters, $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3, \hat{\beta}_4) = \hat{\vec{\beta}}$ and $\hat{\sigma}$. Either maximum likelihood estimation or minimizing $SSE = \sum(y_i - \hat{y}_i)^2$ (where $\hat{y}_i = \hat{\vec{\beta}} \cdot \vec{x}_i$) leads to normal equations that you can guess:

$$\begin{aligned} 0 &= \sum y_i - \hat{y}_i \\ 0 &= \sum (y_i - \hat{y}_i) x_{i1} \\ &\vdots \\ 0 &= \sum (y_i - \hat{y}_i) x_{i4}. \end{aligned}$$

These are linear in the $\hat{\beta}_j$s and can usually be solved uniquely. The actual solution requires more linear algebra than you need to take this course. There is a follow-on course in regression analysis that goes deeper into the solution. For us it is enough to know that these estimators exist and that $\hat{\sigma}^2 = s^2 = SSE/(n-5)$ is an unbiased estimator for the variance of the Y_is.

Further the decomposition of variation carries over without modification: $SST = \sum(y_i - \bar{y})^2$ and $SSR = \sum(\hat{y}_i - \bar{y})^2$ still satisfy $SST = SSE + SSR$, and further $SSE \sim \sigma^2 \chi^2_{n-5}$ and $SSR \sim \sigma^2 \chi^2_4$. Each of the $\hat{\beta}_j$ have variance $\sigma^2_{\hat{\beta}_j}$ that is σ^2 times a function of the x_i, so the standard error of $\hat{\beta}_j$ is the square root of that same function times $s^2 = SSE/(n-5)$. That means $s_{\hat{\beta}_j}$ is a constant times the

square root of $\chi^2_{n-5}/(n-5)$, so, if $H_0 : \beta_j = 0$ is true, $\hat{\beta}_j/s_{\hat{\beta}_j}$ is a t variable with $n-5$ degrees of freedom. We can use these statistics to test whether each of the β_j are actually zero, and can be removed from the model.

Further, we can construct a whole model utility test using an F-statistic just like the one for simple regression:

$$\frac{SSR/4}{SSE/(n-5)} \sim \frac{\chi^2_4/4}{\chi^2_{n-5}/(n-5)} = F_{4,n-5}.$$

Calculation and interpretation of R output

We can put this to work on the data from Example 12.25 from the text, which considers this very situation. Here is R loading the data and generating a summary of the multiple regression model.

```
setwd("~/Documents/336 Spring 18/R/") #Change this to work on your computer
load("CH12/exmp1225.RData")           #Note the data file is in a directory called CH12
attach('exmp12-25')
mreg<-lm(Grade ~ Algebra + actm + actns + hsrank) #Create regression model
summary(mreg)  #View summary of model
```

```
##
## Call:
## lm(formula = Grade ~ Algebra + actm + actns + hsrank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -21.812  -9.057   1.657   6.656  17.947
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.1215    10.7519   3.360 0.001230 **
## Algebra       0.9610     0.2640   3.640 0.000499 ***
## actm          0.2718     0.4535   0.599 0.550737
## actns         0.2161     0.3132   0.690 0.492351
## hsrank        0.1353     0.1036   1.306 0.195677
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.897 on 75 degrees of freedom
## Multiple R-squared:  0.289,  Adjusted R-squared:  0.2511
## F-statistic: 7.622 on 4 and 75 DF,  p-value: 3.297e-05
```

Looking at the `summary` command output, we see that it begins with the command that generated the model, followed by a five number summary of the distribution of the residuals. Given that the y-value is a Calculus grade based on the usual one hundred point scale, we see that these residuals are fairly large.

Next comes a list of information about each coefficient. As in simple regression, the first column contains

the name of the variable for that coefficient. The second column is the $\hat{\beta}_j$ value, our estimate for the coefficient. The third column gives the standard errors for the estimates. The ratio of estimator divided by standard error gives the t statistic in the fourth column. The degrees of freedom for all these t statistics (75, since $n = 80$) is given just below this table. The fifth column gives the p-values for the null hypothesis $H_0 : \beta_j = 0$ against a two-sided alternative, with the final column a visual code for the strength of the evidence that the coefficient is nonzero. Notice that the only coefficients that are clearly not zero are β_0 and β_1. This suggests that this model is too complicated; we will return to this concern after with review the information in the summary.

After the key to the visual code of significance comes the value of $s = \sqrt{SSE/(n-5)}$ and the degrees of freedom $n - 5$. The multiple r^2 value is the coefficient of determination,

$$r^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST}$$

Also provided is the adjusted r^2 value, which is useful in this setting. The idea behind the adjusted r^2 is to try to balance the increased accuracy of fit possible by adding an additional explanatory variable (x variable) to the model with the complication of one more explanatory variable. The idea is simple, each of SST, SSE and SSR has a degrees of freedom parameter that goes with it, $d_{SST} = n-1$, $d_{SSE} = n-5$, and $d_{SSR} = 4$, respectively. Notice that just as $SST = SSE + SSR$, also $d_{SST} = d_{SSE} + d_{SSR}$. we define the mean square quantities MST, MSE, MSR by dividing the sum of square quantities by their degrees of freedom: $MST = SST/(n-1)$, $MSE = SSE/(n-5)$ and $MSR = SSR/4$. Then we define the adjusted r^2 by

$$r^2_{\text{adj}} = 1 - \frac{MSE}{MST} = 1 - \frac{SSE/(n-5)}{SST/(n-1)} = 1 - \frac{n-1}{n-5} \cdot \frac{SSE}{SST}.$$

Notice that this differs from the coefficient of determination by the fraction $(n-1)/(n-5) > 1$ multiplying SSE/SST before subtracting from 1. Thus the difference, the adjusted r^2, is slightly smaller than r^2. This fraction penalizes the model for th fact that it uses four explanatory variables to predict $\hat{y}_i$-values. In general, a model that uses k explanatory variables will have

$$r^2_{\text{adj}} = 1 - \frac{n-1}{n-(k+1)} \cdot \frac{SSE}{SST} \text{ for a model with } k \text{ explanatory variables.}$$

This makes the size of r^2_{adj} a measure of the quality of of the model's fit that offers some compensation for the complexity of the model.

Notice that neither the coefficient of determination r^2 nor the (never larger) adjusted r^2 are very large at all. The model accounts for less than 29% of the total variation in Calculus grades. However, looking at the next line, we see that that even that modest reduction in variation is compelling evidence that the model is helpful in explaining the variation in Calculus grades. The Whole Model Utility Test is based on the f-statistic mentioned above, and its tiny p-value suggests that the model has great value in explaining Calculus grades. The tiny p-value is not an indication of the value of the entire model, but it suggests that, if the model assumptions are reasonable, at least one of the coefficients in the model is nonzero. Recalling the results of the inference for each of the individual β_js, we have strong reason to believe that β_0 and β_1 are nonzero, but the argument for the remaining β_js is less compelling.

The aov command gives a more detailed decomposition of variation for the model. It breaks SSR into components for each of the explanatory variables.

```
aov(mreg)
```

```
## Call:
##    aov(formula = mreg)
##
## Terms:
##                   Algebra     actm    actns   hsrank Residuals
## Sum of Squares  2490.766  258.378   70.044  166.962  7346.050
## Deg. of Freedom        1        1        1        1        75
##
## Residual standard error: 9.896834
## Estimated effects may be unbalanced
```

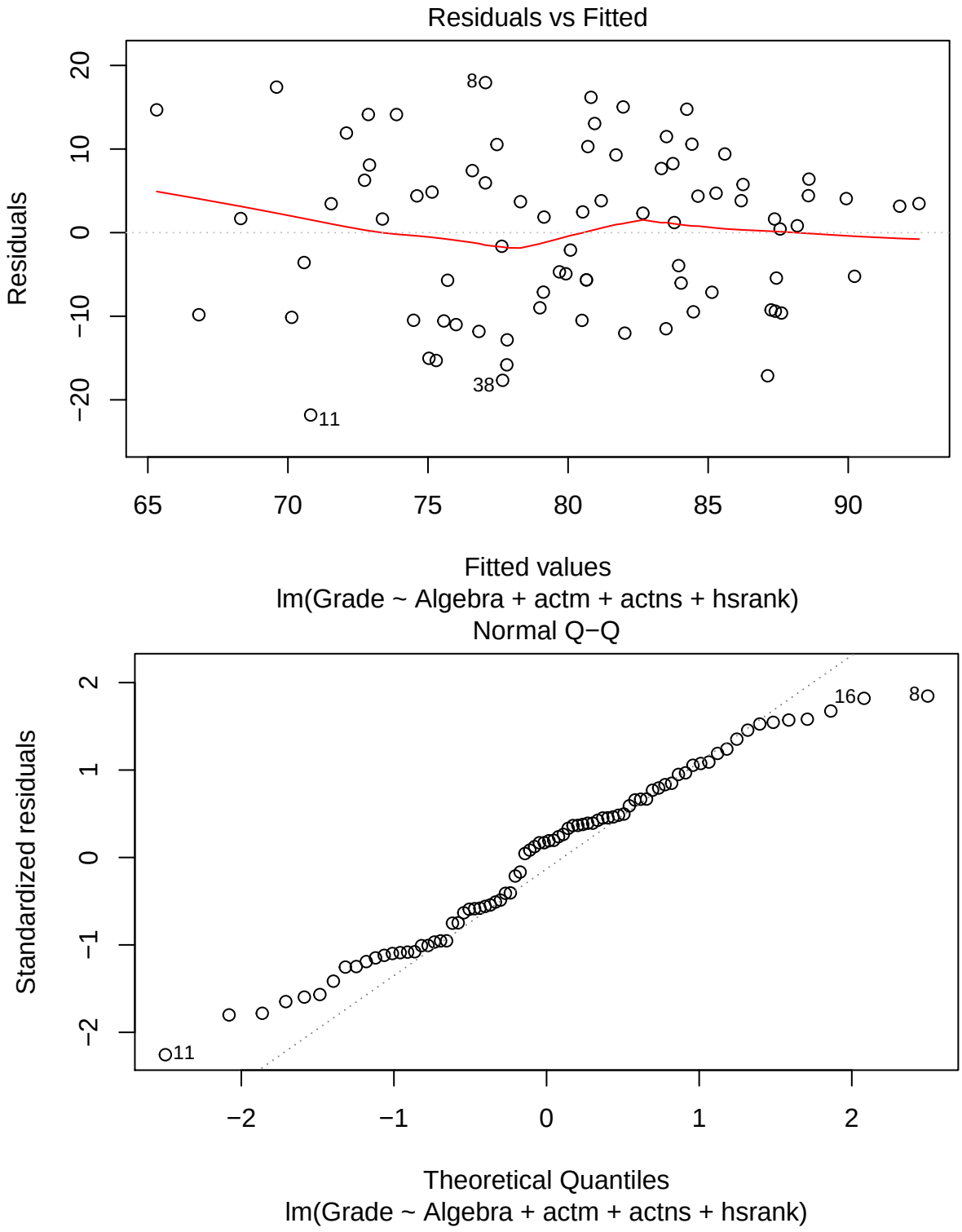
Notice that the sum of squares for the x_1 variable (Algebra grade) is close to ten times larger (or more) than the sum of squares for any other explanatory variable. Note also that the warning that "Estimated effect may be unbalanced" appears, indicating that some combinations of the explanatory variables occur with greater frequency than others. This is to be expected, since all four explanatory variables are likely to be correlated with one another.

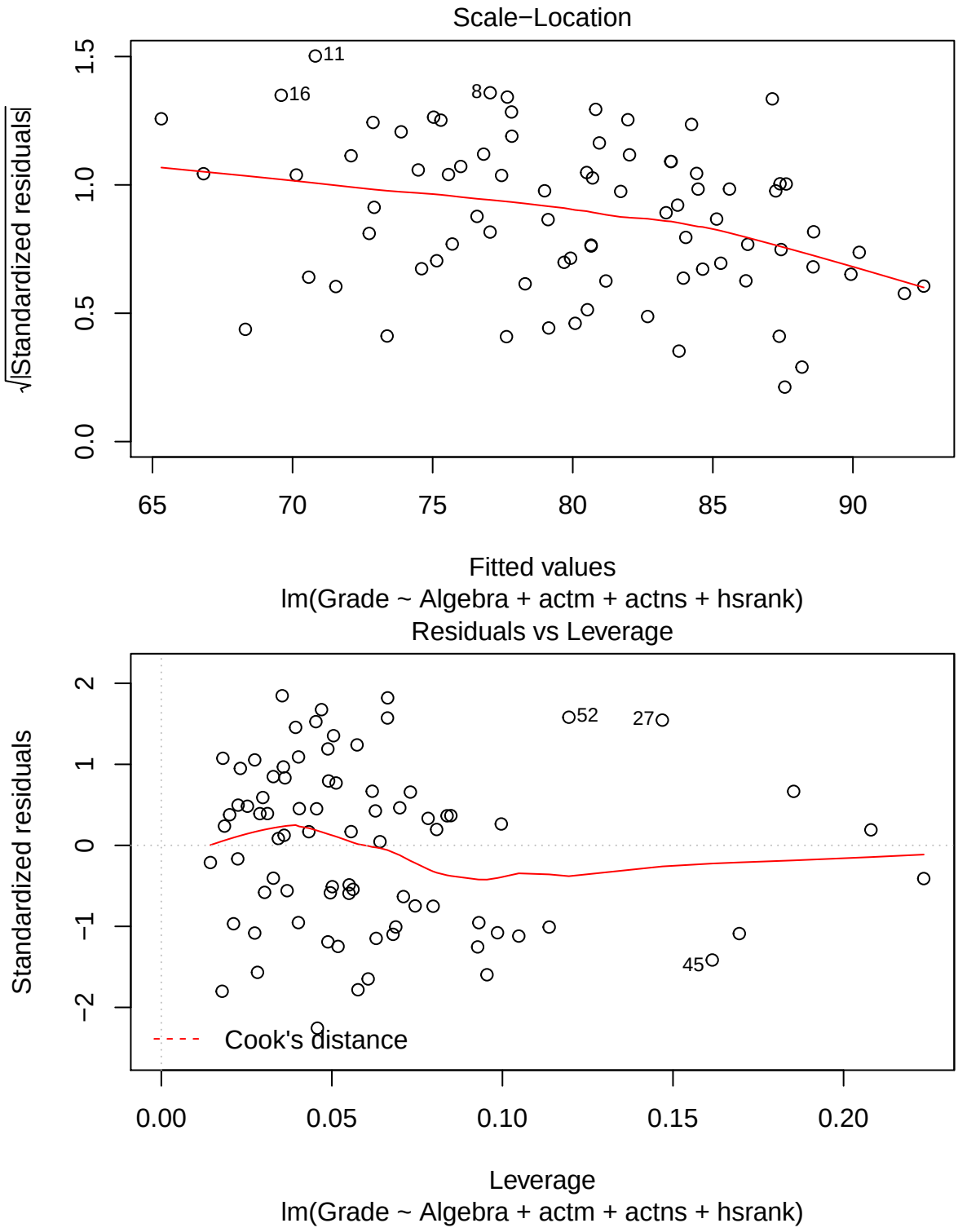
This is called *multicollinearity*, or sometimes just *collinearity*. It can result in drastic changes in the coefficient estimates in the model when there are only small changes in the data. This is not likely to be important in models like this one, where one explanatory variable (here, Algebra grades) is vastly more successful than the others in predicting the response variable (here, Calculus grades). In general, multicollinearity does not harm the predictive capability of the model, but it does render the estimates unstable in response to small changes in the data.

Model checking

We should still look at the diagnostic plots to see if the data generally fits the regression model.

```
plot(mreg)
```


Residuals vs Fitted
Residuals
20
10
0
−10
−20
8
38
11
65
70
75
80
85
90
Fitted values
lm(Grade ~ Algebra + actm + actns + hsrank)
Normal Q−Q
Standardized residuals
2
1
0
−1
−2
16
8
11
−2
−1
0
1
2
Theoretical Quantiles
lm(Grade ~ Algebra + actm + actns + hsrank)



The first plot suggests that there is no evidence of nonlinearity in the residuals. It is worth pointing out that the dependence on more that one explanatory variable can obscure nonlinearity in the dependence on one variable. In this instance, where one coefficient is large in comparison with the others, that is unlikely to be a problem.

The second plot suggests that the residuals deviate from normality by having thin tails. This makes some sense. Positive residuals are constrained by the firm upper bound on Calculus grades of 100%; that is, the largest possible residual for $\hat{y}_i$ is $100 - \hat{y}_i$. This explains the thin right/upper tail. *Survivorship bias* explains the lack of large negative residuals. Students whose performance is drastically worse than they anticipated are likely to drop Calculus, and so not "survive" to contribute a y value to the data set. Changing to logistic regression would allow the model to overcome the lack of large positive residuals, but not fix the survivorship bias. It is probably better to accept the more balanced deviation from normality in the linear model.

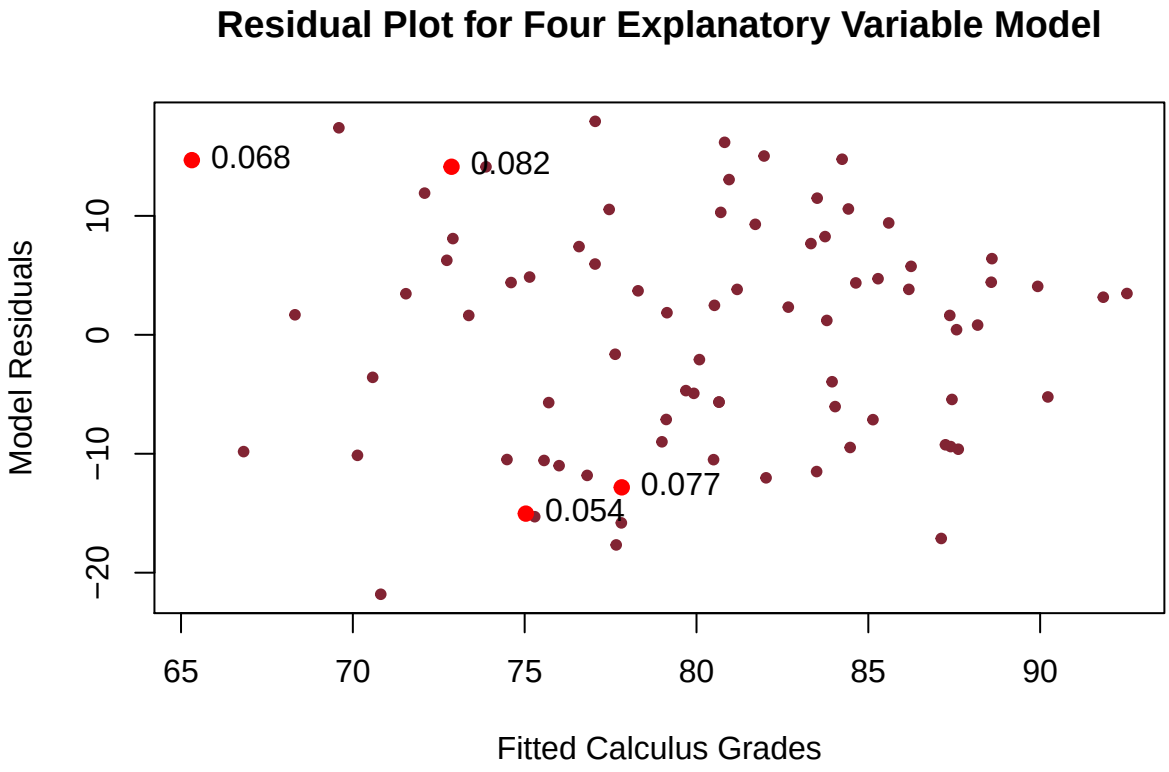
The third plot shows that the data is slightly heteroskedastic. This is because students with higher expected grades are more likely to have their over-performance obscured by the 100% upper bound. Again, this is likely corrected by the logistic transformation, or another transformation with the same effect of expanding distance near $y = 100\%$. As above, we prefer to live with this artifact of the data rather than unbalance the distribution of residuals.

The fourth plot shows that, under the lenient interpretation of Cook's distance, there is no reason to be concerned about any of the data having unusual influence. The contours for Cook's distance at 0.5 and 1 are both literally off the chart. If we use the more stringent interpretation of Cook's distance, the limit is $4/n = 0.05$. Here is an analysis similar to the one offered in last class for this situation. Notice that the scatterplot is for residuals against fitted values, like the first diagnostic plot, since we cannot plot the actual y-values against all four explanatory variables (without a five dimensional plot).

```
cd<-cooks.distance(mreg)
cd[cd>0.05]
```

```
##         25         27         45         52
## 0.05384468 0.08225303 0.07715387 0.06785774
```

```
inflcds<-sort(cd)[77:80]
infli<-as.numeric(names(inflcds))
plot(mreg$fitted.values,mreg$residuals,xlab="Fitted Calculus Grades",ylab="Model Residuals",
     main="Residual Plot for Four Explanatory Variable Model",pch=20,col=rgb(0.51,0.14,0.2))
points(mreg$fitted.values[infli],mreg$residuals[infli],pch=19, col="red")
text(mreg$fitted.values[infli],mreg$residuals[infli],labels=round(cd[infli],3),pos=4)
```



We see that there are four data points with influence large than the stricter threshold of 0.05. Looking at the scatterplot, three of these four seem to have fitted values that are common among many students and residuals that are unusually low or high. These large Cook's distances are likely an artifact of the thin tails in the distribution of residuals. Only data point 58, with a Cook's distance of roughly 0.068 is located in a position where it has great leverage, with the student receiving the lowest fitted Calculus score, but over-performing that prediction by over 10 points. This is not an unusual phenomenon from an empirical point of view. A student with a relatively unimpressive admissions profile arrives at college and substantially exceeds expectations: this is basically the plot of *Legally Blonde*, among many other cultural manifestations. Given the frequency of this as an observed event, it seems wise to allow it its moderately large influence on our model.

Using the model to predict

To summarize, we have a flawed but acceptable model for our data. Right now its principle weakness is that it uses several explanatory variables that seem to be superfluous, rendering the model excessively complicated. We can still use the model quite easily in R. Here are several model predictions associated with one point of the explanatory variables. We choose the point with coordinates the means of all the explanatory variables. This will make the $\hat{y}$-value equal to $\bar{y}$, and result in the narrowest confidence interval for $\hat{y}$ and prediction interval for y.

```
newpt<-data.frame(Algebra=19.425,actm=28.25,actns=26.925,hsrank=87.675) #explanatory variables at their m
predict(mreg,newdata=newpt) #fitted value yhat at point of interest, s.b. ybar
```

```
##      1
```

```
## 80.15
```

```
predict(mreg,newdata=newpt,se.fit=T) #standard error for yhat at point of interest
```

```
## $fit
##     1
## 80.15
##
## $se.fit
## [1] 1.1065
##
## $df
## [1] 75
##
## $residual.scale
## [1] 9.896834
```

```
predict(mreg,newdata=newpt,interval="confidence") #CI for fitted value
```

```
##     fit      lwr      upr
## 1 80.15 77.94574 82.35426
```

```
predict(mreg,newdata=newpt,interval="prediction") #PI for new observation
```

```
##     fit      lwr      upr
## 1 80.15 60.31166 99.98834
```

The second command gives the fitted value, $\hat{y} \approx 80$ associated with the combination of means of explanatory variables indicated in the first command. The third command repeats the fitted value, and gives the standard error in that estimate, together with the degrees of freedom of the model and the value of $s = \sqrt{SSE/(n-5)}$ used in calculating that standard error. The standard error at any other point would be larger. The fourth command gives a 95% confidence interval for $\hat{y}$, and the fifth command gives a 95% prediction interval for a new observation associated with that combination of explanatory variables. It is interesting to note that, while the fitted value is likely to be on the cusp between C and B, between 78 and 82, the prediction is somewhere between 60, a D−, and 100, a perfect A+. Thus all we can safely predict about this student is that they will likely not fail. This is a consequence of the low coefficient of determination of this model. Relatively little of the variation in Calculus grades is accounted for by the correlation with the explanatory variables.

Let's turn next to ways to simplify the model. This cannot improve the problem of weak correlation observed in the prediction interval by simplifying, but we can end up with a simpler model that is nearly as predictive.

Elements of model selection

Backward stepwise regression

We can try to simplify the model by "backward stepwise regression;" removing the coefficient with the highest p-value. Typically, as we do this, the p-values of the other coefficients decrease. First, ACTM has the highest p-value, so we remove it. Because the p-value is greater than 0.5, this seems perfectly reasonable. An estimate at least as large (in absolute value) as the one we calculated is more likely than not if the coefficient β_2 is zero.

```
reg3var<-lm(Grade ~ Algebra + actns + hsrank) #Create regression model
summary(reg3var)  #View summary of model
```

```
##
## Call:
## lm(formula = Grade ~ Algebra + actns + hsrank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.054  -8.453   2.003   7.021  17.885
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.12006    9.47687   4.128 9.29e-05 ***
## Algebra      1.01060    0.24967   4.048 0.000123 ***
## actns        0.29878    0.28001   1.067 0.289336
## hsrank       0.15232    0.09926   1.534 0.129075
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.855 on 76 degrees of freedom
## Multiple R-squared:  0.2856, Adjusted R-squared:  0.2574
## F-statistic: 10.13 on 3 and 76 DF,  p-value: 1.09e-05
```

Looking at the summary of the model with three explanatory variables, we see that the t statistics for all the remaining statistics are larger than they were in the four variable model, and so the p-values are all smaller. The coeffient of determination, r^2, is smaller, but only barely, dropping from 0.289 to 0.2856. This drop is so slight that the adjusted r^2 increases from 0.2511 to 0.2572. This suggests that the simpler model is better. Also, the p-value for the whole model utility test has decreased, also suggesting the three variable model is better than the four variable model.

Taking the next step in backward stepwise regression, we remove the ACT natural science test score, as it has the largest p-value in the three variable model.

```
reg2var<-lm(Grade ~ Algebra + hsrank) #Create regression model
summary(reg2var)  #View summary of model
```

```
##
## Call:
## lm(formula = Grade ~ Algebra + hsrank)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.145  -8.364   1.226   6.531  18.325
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.86850    8.37456   5.238 1.37e-06 ***
## Algebra      1.04993    0.24716   4.248 5.98e-05 ***
## hsrank       0.18120    0.09559   1.896   0.0618 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.864 on 77 degrees of freedom
## Multiple R-squared:  0.2749, Adjusted R-squared:  0.2561
## F-statistic:  14.6 on 2 and 77 DF,  p-value: 4.219e-06
```

In the two variable model we see that the pattern of increasing t statistics and decreasing p-values for the remaining β_js continues. The pattern of decreasing r^2 is also reliable, although the drop this time is larger, from 0.2856 to 0.2749. This accelerating decrease is also usual. This time the drop is large enough that the adjusted r^2 also drops, suggesting that this is a slightly worse model. However, the whole model utility p-value is smaller suggesting that this is a better model.

For the moment, let's assume that we wish to continue simplifying the model, trusting the whole model p-value over the adjusted r^2. (We will shortly talk about more sophisticated statistics for assessing the suitability of competing models.) As we remove the hsrank variable, it is worth reflecting on its p-value of 0.062. This is more than 0.05, so not small enough to allow us to reject the null hypothesis. However, by removing the variable from the model we are actively accepting the null hypothesis. In other words, we are not concerned with the probability of a Type I error, by our action, we stand in danger of making a Type II error, and so the relevant probability is the β value associated with some $\beta_4 \neq 0$. (Note that we are facing a weakness of the book's notation, using β to mean both the probability of a Type II error and also a generic regression coefficient.)

```
reg1var<-lm(Grade ~ Algebra) #Create regression model
summary(reg1var)  #View summary of model
```

```
##
## Call:
## lm(formula = Grade ~ Algebra)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.301  -8.712   2.281   6.983  20.124
```

```
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  56.9991     4.7842  11.914  < 2e-16 ***
## Algebra       1.1918     0.2394   4.978 3.76e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.03 on 78 degrees of freedom
## Multiple R-squared:  0.2411, Adjusted R-squared:  0.2313
## F-statistic: 24.78 on 1 and 78 DF,  p-value: 3.764e-06
```

Looking at the one variable model, notice that the drop in r^2 continues to accelerate, and the drop in adjusted r^2 also accelerates, but there is still a slight decrease in the p-value of the whole model utility test. Should we adopt this model in preference to the two variable model? Should we return to the three variable model? (Nothing suggests that the four variable model would be better than the three variable model.) We need a more reliable way of distinguishing between competing models.

Using an information criterion to choose a model

There is a statistic called an information criterion that serves as a metric for comparing models. It is basically the log likelihood function for the model, penalized by a measure of the number of parameters (β_js) in the model. The log likelihood is multiplied by -1 to make it positive, so smaller is better. The positive penalty is then added and takes two different forms. The small penalty form is called AIC (for Akaike Information Criterion), is more likely to give a more complicated model. The large penalty form is called BIC (for Bayesian Information Criterion) and is more likely to give a less complicated model. There are reasons to use each of these. In our case, we can see this difference in action. Here are the AIC and BIC values for all of the regression models we have created.

```
c(AIC(mreg),AIC(reg3var),AIC(reg2var),AIC(reg1var))
```

```
## [1] 600.6215 599.0037 598.1933 599.8422
```

```
c(BIC(mreg),BIC(reg3var),BIC(reg2var),BIC(reg1var))
```

```
## [1] 614.9136 610.9139 607.7214 606.9883
```

both measures agree that the two variable model is an improvement over the more complicated models, but they differ in comparing the two-variable and one variable models. AIC gives the lowest value (best score) to the two variable model, while BIC gives the best score to the one-variable model. These are the two we should choose between, and the final choice depends on our preference for parsimony (simplicity of the model) or accuracy (larger r^2).

In our example, we need to choose between a simple linear regression model and a slightly more complicated multiple linear regression model. We can think of the high school rank as a rough measure of a student's conscientiousness in school, and the algebra grade as a measure of the student's mathematical background, aptitude, and interest. If this interpretation is accurate, the high school rank offers a predic-

tive measure that is at least a little independent of the the algebra grade that is clearly the primary predictor. The improvement in fit afforded by including it is so modest that it is not clear that it is worth the added complication, but it might be.

From a practical point of view, one can argue that the high school rank coefficient is too small to make much practical difference. We can see in the five number summary below that the middle fifty percent of high school ranks in the data set ranges from 82 to 97, for an IQR of 15 points. The coefficient in the two variable model is about 0.19, making a more from the 25th percentile to the 75th percentile change the expected grade less than 3 points, so less than one grade step. Given that a change in rank that large has so small an effect on the expected grade, it seems safe to ignore its effect as negligible, and use the simpler model. This argument depends on the context of the problem; one can imagine a different situation when a difference of less than 3 in the $\hat{y}$-value would be more significant, and so suggest that inclusion of high school rank is appropriate.

```
fivenum(hsrank)
```

```
## [1] 46 82 92 97 99
```