First Year College

Cody Frisby

March 2, 2016

1) Intent

Colleges have limited resources of financial aid for students. They would like to have high levels of confidence in the students' success at university who are awarded financial aid. If we cannot reward all applicants who are eligible for financial aid then whom should we? Using metrics from the applicants high school perhaps we can have a model where we can select those that are most likely to succeed. Here we have data from the last 10 years where the response is a SRS of student's GPA after their first year of college and the predictors being these metrics from high school.

2) Exploratory Analysis

Using the attached scatter plot matrix we can identify highly correlated predictors of gpa. Also, the correlation coefficients between gpa and the covariates can help us asess if MLR is appropriate. All the correlation coeffs are above 0.6 with gpa. There are some covariates that are highly correlated with each other as well, for example satm & hse. This may be of concern when looking at which variables to include in our MLR model. Here, it appears MLR would be appropriate.

3) MLR model

A) Here we write our regression model:

$$gpa = \beta_0 + \beta_1 satm + \beta_2 satv + \beta_3 hsm + \beta_4 hse + \varepsilon_i$$

and the equation

$$gpa = 0.1615496 + 0.0020102 satm + 0.0012522 satv + 0.1894402 hsm + 0.0875637 hse$$

where *satm* and *satv* $\beta_i s$ are significant at $\alpha = 0.05$. *hsm* and *hse* are not significant. Here, for every unit increase in SAT math scores, we expect a 0.00201 increase in *gpa*, holding all other variables constant.

B) Sigma The estimate for the common standard deviation about the regression line can be found on the Model A printout as Root MSE and the value is 0.26851.

C) R^2 The value for the coefficient of determination can be found on the SAS output for model A. It's value is 0.8528. This means that approx. 85% of the variation in gpa scores can be explained by the predictors, *satm, satv, hsm, & hse.*

D) Cross Validation The summary statistics for the cross validation can be found on page 12 of the document. Here we have $predictedbias = 0.009052918$, $rpmse = 0.31756$, and mean_square_bias = 0.10084. The bias, is small, close to zero, this means our model doesn't have hardly any bias as far as predicted values being close to the observed values. The rpmse value, basically predicted $\hat{\sigma}$, tells us how far off we can expect to be, on average. Here, in gpa, we are off approx. 0.32 points.

4 Model Assumptions

A) Linearity We address this assumption by examining the scatterplot matrix and the correlation coefficients. This assumption appears to hold up well for all the predictors. There appears to be a lineear relationship between them and *gpa*. *hsm* and *hse* being less linear than the others.

B) Independence For this assumption we can use the same plots/summary stats as Linearity. The difference is we are examining multicolinearity among the predictors. There appears to be some. The correlation coefficient between satm and hse is quite high, 0.663. Also, we may need be concerned about satm & hsm, cor = 0.559. Independence is suspect for a few of our predictors. In fact, all the predictors correlation with one another is quite high except hsm & hse.

C) Equality of Variances For this assumption we use all the residual plots for model A on page 2. We visually inspect for any patterns or structure. Basically we'd like the plots to appear to be a SRS from a normal population. All the plots look OK excpet perhaps *hsm* plot. There may be heterscedasticity to be concerned about.

D) Normality This assumption is best examined by inspecting the qq plot and/or the histogram of the model residuals. Both these plots look OK and this assumption is valid.

5) Prediction

$$gpa = \hat{\beta}_0 + \hat{\beta}_1(665) + \hat{\beta}_2(575) + \hat{\beta}_3(2.86) + \hat{\beta}_4(3.05)$$

$$3.0271928 = 0.1615496 + 0.0020102(665) + 0.0012522(575) + 0.1894402(2.86) + 0.0875637(3.05)$$

Here is the predicted value, given the above inputs, with a 95% prediction interval.

fit	lwr	upr
3.027193	2.406578	3.647807

.

6) Model B

A) Coefficients Here we write our regression model:

$$gpa = \beta_0 + \beta_1 satm + \beta_2 satv + \beta_3 hsm + \varepsilon_i$$

and the equation

$$gpa = 0.3342498 + 0.0021849 satm + 0.0013123 satv + 0.1798702 hsm$$

where *satm* and *satv* $\beta_i s$ are significant, again, at $\alpha = 0.05$. *hsm* is not significant at the same α. Here, for every unit increase in SAT math scores, we expect a 0.002185 increase in *gpa*, holding all other variables constant.

B) F test

$$F = \frac{\frac{SSE(RM)-SSE(FM)}{k-m}}{\frac{SSE(FM)}{n-k-1}}$$

$$0.2461373 = \frac{\frac{1.0992453 - 1.0814988}{1}}{\frac{1.0814988}{15}}$$

The t-stat from model A for the *hse* variable is 0.4961223, which, when squared, 0.2461373, is approximately equal to our calculated F statistic. Comparing our F statistic with the critical value from the F distribution, 4.5430772, we conclude that removing *hse* from our model does not adversly affect the explanatory power of our reduced model.

C) Sigma The estimate for $\hat{\sigma}$ from model B is 0.26211.

D) R^2 The coef of determination is 0.8504. This means that approx 85% of the variation in *gpa* can be explained by the predictors *satm*, *satv*, and *hsm*.

E) Cross Validation Here we have $predictedbias = 0.002999431$, $rpmse = 0.30263$, and $mean_square_bias = 0.091584$. The bias, is small, close to zero, and less than the bias from the model A cross validations. The rpmse value, basically predicted $\hat{\sigma}$, tells us how far off we can expect to be, on average. Here, in gpa, we are off approx. 0.30 points. This is slightly less than the previous model.

7) Model Assumptions

A) Linearity This assumption is valid, same as above (scatterplot matrix and correlation coefs), only we have now removed the least linear variable from the model. So we are even better off with the assumption of linearity.

B) Independence For this assumption we can use the same plots/summary stats as Linearity. The difference is we are examining multicolinearity among the predictors. There appears to be some only now we've removed the variable that was most concerning, *hse*. Independence assumption appears to be more valid now.

C) Equality of variances For this assumption we use all the residual plots for model B. We visually inspect for any patterns or structure. Basically we'd like the plots to appear to be a SRS from a normal population. All the plots look OK excpet perhaps *hsm* plot. There may be heterscedasticity to be concerned about.

D) Normality The normaility assumption appears valid. We examine the QQplot and the histogram of the model residuals.

8) Prediction with Model B

$$gpa = \hat{\beta}_0 + \hat{\beta}_1(665) + \hat{\beta}_2(575) + \hat{\beta}_3(2.86)$$

$$3.0562062 = 0.3342498 + 0.0021849(665) + 0.0013123(575) + 0.1798702(2.86)$$

Here is the predicted value, given the above inputs, with a 95% prediction interval.

fit	lwr	upr
3.056206	2.465948	3.646465

This result compares very well with the above prediction using model A except each value has shifted up slightly. The interval for this prediction is a little more narrow than the previous one, model A.

9) Model A vs. B

The model assumptions for B vs. A appear to be more valid. We don't have as many concerns with the linearity and independence assumptions, since we've excluded the top violator in model A from model B. The parameter estimates from both models are very similar, excluding *hse* of course. $\hat{\sigma}$ from both models are very similar as well, model B having a slightly smaller value. R^2 from both models are almost equal, model A value being slighly larger. The predicted bias is smaller for model B. For these reasons, plus model B is simpler, I would select model B. It predicts just as well as model A with slightly smaller bias and variation.

10) Model C

A) Coefficients Here we write our regression model:

$$gpa = \beta_0 + \beta_1 satm + \beta_2 satv + \varepsilon_i$$

and the equation

$$gpa = 0.5071417 + 0.0026056 satm$$

where *satm* and *satv* $\beta_i s$ are significant, again, at $\alpha = 0.05$. Here, for every unit increase in SAT math scores, we expect a 0.00261 increase in *gpa*, holding *satv* constant.

B) F test

$$F = \frac{\frac{SSE(RM)-SSE(FM)}{k-m}}{\frac{SSE(FM)}{n-k-1}}$$

$$4.2085407 = \frac{\frac{1.3883839-1.0992453}{1}}{\frac{1.0992453}{16}}$$

The t-stat from model B for the *hsm* variable is 2.0514728, which, when squared, 4.2085407, is approximately equal to our calculated F statistic. Comparing our F statistic with the critical value from the F distribution, 4.4939985, we conclude that removing *hsm* from our model does not adversly affect the explanatory power of our reduced model C, although it's a close call.

C) Sigma The estimate for $\hat{\sigma}$ from model C is 0.28578.

D) R^2 The coef of determination is 0.811. This means that approx 80% of the variation in *gpa* can be explained by the predictors *satm* and *satv*.

E) Cross Validation Here we have $predictedbias = 0.002418093$, $rpmse = 0.31566$, and $mean_square_bias = 0.099643$. The bias, is small, close to zero, and less than the bias from the models A and B. The rpmse value, basically predicted $\hat{\sigma}$, tells us how far off we can expect to be, on average. Here, in gpa, we are off by approx. 0.316 points. This is slightly more than model B.

11) Model Assumptions

A) Linearity This assumption is valid, same as above (scatterplot matrix and correlation coefs), only we have now removed the least linear variables from the model. So we are even better off with the assumption of linearity than with models A & B.

B) Independence For this assumption we can use the same plots/summary stats as Linearity. The difference is we are examining multicolinearity among the predictors. There appears to be some only now we've removed the variable that was most concerning in model B, *hsm*. Independence assumption appears to be more valid now.

C) Equality of variances For this assumption we use all the residual plots for model C. We visually inspect for any patterns or structure. Basically we'd like the plots to appear to be a SRS from a normal population. This assumption appears to be valid.

D) Normality The normaility assumption appears valid. We examine the QQplot and the histogram of the model residuals. Nothing that need concern us here.

12) Prediction with Model C

$$gpa = \hat{\beta}_0 + \hat{\beta}_1(665) + \hat{\beta}_2(575) + \hat{\beta}_3(2.86)$$
$$3.144997 = 0.5071417 + 0.0026056(665) + 0.0015741(575) + 0.2667266(2.86)$$

Here is the predicted value, given the above inputs, with a 95% prediction interval.

fit	lwr	upr
3.144997	2.512291	3.777703

This result compares very well with the above prediction using model A except each value has shifted up slightly. The interval for this prediction is a little wider than model A and B but closer to A. Based on this result I would prefer model B since it's prediction interval is more narrow.

13) Model A vs. B vs. C

Model C wins based on model assumptions. It has the least concerns. But, it has the lowest R^2 value of all three models. Model B leaves in one variable that is boarderline important. Meaning, it provides more explanation for the variaiton in *gpa* than what we get by leaving it out. This, i believe, is why model B wins for prediction accuracy, *the smallest prediction interval.* Even though the F test for model B and C did not reject the null hypothesis, I prefer model B. It not only had the smallest prediction interval, it also had the lowest *rpmse* from the cross validation tests.

14) Chosen Model

A) Important factors Model B does well identifing the most important factors relating to first year college gpa, *satm* and *satv*. While at the same time it including a variable, *hsm*, that helps reduce $\hat{\sigma}$ giving us a smaller prediction interval.

B) Accurate Predictions Model B wins hands down for being the most accurate. It had the smallest bias, rpmse, and $\hat{\sigma}$ of all three models. It has has the largest adjusted R^2 value at 0.8223.