In 1982, SAT scores were first published on a state-by-state basis. Average scores ranged from 790 (out of 1600) (S. Carolina) to 1088 (Iowa). The variability among states suggested large differences in the effectiveness of public school systems, and so the data were used to support various contradictory claims about secondary education. In response, Powell and Steelman\(^1\) attempted to use the data to identify some of the variables associated with state-to-state variation, and as they said, *assess the extent to which the compositional/demographic and school-structural characteristics are implicated in SAT differences.*

These data are sometimes called ecological data. Such data are aggregate statistics—usually means. The data (means) are not observations on individuals (in this case, they are observations on states). Inferences drawn about *individuals* from these data are potentially misleading or incorrect. In this example, inferences can only be valid for the population of states and do not apply to individual high school students.

Powell and Steelman obtained measurements on the following variables for each state:

1. \(x_{\text{takers}}\) = percentage of total eligible students that took the test,
2. \(x_{\text{income}}\) = median family income of the test-takers (hundreds of dollars),
3. \(x_{\text{years}}\) = average number of years that the takers took formal courses in social sciences, humanities, and natural sciences,
4. \(x_{\text{public}}\) = percentage of takers that attended public secondary school,
5. \(x_{\text{expend}}\) = total state expenditure on secondary schools (hundreds of dollars per student), and
6. \(x_{\text{rank}}\) = median percentile ranking of the test takers among the students in their classes.

---

The figure below plots rank against percent takers.

Midwestern states tend to have low percentages of test takers, principally because most midwestern state universities required that applicants take the ACT test. Presumably, most of the students that took the SAT in these midwestern states were interested in attending college out-of-state. These differences among states may be viewed as inducing selection bias; in other words, the state means were obtained from different populations because of self-selection (students decide on whether they take the SAT). These comments imply that the percentage of test takers may have a substantial role in explaining variation in state SAT means but provide little if any insight into the school-structural characteristics responsible for SAT differences.

It’s expected that median class rank is an important variable for explaining variation in mean state SAT scores. In addition, it is strongly correlated with percent takers ($r = -.943$), and so it may be that both variables may or may not be statistically significant if a fitted model contains both. Both variables reflect self-selection. Differences between states with respect to these variables will confound a comparison of states that is aimed at understanding the success (or lack thereof) of states at preparing students for post-graduate education. The other potential explanatory variables (income, years, expenditure and public) are far more interesting with respect to the characteristics of states that succeed in preparing students for post-graduate education.

Objectives of the multiple regression analysis

1. Determine which variables are associated with state SAT scores after accounting for selection bias.

2. Determine the ranks of the states after adjusting for selection bias.

3. Determine which states’ students perform the best for the amount of money they spend.
There are limitations associated with this study and the data. First, the participating students were self-selected, and the students cannot be viewed as a representative sample of all eligible students. Further, the data are observations on states, not individual students. It’s impossible to draw statistically valid inferences about the students from the aggregated state statistics. However, the data can loosely be viewed as representative of a period of time centered on the sampling year of 1982 and a process generating aggregate statistics for each state. As the aggregate statistics are computed from usually enormous numbers of students, it’s reasonable to assume that the 1982-year data will be representative of some loosely defined time period centered about 1982.

It has already been pointed out that the percentage of test-takers and median class rank of the takers vary substantially among states. Differences in percentage of test-takers probably will affect state mean SAT scores, and so comparisons of SAT means among states with different percentages of takers is unlikely to yield any concrete conclusions. Adjusting for percentage of takers cannot change the population for which inference is appropriate, but it can account for differences in the estimates that are attributable to differences among states with respect to the percentage of takers.

Issues related to many explanatory variables

For most regression problems involving many explanatory variables, researchers are strongly motivated to search among all possible sets of explanatory variables for what is incorrectly believed to be the best set.\(^2\) Two motivations are obvious: simplicity (fewer variables imply a simpler model), and statistical significance (every explanatory variable retained in the model should be statistically significant). Both motivations are good; however, model fitting should be guided by the objectives of the study rather than simply achieving parsimony and significance.

Below, three objectives (both good and bad) of variable selection are discussed along with the implications for model fitting.

Objectives of variable selection

1. Selecting a variable to adjust for its effect is a reasonable and attainable objective. Examples are adjusting for differences among states with respect to the percentage of test takers in the SAT problem and body weight in the analysis of factors associated

\(^2\)Given \(k\) explanatory variables, there are \(2^k\) possible subsets. If \(k = 10\), then there are \(2^{10} = 1024\) possible models to evaluate.
with brain weight. Body weight is an obvious but uninteresting factor. Without accounting for it, analyzing the association between brain weight and gestation time is hopelessly confounded (large-brained animals have longer gestation times simply because it takes longer for the fetus to reach full-term).

2. An objective that is not well-defined (and poor) is to identify a set of variables from among a large set that explain observed variation in the response variable mean. Without elucidating an a priori scientific rationale for including explanatory variables in a model, this objective amounts to fishing for an explanation. The objective is weak because

(a) There may be alternative ways to express or measure the same information carried by the variables, and some important variables may have been overlooked, or not measurable (e.g., educational level of parents).

(b) It is possible that there are unobserved variables that affect the response variable and are manifested through the observed explanatory variables. Then, attributing importance to the observed variables may be misleading. Gestation time presumably reflects the in-utero costs of a complex brain. If so, it is likely to be positively associated with brain weight after accounting for body size but it is simplistic to conclude that gestation time is a causal agent for brain size from biological principles.

(c) Inclusion, or exclusion, of the observed variables is affected by the correlation between them. For example, median class rank and percentage of test takers are strongly correlated and both explain variation is state SAT means. It’s possible that only one of these variables will be significant in a model containing both, but it would be incorrect to conclude that the non-significant variable is not associated with the response.

(d) If there are many variables in the final fitted model, it will often be difficult to interpret the coefficients (since the coefficient describes the estimated change in the response variable mean given a 1 unit change in the associated variable when all other variables are held fixed—a condition that may not make sense). Furthermore, the tests of significance are conditional. For example, a test of $H_0 : \beta_i = 0$ versus $H_a : \beta_i \neq 0$ is interpreted conditionally, and yields a conclusion such as $x_i$ is (is not) significant after accounting for variables $x_1, \ldots, x_{i-1}, x_{i+1} \ldots x_{p-1}$.

3. If the objective of fitting the regression model is prediction (interpretation being of limited interest), then formulaic rules or automatic variable selection algorithms are acceptable because different subsets often provide equivalent and accurate predictions. The emphasis of model evaluation shifts from assessing statistical significance of explanatory variables to estimation of prediction error (e.g., estimating the expected
difference between a future value and its prediction (obtained from the model).
Considering these points, the objective of determining which variables are associated with state SAT scores after accounting for selection bias is not a well-defined objective and a poor use of regression.

Different questions require different approaches towards the explanatory variables

1. If the objective is to compare the educational abilities of students (at the state level), then a simple comparison of the state SAT means is adequate provided that the SAT means accurately reflect the mean ability level of all students. The aim of a regression analysis is to determine whether the unadjusted means are representative, or whether some adjustment is needed to compensate for self-selection. If adjustment is needed, then the regression analysis provides the adjustment.

Suppose that there is a substantial effect of percentage of test takers and median class rank on the response variable mean. Then a comparison of two states with different levels can be accomplished fitting a model with these variables and comparing the residuals for each state. The residuals are, in essence, what is not explained by the model after accounting for differences in the explanatory variables. Thus, the differences between residuals is a comparison that has accounted for differences due to percentage of test takers and median class rank. Other explanatory variables don’t belong in the model.

2. The impact of state expenditures can be assessed if all useful variables are included in the model so that the estimated effect of expenditure does not reflect other variables correlated with expenditure.

3. If prediction is the objective, then all variables should be considered. Prediction is not a reasonable objective in this situation, though.

A strategy for dealing with many explanatory variables

1. Identify the key objectives.

2. Review the available variables and eliminate those variables that are weakly related to the objectives and those that are obviously redundant.

3. Perform exploratory data analysis to understand how the variables are related pair-wise using graphical displays and correlation coefficients.

4. Perform transformations as needed. For example, a pattern of non-constant variance between the response and the explanatory variables suggests a logarithmic transformation of the response variable may be necessary to meet the necessary conditions of hypothesis tests and confidence intervals.
5. Examine residual plots after fitting a rich model (i.e., containing variables that appear to be potentially important explanatory variables). Perform additional transformations and investigate outliers.

6. Select a suitable set of explanatory variables. Retain enough control of the variable selection procedure to insure that the objectives are met.

7. Address the objectives using the fitted model.

Model residuals will play an important role in this analysis since adjusting for percent takers and median class rank is accomplished by fitting a model of SAT means to percent takers and median class rank. The fitted values $\hat{y}_1, \ldots, \hat{y}_{50}$ are the estimated expected SAT means given percent takers and median class rank. The residuals $y_1 - \hat{y}_1, \ldots, y_{50} - \hat{y}_{50}$ measure how much better or worse a state’s students have performed than expected (based on percent takers and median class rank).

**Preliminary analysis of the state SAT scores**

The matrix scatterplot shows that SAT scores are strongly and negatively associated with percent takers and positively associated with class rank. Percent takers and rank are also strongly and negatively associated. No other variable pairs are strongly correlated.

There is one state with large leverage associated with expenditure and another with large leverage associated with the percent public schools variable (below and left). In the figure below and right, the residuals from the regression of state SAT mean on percent takers and median class rank are plotted against expenditure. The regression of the residuals and expenditure with and without Alaska are graphed.
The figure strongly suggests that expenditure is potentially important and Alaska is influential.

Turning to the problem of adjusting for differences among states with respect to state mean SAT on percentage test takers and the median class rank of the takers, Table 1 summarizes the regression of state mean SAT on percentage test takers and the median class rank of the takers. Median class rank is not statistically significant, presumably because of the strong association between percentage test takers and the median class rank. The lack of statistical significance found in the multiple regression model does not imply that median rank doesn’t need to be controlled for; logically, it does need to be controlled for if states are to be fairly compared. It belongs in the model used for the adjusting for state-to-state differences in mean SAT scores due to differences in percentage test takers and the median class rank of the takers.

Table 1: Coefficients and standard errors obtained from linear regression of SAT means on log-takers and median class rank. $R^2 = .815$, $n = 50$.

| Variable            | Coefficient | Std. Error | $t$-statistic | $P(T > |t|)$ |
|---------------------|-------------|------------|---------------|------------|
| Intercept           | 882.1       | 224.1      | 3.94          | .0002      |
| log-takers          | −45.19      | 14.06      | −3.21         | .0024      |
| Median class rank   | 2.40        | 2.33       | 1.03          | .3090      |
The figures above show the studentized residuals and Cook’s distance for each state. South Carolina is notable as its residual is very small; further, Cook’s distance for South Carolina is large—indicating that this state is influential. The residual plot can be used for comparing adjusted SAT means among states. However, the comparison should be based on the simple residuals $y_i - \hat{y}_i$ rather than the studentized residuals $\frac{y_i - \hat{y}_i}{\sigma \sqrt{1 - h_i}}$, each of which has a unique denominator.

The comparison of states after accounting for differences among percent takers and median class rank below retains all 50 observations since there is no clear reason to omit South Carolina. The residual plot (right) visually displays the results. In this situation, a fitted value is a model prediction of the state SAT given the observed values of percent takers and median class rank. The residual is the difference between the state mean SAT score and the prediction based upon the values of percent takers and median class rank for the state. A large residual implies that the students, on average, did better than expected given the percent takers and median class rank of the takers, and a small residual implies that on average the takers did worse than expected given the percent takers and median class rank.
class rank of the takers. The plot shows that there are several states (South Carolina, North Carolina, Georgia and Mississippi) with unusually small residuals showing that for these states, after removing the effect of differences among states with respect to percent takers and median class rank, SAT means for these states are substantially below the rest.

The best-performing states are, in order, New Hampshire, Iowa, Connecticut, Montana, Minnesota and Washington.

The fitted values (not shown) describe the state demographics with respect to percent takers and median class rank. The largest adjustments to the raw SAT means were to Iowa, Arkansas, North Dakota, and South Dakota.³

The next objective is to assess the impact of expenditure on the model (and the residuals), after accounting for any other variables that are informative (e.g., years of formal coursework). There are two observations with unusually large influence: South Carolina and Alaska (because state expenditure in Alaska was much greater than the other 49 states). Arguably, Alaska could be removed because the expenditure for Alaska may not reflect educational commitment but instead a higher cost-of-living.

The Figure below and left plots the studentized residuals against the fitted values obtained from the model containing the log or percent takers and median class rank. After accounting for self-selection, New Hampshire, Connecticut, Washington, Montana, and Iowa scores were larger than other states.

After examining a main-effects model with all variables, the public and the income variables were removed from the model as there was no evidence that either associated parameter was different from zero. The fitted model is shown in Table 2. Table 2 shows that there is conclusive evidence of an effect associated with expenditure (p-value = .0003) and the estimated effect of each additional $100 spent per student per year on secondary education is an increase in mean state SAT of 2.42 points. A 95% confidence interval for the effect is [1.17, 3.67]. The estimated effect size seems small in comparison to the estimated effect of one additional year of formal coursework (an increase in mean SAT score of 17.85 points).

After accounting for the self-selection variables, years of formal coursework and expenditure, Iowa is predicted to have the largest state mean SAT score (it’s fitted value is largest in the Figure below and right); moreover its residual is positive indicating that the observed mean was even larger than predicted.

³The original SAT mean for Iowa was 1088 and the adjusted score was 1047.3.
Table 2: Coefficients and standard errors obtained from linear regression of SAT means on log-takers, median class rank, expenditure and years (average number of years of formal coursework). $R^2 = .892, n = 50.$

| Variable            | Coefficient | Std. Error | $t$-statistic | $P(T > |t|)$ |
|---------------------|-------------|------------|---------------|-------------|
| Intercept           | 388.4       | 253.2      | 1.53          | .1321       |
| log-takers          | −38.0       | 13.0       | −2.93         | .0053       |
| Median class rank   | 4.00        | 2.06       | 1.94          | .0588       |
| Years               | 17.85       | 5.73       | 3.11          | .0032       |
| Expenditure         | 2.42        | .619       | 3.91          | .0003       |

Sequential variable selection techniques

Recall that if there are $p − 1$ variables, then the number of distinct sets of variables that can be formed is $2^{p−1}$. With only six variables, 64 distinct models that can be fit. As an alternative to fitting all possible models, a variety of sequential procedures have been developed. None are a replacement for science-driven model fitting, but they are extensively used. Sequential procedures examine fewer models than the full set in a manner that limits the set without (in principle) overlooking the better models.

Sequential procedures begin with a tentative initial model. The initial model is compared to one or more models that differ by having either one additional, or one fewer, variables. The
initial model is discarded in favor of the new model if the new one is better in some sense. The three traditional methods are

1. **Forward selection**

   The initial model is $E(y|x) = \beta_0$. Each of the following models are fit:

   \[
   E(y|x) = \beta_0 + \beta_1 x_1 + \varepsilon \\
   \vdots \\
   E(y|x) = \beta_0 + \beta_{p-1} x_{p-1}.
   \]

   The model that produces the greatest reduction in the residual sums-of-squares (compared to the initial model) is adopted (provided that the corresponding explanatory variable is statistically significant).

   Using the adopted one-variable model, attempt to introduce each of the remaining variables, one at a time and identify the variable that has a smallest p-value. If this smallest p-value is smaller than the chosen threshold value, conduct further iterations until no further variables are found to be significant. Two test statistics are used to determine whether the additional variable should be retained in the model.

   (a) If the candidate explanatory variable is quantitative and hence represented by a single term, $x_j$, then the $t$-statistic for a test of $H_0 : \beta_j = 0$ versus $H_a : \beta_j \neq 0$, given all previously selected variables, assesses the strength of evidence supporting $H_a : \beta_j \neq 0$.

   (b) If a variable is categorical (i.e., is a factor) with $k$ levels, then $k - 1$ indicator variables are needed to identify the $k$ levels. An extra-sums-of-squares $F$-statistic is used to test the significance of the factor:

   \[
   F = \frac{(SSE_{\text{reduced}} - SSE_{\text{full}})/(df_{\text{reduced}} - df_{\text{full}})}{SSE_{\text{full}}/df_{\text{full}}},
   \]

   where the full model contains the factor and the reduced model does not contain the factor. The p-value is $P(F_{df_{\text{reduced}} - df_{\text{full}}, df_{\text{full}}} > f)$ where $F_{df_{\text{reduced}} - df_{\text{full}}, df_{\text{full}}}$ is a $F$-statistic with $df_{\text{reduced}} - df_{\text{full}}$ numerator and $df_{\text{full}}$ denominator degrees of freedom and $f$ is the observed value of the statistic.

2. **Backward elimination**

   (a) Begin with a large model including all variables of potential interest, including interaction (if interaction is considered to be potentially interesting).

   (b) Remove variables sequentially based on either the $t$ or $F$-statistics (depending on whether the variable in question is quantitative or categorical).
(c) Continue until all nonsignificant variables are removed (except for main effects that are part of a significant interaction variable).

Differences between forward selection and backward elimination

Models fit via backward elimination tend to contain more terms than with forward selection (bad), but is less likely to overlook an important term (good).

Interaction is difficult to accommodate in forward selection. Suppose that $x_1$ enters the model and $x_2$ does not. It may be that the effect of $x_2$ strongly depends on the level of $x_1$. In other words, the interaction variable $x_1 \times x_2$ is important. Without $x_2$ in the model, the interaction variable $x_1 \times x_2$ will never be investigated because interaction should not be allowed in a model in the absence of a main effect ($x_2$ in this case).

3. Stepwise regression

The initial model is $E(y|x) = \beta_0$. Each iteration consists of two steps: one step of forward selection followed by one step of backward elimination (remove any variable that became non-significant as a result of adding a variable in the forward selection sub-step). The number of significance test may be substantially greater than either forward selection or backward elimination which

A hybrid strategy can pursued that uses forward selection but also considers interaction variables and their associated main effects. The motivating principles are parsimony (model simplicity) and coherency (ignoring tests of significance for main effects when interaction is important). Groups of variables can be formed based on similarity and selection carried out within groups. Then the useful variables found in the group-specific regressions can be collected and used in a second model-fitting exercise.

There are two important characteristics of the variable selection methods discussed so far: the selection of a variable is carried out by comparing nested models where two models differ only with respect to a single quantitative variable or factor, and secondly, each variable is evaluated for inclusion or removal using significance test.

Variable selection techniques (2)

Relying on significance tests for variable selection when there are many variables destroys the most useful property of the tests. The significance level of the accept/reject rule becomes

---

4 Model A is nested in model B if every variable in model A appears in model B.

5 A variable may be tested for significance (to enter or be dropped) many times.
meaningless, that is if the rule is to keep a variable in a tentative model if the p-value is less than a pre-defined level $\alpha$ (say, .05), then $\alpha$ is the probability of incorrectly including a variable that has no association with the response given the other variable included in the model.

While the test-wise (Type I) error rate is .05, the experiment-wise error rate (the probability of incorrectly adding or retaining at least one meaningless variable) grows with the number of tests conducted.

In response to this problem, some variable selection techniques avoid significance testing and instead use an alternative criterion. For example, $R^2$ can be used to compare nested models. However, $R^2 = (\text{SSR}_{\text{reduced}} - \text{SSR}_{\text{full}})/\text{SSR}_{\text{full}}$ is always larger when comparing a model with more variables to a model with fewer variables if the models are nested. Therefore, the final model will contain all available variables. Alternatively, the adjusted coefficient of determination, $\text{adj } R^2 = (\hat{\sigma}_2^{\text{reduced}} - \hat{\sigma}_2^{\text{full}})/\hat{\sigma}_2^{\text{full}}$ is better, but also tends to produce models with more terms than appropriate.

Alaike’s and Bayes information criteria (AIC and BIC)

AIC and BIC are model selection statistics that measure the purported information content of a model by contrasting a measure of lack of fit ($n \log(\hat{\sigma}^2)$) and a penalty for number of model terms. Model selection is conducted by comparing the AIC or BIC of two models, one of which is nested within the other. The model with the smaller criterion value is selected. The criteria are

$$
\text{AIC} = n \log(\hat{\sigma}^2) + 2p \\
\text{BIC} = n \log(\hat{\sigma}^2) + p \log(n).
$$

The penalty associated with number of model terms increases with $n$ for BIC but is independent of sample size $n$ for AIC. For large sample sizes, AIC is somewhat more liberal and tends to produce models with more terms than BIC. A purported advantage of AIC and BIC is that no formal hypothesis testing is conducted thereby side-stepping criticisms about uncontrolled experiment-wise Type I error rates that apply to forward selection, backward elimination and step-wise regression.

However, model selection using AIC and BIC does not avoid the problem of uncontrolled and unknown experiment-wise error rates since variable selection requires decisions regarding the inclusion of variables in a models. Every time a decision is made whether to include or exclude a variable, an error might be made. More decisions implies that the number of

---

6The significance level $\alpha$ is the Type I error rate of the test.
7forward selection
8backward elimination
possible errors increases. The criterion used to make the decision does not alter the fact that errors will accumulate as additional decisions are made.