Regression 551 Midterm

Matthew Stoebe

2024-09-21

#Question 1 Discuss issues of reliability and validity for this example, giving two sentences for each.

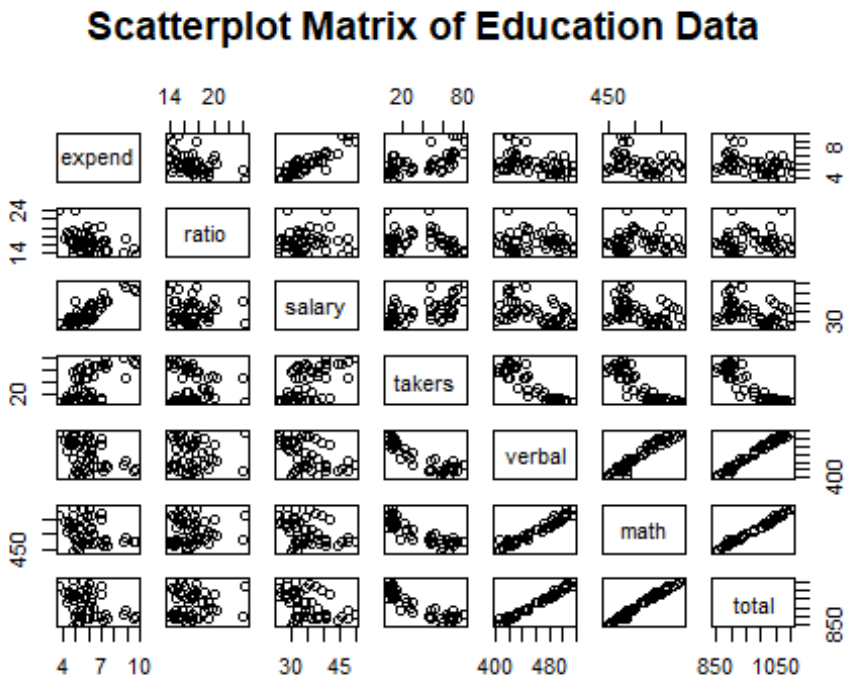
1. Reliability: The reliability of this data depends on the consistency of data collection methods across all states. Since educational expenditure and SAT scores are likely collected and reported using standardized government procedures, we can assume the data is reliable.

2. Validity: The validity of the data pertains to how well the variables measure what they are intended to measure. For example, the percentage of students taking the SAT varies by state, which may affect the validity of using average SAT scores to compare educational achievement across states.

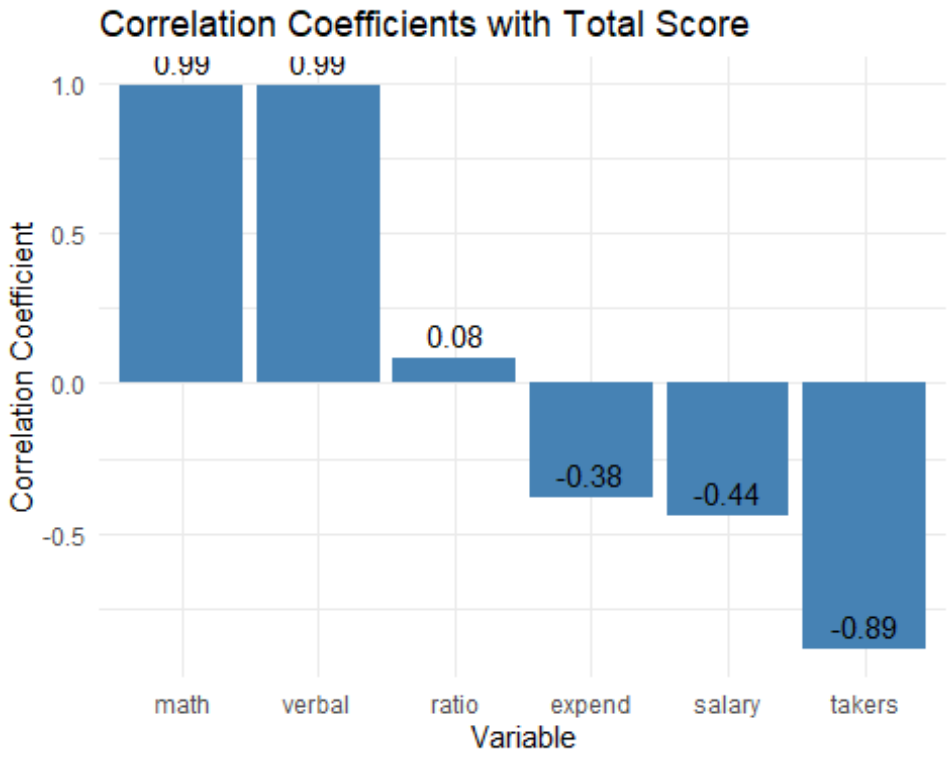
This actually manifests itself in the data below where you see that states with higher participation rates have lower test scores. It is plausible that in states with low partition rates, the students included to take standardize testing are high achievers leading to inflated test scores. This would indicate that test scores may not be the best measure of student success in all cases.

#Question 2 Make a grid of graphs exploring your data, along with a paragraph describing what you have learned. Add another paragraph explaining how you have used the principles of statistical graphics in making your plots, and another paragraph discussing what important aspects of the data you were not able to include in these graphs. Your graphs should be communicate key features of the data clearly.

```
data_path <- "./Other Data/Midterm/gruber.csv"

data <- read.csv(data_path)


pairs(data[, c("expend", "ratio", "salary", "takers", "verbal", "math",
"total")],
      main = "Scatterplot Matrix of Education Data")
```




```r
correlations <- sapply(data[, c("expend", "ratio", "salary", "takers",
"verbal", "math")], function(x) cor(x, data$total))

# Convert to a data frame for plotting
correlation_data <- data.frame(
  Variable = names(correlations),
  Correlation = correlations
)

correlation_data <- correlation_data[order(-correlation_data$Correlation), ]

ggplot(correlation_data, aes(x = reorder(Variable, -Correlation), y =
Correlation)) +
  geom_bar(stat = "identity", fill = "steelblue") +
  geom_text(aes(label = round(Correlation, 2)), vjust = -0.5) +
  theme_minimal() +
  labs(title = "Correlation Coefficients with Total Score",
       y = "Correlation Coefficient",
       x = "Variable")
```



Some things to call out: 1. Verbal and Math Test scores have very high correlation with each other. This would indicate that the sometimes perceived spread between students who are "good at math" and "good at communication" may not actually be as strong as people think. Students with high verbal scores tend to have high math scores, and students with low verbal scores tend to have low math scores.

2. As expected, Verbal scores and Math scores are highly correlated with total scores. This indicates that in building a regression model, we should not choose both the component scores and the total as this may lead to issues with colinearity. This same argument could be made for Verbal and Math Scores, but not as strongly since there is some variation.

3. There is little between expenditure/ teacher salary and test performance (verbal, math, total). However, More spending does seem to lead to a higher participation rate in standardized testing

4. As expected, there is some positive correlation between expenditure and salary which indicates that teachers in states with higher expenditure tend to get paid more, this increase in expenditure is distributed differently between teacher salaries and other programs.

5. There is negative correlation between the percentage of students taking the SAT, and the scores received on the SAT. This is very interesting. There may be some bias in WHO takes the test by state where states with high expenditure / high participation rates have many lower performing students actually taking the test whereas those students may not be taking the test in lower expenditure states.

##Use of Statistical Graphics Principles In creating these plots, I applied principles such as:

- Clarity: Using scatter plots and bar charts with no frills makes it easy to identify relationships between pairs of variables.
- Context: Including variable names and units where appropriate to provide context.

##Limitations Limitations of the charts above

- The plots do not indicate causation, only correlation.
- Factors such as socioeconomic status or education policies are not accounted for.
- We are not accounting for interactions between different features. Only 1:1 relationships
- Data Distribution: The plots do not represent distribution of individual variables well

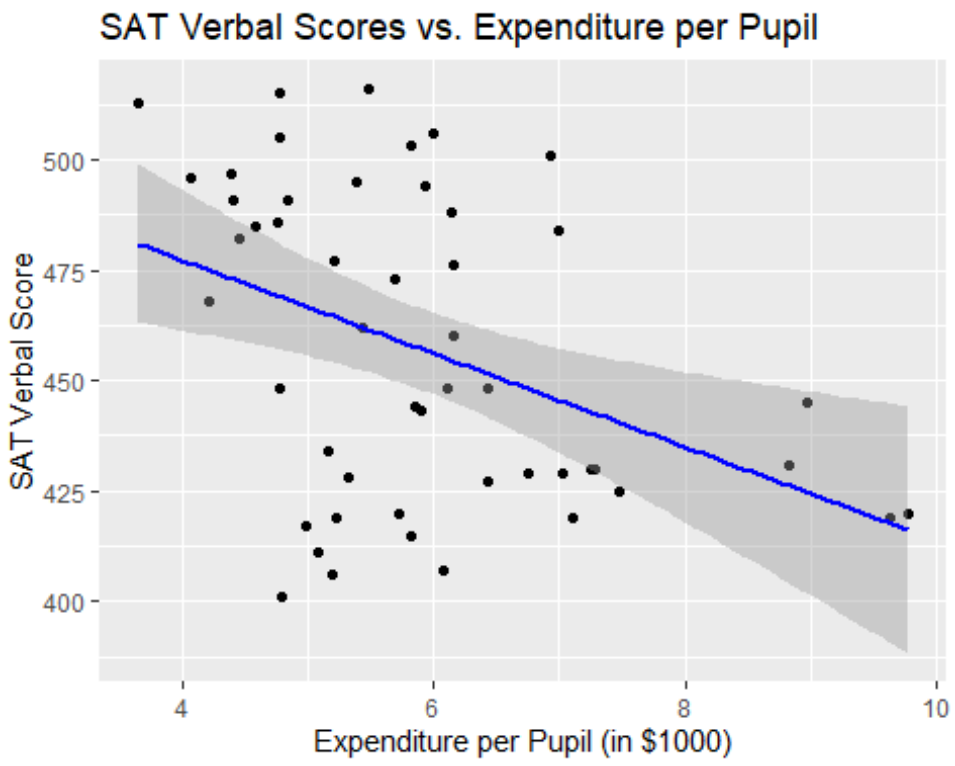
Question 3

```
# Fit the model
model <- stan_lm(verbal ~ expend, data = data, prior = NULL, prior_intercept
= NULL,
                 prior_aux = NULL, prior_PD = FALSE, refresh = 0)

summary(model)

##
## Model Info:
##  function:     stan_lm
##  family:       gaussian [identity]
##  formula:      verbal ~ expend
##  algorithm:    sampling
##  sample:       4000 (posterior sample size)
##  priors:       see help('prior_summary')
##  observations: 50
##  predictors:   2
##
## Estimates:
##                 mean   sd    10%   50%   90%
## (Intercept)   519.7   19.1 495.1 520.3 543.8
## expend        -10.6    3.1 -14.5 -10.7  -6.5
## sigma          32.6    3.3  28.6  32.3  36.9
## log-fit_ratio  -0.3    0.1  -0.5  -0.3  -0.1
## R2              0.4    0.2   0.1   0.3   0.7
##
## Fit Diagnostics:
##            mean   sd    10%   50%   90%
## mean_PPD 457.1    6.4 449.3 457.0 465.2
##
## The mean_ppd is the sample average posterior predictive distribution of
the outcome variable (for details see help('summary.stanreg')).
```

```
##
## MCMC diagnostics
##                mcse Rhat n_eff
## (Intercept)    0.6  1.0  1068
## expend         0.1  1.0  1088
## sigma          0.1  1.0  1316
## log-fit_ratio 0.0  1.0  1041
## R2             0.0  1.0   857
## mean_PPD       0.1  1.0  3472
## log-posterior 0.0  1.0   955
##
## For each parameter, mcse is Monte Carlo standard error, n_eff is a crude
measure of effective sample size, and Rhat is the potential scale reduction
factor on split chains (at convergence Rhat=1).
```

```
ggplot(data, aes(x = expend, y = verbal)) +
  geom_point() +
  geom_smooth(method = "lm", se = TRUE, color = "blue") +
  labs(title = "SAT Verbal Scores vs. Expenditure per Pupil",
       x = "Expenditure per Pupil (in $1000)",
       y = "SAT Verbal Score")
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



##Interpretation of Parameters

Intercept: The estimated intercept, 519.8 is the expected SAT verbal score when the expenditure per pupil is zero. Since zero expenditure is not realistic, the intercept serves as a baseline for the model but is not of practical interest.

Slope : The estimated slope indicates that for each additional $1,000 spent per pupil, the average SAT verbal score decreases by approximately -10.6 points, suggesting a negative association between spending and verbal scores.

Uncertainties: The standard errors provide the uncertainty in the parameter estimates.

#Question 4

```
new_data <- data.frame(expend = 8)
head(new_data)
```

```
##   expend
## 1      8
```

```
# Point estimate and uncertainty for the mean response
pred_mean <- predict(model, newdata = new_data, se.fit = TRUE)

# Display the results
cat("Predict:\n")
```

```
## Predict:
```

```
cat("Estimated mean SAT verbal score:", round(pred_mean$fit, 2), "\n")
```

```
## Estimated mean SAT verbal score: 435
```

```
cat("Standard error:", round(pred_mean$se.fit, 2), "\n")
```

```
## Standard error: 8
```

```
# Posterior distribution of the linear predictor (mean response)
linpred_samples <- posterior_epred(model, newdata = new_data)

# Summarize the posterior samples
linpred_mean <- mean(linpred_samples)
linpred_sd <- sd(linpred_samples)

cat("Posterior_linpred:\n")
```

```
## Posterior_linpred:
```

```
cat("Mean of posterior predictive distribution:", round(linpred_mean, 2),
"\n")
```

```
## Mean of posterior predictive distribution: 435
```

```
cat("Standard deviation:", round(linpred_sd, 2), "\n")
```

```
## Standard deviation: 8
```

```r
# Posterior predictive distribution (includes residual error)
predict_samples <- posterior_predict(model, newdata = new_data)

# Summarize the posterior predictive samples
predict_mean <- mean(predict_samples)
predict_sd <- sd(predict_samples)

# Display the results
cat("Posterior_predict:\n")

## Posterior_predict:

cat("Mean of posterior predictive distribution:", round(predict_mean, 2),
"\n")

## Mean of posterior predictive distribution: 434.21

cat("Standard deviation:", round(predict_sd, 2), "\n")

## Standard deviation: 33.63
```

Discussion

The main difference appears between posterior_linepred and posterior_pred in that the standard deviation is much higher for the latter. This is because linepredict is focused on predicting the the expectation or average and its stdev while posteriorpredict is predicting the individual point. This leads to much higher error on the individual prediction. Point estimates are the same accross all 3 predictions.

The Standard deviations in both cases are reasonable, but I do not believe that this data set as a whole is a good method of predicting test scores. While it does show some interesting relationships, the issues related to participation skewing the students taking tests, in my mind brings the broader analysis into question.

Additionally, data for one year is limited and I think a more interesting question would be to look at significant changes in test scores and participation rates over time and search to understand what caused them. It is also possible that comparing states is not the most effective method for forming opinions as there are too many confounding factors to make it an apples to apples comparison.Instead, we should benchmark states to previous performance and look for things that helped improve or worsen test scores. These trends could then be abstracted across states to see if the same types of effects hold true.