Homework 4

STAT3530 Applied Linear Models

Due Thursday, October 8, in class.

You may collaborate, but write up your own solutions. Submit a paper copy of this document with handwritten solutions below each prompt. Please attach any plots you are asked to make and label them by problem number.

Each dataset loads from the course site; the line is given at the start of each problem.

Occupational prestige

In class we fit a model for the `prestige` data: 102 Canadian occupations rated for prestige in 1971, paired with the average `education` in years, the average `income` in dollars, and the percentage `women` in each occupation. We found that income belongs on the log scale, and checked that the resulting model satisfies the assumptions, so take the model

$$\text{prestige}_i = \beta_0 + \beta_1 \,\text{education}_i + \beta_2 \log(\text{income}_i) + \beta_3 \,\text{women}_i + \epsilon_i$$

as correctly specified.

```
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/prestige.RData"))

fit <- lm(prestige ~ education + log(income) + women, data = prestige)
```

1. Interpret each of the three partial significance tests. For each predictor, say what is being held fixed and what the test does or does not establish.

   i. `education`

   ii. `log(income)`

   iii. `women`

2. Report a 95% confidence interval for each coefficient and interpret it in context.

   i. `education`

   ii. `log(income)`. State what the interval means on the original dollar scale rather than the log scale.

   iii. `women`

3. Plot leverage, studentized residuals, and Cook's distance, and attach them. The two most influential occupations are `ministers` and `babysitters`. Fill in the table, using $h_{ii} > 2p/n$ for high leverage and $|r_i| > 2$ for an outlier.

   To pull out the values for one observation, find its row number and index the diagnostic by it:

```
k <- match(c('ministers', 'babysitters'), prestige$occupation)   # row numbers

hatvalues(fit)[k]
rstandard(fit)[k]
cooks.distance(fit)[k]
```

	h_{ii}	r_i	D_i	high leverage?	outlier?
`ministers`	_____	_____	_____	_____	_____
`babysitters`	_____	_____	_____	_____	_____

4. Explain *why* each occupation is unusual in the way the table says it is. The two are unusual for different reasons.

   Look at the two rows themselves, and compare them to the rest of the data:

```
prestige[k, ]                # the rows you found in (3)

summary(prestige)            # for comparison
```

   i. `ministers`

   ii. `babysitters`

5. Refit the model three times: without `ministers`, without `babysitters`, and without both. Report the coefficient on `women` and its p-value each time.

	$\hat{\beta}_{\text{women}}$	p-value
all 102 occupations	_____	_____
without `ministers`	_____	_____
without `babysitters`	_____	_____
without both	_____	_____

6. Does the inference on `women` change? Is the conclusion sensitive to one of these occupations or to both together?

7. Given your answer to (6), what would you report to someone who asked whether the percentage of women in an occupation is associated with its prestige?

Educational expenditure and SAT scores

The `sat` data record, for each of the 50 states in 1994–95, average spending per pupil `expend` in thousands of dollars, the pupil/teacher `ratio`, the average teacher `salary` in thousands, the percentage `takers` of eligible students who sat the exam, and the average `total` SAT score.

```
load(url("https://tdruiz-stat3530.share.connect.posit.cloud/_data/sat.RData"))
```

A state legislator claims that spending more money on schools does not improve outcomes.

8. Regress `total` on `expend`, checking that the model is well specified. Report the estimated slope and its p-value, and state what a person would conclude from this fit alone.

9. Now adjust for `takers`, again checking specification. Attach the diagnostics for your final model and report the adjusted association between spending and SAT total.

10. The two answers disagree. Plot `takers` against `expend`, attach your plot, and use it to explain why.

11. Use your model from (9) to predict the mean SAT total for a state spending \$6,000 per pupil where 30% of students take the exam, and again for a state spending \$7,000 per pupil where 30% take the exam.

    i. Report both predictions with 95% intervals.

    ii. Compare the difference between them to the coefficient you reported in (9). Can you explain why they are identical?

12. Estimate the root mean square prediction error of the model in (9) and interpret it in context.

13. Add `salary` and `ratio` to the model and compute the same quantity. Do predictions improve?

14. Now compare the same two models in sample, using the average squared residual. Is it the same story? Explain which comparison you would trust and why.