Homework 5

STAT3530 Applied Linear Models

Due Thursday, October 15, 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 or output you are asked to produce and label them by problem number.

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

Seat position

The car seat data is familiar from class. For each of 38 drivers, observations comprise the horizontal distance `hipcenter` of the hips from a fixed point in the car, along with `Age`, `Weight`, and six body measurements: `HtShoes`, `Ht`, `Seated`, `Arm`, `Thigh`, and `Leg`.

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

full <- lm(hipcenter ~ ., data = seatpos)
```

Suppose an engineer designing a seat explains to you that an intern who previously analyzed this data concluded that body measurements are helpful in explaining horizontal position of the seat, but claimed that since body measurements are all related there's no way to determine which ones are important and it's best to just collect as many as possible. The engineer asks you whether you agree and if not, what data you think they should collect and why.

You decide to consider a **reduced model** that keeps one measure of overall length, one measure of girth that is least related to length (`Thigh`, correlated 0.73 with `Ht`), and `Age`, which is essentially unrelated to body size:

$$\texttt{hipcenter} = \beta_0 + \beta_1\,\texttt{Age} + \beta_2\,\texttt{Ht} + \beta_3\,\texttt{Thigh} + \epsilon$$

1. Test whether the five dropped predictors matter. Write the null and alternative hypotheses, numerator and denominator degrees of freedom for the F test, and report the conclusion in context.

2. Compare the partial inference on height (without shoes) in the full and reduced models: report the conclusion of each test in context. If the models give different results, explain why.

3. In the reduced model, height is now the most significant predictor. Test whether `Age` and `Thigh` add anything to a model that already has `Ht`. Write the null and alternative hypotheses, numerator and denominator degrees of freedom for the F test, and report the conclusion in context.

4. Now compare the models on prediction rather than on significance. Estimate the root mean square prediction error of each model and fill in the table below. Does collecting more than one body measurement appear to help?

   Recall that estimating RMSPE by a leave-one-out average does not require refitting:

```
rmspe <- function(m) sqrt(mean((resid(m) / (1 - hatvalues(m)))^2))
```

model	RMSPE
full, all eight predictors	_____
reduced, `Age + Ht + Thigh`	_____
`Ht` alone	_____

5. Does it matter *which* body measurement you collect? Fit a model with `Thigh` as the only predictor and compute its RMSPE. Compare with the `Ht`-only model.

6. In a few sentences, provide a recommendation and justification for the engineer.

Savings rates

The savings data record, for each of 50 countries averaged over 1960–1970, the personal savings rate `sr`, the percentages `pop15` and `pop75` of the population under 15 and over 75, per-capita disposable income `dpi`, and its percentage growth rate `ddpi`. The first two describe the age structure of a population; the last two describe its economy.

Fit a model with the four predictors standardized so that they are on comparable scales:

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

sv <- data.frame(sr = savings$sr,
                 scale(savings[, c('pop15', 'pop75', 'dpi', 'ddpi')]))

fit <- lm(sr ~ ., data = sv)
```

Write $\beta = (\beta_0, \beta_{15}, \beta_{75}, \beta_{\texttt{dpi}}, \beta_{\texttt{ddpi}})^T$ for the coefficient vector of this model. For each problem below, answer the question by testing a contrast of the form $c^T\beta$; write the null and alternative hypotheses, write the vector c^T of contrast coefficients, and report the conclusion of the test in context.

7. Do the two demographic variables matter equally?

8. Do the two economic variables matter equally?

9. Do age structure and the economy matter equally? (The question is whether the demographic effects together equal the economic effects together.)

10. Compute and interpret a point estimate for the contrast from (9).

Prostate cancer

Ninety-seven men about to undergo surgery for prostate cancer. The response `lpsa` is the log of prostate specific antigen, a protein measured from a blood sample and used to track the disease. Everything else requires tissue, so understanding what drives PSA is of practical interest.

Three of the predictors describe the **patient**: `lweight` (log prostate weight), `age`, and `lbph` (log amount of benign prostatic hyperplasia, a non-cancerous enlargement). Five describe the **tumor**: `lcavol` (log cancer volume), `lcp` (log capsular penetration, how far the cancer has grown through the wall of the prostate), `svi` (whether it has invaded the seminal vesicles, 0 or 1), `gleason` (a 6–9 grade of how abnormal the cells look), and `pgg45` (the percentage of tissue graded 4 or 5).

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

fit <- lm(lpsa ~ ., data = prostate)
```

11. Test whether any of the patient or tumor measurements explain (log) PSA. State the null and alternative hypotheses, degrees of freedom for the F test, and report the result in context.

12. Test whether the three patient variables explain (log) PSA.

    i. Write the null and alternative hypotheses, write the degrees of freedom for the F test, and report the result in context.

    ii. Compare and contrast the inference in (i) with inference by partial t tests. Are all of the variables individually significant? If not, is this a discrepancy? Explain in a sentence or two.