NCSU ST 503 Discussion 5

Probem 4.5 Faraway, Julian J. Linear Models with R CRC Press.

Bruce Campbell

Comparing models of body fat measurement

For the fat data used in this chapter, a smaller model using only age, weight, height and abdom was proposed on the grounds that these predictors are either known by the individual or easily measured.

(a) Compare this model to the full thirteen-predictor model used earlier in the chapter. Is it justifiable to use the smaller model?

$brozek \sim age + weight + height + abdom$

```
##          term      estimate std.error  statistic      p.value
## 1 (Intercept) -32.769635854 6.54190241 -5.0091906 1.041540e-06
## 2         age  -0.007051258 0.02434164 -0.2896789 7.723049e-01
## 3      weight  -0.123721774 0.02504553 -4.9398736 1.441726e-06
## 4      height  -0.116693958 0.08272693 -1.4105921 1.596231e-01
## 5       abdom   0.889704097 0.06726722 13.2264134 1.492006e-30
```

rsquared
0.7211

$brozek \sim age + weight + height + neck + chest + abdom + hip + thigh + knee + ankle + biceps + forearm + wrist$

```
##           term     estimate   std.error   statistic      p.value
## 1  (Intercept) -15.29254907 16.06992071 -0.95162567 3.422523e-01
## 2          age   0.05678616  0.02996465  1.89510481 5.929042e-02
## 3       weight  -0.08030986  0.04958051 -1.61978675 1.066023e-01
## 4       height  -0.06460028  0.08893033 -0.72641448 4.682985e-01
## 5         neck  -0.43754090  0.21533372 -2.03192006 4.327265e-02
## 6        chest  -0.02360333  0.09183940 -0.25700662 7.973957e-01
## 7        abdom   0.88542903  0.08007684 11.05724248 3.306570e-23
## 8          hip  -0.19841862  0.13515624 -1.46806848 1.434060e-01
## 9        thigh   0.23189542  0.13371812  1.73421094 8.417548e-02
## 10        knee  -0.01167679  0.22414282 -0.05209531 9.584964e-01
## 11       ankle   0.16353590  0.20514349  0.79717810 4.261422e-01
## 12      biceps   0.15279894  0.15851276  0.96395360 3.360476e-01
## 13     forearm   0.43048875  0.18445247  2.33387361 2.043567e-02
## 14       wrist  -1.47653692  0.49551887 -2.97977942 3.183449e-03
```

rsquared
0.749

The R^2 is slightly higher for the full model, but we claim that based on practical model deployment considerations it's justifiable to use the smaller model. If the measurements for the full model were made in a laboratory setting, one could imagine a scenario where the full model would perform worse in deployment due to poor measurement of the extra variables.

(b) Compute a 95% prediction interval for median predictor values and compare to the results to the interval for the full model. Do the intervals differ by a practically important amount?

Subset model prediction interval

```
##        fit      lwr      upr
## 1 17.84028 9.696631 25.98392
```

pi.width
16.29

Full model prediction interval

```
##        fit     lwr      upr
## 1 17.49322 9.61783 25.36861
```

pi.width
15.75

The full model does have a smaller prediction window. We don't see a big difference for the prediction intervals for the 2 models.

(c) For the smaller model, examine all the observations from case numbers 25 to 50. Which two observations seem particularly anomalous?

We plotted the features and examined the raw data and determined that case numbers 39 and 42 are potential anomalies or represent extreme values for the predictors.

Table 5: Possible outliers in dataset fat.

	age	weight	height	abdom
39	46	363.1	72.25	148.1
42	44	205	29.5	104.3

(d) Recompute the 95% prediction interval for median predictor values after these two anomalous cases have been excluded from the data. Did this make much difference to the outcome?

```
##       fit      lwr      upr
## 1 17.9033 9.887851 25.91874
```

pi.width
16.03

The prediction interval has gotten smaller but the removal of the outliers has not changed the size of the prediction interval by a lot. If we look at what happends when we perform this at the extreme values of the model parameters we might get another answer to this question.

Prediction interval for $brozek \sim age + weight + height + abdom$ at max of predictors

```
##        fit      lwr    upr
## 1 44.42187 35.41174 53.432
```

pi.width
18.02

Prediction interval for $brozek \sim age + weight + height + abdom$ at max of predictors. Model fit without outliers.

```
##        fit      lwr      upr
## 1 38.01869 29.61016 46.42723
```

pi.width
16.82

We see that the differnce prediction interval sizes is greater at the extremes of the predictors.