NCSU ST 503 HW 6

Probems 7.4, 7.6, 7.8 Faraway, Julian J. Linear Models with R, Second Edition Chapman & Hall / CRC Press.

Bruce Campbell

10 October, 2017

7.4 longley data analyis

```
##                 Estimate  Std. Error t value  Pr(>|t|)
## (Intercept)  -3.4823e+03  8.9042e+02 -3.9108 0.0035604
## GNP.deflator  1.5062e-02  8.4915e-02  0.1774 0.8631408
## GNP          -3.5819e-02  3.3491e-02 -1.0695 0.3126811
## Unemployed   -2.0202e-02  4.8840e-03 -4.1364 0.0025351
## Armed.Forces -1.0332e-02  2.1427e-03 -4.8220 0.0009444
## Population   -5.1104e-02  2.2607e-01 -0.2261 0.8262118
## Year          1.8292e+00  4.5548e-01  4.0159 0.0030368
##
## n = 16, p = 7, Residual SE = 0.30485, R-Squared = 1
```

(a) Compute and comment on the condition numbers.

We were not asked to do this, but it's interesting - so we give it a try.

Table 1: eigenvalues

ev.1	ev.2	ev.3	ev.4	ev.5	ev.6
66652993	209073	105355	18040	24.56	2.015

Table 2: condition numbers

rcond.1	rcond.2	rcond.3	rcond.4	rcond.5	rcond.6
1	17.86	25.15	60.78	1647	5751

We see a high condition number κ, and note that $\sqrt{\frac{\lambda_1}{\lambda_i}} > 30$ for $i = 4, 5$ as well where λ_i denotes the sorted eigenvalues.

(b) Compute and comment on the correlations between the predictors.

Table 3: Correlation (continued below)

	corr.mat.GNP.deflator	corr.mat.GNP	corr.mat.Unemployed
GNP.deflator	1	0.99	0.62
GNP	0.99	1	0.6
Unemployed	0.62	0.6	1
Armed.Forces	0.46	0.45	-0.18
Population	0.98	0.99	0.69
Year	0.99	1	0.67

	corr.mat.Armed.Forces	corr.mat.Population	corr.mat.Year
GNP.deflator	0.46	0.98	0.99
GNP	0.45	0.99	1
Unemployed	-0.18	0.69	0.67
Armed.Forces	1	0.36	0.42
Population	0.36	1	0.99
Year	0.42	0.99	1

We see a significant amount of correlation between the predictors.

(c) Compute the variance inflation factors.

Table 5: Variance Inflation Factors (continued below)

VIF.GNP.deflator	VIF.GNP	VIF.Unemployed	VIF.Armed.Forces
135.5	1789	33.62	3.589

VIF.Population	VIF.Year
399.2	759

The variance inflation factor for all but the Armed.Forces predictor is large.

We look at a reduced model below. This was iteratively defined by removing predictors with high condition numbers and VIF factors. We note that the R^2 is comparable to the original model.

```
##                Estimate Std. Error t value  Pr(>|t|)
```

```
## (Intercept)  30.3934125  1.8642850 16.3030 1.494e-09
## GNP.deflator  0.3980758  0.0309950 12.8432 2.261e-08
## Unemployed   -0.0107250  0.0032205 -3.3303  0.005995
## Armed.Forces -0.0081653  0.0038294 -2.1322  0.054337
##
## n = 16, p = 4, Residual SE = 0.67763, R-Squared = 0.97
```

Table 7: eigenvalues

ev.1	ev.2	ev.3
2975066	112877	1589

Table 8: condition numbers

rcond.1	rcond.2	rcond.3
1	5.134	43.28

	GNP.deflator	Unemployed	Armed.Forces
GNP.deflator	1.00	0.62	0.46
Unemployed	0.62	1.00	-0.18
Armed.Forces	0.46	-0.18	1.00

Table 10: Variance Inflation Factors

VIF.GNP.deflator	VIF.Unemployed	VIF.Armed.Forces
3.655	2.959	2.32

7.6 cheddar dataset analysis

Using the cheddar data, fit a linear model with taste as the response and the other three variables as predictors.

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -28.87677   19.73542 -1.4632 0.155399
## Acetic        0.32774    4.45976  0.0735 0.941980
## H2S           3.91184    1.24843  3.1334 0.004247
## Lactic       19.67054    8.62905  2.2796 0.031079
##
## n = 30, p = 4, Residual SE = 10.13071, R-Squared = 0.65
```

(a) Is the predictor Lactic statistically significant in this model?

We see that lactic is statistically significant at a level of $\alpha = 0.05$ with a p-value of 0.031

(b) Give the R command to extract the p-value for the test of $\beta_{lactic} = 0$. Hint: look at faraway::sumary()$coef.

After some trial and error we got the command below.

```
summary(lm.fit)$coefficients[4, 4]
```

```
## [1] 0.03107948
```

We really do not like to index model parameters by the numerical index. If this were production code we'd look for a way to use the predictor name directly. StackOverflow provided us the hint for the code below.

```
coef(summary(lm.fit))["Lactic", "Pr(>|t|)"]
```

```
## [1] 0.03107948
```

(c) Add normally distributed errors to Lactic with mean zero and standard deviation 0.01 and refit the model. Now what is the p-value for the previous test?

```
## [1] 0.03014723
```

(d) Repeat this same calculation of adding errors to Lactic 1000 times within for loop. Save the p-values into a vector. Report on the average p-value. Does this much measurement error make a qualitative difference to the conclusions?

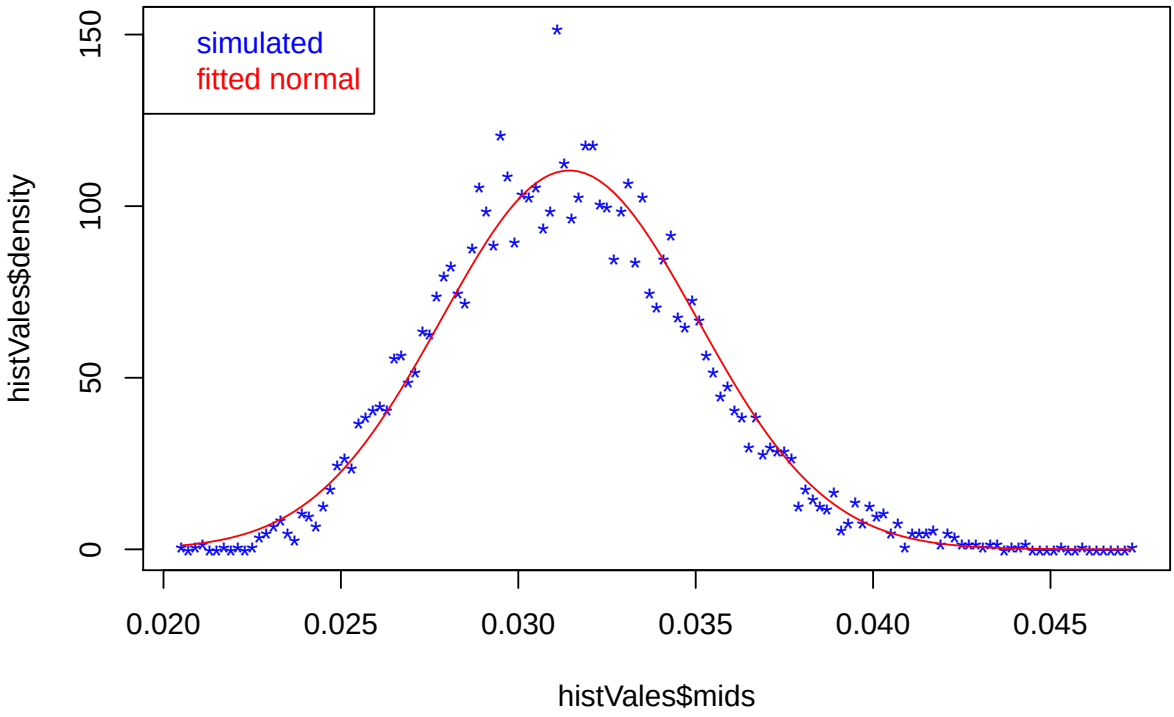


Table 11: Mean and sd of Lactic pvalues from simulation

mean.bval	sd.bval
0.03144	0.003614

We see that the p-values are not dramatically affected by the addition of noise. Above we have plotted the empirical distribution of the p-values and a normal with the same mean and standard deviation.

(e) Repeat the previous question but with a standard deviation of 0.1. Does this much measurement error make an important difference?

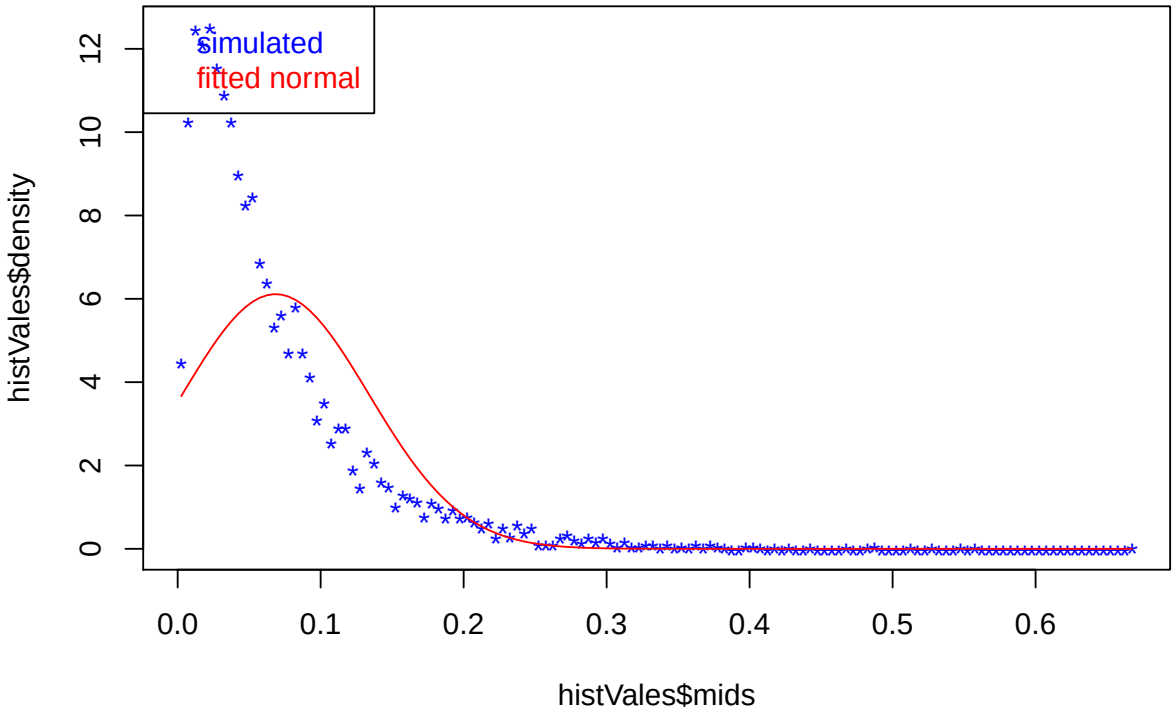


Table 12: Mean and sd of Lactic pvalues from simulation

mean.bval	sd.bval
0.06852	0.06527

We see that the p-value is significantly affected at this level of additional noise in the predictor.

7.8 fat data analysis

Use the fat data, fitting the model described in Section 4.2.

```
##                Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -15.292549  16.069921 -0.9516  0.342252
## age           0.056786   0.029965  1.8951  0.059290
## weight       -0.080310   0.049581 -1.6198  0.106602
```

```
## height       -0.064600   0.088930 -0.7264  0.468299
## neck         -0.437541   0.215334 -2.0319  0.043273
## chest        -0.023603   0.091839 -0.2570  0.797396
## abdom         0.885429   0.080077 11.0572 < 2.2e-16
## hip          -0.198419   0.135156 -1.4681  0.143406
## thigh         0.231895   0.133718  1.7342  0.084175
## knee         -0.011677   0.224143 -0.0521  0.958496
## ankle         0.163536   0.205143  0.7972  0.426142
## biceps        0.152799   0.158513  0.9640  0.336048
## forearm       0.430489   0.184452  2.3339  0.020436
## wrist        -1.476537   0.495519 -2.9798  0.003183
##
## n = 252, p = 14, Residual SE = 3.98797, R-Squared = 0.75
```

(a) Compute the condition numbers and variance inflation factors. Comment on the degree of collinearity observed in the data.

Table 13: eigenvalues (continued below)

ev.1	ev.2	ev.3	ev.4	ev.5	ev.6	ev.7	ev.8	ev.9	ev.10
19592555	64185	30597	5704	2804	1935	1030	637.7	528.1	431.8

ev.11	ev.12	ev.13
376.4	272.4	63.45

Table 15: condition numbers (continued below)

rcond.1	rcond.2	rcond.3	rcond.4	rcond.5	rcond.6	rcond.7	rcond.8
1	17.47	25.3	58.61	83.59	100.6	137.9	175.3

rcond.9	rcond.10	rcond.11	rcond.12	rcond.13
192.6	213	228.2	268.2	555.7

We note a high condition number for the model matrix, and a number of the individual predictors have a large value of $\frac{\lambda_1}{\lambda_i}$

Table 17: Variance Inflation Factors (continued below)

VIF.age	VIF.weight	VIF.height	VIF.neck	VIF.chest	VIF.abdom	VIF.hip
2.25	33.51	1.675	4.324	9.461	11.77	14.8

VIF.thigh	VIF.knee	VIF.ankle	VIF.biceps	VIF.forearm	VIF.wrist
7.778	4.612	1.908	3.62	2.192	3.378

We see weight and abdom - and marginally chest - have VIF values indicating colinearity with other predictors.

(b) Cases 39 and 42 are unusual. Refit the model without these two cases and recompute the collinearity diagnostics. Comment on the differences observed from the full data fit.

```
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept)  2.622509  21.654912  0.1211  0.903711
## age          0.065827   0.029806  2.2085  0.028168
## weight      -0.015960   0.062198 -0.2566  0.797713
## height      -0.223547   0.177083 -1.2624  0.208057
## neck        -0.359262   0.217585 -1.6511  0.100041
## chest       -0.110593   0.100290 -1.1027  0.271267
## abdom        0.839876   0.084675  9.9188 < 2.2e-16
## hip         -0.153126   0.135132 -1.1332  0.258298
## thigh        0.174473   0.136035  1.2826  0.200902
## knee        -0.069409   0.227540 -0.3050  0.760603
## ankle        0.174260   0.203676  0.8556  0.393100
## biceps       0.151605   0.157859  0.9604  0.337846
## forearm      0.268272   0.191700  1.3994  0.162995
## wrist       -1.642878   0.493838 -3.3268  0.001019
##
## n = 250, p = 14, Residual SE = 3.94247, R-Squared = 0.75
```

We see fewer significant predictors in the fit without the outliers. This makes intuitive sense - if the outlier(s) is(are) related to abnormal values of one of the predictors then it possible that predictor will have undue influence on the fit through the outliers. Removing the outliers eliminates the influence and the significance. Neck and thigh are predictors that we'd consider for this effect, and indeed inspecting the data confirms that is the case for one (39) of the redacted data points. The other outlier (42) has a very small value for the height predictor. Although it is not significant in the model fit with the redacted data, the p-value is half that for the mode with the outliers.

(c) Fit a model with brozek as the response and just age, weight and height as predictors. Compute the collinearity diagnostics and compare to the full data fit.

Table 19: eigenvalues

ev.1	ev.2	ev.3
9824566	51002	15672

Table 20: condition numbers

rcond.1	rcond.2	rcond.3
1	13.88	25.04

Table 21: Variance Inflation Factors

VIF.age	VIF.weight	VIF.height
1.083	1.381	1.47

We see the colinearity diagnostics all indicate that there is no linear association among the predictors.

(d) Compute a 95% prediction interval for brozek for the median values of age, weight and height.

Table 22: Median Value of Predictors

median.age	median.weight	median.height
43	176.1	70

Table 23: 95% Prediction Interval For Univariate Median of Predictors

fit	lwr	upr
18.49	8.648	28.33

Table 24: 95% Prediction Interval Width For Univariate Median of Predictors

pi.width
19.68

(e) Compute a 95% prediction interval for brozek for age=40, weight=200 and height=73. How does the interval compare to the previous prediction?

Table 25: 95% Prediction Interval For (age=40, weight=200 and height=73)

fit	lwr	upr
20.18	10.32	30.05

Table 26: 95% Prediction Interval Width For(age=40, weight=200 and height=73)

pi.width
19.73

This interval does not differ in width from the interval calculated from the median predictor values.

(f) Compute a 95% prediction interval for brozek for age=40, weight=130 and height=73. Are the values of predictors unusual? Comment on how the interval compares to the previous two answers.

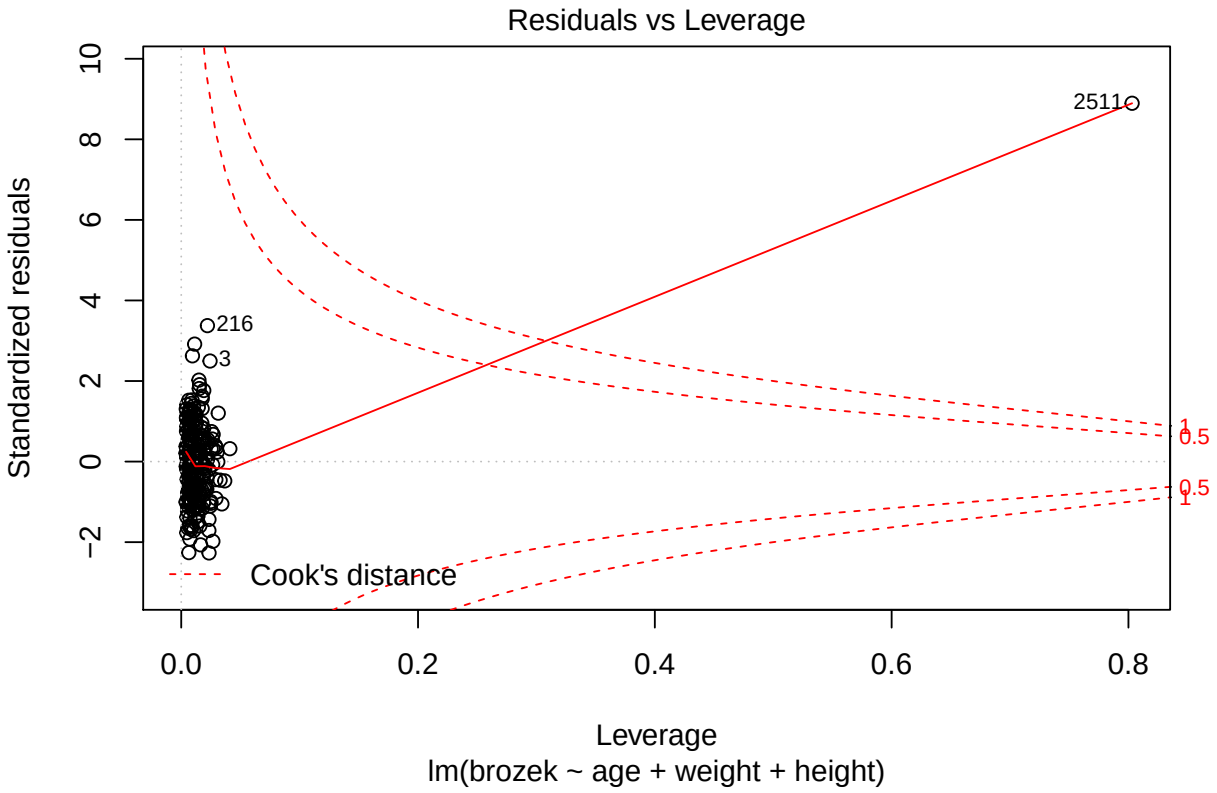
Table 27: 95% Prediction Interval For (age=40, weight=200 and height=73)

fit	lwr	upr
3.72	-6.282	13.72

Table 28: 95% Prediction Interval Width For(age=40, weight=200 and height=73)

pi.width
20

The prediction interval with is larger for this example, and the predicted body fat is a very low value. Due to the weight, this data points is likely a high leverage points. We can add it to the training set and see.



Indeed - the added point (251) is a high leverage point in the model with