

Statistics Final Exam

Consider the following analysis of a dataset from economics. Each line (observation) in the dataset corresponds to a fortune 500 company. The analysis concerns the following variables:

- *lsales* – the natural log of the company’s sales
- *lassets* – the natural log of the company’s assets
- *lmarket.value* – the natural log of the company’s current market value
- *lemployees* – the natural log of the number of employees of the company
- *energy* – an indicator variable, 1 if the company is in the energy sector, 0 otherwise

Here is some of the *R* session for the analysis:

```
> m1 <- lm(lsales ~ lassets + lmarket.value + lemployees + energy)
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.94648	0.45925	8.593	8.85e-13
lassets	0.19156	0.05282	3.627	0.00052
lmarket.value	0.07861	0.06783	1.159	0.06015
lemployees	0.60392	0.05462	11.056	0.03824
energy	0.61879	0.14339	4.316	4.86e-05

Residual standard error: 0.4833 on 74 degrees of freedom

Multiple R-squared: 0.845, Adjusted R-squared: 0.8388

F-statistic: 136.3 on 3 and 75 DF, p-value: < 2.2e-16

1. Based on this model, what would be the effect of a one point increase in *lassets* on *lsales*?

7. What are the two models being tested above?

8. What are the null and alternative hypotheses being tested by the p-value $< 2.2e-16$?

9. What is the conclusion of the test above?

10. What are the two models being tested above?

11. How many of the confidence intervals computed by the command:
`> confint(m1)`
would contain 0 (which ones, if any, and why)?

12. What percentage of the variability in *lsales* is explained by the model?

13. What is the model for energy sector employees? (Write out the actual numerical coefficients for the β_i . Feel free to round to two decimal places if you like.)

14. Suppose that after the exam, I pass a collection plate and our class uses the proceeds to purchase a U.S. automobile manufacturer. Our ability to secure governmental bailout funding will be proportional to our potential sales, and hence a need to predict future sales.

```
> predict(m1,newdata=our.co,interval='confidence',level=.95)
```

```
      fit      lwr      upr
8.725756 8.5666 8.88491
```

(a) What does it mean that we are 95% confident with this 95% confidence interval?

(b) What, exactly, is it that we are 95% confident that this 95% confidence interval contains?

15. For each of the diagnostic plots listed below, tell what, exactly, is being (informally) tested for and the implications of the specific problems that might be found by that plot for future analysis.

(a) The residual plot.

(b) The jackknife residual plot.

(c) The hat value plot.

(d) The diffits plot.

(e) Collectively, the dfbetas plots.

16. Looking at the added variable plot for lemployees, make predictions for the output of the summary command below. In particular comment on what you think will likely change (relative to m1) regarding r^2 , the statistical significance of the slopes, and on

the p -value for the slope of $l\text{employees}^2$.

```
> m2 <- lm(lsales ~ lassets + lmarket.value + lemployees + I(lemployees2))
> summary(m2)
```

17. The analyst decides to explore a further model (m3). What are the null and alternative hypotheses being tested by the anova command?

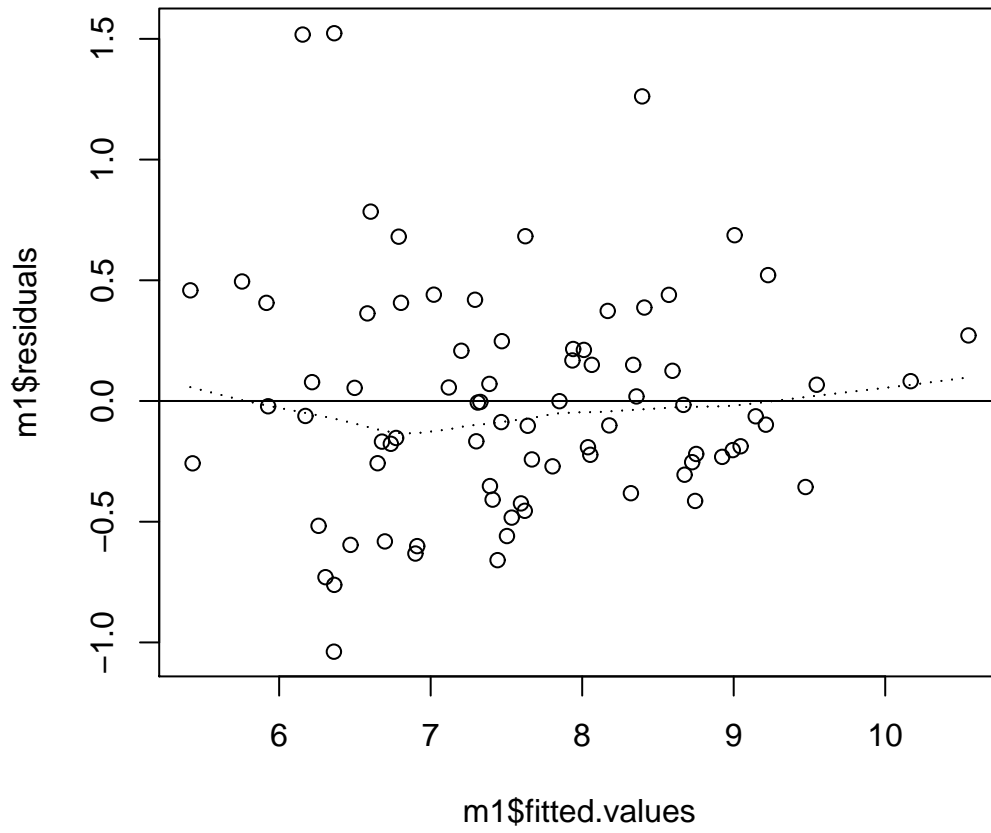
```
> m3 <- lm(lsales ~ lassets + lemployees + lmarket.value + energy
           lassets*energy + lmarket.value*energy + lemployees*energy)
```

```
> anova(m1,m3)
```

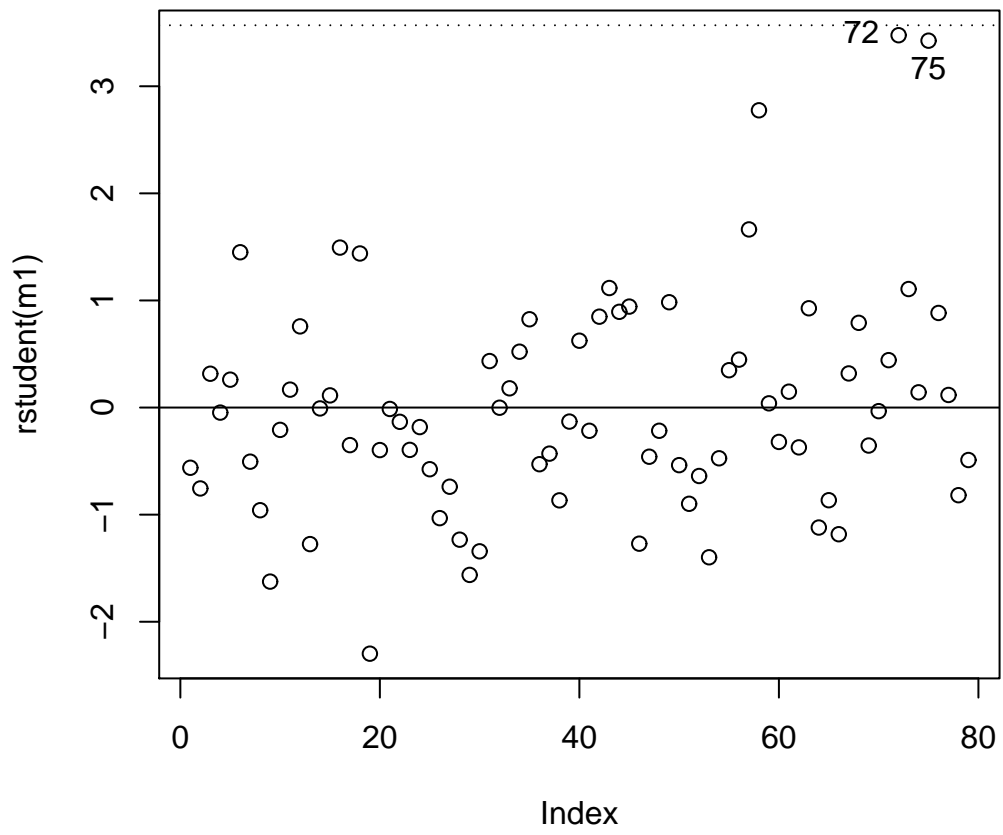
	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	75	13.9979				
2	71	11.9582	4	2.0397	3.0276	0.02304

18. What is the result of this test? What does that mean about the relationship between a firm's being in the energy sector, its lassets, lemployees, lmarket.value, and the y -variable lsales?

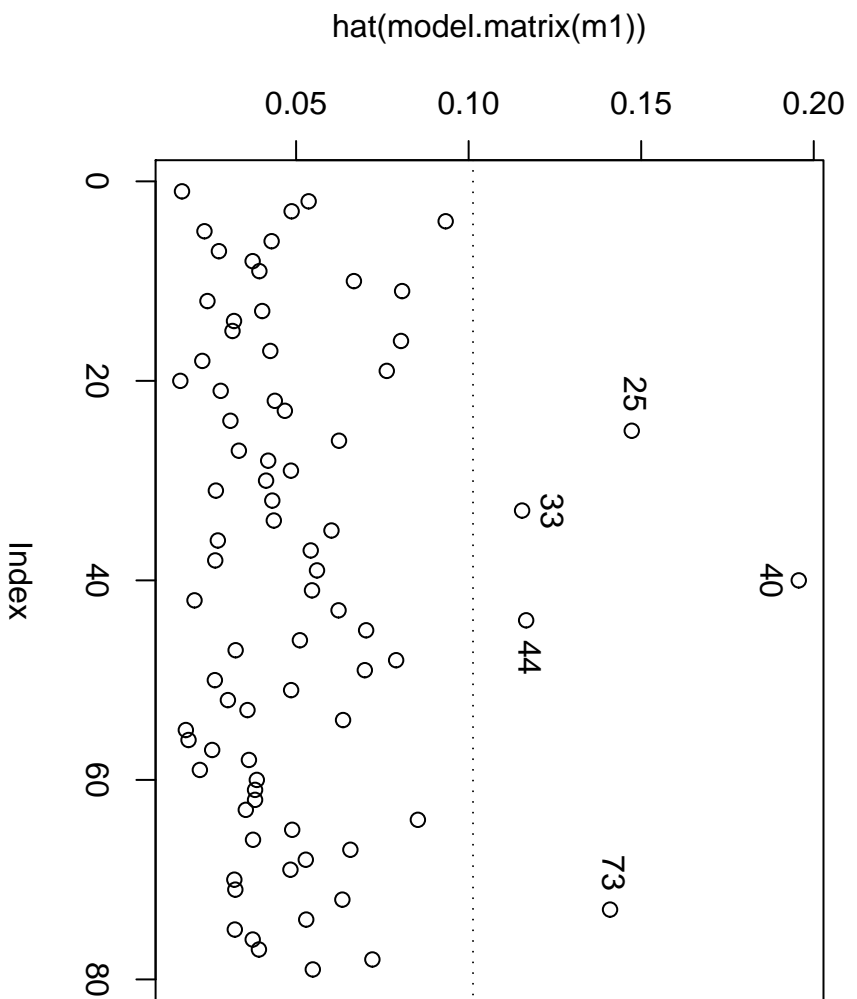
m1 residual plot



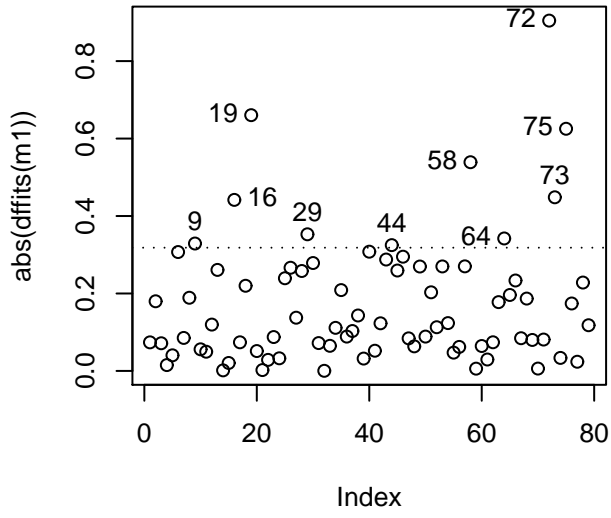
m1 jackknife residual plot



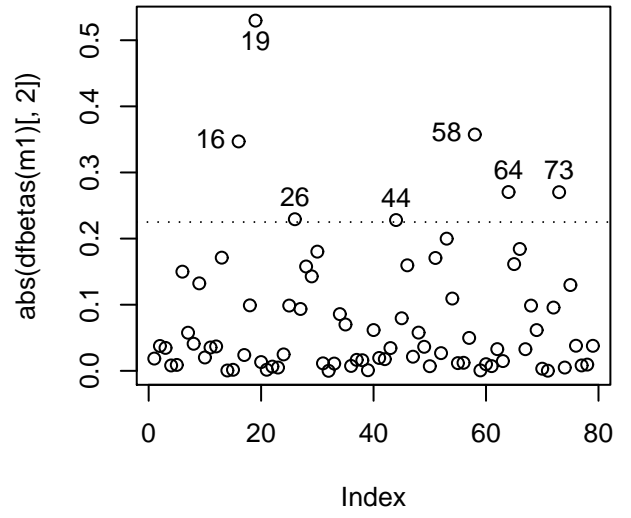
m1 hat plot



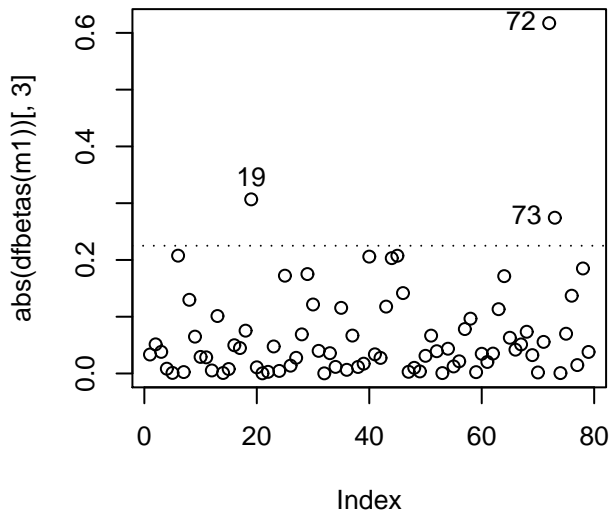
dffits m1



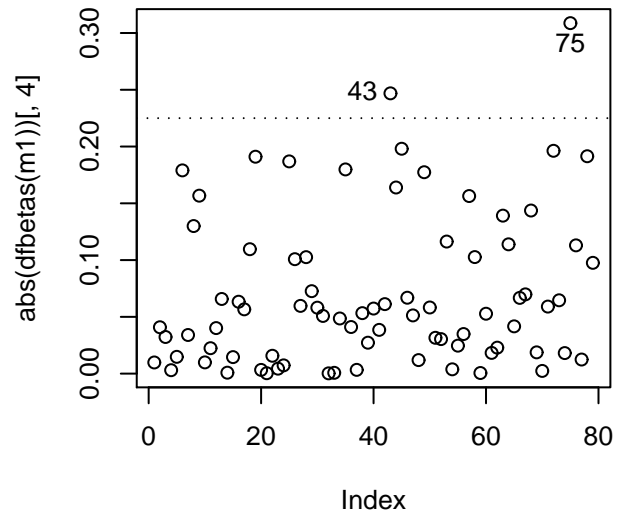
lassets dfbetas



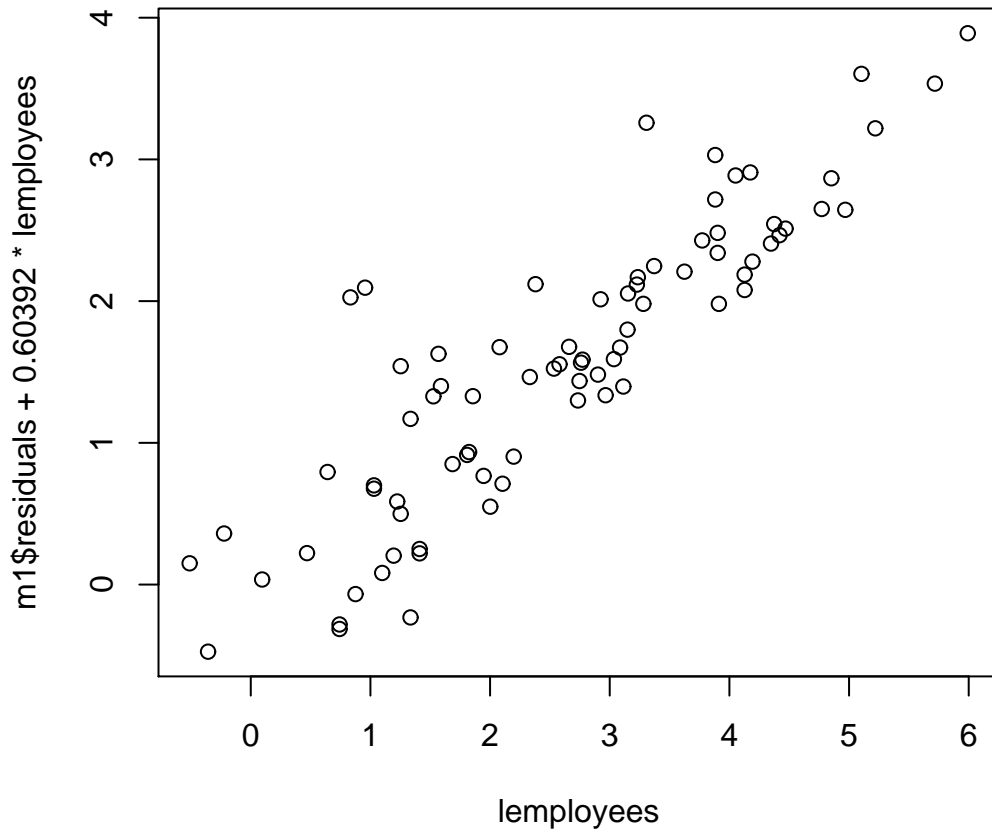
lmarket.value dfbetas



lemployees dfbetas



employees added variable plot



```
> m1 <- lm(lsales ~ lassets + lmarket.value + lemloyees + energy)
> summary(m1)
```

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	3.94648	0.45925	8.593	8.85e-13
lassets	0.19156	0.05282	3.627	0.00052
lmarket.value	0.07861	0.06783	1.159	0.06015
lemloyees	0.60392	0.05462	11.056	0.03824
energy	0.61879	0.14339	4.316	4.86e-05

Residual standard error: 0.4833 on 74 degrees of freedom

Multiple R-squared: 0.845, Adjusted R-squared: 0.8388

F-statistic: 136.3 on 3 and 75 DF, p-value: < 2.2e-16

```
> predict(m1,newdata=our.co,interval='confidence',level=.95)
```

	fit	lwr	upr
	8.725756	8.5666	8.88491

```
> m3 <- lm(lsales ~ lassets + lemloyees + lmarket.value + energy
           lassets*energy + lmarket.value*energy + lemloyees*energy)
```

```
> anova(m1,m3)
```

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	75	13.9979				
2	71	11.9582	4	2.0397	3.0276	0.02304