

Exercise 7.1

We have a sample of 200 people, with age (in years) and bone mineral density (**bmd**) for each person. A simple linear regression model was run using age to predict BMD. Output is below.

```
. regress bmd age
```

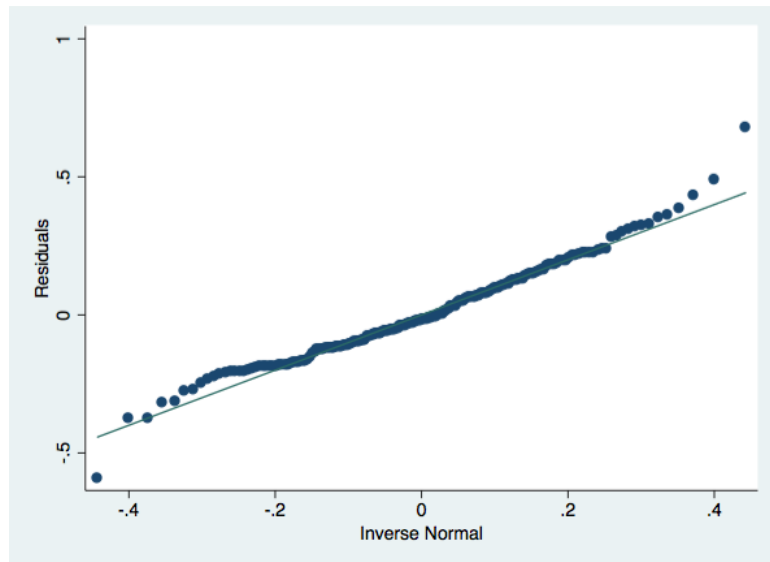
Source	SS	df	MS	Number of obs	=	200
Model	1.62569273	1	1.62569273	F(1, 198)	=	54.86
Residual	5.86753243	198	.029634002	Prob > F	=	0.0000
Total	7.49322516	199	.037654398	R-squared	=	0.2170
				Adj R-squared	=	0.2130
				Root MSE	=	.17215

bmd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.0046087	.0006222	-7.41	0.000	-.0058358	-.0033817
_cons	1.164711	.0332406	35.04	0.000	1.09916	1.230262

(a) Interpret the effect of age on BMD based on this model. Is there evidence of a significant effect?

(b) Report and interpret the R^2 value for this model.

(c) In checking the model assumptions, one useful plot is a Q-Q plot of the residuals. This plot is below. Do there appear to be any outliers based on this plot?

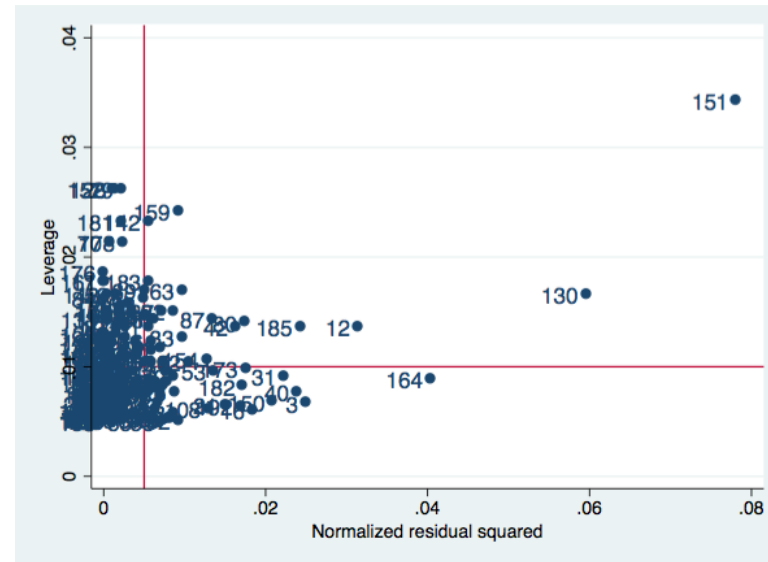
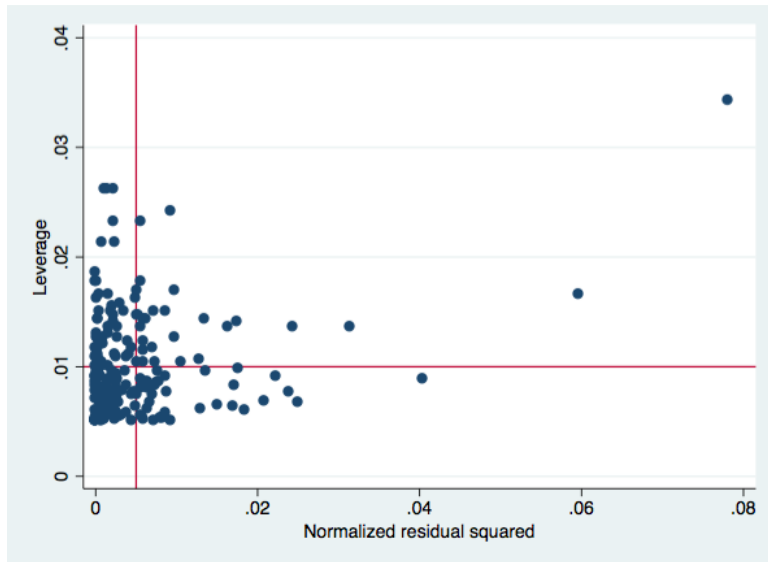


(d) As a next step, the studentized residuals were calculated. Using the output below, identify what (if any) IDs you are most concerned about as potential outliers.

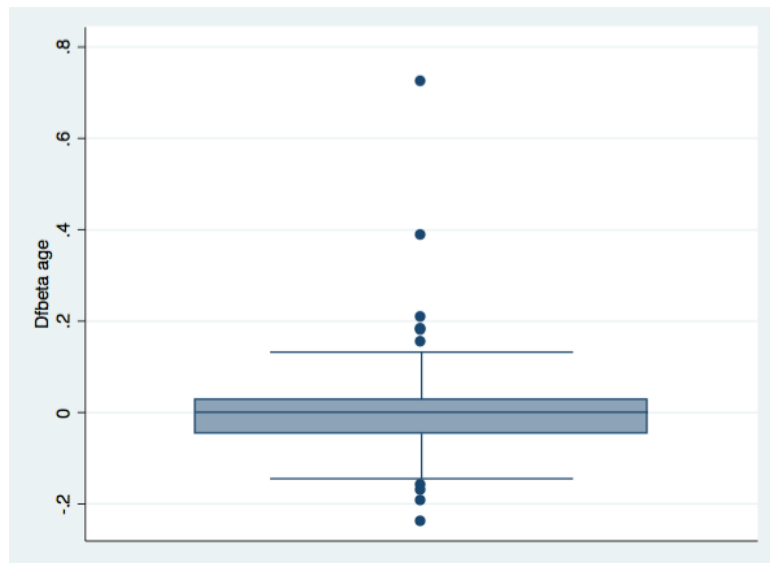
```
. list id rhat if rhat > 2 | rhat < -2
```

	id	rhat
3.	3	2.257787
12.	12	2.549181
31.	31	2.12993
40.	40	-2.206589
130.	130	-3.566072
150.	150	2.059209
151.	151	4.165857
164.	164	2.8946
185.	185	-2.233705

(e) Next, a leverage-versus-squared-residuals plot was generated. Two versions of the plot are shown below – one with and one without the IDs superimposed. Are there any points that you think might be highly influential based on this plot? If so, report the ID(s) and explain why they might be highly influential.



(f) The DFBETAs for the model were calculated and plotted as shown below (once with and once without the IDs superimposed). Do the DFBETAs support your conclusion in part (e)?



The regression model was run with again, removing the one influential point. Output is below.

```
. regress bmd age if id ~= 151
```

Source	SS	df	MS	Number of obs	=	199
				F(1, 197)	=	69.00
Model	1.88863492	1	1.88863492	Prob > F	=	0.0000
Residual	5.39249059	197	.027373049	R-squared	=	0.2594
				Adj R-squared	=	0.2556
Total	7.28112551	198	.036773361	Root MSE	=	.16545

bmd	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
age	-.0050421	.000607	-8.31	0.000	-.0062392	-.003845
_cons	1.182745	.0322393	36.69	0.000	1.119167	1.246324

(g) Interpret the effect of age on BMD based on this model with the influential point removed. Is there evidence of a significant effect? How do these results compare to the original results?

(h) Report and interpret the R^2 value for this model. Compare this to the R^2 from the original results.

Exercise 7.2

We have measured the IL-6, an inflammatory cytokine in the blood (units: pg/ml), in a sample of 85 adults. We also know the age (years) and sex (1=male, 0=female) of each person. We are interested whether higher IL-6 is associated with depression, with depression measured using a scale where higher means worse depression. Initially, we fit the model as shown below in Stata.

```
. regress depress il6 age
```

Source	SS	df	MS	Number of obs	=	85
				F(2, 82)	=	10.57
Model	94.9158104	2	47.4579052	Prob > F	=	0.0001
Residual	368.331248	82	4.49184449	R-squared	=	0.2049
				Adj R-squared	=	0.1855
Total	463.247059	84	5.51484594	Root MSE	=	2.1194

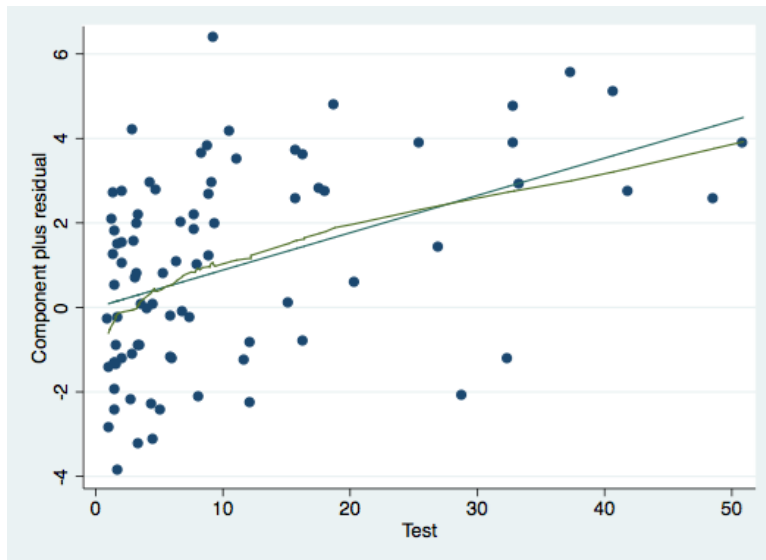
depress	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
il6	.0883511	.0196082	4.51	0.000	.0493442	.1273579
age	.0537898	.0556169	0.97	0.336	-.05685	.1644296
_cons	2.470479	2.803275	0.88	0.381	-3.106127	8.047085

(a) Interpret the effect of IL-6 in this model. Is there evidence of a significant effect? (assume $\alpha=0.05$)

(b) Interpret the effect of age in this model. Is there evidence of a significant effect? (assume $\alpha=0.05$)

(c) What is the R^2 for this model? Interpret this quantity.

Based on the CPR plot below, it appears that there may be a non-linear effect of IL-6, and a log-transformation may be appropriate. Thus, The model was run using natural log-transformed IL-6. Output is on the next page.




```
. generate ln_il6 = ln(il6)
. regress depress ln_il6 age
```

Source	SS	df	MS	Number of obs	=	85
				F(2, 82)	=	12.00
Model	104.872944	2	52.4364722	Prob > F	=	0.0000
Residual	358.374115	82	4.37041603	R-squared	=	0.2264
				Adj R-squared	=	0.2075
Total	463.247059	84	5.51484594	Root MSE	=	2.0906

depress	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
ln_il6	1.034924	.2151201	4.81	0.000	.6069821	1.462867
age	.055592	.0548654	1.01	0.314	-.0535527	.1647368
_cons	1.440774	2.789832	0.52	0.607	-4.10909	6.990638

(d) Interpret the effect of IL-6 in this model. Is there evidence of a significant effect? (assume $\alpha=0.05$)

(e) How does depression score change if IL-6 increases by 10%?

(f) Interpret the effect of age in this model. Is there evidence of a significant effect? (assume $\alpha=0.05$)

(g) What is the R^2 for this model? Interpret this quantity, and compare to the R^2 in the original model.

Exercise 7.3

In the same sample of 85 people, we have another cytokine measured, called TNF (units: pg/ml). We would like to know if age and sex are associated with TNF. A regression was run using age and sex to predict TNF. Output is below.

```
. regress tn timer sex age
```

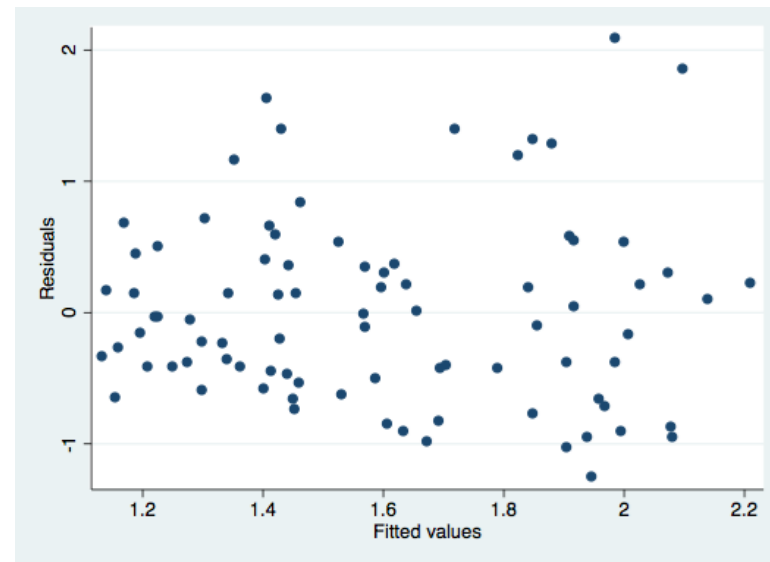
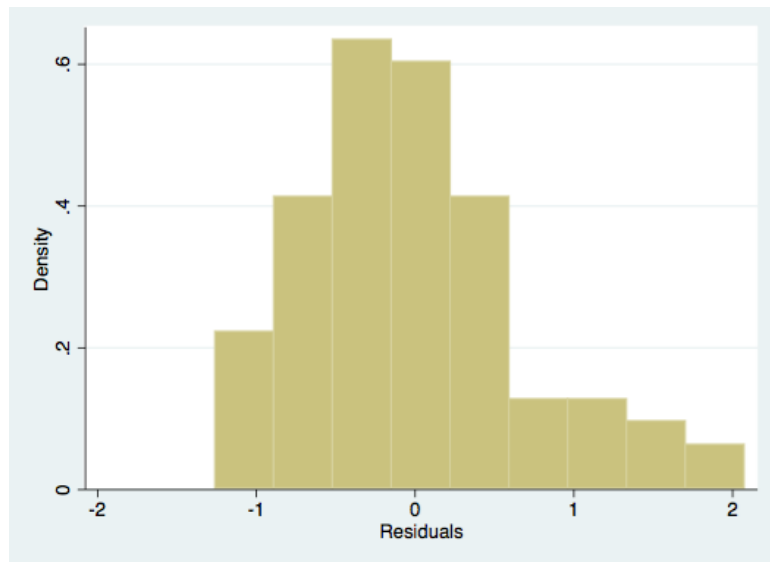
Source		SS	df	MS	Number of obs	=	85
-----+-----							
Model		7.48742689	2	3.74371345	F(2, 82)	=	7.33
Residual		41.8874142	82	.510822124	Prob > F	=	0.0012
-----+-----							
Total		49.3748411	84	.587795727	R-squared	=	0.1516
					Adj R-squared	=	0.1310
					Root MSE	=	.71472

tnf		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----							
sex		.5026441	.1559238	3.22	0.002	.192462	.8128262
age		.0383495	.0187544	2.04	0.044	.001041	.075658
_cons		-.5466748	.943964	-0.58	0.564	-2.42452	1.33117
-----+-----							

(a) Interpret the effect of age in this model. Is there evidence of a significant effect? (assume $\alpha=0.05$)

(b) Interpret the effect of sex in this model. Is there evidence of a significant effect? (assume $\alpha=0.05$)

(c) Based on the histogram of residuals from this model and the RVF plot (both shown below), do there appear to be any problems with model assumptions?



The model was re-run, using natural log-transformed TNF as the outcome. Output is below.

```
. generate ln_tnf = ln(tnf)
```

```
. regress ln_tnf sex age
```

Source		SS	df	MS	Number of obs	=	85	
-----+-----								
Model		2.58442832	2	1.29221416	F(2, 82)	=	6.83	
Residual		15.5232379	82	.189307779	Prob > F	=	0.0018	
-----+-----								
Total		18.1076662	84	.215567454	R-squared	=	0.1427	
-----+-----								
						Adj R-squared	=	0.1218
						Root MSE	=	.4351

ln_tnf		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
-----+-----							
sex		.2957837	.0949209	3.12	0.003	.1069559	.4846116
age		.022441	.011417	1.97	0.053	-.0002711	.0451531
_cons		-.8939395	.5746519	-1.56	0.124	-2.037105	.2492261
-----+-----							

(d) Interpret the effect of age in this model. Is there evidence of a significant effect? (assume $\alpha=0.05$)

(e) Interpret the effect of sex in this model. Is there evidence of a significant effect? (assume $\alpha=0.05$)