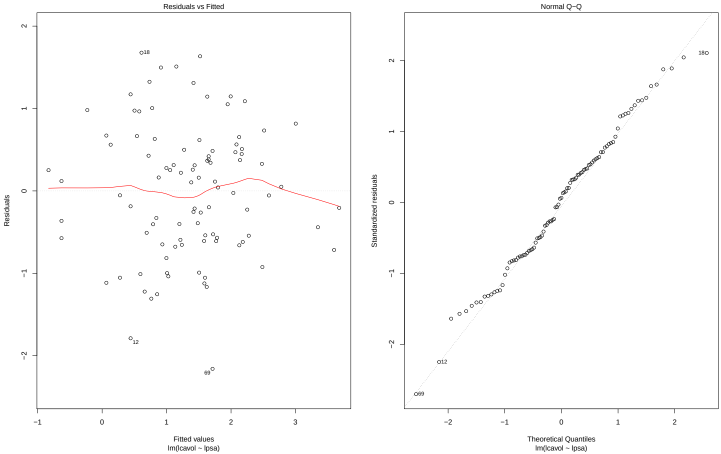
Tutorial 5

YANG YANG

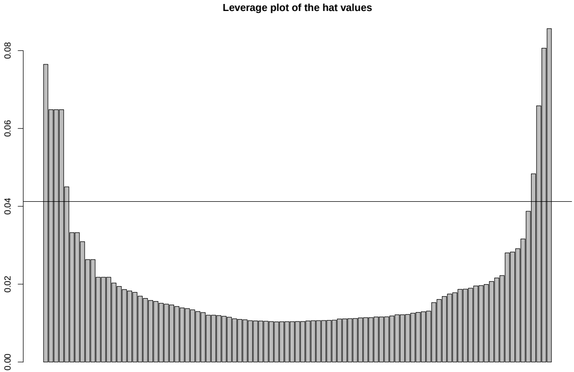
The Australian National University

Week 6, 2017

Q2 (b) Residual plot; Q-Q plot



Q2 (b) Leverage barplot




```
barplot(hat(lpsa),main="leverage plot of the hat values")
abline(h=4/length(lpsa))
```

Q2 (d)

```
> summary(prostate.lm)

Call:
lm(formula = lcavol ~ lpsa)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.15948 -0.59383  0.05034  0.50826  1.67751 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) -0.50858    0.19419  -2.619   0.0103 *  
lpsa         0.74992    0.07109  10.548   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8041 on 95 degrees of freedom
Multiple R-squared:  0.5394,	Adjusted R-squared:  0.5346 
F-statistic: 111.3 on 1 and 95 DF,  p-value: < 2.2e-16
```

$$\text{Model}: \texttt{lcavol} = \beta_0 + \beta_1 \texttt{lpsa} + \varepsilon \quad \varepsilon \sim i.i.d.\ N\left(0, \sigma^2\right)$$

$$\text{H}_0 : \beta_1 = 0 \quad \text{H}_A : \beta_1 \neq 0$$

$t_{95} = 10.5$, $p << 0.05$, so reject H_0 in favour of H_A and conclude that the slope coefficient of `lpsa` is significantly different from 0, implying there is a significant linear relationship between `lcavol` and `lpsa`. Note this test is again directly equivalent to the tests in parts (a) and (c).

Q2 (e)

```
> range(lpsa)
[1] -0.43078  5.58293
> lpsa.values <- -20:120/20
> lpsa.values
  [1] -1.00 -0.95 -0.90 -0.85 -0.80 -0.75 -0.70 -0.65 -0.60 -0.55 -0.50 -0.45 -0.40 -0.35 -0.30 -0.25 -0.20 -0.15 -0.10
 [20] -0.05  0.00  0.05  0.10  0.15  0.20  0.25  0.30  0.35  0.40  0.45  0.50  0.55  0.60  0.65  0.70  0.75  0.80  0.85
 [39]  0.90  0.95  1.00  1.05  1.10  1.15  1.20  1.25  1.30  1.35  1.40  1.45  1.50  1.55  1.60  1.65  1.70  1.75  1.80
 [58]  1.85  1.90  1.95  2.00  2.05  2.10  2.15  2.20  2.25  2.30  2.35  2.40  2.45  2.50  2.55  2.60  2.65  2.70  2.75
 [77]  2.80  2.85  2.90  2.95  3.00  3.05  3.10  3.15  3.20  3.25  3.30  3.35  3.40  3.45  3.50  3.55  3.60  3.65  3.70
 [96]  3.75  3.80  3.85  3.90  3.95  4.00  4.05  4.10  4.15  4.20  4.25  4.30  4.35  4.40  4.45  4.50  4.55  4.60  4.65
[115]  4.70  4.75  4.80  4.85  4.90  4.95  5.00  5.05  5.10  5.15  5.20  5.25  5.30  5.35  5.40  5.45  5.50  5.55  5.60
[134]  5.65  5.70  5.75  5.80  5.85  5.90  5.95  6.00
```

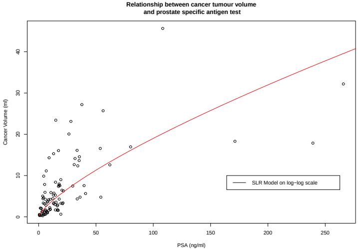
- `range(lpsa)` gives the range of x-axis in the scatter plot.
- `lpsa.values <- -20:120/20` creates a sequence of values that cover the full range of lpsa
- The more points we have, the smoother our curve would look like. Normally 50 points are enough for a good plot.

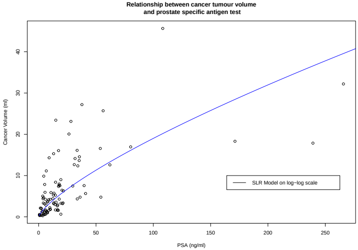
Q2 (e) extra plot

```
# Using predict() to plot the SLR model which is back-transformed to the original (non-log) scale:
cintervals <- predict(prostate.lm, newdata=data.frame(lpsa=lpsa.values), interval="confidence")

plot(exp(lpsa), exp(lcavol), main="Relationship between cancer tumour volume\n and prostate specific antigen test",
     xlab="PSA (ng/ml)", ylab="Cancer Volume (ml)")
lines(exp(lpsa.values), exp(cintervals[,"fit"]),col="red")
legend(164, 10, c("SLR Model on log-log scale"), lty=1)

# Manually typing in SLR model:
coef(prostate.lm)

yi <- coef(prostate.lm)[1]+coef(prostate.lm)[2]*lpsa.values
plot(exp(lpsa), exp(lcavol), main="Relationship between cancer tumour volume\n and prostate specific antigen test",
     xlab="PSA (ng/ml)", ylab="Cancer Volume (ml)")
lines(exp(lpsa.values), exp(yi),col="red")
legend(164, 10, c("SLR Model on log-log scale"), lty=1)
```

- Put `lpsa.values` into the SLR modelt to calculate a sequence of response $\hat{Y}_i$.
- We can use either `predict()` function or do the multiplication manually.
- Match the scale of axes. Don't forget to do back transformation.

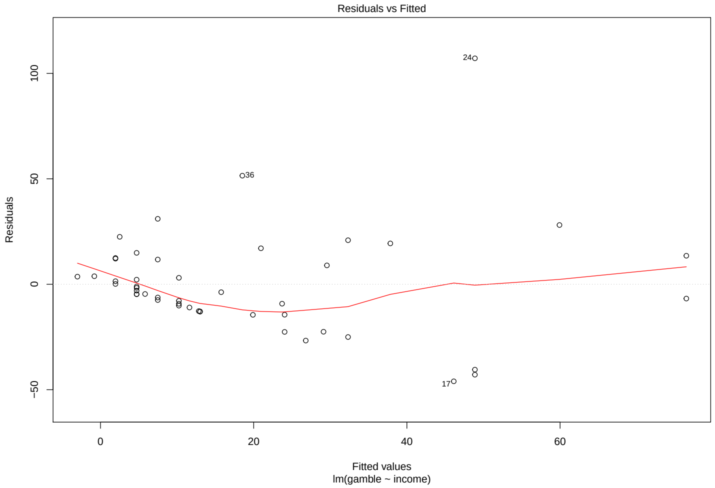
Q2 (e) extra plot





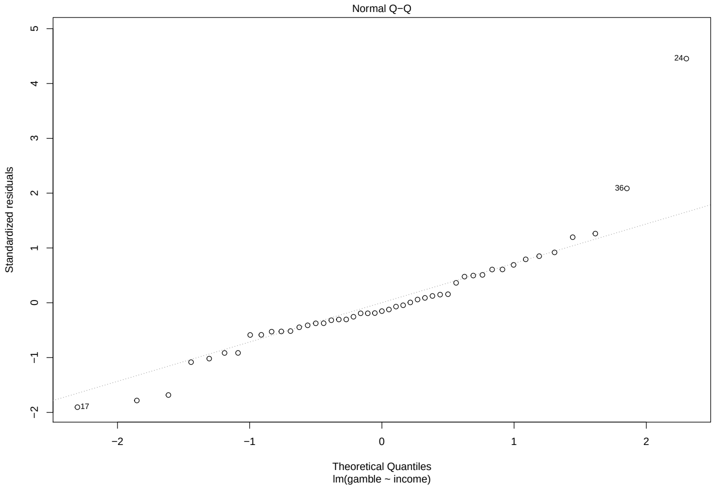
Both methods create exactly the same curves. The `predict()` function is recommended as it can be used to get CI and PI.

Q3 (b)



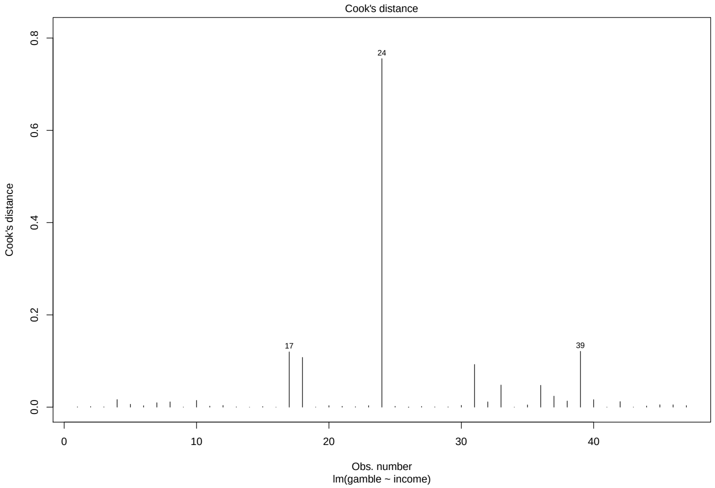
Residuals vs Fitted
Residuals
Fitted values
lm(gamble ~ income)
100
50
0
−50
0
20
40
60
24
36
17

Q3 (b)



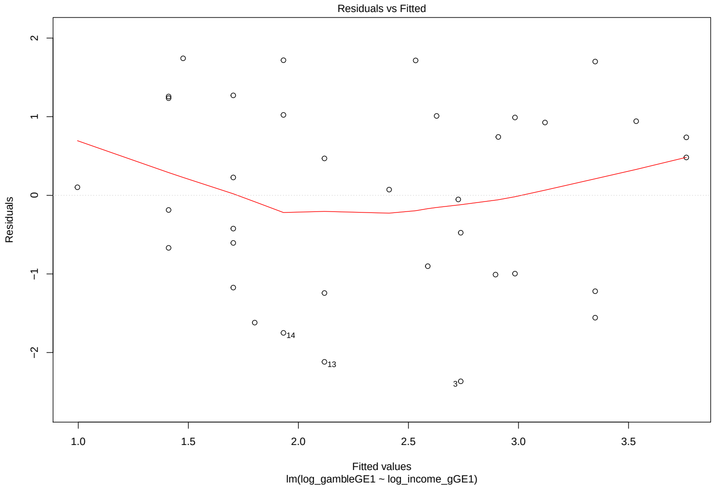
Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(gamble ~ income)
24
36
17
−2
−1
0
1
2
3
4
5

Q3 (b)



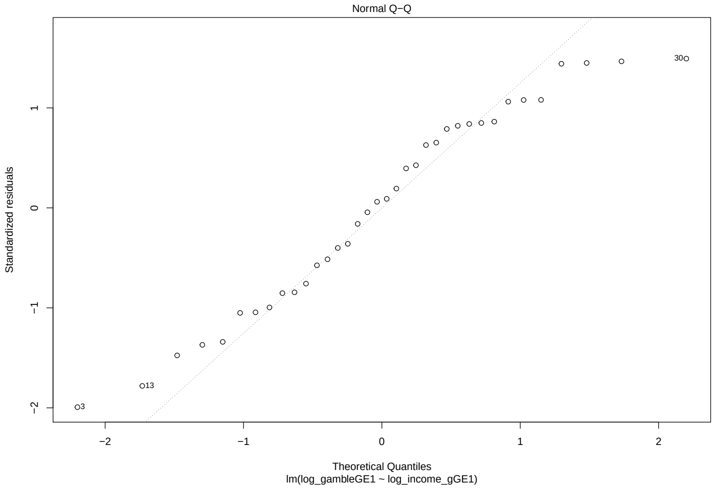
Cook's distance
Cook's distance
0.0
0.2
0.4
0.6
0.8
24
17
39
0
10
20
30
40
Obs. number
lm(gamble ~ income)

Q3 (c)



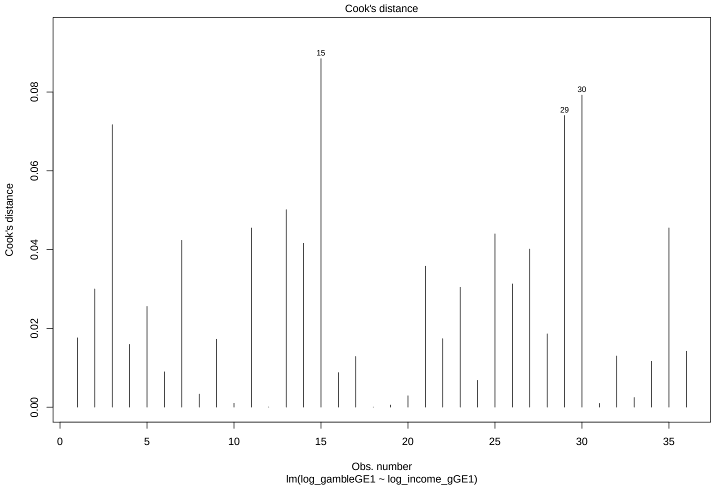
Residuals vs Fitted
Residuals
2
1
0
−1
−2
14
13
3
1.0
1.5
2.0
2.5
3.0
3.5
Fitted values
lm(log_gambleGE1 ~ log_income_gGE1)

Q3 (c)



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(log_gambleGE1 ~ log_income_gGE1)
30
13
3
−2
−1
0
1
2

Q3 (c)


Cook's distance
Cook's distance
0.00
0.02
0.04
0.06
0.08
15
30
29
0
5
10
15
20
25
30
35
Obs. number
lm(log_gambleGE1 ~ log_income_gGE1)