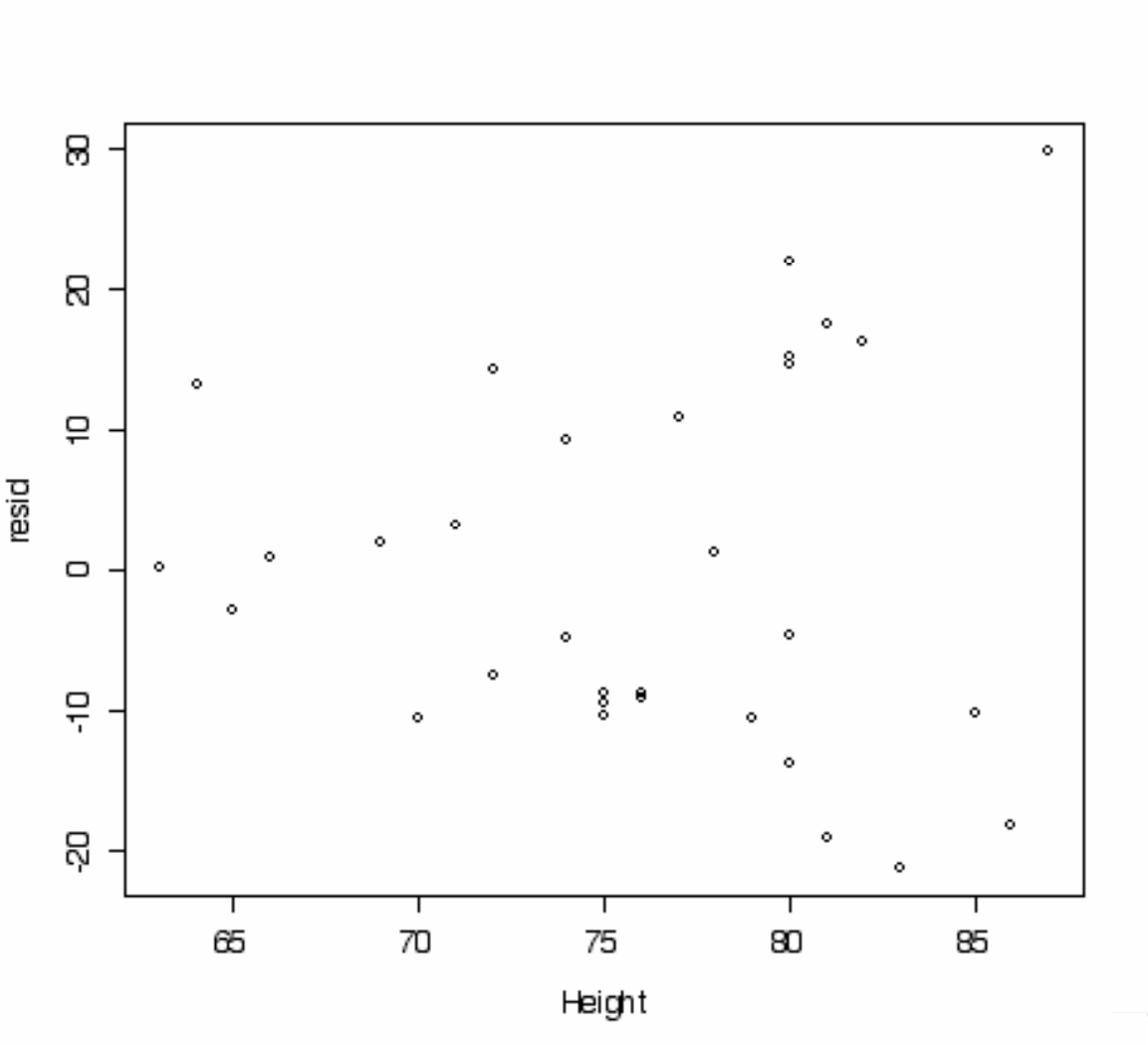
R Notes for Lecture 7

Get tree data.

```
> treeframe <- read.table("tree.txt", header=T, sep="\t")
> names(treeframe)
[1] "Diameter" "Height"   "Volume"
> attach(treeframe)
> plot(Height, Volume)
> lines(lowess(Height, Volume))
```

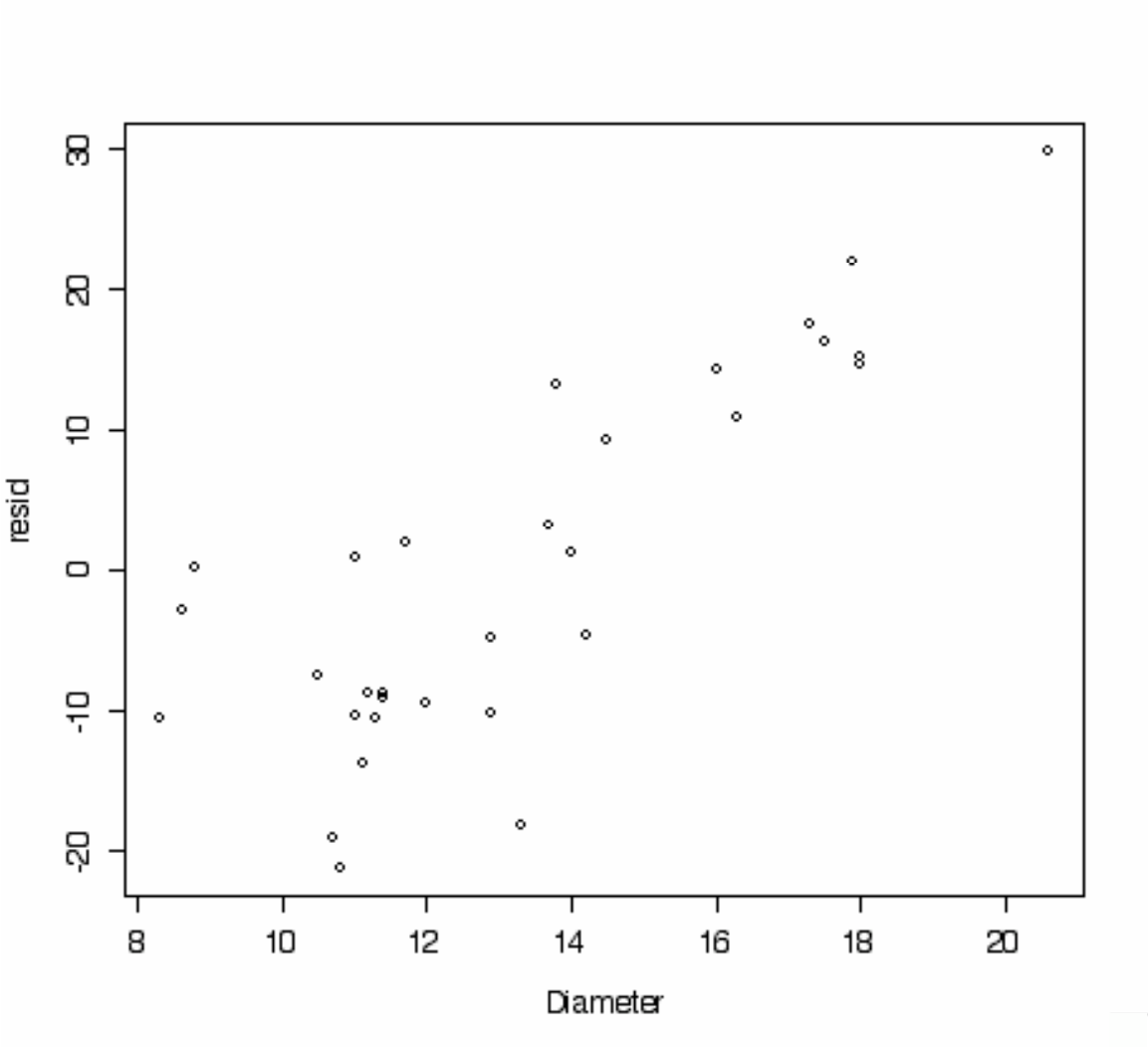
Fit model

```
> lm.height <- lm(Volume ~ Height)
> abline(lm.height)
> summary(lm.height)

Call:
lm(formula = Volume ~ Height)

Residuals:
    Min      1Q  Median      3Q     Max
-21.274  -9.894  -2.894  12.067  29.852

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -87.1236    29.2731  -2.976 0.005835 **
Height        1.5433     0.3839   4.021 0.000378 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 13.4 on 29 degrees of freedom
Multiple R-Squared: 0.3579,   Adjusted R-squared: 0.3358
F-statistic: 16.16 on 1 and 29 DF,  p-value: 0.0003784

>
Find and plot residuals
> resid <- residuals(lm.height)
> plot(Height, resid)
```




```
plot(Diameter, resid)
```

(This shows that some of the left over information is in the Diameter.)



Now fit a model E(Height|Diameter) = a + b*diameter

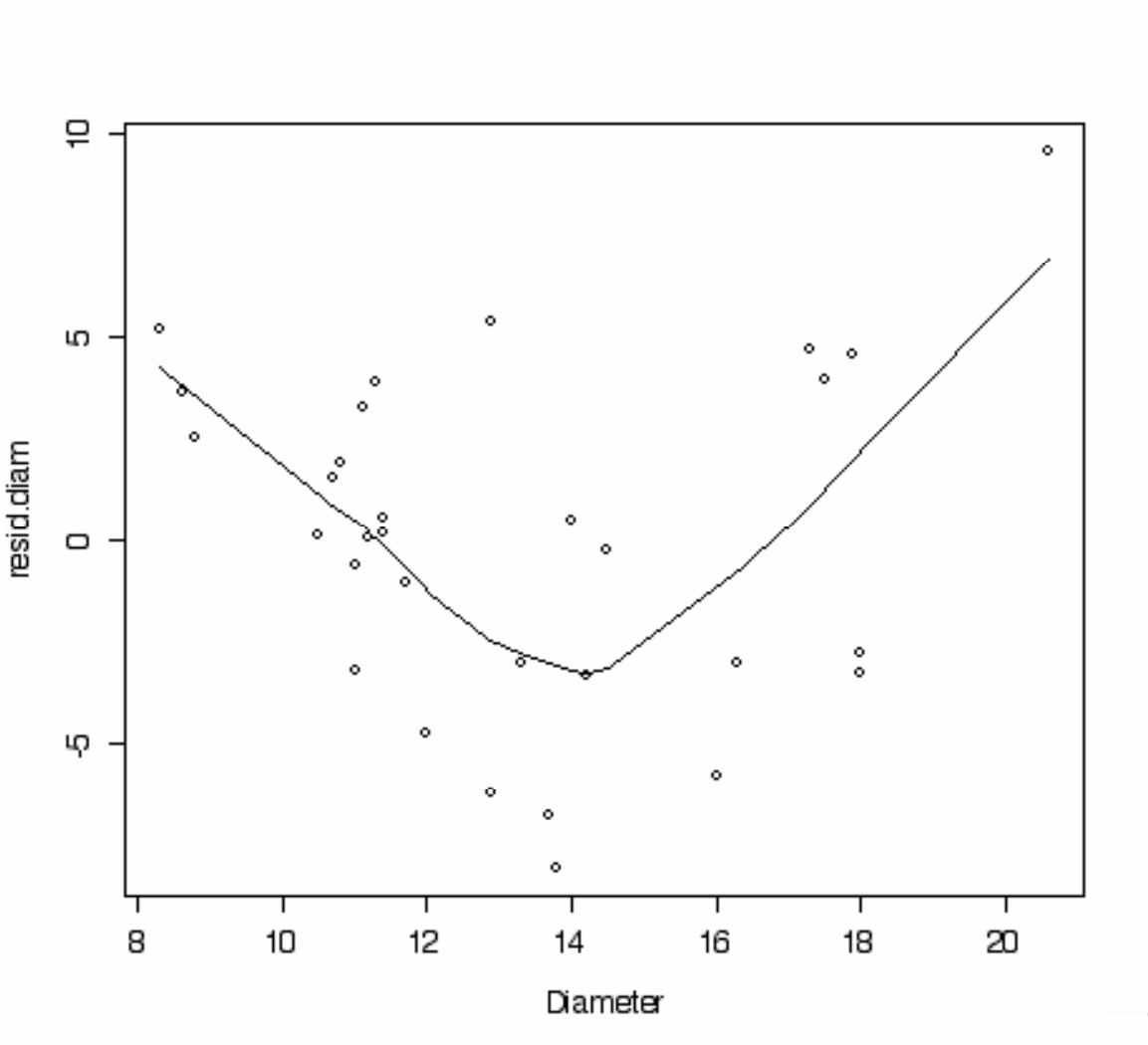
```
> lm.diameter <- lm(Volume ~ Diameter)
> summary(lm.diameter)

Call:
lm(formula = Volume ~ Diameter)

Residuals:
    Min      1Q  Median      3Q     Max
-8.0654 -3.1067  0.1520  3.4948  9.5868

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

```
(Intercept) -36.9435     3.3651  -10.98 7.62e-12 ***
Diameter      5.0659     0.2474   20.48  < 2e-16 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1

Residual standard error: 4.252 on 29 degrees of freedom
Multiple R-Squared: 0.9353,   Adjusted R-squared: 0.9331
F-statistic: 419.4 on 1 and 29 DF,  p-value: < 2.2e-16
```

Predict height = -36.9435 + 5.0659*Diameter

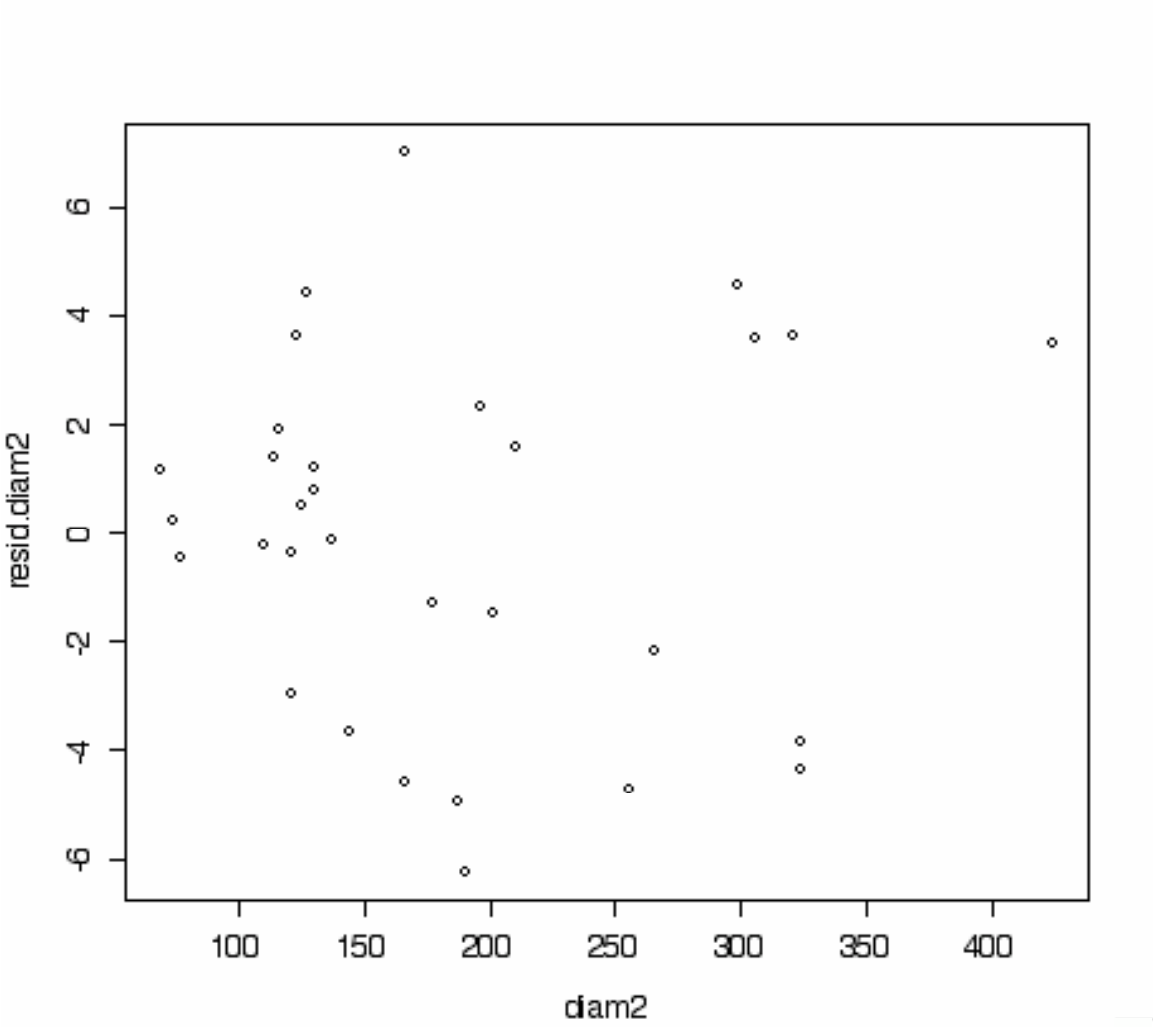
The residual plot has a pretty clear curve in it,
suggesting that the "real" relationship might be quadratic:

```
> resid.diam <- residuals(lm.diameter)
> plot(Diameter, resid.diam)
> lines(lowess(Diameter, resid.diam))
>
```



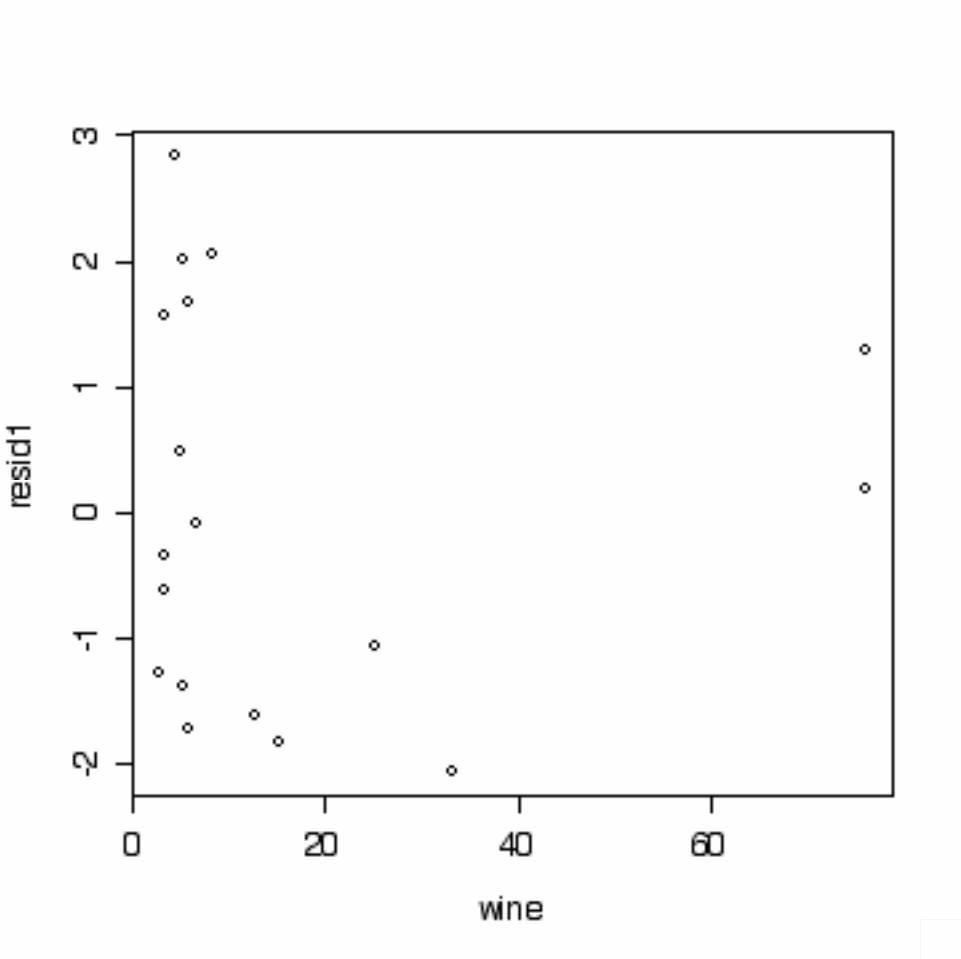
Fitting E(Volume|diameter) = a + b(diameter)^2 gives a better (more linear) fit:

```
> diam2 <- Diameter^2
> lm.diam2 <- lm(Volume ~ diam2)
> resid.diam2 <- residuals(lm.diam2)
> plot(diam2, resid.diam2)
>
```



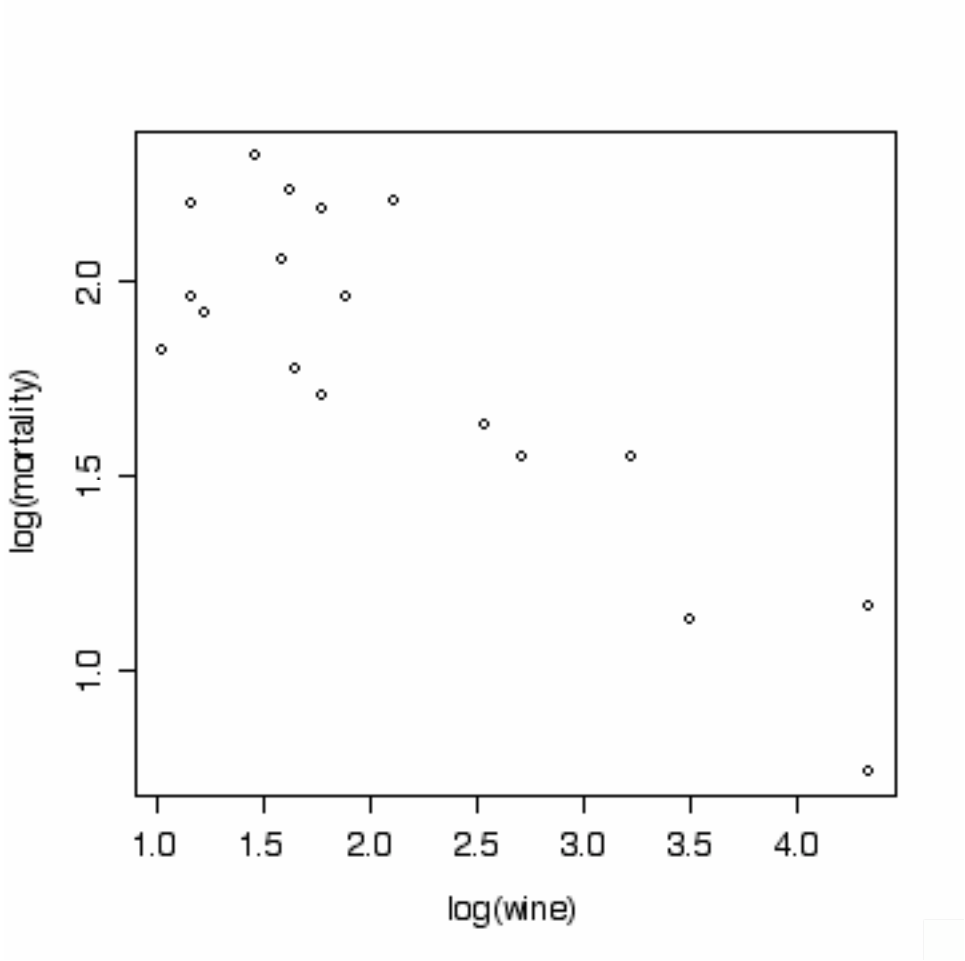
Wine Data

```
> winetable <- read.table("wine.txt", header=T, sep="\t")
> attach(winetable)
> names(winetable)
[1] "wine"      "mortality" "country"
> plot(mortality, wine)
> plot(wine, mortality)
> lm1 <- lm(mortality~wine)
> resid1 <- residuals(lm1)
> plot(wine, resid1)
```

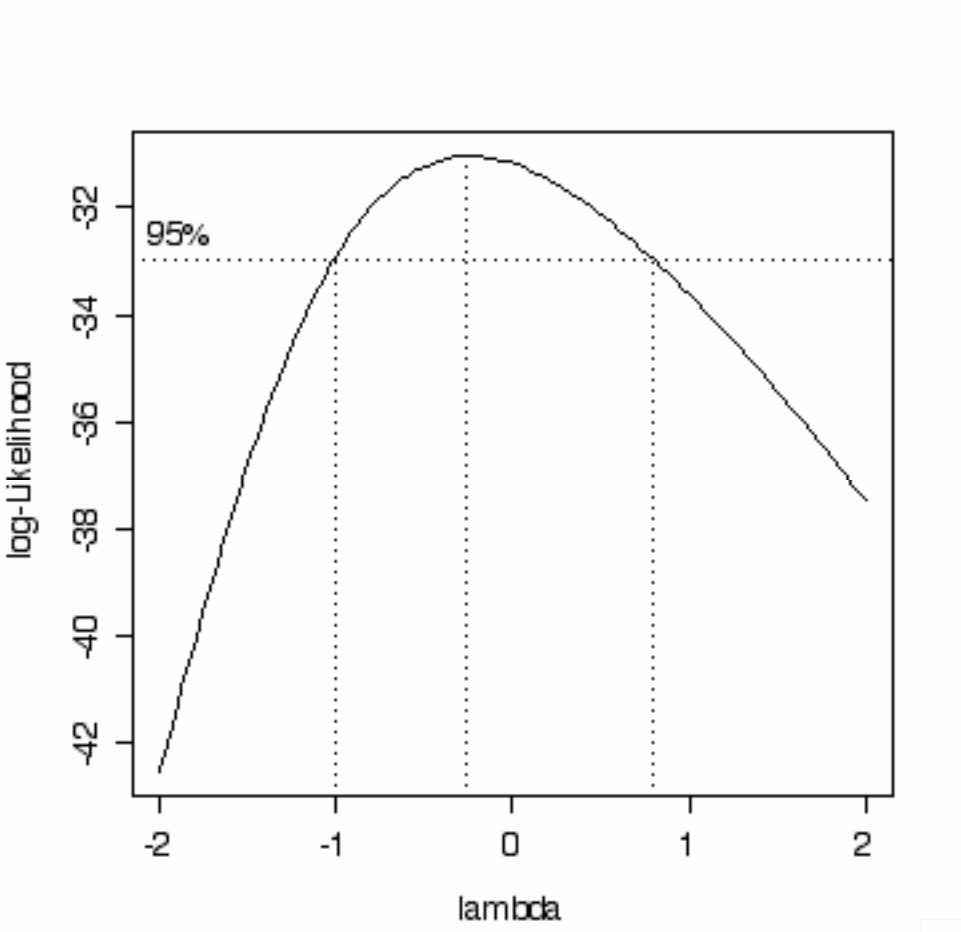


So we try various transformations. Only the last is shown:

```
plot(wine, mortality)
> plot(wine, sqrt(mortality))
> plot(wine, log(mortality))
> plot(sqrt(wine), mortality)
> plot(log(wine), mortality)
> plot(log(wine), log(mortality))
```



Or we can look at boxcox transforms:
boxcox(lm1)
The plot suggests something less than 1 and bigger than -1
Note that 0 is really close to the center of this interval
-- this suggests that the log transform might be really
good.

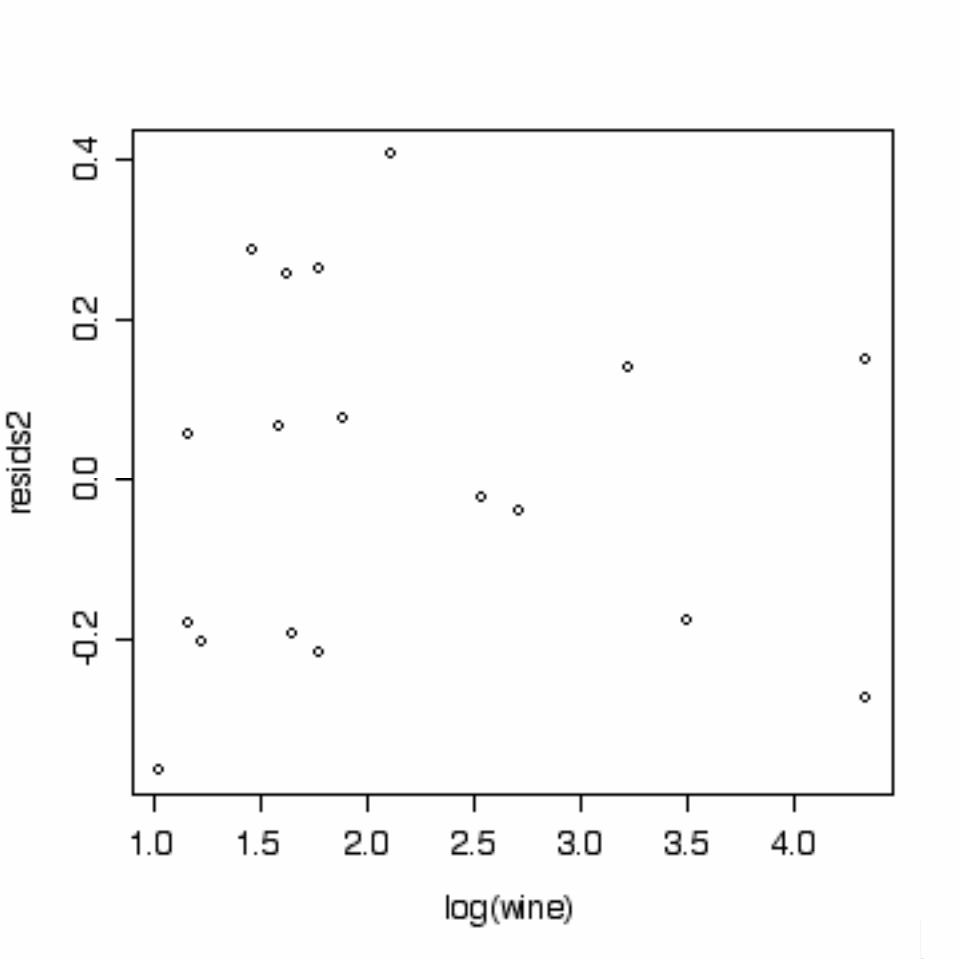



```
> lm2 <- lm(log(mortality)~ log(wine))
> summary(lm2)

Call:
lm(formula = log(mortality) ~ log(wine))

Residuals:
     Min       1Q   Median       3Q      Max
-0.36487 -0.19122  0.01497  0.14485  0.40525

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.55555    0.12690  20.139 8.60e-13 ***
log(wine)   -0.35560    0.05291  -6.721 4.91e-06 ***
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 `
' 1
```

```
Residual standard error: 0.2285 on 16 degrees of freedom
Multiple R-Squared: 0.7384,   Adjusted R-squared: 0.7221
F-statistic: 45.17 on 1 and 16 DF,  p-value: 4.914e-06
```

Before interpreting this, though, better check the residuals:

```
> resids2 <- residuals(lm2)
> plot(log(wine), resids2)
```



Much better!

We now have a model that says
Log(predicted mortality) = 2.5555 - .35556*log(wine)