Arsenic report

Randy L Coryell

June 17, 2016

Contents

Overview of Analysis

This report outlines a *regression* analysis of arsenic concentration as a function of well depth. it woll only include some basic assumption checks of normality and equal variance of the error terms.

Note: Arsenic is measured in ppb; well depth is measured in feet.

As we proceed in our analysis we shall endeavor to remember the following:

> The most that can be expected from any model is that it can supply a useful approximation to reality: All models are wrong; some models are useful.
>
> *George Box*

Read the Data

```
file.name <- "arsenic-03.data"
arsenic.data <- read.table(file=file.name, sep="\t", skip=5,
                           header=TRUE, na.strings=c("."))
dim(arsenic.data)
```

```
## [1] 200   2
```

```r
head(arsenic.data, 4)
```

```
##   arsenic depth
## 1     1.2  9.82
## 2     2.5 10.20
## 3     3.4 10.47
## 4     4.3 10.97
```

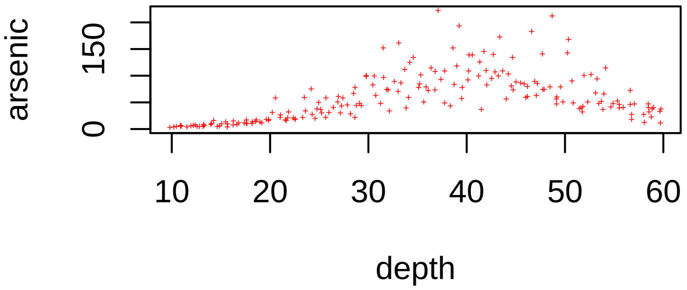
The data has 0 observations with missing values out of a total of 200 observations.

Analysis

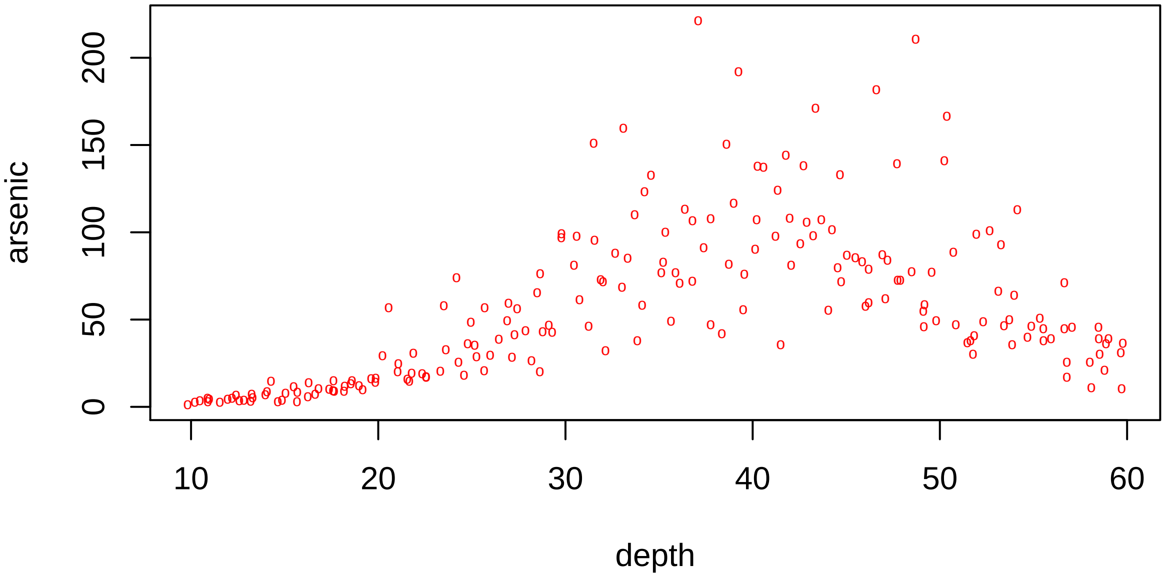
Visual Inspection of the Relationship

We wish to examine the data visually.

```r
plot(arsenic ~ depth, data=arsenic.data, cex=0.3, pch="+", col="red")
```



```r
plot(arsenic ~ depth, data=arsenic.data, cex=.5, pch="o", col="red")
```



Perhaps a cubic relationship would fit the trend as it first curves upwards and then curves downwards. It also appears that the variability increases in the middle.

We will try a regression fit of arsenic using a cubic polynomial in *depth*. We first create quadratic and cubic *depth* variables.

```
arsenic.data <- within(arsenic.data, {
depth2 <- depth^2
depth3 <- depth^3
})
head(arsenic.data, 3)
```

```
##   arsenic depth    depth3   depth2
## 1     1.2  9.82  946.9662  96.4324
## 2     2.5 10.20 1061.2080 104.0400
## 3     3.4 10.47 1147.7308 109.6209
```

```
reg01 <- lm(arsenic ~ depth + depth2 + depth3,
            data=arsenic.data)
summary(reg01)$coef
```

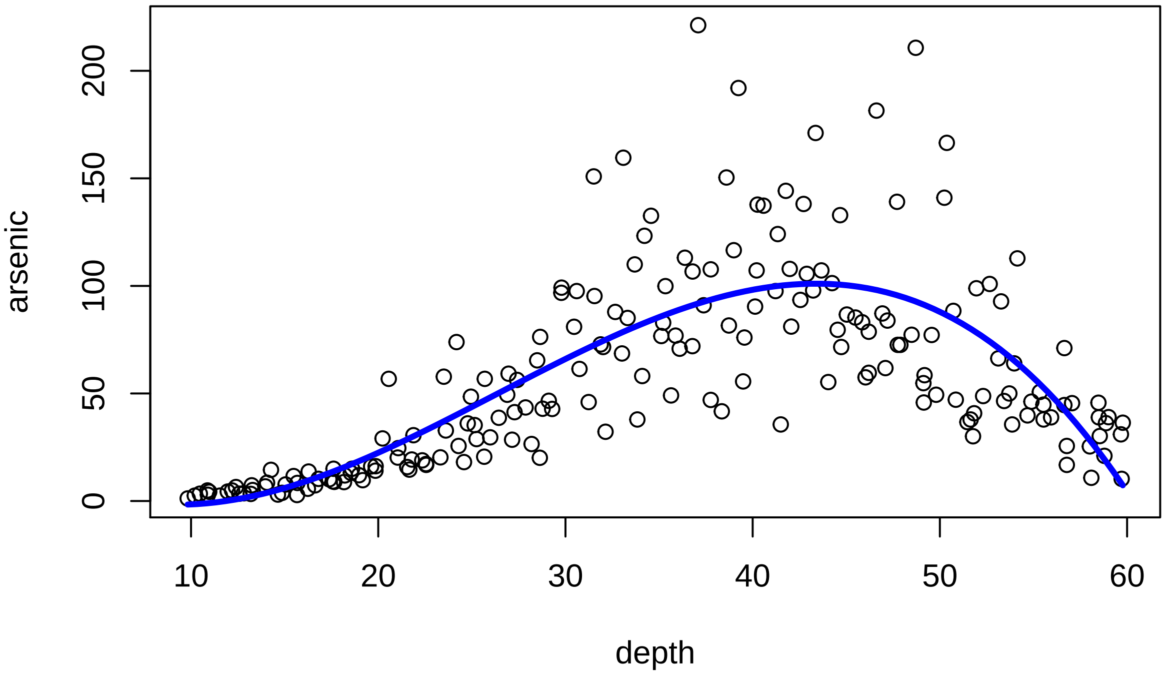
```
##                 Estimate   Std. Error    t value     Pr(>|t|)
## (Intercept) 24.984740904 28.279682077  0.8834873 3.780554e-01
## depth       -6.216938338  2.931700817 -2.1205910 3.521358e-02
## depth2       0.407811186  0.090586884  4.5018789 1.154290e-05
## depth3      -0.005165608  0.000855654 -6.0370293 7.728324e-09
```

Table of Regression Coefficients

Variable	Coefficient	Significant
Intercept	24.985	No
Depth	-6.217	Yes
Depth2	0.408	Yes
Depth3	0.005	Yes

We add the fitted curve to the data plot to see how well it appears to fit.

```
plot(arsenic ~ depth, data=arsenic.data)
x.range <- range(arsenic.data$depth)
x <- seq(x.range[1], x.range[2], length.out=100)
curve.data <- data.frame(depth=x, depth2=x^2, depth3=x^3)
curve.data$arsenic <- predict(reg01, curve.data)
lines(x=curve.data$depth, y=curve.data$arsenic,
      col="blue", lwd=3)
```



The fit looks fairly decent.

Assumptions

We now check assumptions about:

- error variances
  - should be equal
- error distribution
  - should be normal

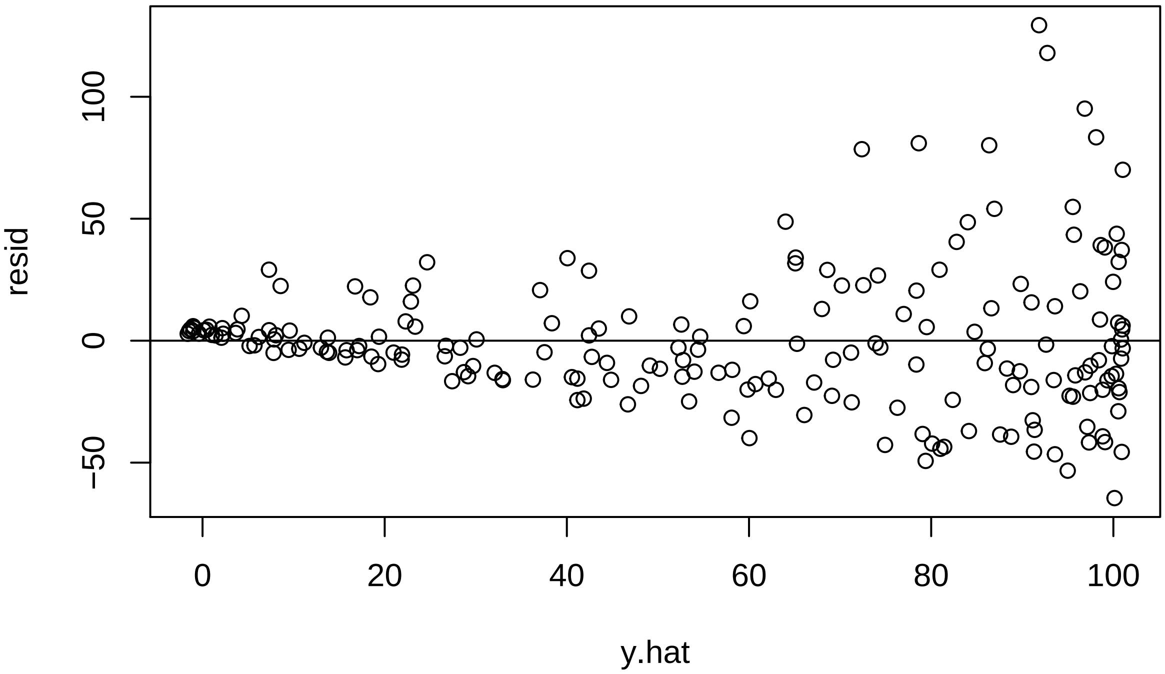
We will do this with:

1. plotting e_i versus $\hat{y}_i$
2. normal quantile plot

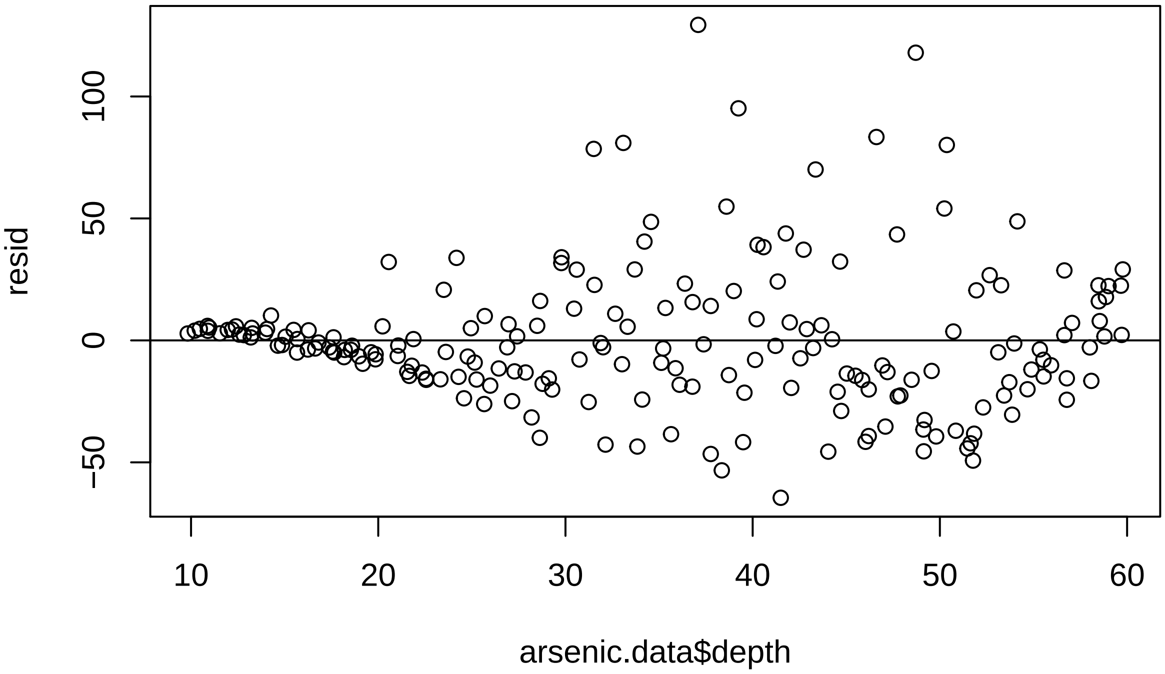
Equal Variance of Errors

We visually investigate the assumption of equal variances.

```
resid <- residuals(reg01)
y.hat <- predict(reg01, arsenic.data)
plot(resid ~ y.hat)
abline(h=0)
```



```
plot(resid ~ arsenic.data$depth)
abline(h=0)
```

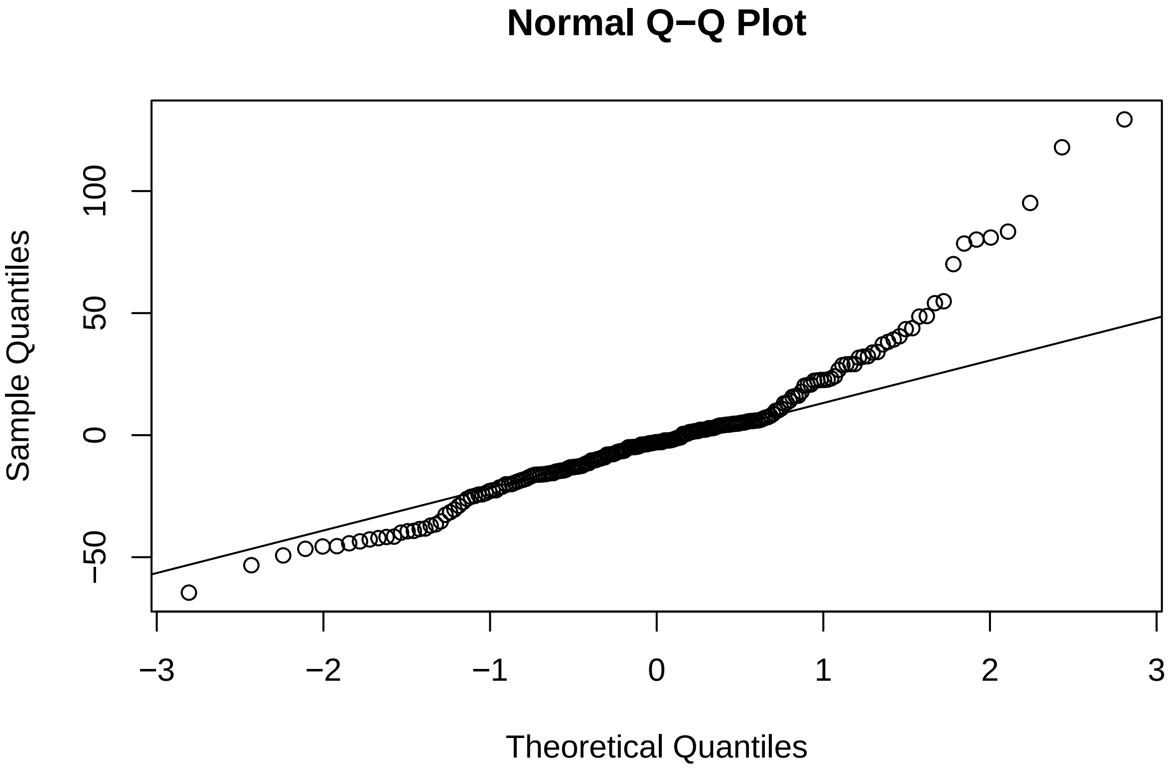


Clearly the assumption of equal variances of the error terms is not met.

Normality of Errors

We visually investigate the assumption of normality.

```
qqnorm(resid)
qqline(resid)
```



Clearly the normality assumption on the error terms is also not met.

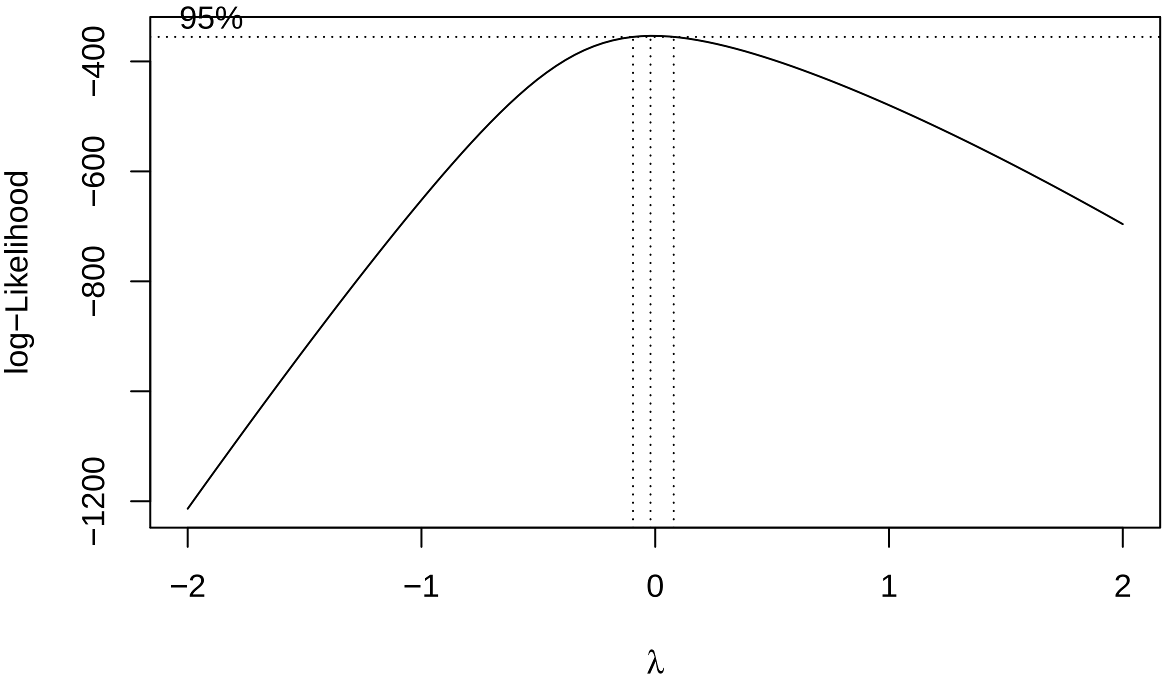
Transformation of the Response

To address the equal variance and normality of error term assumption violations, we attempt to find a transformation for the response variable using the BoxCox transformation method.

```
library("MASS")
```

```
## Warning: package 'MASS' was built under R version 3.2.2
```

```
boxcox(reg01)
```

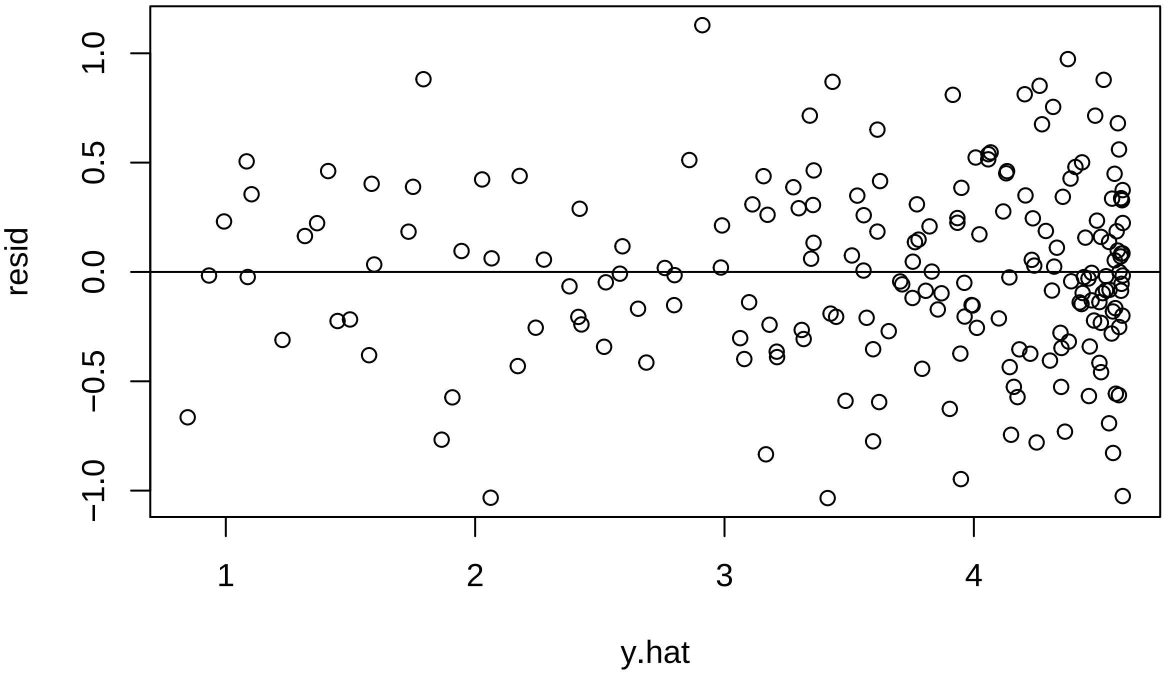


From the plot we see that the BoxCox technique points to a log-transformation (since $\lambda = 0$).

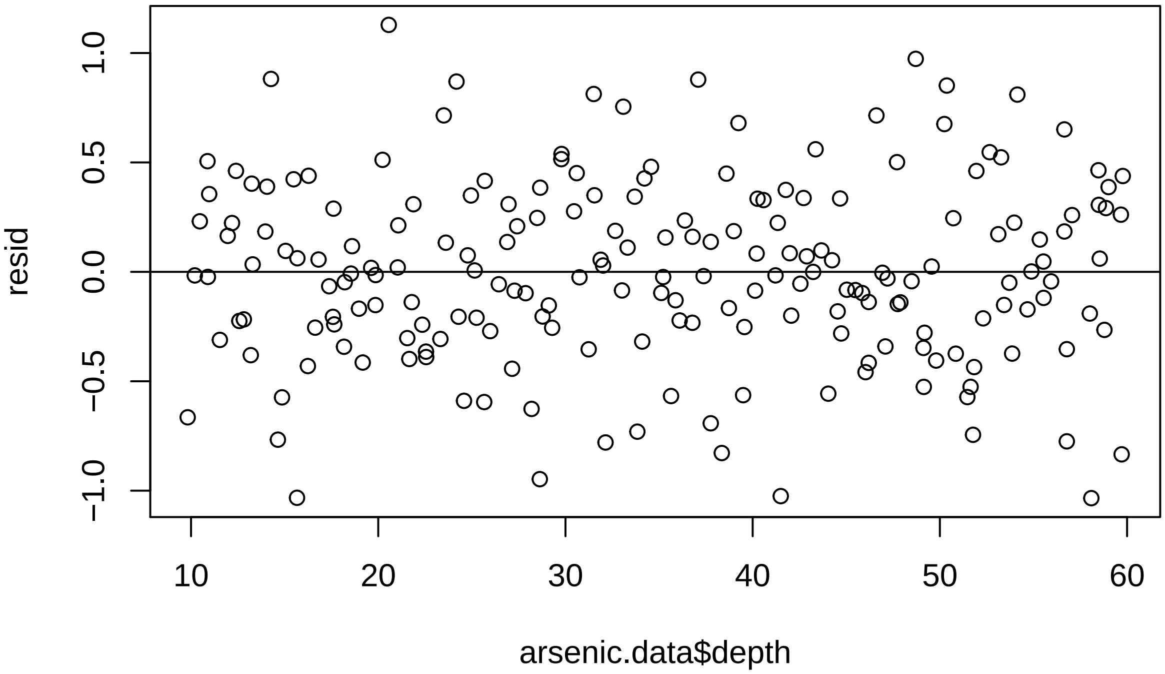
Fit of Transformed Response

We try fitting the data with the log-transformed response, and then perform assumption checks again.

```
arsenic.data$log.arsenic <- log(arsenic.data$arsenic)
reg02 <- lm(log.arsenic ~ depth + depth2 + depth3, data=arsenic.data)
resid <- residuals(reg02)
y.hat <- predict(reg02, arsenic.data)
plot(resid ~ y.hat)
abline(h=0)
```



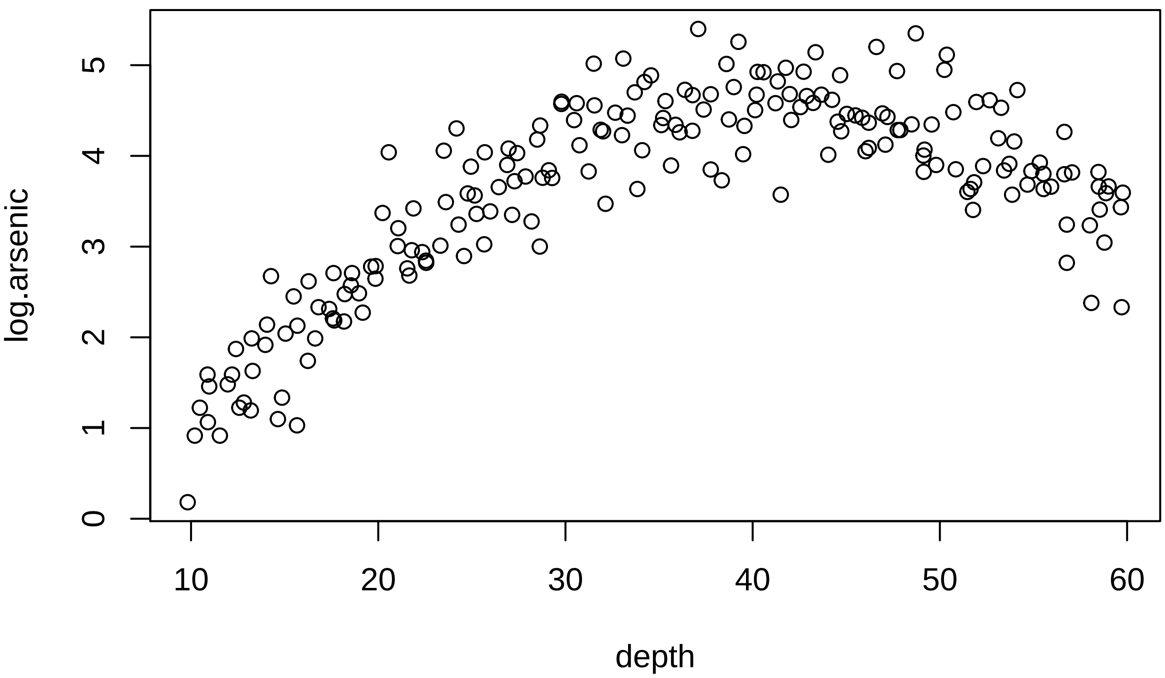
```
plot(resid ~ arsenic.data$depth)
abline(h=0)
```



There does not appear to be any equal variance assumption violation with the log-transformed response.

We examine a plot of the log-transformed response versus depth.

```
plot(log.arsenic ~ depth, data=arsenic.data)
```



The data appear much more consistent and looks as though a quadratic fit may be adequate.

Examination of the Fit with Transformed Response

We examine the regression fit using the transformed response.

```
summary(reg02)
```

```
##
## Call:
## lm(formula = log.arsenic ~ depth + depth2 + depth3, data = arsenic.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.03413 -0.25519 -0.01541  0.28952  1.12849
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.661e+00  4.052e-01  -4.099 6.07e-05 ***
## depth        2.834e-01  4.201e-02   6.747 1.66e-10 ***
## depth2      -2.747e-03  1.298e-03  -2.116   0.0356 *
## depth3      -1.082e-05  1.226e-05  -0.882   0.3786
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4183 on 196 degrees of freedom
## Multiple R-squared:  0.8538, Adjusted R-squared:  0.8516
## F-statistic: 381.7 on 3 and 196 DF,  p-value: < 2.2e-16
```

Indeed, the cubic term is not significant, so we refit the model without the cubic term.

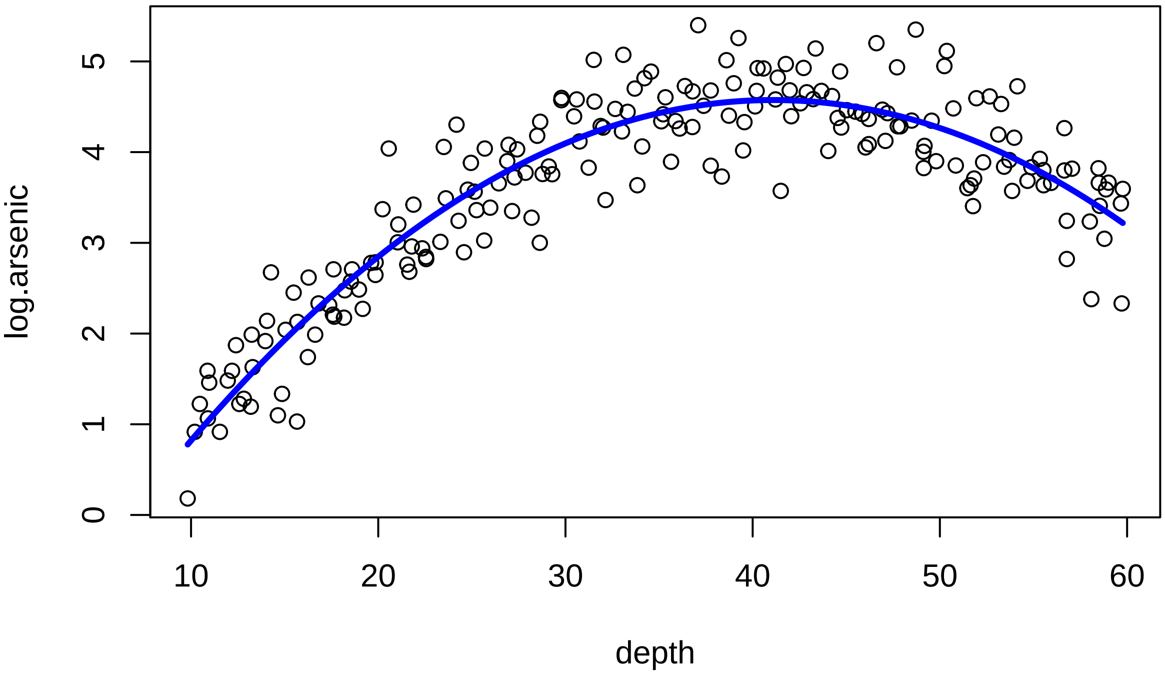
```
reg03 <- lm(log.arsenic ~ depth + depth2, data=arsenic.data)
summary(reg03)
```

```
##
## Call:
## lm(formula = log.arsenic ~ depth + depth2, data = arsenic.data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.07227 -0.25674 -0.00649  0.26950  1.10195
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.982502   0.177325  -11.18   <2e-16 ***
## depth        0.319152   0.011184   28.54   <2e-16 ***
## depth2      -0.003884   0.000157  -24.74   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4181 on 197 degrees of freedom
## Multiple R-squared:  0.8533, Adjusted R-squared:  0.8518
## F-statistic: 572.7 on 2 and 197 DF,  p-value: < 2.2e-16
```

And the quadratic fit has all the terms highly significant (p-value < 0.0001).

We now plot the fit with the data.

```
plot(log.arsenic ~ depth, data=arsenic.data)
x <- seq(min(arsenic.data$depth), max(arsenic.data$depth),
             length.out=100)
curve.data$log.arsenic <-predict(reg03, curve.data)
lines(x=curve.data$depth, y=curve.data$log.arsenic,
      col="blue", lwd=3)
```


log.arsenic
0
1
2
3
4
5
10
20
30
40
50
60
depth