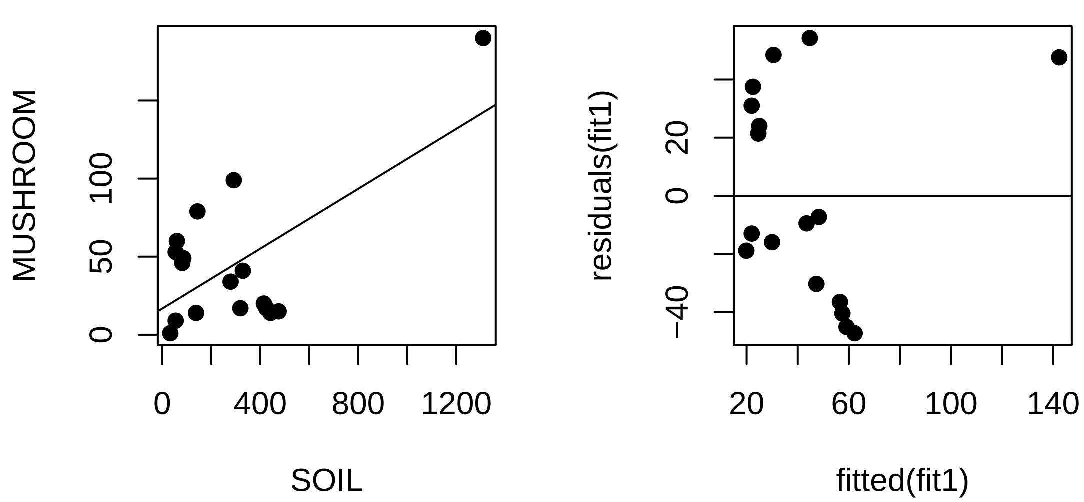
This is an annotated analysis of a data set on the concentrations of radioactive cesium in mushrooms and soil sampled from 17 wooded locations in central Italy at times within a couple years of the nuclear reactor accident in Chernobyl. The purpose of the study was to examine the uptake of radioactive cesium by mushrooms from the surrounding soil.

Here is the actual data set with concentrations of the radioactive isotopes of cesium (in Bq/kg) in soil and mushrooms.

```
> ex0818

   MUSHROOM SOIL
1         1   33
2         9   55
3        14  138
4        17  319
5        20  415
6        17  425
7        14  442
8        15  475
9        34  279
10       41  329
11       46   82
12       49   86
13       53   55
14       60   60
15       79  144
16       99  292
17      190 1310
```

Here is a plot of the data with a regression line fitted and a residual plot versus the fitted values.

```
> par(mfrow = c(1, 2))
> plot(SOIL, MUSHROOM, pch = 16)
> fit1 <- lm(MUSHROOM ~ SOIL)
> abline(fit1)
> plot(fitted(fit1), residuals(fit1), pch = 16)
> abline(h = 0)
```



We notice that the 17th observation is an outlier. It has unusually large measurements in both variables. The residual is not that unusual, but this can be an artifact. This one point will be highly influential and will tend to pull the regression line close to it to avoid having a single enormous squared residual.

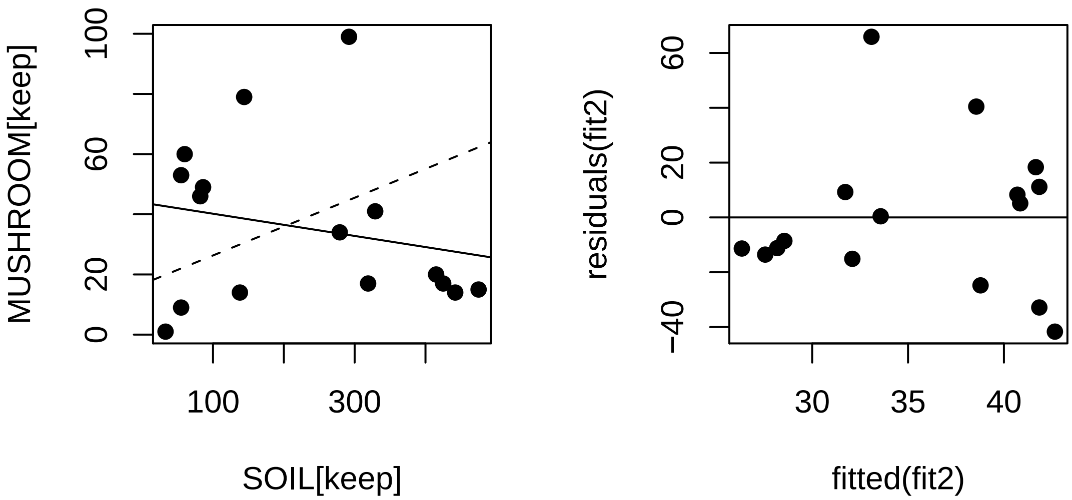
Here is a summary of the regression and a 95% confidence interval for the slope.

```
> sum1 <- summary(fit1)
> sum1

Call:
lm(formula = MUSHROOM ~ SOIL)

Residuals:
    Min      1Q  Median      3Q     Max 
-47.279 -30.319  -9.483  31.000  54.271 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept) 16.72569   12.41954   1.347  0.19807   
SOIL         0.09590    0.02993   3.205  0.00591 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 

Residual standard error: 36.56 on 15 degrees of freedom
Multiple R-Squared: 0.4064,       Adjusted R-squared: 0.3668 
F-statistic: 10.27 on 1 and 15 DF,  p-value: 0.005909 

> df1 <- df.residual(fit1)
> t.crit <- qt(0.975, df1)
> t.crit

[1] 2.131450

> t.crit * (coef(sum1))[2, 2]

[1] 0.06378854
```

A 95% confidence interval for the slope $\beta_1$ is $0.096 \pm 0.064$.

We ought to be concerned about the effect of the outlier. One possibility to consider is to see if the estimated regression line is substantially different if the point is excluded from the analysis.

```
> par(mfrow = c(1, 2))
> keep <- 1:16
> plot(SOIL[keep], MUSHROOM[keep], pch = 16)
> fit2 <- lm(MUSHROOM[keep] ~ SOIL[keep])
> abline(fit2)
> abline(fit1, lty = 2)
> plot(fitted(fit2), residuals(fit2), pch = 16)
> abline(h = 0)
```




Notice that the new regression line is substantially different from the previous regression line. Here is a summary of the regression and a new 95% confidence interval for the slope.

```
> sum2 <- summary(fit2)
> sum2

Call:
lm(formula = MUSHROOM[keep] ~ SOIL[keep])

Residuals:
    Min      1Q  Median      3Q     Max
-41.658 -13.938  -4.061   9.744  65.908

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 43.87726   12.25571   3.580  0.00301 **
SOIL[keep]  -0.03693    0.04454  -0.829  0.42085
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 27.76 on 14 degrees of freedom
Multiple R-Squared: 0.04682,        Adjusted R-squared: -0.02126
F-statistic: 0.6877 on 1 and 14 DF,  p-value: 0.4208

> df2 <- df.residual(fit2)
> t.crit <- qt(0.975, df2)
> t.crit

[1] 2.144787

> t.crit * (coef(sum2))[2, 2]

[1] 0.0955225
```

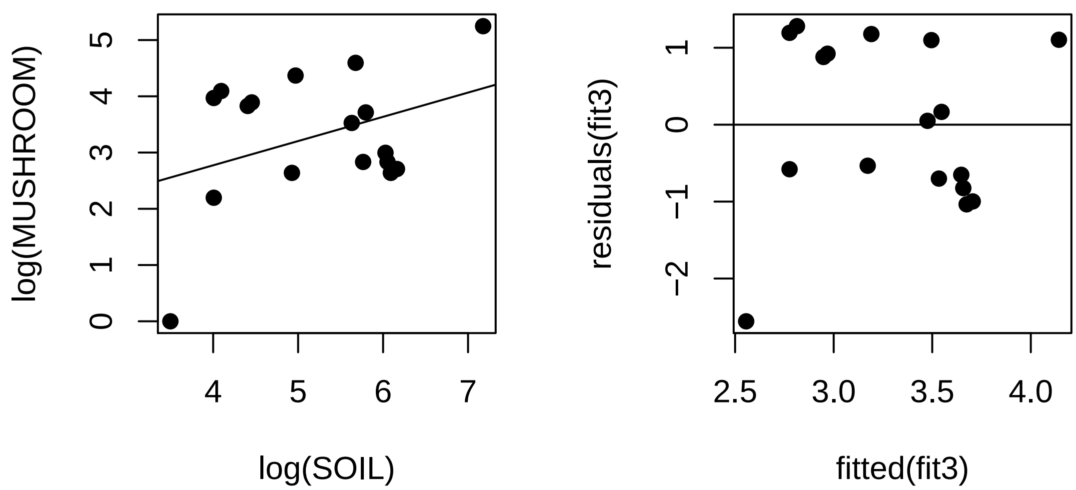
A 95% confidence interval for the slope $\beta_1$ is $-0.037 \pm 0.096$.

The inferences on the relationship between radioactive cesium concentration in soil and in mushrooms is now completely different. With the outlier included, there is strong evidence that the concentration of radioactive cesium in mushrooms increases as the concentration in the soil increases. When this point is excluded, there is no evidence of a relationship.

All we can really conclude is that for the one data point in which the soil concentration was quite high, the mushroom concentration was very high as well. For lower concentrations, there is very little evidence of a strong relationship.

An alternative analysis would be to analyze concentrations on the log scale. Here are the plots.

```
> par(mfrow = c(1, 2))
> plot(log(SOIL), log(MUSHROOM), pch = 16)
> fit3 <- lm(log(MUSHROOM) ~ log(SOIL))
> abline(fit3)
> plot(fitted(fit3), residuals(fit3), pch = 16)
> abline(h = 0)
```




On this scale, the sample with the smallest soil concentration is a bit of an outlier, but we do not see much of a relationship. The slope is not significant on the log scale.

```
> summary(fit3)

Call:
lm(formula = log(MUSHROOM) ~ log(SOIL))

Residuals:
     Min       1Q   Median       3Q      Max 
-2.55598 -0.70057  0.05032  1.09945  1.28070 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.0490     1.4896   0.704    0.492
log(SOIL)     0.4310     0.2804   1.537    0.145

Residual standard error: 1.136 on 15 degrees of freedom
Multiple R-Squared: 0.1361,        Adjusted R-squared: 0.07847 
F-statistic: 2.362 on 1 and 15 DF,  p-value: 0.1451
```