MIE237

Neil Montgomery

2016-02-05

regression

another example

From the 2015 exam: "A gas distribution company is concerned with the health of the gas meters being used by its industrial customers. There are over 20,000 such customers, which is too many to visit to examine each gas meter. So they select a sample of n=400 meters from the database and send technicians to visit only these meters. The technicians perform an analysis of each meter, which includes some testing, and record the following data (along with the meter ID)"

- `max_kpa`: maximum test pressure in kPa
- `volts`: result of an electric test run through meter in V
- …a bunch of others.

(almost) the question

The simple regression model with `volts` as the response and `max_kpa` as the input is fit resulting in the following output with some numbers removed. Recall that the sample variance of the `volts` readings is 1.4011 and the sample size is 400.

Questions: fill in the ANOVA table (sum of squares decomposition etc.) and do the hypothesis test for $H_0 : \beta_1 = 0$ versus $H_1 : \beta_1 \neq 0$.

the modified output

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.761059   1.054718 -14.943   <2e-16
max_kpa       MISSING    0.004506 MISSING  MISSING
---

Residual standard error: 1.066 on 398 degrees of freedom
Multiple R-squared:  0.1913,    Adjusted R-squared:  0.1893
F-statistic: 94.14 on MISSING and MISSING DF,  p-value: MISSING
```

And I gave an "analyis of variance" table (with the sum of squares decomposition etc.) with missing.

unmodified output

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -15.761059   1.054718 -14.943   <2e-16 ***
max_kpa       0.043719   0.004506   9.703   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.066 on 398 degrees of freedom
Multiple R-squared:  0.1913,    Adjusted R-squared:  0.1893
F-statistic: 94.14 on 1 and 398 DF,  p-value: < 2.2e-16


## Analysis of Variance Table
##
## Response: volts
##            Df Sum Sq Mean Sq F value    Pr(>F)
## max_kpa     1 106.94 106.940  94.142 < 2.2e-16 ***
## Residuals 398 452.11   1.136
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Model assumptions and diagnostic plots

Model: $Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ with $\varepsilon_i \sim N(0, \sigma^2)$

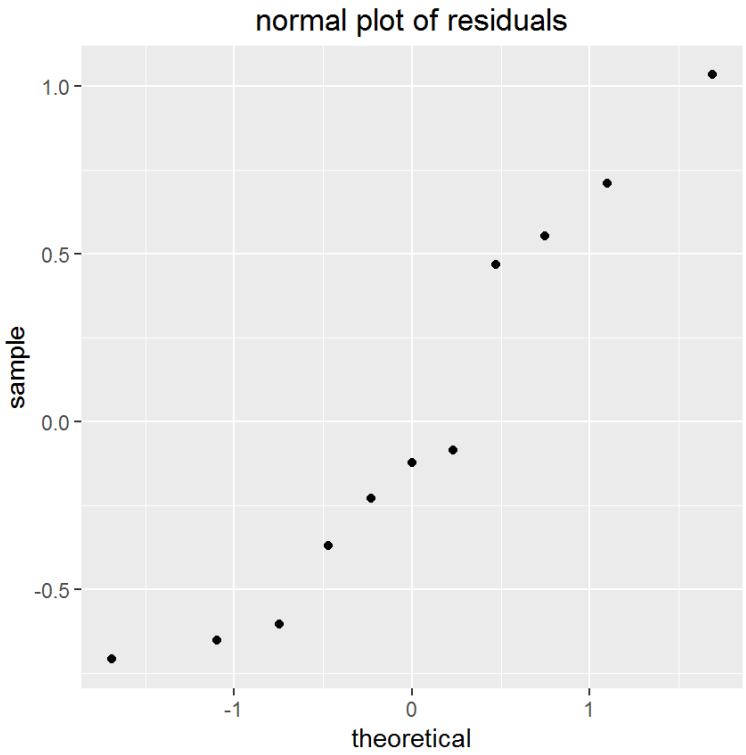
Assumption 1 Normal error. Evaluate using normal quantile plot of the residuals $\hat{e}_i = y_i - \hat{y}_i$.

Assumption 2 Equal variance. Evaluate using scatterplot of residuals $\hat{e}_i$ (horizontal) versus fitted values $\hat{y}_i$ (vertical)

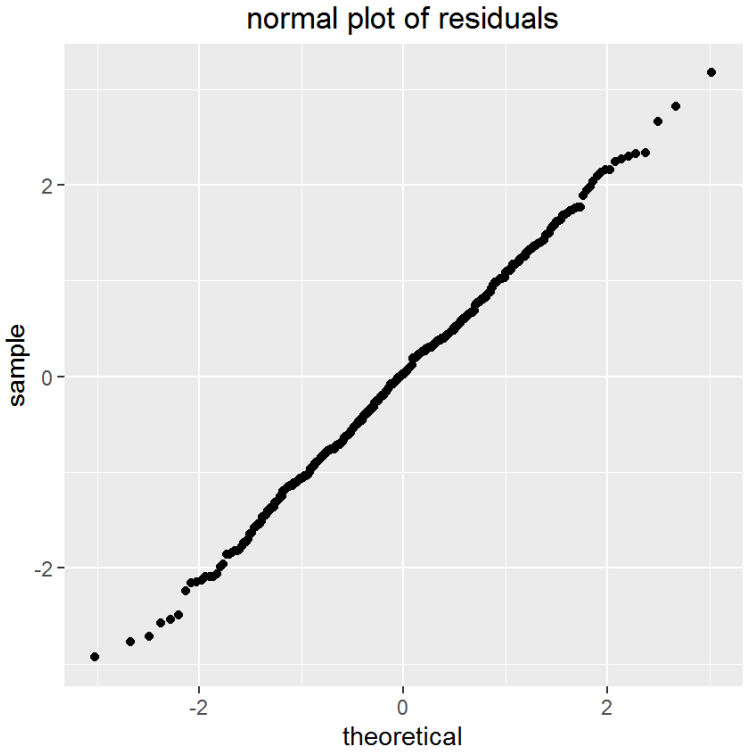
Assumption 3 Linear relationship between x and y. (Not exactly an assumption, per se, but we'll call it one.) Evaluation using scatterplot of residuals $\hat{e}_i$ (vertical) versus fitted values $\hat{y}_i$ (horizontal)

use the plot to detect problems.

Sugar example

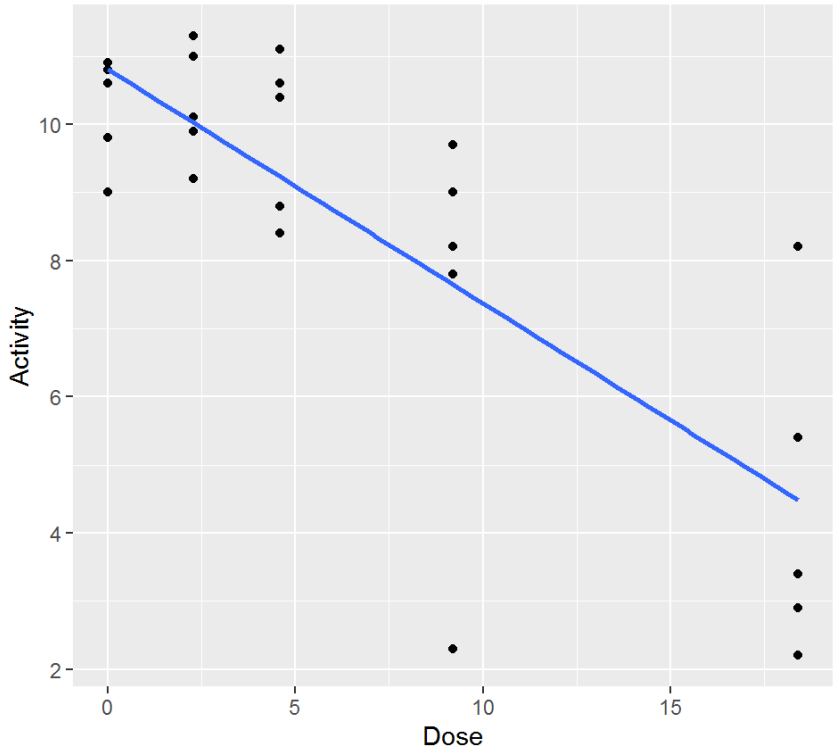


Exam example `volts` versus `kpa_max`



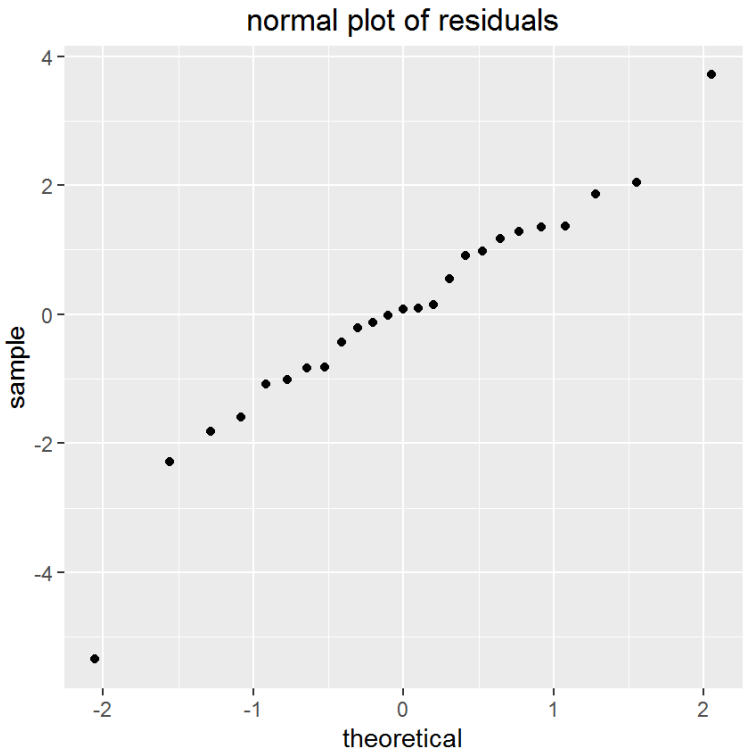
11.39 from the textbook

Data: textbook question 11.39. Studies the effect of organophosphate doseon mouse brain activity.

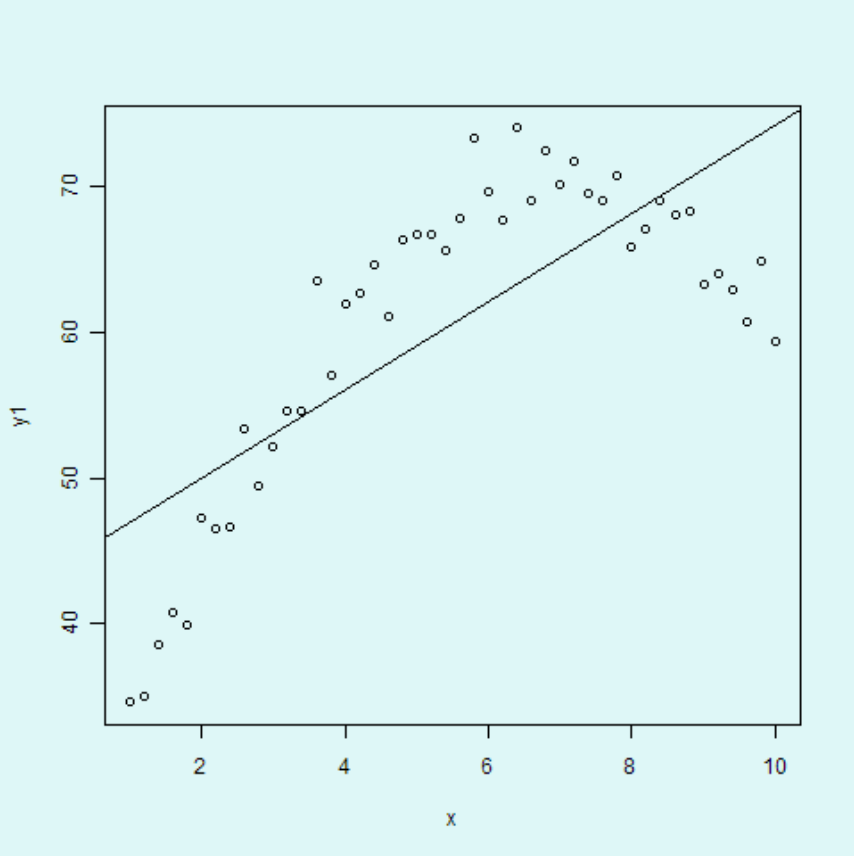


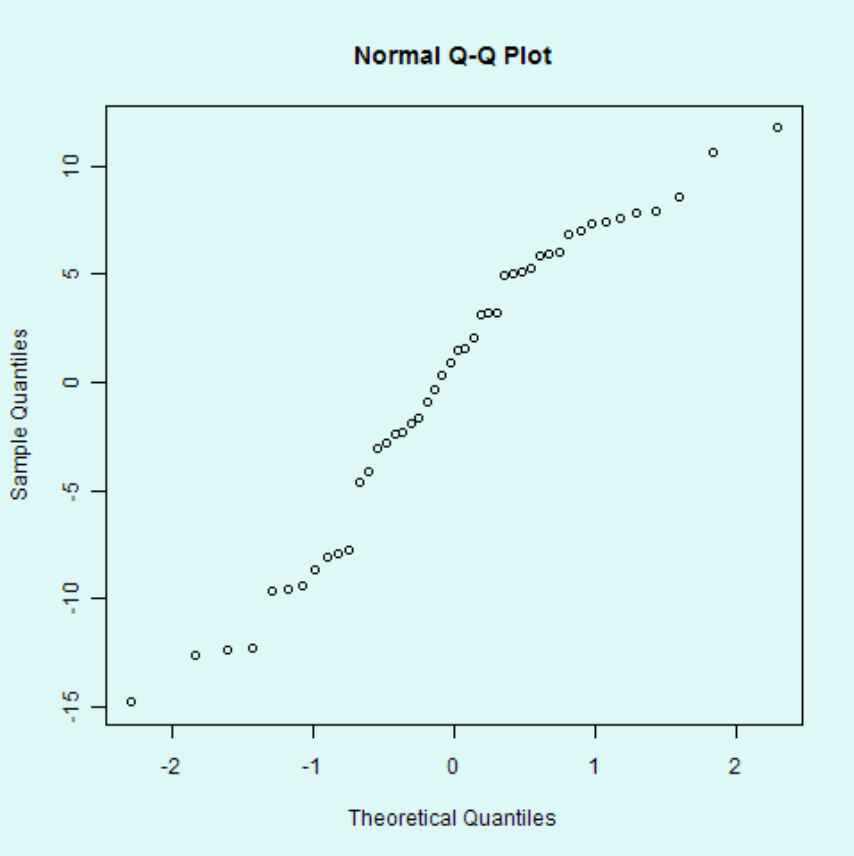
term	estimate	std.error	statistic	p.value
(Intercept)	10.811	0.522	20.705	0.000
Dose	-0.344	0.055	-6.242	0.000

mouse data "residual plots"

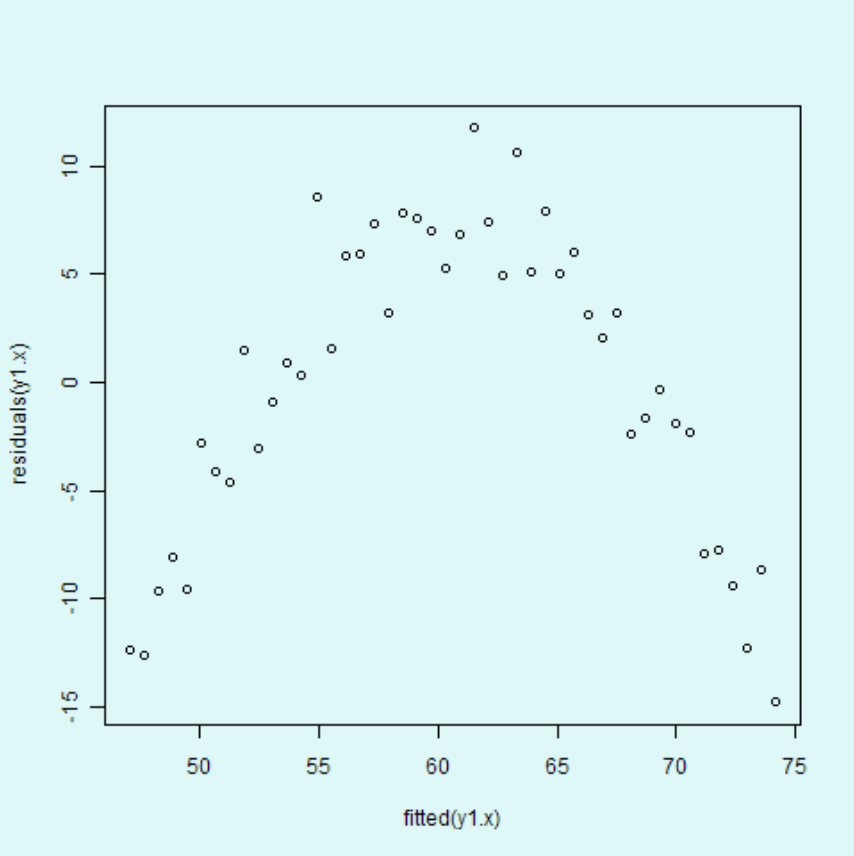


More plots - dataset I (obvious nonlinear)

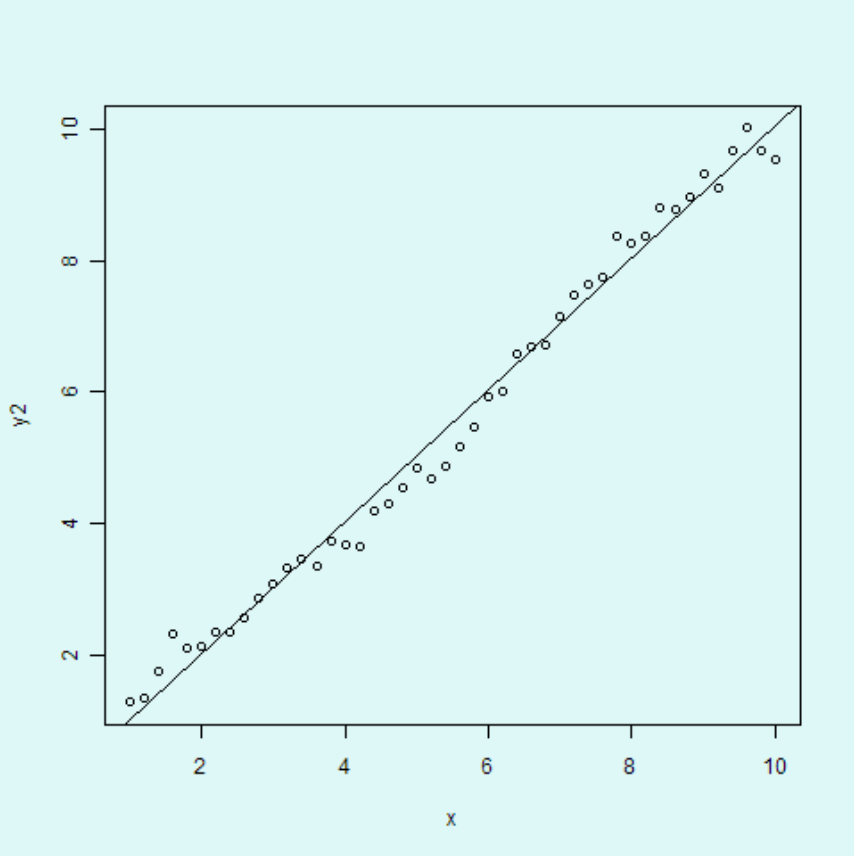


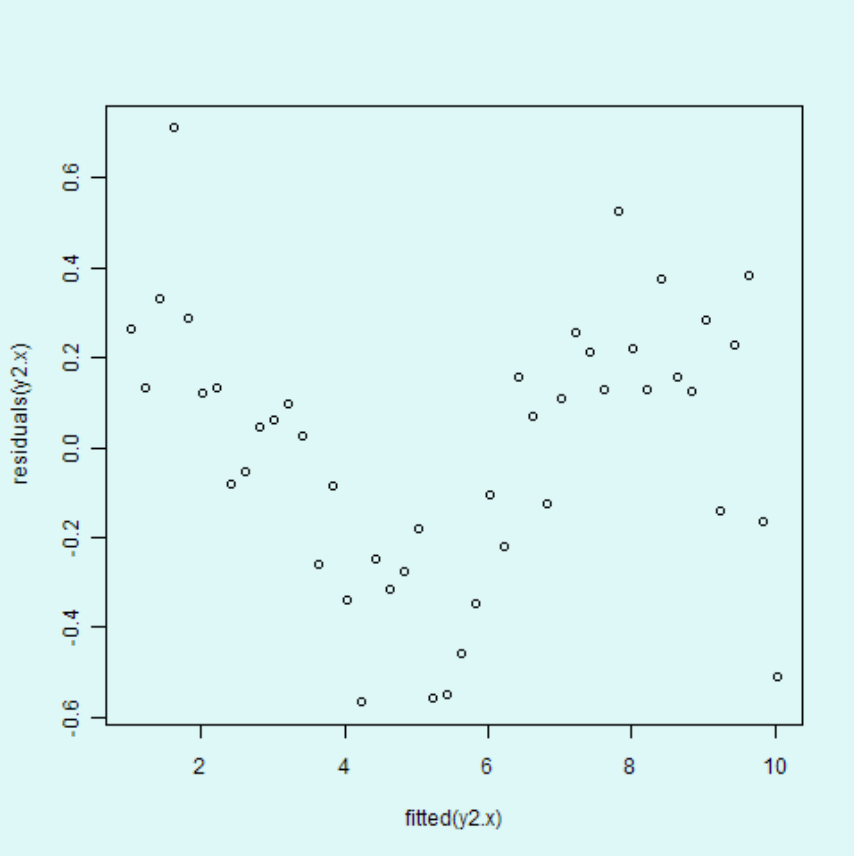


More plots - dataset I



More plots - dataset II





More plots - dataset III (unequal variance)