UPPSALA UNIVERSITY
Department of Mathematics
Rolf Larsson

Exam in Mathematical Statistics
Inference 1 1MS035
2021–01–11
Solutions

1. We have a random sample 1,3,4,3,4 from a random variable X with probability function

$$P(X=k)=\begin{cases}1-10\theta, & k=0,\\ \theta k, & k=1,2,3,4,\\ 0, & \text{otherwise,}\end{cases}$$

where $0\le\theta\le 0.1$. Estimate θ using

(a) the method of moments, (1p)

Solution: We observe $\bar{x}=15/5=3$, and we have

$$m(\theta)=E(X)=\sum_{k=0}^{4}kP(X=k)=1*\theta+2*(2\theta)+3*(3\theta)+4*(4\theta)=30\theta.$$

Hence, solving $3=30\theta$ gives the moment estimate $\theta^*=0.1$.

(b) the least squares method, (2p)

Solution: Say that the observations are $x_1,...,x_n$. Because $m(\theta)=30\theta$, we need to minimize

$$Q(\theta)=\sum_{i=1}^{n}(x_i-30\theta)^2.$$

To this end, we calculate the derivatives

$$Q'(\theta)=-60\sum_{i=1}^{n}(x_i-30\theta)=-60n\bar{x}+1800n\theta$$

and

$$Q''(\theta)=1800n.$$

Because $Q''(\theta)>0$, we get a minimum by solving $0=Q'(\theta)$,
i.e. $\theta=\bar{x}/30$.
Hence, in our case, the LSE is $\theta^*=3/30=0.1$, the same as the moment estimate.

(c) maximum likelihood. (2p)

Solution: Because we have a discrete distribution, the likelihood equals the probability of observing the sample. Hence, it is, by independence,

$$\begin{aligned}L(\theta)&=P(X_1=1,X_2=3,X_3=4,X_4=3,X_5=4)\\&=\theta*(3\theta)*(4\theta)*(3\theta)*(4\theta)=144\theta^5.\end{aligned}$$

Observe that by assumption, $0 \leq \theta \leq 0.1$. Since the likelihood is an increasing function of θ, we must have that its maximum in this interval is attained at its right endpoint, i.e. at $\theta = 0.1$. Hence, the MLE is $\theta^* = 0.1$, the same as the moment estimate and the LSE.

2. We have a random sample x_1, x_2 from a random variable X with expectation μ and variance σ^2, and another random sample $y_1, y_2, ..., y_5$ from a random variable Y with expectation 5μ and variance $5\sigma^2$. We may assume that X and Y are independent. The sample means are denoted $\bar{x}$ and $\bar{y}$.

   The following estimates of μ are proposed:

$$\mu_1^* = \frac{\bar{x} + \bar{y}}{6}, \quad \mu_2^* = \frac{5\bar{x} + \bar{y}}{10}.$$

   (a) Show that μ_1^* and μ_2^* are both unbiased. (1p)

   Solution: The expectations of the corresponding estimators are

$$E(\mu_1^*) = E\left(\frac{1}{6}\overline{X} + \frac{1}{6}\overline{Y}\right) = \frac{1}{6}E(\overline{X}) + \frac{1}{6}E(\overline{Y}) = \frac{1}{6} * \mu + \frac{1}{6} * 5\mu = \mu$$

   and

$$E(\mu_2^*) = E\left(\frac{5}{10}\overline{X} + \frac{1}{10}\overline{Y}\right) = \frac{5}{10}E(\overline{X}) + \frac{1}{10}E(\overline{Y}) = \frac{5}{10} * \mu + \frac{1}{10} * 5\mu = \mu.$$

   Hence, both are unbiased.

   (b) Which one of μ_1^* and μ_2^* is most efficient? (2p)

   Solution: Observe that

$$V(\overline{X}) = \frac{V(X)}{2} = \frac{1}{2}\sigma^2$$

   and

$$V(\overline{Y}) = \frac{V(Y)}{5} = \frac{5\sigma^2}{5} = \sigma^2.$$

   Hence, the variances of the estimators are

$$\begin{aligned} V(\mu_1^*) &= V\left(\frac{1}{6}\overline{X} + \frac{1}{6}\overline{Y}\right) = \left(\frac{1}{6}\right)^2 V(\overline{X}) + \left(\frac{1}{6}\right)^2 V(\overline{Y}) \\ &= \frac{1}{36} * \frac{1}{2}\sigma^2 + \frac{1}{36} * \sigma^2 = \frac{3}{72}\sigma^2 \approx 0.042\sigma^2 \end{aligned}$$

   and

$$\begin{aligned} V(\mu_2^*) &= V\left(\frac{5}{10}\overline{X} + \frac{1}{10}\overline{Y}\right) = \left(\frac{5}{10}\right)^2 V(\overline{X}) + \left(\frac{1}{10}\right)^2 V(\overline{Y}) \\ &= \frac{25}{100} * \frac{1}{2}\sigma^2 + \frac{1}{100} * \sigma^2 = \frac{27}{200}\sigma^2 = 0.135\sigma^2. \end{aligned}$$

   Hence, $V(\mu_1^*) < V(\mu_2^*)$ for all σ^2, and so, μ_1^* is most efficient.

(c) Is there any other unbiased estimate on the form $a\bar{x} + b\bar{y}$ which is more efficient than both μ_1^* and μ_2^*? (2p)

Solution: At first, observe that

$$E(a\overline{X} + b\overline{Y}) = aE(\overline{X}) + bE(\overline{Y}) = a\mu + b * 5\mu = (a + 5b)\mu,$$

so for this to equal μ, we must have $a = 1 - 5b$. Moreover,

$$V(a\overline{X} + b\overline{Y}) = a^2V(\overline{X}) + b^2V(\overline{Y}) = a^2 * \frac{1}{2}\sigma^2 + b^2\sigma^2 = \left(\frac{a^2}{2} + b^2\right)\sigma^2,$$

and with $a = 1 - 5b$, we have $V(a\overline{X} + b\overline{Y}) = g(b)\sigma^2$, where

$$g(b) = \frac{1}{2}(1 - 5b)^2 + b^2 = \frac{1}{2} - 5b + \frac{27}{2}b^2.$$

Now, find the b that minimizes $g(b)$. Differentiation yields

$$g'(b) = -5 + 27b, \quad g''(b) = 27 > 0,$$

and so, we get the minimum by solving $0 = g'(b)$, which yields $b = 5/27$. This means $a = 1 - 5b = 2/27$, and we get the estimate

$$\hat{\mu} = \frac{2\bar{x} + 5\bar{y}}{27}.$$

This is distinct from μ_1^* and μ_2^*, so the answer is no!
Also, observe that for the corresponding estimator, we have

$$\begin{aligned} V(\hat{\mu}) &= V\left(\frac{2}{27}\overline{X} + \frac{5}{27}\overline{Y}\right) = \left(\frac{2}{27}\right)^2 V(\overline{X}) + \left(\frac{5}{27}\right)^2 V(\overline{Y}) \\ &= \frac{4}{27^2} * \frac{1}{2}\sigma^2 + \frac{25}{27^2} * \sigma^2 = \frac{1}{27}\sigma^2 \approx 0.037\sigma^2, \end{aligned}$$

which is seen to be smaller than the variances in (b).

3. Barry arranges a lottery with a large number of ballots (lotter). He claims that on average, one ballot out of five is a win.

   Penny buys 25 ballots, but no one of them is a win.

   (a) Is Barry telling the truth, or is he cheating? Try to find this out by performing a hypothesis test. (1p)

   Solution: Because the total number of ballots is large, it is reasonable to say that the number of wins for Penny is $X \sim \text{Bin}(25, p)$.

   According to Barry, $p = 1/5 = 0.2$. If he is cheating, we have $p < 0.2$. Hence, it is natural to test H_0: $p = 0.2$ vs H_1: $p < 0.2$.

   The observation is $x = 0$. By the direct method, the P value is

   $$P(X \leq 0; p = 0.2) = P(X = 0; p = 0.2) = 0.8^{25} \approx 0.0038.$$

   Because $0.0038 < 0.01$, we may reject H_0 at the 1% level. On this risk level, we have proof that Barry is cheating.

   (b) What is the critical region of the test in (a) if it is performed on the 5% level? (2p)

   Solution: Because we reject for small x, the critical region must have the form $C = \{x \leq K\}$ for some K. For a 5% level test, K is the largest integer such that $P(X \leq K; p = 0.2) \leq 0.05$.

   We saw in (a) that $K = 0$ fulfills the condition. Moreover, we have

   $$P(X \leq 1; p = 0.2) = 0.8^{25} + 25 * 0.2 * 0.8^{24} \approx 0.027 < 0.05,$$

   but

   $$P(X \leq 2; p = 0.2) = 0.8^{25} + 25 * 0.2 * 0.8^{24} + \binom{25}{2} * 0.2^2 * 0.8^{23}$$
   $$\approx 0.098 > 0.05.$$

   Hence, we choose $K = 1$, i.e. the critical region is $C = \{x \leq 1\}$.

   (c) What is the power of the 5% level test if in fact, on average only one ballot out of 50 is a win? (2p)

   Solution: This corresponds to $p = 1/50 = 0.02$. The power is the probability to reject H_0 for this p, i.e.

   $$P(X \leq 1; p = 0.02) = 0.98^{25} + 25 * 0.02 * 0.98^{24} \approx 0.91.$$

4. Helga goes by bicycle to work. She has two routes, A and B. She wants to know if the routes are equally fast. Everyday, she flips a coin to decide which route to take. After ending the experiment, she got the data given in the table below. The times are given in minutes and parts of minutes (hence not seconds).

Route A	10.27	10.72	10.36	10.01	10.25
Route B	9.86	9.99	10.02	10.22	9.52

Are the routes equally fast? Try to answer this question by performing a suitable statistical test. Make sure to specify all your assumptions. (5p)

Solution: We have two random samples that are not related pairwise (the routes are taken on different days). It is reasonable to assume that the times are normally distrubuted.

So, we have a random sample $x_1, ..., x_5$ from $X \sim N(\mu_1, \sigma_1^2)$ and a random sample $y_1, ..., y_5$ from $Y \sim N(\mu_2, \sigma_2^2)$, where X and Y are independent. The variances σ_1^2 and σ_2^2 are unknown but considered equal, i.e. $\sigma_1^2 = \sigma_2^2 = \sigma^2$, where σ^2 is unknown.

We want to test H_0: $\mu_1 = \mu_2$ vs H_1: $\mu_1 \neq \mu_2$. (There is nothing in the formulation of the problem that indicates that we should use a one-sided test.) We estimate s^2 by the pooled variance s_p^2, where since $s_x^2 = 0.06627$ and $s_y^2 = 0.06712$, we get ($n_1 = n_2 = 5$)

$$s_p^2 = \frac{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}{n_1 + n_2 - 2} = \frac{4 * 0.06627 + 4 * 0.06712}{8} = 0.066695.$$

Moreover, $\bar{x} = 10.322$ and $\bar{y} = 9.922$.

The observed test statistic is

$$T_{obs} = \frac{\bar{x} - \bar{y}}{s_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{10.322 - 9.922}{\sqrt{0.066695}\sqrt{\frac{1}{5} + \frac{1}{5}}} \approx 2.45 > t_{0.025}(8) = 2.3060,$$

and so, we may reject H_0 at the 5% level.

We have evidence at the 5% level that the routes are not equally fast.

5. A company produces a kind of electrical units. The number of defect units for one day is supposed to follow the Poisson distribution with parameter μ.

   One day, the company produced 45 defective units.

   (a) Calculate a 99% confidence interval for μ. (2p)

   Solution: We have one observation $x = 45$ of $X \sim \text{Po}(\mu)$. Because $E(X) = \mu = V(X)$, the reference variable is

   $$\frac{X - \mu}{\sqrt{\mu^*}} \approx N(0, 1)$$

   where $\mu^* = x$. This gives the 99% confidence interval ($\lambda_{0.005} = 2.5758$)

   $$I_\mu = x \pm \lambda_{0.005}\sqrt{x} = 45 \pm 2.5758 * \sqrt{45} = 45 \pm 17.3 = (27.7,\ 62.3).$$

   Finally, we check the rule-of-thumb for normal approximation, which is that the lower endpoint of the interval, 27.7, is greater than 15.

   (b) Test H_0: $\mu = 30$ vs H_1: $\mu \neq 30$ at the level 1% by using the test variable method. (2p)

   Solution: With $\mu_0 = 30$, the test variable is

   $$T = \frac{X - \mu_0}{\sqrt{\mu}_0} \approx N(0, 1),$$

   where the rule-of-thumb for normal approximation is fulfilled here, since we have $\mu_0 = 30 > 15$. Hence, we get

   $$T_{obs} = \frac{x - \mu_0}{\sqrt{\mu}_0} = \frac{45 - 30}{\sqrt{30}} \approx 2.74 > 2.5758 = \lambda_{0.005},$$

   and so, we reject H_0 at the 1% level.

   (c) Compare the results in (a) and (b). Do they lead to the same conclusion? Explain. (1p)

   Solution: We could have performed the test in (b) by using the confidence method, i.e. by checking if $\mu = 30$ belongs to the confidence interval. In fact, it does, and this corresponds to not rejecting H_0, which contradicts the result in (b)!

   Sometimes, illogical things like this may happen. The reason is that we have performed normal approximation differently in (a) and (b).

   Using methods which do not involve any distribution approximations, we should not run into contradictions like this one.

6. Before the US president election, opinion polls were held in different states. For the state of Arizona, the poll of Survey/Minkey/Axios, Oct 20- Nov 2, gave Donald Trump 46% of the sympathies. The number of asked voters was 4278. At the same time, a corresponding poll in New Jersey gave Donald Trump 38% of the sympathies, where the number of asked voters were 3870.

   (a) Calculate a 95% confidence interval for the difference of proportions of Trump voters for the two states. (4p)

   Solution: Let X and Y be the numbers of Trump voters in the polls of Arizona and New Jersey, respectively. Because the numbers of potential voters are so large, it is reasonable to say that $X \sim \text{Bin}(n_1, p_1)$ and $Y \sim \text{Bin}(n_2, p_2)$, where $n_1 = 4278$, $n_2 = 3870$ and p_1 and p_2 are the unknown proportions of Trump voters in the two states.

   We want to construct a 95% confidence interval for $p_1 - p_2$. To this end, we may use the reference variable

$$\frac{\frac{X}{n_1} - \frac{Y}{n_2} - (p_1 - p_2)}{d} \approx N(0,1),$$

   where

$$d = \sqrt{\frac{p_1^*(1-p_1^*)}{n_1} + \frac{p_2^*(1-p_2^*)}{n_2}},$$

   with $p_1^* = x/n_1$, $p_2^* = y/n_2$. In our case, we are given the estimates $p_1^* = 0.46$ and $p_2^* = 0.38$. Inserting these, we get $d \approx 0.0109$.

   Checking the rule-of-thumb for normal approximation yields
   $n_1 p_1^*(1-p_1^*) = 4278 * 0.46 * 0.54 \approx 1063 > 5$ and
   $n_2 p_2^*(1-p_2^*) = 3870 * 0.38 * 0.62 \approx 912 > 5$.

   Hence, normal approximation is permitted.

   We get the 95% confidence interval

$$\begin{aligned} I_{p_1-p_2} &= p_1^* - p_2^* \pm \lambda_{0.025} d = 0.46 - 0.38 \pm 1.96 * 0.0109 \\ &= 0.08 \pm 0.02 = (0.06,\ 0.10). \end{aligned}$$

   (b) Were the proportions of Trump voters different in the two states? (1p)

   Solution: Because 0 does not belong to the confidence interval, we have evidence that the proportions were different on the risk level 5%.

   Comment: The election later in November gave the (preliminary) results 49.1% for Trump in Arizona and 41.4% in New Jersey. Note that both numbers were severely underestimated by the polls, but the difference remained about the same.

7. At the North Pole, where Santa Claus (jultomten) lives, his pixies (tomtenissar) have produced 10 000 baby dolls. These dolls are supposed to be identical, but nevertheless, their weights vary a little bit. The weights of the 10 000 baby dolls can be seen as a random sample with observed mean $\bar{x} = 237.00$ g and observed standard deviation $s = 1.20$ g.

   (a) Calculate a 95% confidence interval for the weight of a baby doll of the type produced by the pixies. (1p)

   Solution: We have a random sample of observed weights, $x_1, ..., x_n$, where $n = 10000$, from X, which is a random variable with $E(X) = \mu$ and $V(X) = \sigma^2$, say. We may estimate σ^2 by $s^2 = 1.2^2 = 1.44$. By normal approximation, we have that the reference variable

$$\frac{\bar{X} - \mu}{s/\sqrt{n}} \approx N(0, 1).$$

   (The advice in this kind of situtations is to use the $t(n-1)$ distribution, but since n is so large here, this distribution is extremely close to normal.) With $\bar{x} = 237.00$, this gives us the 95% confidence interval

$$I_\mu = \bar{x} \pm \lambda_{0.025} \frac{s}{\sqrt{n}} = 237 \pm 1.96 * 1.2\sqrt{10000} = 237 \pm 0.02 = (236.98,\ 237.02).$$

   (b) Santa gets angry with the pixies if the proportion of dolls with smaller weight than 235 g is greater than 10%. Calculate a suitable confidence interval to try to conclude if Santa will get angry this time. You may assume that the weight of a baby doll is normally distributed. (4p)

   Solution: Let $p = P(X < 235)$. The interesting thing here is to see if we can exclude the possibility that $p > 0.1$. One way to do this is to construct a one-sided confidence interval for p of the type $(0, p_u)$ and then check if $p_u < 0.1$, in which case we have evidence that we do not have $p > 0.1$.
   To accomplish this, we start by constructing a one-sided confidence interval for μ as (μ_l, ∞), where

$$\mu_l = \bar{x} - \lambda_{0.05} \frac{s}{\sqrt{n}} = 237 - 1.6449 * \frac{1.2}{\sqrt{10000}} \approx 236.98.$$

   We may assume that $X \sim N(\mu, \sigma^2)$, where we may safely insert the approximation $\sigma^2 \approx s^2 = 1.44 = 1.2^2$. Now, note that

$$p = P(X < 235) \approx P\left(\frac{X - \mu}{1.2} \leq \frac{235 - \mu}{1.2}\right) \approx \Phi\left(\frac{235 - \mu}{1.2}\right).$$

   Since this is a monotonely decreasing function of μ, we have that $\mu_l \leq \mu$ corresponds to

$$p \approx \Phi\left(\frac{235 - \mu}{1.2}\right) \leq \Phi\left(\frac{235 - \mu_l}{1.2}\right) = p_u,$$

and inserting $\mu_l = 236.98$ from above, we obtain

$$p_u = \Phi\left(\frac{235 - 236.98}{1.2}\right) = \Phi(-1.65) \approx 0.05.$$

Hence, we have that $(0,\ 0.05)$ is a one-sided confidence interval for p with approximate confidence level 95%. So on the risk level 5%, we find no evidence that $p > 0.1$. Santa should not get angry for this reason.

8. The production of wind energy in Sweden (in GWh) over the years 1993-2015 is plotted in figure 1. As is seen from this figure, the production increases exponentially over time. In figure 2, instead the (natural) log of the production is plotted.

   Let Y_t be the log of the wind energy production in Sweden for year t, and consider the regression model

   $$Y_t = \alpha + \beta t + \varepsilon_t,$$

   where $t = 1993, 1994, ..., 2015$. The ε_t are assumed to be independent $N(0, \sigma^2)$.

   This model was run for the data, and the following R output was produced:

```
Call:
lm(formula = y ~ t)

Residuals:
     Min       1Q   Median       3Q      Max
-0.42137 -0.16032  0.05747  0.17785  0.34174

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -476.30335   13.14537  -36.23   <2e-16 ***
t              0.24109    0.00656   36.75   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2087 on 21 degrees of freedom
Multiple R-squared:  0.9847,Adjusted R-squared:  0.984
F-statistic:  1351 on 1 and 21 DF,  p-value: < 2.2e-16
```

   (a) Which are the estimates of the α and β parameters, and what is the coefficient of determination here? (2p)

   Solution: From the R output, we can read off the estimates $\alpha^* \approx -476.3$, $\beta^* \approx 0.241$, and the coefficient of determination as `Multiple R-squared`, $R^2 \approx 0.98 = 98\%$.

   (b) Is there evidence that the wind energy production in Sweden changes over time? Motivate your answer. (1p)

   Solution: Test H_0: $\beta = 0$ vs H_1: $\beta \neq 0$. In the end of the `t` row, the p value of this test is given as $< 2.2 * 10^{-16}$. Hence, we can reject H_0 at all reasonable levels. There is strong evidence that the wind energy production in Sweden changes over time.

(c) In figures 3-4, a histogram and a qq plot of the model residuals are given. Figure 5 plots the residuals (on the y axis) vs the fitted values (on the x axis). Based on these plots, do you think that the model is a good description of the data, and in that case, why (or why not)? (2p)

Solution: The histogram does not look very normal distribution like, but this can have to do with the choice of bins. In the qq plot, for normality the points should lie close to a straight line, and this is about what we see. So it seems fair to conclude that the residuals are about normal.

The last plot should show no specific pattern and a constant vertical spread of points. However, there seems to be a pattern (up-down-up) that casts doubt on if the model is good enough to describe our data.

Appendix: figures

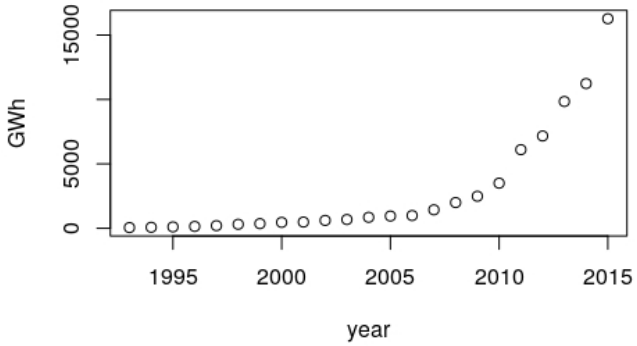

Figure 1: Swedish Wind energy production vs time.

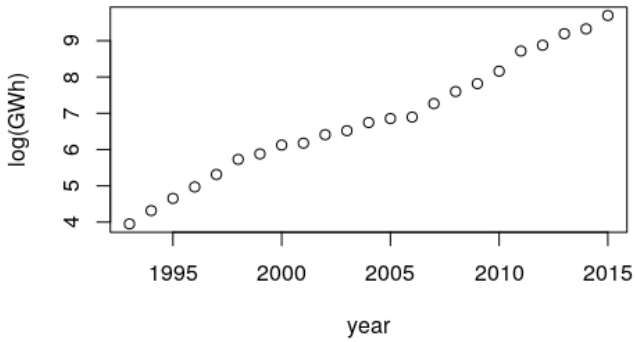

Figure 2: Log of Swedish Wind energy production vs time.

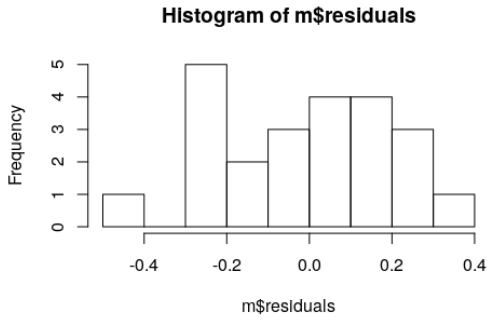

Figure 3: Histogram of residuals.

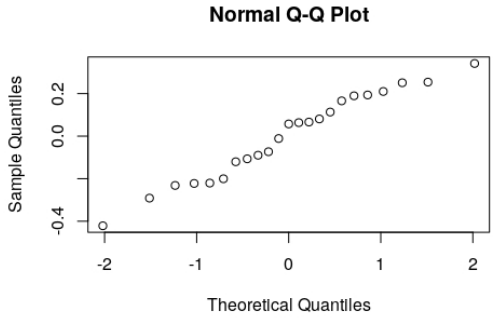

Figure 4: QQ plot of residuals.

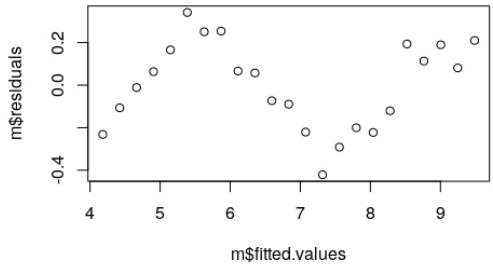

Figure 5: Residuals vs fitted values.