# Solutions for Simple Linear Regression Problems

S520

Arturo Valdivia

4/27/2023

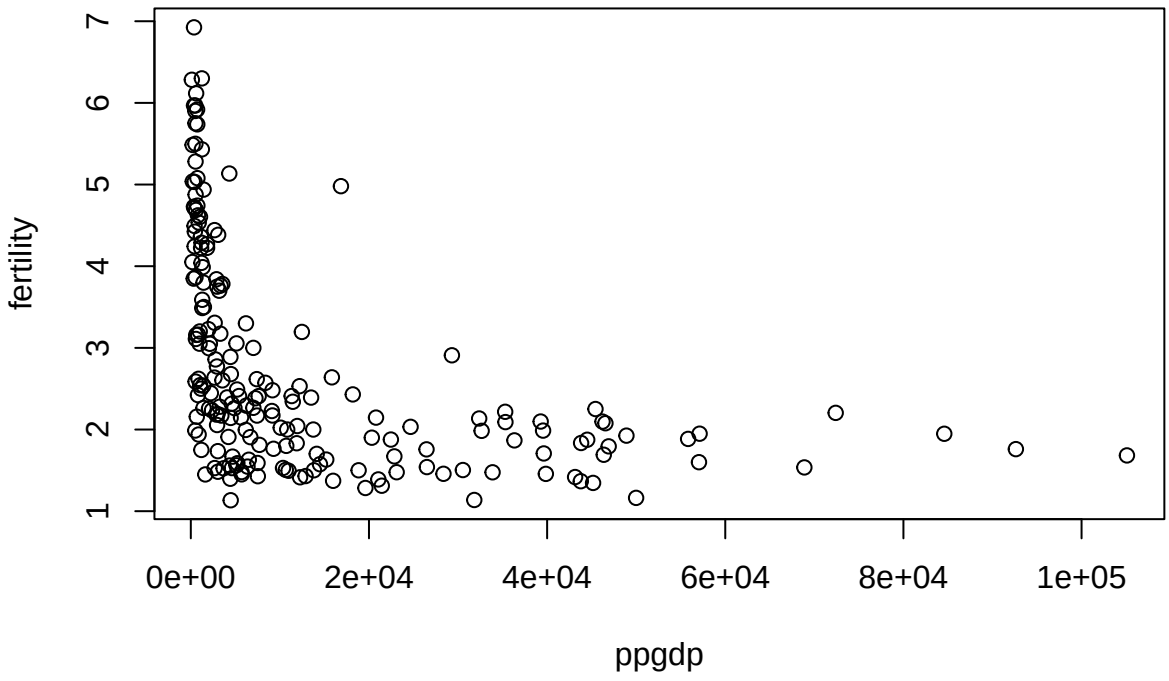
## 1

### 1a.

The predictor is ppgdp (Per capita gross domestic product in US dollars) and the response is fertility (number of children per woman).

### 1b.

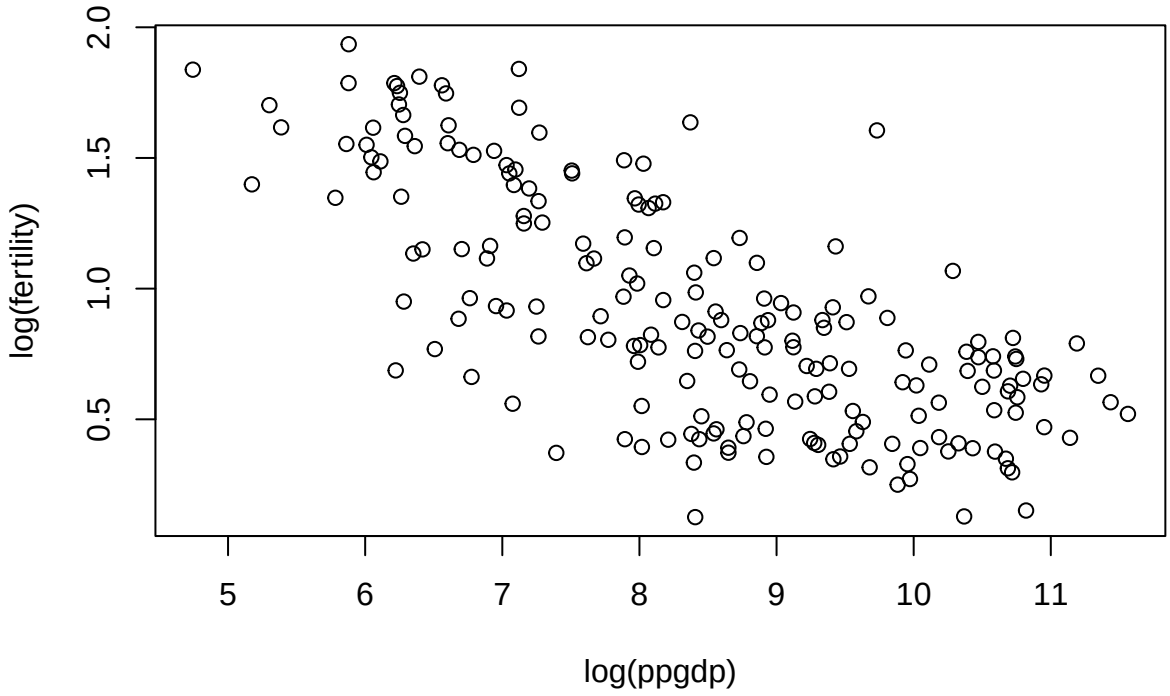
```
library(alr4)
plot(fertility ~ ppgdp, data = UN11)
```



The relationship shows some clear curvature. A straight line would do a poor job summarizing this relationship.

**1c.**

```
plot(log(fertility) ~ log(ppgdp), data = UN11)
```



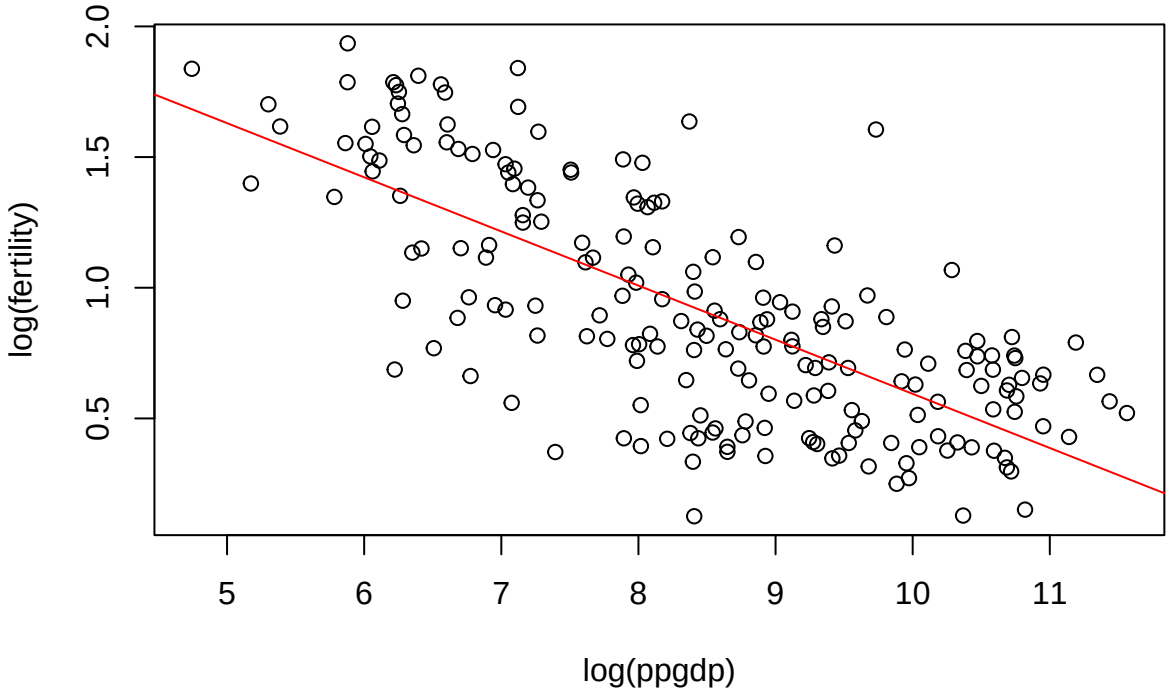

Yes, the relationship appears linear now. Using log transformations seem to be a better choice.

**1d.**

```
m1 = lm(log(fertility) ~ log(ppgdp), data = UN11)
```

**1e.**

```
plot(log(fertility) ~ log(ppgdp), data = UN11)
abline(m1, col="red")
```




**1f.**

Now $H_0 : \beta_1 \geq 0$ vs $H_1 : \beta_1 < 0$. We use the summary of `m1`, the model obtained in part d. Observe that the t statistic is the same as in the summary output (as $\beta_1 = 0$ under $H_0$), but the p-value is different as this is a left-tailed test

```
tt <- summary(m1)$coef[2,3]
df.m1 <- summary(m1)$df[2] # degrees of freedom from summary output
pt(tt, df.m1) # p-value obtained from the summary
```

```
## [1] 4.531178e-34
```

The p-value is close to zero. We reject the null hypothesis and conclude that the slope is negative. Observe that the data frame contain some missing values for the variables of interest and you need to take this into account when determining the degrees of freedom.

**1g.**

```
summary(m1)$r.squared
```

```
## [1] 0.525985
```

The model helps explain about 52.6% of the variation in `fertility` (when changes in `ppgdp` happen).

**1h.**

```
ci.log.fer = predict(m1, newdata = data.frame(ppgdp = 1000), interval = "prediction", level = .95)
exp(ci.log.fer)
```

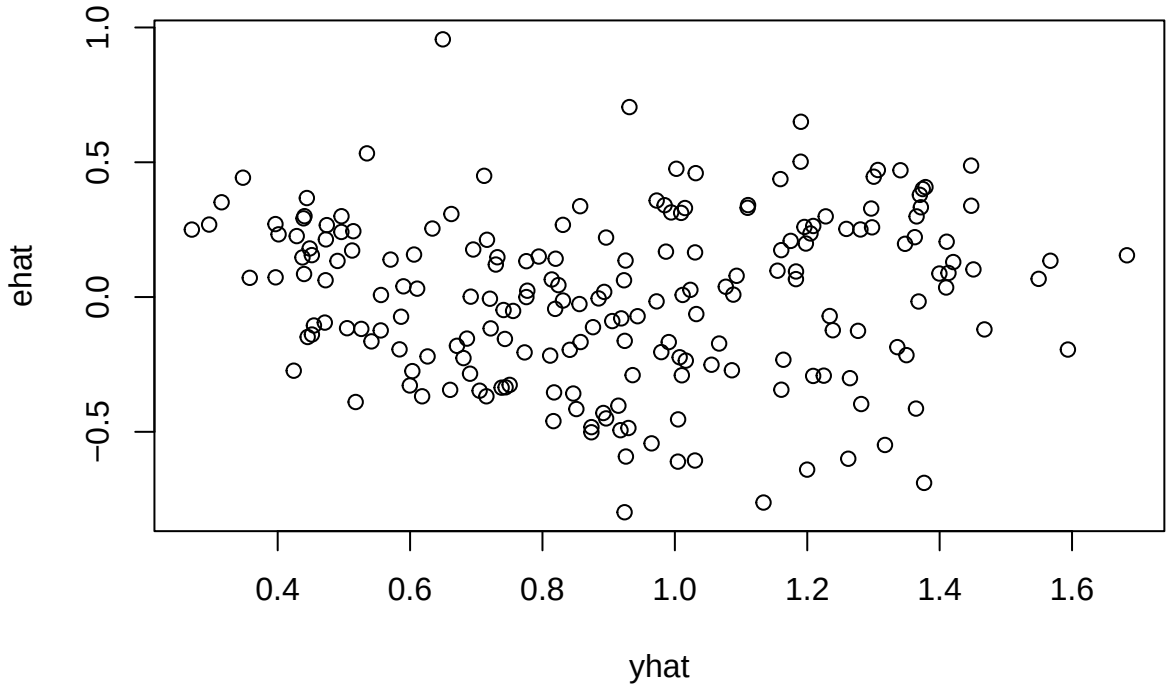
```
##        fit      lwr     upr
## 1 3.436891 1.869889 6.31707
```

If `ppgdp = 1000` We are 95% confident that fertility is a number between 1.87 and 6.32.

**1i.**

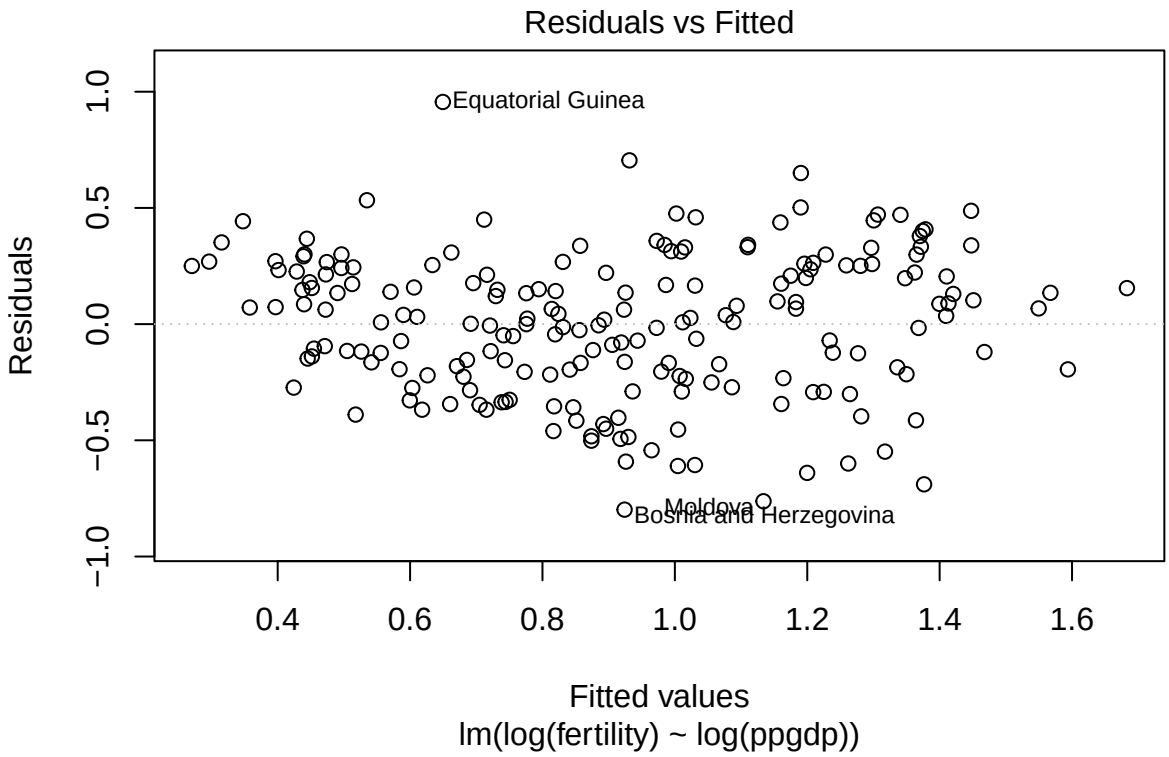
Let's construct this plot first manually:

```
x = log(UN11$ppgdp)
y = log(UN11$fertility)

b1 = sum((x - mean(x))*(y - mean(y)))/sum((x - mean(x))^2)
b0 = mean(y) - mean(x)*b1

# Fitted values and residuals
yhat = b0 + b1*x
ehat = y - yhat # residuals

## Residual against fitted values
plot(ehat ~ yhat)
```




Recall that you can also get this plot directly from `m1`:

```
# Or simply plot with m1
plot(m1,1, add.smooth = F)
```

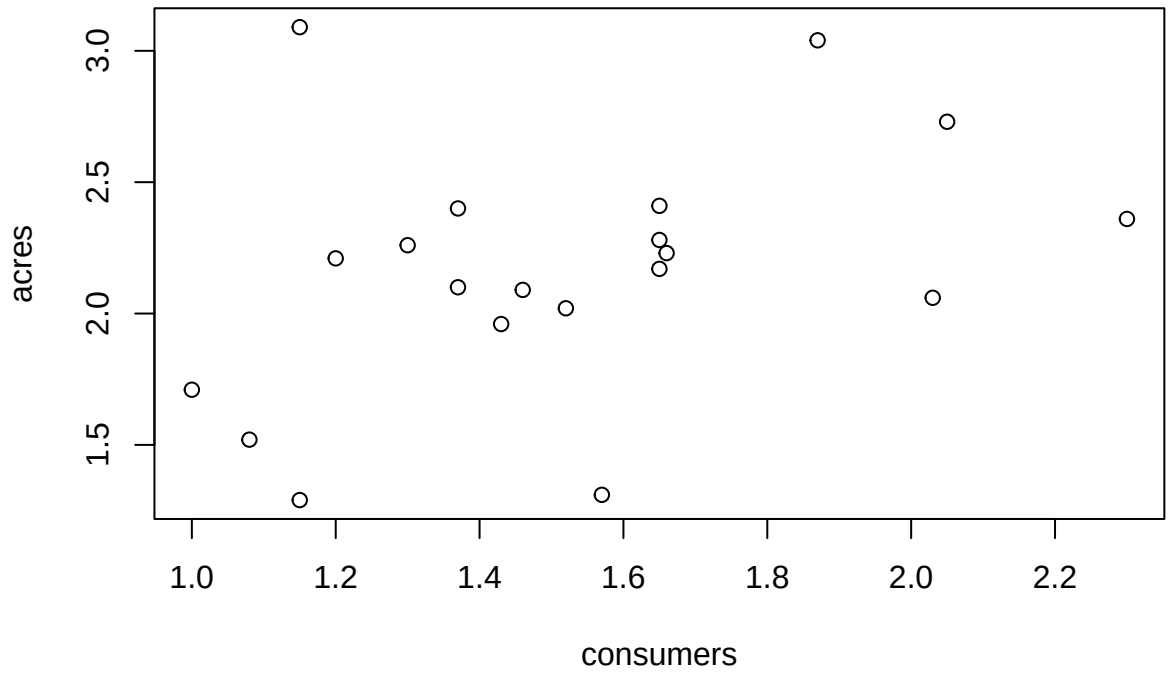



Residuals versus fitted values show a relationship close to a null plot, however plot seems to suggest a mild in change of dispersion (larger) for larger fitter values. There are also a couple of outliers, but overall seems that no obvious violations to the model assumptions are present. In addition, the QQ plot of residuals does not perfectly overlap with the normal curve, but deviations are not too concerning.

## 2

### 2a.

```
# We need to import the data into R first
sahlins <- read.table("Sahlins.txt", header=T)
# Now, we can create the plot
plot(acres ~ consumers, data=sahlins)
```




Although there are fairly few points and some of them may be outliers, there seems to be a weakly positive linear relationship between acres/gardener and consumers/gardener and a line could be an acceptable representation of this association.

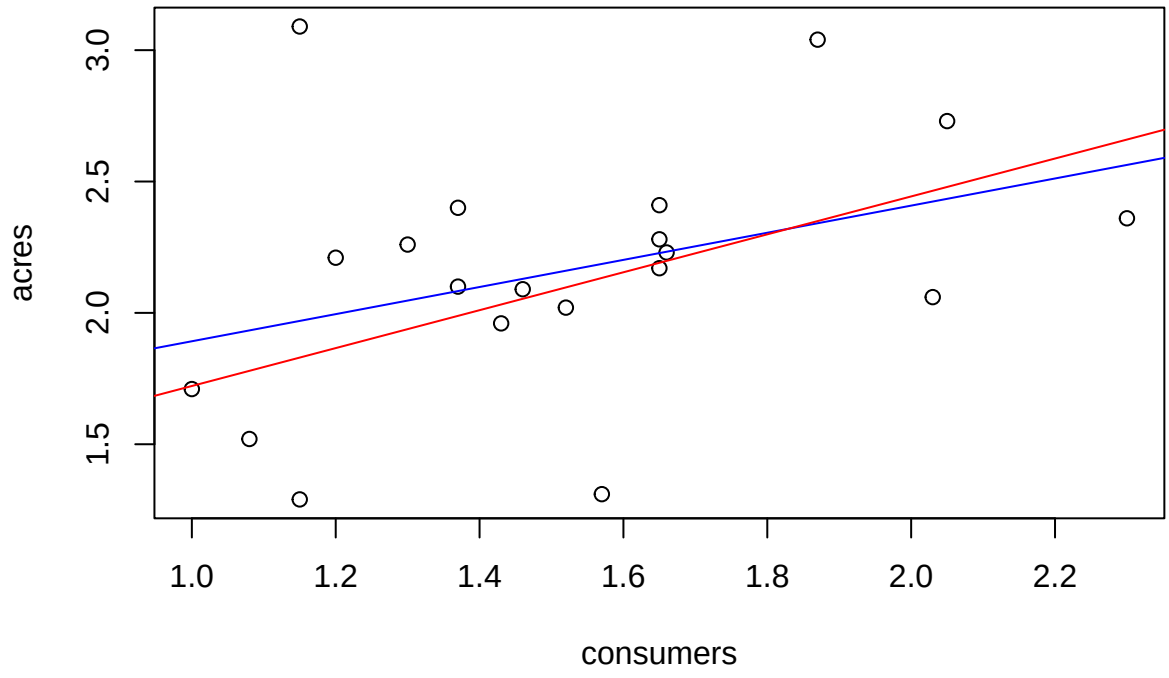
### 2b.

```
m2 <- lm(acres ~ consumers, sahlins)
m2.no4 = lm(acres ~ consumers, sahlins, subset = -4)
summary(m2)
```

```
##
## Call:
## lm(formula = acres ~ consumers, data = sahlins)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8763 -0.1873 -0.0211  0.2135  1.1206
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.3756     0.4684   2.937  0.00881 **
## consumers     0.5163     0.3002   1.720  0.10263
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4543 on 18 degrees of freedom
## Multiple R-squared:  0.1411, Adjusted R-squared:  0.0934
## F-statistic: 2.957 on 1 and 18 DF,  p-value: 0.1026
```

```
plot(acres ~ consumers, sahlins)
abline(m2, col="blue")
abline(m2.no4, col="red")
```




The intercept is 1.3756 acres/gardener. No interpretation for the intercept is given, because there is no meeting of having the regressor to be 0. In terms of the slope, for each additional consumer/gardender, the ratio for acres/gardener increases by 0.5163 units. In addition, the estimated standard deviation for the model is 0.4543. The estimated model can be written as

$$\hat{y} = 1.3756 + 0.5163x$$

The second regression (without the 4th observation) seems to do a better job. Unfortunately, we cannot

simply remove observation because is more convenient (in particular when you don't know if this was a typo or actually is capturing some valuable information that we should take into account.)

**2c.**

The standard errors are given in the summary output of the model:

```
# Standard errors of intercept and slope
c(summary(m2)$coef[1,2], summary(m2)$coef[2,2])
```

```
## [1] 0.4684047 0.3002335
```

Test for the intercept: $H_0 : \beta_0 = 0,\ H_1 : \beta_0 > 0$

```
# Test for the intercept
beta0hat = coef(m2)[1]
k = 0
se.beta0hat = summary(m2)$coef[1,2]
t.stat0 = (beta0hat - k)/se.beta0hat
t.stat0
```

```
## (Intercept)
##    2.936872
```

```
n = dim(sahlins)[1] #number of rows in the data set
1 - pt(abs(t.stat0), n - 2)
```

```
## (Intercept)
## 0.004406897
```

For a significance level $\alpha = 0.01$, we reject the null and conclude slope is positive.

Test for the slope: $H_0 : \beta_1 = 0,\ H_1 : \beta_1 > 0$

```
# Test for the slope
beta1hat = coef(m2)[2]
se.beta1hat = summary(m2)$coef[2,2]
t.stat1 = (beta1hat - k)/se.beta1hat
t.stat1
```

```
## consumers
##  1.719728
```

```
1 - pt(abs(t.stat1), n - 2)
```

```
##  consumers
## 0.05131463
```

For a small $\alpha = 0.01$, we clearly fail to reject the null. There is no evidence that this slope is different han zero.

Now, using the data without the fourth household and performing: $H_0 : \beta_0 = 0,\ H_1 : \beta_0 > 0$

```r
# Test for the intercept
beta0hat.no4 = coef(m2.no4)[1]
se.beta0hat.no4 = summary(m2.no4)$coef[1,2]
t.stat0.no4 = (beta0hat.no4 - k)/se.beta0hat.no4
t.stat0.no4
```

```
## (Intercept)
##    2.519375
```

```r
1 - pt(abs(t.stat0.no4), n - 3)
```

```
## (Intercept)
##  0.01102734
```

We now test for the slope without the fourth household: $H_0 : \beta_1 = 0$, $H_1 : \beta_1 \neq 0$

```r
# Test for the slope
beta1hat.no4 = coef(m2.no4)[2]
se.beta1hat.no4 = summary(m2.no4)$coef[2,2]
t.stat1.no4 = (beta1hat.no4 - k)/se.beta1hat.no4
t.stat1.no4
```

```
## consumers
##  2.870143
```

```r
1 - pt(abs(t.stat1.no4), n - 3)
```

```
##   consumers
## 0.005306328
```

We can not obtain the opposite conclusion, there seem to be evidence to reject the null in conclude that the slope is greater than zero. The 95% confidence intervals are given below:

```r
confint(m2, level=.95)
```

```
##                  2.5 %   97.5 %
## (Intercept)  0.3915628 2.359726
## consumers   -0.1144471 1.147087
```

```r
confint(m2.no4, level=.95)
```

```
##                 2.5 %   97.5 %
## (Intercept) 0.1625647 1.837443
## consumers   0.1911570 1.252031
```

When the entire data was employed, the test for the intercept reject the null hypothesis with p-value 0.0088, which means there is strong evidence that the intercept is not 0. However, we fail to reject the null hypothesis in testing the slope. There is not enough evidence against $\beta_1 = 0$ with the whole data. We can identify that the confidence interval of $\beta_1$ includes 0 which matches the test result.

While we have the same test result to reject the null hypothesis as all the data was used for the intercept, the test for the slope without the fourth household rejects the null hypothesis. There is enough evidence that this is not a primitive communist society.

**2d.**

```
exp(predict(m2, newdata = data.frame(consumers = 1.5),
            interval = "predict",
            level = .98))
```

```
##        fit      lwr      upr
## 1 8.585928 2.616313 28.17636
```

```
exp(predict(m2, newdata = data.frame(consumers = 1.5),
            interval = "confidence",
            level = .98))
```

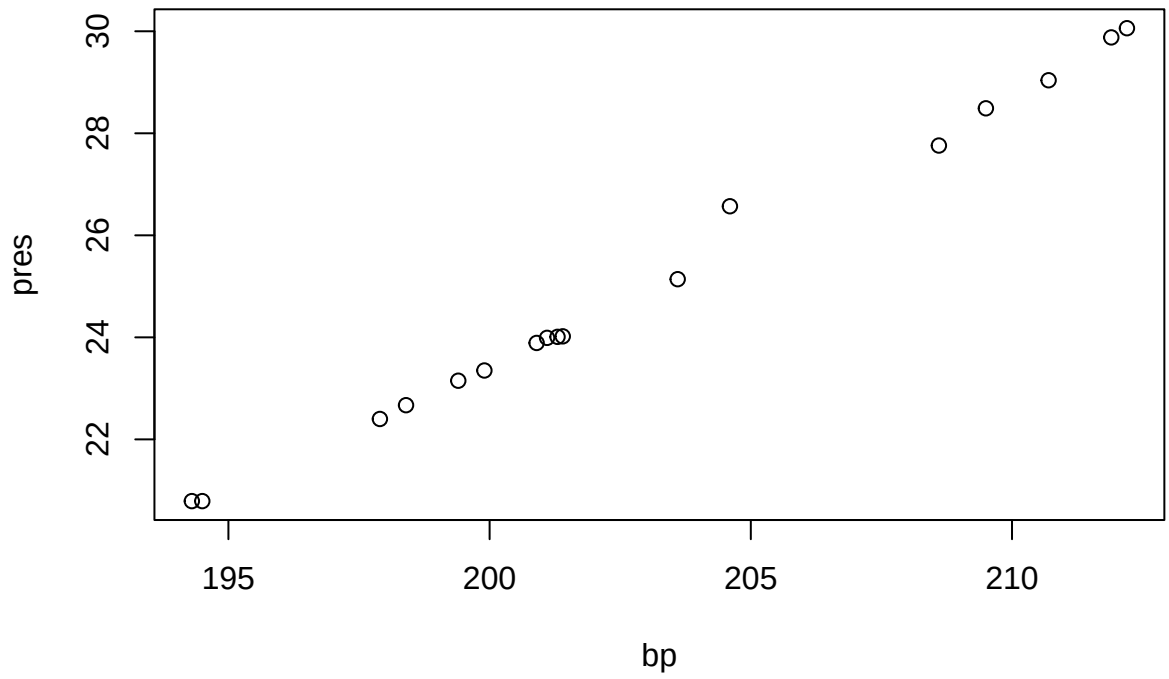
```
##        fit      lwr      upr
## 1 8.585928 6.620916 11.13413
```

Yes, the two questions correspond to two different problems and the answers will changed. A prediction interval for a new response when the regressor is consumers/gardener = 1.5 needs to accounts for two sources of variation, while the interval for the expected response only needs to account for one source of variation (check class notes for details). This leads to a wider prediction interval.

## 3

### 3a.

```
library(alr4)
plot(pres ~ bp, Forbes)
```




It appears to be a strong positive linear relationship between adjusted boiling point of water and atmospheric pressure. Observe, however, that there is one point that doesn't follow this relationship as well as the rest.

### 3b.

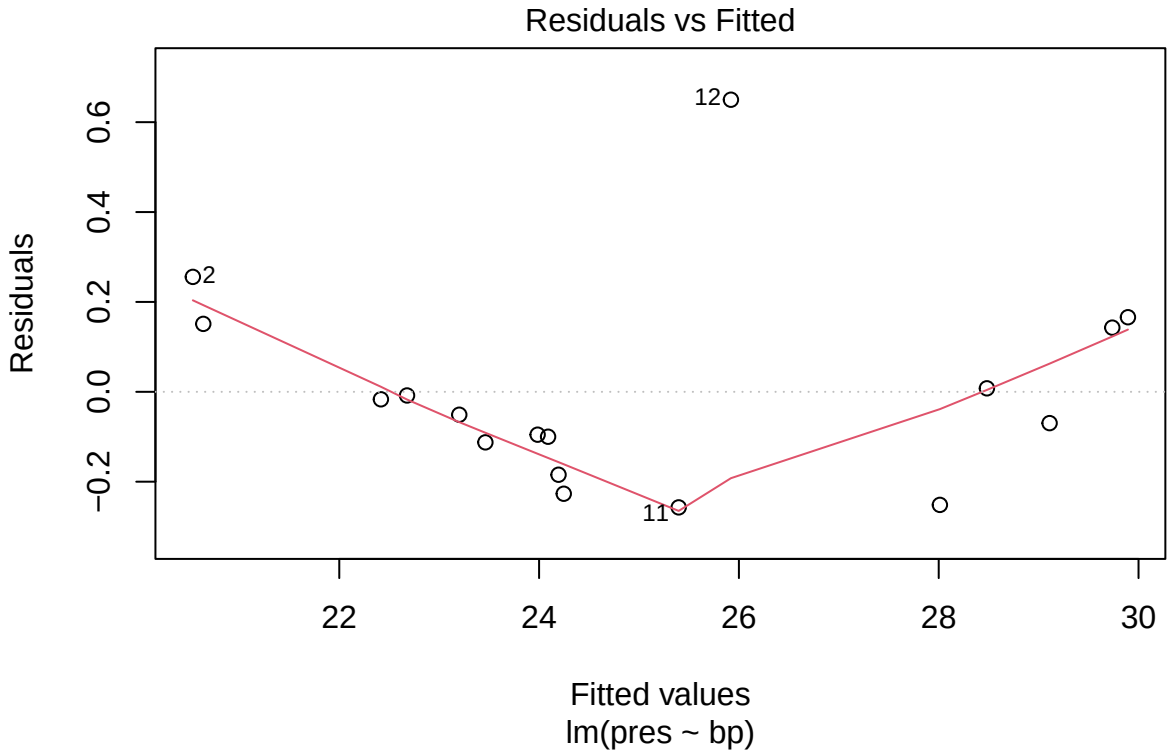
```
m3 <- lm(pres ~ bp, Forbes)
m3
```

```
##
## Call:
## lm(formula = pres ~ bp, data = Forbes)
##
## Coefficients:
## (Intercept)           bp
##    -81.0637       0.5229
```

The range of values for `bp` doesn't include 0; therefore, the intercept doesn't have a valid interpretation. Every one extra degree(F) of boiling point of water leads to a 0.5229 increase in atmospheric pressure in inches of Mercury, on average.

**3c.**

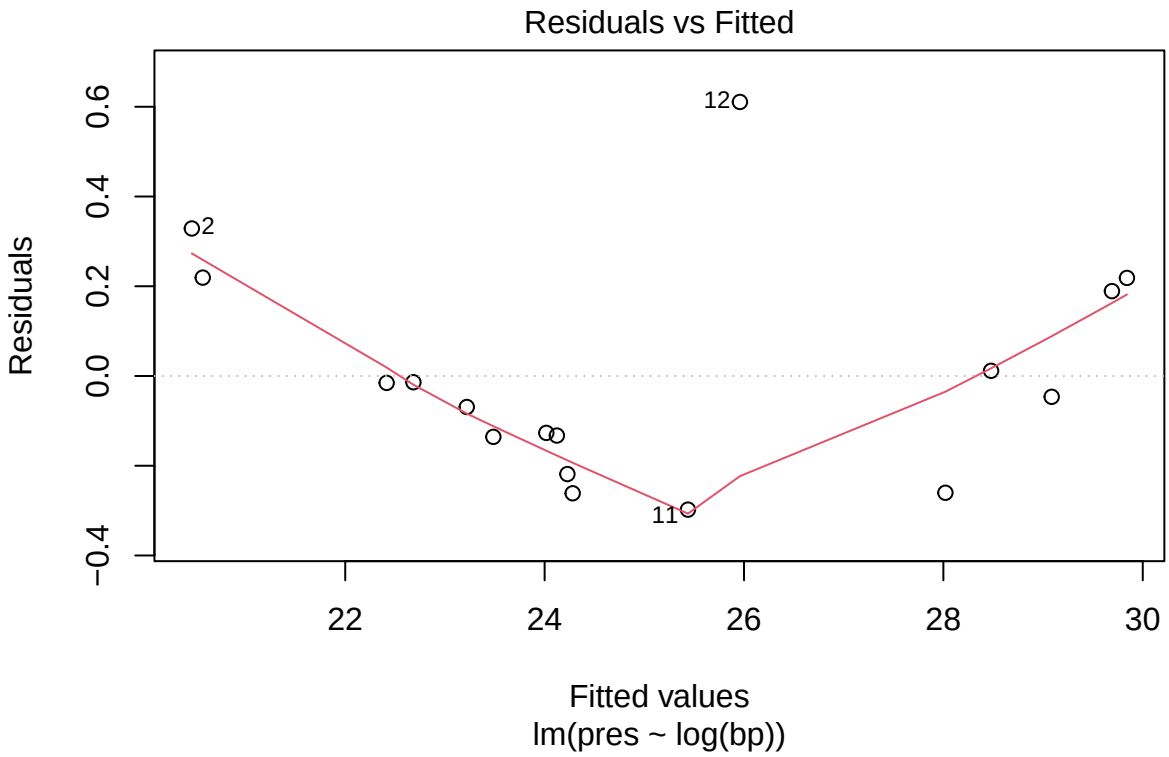
```
plot(m3,1)
```



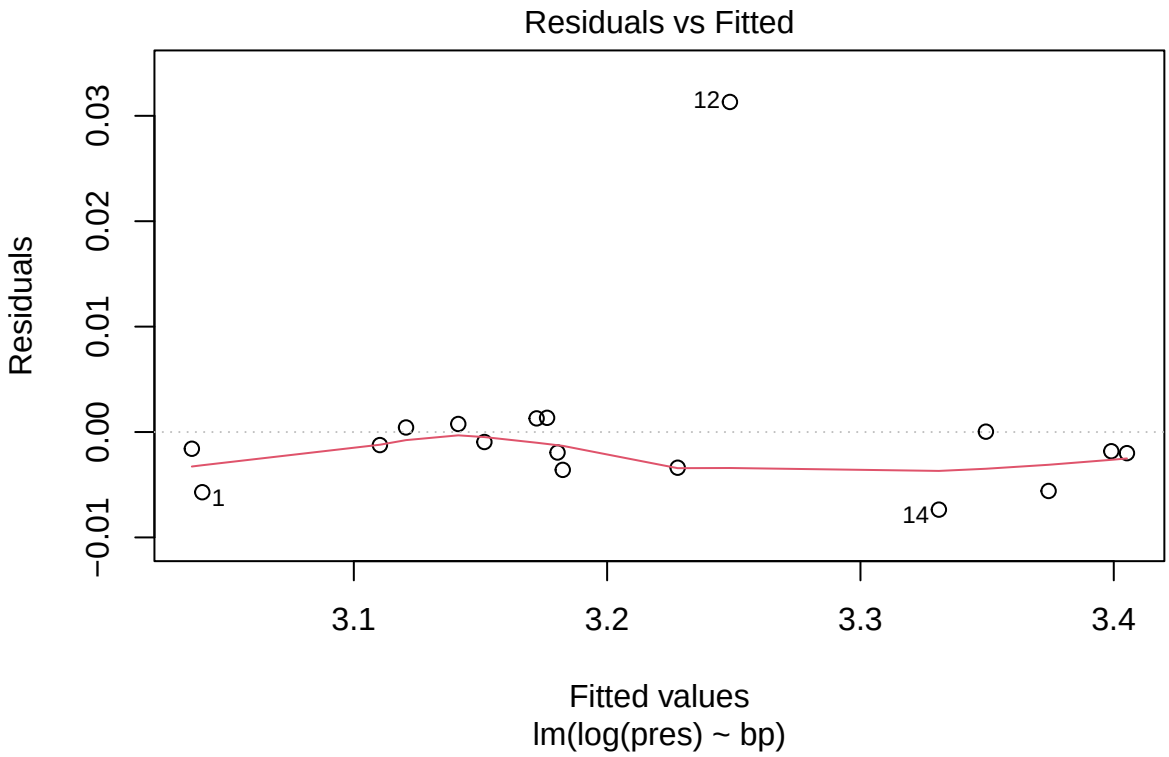

The plot residuals against fitted values is not a null plot. Residuals go down and then go up. In addition, observation 12 doesn't follow the general trend at all.

**3d.**

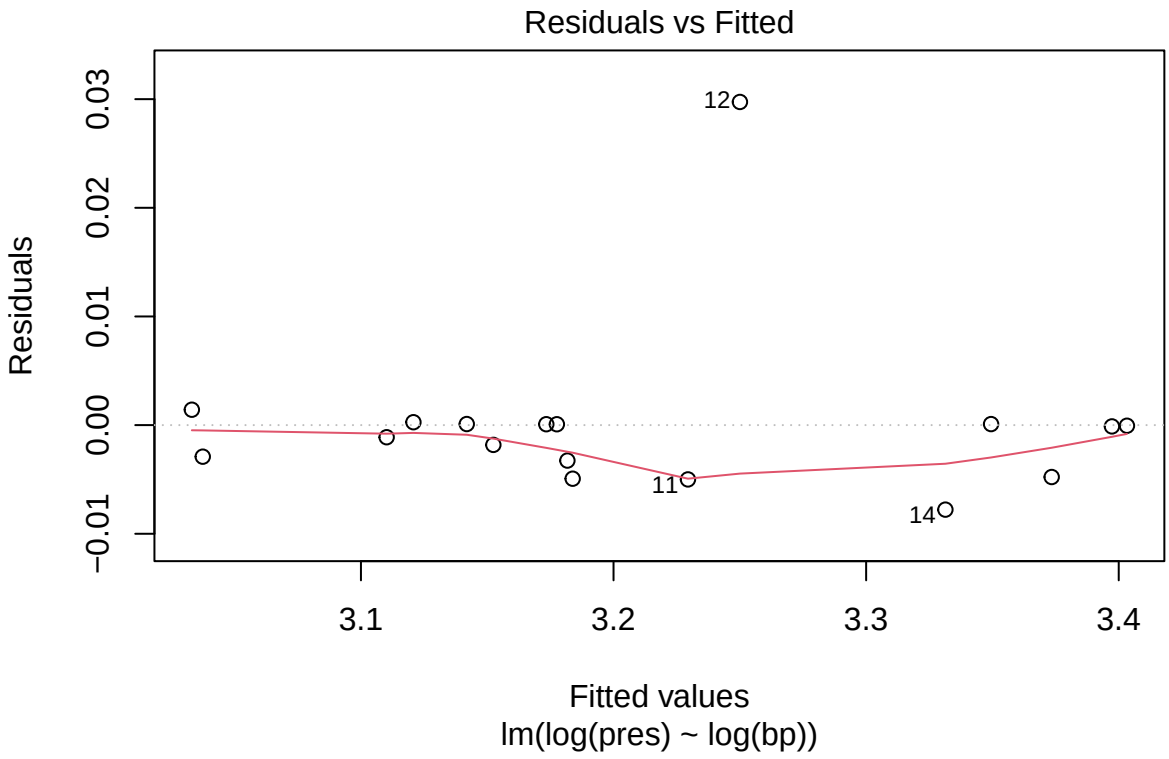
```
m3.1 <- lm(pres ~ log(bp), Forbes)
m3.2 <- lm(log(pres) ~ bp, Forbes)
m3.3 <- lm(log(pres) ~ log(bp), Forbes)
plot(m3.1,1)
```


Residuals vs Fitted
Residuals
0.6
0.4
0.2
0.0
−0.4
12
2
11
22
24
26
28
30
Fitted values
lm(pres ~ log(bp))

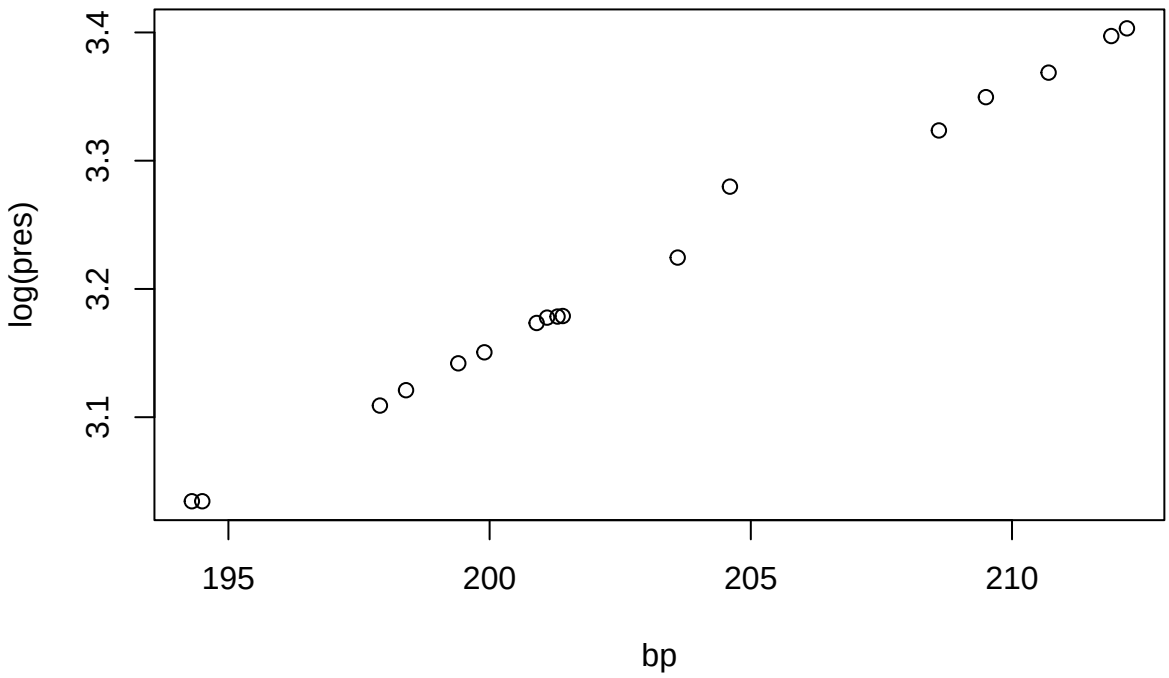
```
plot(m3.2,1)
```


Residuals vs Fitted
Residuals
−0.01
0.00
0.01
0.02
0.03
12
1
14
3.1
3.2
3.3
3.4
Fitted values
lm(log(pres) ~ bp)

```
plot(m3.3,1)
```




```
plot(log(pres) ~ bp, Forbes)
```




Observation 12 prevents to clearly compare the plots. That said, the model that uses the log transformation of the response only, seems to be the most appropriate representation (the residual against fitted values plot looks closer to a null plot if we do not take into account observation 12).

**3e.**

```
confint(m3.2, level=.97)
```

```
##                    1.5 %     98.5 %
## (Intercept) -1.15528615 -0.7864463
## bp           0.01971402  0.0215307
```

We are 97% confident that, if the boiling point of water increases by 1F, the log atmospheric pressure increases some value in the interval (0.0197, 0.0215).

**3f.**

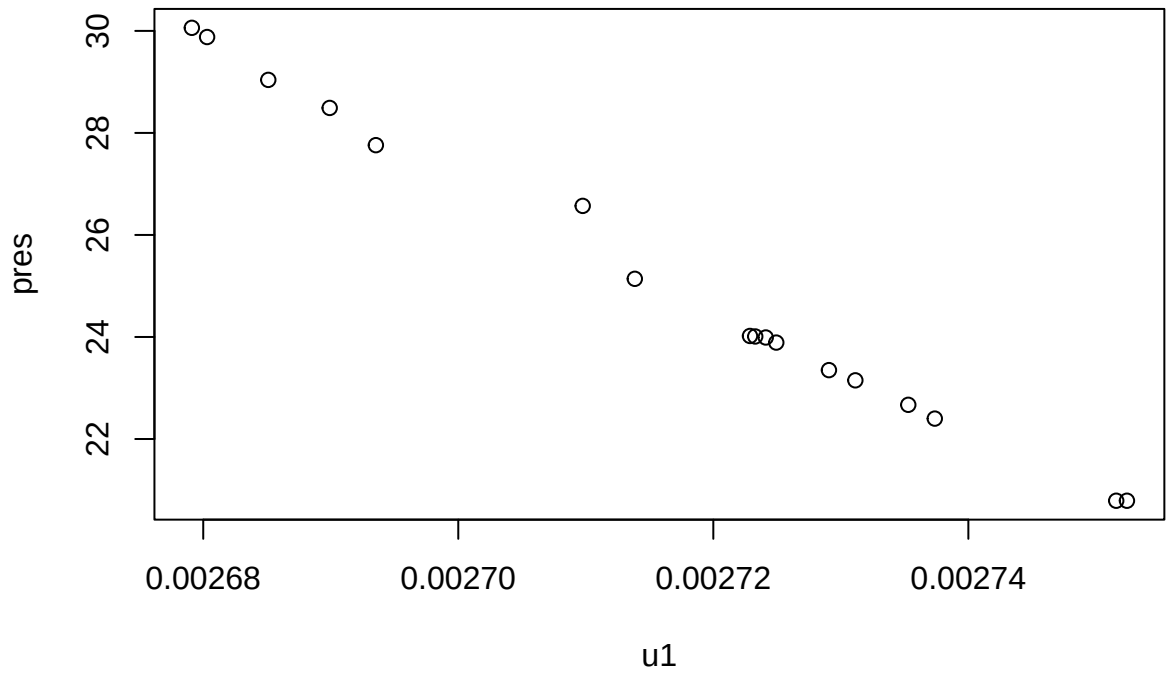
```
exp(predict(m3.2, newdata = data.frame(bp = 200),
            interval = "predict",
            level = .94))
```

```
##        fit      lwr      upr
## 1 23.42037 22.99299 23.85568
```

## 4

### 4a.

```
Forbes$u1 <- 1/((5/9)*Forbes$bp + 255.37)
plot(pres ~ u1, Forbes)
```




Because we take the inverse of bp, the updated plot with `u1` displays a negative slope.

### 4b.

```
m4 <- lm(pres ~ u1, Forbes)
summary(m4)
```

```
##
## Call:
## lm(formula = pres ~ u1, data = Forbes)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.28216 -0.12643 -0.05569  0.17111  0.62569
##
## Coefficients:
```

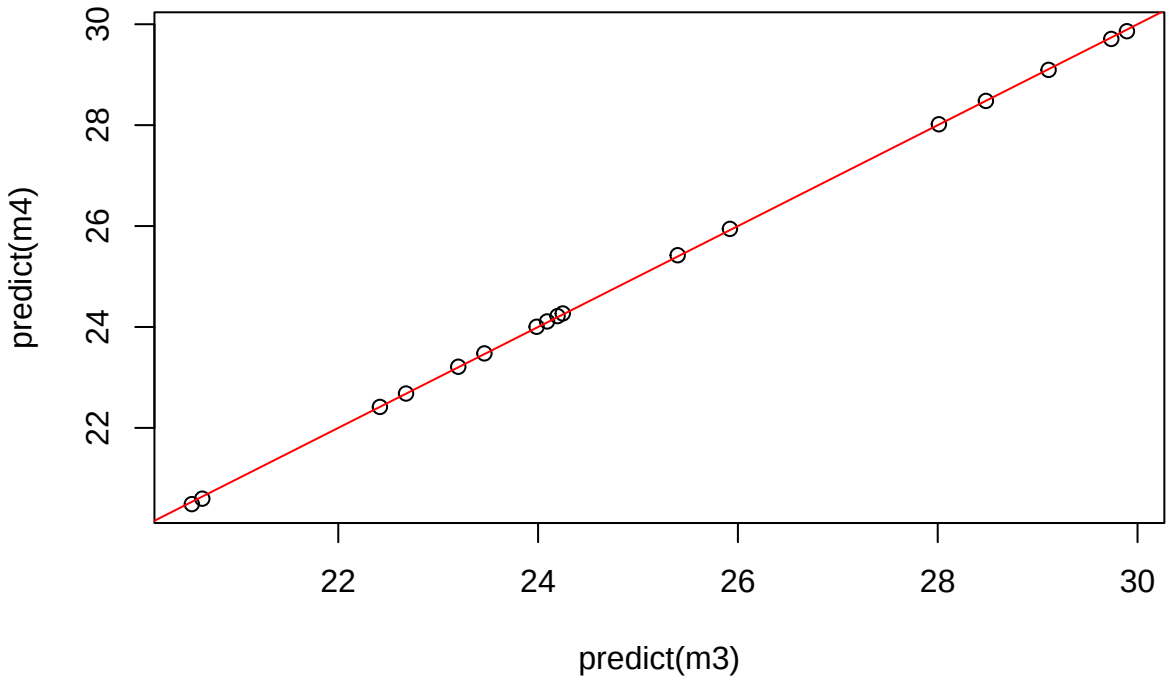
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.723e+02  7.013e+00   53.08   <2e-16 ***
## u1          -1.278e+05  2.581e+03  -49.51   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2433 on 15 degrees of freedom
## Multiple R-squared:  0.9939, Adjusted R-squared:  0.9935
## F-statistic:  2451 on 1 and 15 DF,  p-value: < 2.2e-16
```

There is strong evidence that neither the intercept nor the slope are 0. For plots, there is still a pattern in the plot of residuals vs. fitted values and there are a couple of abnormal points.

**4c.**

Let's obtain the desired plot. We can use the function `predict()` for this purpose on each regression line:

```
plot(predict(m4) ~ predict(m3))
abline(a = 0, b = 1, col="red")
```




Fitted values by m3 (obtained from 3b) and m4 (obtained from 4b) look almost the same as they are on x=y. Both regression are almost equivalent and it would be difficult to determine if one is better than the other.

**4d.**

The structure of this problem is quite similar to a 2-sample problem we encountered earlier (ISI, Chapter 11): we are not comparing two means, $\mu_1 - \mu_2$, rather two slopes, $\beta_{1(Forbes)} - \beta_{1(Hooker)}$, but we can proceed as before. We use a t-test to compare the two slopes. $H_0 : \beta_{1(Forbes)} - \beta_{1(Hooker)} = 0$ versuse $H_1 : \beta_{1(Forbes)} - \beta_{1(Hooker)} \neq 0$. In addition, we use properties of the variance (the variance of the difference is the sum of the variances) and for the degrees of freedom we use $n_1 + n_2 - 4$. The degrees of freedom can be found in other ways, but the results would be fairly similar assuming the samples are not too small. Here is the code:

```
m4.h <- lm(pres ~ bp, Hooker)
summ.m3 <- summary(m3)
summ.m4.h <- summary(m4.h)

s1 = summ.m3$coefficients[2,2]
s2 = summ.m4.h$coefficients[2,2]
se = sqrt(s1^2 + s2^2)
n1 <- dim(Forbes)[1]
n2 <- dim(Hooker)[1]
ts = (coef(m3)[2]-coef(m4.h)[2] - 0)/se
ts # test statistic
```

```
##       bp
## 6.581518
```

```
df = n1+n2-4
2*(1 - pt(ts, df)) # p-value
```

```
##           bp
## 4.706412e-08
```

The test rejects the null hypothesis so we can conclude there is enough evidence that the two slopes are different.