S631 HW4

Erik Parker

September 20, 2017

1. ALR 2.16 United Nations Data

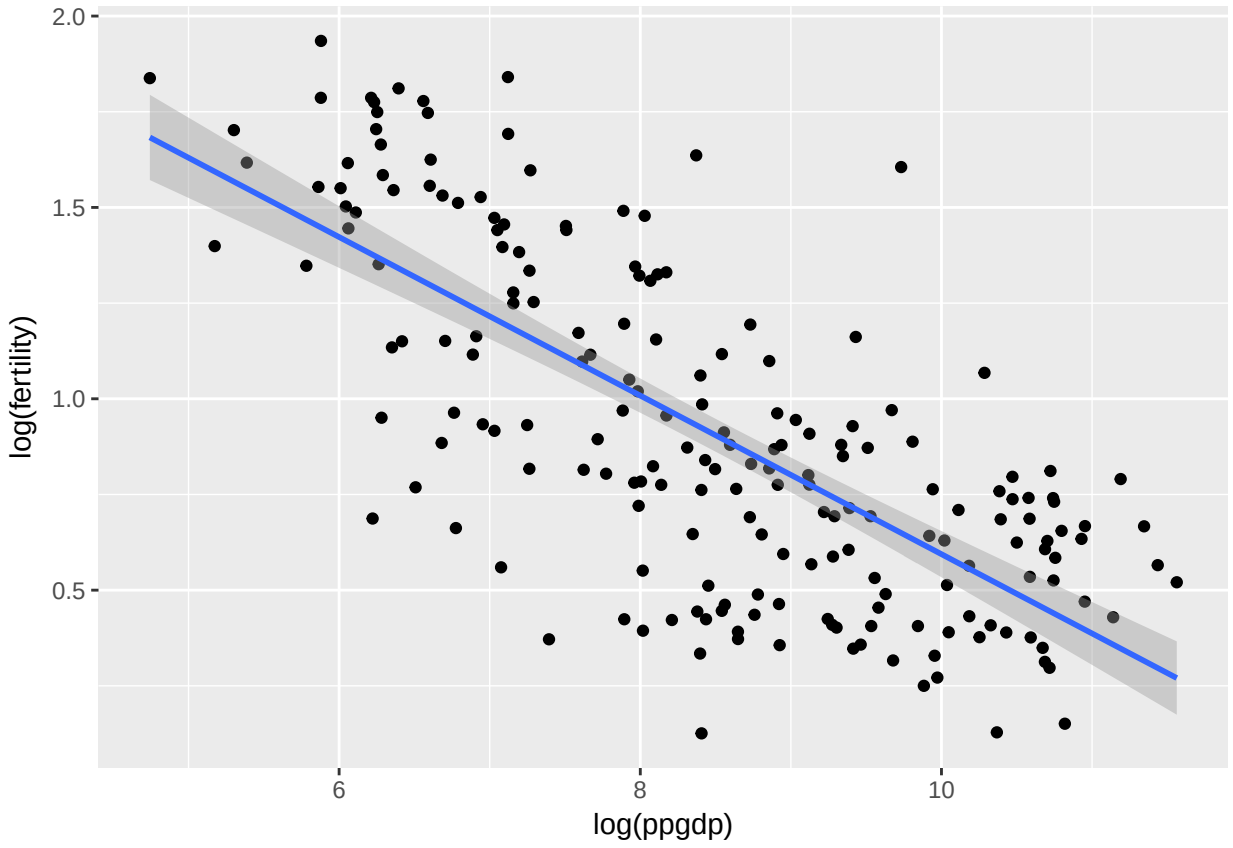
a. Compute the simple linear regression model corresponding to the graph in problem 1.1.3 (log(fertility) vs log(ppgdp))

```
library(alr4)
library(ggplot2)

UN <- UN11

lm1 <- lm(formula = log(fertility) ~ log(ppgdp), data = UN)

lm1
```

```
##
## Call:
## lm(formula = log(fertility) ~ log(ppgdp), data = UN)
##
## Coefficients:
## (Intercept)   log(ppgdp)
##      2.6655      -0.2071
```

b. Draw a graph of log(fertility) versus log(ppgdp) and add the fitted line to it

```
ggplot(UN, aes(log(ppgdp), log(fertility))) + geom_point() + geom_smooth(method = "lm",
    formula = y ~ x)
```



c. Test the hypothesis that the slope is 0 vs the alternative that it is negative (a one-sided test). Give the significance level of the result and a sentence that summarizes the result.

```
# t = B1hat - B1 / Se(B1hat|X)

B1hat <- coef(summary(lm1))[2, 1]

B1 <- 0

SeB1hat <- coef(summary(lm1))[2, 2]

t <- (B1hat - B1)/SeB1hat


twosidep <- coef(summary(lm1))[2, 4]

onesidep <- twosidep/2

onesidep
```

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

With a p-value of 4.53e-34 from our one sided test, we can say that there is an extremely low probability of finding the slope we did ($\hat{\beta}_1$ = -0.207) given that the real slope (β_1) was zero.

d. Give the value of the coeffiecient of determination and explain its meaning.

```r
summary(lm1)$r.squared
```

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

The coefficient of determination, denoted in the lm() ourput as R-squared, is a measure of goodness of fit of a model to the data. Specifically, it is the proportion of the variance in the dependent variable (here, fertility) that is explained by the independent variable (here, ppgdp).

e. Obtain a point prediction and 95% prediction interval for log(fertility) for a locality not in the data with ppgdp = 1000.

To obtain a prediction using our existing values for $\hat{\beta}_1$ and $\hat{\beta}_0$, we can use the following formula:

$$\tilde{y}_* = \hat{\beta}_0 + \hat{\beta}_1 x_*$$

where x_* is the observed value of our predictor (here a locality with ppgdp = 1000), and $\tilde{y}_*$ is the prediction for our unobserved value of y.

```r
B0hat <- coef(summary(lm1))[1, 1]

ytilde <- B0hat + B1hat * log(1000)

ytilde
```

```
## [1] 1.234567
```

So, a locality with a (log transformed) ppgdp of 1000 is predicted to have a (log transformed) fertility of 1.235. To find the 95% prediction interval for log(fertility), we can use the predict() command as shown below.

```r
newdata <- data.frame(ppgdp = 1000)

predicted <- predict(lm1, newdata, interval = "predict")
predicted
```

```
##        fit       lwr      upr
## 1 1.234567 0.6258791 1.843256
```

```r
lowernolog <- exp(predicted[1, 2])
uppernolog <- exp(predicted[1, 3])

lowernolog
```

```
## [1] 1.869889
```

```r
uppernolog
```

```
## [1] 6.31707
```

From this, we see both that our point prediction calculated above was correct, and also that the 95% prediction interval of log(fertility) given a locality with ppgdp = 1000 is between 0.626 and 1.843. To transform this to a 95% prediction interval for fertility on the original scale, we can exponentiate both the lower and upper bounds of our confidence interval which gives us a 95% confidence interval for fertility between 1.87 and 6.32.

f. Identify the following:

i and ii. The locality with the highest, and lowest values of fertility

```
head(UN[order(-UN$fertility), ])
```

```
##             region  group fertility  ppgdp lifeExpF pctUrban
## Niger       Africa africa     6.925  357.7    55.77       17
## Zambia      Africa africa     6.300 1237.8    50.04       36
## Somalia     Africa africa     6.283  114.8    53.38       38
## Mali        Africa africa     6.117  598.8    53.14       37
## Afghanistan   Asia  other     5.968  499.0    49.49       23
## Malawi      Africa africa     5.968  357.4    55.17       20
```

```
head(UN[order(UN$fertility), ])
```

```
##                         region group fertility   ppgdp lifeExpF pctUrban
## Bosnia and Herzegovina Europe other     1.134  4477.7    78.40       49
## Hong Kong                Asia other     1.137 31823.7    86.35      100
## Macao                    Asia other     1.163 49990.2    83.80      100
## Malta                  Europe other     1.284 19599.2    82.29       95
## Portugal               Europe  oecd     1.312 21437.6    82.76       61
## Austria                Europe  oecd     1.346 45158.8    83.55       68
```

The locality with the highest fertility is Niger, and the locality with the lowest is Bosnia and Herzegovina.

iii. The two localities with the largest positive residuals from the regression with both variables in log scale, and the two with the largest negative residuals in log scale.

```
residuals <- lm1$residuals
residuals <- as.data.frame(residuals)

residuals <- residuals[order(-residuals), , drop = FALSE]

head(residuals)
```

```
##                   residuals
## Equatorial Guinea 0.9559557
## Angola            0.7047167
## Zambia            0.6501748
## Israel            0.5329906
## Nigeria           0.5020828
## Niger             0.4876082
```

```
tail(residuals)
```

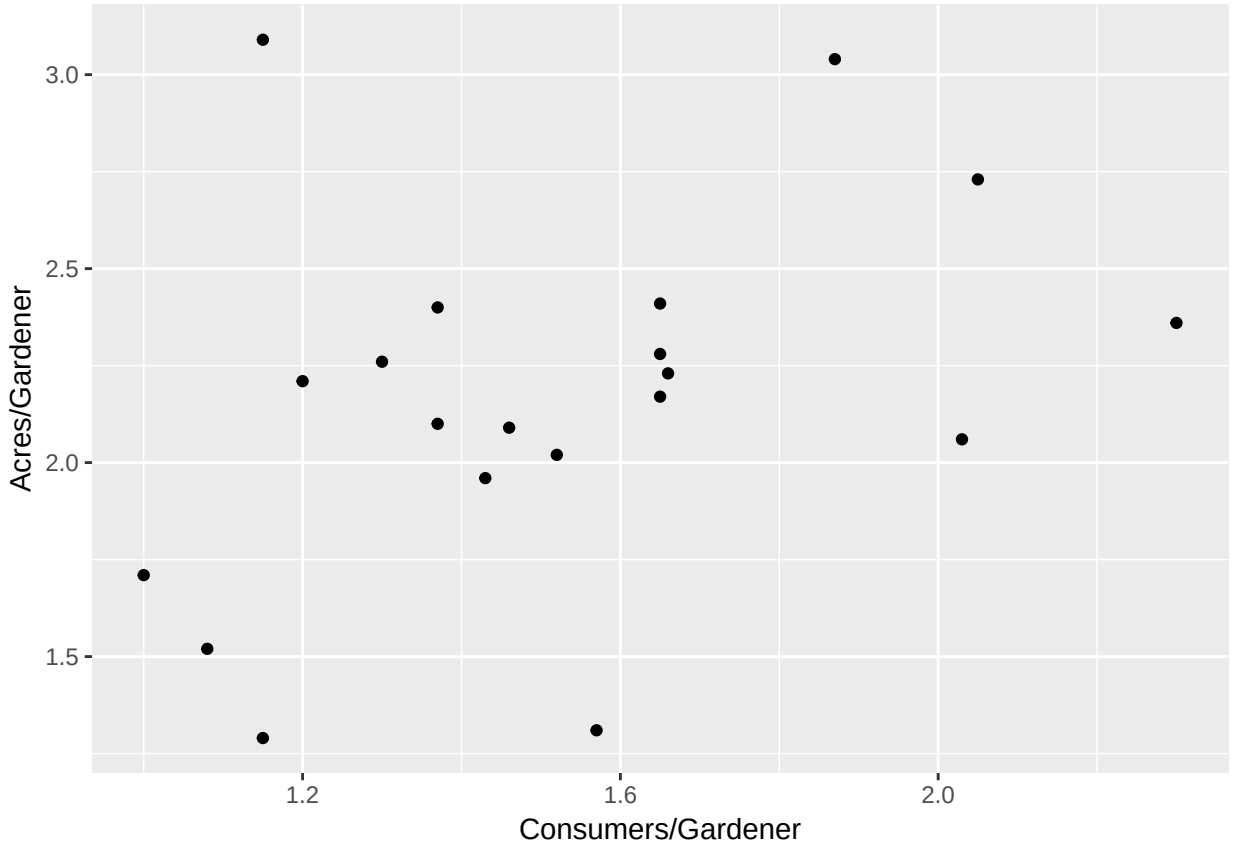
```
##                          residuals
## Georgia                -0.6063727
## Ukraine                -0.6105202
## Viet Nam               -0.6401918
## North Korea            -0.6893729
## Moldova                -0.7623290
## Bosnia and Herzegovina -0.7982759
```

So, from this we can see that the two localities with the largest positive residuals are Equatorial Guinea and Angola, while the two with the largest negative residuals are Bosnia and Herzegovina and Moldova.

2. Using the data Sahlins.txt, which describes agricultura production from a village in Central Africa, examine the relationship between the explanatory variable Consumer/Gardener (the ratio of consumers to productive individuals per household) and the response variable Acres/Gardener (amount of land cultivated per household).

a. Draw a scatterplot of Acres/Gardener vs. Consumers/Gardener. What relationship can be seen? Anything else noteworthy from the plot?

```
Sahlins <- read.table("Sahlins.txt", header = TRUE)

ggplot(Sahlins, aes(x = consumers, y = acres), ) + geom_point() + labs(x = "Consumers/Gardener",
    y = "Acres/Gardener")
```



Naively, based on the identities of the two variables, I would predict that we should see a positive relationship between Consumers and Acres, because if a household is consuming more resources on average, they should also be producing more so they can support themselves. Though, I could also imagine that the inverse would be true (a household with more consumers might contain more childern/elderly individuals who need to be cared for, so there are less hands free to produce food), or that there would be no relationship (The communtiy shares food based on need, so there is no real incentive for one household to produce much more than another).

After plotting the data though,it seems to me that there is a slightly positive relationship between these two variables. Houses with more consumers do seem to be cultivating more land on average than houses with fewer consumers. That said there are a number of outliers and strong leverage points which could be swaying my interpretation of this plot. For example, there is a cluster of three points at the lower left end of the plot which suggest that low consumer households produce less, along with a cluster of points in the top right which suggests that higher consumer household produce more. As previously stated though, there are a number of individual outliers

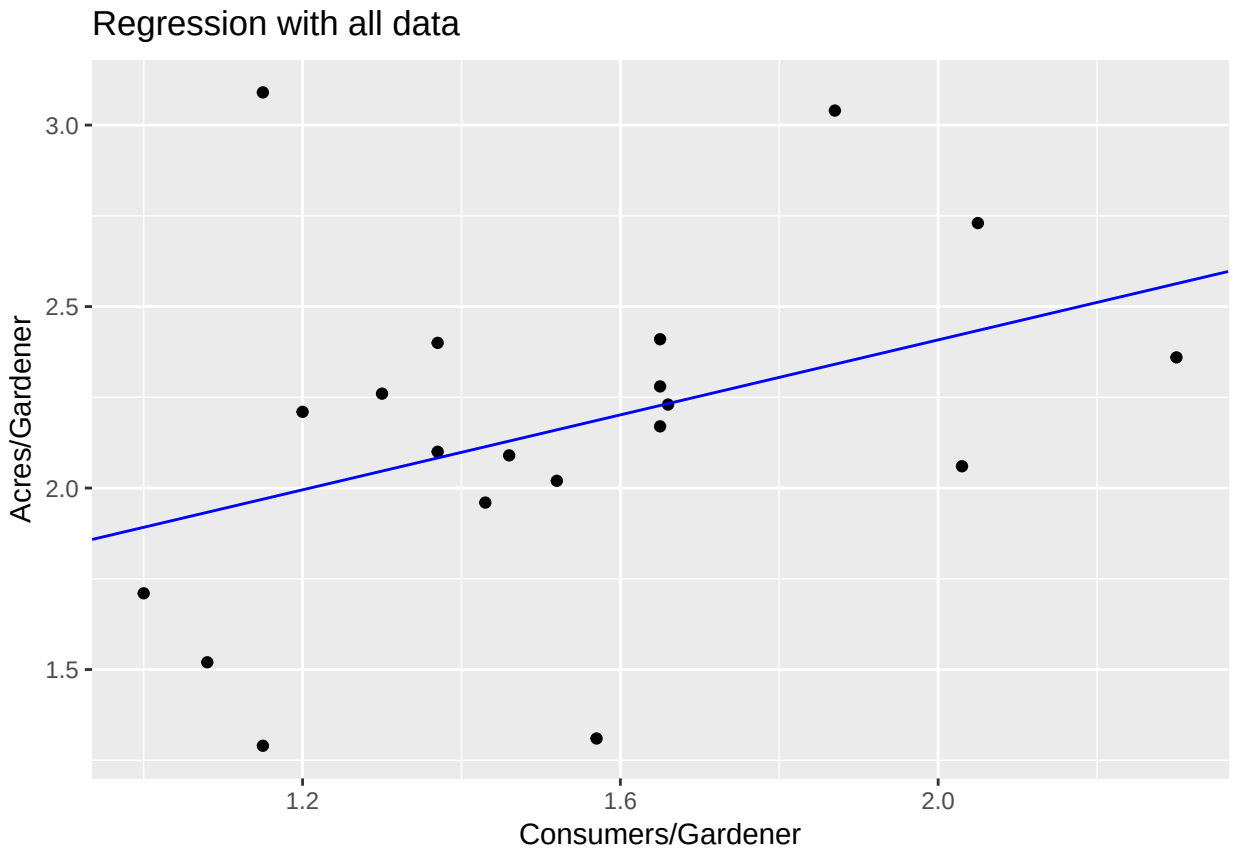
that complicate things, such as a single household which shows the hardest work (highest Acres value), but a very low consumer to gardener ratio.

b. Perform a regression of Acres/Gardener on Consumers/Gardener and analyze the results.

```
model1 <- lm(data = Sahlins, acres ~ consumers)
summary(model1)
```

```
##
## 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
```

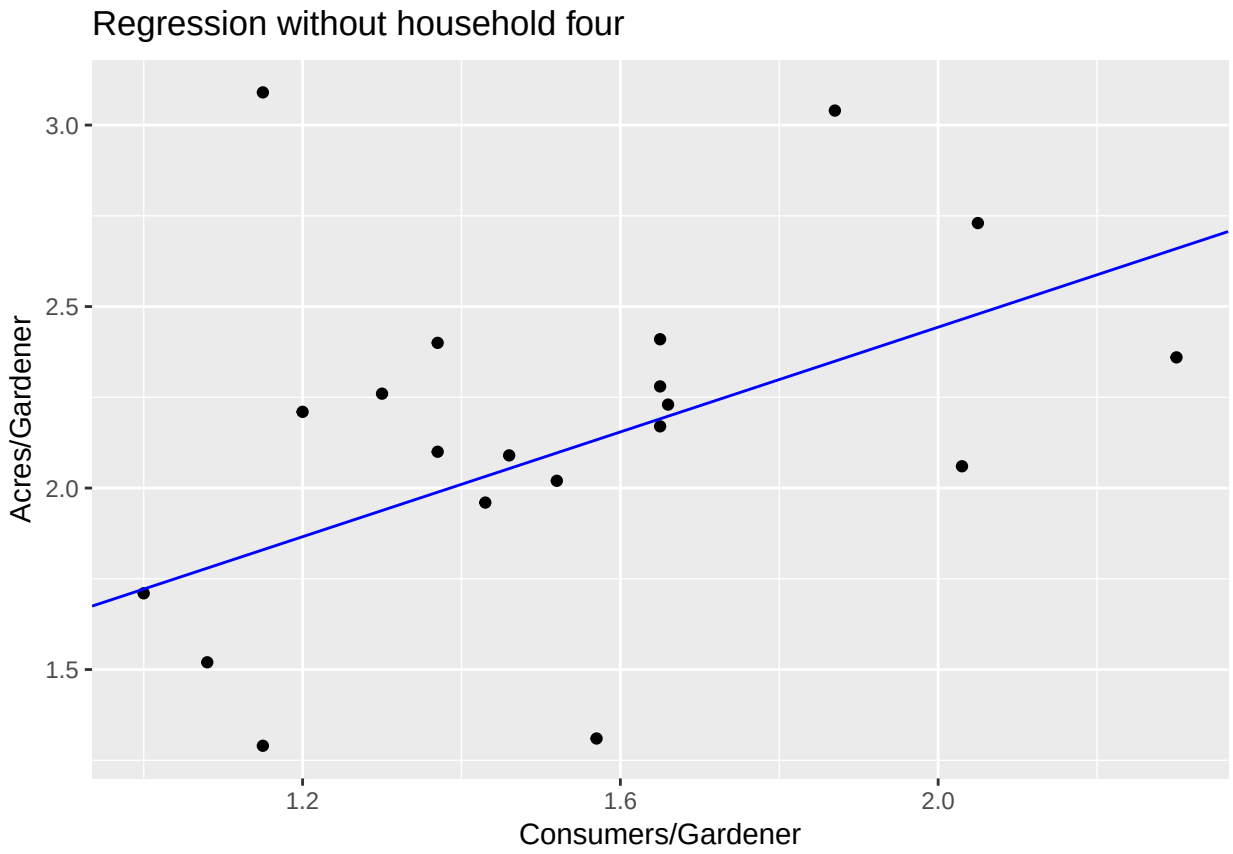
```
ggplot(Sahlins, aes(x = consumers, y = acres), ) + geom_point() + labs(x = "Consumers/Gardener",
    y = "Acres/Gardener", title = "Regression with all data") + geom_abline(intercept = 1.3756,
    slope = 0.5163, col = "blue")
```




```
model2 <- lm(data = Sahlins[-4, ], acres ~ consumers)
summary(model2)
```

```
##
## Call:
## lm(formula = acres ~ consumers, data = Sahlins[-4, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82291 -0.16808  0.03215  0.23505  0.69061
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.0000     0.3969   2.519   0.0221 *
## consumers     0.7216     0.2514   2.870   0.0106 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3681 on 17 degrees of freedom
## Multiple R-squared:  0.3264, Adjusted R-squared:  0.2868
## F-statistic: 8.238 on 1 and 17 DF,  p-value: 0.01061
```

```
ggplot(Sahlins, aes(x = consumers, y = acres), ) + geom_point() + labs(x = "Consumers/Gardener",
    y = "Acres/Gardener", title = "Regression without household four") + geom_abline(intercept = 1,
    slope = 0.7216, col = "blue")
```



From the first regression, containing all of the datapoints, we can see that the intercept, $\hat{\beta}_0$, is not 0 as might be predicted if this was a totally market based economy, but also the slope, $\hat{\beta}_1$ is not 0 as would be predicted if this was a completely communist economy. In fact, the intercept and slope are both positive. This suggests that there is both redistribution of food to needy households (positive $\hat{\beta}_0$ means unproductive households still consume more than they produce), and also a market economy influenced drive to produce more than the houshold's base consumption needs (positive $\hat{\beta}_1$). This means that, according to the model, this society is neither purely communist nor market capatilist, but rather shows a mix of both systems where households with higher needs are working harder per person, but those households which don't work as hard are also supported.

When the regression is performed without the fourth, outlying household, the model seems to do a better job representing the data. This can be seen graphically, with the regression line seeming to fit the data better, as well as through analysis of the R^2 values, and the residual standard errors and σ^2 values. In the case of R^2, we can see that the value is larger for the second model, implying better fit, while the value of σ^2 is smaller for the second model, indicating that there is less residual error and variance left unexplained in the response variable after the model is fit.

Overall, the second model, with household 4 excluded, seems to do a reasonable job in explaning the relationship between Acres/Gardener and Consumers/Gardener, though there is definitely still room for improvement.

c. Find the standard errors of the intercept and slope. Can we conclude that the population slope is greater than zero? Can we conclude that the intercept is greater than zero?

```
seb01 <- coef(summary(model1))[1, 2]
seb11 <- coef(summary(model1))[2, 2]

seb02 <- coef(summary(model2))[1, 2]
```

```r
seb12 <- coef(summary(model2))[2, 2]

seb01
```

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

```r
seb02
```

```
## [1] 0.3969254
```

```r
seb11
```

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

```r
seb12
```

```
## [1] 0.251414
```

```r
# can compute confidence intervals from lm() parameters using confint()

conf1 <- confint(model1, level = 0.95)
conf1
```

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

```r
conf2 <- confint(model2, level = 0.95)
conf2
```

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

```r
# Can find one tailed p-values corresponding to one-tailed hypothesis tests
# by taking the p-values generated by the lm() command and dividing them by
# two.

pb01 <- coef(summary(model1))[1, 4]/2
pb11 <- coef(summary(model1))[2, 4]/2

pb02 <- coef(summary(model2))[1, 4]/2
pb12 <- coef(summary(model2))[2, 4]/2

pb01
```

```
## [1] 0.004406897
```

```r
pb11
```

```
## [1] 0.05131463
```

```r
pb02
```

```
## [1] 0.01102734
```

```r
pb12
```

```
## [1] 0.005306328
```

After all this analysis, we can conclude that, for both models (household 4 included and excluded) the intercept is very likely greater than zero. This is shown by an intercept ($\hat{\beta}_0$) of 0 not being

included in the 95% confidence intervals for either model, meaning that we are 95% confident that the true intercept is contained in these intervals and so we are 95% confident β_0 is not 0. A similar conclusion is reached when our p-values are examined, and we see that for both models there is a very low probability (much lower than $\alpha = 0.05$) of finding a value for $\hat{\beta}_1$ at least as extreme as the one seen given that our null hypothesis (the intercept <= 0) is true. This means we can feel comfortable rejecting our null hypothesis and concluding the intercept is greater than zero.

When we consider the null hypothesis that the slope ($\hat{\beta}_1$) is less than or equal to 0, things are a little more complicated. Using the same confidence interval and hypothesis testing techniques as before, we see here that we can be most confident that the second model, with household 4 removed, has a slope greater than zero as the p-value is quite low and the 95% confidence interval does not contain 0. The confidencce interval of the first model though does just barely contain 0, and the p-value is more marginal at a level of $\alpha = 0.05$. So for $\hat{\beta}_1$ it is safer to conclude that the population slope is greater than zero when the obvious outlier is removed.

d. Using the regression for the entire dataset, predict the Acres/gardener ratio for a household with a Consumer/Gardener ratio of 1.5. Obtain an interval with a 98% confidence level.

```
predictacres <- data.frame(consumers = 1.5)

predict(model1, predictacres, interval = "predict", level = 0.98)
```

```
##        fit      lwr      upr
## 1 2.150125 0.961766 3.338483
```

Using the predict() command to generate a point prediction a 98% confidence interval for the prediction, we see that our point estimate for Acres/Gardener when the Consumers/Gardener ratio is equal to 1.5 is 2.15. Though this value is just the point on the regression line and doesn't give us as much information as the prediction interval, which is the estimate of the interval (here with 98% confidence) in which future observations of the acres ratio will fall when the consumer ratio is equal to 1.5. This interval here is quite wide and ranges from 0.962 to 3.338.

If we were instead interested in finding the mean Acres/Gardener ratio for all households with a Consumer/Gardener ratio equal to 1.5, these predictions would be quite different. This is because when we find fitted values and fitted value confidence intervals, there is no additional error term for the prediction error variability that arises when we estimate the coefficients like when we find prediction intervals. This difference can be easily seen by analyzing the formulas for the standard error of the predicted and fitted values.

$$sepred = \sqrt{\sigma^2(1 + \frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})} \quad and \quad sefit = \sqrt{\sigma^2(\frac{1}{n} + \frac{(x_* - \bar{x})^2}{SXX})}$$

So, the interval found for the mean Acres/Gardener ratio for all households with a Consumers/Gardener ratio equal to 1.5 would be smaller as there is one less error term to consider.