Multiple regression

KEY

Wed Mar 10 13:54:52 2021

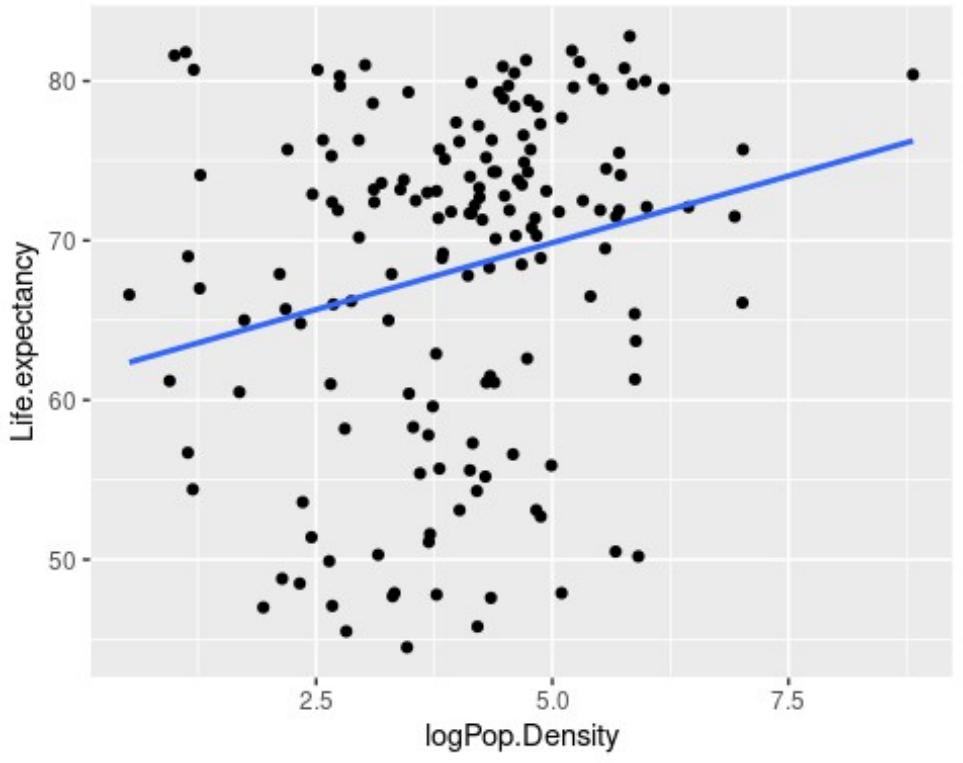
Import the GapMinder data:

```
library(readxl)
gapminder <- read_excel("gapminder.xlsx", "gapminder")
```

Case 1 - multiple regression enhances the significance of the relationship between variables

Make a scatterplot of life expectancy against the log of population density:

```
library(ggplot2)

ggplot(gapminder, aes(x = logPop.Density, y = Life.expectancy)) +
geom_point() + geom_smooth(method = "lm", se = F)

## `geom_smooth()` using formula 'y ~ x'
```



Question: does it appear from the graph that there is a relationship between life expectancy and the log of population density? Is it positive or negative?

Yes, the line has a positive slope, so it appears there is a positive relationship.

Run a linear model of life expectancy on log of population density - with just one predictor this is a simple linear regression model. Get the summary of the fitted model, and an ANOvA table.

```
life.logpop.lm <- lm(Life.expectancy ~ logPop.Density, data = gapminder)
summary(life.logpop.lm)

##
## Call:
## lm(formula = Life.expectancy ~ logPop.Density, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -22.786  -7.227   2.835   7.456  18.436
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     61.4903     2.4365  25.237  < 2e-16 ***
## logPop.Density   1.6739     0.5744   2.914  0.00408 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.12 on 160 degrees of freedom
## Multiple R-squared:  0.05041,    Adjusted R-squared:  0.04447
## F-statistic: 8.493 on 1 and 160 DF,  p-value: 0.004075

anova(life.logpop.lm)

## Analysis of Variance Table
##
## Response: Life.expectancy
##                 Df  Sum Sq Mean Sq F value   Pr(>F)
## logPop.Density   1   869.7  869.71  8.4933 0.004075 **
## Residuals      160 16383.8  102.40
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

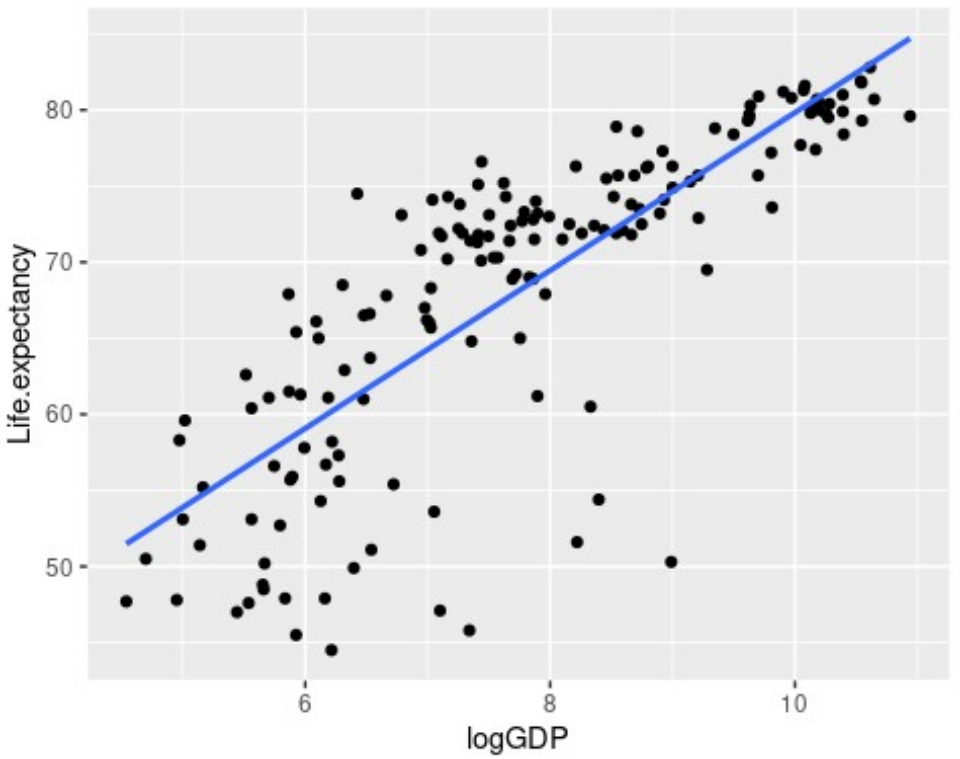
Question: is logPop.Density statistically significant? Is it a good predictor of variation in life expectancy? What statistics did you use to answer each of these questions?

Yes it is statistically significant, but the R^2 is low (0.05). The p-value indicates significance, and R^2 is a measure of how well the predictor accounts for variation in the response.

Now make a scatterplot of life expectancy against the log of gross domestic product:

```
ggplot(gapminder, aes(x = logGDP, y = Life.expectancy)) + geom_point() +
geom_smooth(method = "lm", se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Run a linear model of life expectancy on log of GDP - this too will be a simple linear regression. Get the summary of the fitted model, and an ANOvA table.

```
life.loggdp.lm <- lm(Life.expectancy ~ logGDP, data = gapminder)
summary(life.loggdp.lm)

##
## Call:
## lm(formula = Life.expectancy ~ logGDP, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -24.288  -2.462   1.064   3.842  13.220
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  27.9361     2.4063   11.61   <2e-16 ***
## logGDP        5.1897     0.3038   17.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.179 on 160 degrees of freedom
## Multiple R-squared:  0.6459, Adjusted R-squared:  0.6437
## F-statistic: 291.9 on 1 and 160 DF,  p-value: < 2.2e-16
```

```
anova(life.loggdp.lm)
```

```
## Analysis of Variance Table
##
## Response: Life.expectancy
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## logGDP      1 11144.5 11144.5  291.89 < 2.2e-16 ***
## Residuals 160  6108.9    38.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: is logGDP statistically significant? Is it a good predictor of variation in life expectancy (at least, does it seem to be a better predictor than logPop.Density)? How do you know?

Yes it is statistically significant. It is a much better predictor of life expectancy than logPop.Density was because R^2 is 0.65.

Correlate the two predictors before using them in a model together:

```
with(gapminder, cor(logPop.Density, logGDP))
```

```
## [1] 0.09463973
```

Question: is there a large correlation between logGDP and logPop.Density?

No, 0.09 is a weak relationship.

Now run a multiple regression, using both logPop.Density and logGDP as predictors. Get the summary of the fitted model and the ANOVA table (using the anova() function will give you Type I SS).

```
life.logpop.loggdp.lm <- lm(Life.expectancy ~ logPop.Density + logGDP, data =
gapminder)
summary(life.logpop.loggdp.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ logPop.Density + logGDP, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.2227  -2.5698   0.7908   3.9722  11.3545
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     24.1680     2.6064   9.273  < 2e-16 ***
## logPop.Density   1.1168     0.3421   3.264  0.00134 **
## logGDP           5.0982     0.2963  17.205  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 6.001 on 159 degrees of freedom
## Multiple R-squared:  0.6682, Adjusted R-squared:  0.664
## F-statistic: 160.1 on 2 and 159 DF,  p-value: < 2.2e-16

anova(life.logpop.loggdp.lm)

## Analysis of Variance Table
##
## Response: Life.expectancy
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## logPop.Density   1   869.7   869.7  24.153 2.196e-06 ***
## logGDP           1 10658.5 10658.5 296.006 < 2.2e-16 ***
## Residuals      159  5725.2    36.0
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: how did the p-value for logPop.Density and logGDP change when you included both in the model together?

The p-value for logPop.Density got smaller. It didn't obviously change for logGDP because it was already at < 2e-16, but the t-value for logGDP got slightly bigger.

Question: how did the slope coefficients change, compared to their values when you included just one of the predictors at a time?

The slope coefficients got smaller for both variables.

Question: which of these three models (the two that included only a single predictor, and the one that included both at once) had the largest R^2? What does that tell you about which model did the best job of predicting life expectancy?

The model with both predictors had the highest R^2, 0.6682. This tells us that the model with both predictors is better at explaining variation in life expectancy than either predictor alone.

Now get the Type II SS ANOVA table for this multiple regression:

```
library(car)

## Loading required package: carData

Anova(life.logpop.loggdp.lm)

## Anova Table (Type II tests)
##
## Response: Life.expectancy
##                 Sum Sq  Df F value    Pr(>F)
## logPop.Density   383.7   1  10.655  0.001344 **
## logGDP         10658.5   1 296.006 < 2.2e-16 ***
## Residuals       5725.2 159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Refer to the 3D plot of these data, showing the plane that represents the model predicted values and residuals indicating how far each country is from the plane.

Question: based on your interpretation of the residuals on the 3D graph on the course web site, which countries had high life expectancies for their logPop.Density and logGDP? Which had short life expectances for their logPop.Density and logGDP?

> Big positive residuals indicate higher life expectancies than expected given the country's logPop.Density and logGDP. The biggest positive residuals were from Albania, Vietnam, and Kyrgystan. Big negative residuals indicate lower than expected life expectancies, which were found with Swaziland, Equatorial Guinea, and South Africa.

Calculate the standardized coefficients:

```
library(biol531)
stdcoeff(life.logpop.loggdp.lm)

## logPop.Density         logGDP
##      0.1497962      0.7895216
```

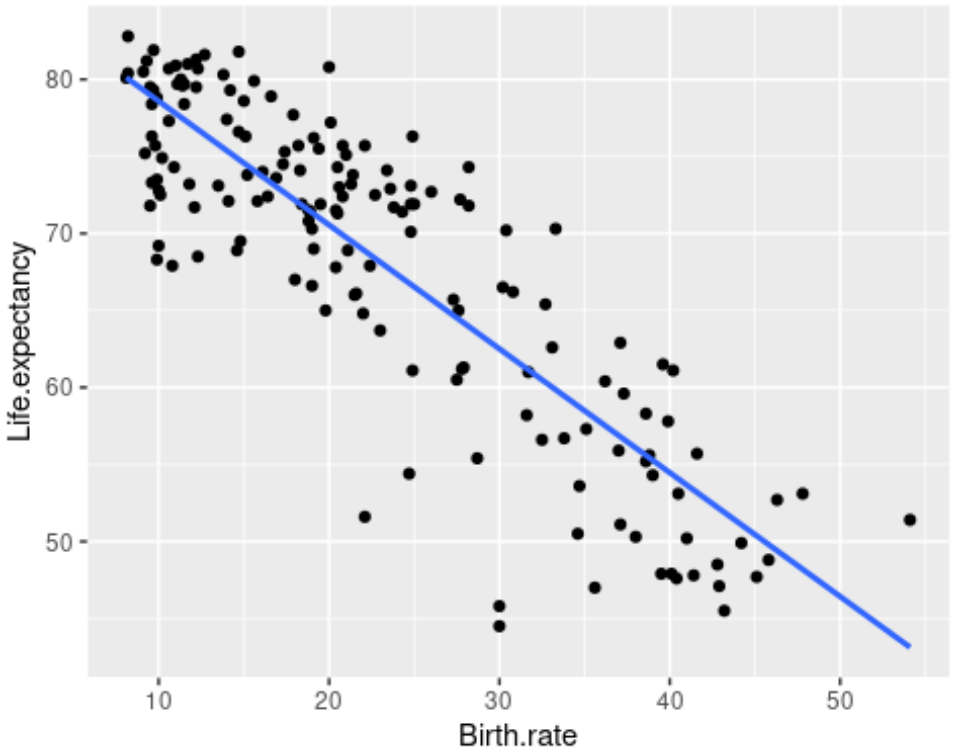
Question: which predictor is better at predicting life expectancy? Do you need to worry about the fact that poulation density is measured in people per square mile, and that GDP is measured in dollars per person? Why or why not (that is, do standardized coefficients correct for the problem)?

> The standardized coefficient for logGDP is 5.3 times bigger than the coefficient for logPop.Density. Using standardized coefficients expresses the slopes in standard deviation units, which accounts for the differences in measurement units between the predictors.

Effects of strongly correlated predictors can't be told apart

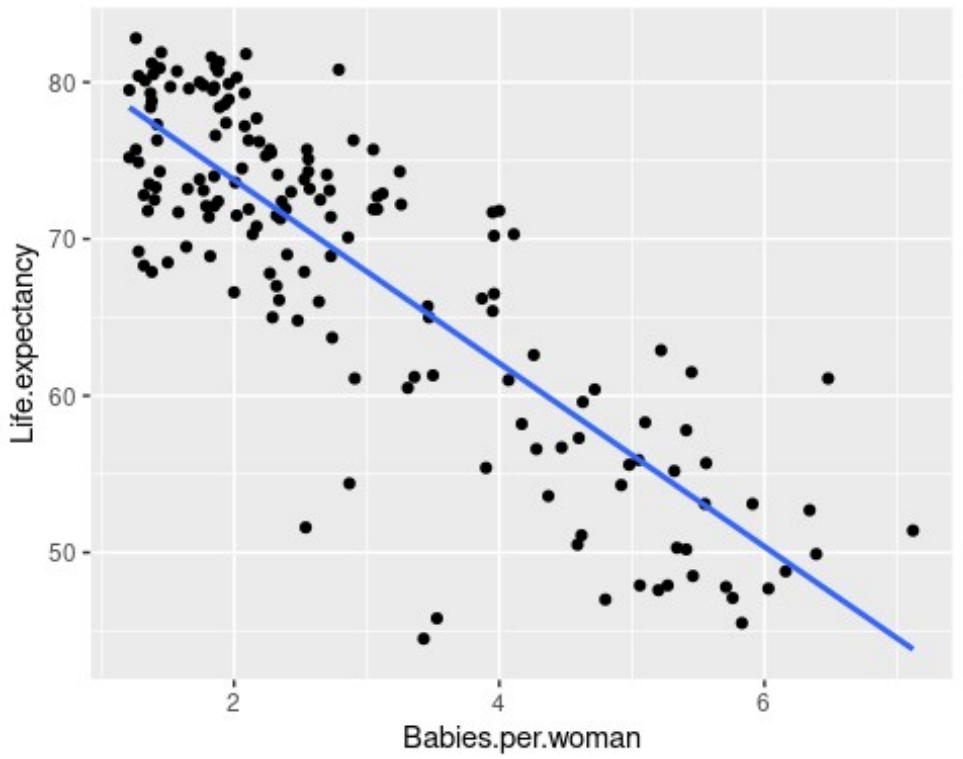
Make a scatter plot of life expectancy on birth rate:

```
ggplot(gapminder, aes(x = Birth.rate, y = Life.expectancy)) + geom_point() +
geom_smooth(method = "lm", se = F)

## `geom_smooth()` using formula 'y ~ x'
```



Now make a scatter plot of life expectancy on babies per woman:

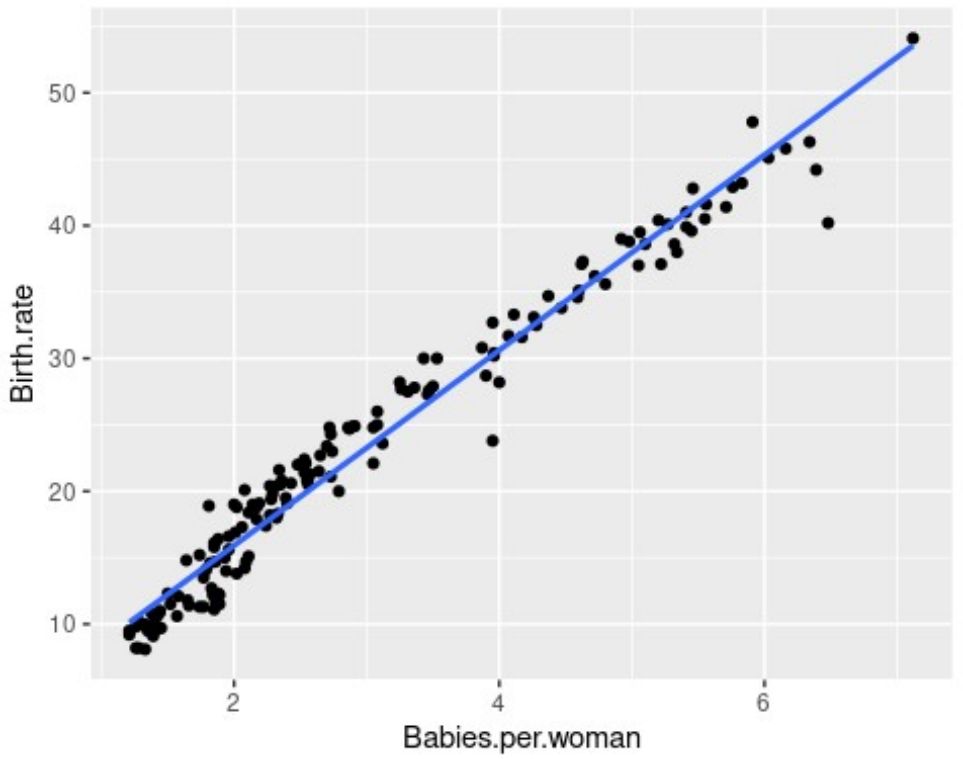
```
ggplot(gapminder, aes(x = Babies.per.woman, y = Life.expectancy)) +
geom_point() + geom_smooth(method = "lm", se = F)

## `geom_smooth()` using formula 'y ~ x'
```



And, make a scatter plot of birth rate on babies per woman:

```
ggplot(gapminder, aes(x = Babies.per.woman, y = Birth.rate)) + geom_point() +
geom_smooth(method = "lm", se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
with(gapminder, cor(Birth.rate, Babies.per.woman))
```

```
## [1] 0.9836564
```

Run a linear model of life expectancy on birth rate alone, one with babies per woman alone, and one with both birth rate and babies per woman.

First birth rate alone:

```
life.birth.rate.lm <- lm(Life.expectancy ~ Birth.rate, data = gapminder)
summary(life.birth.rate.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Birth.rate, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.0074  -3.5135   0.4865   4.0114  10.4465
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 86.63401    0.96515   89.76   <2e-16 ***
## Birth.rate  -0.80422    0.03795  -21.19   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 5.323 on 160 degrees of freedom
## Multiple R-squared:  0.7373, Adjusted R-squared:  0.7356
## F-statistic:   449 on 1 and 160 DF,  p-value: < 2.2e-16
```

```
Anova(life.birth.rate.lm)
```

```
## Anova Table (Type II tests)
##
## Response: Life.expectancy
##            Sum Sq  Df F value    Pr(>F)
## Birth.rate  12720   1  448.99 < 2.2e-16 ***
## Residuals    4533 160
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: is birth rate a good predictor of life expectancy? How do you know?

Yes, an excellent predictor - the p-value is less than 0.05, and R^2 is 0.74.

Now run a linear model with babies per woman alone:

```
life.bpw.lm <- lm(Life.expectancy ~ Babies.per.woman, data = gapminder)
summary(life.bpw.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Babies.per.woman, data = gapminder)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -20.906  -3.800   0.792   3.984  13.533
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        85.467      1.003   85.21   <2e-16 ***
## Babies.per.woman   -5.849      0.304  -19.24   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.705 on 160 degrees of freedom
## Multiple R-squared:  0.6982, Adjusted R-squared:  0.6963
## F-statistic: 370.2 on 1 and 160 DF,  p-value: < 2.2e-16
```

```
Anova(life.bpw.lm)
```

```
## Anova Table (Type II tests)
##
## Response: Life.expectancy
##                   Sum Sq  Df F value    Pr(>F)
## Babies.per.woman 12046.4   1  370.15 < 2.2e-16 ***
## Residuals         5207.1 160
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: is babies per woman a good predictor of life expectancy? How do you know?

Also yes, the p-value is less than 0.05, and R^2 is 0.698.

Finally, run a linear model with both predictors together:

```
life.bpw.br.lm <- lm(Life.expectancy ~ Babies.per.woman + Birth.rate, data =
gapminder)
summary(life.bpw.br.lm)

##
## Call:
## lm(formula = Life.expectancy ~ Babies.per.woman + Birth.rate,
##     data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1226  -3.4903   0.6979   3.8153  11.1189
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       86.7603     0.9689  89.544  < 2e-16 ***
## Babies.per.woman   1.9491     1.5726   1.239    0.217
## Birth.rate        -1.0608     0.2104  -5.041 1.25e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.314 on 159 degrees of freedom
## Multiple R-squared:  0.7398, Adjusted R-squared:  0.7365
## F-statistic:   226 on 2 and 159 DF,  p-value: < 2.2e-16

Anova(life.bpw.br.lm)

## Anova Table (Type II tests)
##
## Response: Life.expectancy
##                  Sum Sq  Df F value    Pr(>F)
## Babies.per.woman   43.4   1  1.5361     0.217
## Birth.rate        717.5   1 25.4098 1.249e-06 ***
## Residuals        4489.6 159
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

anova(life.bpw.br.lm)

## Analysis of Variance Table
##
## Response: Life.expectancy
```

```
##                    Df  Sum Sq Mean Sq F value    Pr(>F)
## Babies.per.woman    1 12046.4 12046.4  426.62 < 2.2e-16 ***
## Birth.rate          1   717.5   717.5   25.41 1.249e-06 ***
## Residuals         159  4489.6    28.2
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: are both of the predictors still statistically significant when you include both in the same model?

No, babies per woman becomes non-significant.

Question: Birth.rate is statistically significant, but Babies.per.woman is not. Does this mean that Birth.rate is an excellent predictor of life expectancy, but Babies.per.woman is not? Why or why not?

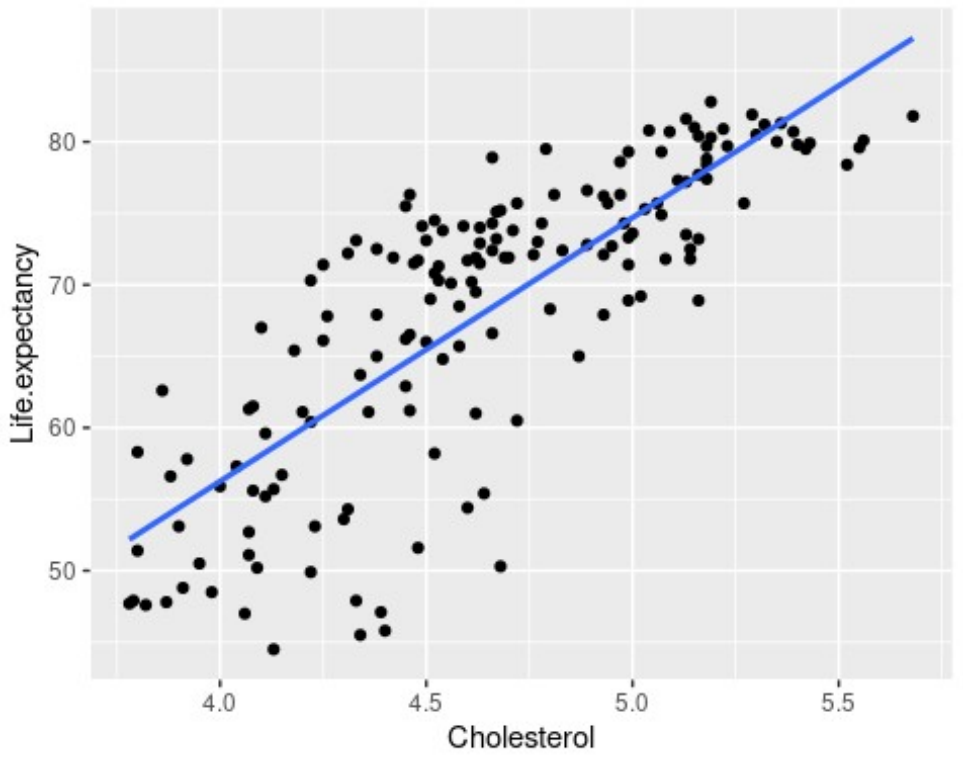
No, most of the variation that birth rate explains is shared with babies per woman, and only a little bit is uniquely attributable to birth rate. Rather than basing a firm conclusion on such a small amount of independent variation, it is better to conclude that while reproduction seems to be associated with lowered life expectancy, we can't attribute the relationship to one of our reproductive measures or the other.

Multiple regression can identify spurious relationships

First, make a graph of life expectancy on cholesterol:

```
ggplot(gapminder, aes(x = Cholesterol, y = Life.expectancy)) + geom_point() +
geom_smooth(method = "lm", se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Now run a regression of life expectancy on cholesterol:

```
life.cholesterol.lm <- lm(Life.expectancy ~ Cholesterol, data = gapminder)
summary(life.cholesterol.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Cholesterol, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -18.4935  -3.8816   0.5874   4.3725  11.5649
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -17.539      5.059  -3.467 0.000676 ***
## Cholesterol   18.447      1.083  17.028  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.192 on 160 degrees of freedom
## Multiple R-squared:  0.6444, Adjusted R-squared:  0.6422
## F-statistic:   290 on 1 and 160 DF,  p-value: < 2.2e-16
```

Question: according to the slope coefficient, what is the relationship between cholesterol level and life expectancy across different countries of the world? Is this what you would expect, given the effect of cholesterol on cardiovascular health?

The slope is 18.447, which is positive - this is indicating that an increase in cholesterol in the diet is associated with longer life. This is not expected, cholesterol is supposed to be bad for us.

We expect that cholesterol is correlated with wealth (as indicated by logGDP) and healthcare (as indicated by TB rates and maternal mortality). Get the correlation coefficients:

```
cor(gapminder[,c("Cholesterol", "logGDP","Maternal.mortality","TB")])
```

```
##                     Cholesterol     logGDP Maternal.mortality         TB
## Cholesterol           1.0000000  0.8922762         -0.7127377 -0.6653718
## logGDP                0.8922762  1.0000000         -0.6947494 -0.6537176
## Maternal.mortality   -0.7127377 -0.6947494          1.0000000  0.7400895
## TB                   -0.6653718 -0.6537176          0.7400895  1.0000000
```

Question: these are fairly large correlations, but do these correlations measure all of the confounded variation between these variables?

No, these are only correlations between pairs of variables. Combinations of several variables can be even more inter-related than this.

Now run a model with indicators of wealth included along with cholesterol - use cholesterol, logGDP, Maternal.mortality, and TB as predictors of life expectancy:

```
life.cholesterol.wealth.lm <- lm(Life.expectancy ~ Cholesterol + logGDP +
Maternal.mortality + TB, data = gapminder)
summary(life.cholesterol.wealth.lm)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Cholesterol + logGDP + Maternal.mortality +

##     TB, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2809  -2.0207   0.3885   2.3027  11.8985
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)        48.412967   5.581961   8.673 4.98e-15 ***
## Cholesterol         2.973237   1.613061   1.843 0.067181 .
## logGDP              1.490412   0.441504   3.376 0.000928 ***
## Maternal.mortality -0.011630   0.001791  -6.492 1.05e-09 ***
## TB                 -0.016278   0.002175  -7.483 4.88e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.965 on 157 degrees of freedom
```

```
## Multiple R-squared:  0.857,  Adjusted R-squared:  0.8533
## F-statistic: 235.1 on 4 and 157 DF,  p-value: < 2.2e-16
```

Run the anova() command to get Type I SS:

```
anova(life.cholesterol.wealth.lm)
```

```
## Analysis of Variance Table
##
## Response: Life.expectancy
##                      Df  Sum Sq Mean Sq F value    Pr(>F)
## Cholesterol           1 11118.4 11118.4 707.310 < 2.2e-16 ***
## logGDP                1   646.8   646.8  41.148 1.582e-09 ***
## Maternal.mortality    1  2140.2  2140.2 136.149 < 2.2e-16 ***
## TB                    1   880.1   880.1  55.990 4.876e-12 ***
## Residuals           157  2467.9    15.7
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Now get the Type II SS with the Anova() command from the car library:

```
library(car)
Anova(life.cholesterol.wealth.lm)
```

```
## Anova Table (Type II tests)
##
## Response: Life.expectancy
##                      Sum Sq  Df F value    Pr(>F)
## Cholesterol           53.41   1  3.3975 0.0671814 .
## logGDP               179.13   1 11.3957 0.0009278 ***
## Maternal.mortality   662.58   1 42.1509 1.055e-09 ***
## TB                   880.12   1 55.9898 4.876e-12 ***
## Residuals           2467.93 157
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: cholesterol is significant in the Type I SS table, but not in the Type II SS table. Why?

> Cholesterol is confounded with the other variables, so when it is tested first in a Type I ANOVA it is statistically significant because it is assigned the variation it shares with the other predictors. In a Type II ANOVA it is only assigned the variation that is uniquely attributable to it, and there is not much of that (not enough to be significant).

Calculate the residuals for life expectancy, and for cholesterol, accounting for logGDP, TB, and Maternal.mortality:

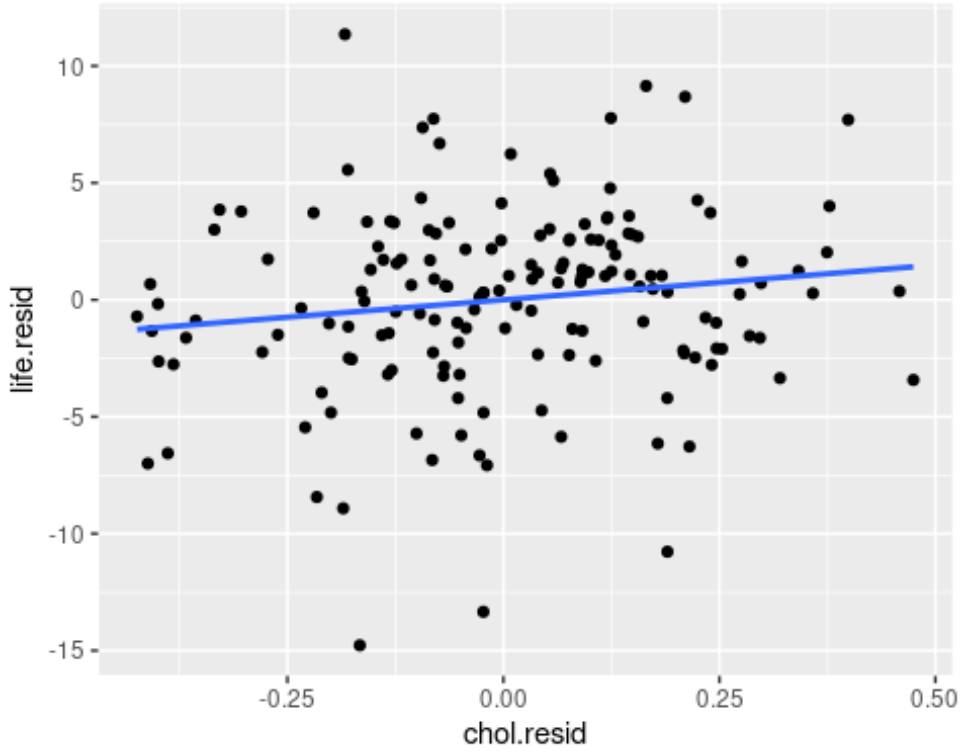
```
gapminder$life.resid <- residuals(lm(Life.expectancy ~ logGDP +
Maternal.mortality + TB, data = gapminder))
```

```
gapminder$chol.resid <- residuals(lm(Cholesterol ~ logGDP +
Maternal.mortality + TB, data = gapminder))
```

Graph life.resid against chol.resid to see how much relationship there is between life expectancy and cholesterol after logGDP, Maternal.mortality, and TB have been eliminated:

```
ggplot(gapminder, aes(x = chol.resid, y = life.resid)) + geom_point() +
geom_smooth(method = "lm", se = F)
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Run a linear model that explains life.resid with chol.resid:

```
life.chol.resid.lm <- lm(life.resid ~ chol.resid, data = gapminder)
summary(life.chol.resid.lm)
```

```
##
## Call:
## lm(formula = life.resid ~ chol.resid, data = gapminder)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -14.2809  -2.0207   0.3885   2.3027  11.8985
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -4.486e-16  3.086e-01   0.000   1.0000
## chol.resid   2.973e+00  1.598e+00   1.861   0.0646 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.927 on 160 degrees of freedom
## Multiple R-squared:  0.02118,    Adjusted R-squared:  0.01506
## F-statistic: 3.462 on 1 and 160 DF,  p-value: 0.06461
```

```
anova(life.chol.resid.lm)
```

```
## Analysis of Variance Table
##
## Response: life.resid
##             Df  Sum Sq Mean Sq F value  Pr(>F)
## chol.resid   1   53.41  53.406  3.4624 0.06461 .
## Residuals  160 2467.93  15.425
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Question: is cholesterol still a good predictor of life expectancy when it is included with measures of wealth?

> No, it is not statistically significant when it is included with the predictors that indicate wealth.

Question: does this prove that the relationship between life expectancy and cholesterol was spurious? Why might we think it's spurious, even though the statistical evidence can't tell us this for sure?

> It does not prove that the relationship was spurious by itself, because there are other possible explanations - it could be that rich countries are able to afford cholesterol-rich foods which are healthy and contribute to long life. All we know from the statistics is that cholesterol's effects on life expectancy are not independent of wealth. But, because of what we know about the effects of cholesterol on the cardiovascular system, and what we know about the effects of wealth on life expectancy, we can feel fairly confident that the relationship is spurious, and we do not need to re-write the medical books based on these data.