Log-Transforming the Predictor

2017-12-06

Read in Data

The data we will use in this set of notes, *infant-mortality.csv*, contains country-level data on the infant mortality rates and risk factors for several countries.

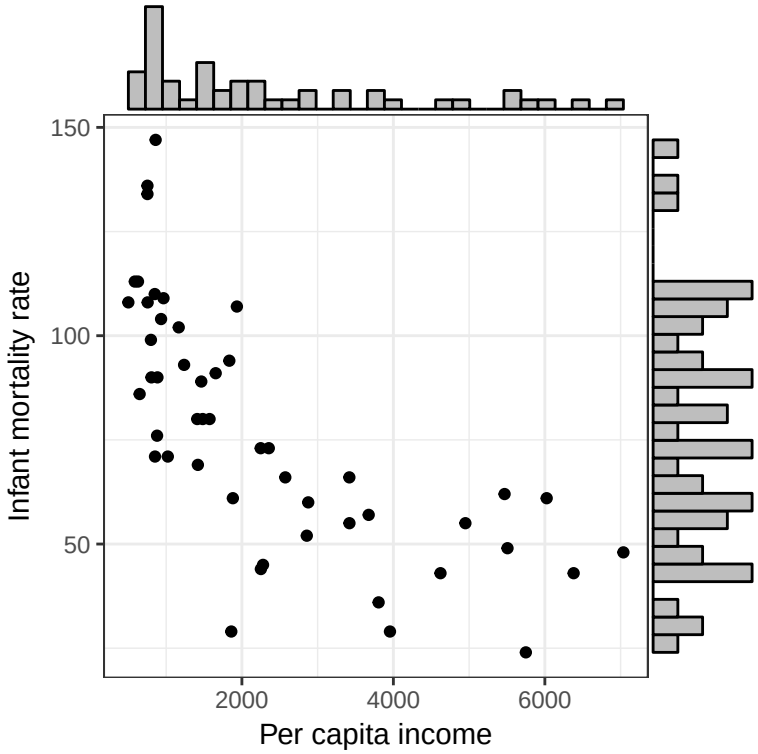
```
# Load libraries
library(broom)
library(dplyr)
library(ggplot2)
library(ggExtra)
library(readr)
library(sm)

# Read in data
infant = read_csv(file = "~/Dropbox/epsy-8251/data/infant-mortality.csv")
head(infant)
```

```
# A tibble: 6 x 6
       country mortality   pci youngmom oldmom close
         <chr>     <int> <int>    <int>  <int> <int>
1   Bangladesh        80  1483       27      2    16
2        Benin       104   933       16      5    17
3      Bolivia        73  2355       13      5    28
4       Brazil        48  7037       19      3    29
5 Burkina Faso       109   965       17      5    17
6     Cameroon        80  1573       21      4    25
```

Relationship between Per Capita Income and Infant Mortality Rate

The scatterplot of Per Capita Income (PCI) and infant mortality rates suggests that the relationship between these variables may be curvilinear.

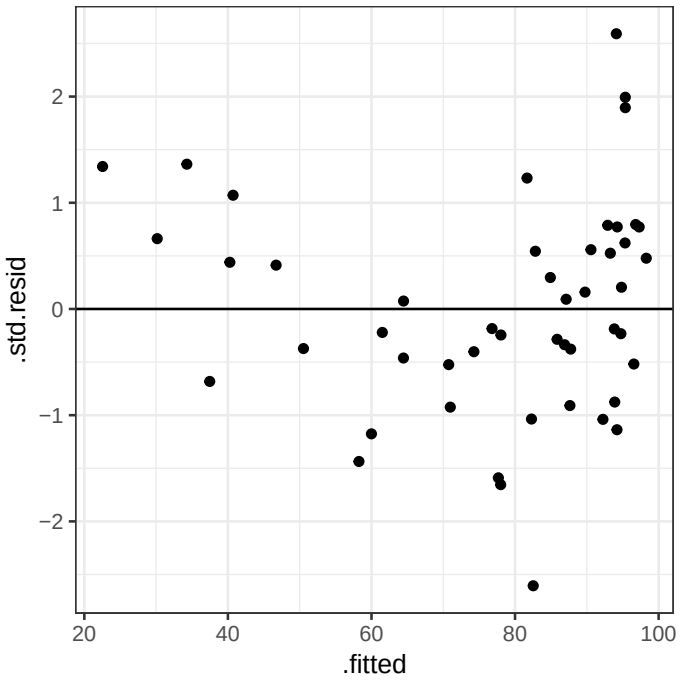


If we are unsure, we can fit the regression model and examine the residuals.

```
lm.1 = lm(mortality ~ 1 + pci, data = infant)

out1 = augment(lm.1)

ggplot(data = out1, aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw()
```



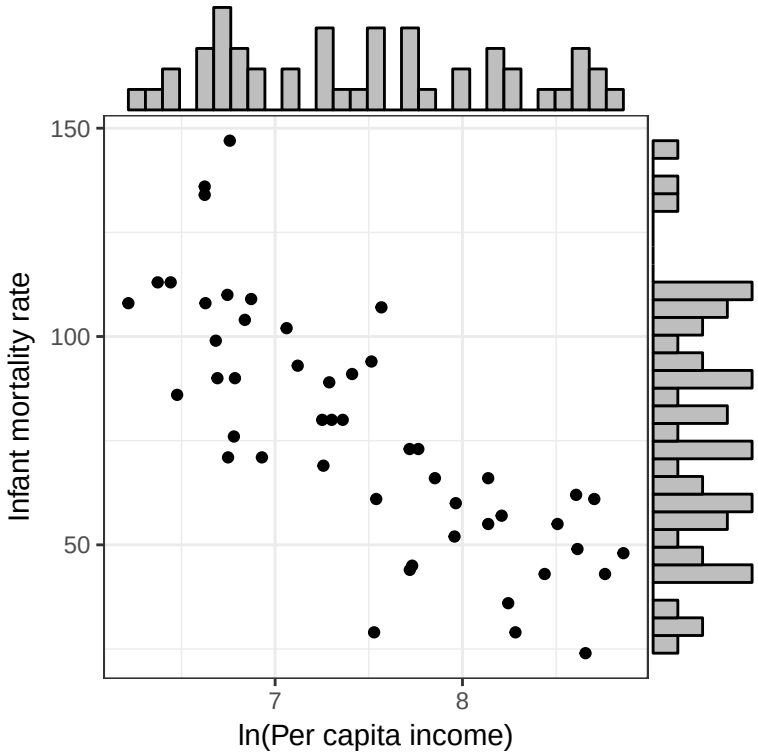
The residuals suggest a non-linear realationship.

Log-Transforming the Predictor

Looking at the histogram of PCI, we see the distribution is right-skewed. Furthermore, the relationship shows an *exponential decay* function. This might be alleviated if we log-transform right-skewed variables. To do this, since the X-variable is skewed, we fit a model that predicts Y using the logarithm of X.

```
infant = infant %>% mutate(Lpci = log(pci))
head(infant)
```

```
# A tibble: 6 x 7
       country mortality   pci youngmom oldmom close     Lpci
         <chr>     <int> <int>    <int>  <int> <int>    <dbl>
1   Bangladesh        80  1483       27      2    16 7.301822
2        Benin       104   933       16      5    17 6.838405
3      Bolivia        73  2355       13      5    28 7.764296
4       Brazil        48  7037       19      3    29 8.858937
5 Burkina Faso       109   965       17      5    17 6.872128
6     Cameroon        80  1573       21      4    25 7.360740
```



- The log-transformed PCI variable is reasonably symmetric.
- The relationship between `Lpci` and `mortality` is reasonably linear.

Fitting the Regression Model

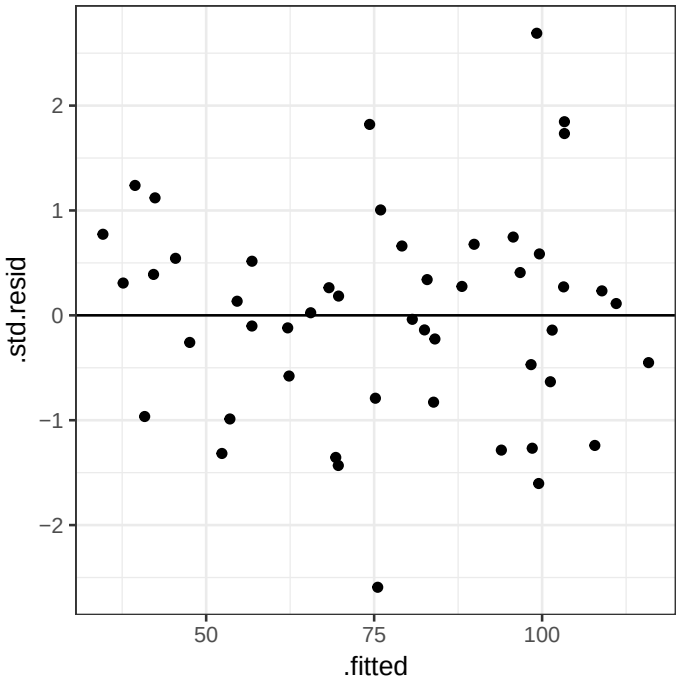
We now fit the model, using the predictor `Lpci`.

```
lm.2 = lm(mortality ~ 1 + Lpci, data = infant)
```

Before examining the coefficients, we can scrutinize the residuals to see whether the log-transformation helped us meet the assumption of linearity.

```
# Obtain residuals
out = augment(lm.2)

# Check linearity assumptions
ggplot(data = out, aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw()
```



The assumption looks reasonably met as the horizontal line of $y = 0$ is encompassed in the confidence envelope of the loess smoother.

Interpret the Regression Results

We can now look at the `summary()` output and interpret the output.

```
summary(lm.2)
```

```
Call:
lm(formula = mortality ~ 1 + Lpci, data = infant)

Residuals:
    Min      1Q  Median      3Q     Max
-46.531 -11.231   1.972   9.557  47.797

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  306.899     26.240  11.696 1.61e-15 ***
Lpci         -30.733      3.493  -8.798 1.69e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.13 on 47 degrees of freedom
Multiple R-squared:  0.6222,    Adjusted R-squared:  0.6142
F-statistic: 77.41 on 1 and 47 DF,  p-value: 1.685e-11
```

Examining the model-level output, we see that differences in ln(PCI) explain 62.2% of the variation in infant mortality rates. This is statistically significant, $F(1,\ 47) = 77.41$, $p < .001$. Since differences in ln(PCI) imply that there are differences in the raw PCI values, we would typically just say that "differences in per capita income explains 62.2% of the variation in infant mortality rates."

Moving to the coefficient-level output, we can write the fitted equation as,

$$\widehat{\text{Infant Mortality Rate}} = 306.9 - 30.7\Big[\ln(\text{PCI})\Big]$$

We can interpret the coefficients as we always do, recognizing that these interpretation are based on the log-transformed predictor.

- The intercept value of 306.9 is the predicted average infant mortality rate for countries with a ln(PCI) value of 0.
- The slope value of -30.7 indicates that each one-unit difference in ln(PCI) is associated with a -30.7-unit difference in infant mortality rate, on average.

Better Interpretations: Back-transforming

While these interpretations are technically correct, it is more helpful to your readers (and more conventional) to interpret any regression results in the metric of PCI rather than log-transformed PCI. This means we have to back-transform the interpretations. To back-transform a logarithm, we use its inverse function; exponentiation.

We interpreted the intercept as, "the predicted average infant mortality rate for all countries with a ln(PCI) value of 0". To interpret this using the metric of raw PCI, we have to recall that ln(PCI) = 0 can be re-expressed as $e^0 = \text{PCI}$, which implies that PCI = 1. Thus, rather than using the log-transformed interpretation, we can, instead, interpret the intercept as,

- The predicted average infant mortality rate for all countriess with a PCI of 1 is 306.9.

Since there are no countries in our data that have a PCI of 1, this is extrapolation.

What about the slope? Our interpretation was that "each one-unit difference in ln(PCI) is associated with a -30.7-unit difference in infant mortality rate, on average."

Remember that if we use the natural logarithm, we can think about the raw-variable in terms of "percent change". In this example, we transformed X. Instead of talking about a one-unit difference in ln(PCI) we can refer to a "one-percent difference in PCI".

Consider three countries, each having a PCI that differs by 1%; say these countries have PCI values of 1000, 1010, 1020.1. Using the fitted equation, we can compute the predicted infant mortality rate for each of these hypothetical scountries. The PCI values and predicted infant mortality rates for these countries are given below:

pci	mortality
1000.0	94.83191
1010.0	94.52644
1020.1	94.22096

Examine the differences between each subsequent pair of predicted infant mortality rates.

```
94.22096 - 94.52644
```

```
[1] -0.30548
```

```
94.52644 - 94.83191
```

```
[1] -0.30547
```

This difference is -0.306. In other words, for countries whose PCI differs by 1%, their predicted infant mortality rate differs by -0.306, on average. This is essentially the slope term divided by 100.

$$\frac{\hat{\beta}_1}{100}$$

Now we have an interpretation for our slope coefficient using the raw metric of SAT score:

- Each 1% difference in PCI is associated with a $-.306$-unit difference in predicted infant mortality rate, on average.

Plot of the Back-Transformed Model

Since the predictor in the model is ln(PCI), we set up a data frame that includes a sequence of ln(PCI) values. Based on the raw data, a reasonable range for the ln(PCI) values is 6.217 to 8.859. Then we can predict using the `lm.2` fitted model.

```
plotdata = data.frame(
  Lpci = seq(from = 6.217, to = 8.859, by = .001)
) %>%
  mutate(yhat = predict(lm.2, newdata = .))

head(plotdata)
```

```
   Lpci     yhat
1 6.217 115.8320
2 6.218 115.8012
3 6.219 115.7705
```

```
4 6.220 115.7398
5 6.221 115.7090
6 6.222 115.6783
```

Once we have predicted the infant mortality rates, we can back-transform the `Lpci` predictor to raw PCI.

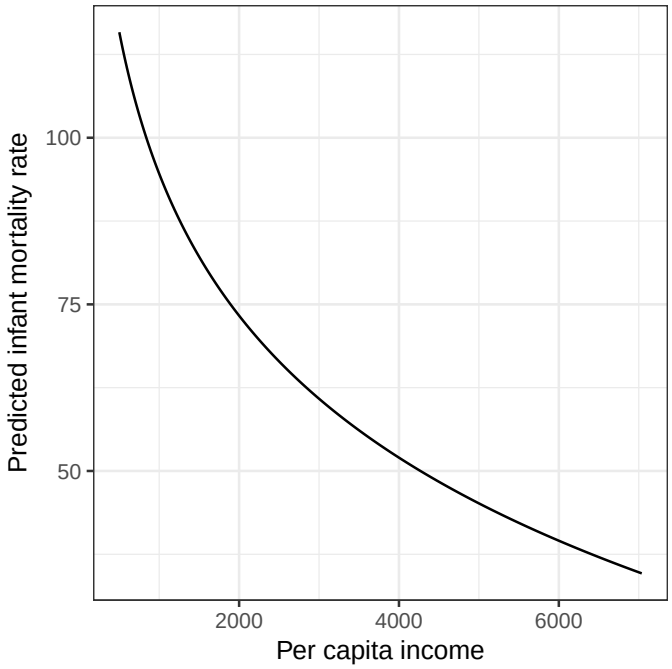
```
plotdata = plotdata %>%
  mutate( pci = exp(Lpci) )

head(plotdata)
```

```
   Lpci     yhat      pci
1 6.217 115.8320 501.1974
2 6.218 115.8012 501.6988
3 6.219 115.7705 502.2008
4 6.220 115.7398 502.7032
5 6.221 115.7090 503.2062
6 6.222 115.6783 503.7096
```

Now we plot using the raw SAT scores on the x-axis.

```
ggplot(data = plotdata, aes(x = pci, y = yhat)) +
  geom_line() +
  theme_bw() +
  xlab("Per capita income") +
  ylab("Predicted infant mortality rate")
```



This plot shows the non-linearity in the relationship (exponential decay) between income and infant mortality rate.

Log-Transforming Both Predictor and Outcome

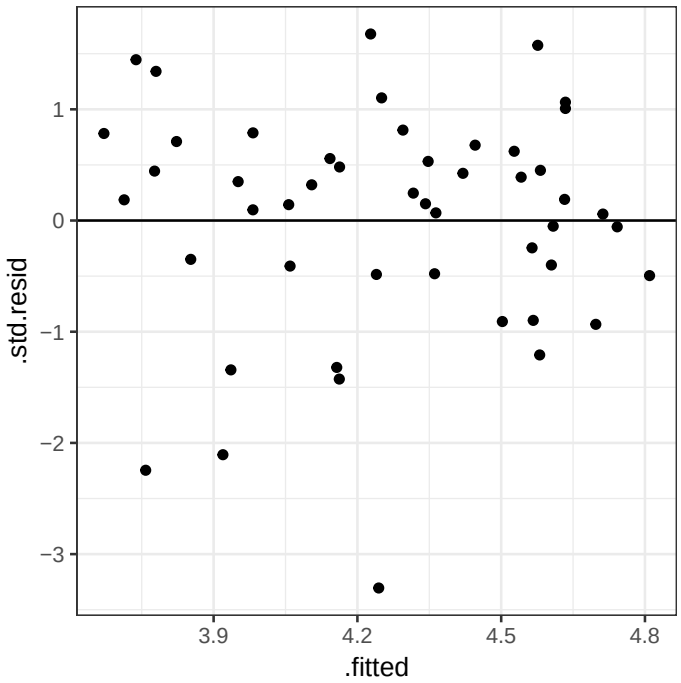
Looking back at the model residuals, you may wonder whether the assumption of heterogeneity of variance is met. We can try to log-transform the outcome to alleviate this.

```r
infant = infant %>%
  mutate(Lmortality = log(mortality))

lm.3 = lm(Lmortality ~ 1 + Lpci, data = infant)

# Obtain residuals
out = augment(lm.3)

# Check linearity assumptions
ggplot(data = out, aes(x = .fitted, y = .std.resid)) +
  geom_point() +
  geom_hline(yintercept = 0) +
  theme_bw()
```



This seems quite reasonable. Now we will look at interpreting the regression coefficients:

```r
summary(lm.3)
```

```
Call:
lm(formula = Lmortality ~ 1 + Lpci, data = infant)

Residuals:
     Min       1Q   Median       3Q      Max
-0.87716 -0.12695  0.04784  0.16416  0.44519

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.48820    0.38814  19.293  < 2e-16 ***
Lpci        -0.43087    0.05167  -8.339 7.98e-11 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 0.2682 on 47 degrees of freedom
Multiple R-squared:  0.5967,    Adjusted R-squared:  0.5881
F-statistic: 69.54 on 1 and 47 DF,  p-value: 7.982e-11
```

Writing the fitted equation,

$$\widehat{\ln(\text{Infant Mortality Rate})} = 7.48 - 0.43\Big[\ln(\text{PCI})\Big]$$

We can interpret the coefficients as we always do, recognizing that these interpretation are based on the log-transformed predictor AND log-transformed outcome.

- The intercept value of 7.48 is the predicted average ln(infant mortality) rate for countries with a ln(PCI) value of 0.
- The slope value of -0.43 indicates that each one-unit difference in ln(PCI) is associated with a -0.43-unit difference in ln(infant mortality), on average.

We can also back-transform to obtain better interpretations.

- The back-transformed intercept value of 1772.2 is the predicted average infant~mortality rate for countries with a PCI value of 1 (extrapolation).
- Each one-percent difference in PCI is associated with a 0.43% decrease in infant mortality rate on average.

You can see this by considering again our three hypothetical countries that have PCIs that differ by one percent.

```
new = data.frame(
  pci = c(1000, 1010, 1020.1)
) %>%
  mutate(
    Lpci = log(pci)
    ) %>%
  mutate(
    Lmortality = predict(lm.3, newdata = .)
    ) %>%
  mutate(mortality = exp(Lmortality))

knitr::kable(new)
```

pci	Lpci	Lmortality	mortality
1000.0	6.907755	4.511841	91.08938
1010.0	6.917706	4.507554	90.69968
1020.1	6.927656	4.503266	90.31166