SOC6707 Intermediate Data Analysis

Monica Alexander

Week 6: Linear Regression II

Announcements

- Assignment 2 and EDA released

Overview

- Explained v unexplained variation
- Hypothesis testing of coefficients
- Log transforms

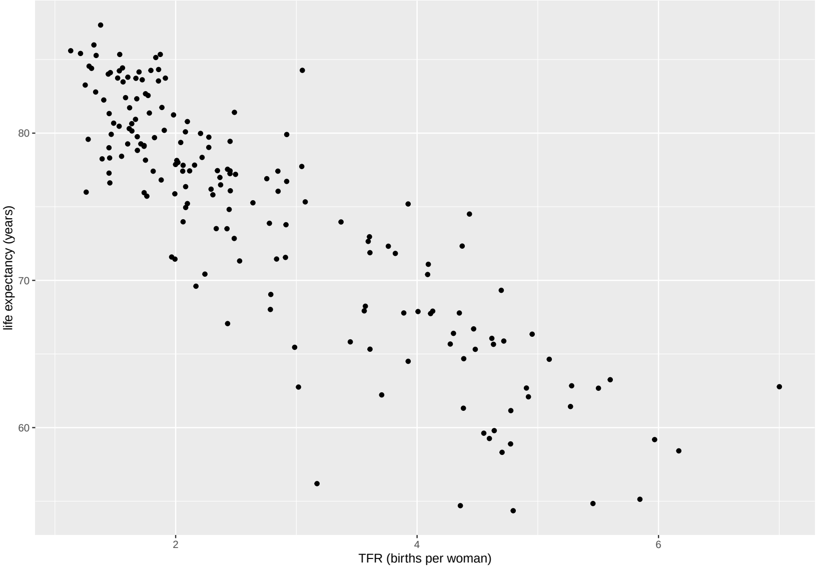
Review of SLR set-up

- Y_i is the response variable, and X_i is the explanatory variable

Example:

- Research question: In 2017, how does the expected value of life expectancy differ or change across countries with different levels of fertility?
- In other words, is life expectancy associated with fertility, and if so, how?

Scatter plot



Fit SLR in R

```r
country_ind_2017 <- country_ind %>% filter(year==2017)
slr_mod <- lm(life_expectancy~tfr, data = country_ind_2017)
summary(slr_mod)
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0718  -2.3864   0.3132   2.6537  11.3498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.2394     0.7085  125.95   <2e-16 ***
## tfr          -5.3526     0.2326  -23.02   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.994 on 174 degrees of freedom
## Multiple R-squared:  0.7527, Adjusted R-squared:  0.7513
## F-statistic: 529.7 on 1 and 174 DF,  p-value: < 2.2e-16
```

How much variation does our model explain: R^2

Thinking about variation

- So far we've been mostly concerned about conditional expectations, that is, population means for different subgroups/populations of different characteristics
- Let's think about variation in Y_i around measures of central tendency for a moment

What sorts of variation may we be interested in?

- Variation of data Y_i around the observed mean $\bar{Y}_i$
- Variation of fitted values $\hat{Y}_i$ around observed mean $\bar{Y}_i$
- Variation of data Y_i around fitted values $\hat{Y}_i$

Sums of squares

- Variation of data Y_i around the observed mean $\bar{Y}_i$
  - Total sum of squares SST: $(Y_i - \bar{Y}_i)^2$
- Variation of fitted values $\hat{Y}_i$ around observed mean $\bar{Y}_i$
  - Model sum of squares SSM: $(\hat{Y}_i - \bar{Y}_i)^2$
- Variation of data Y_i around fitted values $\hat{Y}_i$
  - Residual sum of squares SSR: $(Y_i - \hat{Y}_i)^2$

Sums of squares

- Variation of data Y_i around the observed mean $\bar{Y}_i$
  - Total sum of squares SST: $(Y_i - \bar{Y}_i)^2$
  - Total variation in Y_i
- Variation of fitted values $\hat{Y}_i$ around observed mean $\bar{Y}_i$
  - Model sum of squares SSM: $(\hat{Y}_i - \bar{Y}_i)^2$
  - Variation explained by our X's
- Variation of data Y_i around fitted values $\hat{Y}_i$
  - Residual sum of squares SSR: $(Y_i - \hat{Y}_i)^2$
  - Variation not explained by X's

$$SST = SSM + SSR$$

R^2

$$SST = SSM + SSR$$

$$R^2 = \frac{SSM}{SST} = 1 - \frac{SSR}{SST}$$

The proportion of total variation in Y_i explained by covariates X_i.

Hypothesis testing

Fit SLR in R

```
country_ind_2017 <- country_ind %>% filter(year==2017)
slr_mod <- lm(life_expectancy~tfr, data = country_ind_2017)
summary(slr_mod)
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0718  -2.3864   0.3132   2.6537  11.3498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.2394     0.7085  125.95   <2e-16 ***
## tfr          -5.3526     0.2326  -23.02   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.994 on 174 degrees of freedom
## Multiple R-squared:  0.7527, Adjusted R-squared:  0.7513
## F-statistic: 529.7 on 1 and 174 DF,  p-value: < 2.2e-16
```

SLR fit

- The estimate of $\hat{\beta}_1$ tells us there's a negative association between TFR and life expectancy as estimated from the data
- But how sure of this are we? It's not a perfect relationship, and there is some noise
- It was reasonably clear from our scatterplot, but what if our scatter plot had looked different?

Intuition of hypothesis testing

- We are assuming there's some underlying β_1 that we're trying to find (this assumes the truth is a linear relationship)
- We get an estimate of β_1 (called $\hat{\beta_1}$) based on data we collect
- But this estimate could be right, almost right, or completely wrong compared to the truth
- In regression we are usually interested in deciding whether we believe β_1 is non-zero (i.e. there is a linear association between our two variables)
- The degree to which we believe this depends on what the data look like

Intuition of hypothesis testing

- If the data look a lot like a linear relationship, then we conclude that there's enough evidence to suggest a non-zero relationship and that our estimate is probably right
- The more randomness there is in the data, the less likely we are to believe our estimate is the truth
- Hypothesis testing (based on t-tests) is a way of accounting for this uncertainty and making inferences about the relationships between variables

Intuition of hypothesis testing

How do we account for the uncertainty in the data before making decisions about whether β_1 is zero or not?

- The regression model has a bunch (five) of assumptions underlying it
- If we assume these are true, then it turns out we know what the probability distribution of possible values of $\hat{\beta}_1$ look like
- If a lot of probability density in this distribution is near zero (read: if zero is likely), then we would conclude there's not enough evidence to suggest a linear relationship
- And vice versa

To dos

To do:

- Learn assumptions
- Write down distribution for $\hat{\beta}_1$
- Do hypothesis testing
- Celebrate, eat cake, graduate

The MLR assumptions

The five assumptions of multiple linear regression:

1. no model misspecification
2. there is independent variation in all of the explanatory variables
   - In other words, none of the explanatory variables are constants, and there are no perfect linear relationships among the explanatory variables
   - e.g. can't have $X_{i1} = X_{i2} + X_{i3}$
3. All variables are from a simple random sample
   - This assumption implies that all members of a population have an equal probability of selection, that all possible samples of size n have an equal probability of selection, and that each observation is independent of all the others

The MLR assumptions

4. The variance of $\varepsilon_i = Y_i - E\left(Y_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}\right)$ is the same across all values of the explanatory variables i.e. $\text{Var}\left(\varepsilon_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}\right) = \sigma^2$
   - This is called homoskedasticity
5. The normality assumption $\varepsilon_i = Y_i - E\left(Y_i \mid X_{i1}, X_{i2}, \ldots, X_{ik}\right)$ is normally distributed

Assume a spherical elephant

- If we take the five assumptions above as given, it turns out that the distribution of possible values of our estimate $\hat{\beta}_1$ around the true value β_1 is knowm
- In particular, we are going to look at a transformed version of $\hat{\beta}_1$:

$$\frac{\widehat{\beta}_1 - \beta_1}{se\left(\widehat{\beta}_1\right)}$$

  where $se\left(\widehat{\beta}_1\right)$ is the standard error of $\hat{\beta}_1$.
- This should look vaguely familiar, from when we calculated Z-scores.

The t-statistic

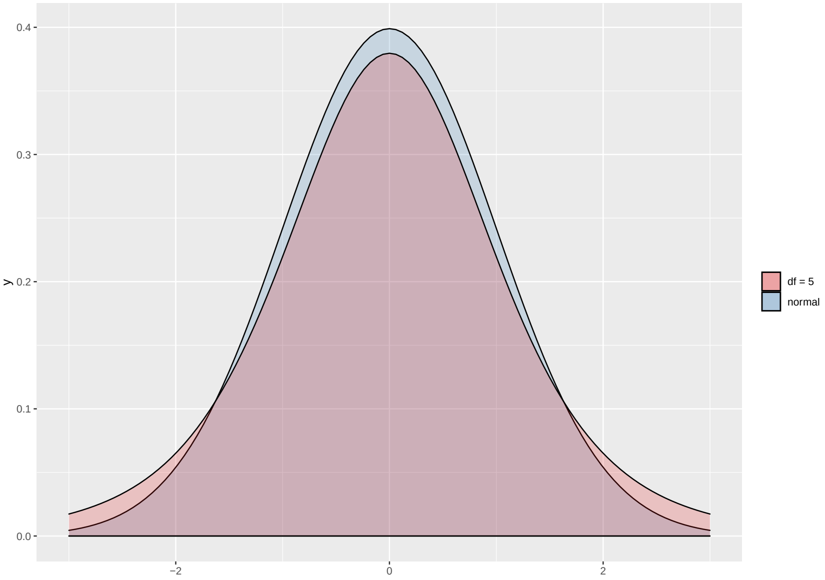
Let's give this quantity a name:

$$T_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - \beta_1}{se\left(\widehat{\beta}_1\right)}$$

Given the five assumptions discussed, this follows a t-distribution with $n-(k+1)$ degrees of freedom.

- The t-distribution looks similar to the standard normal distribution, but has 'heavier tails' when $df < 120$ (i.e. there's more probability mass further away from the mean)
- for $df \geq 120$ the t-distribution converges to a standard normal distribution.

The t-distribution



What is the standard error?

The standard error of $\hat{\beta}_1$, is

$$\text{se}\left(\hat{\beta}_1\right) = \sqrt{\frac{\widehat{\sigma}^2}{\sum_i \left(X_{i1} - \bar{X}_{i1}\right)^2 (1 - R_1^2)}}$$

where

$$\hat{\sigma}^2 = \frac{\sum_i \hat{\varepsilon}_i^2}{n - (k+1)} = \frac{SSR}{df}$$

and R_1^2 is the R^2 from a regression of X_1 against all other variables in the model

What is the standard error?

Don't try and remember the formulas from the previous slide. Just remember that the standard error of $\hat{\beta}_1$ is proportional to the sum of squares of residuals.

- a larger error variance (i.e., greater unexplained variation in the outcome) is associated with a larger se$\left(\hat{\beta}_1\right)$ and vice versa

What does the standard error do to the distribution of $T_{\widehat{\beta}_1}$?

Hypothesis testing: more intuition

- In regression, we are interested to see if there's evidence to suggest that β_1 is different enough from zero.
- Pretend for a moment that the true value of β_1 is zero. In this world (the null hypothesis world), our $T_{\widehat{\beta}_1}$ is just

$$\frac{\widehat{\beta}_1}{se\left(\widehat{\beta}_1\right)}$$

- In the null hypothesis world, this thing should be t-distributed (i.e. centered at zero with some variation around that)
- So if we calculate this thing and it's really different from zero (i.e. where the distribution is centered), then it's unlikely it came from this distribution, and we can probably reject the world in which β_1 is zero
- If this thing is not very different from zero, then we may not reject this world

Hypothesis testing: more intuition

- We are dealing with randomness, and so there's always a chance that the value we see is from the null hypothesis world in which β_1 is zero
- But the farther away it is from zero, the less likely that's true
- The size of $T_{\widehat{\beta}_1}$ depends not only on the magnitude of $\widehat{\beta}_1$ but also the magnitude of the standard error of $\widehat{\beta}_1$
- So the stronger the relationship (the bigger the $\widehat{\beta}_1$) the less likely we are going to believe the null hypothesis
- But also for less noisy data (the smaller the standard error) the less likely we are going to believe the null hypothesis

Hypothesis testing: more formal language

Say we run an SLR.

- The slope coefficient β_1 is an unknown population quantity, which we have estimated with data from a random sample of that population
- We can test hypotheses about this unknown population quantity based on the fact that the $T_{\widehat{\beta_1}}$ follows a t-distribution with $n-2$ degrees of freedom
- With knowledge of the probability distribution of $T_{\widehat{\beta_1}}$ we can make probabilistic statements about the chances of observing any particular value of $T_{\widehat{\beta_1}}$ given a hypothesized value for the unknown parameter
- In particular, we are often interested in testing to see whether there is evidence to suggest that $\beta_1 \neq 0$ i.e. the slope coefficient is not zero i.e. there is evidence of a relationship between our dependent and independent variable

The t-test steps

To test hypotheses about the value of β_1, we use a t-test (as the SE-standardized estimate follows a t-distribution). The steps of a t-test are:

1. State your null and alternative hypotheses about β_1

- The null hypothesis is denoted H_0
- The alternative hypothesis is denoted H_1
- e.g. $H_0 : \beta_1 = b$ and $H_1 : \beta_1 \neq b$

2. Choose the level of type-I error, α, which gives the probability of rejecting the null hypothesis when it is actually true

- For example, α is most commonly chosen to be 0.05 i.e. the type-I error rate is 5%

The t-test steps (ctd)

3. Compute the t-test statistic

$$t_{\widehat{\beta}_1} = \frac{\left(\widehat{\beta}_1 - b\right)}{\mathsf{se}\left(\widehat{\beta}_1\right)}$$

4. Compute the p-value, which gives the probability of observing a test statistic as or even more extreme than $t_{\widehat{\beta}_1}$ under the assumption that the null hypothesis is true
5. Make a decision (reject the null if the p-value is less than α, and fail to reject otherwise)

The t-test in R

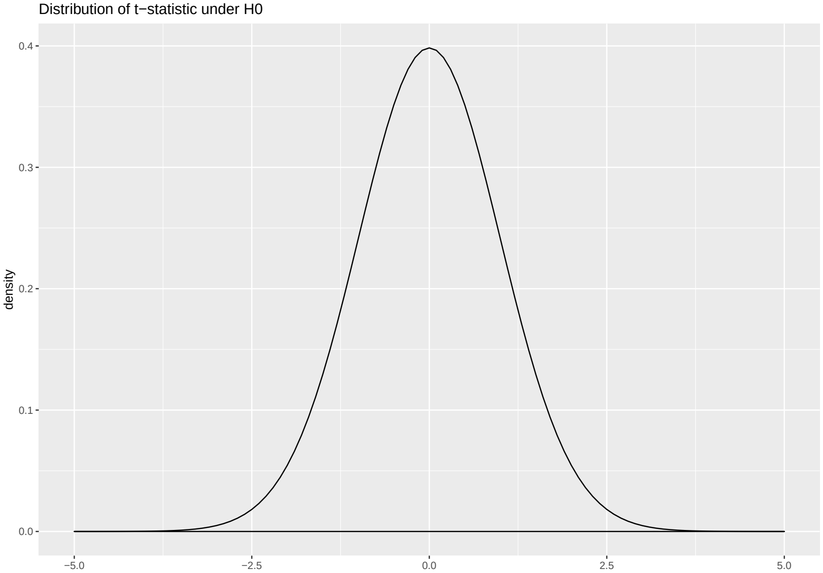
The `lm` summary put put shows the calculations for $t_{\widehat{\beta}_1}$ and corresponding p-value. Specifically these calculations test whether $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$.

```
slr_mod <- lm(life_expectancy~tfr, data = country_ind_2017)
summary(slr_mod)
```

```
##
## Call:
## lm(formula = life_expectancy ~ tfr, data = country_ind_2017)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.0718  -2.3864   0.3132   2.6537  11.3498
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  89.2394     0.7085  125.95   <2e-16 ***
## tfr          -5.3526     0.2326  -23.02   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 3.994 on 174 degrees of freedom
## Multiple R-squared:  0.7527, Adjusted R-squared:  0.7513
## F-statistic: 529.7 on 1 and 174 DF,  p-value: < 2.2e-16
```

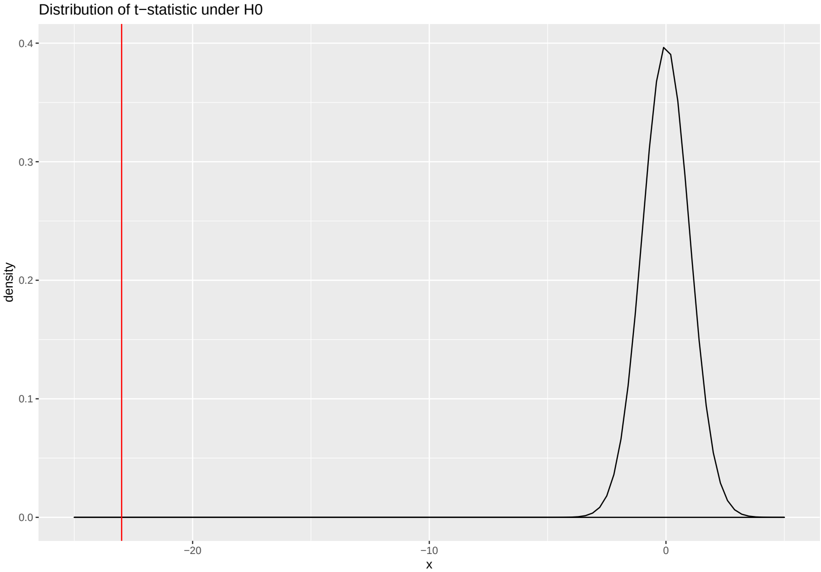
What should we conclude?

Logic of the t-test



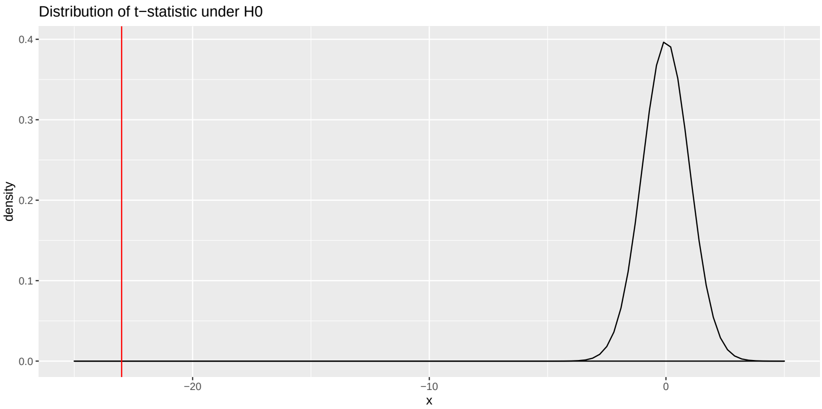
Logic of the t-test

We calculated $t_{\widehat{\beta_1}} = -23$



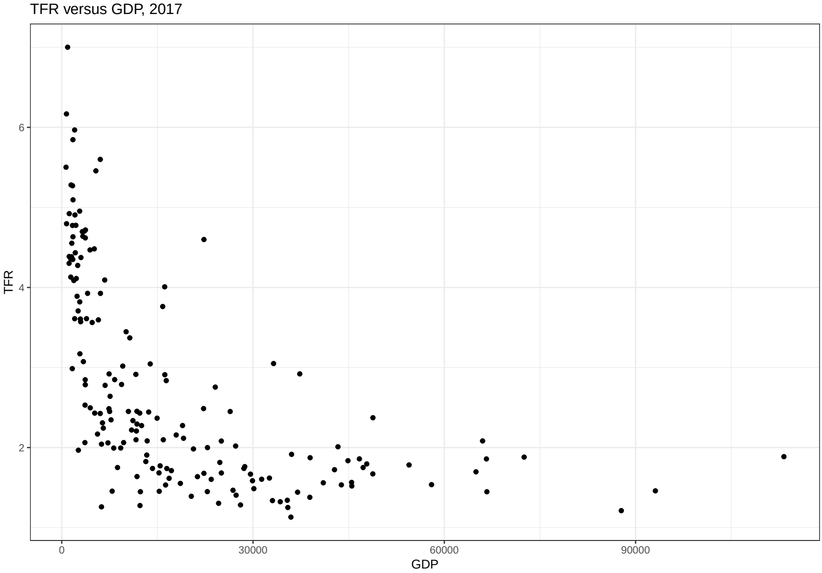
Logic of the t-test

- We calculated $t_{\widehat{\beta}_1} = -23$
- Under the null hypothesis, the probability of observing this value is very small—thus, we conclude the null hypothesis is likely false

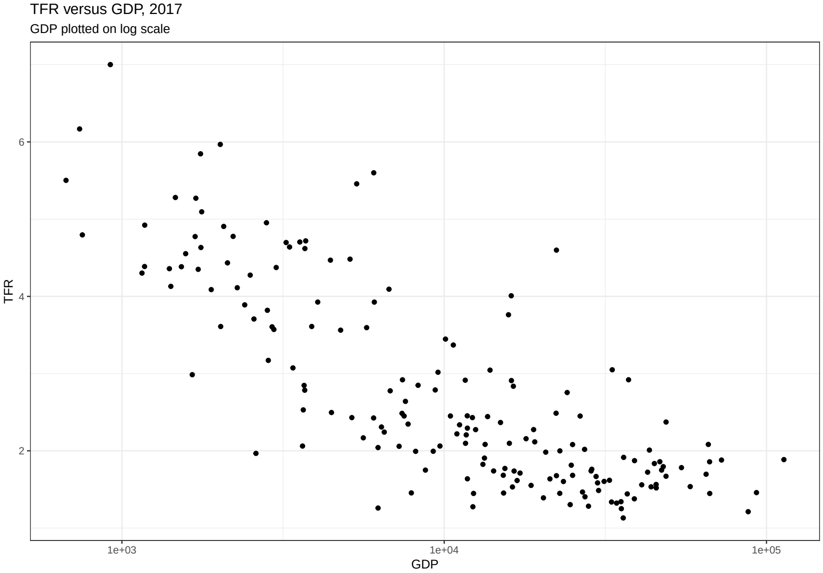


Regression with transformed variables

Motivation



Motivation



TFR versus GDP, 2017
GDP plotted on log scale
TFR
6
4
2
1e+03
1e+04
1e+05
GDP

Variable transformations

- Sometimes we may want to allow for nonlinearities in our models
- A common way to deal with this is to perform a nonlinear transformation on one or more of the explanatory variables **AND/OR** on the response variable
- The interpretation of parameter estimates is less intuitive after transforming the explanatory variables and/or the response variable, although some transformations lend themselves to simple interpretations (i.e., the log transform)

Log transforms

- By far the most common transformation is the natural log transform
- Either $\log Y$ or $\log X$ (or both)
- Luckily, the log transform has a meaningful coefficient interpretation

We will look at

- $logY_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$
- $Y_i = \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$
- $logY_i = \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$

Log transforms: response variable

For response variables, when the model is

$$\log Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i$$

A one unit change in X_{ik} leads to a $(\exp(\beta_k) - 1)\,100$ percent change in Y_i, on average, holding other factors constant.

Log transforms: response variable (SHOW WHY)

Response variable: approximation

It turns out that $\exp(z) \approx 1 + z$ for small values of z.

So an approximate interpretation is

$$100\beta_k \left(\Delta X_{ik}\right) = \%\Delta Y_i$$

where Δ stands for "change".

- Thus, a one unit increase in X_k is associated with a $100 \cdot \beta_k\%$ change in Y_i, on average, holding other factors constant

Log transforms: expanatory variables

For explanatory variables, when the model is

$$\begin{aligned} Y_i &= E\left(Y_i \mid \log X_{i1}, X_{i2}, \ldots, X_{ik}\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i \end{aligned}$$

The interpretation is

$$\frac{\beta_k}{100}\left(\%\Delta X_{ik}\right) = \Delta Y_i$$

where Δ stands for “change”.

- Thus, a one percent (1%) increase in X_k is associated with a $\frac{\beta_k}{100}$ unit change in Y_i, on average, holding other factors constant

Log transforms: both variables

When both the response and explanatory variable is transformed, so the model is

$$\begin{aligned}\log Y_i &= E\left(Y_i \mid \log X_{i1}, X_{i2}, \ldots, X_{ik}\right) + \varepsilon_i \\ &= \beta_0 + \beta_1 \log X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varepsilon_i\end{aligned}$$

We are going to utilize the first approximation here, and say

The interpretation is

$$\beta_k \left(\%\Delta X_{ik}\right) = \%\Delta Y_i$$

- Thus, a one percent (1%) increase in X_k is associated with a β_k % change in Y_i, on average, holding other factors constant

Example

```
country_ind <- country_ind %>%
  mutate(log_tfr = log(tfr)) # log of GDP

summary(lm(log_tfr ~ child_mort + gdp, data = country_ind))
```

```
##
## Call:
## lm(formula = log_tfr ~ child_mort + gdp, data = country_ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66609 -0.18599  0.00086  0.15314  0.64842
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  6.497e-01  1.337e-02  48.599   <2e-16 ***
## child_mort   1.021e-02  2.018e-04  50.586   <2e-16 ***
## gdp         -3.453e-06  3.749e-07  -9.211   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.2372 on 1581 degrees of freedom
## Multiple R-squared:  0.7396, Adjusted R-squared:  0.7393
## F-statistic:  2246 on 2 and 1581 DF,  p-value: < 2.2e-16
```

- A 10ˆ5 unit increase in GDP is associated with a 30% decrease in TFR, holding child mortality constant

Example

```r
country_ind <- country_ind %>%
  mutate(log_gdp = log(gdp)) # log of GDP

summary(lm(tfr ~ child_mort + log_gdp, data = country_ind))
```

```
##
## Call:
## lm(formula = tfr ~ child_mort + log_gdp, data = country_ind)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.0606 -0.3750 -0.0369  0.3388  2.0084
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.5468550  0.2143498   21.21   <2e-16 ***
## child_mort   0.0278231  0.0007136   38.99   <2e-16 ***
## log_gdp     -0.2882433  0.0211493  -13.63   <2e-16 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 0.6295 on 1581 degrees of freedom
## Multiple R-squared:  0.802,  Adjusted R-squared:  0.8018
## F-statistic:  3202 on 2 and 1581 DF,  p-value: < 2.2e-16
```

- A 1% increase in GDP is associated with a decrease of 0.003 children in TFR, holding child mortality constant

Example

```
summary(lm(log_tfr ~ child_mort + log_gdp, data = country_ind))
```

```
##
## Call:
## lm(formula = log_tfr ~ child_mort + log_gdp, data = country_ind)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.66781 -0.16460  0.00366  0.15027  0.58812
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.7755424  0.0769391   23.08   <2e-16 ***
## child_mort   0.0080449  0.0002561   31.41   <2e-16 ***
## log_gdp     -0.1211787  0.0075914  -15.96   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2259 on 1581 degrees of freedom
## Multiple R-squared:  0.7637, Adjusted R-squared:  0.7634
## F-statistic:  2555 on 2 and 1581 DF,  p-value: < 2.2e-16
```

- A 1% increase in GDP is associated with a 0.12% decrease in TFR, holding child mortality constant
- A 10% increase in GDP is associated with a 1.2% decrease in TFR, holding child mortality constant

Summary

- Often we may want to transform dependent or independent variables to make relationships more linear
- Log transforms are by far the most common
- This is because many variables are naturally log-normally distributed, e.g. income and GDP