SOC6707 Intermediate Data Analysis

Monica Alexander

Week 4: Linear Regression II

Overview

- Hypothesis testing of coefficients
- Confidence intervals
- 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?

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
```

```
##          1          2          3          4          5          6
##  1.2147154 -0.3253063  4.7935973  3.9875228 -0.6667741  2.6639127
```

```
##        1        2        3        4        5        6
## 64.44128 80.47331 72.94140 59.26448 78.53777 77.06209
```

Sampling distribution of SE-standardized $\hat{\beta}_1$

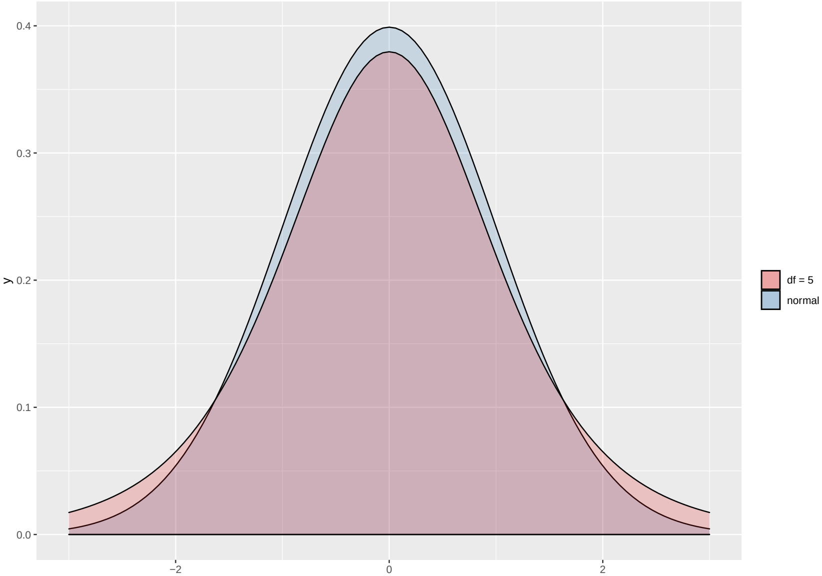
Under the five assumption discussed, the SE-standardized $\hat{\beta}_k$

$$T_{\widehat{\beta}_k} = \frac{\widehat{\beta}_k - \beta_k}{se\left(\widehat{\beta}_k\right)}$$

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



Hypothesis testing

Hypothesis testing

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 SE-standardized estimate 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

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

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)

Logic of the t-test

- Under the 5 assumptions discussed earlier, if the null hypothesis that $\beta_1 = b$ were in fact true, then $T_{\widehat{\beta}_1} = \frac{\widehat{\beta}_1 - b}{se\left(\widehat{\beta}_1\right)}$ would be t-distributed with $n-2$ df.
- We can use this result to make probabilistic statements about the chances of observing different values of $T_{\widehat{\beta}_1}$ in any given sample
- If the probability of observing a test statistic as or even more extreme than the value we actually observe in our sample is very small, then we conclude that the null hypothesis is not likely true

The t-test in R

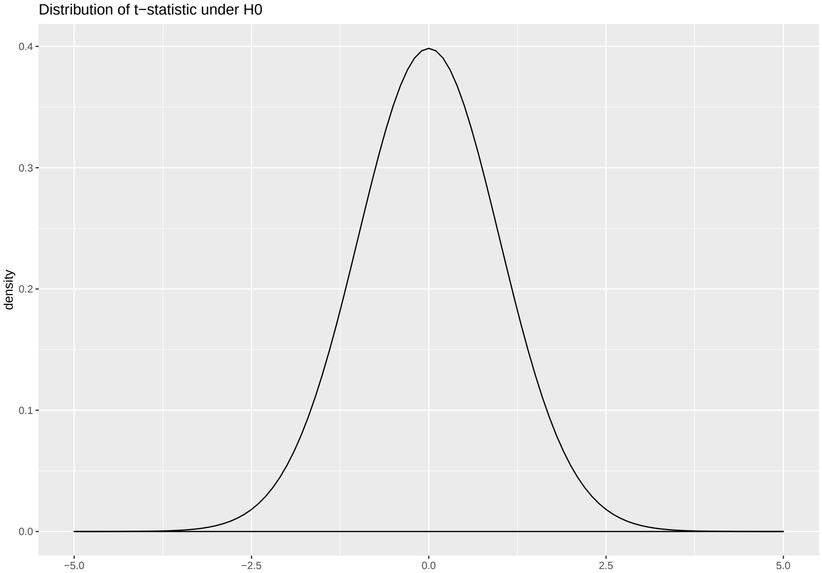
The `lm` summary put put shows the calculations for $t_{\widehat{\beta}_1}$ and corresponding p-value. Specifically these calculations test whether $H_0 : \beta_1 = 0$ and $H_1 : \beta_1 \neq 0$.

```
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
```

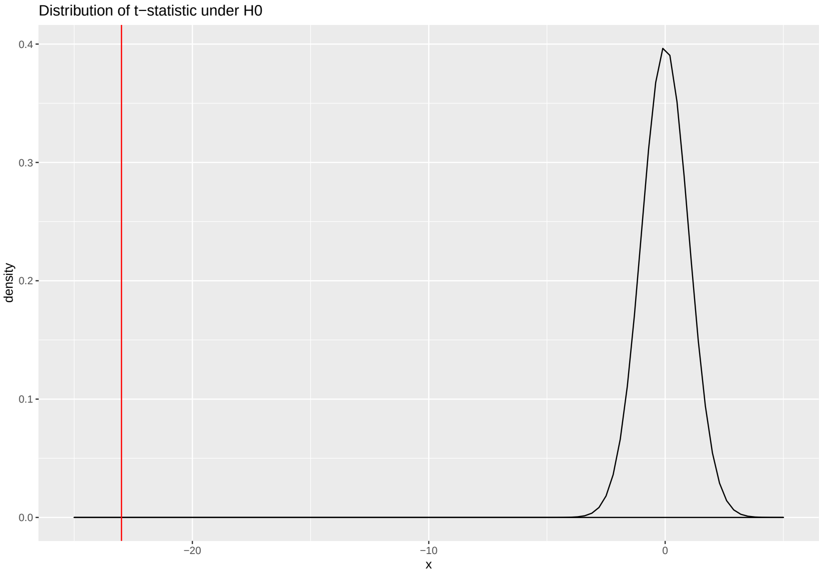
What should we conclude?

Logic of the t-test



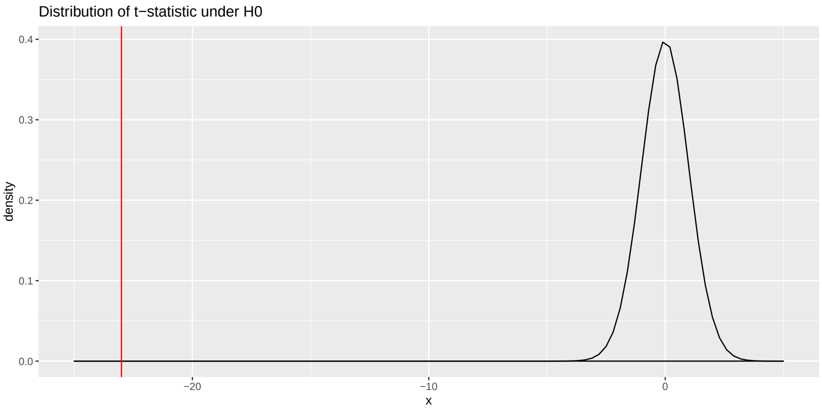
Logic of the t-test

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



Confidence intervals

Interval estimation

- So far, we have focused on point estimation of regression parameters, which involves assigning a single value to these parameters that minimizes the sum of squared residuals
- Interval estimation refers to computing confidence intervals for parameters, which provide a range of values that contain the true value of the parameter with known probability in repeated sampling

Interval estimation steps

1. Choose your confidence level (i.e., the probability that the interval estimate will cover the parameter of interest in repeated sampling)

- Usually choose $\nu = 1 - \alpha$ with $\alpha = 0.05$ so the confidence level is $\nu = 0.95$ or 95%

2. find the critical value, t_α, of the t-distribution with $n - (k+1)$ degrees of freedom for which $P(|T| > t_\alpha) = 1 - \nu = \alpha$

- In words, the probability of the absolute value of our T statistic of interest being greater than the critical value (i.e. outside the bounds defined by t_α) is α (e.g. 0.05 or 5%)

Interval estimation steps (ctd)

3. Compute the limits of the confidence interval

- Lower limit: $\hat{\beta}_1 - \left(t_\alpha \times se\left(\hat{\beta}_1\right)\right)$
- Upper limit: $\hat{\beta}_1 + \left(t_\alpha \times se\left(\hat{\beta}_1\right)\right)$

4. Interpret.

- if random samples were repeatedly collected and confidence intervals were computed as outlined above for each sample, the true value of the parameter, β_1, would lie in the confidence interval in $\nu \times 100$ percent of the samples

Confidence intervals in R

```
# extract beta1 hat and se
summary(slr_mod)$coefficients
```

```
##              Estimate Std. Error   t value      Pr(>|t|)
## (Intercept) 89.239400  0.7085056 125.95440 7.276023e-173
## tfr         -5.352575  0.2325733 -23.01458  1.118294e-54
```

```
b1_hat <- summary(slr_mod)$coefficients[2,1]
se_b1_hat <- summary(slr_mod)$coefficients[2,2]

# choose a confidence level
alpha <- 0.05
v <- 1-alpha
n <- nrow(country_ind_2017)

# calculate critical value
t_alpha <- abs(qt(p = alpha/2, df = n-2))
t_alpha
```

```
## [1] 1.973691
```

```
# calculate confidence interval

# lower
b1_hat - t_alpha*se_b1_hat
```

```
## [1] -5.811603
```

```
# upper
b1_hat + t_alpha*se_b1_hat
```

```
## [1] -4.893547
```

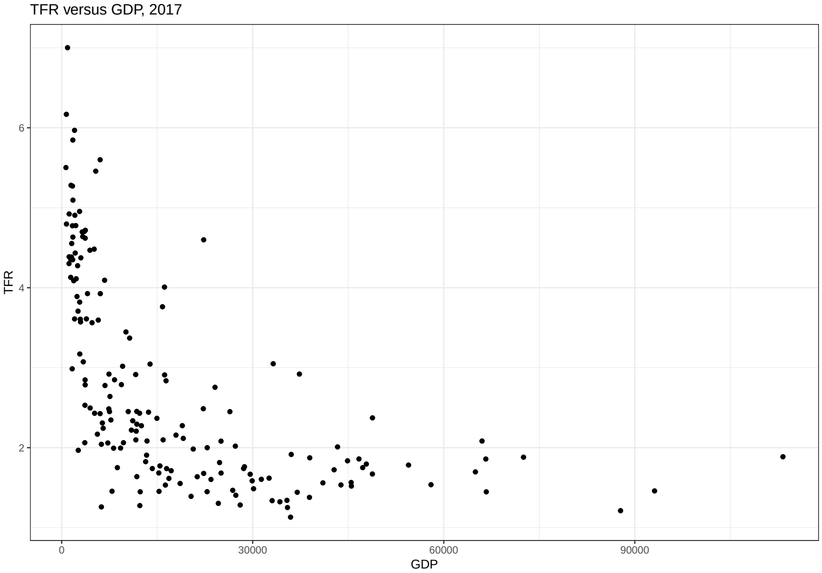
Diagram to explain critical value

Summary

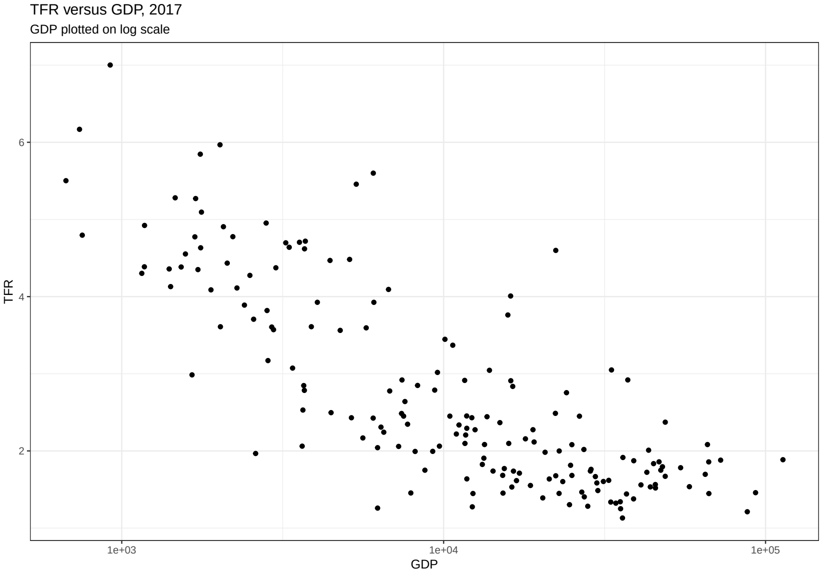
- Under a set of assumptions, the SE-standardized estimator $\hat{\beta}_k$ is t-distributed
- We can use this information to test null hypotheses about whether or not the coefficients are zero, and to create confidence intervals of the likely range of values of β_k
- Note that a t-test of the null hypothesis that the coefficient in an MLR model is zero is a test of statistical independence between the dependent and the independent variable

Regression with transformed variables

Motivation



Motivation

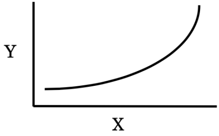
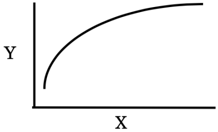


TFR versus GDP, 2017
GDP plotted on log scale
TFR
6
4
2
1e+03
1e+04
1e+05
GDP

Variable transformations

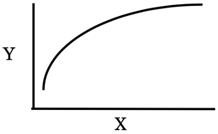
- Sometimes we may want to allow for nonlinearities in our models
- A common way to deal with this is to perform a nonlinear transformation on one or more of the explanatory variables **AND/OR** on the response variable
- The interpretation of parameter estimates is less intuitive after transforming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

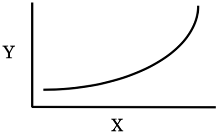
Response variable

Transformation	Name	Type of Nonlinear Relationship
Y^(1/3)	cube root	
Y^(1/2)	sqaure root	
log(Y)	natural logrithm	
Y^3	cubic	
Y^2	quadratic	
exp(Y)	exponentional	

Explanatory variable

Transformation	Name	Type of Nonlinear Relationship
X^(1/3)	cube root	
X^(1/2)	sqaure root	
log(X)	natural logrithm	
X^3	cubic	
X^2	quadratic	
exp(X)	exponentional	





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

For response variables, when the model is

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

The 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

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

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

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

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

```r
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 10^5 unit increase in GDP is associated with a 30% decrease in TFR, holding child mortality constant

Example

```
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

```r
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