Precept 5: Simple OLS

Soc 500: Applied Social Statistics

Shay O'Brien[1]

Princeton University

October 11, 2018

[1]These slides draw on material from Ziyao Tian, Simone Zhang and Matt Blackwell.

Today's Agenda

- Chit chat
  - Review of common pset3 issues
  - How was pset4?
- Slides
  - OLS mechanics and assumptions
  - Hypothesis tests meet regression
  - Residuals and friends
  - Lemma 1: Why does everyone keep logging stuff??
- RStudio
  - Lemma 2: Lists
  - Regression in R

Population and Sample Linear Regression Function

- The population simple linear regression model can be stated as the following:

$$r(x) = E[Y|X = x] = \beta_0 + \beta_1 x$$

- $\beta_0, \beta_1 =$ population intercept and population slope (what we want to estimate)

Population and Sample Linear Regression Function

- The population simple linear regression model can be stated as the following:

$$r(x) = E[Y|X = x] = \beta_0 + \beta_1 x$$

- β_0, β_1 = population intercept and population slope (what we want to estimate)
- The **estimated** or sample regression function is:

$$\widehat{r}(X_i) = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

- $\widehat{\beta}_0, \widehat{\beta}_1$ are the estimated intercept and slope

Population and Sample Linear Regression Function

- The population simple linear regression model can be stated as the following:

$$r(x) = E[Y|X = x] = \beta_0 + \beta_1 x$$

- $\beta_0, \beta_1 =$ population intercept and population slope (what we want to estimate)
- The **estimated** or sample regression function is:

$$\widehat{r}(X_i) = \widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

- $\widehat{\beta}_0, \widehat{\beta}_1$ are the estimated intercept and slope
- the **Ordinary Least Squares** (OLS) estimates are the intercept and slope that minimize the sum of the squared residuals:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \underset{b_0, b_1}{\arg\min} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

Equations for OLS $\hat{\beta}_0$ and $\hat{\beta}_1$

$$\widehat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \\
&= \frac{\frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{n}}{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n}} \\
&= \frac{E[(X - \overline{X})(Y - \overline{Y})]}{E[(X - \overline{X})^2]} \\
&= \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}
\end{aligned}$$

OLS slope as the sum of a random variable

$$\begin{aligned}
\widehat{\beta}_1 &= \frac{\sum_{i=1}^n (X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} - \frac{\sum_{i=1}^n (X_i - \overline{X})\overline{Y}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \sum_{i=1}^n W_i Y_i
\end{aligned}$$

Where here we have the weights, W_i as:

$$W_i = \frac{(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$

Sampling Distributions of Random Variables $\hat{\beta}_1$ & $\hat{\beta}_0$

- $\hat{\beta}_1$ can be seen as **the sum of a RV** (which is also a RV), or a weighted sum of the outcomes.
- Seeing $\hat{\beta}_1$ & $\hat{\beta}_0$ as RV, we want to know their sampling distributions. How?

Sampling Distributions of Random Variables $\hat{\beta}_1$ & $\hat{\beta}_0$

- $\hat{\beta}_1$ can be seen as **the sum of a RV** (which is also a RV), or a weighted sum of the outcomes.
- Seeing $\hat{\beta}_1$ & $\hat{\beta}_0$ as RV, we want to know their sampling distributions. How?
- We need assumptions to learn about their sampling distributions. In other worsd, under what conditions will they look like ...???

OLS Assumptions

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.

OLS Assumptions

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.
2. **Random Sampling**: The observed data represent a random sample from the population described by the model.

OLS Assumptions

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.
2. **Random Sampling**: The observed data represent a random sample from the population described by the model.
3. **Variation in X**: There is variation in the explanatory variable.

OLS Assumptions

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.
2. **Random Sampling**: The observed data represent a random sample from the population described by the model.
3. **Variation in X**: There is variation in the explanatory variable.
4. **Zero conditional mean**: Expected value of the error term is zero conditional on all values of the explanatory variable.

OLS Assumptions

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.
2. **Random Sampling**: The observed data represent a random sample from the population described by the model.
3. **Variation in X**: There is variation in the explanatory variable.
4. **Zero conditional mean**: Expected value of the error term is zero conditional on all values of the explanatory variable.
5. **Homoskedasticity**: The error term has the same variance conditional on all values of the explanatory variable.

OLS Assumptions

1. **Linearity in Parameters**: The population model is linear in its parameters and correctly specified.
2. **Random Sampling**: The observed data represent a random sample from the population described by the model.
3. **Variation in X**: There is variation in the explanatory variable.
4. **Zero conditional mean**: Expected value of the error term is zero conditional on all values of the explanatory variable.
5. **Homoskedasticity**: The error term has the same variance conditional on all values of the explanatory variable.
6. **Normality**: The error term is independent of the explanatory variables and normally distributed.

Assumptions and Sampling Distribution

- Under Assumptions 1-6, we know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) \text{ or } \frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

Assumptions and Sampling Distribution

- Under Assumptions 1-6, we know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) \text{ or } \frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

- Under Assumptions 1-6 and in any sample, we know that

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-2}$$

Assumptions and Sampling Distribution

- Under Assumptions 1-6, we know that

$$\widehat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_u^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right) \text{ or } \frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

- Under Assumptions 1-6 and in any sample, we know that

$$\frac{\widehat{\beta}_1 - \beta_1}{\widehat{SE}[\widehat{\beta}_1]} \sim t_{n-2}$$

- Under Assumptions 1-5 and in large samples, we know that

$$\frac{\widehat{\beta}_1 - \beta_1}{SE[\widehat{\beta}_1]} \sim N(0,1)$$

Three ways of making statistical inference out of regression

1. **Point Estimation**: Consider the sampling distribution of our point estimator $\hat{\beta}_1$ to infer β_1

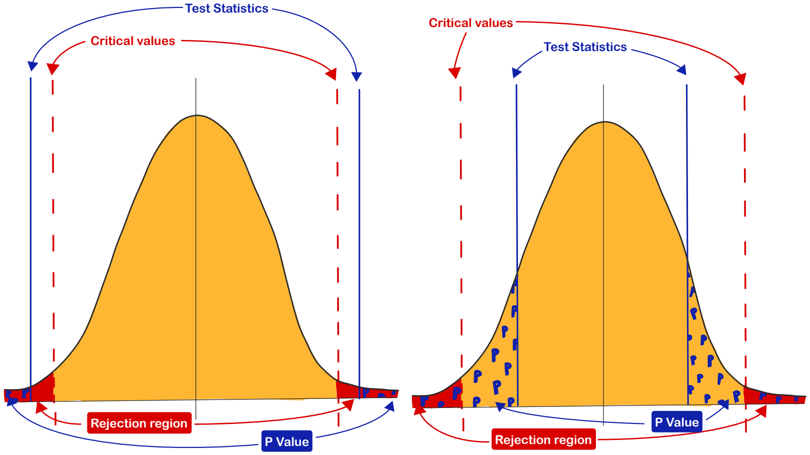
Three ways of making statistical inference out of regression

1. **Point Estimation**: Consider the sampling distribution of our point estimator $\hat{\beta}_1$ to infer β_1
2. **Hypothesis Testing**: Consider the sampling distribution of a test statistic to test hypothesis about β_1 at the α level

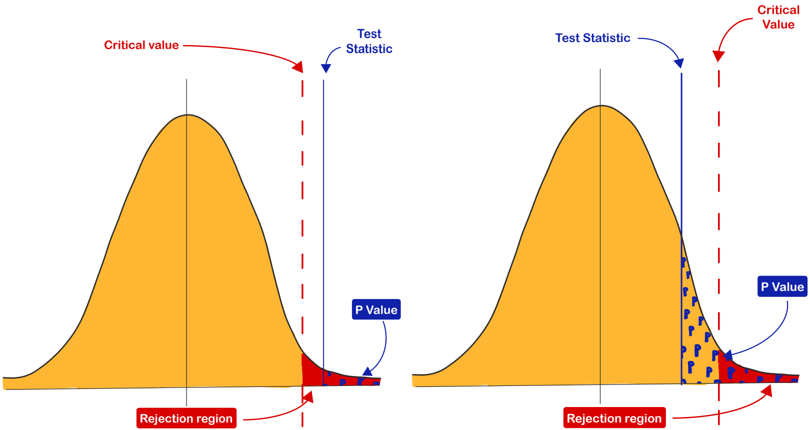
Three ways of making statistical inference out of regression

1. **Point Estimation**: Consider the sampling distribution of our point estimator $\hat{\beta}_1$ to infer β_1
2. **Hypothesis Testing**: Consider the sampling distribution of a test statistic to test hypothesis about β_1 at the α level
3. **Interval Estimation**: Consider the sampling distribution of an interval estimator to construct intervals that will contain β_1 at least $100(1-\alpha)\%$ of the time.

Two-sided hypothesis test



One-sided hypothesis test



You can reject the null hypothesis if:

P-value < Alpha

|Test statistic| > |Critical value|

Otherwise, you have to retain the null*.

*But it's still a hypothesis! You haven't proved that it's true!

You can reject the null hypothesis if:

P-value < Alpha

|Test statistic| > |Critical value|

If one of these comparison statements is true/false,
so is the other and vice versa.

Otherwise, you have to retain the null*.

***But it's still a hypothesis! You haven't proved that it's true!**

You can reject the null hypothesis if:

Translate your estimate into the null distribution		*Translate your certainty level into the null distribution*
P-value	<	Alpha
|Test statistic|	>	|Critical value|

If one of these comparison statements is true/false, so is the other and vice versa.

Otherwise, you have to retain the null*.

***But it's still a hypothesis! You haven't proved that it's true!**

You can reject the null hypothesis if:

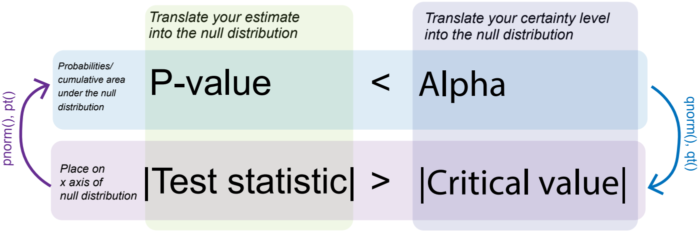
	Translate your estimate into the null distribution		*Translate your certainty level into the null distribution*
Probabilities/ cumulative area under the null distribution	P-value	$<$	Alpha
Place on x axis of null distribution	$\lvert$Test statistic$\rvert$	$>$	$\lvert$Critical value$\rvert$

If one of these comparison statements is true/false, so is the other and vice versa.

Otherwise, you have to retain the null*.

***But it's still a hypothesis! You haven't proved that it's true!**

You can reject the null hypothesis if:



If one of these comparison statements is true/false, so is the other and vice versa.

Otherwise, you have to retain the null*.

***But it's still a hypothesis! You haven't proved that it's true!**

Equations for β_0 and β_1

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}.$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{\text{Sample Covariance between } X \text{ and } Y}{\text{Sample Variance of } X}$$

Null and alternative hypotheses in regression

- Null: $H_0 : \beta_0 = 0$; $H_0 : \beta_1 = 0$
  - The null is the straw man we want to knock down.
  - With regression, almost always null of no relationship
- Alternative: $H_a : \beta_0 \neq 0$; $H_a : \beta_1 \neq 0$
  - Claim we want to test
  - Almost always "some effect"
- Notice that these have no hats! We're talking about the population parameters, not our OLS estimates. Only estimates get hats.

Test statistic

- Under the null of $H_0 : \beta_1 = c$, we can use the following familiar test statistic:

$$T = \frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}$$

where

$$\widehat{SE}[\hat{\beta}_1] = \frac{\hat{\sigma}_u}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

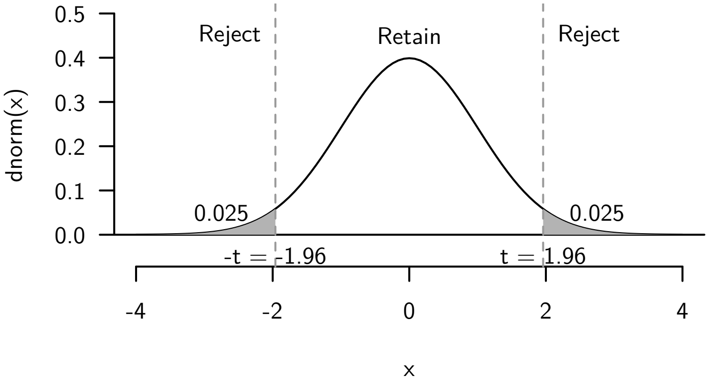
- If the errors are conditionally Normal, then under the null hypothesis we have:

$$T \sim t_{n-2}$$

Rejection region

- Choose a level of the test, α, and find rejection regions that correspond to that value under the null distribution:

$$P(-t_{\alpha/2,n-2} < T < t_{\alpha/2,n-2}) = 1 - \alpha$$



p-value

- The interpretation of the p-value is the same: *the probability of seeing a test statistic at least this extreme if the null hypothesis were true*
- Mathematically:

$$P\left(\left|\frac{\widehat{\beta}_1 - c}{\widehat{SE}[\widehat{\beta}_1]}\right| \geq |T_{obs}|\right)$$

- If the p-value is less than α we would reject the null at the α level.

Fitted values and residuals

- The **estimated** or sample regression function is:

$$\widehat{Y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

- $\widehat{\beta}_0, \widehat{\beta}_1$ are the estimated intercept and slope
- $\widehat{Y}_i$ is the fitted/predicted value
- We also have the residuals, $\widehat{u}_i$ which are the differences between the true values of Y and the predicted value:

$$\widehat{u}_i = Y_i - \widehat{Y}_i$$

- You can think of the residuals as the prediction errors of our estimates.

Prediction error

- Prediction errors without X: best prediction is the mean, so our squared errors, or the **total sum of squares** (SS_{tot}) would be:

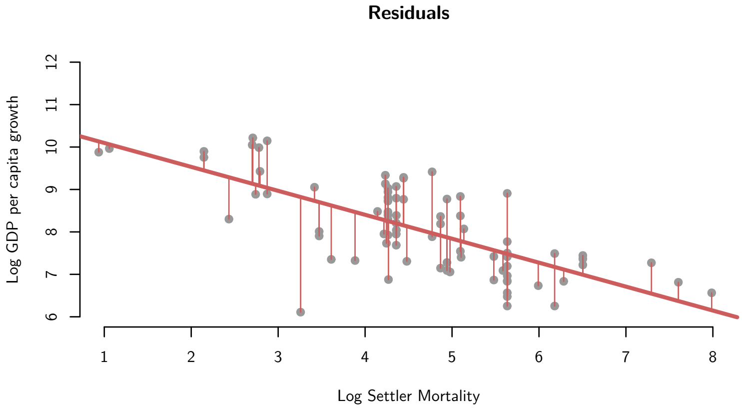
$$SS_{tot} = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$$

- Once we have estimated our model, we have new prediction errors, which are just the sum of the squared residuals or SS_{res}:

$$SS_{res} = \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

Sum of Squares

Sum of Squares



R-square

- **Coefficient of determination** or R^2:

$$R^2 = \frac{SS_{tot} - SS_{res}}{SS_{tot}} = 1 - \frac{SS_{res}}{SS_{tot}}$$

- This is the fraction of the total prediction error eliminated by providing information on X.
- Alternatively, this is the fraction of the variation in Y is "explained by" X.
- $R^2 = 0$ means no relationship
- $R^2 = 1$ implies perfect linear fit

Fun with Non-Linearities

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first **transforming** X appropriately

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first **transforming** X appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first **transforming** X appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of β_1:

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first **transforming** X appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of β_1:
  - Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates **percent increase** in Y associated with one unit increase in X

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first **transforming** X appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of β_1:
  - Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates **percent increase** in Y associated with one unit increase in X
  - Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a **percent increase** in X

Fun with Non-Linearities

- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first **transforming** X appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of β_1:
  - Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates **percent increase** in Y associated with one unit increase in X
  - Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a **percent increase** in X
  - Note that these approximations work only for small increments

Fun with Non-Linearities

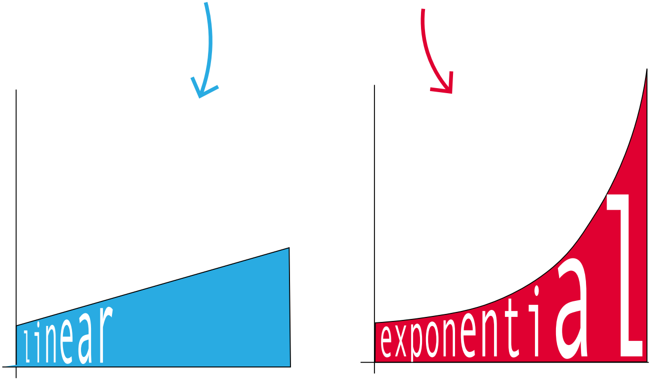
- The linear regression model *can* accommodate non-linearity in X (but not in β)
- We do this by first **transforming** X appropriately

- A useful transformation when variables are positive and right-skewed is the (natural) logarithm

- The log transformation changes the interpretation of β_1:
    - Regress $\log(Y)$ on $X \longrightarrow \beta_1$ approximates **percent increase** in Y associated with one unit increase in X
    - Regress Y on $\log(X) \longrightarrow \beta_1$ approximates increase in Y associated with a **percent increase** in X
    - Note that these approximations work only for small increments
    - In particular, they do not work when X is a discrete random variable

Why does everyone keep logging stuff??

Logs linearize exponential growth.



grows by a fixed amount.

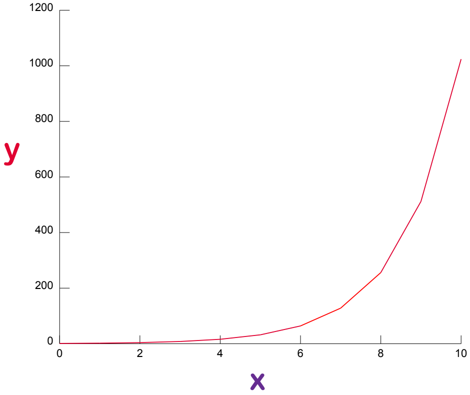
grows by a fixed percent.

How? Let's look.

First, here's a graph showing exponential growth.

We're going to use $y = 2^x$, but any other exponent will work

x	$y=(2^x)$
0	1
1	2
2	4
3	8
4	16
5	32
6	64
7	128
8	256
9	512
10	1024

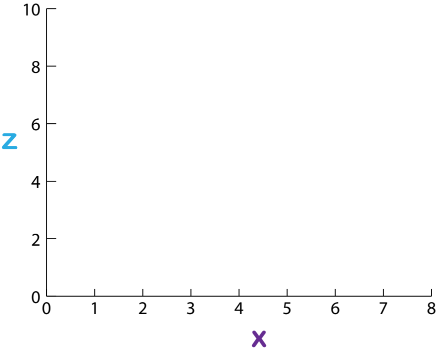


What happens when we take the log of y?

$\log y = z$ $\qquad$ $e^{z} = y$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	
1	2	
2	4	
3	8	
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

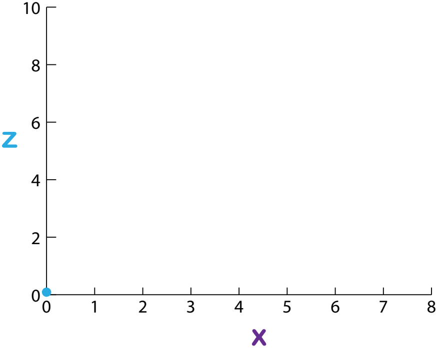


What happens when we take the log of y?

$\log 1 = 0$ $e^{0} = 1$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	0
1	2	
2	4	
3	8	
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

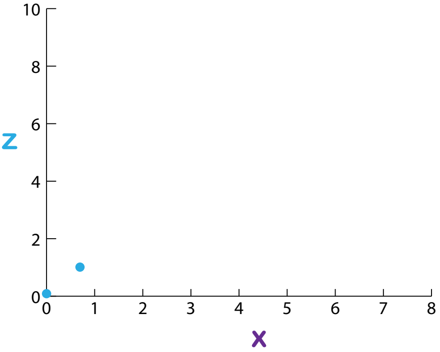


What happens when we take the log of y?

$\log 2 = .69 \qquad e^{.69} = 2$

We're going to use y = 2^x, but any other exponent will work

x	y	z
0	1	0
1	2	.69
2	4	
3	8	
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

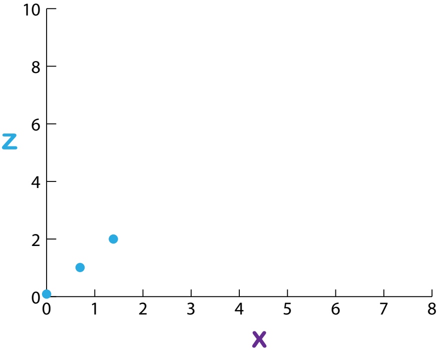


What happens when we take the log of y?

$$\log 4 = 1.39 \qquad e^{1.39} = 4$$

We're going to use y = 2^x, but any other exponent will work

x	y	z
0	1	0
1	2	.69
2	4	1.39
3	8	
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

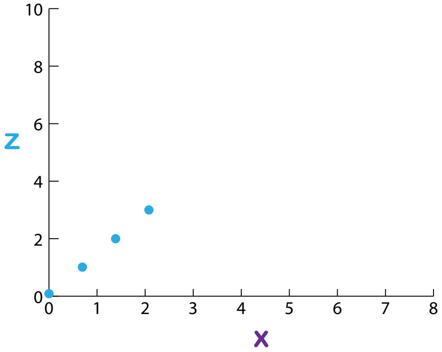


What happens when we take the log of y?

$\log 8 = 2.08 \qquad e^{2.08} = 8$

We're going to use y = 2^x, but any other exponent will work

x	y	z
0	1	0
1	2	.69
2	4	1.39
3	8	2.08
4	16	
5	32	
6	64	
7	128	
8	256	
9	512	
10	1024	

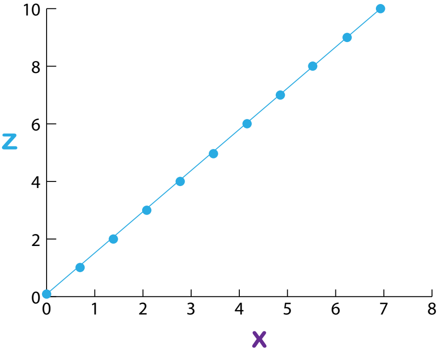


What happens when we take the log of y?

$\log y = z \qquad e^z = y$

We're going to use $y = 2^x$, but any other exponent will work

x	y	z
0	1	0
1	2	.69
2	4	1.39
3	8	2.08
4	16	2.77
5	32	3.47
6	64	4.16
7	128	4.85
8	256	5.55
9	512	6.24
10	1024	6.93



Questions?