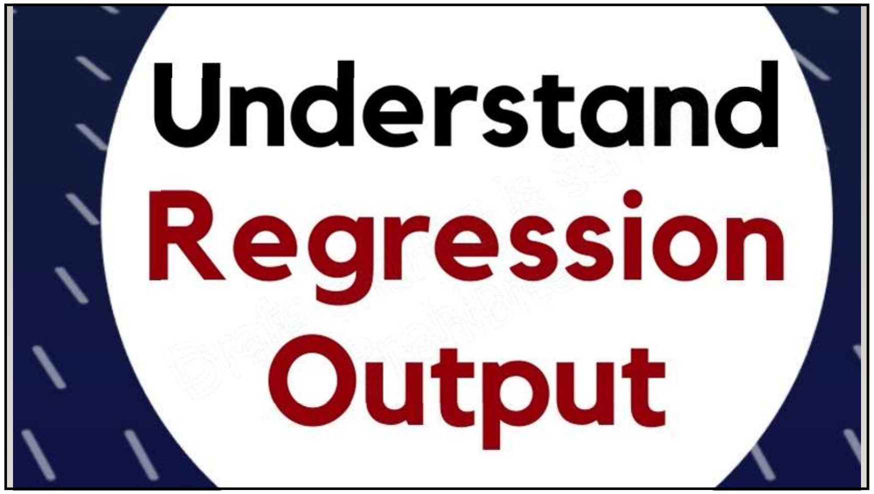


Example!

The Winter Olympics

Does a country's latitude affect their medal tally?

Winter Olympics Medal Tally

Rank	Country	Gold	Silver	Bronze	Total
1	Russian Fed.	13	11	9	33
2	Norway	11	5	10	26
3	Canada	10	10	5	25
4	United States	9	7	12	28
5	Netherlands	8	7	9	24
6	Germany	8	6	5	19
7	Switzerland	6	3	2	11
8	Belarus	5	0	1	6
9	Austria	4	8	5	17
10	France	4	4	7	15
11	Poland	4	1	1	6
12	China	3	4	2	9
13	Korea	3	3	2	8

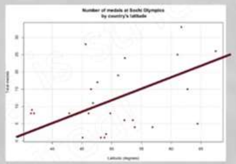
Winter Olympics Medal Tally

Rank	Country	Gold	Silver	Bronze	Total
1	Russian Fed.	13	11	9	**33**
2	Norway	11	5	10	**26**
3	Canada	10	10	5	**25**
4	United States	9	7	12	**28**
5	Netherlands	8	7	9	**24**
6	Germany	8	6	5	**19**
7	Switzerland	6	3	2	**11**
8	Belarus	5	0	1	**6**
9	Austria	4	8	5	**17**
10	France	4	4	7	**15**
11	Poland	4	1	1	**6**
12	China	3	4	2	**9**
13	Korea	3	3	2	**8**

Y variable: Number of medals ***X variables: Latitude***
Average elevation
Log population

Inference & Significance

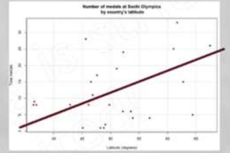
$Medals_i = \beta_0 + \beta_1(Latitude_i)$



- We can NEVER know the true slope (β_1)

Inference & Significance

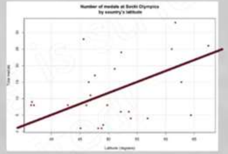
$Medals_i = \beta_0 + \beta_1(Latitude_i)$



- We can NEVER know the true slope (β_1)
- Instead, we calculate the sample slope (b_1), and make inferences about β_1

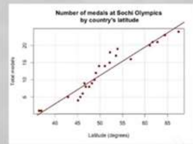
Inference & Significance

$Medals_i = \beta_0 + \beta_1(Latitude_i)$



- We can NEVER know the true slope (β_1)
- Instead, we calculate the sample slope (b_1), and make inferences about β_1
- From our sample, can we **INFER** that the true effect is positive?

Can we infer a positive relationship?

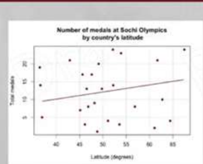


gradient of sample = 0.8

estimated gradient of population,
b = 0.8

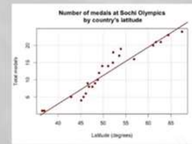
confidence interval:
0.65 < β < 0.95

Yes!



No!

Can we infer a positive relationship?

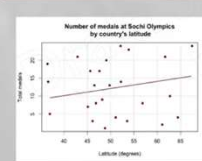


gradient of sample = 0.8

estimated gradient of population,
b = 0.8

confidence interval:
0.65 < β < 0.95

Yes!



gradient of sample = 0.2

estimated gradient of population,
b = 0.2

confidence interval:
-0.5 < β < 0.8

No!

What is "significance"?

- Start with hypothesis that the gradient is 0
  (ie. there is no relationship)

  H0: β = 0

- Use a sample to see if there is enough
  evidence to reject this null hypothesis.

  If so, we can infer:

  H1: β ≠ 0

(ie. we infer that the variable is significant!)

The Winter Olympics!

Can we infer a relationship between

Number of medals won by a country → AND →

1. The country's latitude
2. The country's average elevation
3. The country's population

$$number\ of\ medals_i = \beta_0 + \beta_1(latitude_i) + \beta_2(elevation_i) + \beta_3(\log population_i)$$

The ANOVA section

$$\widehat{number\ of\ medals}_i = b_0 + b_1(latitude_i) + b_2(elevation_i) + b_3(\log population_i)$$

Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

Number of obs =	25
F(3, 21) =	3.22
Prob > F =	0.0434
R-squared =	0.3152
Adj R-squared =	0.2173
Root MSE =	6.7418

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

Prezi

The ANOVA section

How much variation is there in the dependent variable?

Total medals: 33,28,26,25,... 3,1,1,1

Average: 11.3 medals

$$SS = \Sigma(X_i - \bar{X})^2$$
$$= (33-11.3)^2 + (28-11.3)^2 + \ldots$$
$$= 1393.76$$

Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

Prezi

The ANOVA section

Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

How much "explaining" is the model doing?

$R^2 =$

Prezi

The ANOVA section

Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

How much "explaining" is the model doing?

$R^2 = 439.27/1393.76$
$= 0.315$

Prezi

The ANOVA section

Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

How much "explaining" is the model doing?

$R^2 = 439.27/1393.76$
$= 0.315$

Is this model with 3 explanatory variables better than a model with 0 explanatory variables?

The ANOVA section

Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

How much "explaining" is the model doing?

$R^2 = 439.27/1393.76$
$= 0.315$

Is this model with 3 explanatory variables better than a model with 0 explanatory variables?

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

The ANOVA section

Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

How much "explaining" is the model doing?

$R^2 = 439.27/1393.76$
$= 0.315$

Is this model with 3 explanatory variables better than a model with 0 explanatory variables?

$H_0 : \beta_1 = \beta_2 = \beta_3 = 0$

$F_{3,21} = 146.42/45.45$
$= 3.22$

Reject H_0
at 5% level of sig.

The ANOVA section

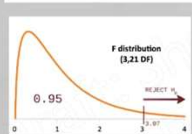
Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

Number of obs =	25
F(3, 21) =	3.22
Prob > F =	0.0434
R-squared =	0.3152
Adj R-squared =	0.2173
Root MSE =	6.7418

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

The ANOVA section

Number of obs =	25
F(3, 21) =	3.22
Prob > F =	0.0434
R-squared =	0.3152
Adj R-squared =	0.2173
Root MSE =	6.7418

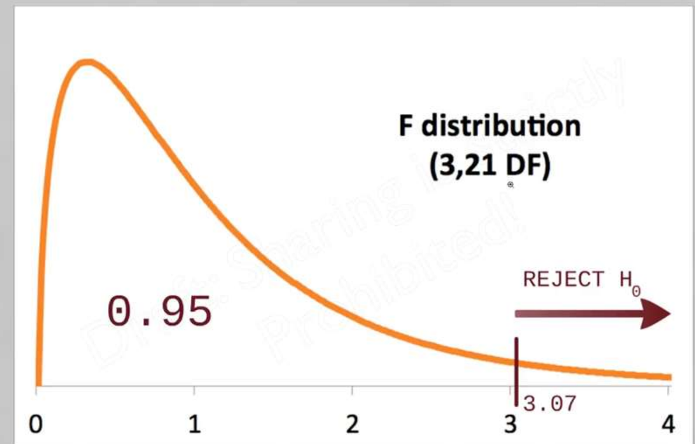


Is this model with 3 explanatory variables better than a model with 0 explanatory variables?

Prob > F = 0.0434

At 10% -> YES!
At 5% -> YES!
At 1% -> NO!

There is a 4.34% probability that the improvements we are seeing with our 3 variable model is due to random chance alone.



The Variables section

$$\widehat{number\ of\ medals_i} = b_0 + b_1(latitude_i) + b_2(elevation_i) + b_3(\log population_i)$$

Source	SS	df	MS
Model	439.274821	3	146.42494
Residual	954.485179	21	45.4516752
Total	1393.76	24	58.0733333

Number of obs =	25
F(3, 21) =	3.22
Prob > F =	0.0434
R-squared =	0.3152
Adj R-squared =	0.2173
Root MSE =	6.7418

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

$$\widehat{number\ of\ medals_i} = -54.528 + 0.523(latitude_i) + 0.003(elevation_i) + 2.146(\log population_i)$$

The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

$$\widehat{number\ of\ medals_i} = -54.528 + 0.523(latitude_i) + 0.003(elevation_i) + 2.146(\log population_i)$$

Interpretation

For every additional degree of latitude, the expected number of medals increases by 0.523 on average, holding all other variables constant.

The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

$$\widehat{number\ of\ medals_i} = -54.528 + 0.523(latitude_i) + 0.003(elevation_i) + 2.146(\log population_i)$$

Interpretation

For every additional degree of latitude, the expected number of medals increases by 0.523 on average, holding all other variables constant.

For every additional metre of average elevation, the expected number of medals increases by 0.003 on average, holding all other variables constant.

The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

$$\widehat{number\ of\ medals_i} = -54.528 + 0.523(latitude_i) + 0.003(elevation_i) + 2.146(\log population_i)$$

Estimate for Netherlands:

Latitude: 52.2 Elevation: 30.1m Pop: 16,500,000
Log pop: 16.62

The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

$$\widehat{number\ of\ medals_i} = -54.528 + 0.523(latitude_i) + 0.003(elevation_i) + 2.146(\log population_i)$$

Estimate for Netherlands:

Latitude: 52.2 Elevation: 30.1m Pop: 16,500,000
Log pop: 16.62

$$\widehat{number\ of\ medals_{NED}} = -54.528 + 0.523(52.2) + 0.003(30.1) + 2.146(16.6) = 8.557$$

The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

$$\widehat{number\ of\ medals_i} = -54.528 + 0.523(latitude_i) + 0.003(elevation_i) + 2.146(\log population_i)$$

Estimate for Netherlands:

Latitude: 52.2 Elevation: 30.1m Pop: 16,500,000

Log pop: 16.62

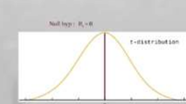
$$\widehat{number\ of\ medals_{NED}} = -54.528 + 0.523(52.2) + 0.003(30.1) + 2.146(16.6) = 8.557$$

Error(NED) = 24 - 8.6 = +15.4

Prezi

The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715



Prezi

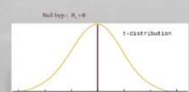
The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf. Interval]	
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

Latitude -> $t_1 = b_1 / SE_1$

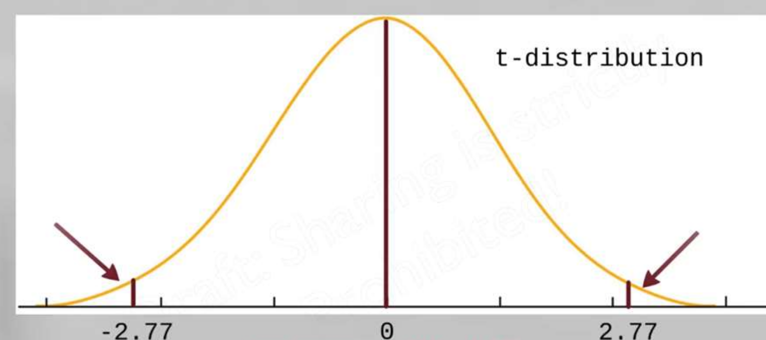
$= 0.522/0.189$

$= 2.77$

$p_1 =$ 

Prezi

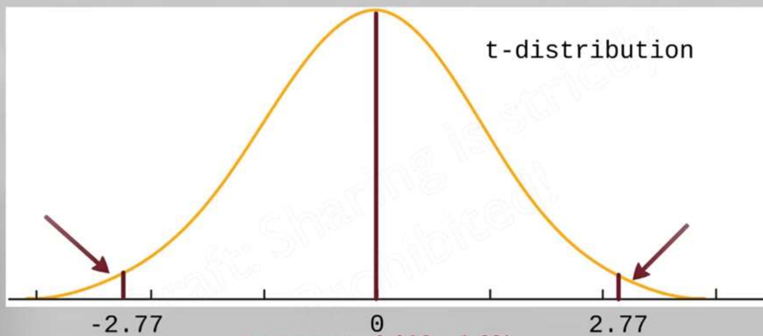
Null hyp : $B_1 = 0$ $b_1 = 0.522$ $t_1 = 2.77$



two areas = 0.012 = 1.2%

Prezi

Null hyp : $B_1 = 0$ $b_1 = 0.522$ $t_1 = 2.77$

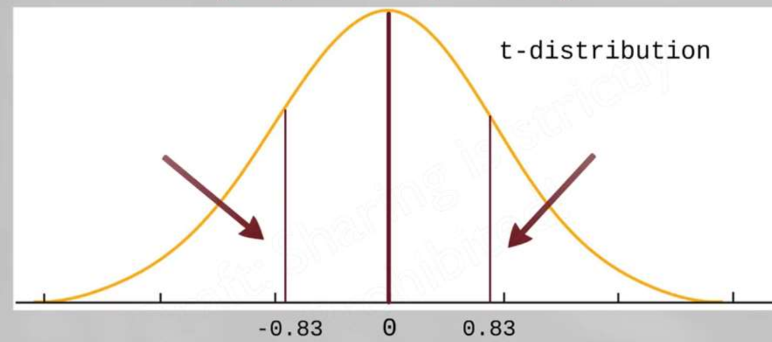


two areas = 0.012 = 1.2%

If the null hypothesis is true ($B_1 = 0$), the chance of us getting a sample AS extreme as we did ($b_1 = 0.522$), is 1.2%

Prezi

Null Hyp: $B_2 = 0$ $b_2 = 0.0317$ $t_2 = 0.83$



Two areas = 41.5%

If the null hypothesis is true ($B_2 = 0$), the chance of us getting a sample AS extreme as we did ($b_2 = 0.0317$), is 41.5%

Prezi

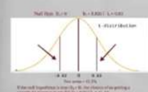
The Variables section

totalmedal	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
cen_lat	.522752	.1889091	2.77	0.012	.129894	.9156099
elev	.003171	.0038126	0.83	0.415	-.0047577	.0110996
logpop	2.146452	.9968635	2.15	0.043	.0733606	4.219543
_cons	-54.52767	21.97521	-2.48	0.022	-100.2276	-8.827715

Elevation -> $t_2 = b_2 / SE_2$

$= 0.0317/0.0038$

$= 0.83$

$p_2 =$ 

Prezi

THANK YOU!