DATA SCIENCE WITH R

DATA SCIENCE WITH R

REGRESSION ANALYSIS

Overview

Simple Linear Regression

Multiple Linear Regression

Regression Assumptions

Implementation in SAS

Regression

SIMPLE LINEAR REGRESSION

- ✓ Concepts - OLS
- ✓ How to Run
- ✓ **Interpret Results**

OLS Results : Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702085646
R Square	0.492924254
Adjusted R Square	0.49246866
Standard Error	451.3259178
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	220385522.7	2.2E+08	1081.938347	2.54E-166
Residual	1113	226712628.6	203695.1		
Total	1114	447098151.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
lIntercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-3632.001323	-2858.891455
ggestate	166.4462854	5.060260218	32.89283	2.54E-166	156.5175606	176.3750103

OLS Results : Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702085646
R Square	0.492924254
Adjusted R Square	0.49246866
Standard Error	451.3259178
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	220385522.7	2.2E+08	1081.938347	2.54E-166
Residual	1113	226712628.6	203695.1		
Total	1114	447098151.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
lIntercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-3632.001323	-2858.891455
ggestate	166.4462854	5.060260218	32.89283	2.54E-166	156.5175606	176.3750103

OLS Results : Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702085646
R Square	0.492924254
Adjusted R Square	0.49246866
Standard Error	451.3259178
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	220385522.7	2.2E+08	1081.938347	2.54E-166
Residual	1113	226712628.6	203695.1		
Total	1114	447098151.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
lIntercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-3632.001323	-2858.891455
ggestate	166.4462854	5.060260218	32.89283	2.54E-166	156.5175606	176.3750103

OLS Results : Excel

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.702085646
R Square	0.492924254
Adjusted R Square	0.49246866
Standard Error	451.3259178
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	220385522.7	2.2E+08	1081.938347	2.54E-166
Residual	1113	226712628.6	203695.1		
Total	1114	447098151.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
lIntercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-3632.001323	-2858.891455
ggestate	166.4462854	5.060260218	32.89283	2.54E-166	156.5175606	176.3750103

OLS Results Interpretation

Understanding the output

Starting with the bottom most table:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

OLS Results Interpretation

Understanding the output

Starting with the bottom most table:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

We want to estimate a straight line that best captures the relationship between Birthweight and Gestation period –

OLS Results Interpretation

Understanding the output

Starting with the bottom most table:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

We want to estimate a straight line that best captures the relationship between Birthweight and Gestation period –

As per this model that straight line is:

Birthweight = -3245.44 + 166 * Gestate

OLS Results Interpretation

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

Birthweight = -3245.44 + 166 * Gestate

OLS Results Interpretation

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

Birthweight = -3245.44 + 166 * Gestate

Here, the intercept is -3245. 44

the beta coefficient on Gestate is 166

OLS Results Interpretation

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

Birthweight = -3245.44 + 166 * Gestate

Here, the intercept is -3245. 44

the beta coefficient on Gestate is 166

How do we interpret the Beta Coefficient?

OLS Results Interpretation

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

Birthweight = -3245.44 + 166 * Gestate

Here, the intercept is -3245. 44

the beta coefficient on Gestate is 166

How do we interpret the Beta Coefficient?

For a unit increase in gestation period (1 week), the average increase in birthweight is 166

OLS Results Interpretation

BETA Coefficient:

- For every unit increase in Gestation Period, we expect to see an increase in Birthweight by 166 grams

OLS Results Interpretation

BETA Coefficient:

- For every unit increase in Gestation Period, we expect to see an increase in Birthweight by 166 grams
- Positive sign on the coefficient on gestation implies a positive relationship between Gestation Period and Birthweight

OLS Results Interpretation

BETA Coefficient:

- For every unit increase in Gestation Period, we expect to see an increase in Birthweight by 166 grams
- Positive sign on the coefficient on gestation implies a positive relationship between Gestation Period and Birthweight
- What does unit increase mean?

OLS Results Interpretation

BETA Coefficient:

- For every unit increase in Gestation Period, we expect to see an increase in Birthweight by 166 grams
- Positive sign on the coefficient on gestation implies a positive relationship between Gestation Period and Birthweight
- What does unit increase mean?
- Will every additional week of gestation automatically add 166 grams of birthweight to every baby?

OLS Results Interpretation

How do we interpret the estimated regression function?

Birthweight = -3245.44 + 166 * Gestate

INTERCEPT

OLS Results Interpretation

How do we interpret the estimated regression function?

Birthweight = -3245.44 + 166 * Gestate

INTERCEPT

With zero gestation weeks, we expect birthweight to be Negative

OLS Results Interpretation

How do we interpret the estimated regression function?

Birthweight = -3245.44 + 166 * Gestate

INTERCEPT

With zero gestation weeks, we expect birthweight to be Negative

- It doesn't really make sense to talk about birthweight at Zero weeks

OLS Results Interpretation

How do we interpret the estimated regression function?

Birthweight = -3245.44 + 166 * Gestate

INTERCEPT

With zero gestation weeks, we expect birthweight to be Negative

- It doesn't really make sense to talk about birthweight at Zero weeks
- Provides a baseline

OLS Results Interpretation

In our example:

Can we say that if birthweight increases 166, gestation will increase by 1?

OLS Results Interpretation

In our example:

Can we say that if birthweight increases 166, gestation will increase by 1?

What about the error term? How does it influence the interpretation of results?

OLS Results Interpretation

The second piece of critical information from the coefficients table is the P values

OLS Results Interpretation

The second piece of critical information from the coefficients table is the P values

P-values denote the probability of rejecting the null hypothesis when it is in fact true

OLS Results Interpretation

The second piece of critical information from the coefficients table is the P values

P-values denote the probability of rejecting the null hypothesis when it is in fact true

- H0: Beta coefficient = 0 (that is, independent variable has no impact on the dependent variable)

OLS Results Interpretation

The second piece of critical information from the coefficients table is the P values

P-values denote the probability of rejecting the null hypothesis when it is in fact true

- H0: Beta coefficient = 0 (that is, independent variable has no impact on the dependent variable)

Lower the p-value?

OLS Results Interpretation

How do we actually test the hypothesis?

OLS Results Interpretation

How do we actually test the hypothesis?

One main point to remember is that we can show the distribution of

$$(\hat{\beta}_j - \beta_j) / se(\hat{\beta}_j) \sim t_{n-k-1}$$

OLS Results Interpretation

How do we actually test the hypothesis?

One main point to remember is that we can show the distribution of

$$(\hat{\beta}_j - \beta_j) / se(\hat{\beta}_j) \sim t_{n-k-1}$$

What does the above equation mean?

OLS Results Interpretation

How do we actually test the hypothesis?

One main point to remember is that we can show the distribution of

$$(\hat{\beta}_j - \beta_j) / se(\hat{\beta}_j) \sim t_{n-k-1}$$

What does the above equation mean?

The difference between the estimated coefficient and the actual value in the population divided by the standard error of the estimated population is distributed as a t-distribution with n-k-1 degrees of freedom, where k+1 are the number of unknown parameters in the population model

OLS Results Interpretation

In the results table for the simple regression we have run, the p-value on the drivers variable is extremely low

	Coefficients	*Standard Error*	*t Stat*	*P-value*	
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-
gestate	166.4462854	5.060260218	32.89283	2.54E-166	

- We should accept the alternate hypothesis that as gestation weeks increase, birthweight will increase
- i.e., gestation period is a statistically significant influencer of birthweight

OLS Results: Confidence Levels

What about reliability or confidence in the results?

- If we see a beta coefficient of 166, are we certain that 100% of the time that if gestation increases by 1 week, then birthweight will increase by 166?
- Remember, the beta estimate is true of the sample on which the model has been built

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	-3245.446394	197.0110519	-16.4734	9.95259E-55	-3632.001323	-2858.891465
gestate	166.4462854	5.060260218	32.89283	2.54E-166	156.5175606	176.3750103

To Be Continued

Regression Analysis

Simple Linear Regression

THANK YOU