DATA SCIENCE WITH R

DATA SCIENCE WITH R

REGRESSION ANALYSIS

Overview

Simple Linear Regression

Multiple Linear Regression

Regression Assumptions

Implementation in SAS

© Jigsaw Academy Education Pvt Ltd

Regression

MULTIPLE LINEAR REGRESSION

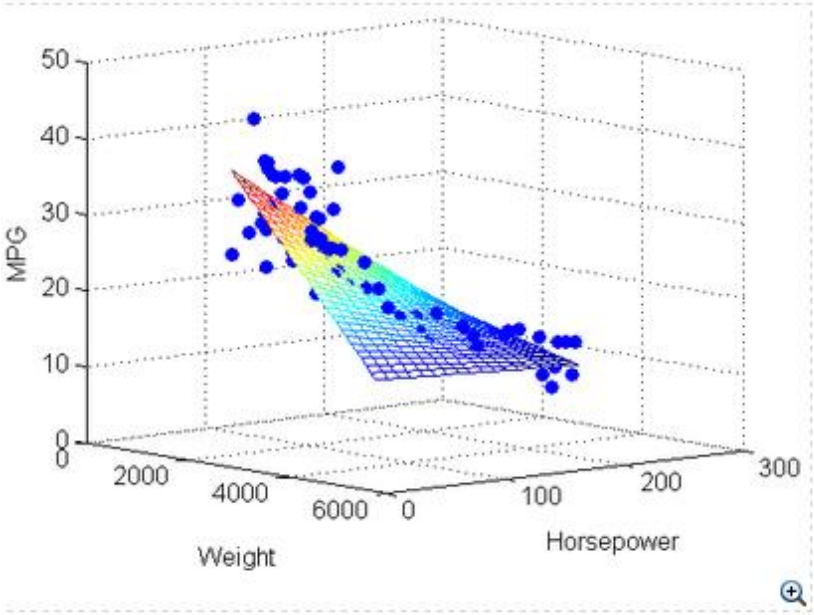
- ✓ **Concepts - OLS**
- ✓ **How to Run**
- ✓ Interpret Results

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

In real life business situations:

- expect multiple independent variables simultaneously impacting the dependent

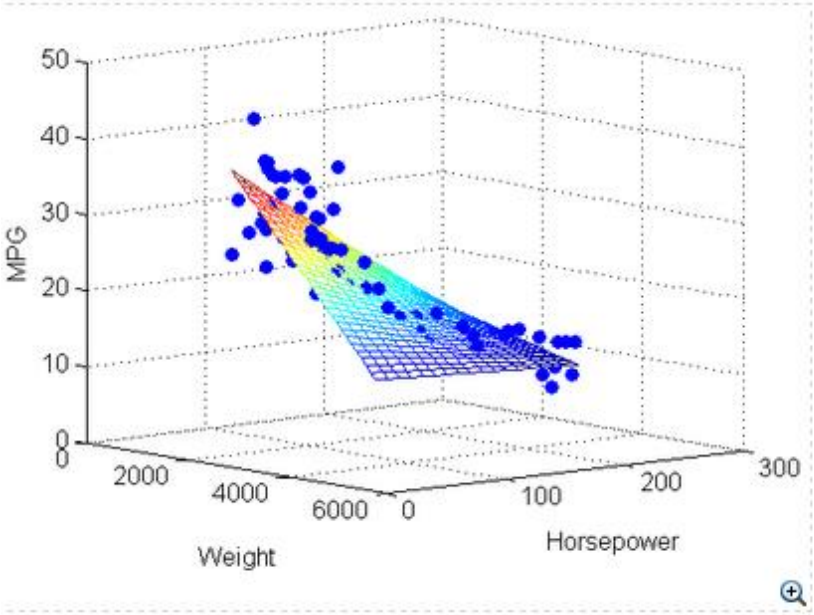


© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

In real life business situations:

- expect multiple independent variables simultaneously impacting the dependent
- We would again estimate the line across multiple dimensions that would minimize the sum of squared residuals



© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

OLS ESTIMATES

- Gestation period is not the only explanatory variable to explain differences in birthweight

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

OLS ESTIMATES

- Gestation period is not the only explanatory variable to explain differences in birthweight

- We had data on other independent variables: Mother's education level, Race, and Smoking Status

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

OLS ESTIMATES

- Gestation period is not the only explanatory variable to explain differences in birthweight
- We had data on other independent variables: Mother's education level, Race, and Smoking Status
- So the actual regression equation is:

$$Birthweight = \beta_0 + \beta_1 * Gestation + \beta_2 * Years\ Of\ Education + \beta_3 * Race + \beta_4 * Smoking$$

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

OLS ESTIMATES

$$Birthweight = \beta_0 + \beta_1 * Gestation + \beta_2 * Years\ Of\ Education + \beta_3 * Race + \beta_4 * Smoking$$

- So, we would now need to estimate 4 beta coefficients
- We would use the same OLS approach of minimizing sum of squared residuals across multiple dimensions
- It is difficult to visualize the actual process of minimizing errors in multiple dimensions (as we can do easily in 2 dimensions), but the logic of minimizing residuals is identical

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.726512578
R Square	0.527820526
Adjusted R Square	0.526118978
Standard Error	436.1074441
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	235987581.2	58996895.29	310.2002602	3.9601E-179
Residual	1110	211110570.1	190189.7028		
Total	1114	447098151.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?

  Coefficient signs, and p-values

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?

  Coefficient signs, and p-values

- Are all the coefficient signs as expected? What should be “expected?”

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?

  Coefficient signs, and p-values

- Are all the coefficient signs as expected? What should be “expected?”

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?

  Coefficient signs, and p-values

- Are all the coefficient signs as expected? What should be “expected?”

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?

  Coefficient signs, and p-values

- Are all the coefficient signs as expected? What should be “expected?”

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?

  Coefficient signs, and p-values

- Are all the coefficient signs as expected? What should be “expected?”

- Which of the independent variables are significant?

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- What are key things to look in this output?

  Coefficient signs, and p-values

- Are all the coefficient signs as expected? What should be “expected?”

- Which of the independent variables are significant?

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

- Relationships captured in this model between the IVs and the DV seem intuitively “correct”

- YearsEducation is insignificant at the 5% level of significance

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.726512578
R Square	0.527820526
Adjusted R Square	0.526118978
Standard Error	436.1074441
Observations	1115

R2 is only 52%, even with the introduction of additional IVs

© Jigsaw Academy Education Pvt Ltd

Multiple Linear Regression

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.726512578
R Square	0.527820526
Adjusted R Square	0.526118978
Standard Error	436.1074441
Observations	1115

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	4	235987581.2	58996895.29	310.2002602	3.9601E-179
Residual	1110	211110570.1	190189.7028		
Total	1114	447098151.3			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*
Intercept	-2834.485706	215.6221999	-13.14561166	8.64385E-37	-3257.55877
YearsEduc	9.571829193	6.458197303	1.482120899	0.138591957	-3.09982207
Race (1=b)	-168.9683577	27.26026985	-6.198337677	8.03139E-10	-222.4558274
Smoke	-174.8128917	31.62426255	-5.527809271	4.04011E-08	-236.8629666
gestate	156.5115539	5.013557387	31.21766478	4.4373E-154	146.6744356

© Jigsaw Academy Education Pvt Ltd

Regression Results: Model Fit

- R^2
- Fit Chart - Actual v/s Fitted Values
- MAPE – Mean Absolute Percentage Error

© Jigsaw Academy Education Pvt Ltd

To Be Continued

Regression Analysis

Multiple Linear Regression

© Jigsaw Academy Education Pvt Ltd

THANK YOU

© Jigsaw Academy Education Pvt Ltd