DATA SCIENCE WITH R

REGRESSION ANALYSIS

Overview

➔ **Simple Linear Regression**

Multiple Linear Regression

Regression Assumptions

Implementation in SAS

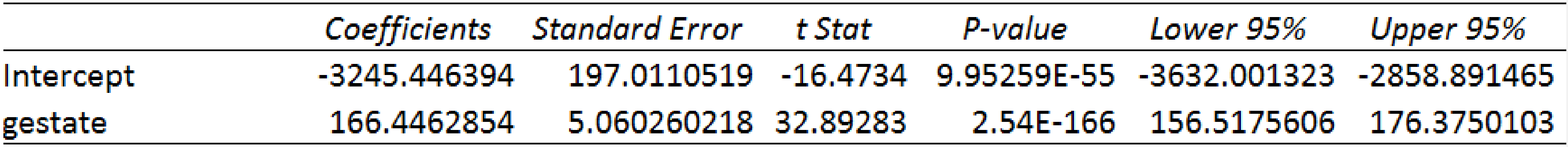
Regression

SIMPLE LINEAR REGRESSION

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

OLS Results: Confidence Levels

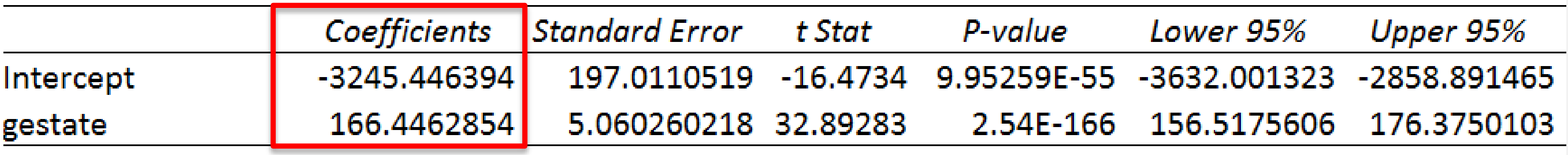
The 95% confidence interval tell us – for a 1 week increase in gestation period, we expect to see an increase in birthweight of between 156.5 and 176.4 gms 95% of the time.



	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

OLS Results: Confidence Levels

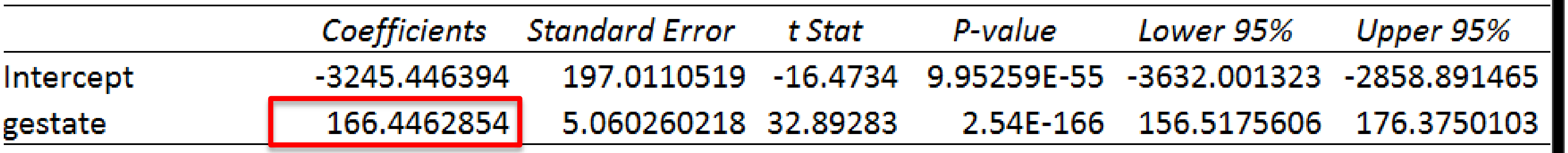
The 95% confidence interval tell us – for a 1 week increase in gestation period, we expect to see an increase in birthweight of between 156.5 and 176.4 gms 95% of the time.



	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

OLS Results: Confidence Levels

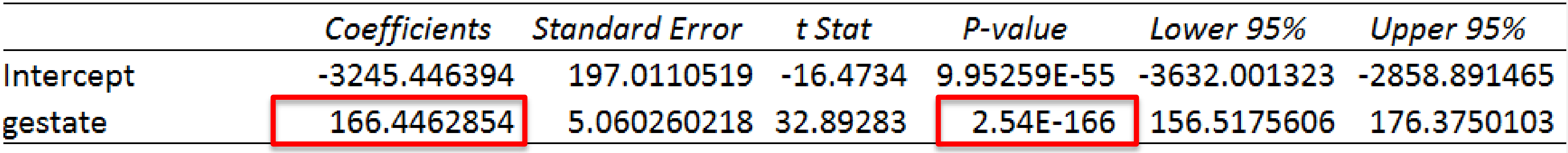
The 95% confidence interval tell us – for a 1 week increase in gestation period, we expect to see an increase in birthweight of between 156.5 and 176.4 gms 95% of the time.



	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

OLS Results: Confidence Levels

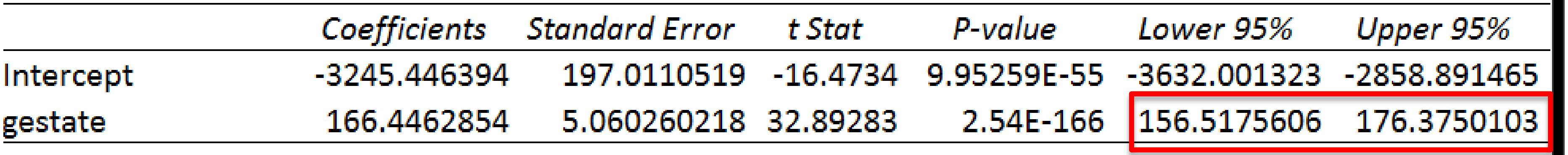
The 95% confidence interval tell us – for a 1 week increase in gestation period, we expect to see an increase in birthweight of between 156.5 and 176.4 gms 95% of the time.



	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

OLS Results: Confidence Levels

The 95% confidence interval tell us – for a 1 week increase in gestation period, we expect to see an increase in birthweight of between 156.5 and 176.4 gms 95% of the time.



	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

If we run regression models on multiple random samples from the same population many times, then 95% of the time the point estimate of the coefficient on the independent variable of interest will lie within the lower and upper bounds calculated

OLS Results Interpretation

While the regression equation is the best straight line equation possible, how do we assess the effectiveness of the overall model?

OLS Results Interpretation

While the regression equation is the best straight line equation possible, how do we assess the effectiveness of the overall model?

One way is to look at a measure of “explainability”; i.e., how much of the dependent variable Y is explained by X?

Or, a better way to put it is, how much of the variance in Y is explained by X?

OLS Results Interpretation

While the regression equation is the best straight line equation possible, how do we assess the effectiveness of the overall model?

One way is to look at a measure of “explainability”; i.e., how much of the dependent variable Y is explained by X?
Or, a better way to put it is, how much of the variance in Y is explained by X?

The mathematical calculation is:

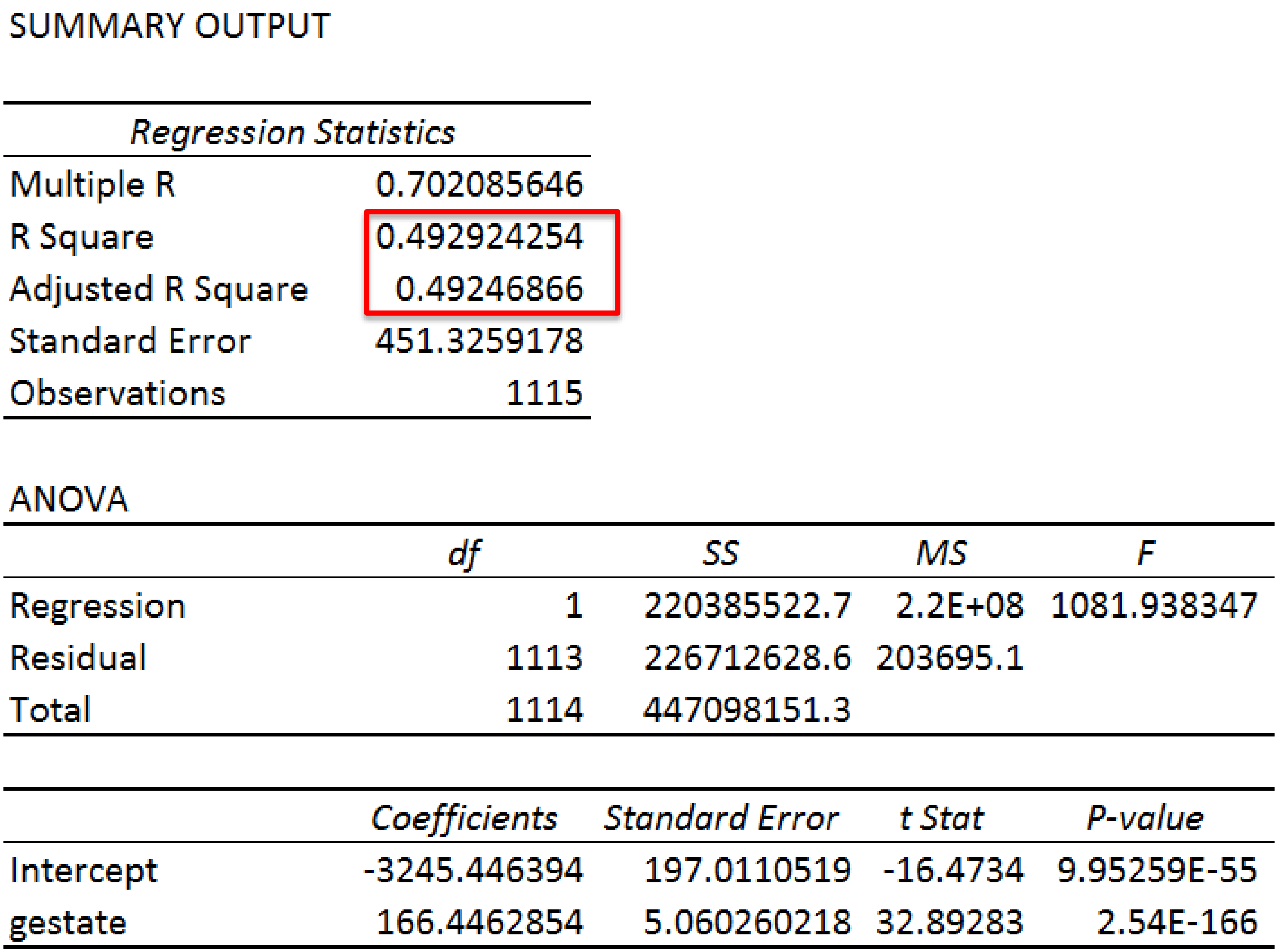
$$R^2 \equiv 1 - \frac{SS_{\text{err}}}{SS_{\text{tot}}}$$

Where,

$$SS_{\text{tot}} = \sum_i (y_i - \bar{y})^2$$

$$SS_{\text{err}} = \sum_i (y_i - f_i)^2$$

OLS Results Interpretation



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*
Regression	1	220385522.7	2.2E+08	1081.938347
Residual	1113	226712628.6	203695.1	
Total	1114	447098151.3		

	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

The R2 estimate is 49%, which implies that 49% of the variation in birthweight is captured or explained by variation in the gestation weeks variable

OLS Results Interpretation

The R2 estimate is 49%, which implies that 49% of the variation in birthweight is captured or explained by variation in the gestation weeks variable

- Clearly, the higher the R^2 the better the model

OLS Results Interpretation

The R2 estimate is 49%, which implies that 49% of the variation in birthweight is captured or explained by variation in the gestation weeks variable

- Clearly, the higher the R^2 the better the model
- However, R^2 is not the only indicator of model fit

OLS Results Interpretation

The R2 estimate is 49%, which implies that 49% of the variation in birthweight is captured or explained by variation in the gestation weeks variable

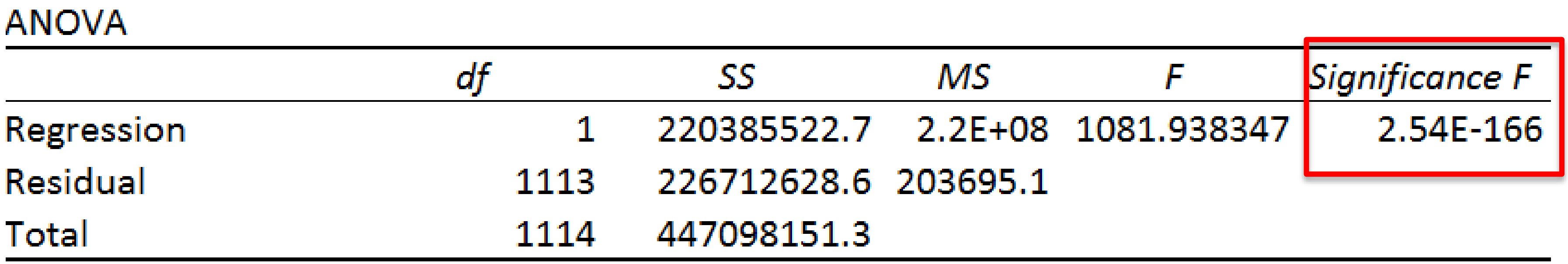
- Clearly, the higher the R^2 the better the model
- However, R^2 is not the only indicator of model fit
- It is possible to have the same R^2 but different models with different fit

OLS Results Interpretation

The R2 estimate is 49%, which implies that 49% of the variation in birthweight is captured or explained by variation in the gestation weeks variable

- Clearly, the higher the R^2 the better the model
- However, R^2 is not the only indicator of model fit
- It is possible to have the same R^2 but different models with different fit
- R^2 also increases with addition of variables, whether relevant or not, it is better to use the adjusted R2 measure

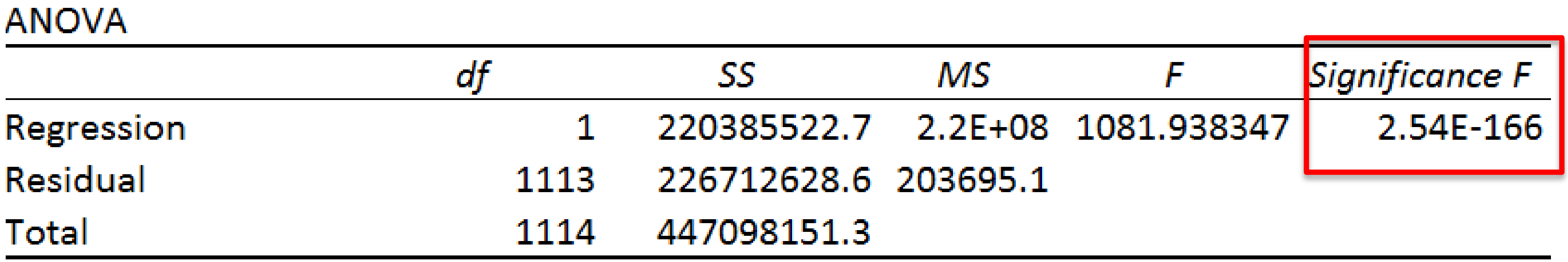
OLS Results Interpretation



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			

OLS Results Interpretation

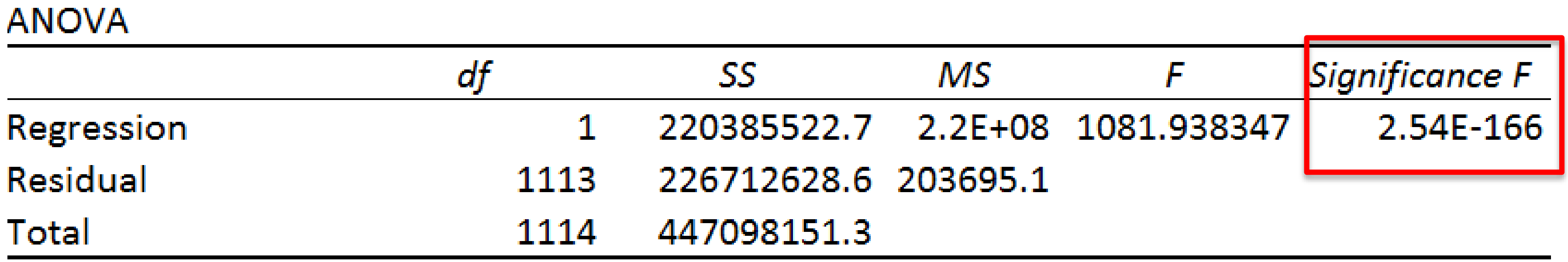


ANOVA

	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			

The ANOVA table shows us the output of the test of the hypothesis that at least one of the beta coefficients is different from zero

OLS Results Interpretation



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			

The ANOVA table shows us the output of the test of the hypothesis that at least one of the beta coefficients is different from zero

In this example, p value < 0.05, so we conclude that at least one of the beta coefficients is significant (in this case, we have only one beta)

Coming Up

Regression Analysis

Multiple Linear Regression

THANK YOU