DATA SCIENCE WITH R

DATA SCIENCE WITH R

REGRESSION

★ Regression Analysis ★

REGRESSION ANALYSIS

Overview

Simple Linear Regression

➔ **Multiple Linear Regression**

Regression Assumptions

Implementation in SAS

Regression

MULTIPLE LINEAR REGRESSION

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

Regression Results: Model Fit

Regression Results: Model Fit

How do we validate the model?

Regression Results: Model Fit

How do we validate the model?

- R^2

Regression Results: Model Fit

How do we validate the model?

- R^2
- Fit Chart - Actual vs Fitted Values

Regression Results: Model Fit

How do we validate the model?

- R^2
- Fit Chart - Actual vs Fitted Values
- MAPE – Mean Absolute Percentage Error

Regression Results: Model Fit

How do we validate the model?

- R^2
- Fit Chart - Actual vs Fitted Values

Regression Results: Model Fit

How do we validate the model?

- R^2
- **Fit Chart - Actual vs Fitted Values**

*Birthweight = - 2834 + 156.51*Gestation + 9.57* Years Of Education -168.9 *Race - 174.8 *Smoking*

Regression Results: Model Fit

How do we validate the model?

- R^2
- **Fit Chart - Actual vs Fitted Values**

Fitted values are values of the Dependent variable (Birthweight) according to the model equation

*Birthweight = - 2834 + 156.51*Gestation + 9.57* Years Of Education -168.9 *Race - 174.8 *Smoking*

Regression Results: Model Fit

How do we validate the model?

- R^2
- **Fit Chart - Actual vs Fitted Values**

Fitted values are values of the Dependent variable (Birthweight) according to the model equation

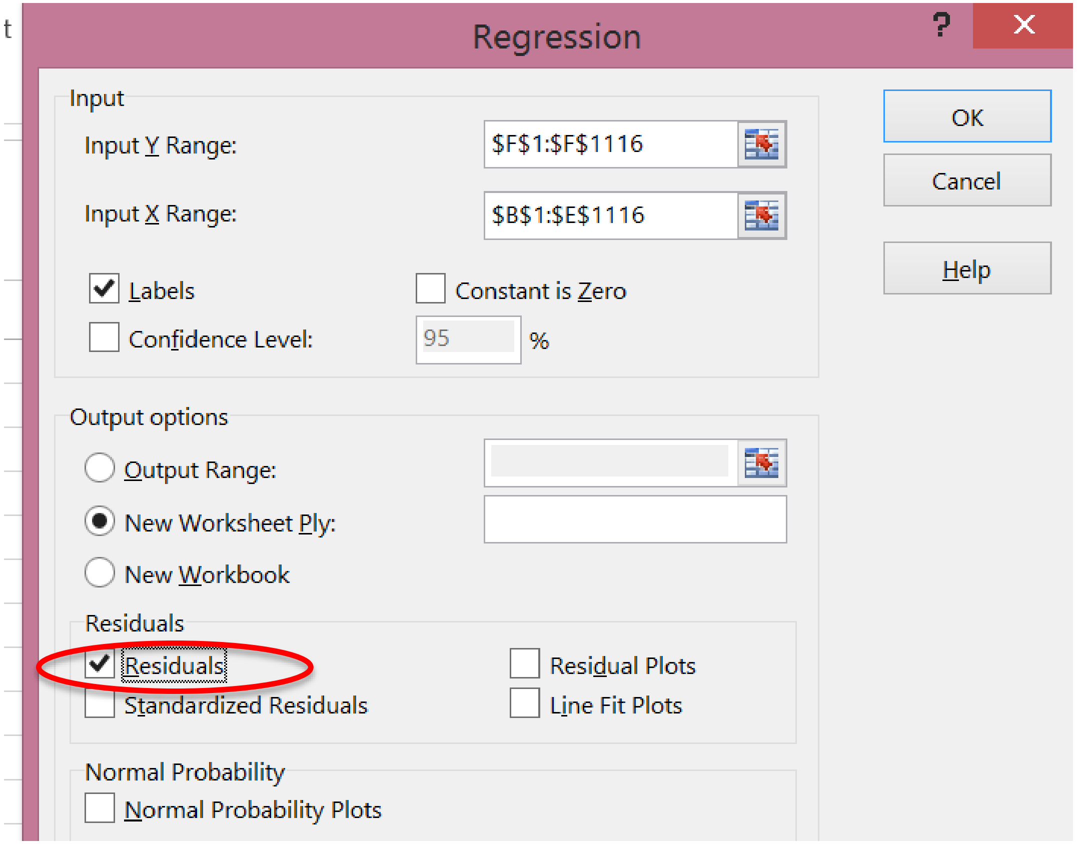
$$\text{Birthweight} = -2834 + 156.51*\text{Gestation} + 9.57*\text{Years Of Education} - 168.9*\text{Race} - 174.8*\text{Smoking}$$

Given values of the X’s (IVs), we can come up with a Fitted value for Y (DV)

Regression Results: Model Fit

*Birthweight = - 2834 + 156.51*Gestation + 9.57* Years Of Education -168.9 *Race - 174.8 *Smoking*

We can automatically generate the fitted values in Excel, using the actual data values for the X variable values:



Regression

Input

Input Y Range: $F$1:$F$1116

Input X Range: $B$1:$E$1116

☑ Labels □ Constant is Zero

□ Confidence Level: 95 %

Output options

○ Output Range:

● New Worksheet Ply:

○ New Workbook

Residuals

☑ Residuals □ Residual Plots

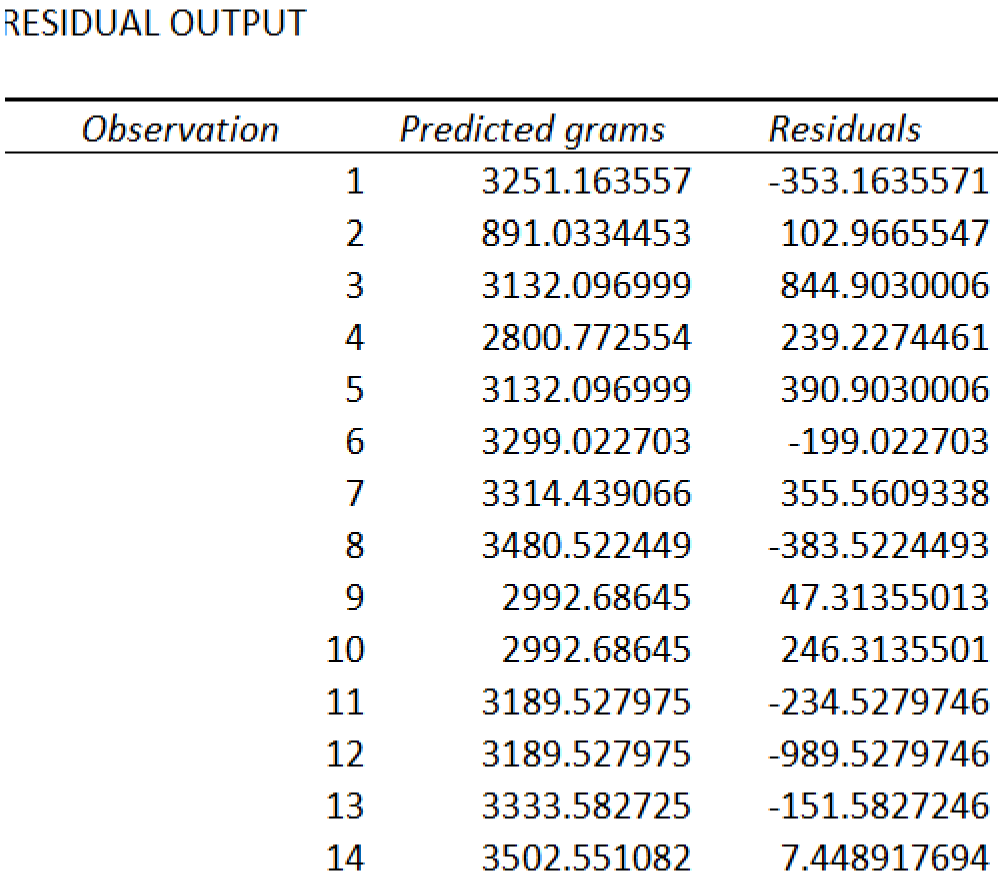
□ Standardized Residuals □ Line Fit Plots

Normal Probability

□ Normal Probability Plots

Regression Results: Model Fit

This will generate predicted (fitted) values, and residuals



RESIDUAL OUTPUT

Observation	*Predicted grams*	*Residuals*
1	3251.163557	-353.1635571
2	891.0334453	102.9665547
3	3132.096999	844.9030006
4	2800.772554	239.2274461
5	3132.096999	390.9030006
6	3299.022703	-199.022703
7	3314.439066	355.5609338
8	3480.522449	-383.5224493
9	2992.68645	47.31355013
10	2992.68645	246.3135501
11	3189.527975	-234.5279746
12	3189.527975	-989.5279746
13	3333.582725	-151.5827246
14	3502.551082	7.448917694

Predicted Values are the fitted values

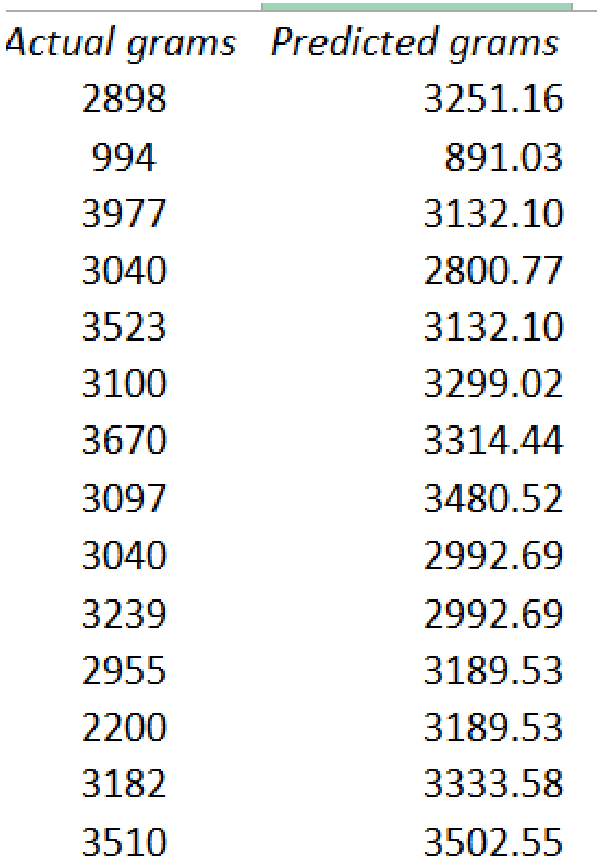
Residuals are the difference between Predicted Values of Y and the Actual Values of Y

How many predicted values will be obtained?

Regression Results: Model Fit

How are the predicted values a measure of model validation?

We can compare the actual Y values to the predicted Y values



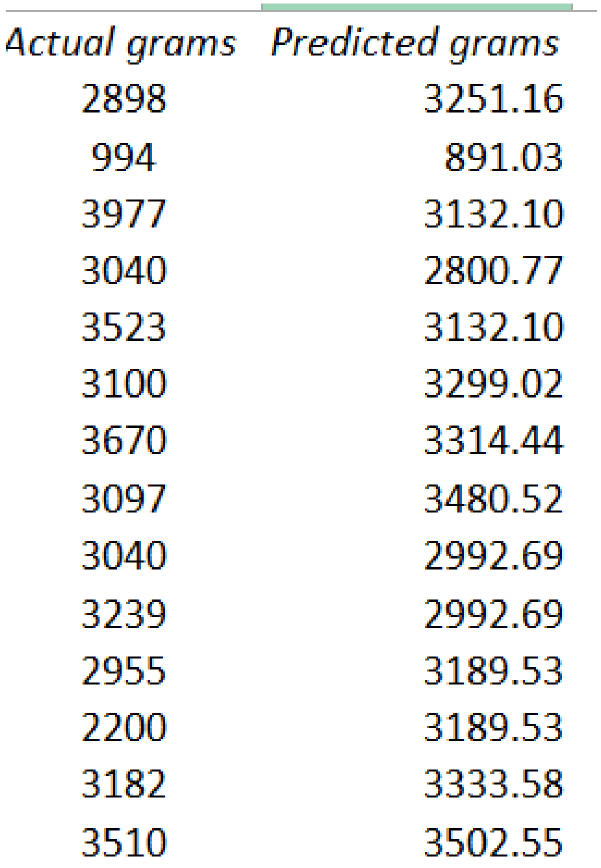
Actual grams	Predicted grams
2898	3251.16
994	891.03
3977	3132.10
3040	2800.77
3523	3132.10
3100	3299.02
3670	3314.44
3097	3480.52
3040	2992.69
3239	2992.69
2955	3189.53
2200	3189.53
3182	3333.58
3510	3502.55

Regression Results: Model Fit

How are the predicted values a measure of model validation?

We can compare the actual Y values to the predicted Y values

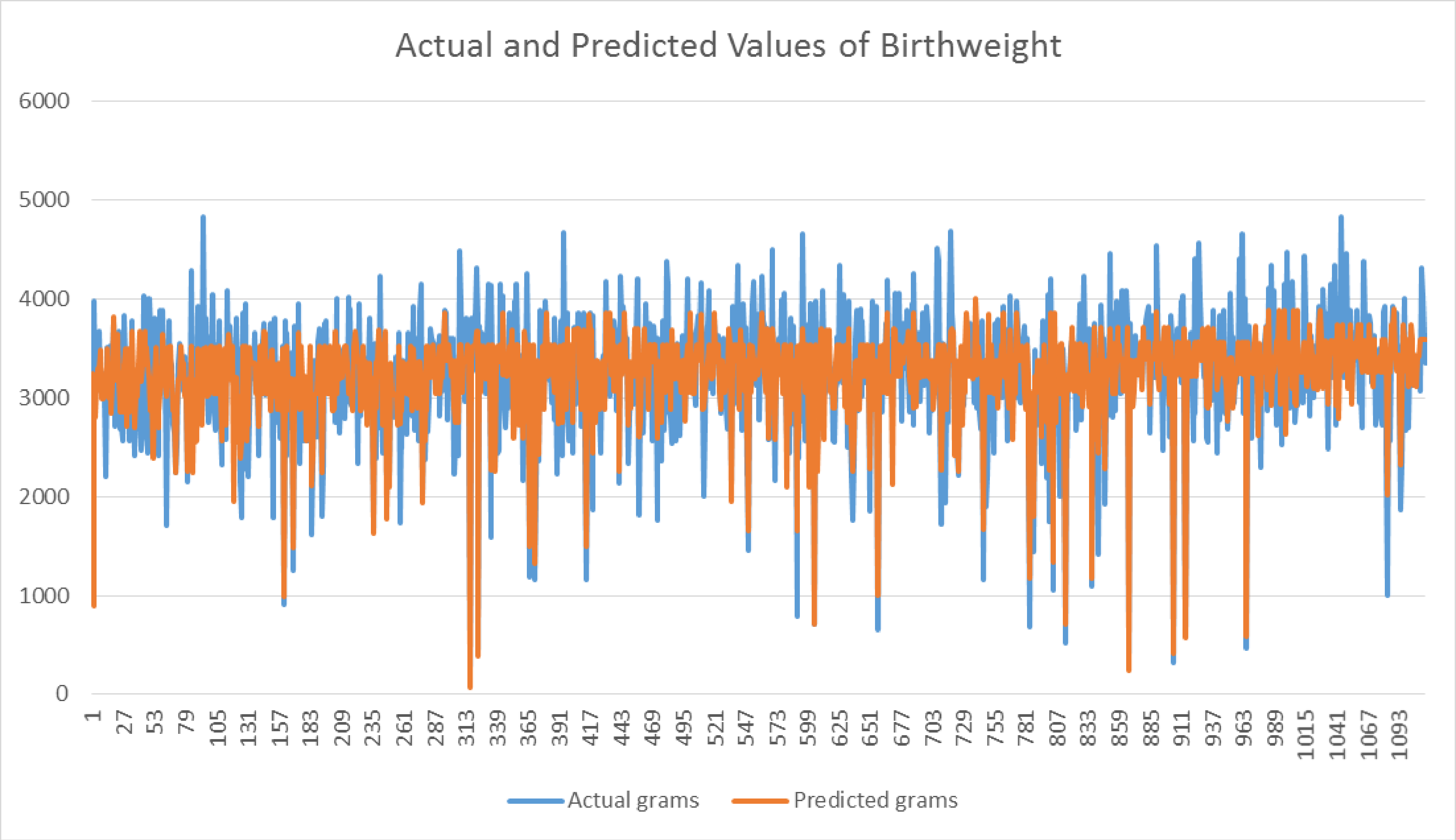
For a good model, what would be expect in the comparision?



Actual grams	Predicted grams
2898	3251.16
994	891.03
3977	3132.10
3040	2800.77
3523	3132.10
3100	3299.02
3670	3314.44
3097	3480.52
3040	2992.69
3239	2992.69
2955	3189.53
2200	3189.53
3182	3333.58
3510	3502.55

Regression Results: Model Fit

Visual Comparision of Actual and Predicted Values: FIT CHART



Regression Results: Model Fit

How do we validate the model?

- R^2
- Fit Chart - Actual vs Fitted Values
- **MAPE – Mean Absolute Percentage Error**

Regression Results: Model Fit

How do we validate the model?

- R^2
- Fit Chart - Actual vs Fitted Values
- **MAPE – Mean Absolute Percentage Error**

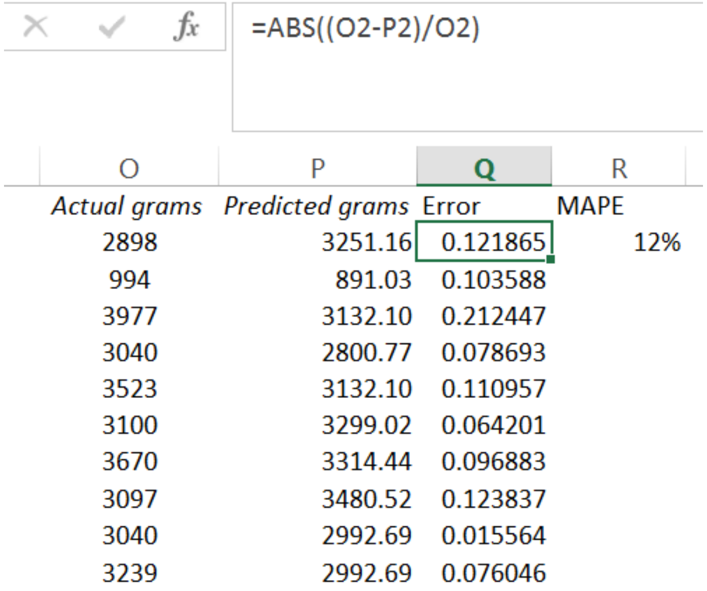
The average absolute difference between Actual and Predicted values generates the MAPE

Regression Results: Model Fit

How do we validate the model?

- R^2
- Fit Chart - Actual vs Fitted Values
- **MAPE – Mean Absolute Percentage Error**

The average absolute difference between Actual and Predicted values generates the MAPE



=ABS((O2-P2)/O2)

O	P	Q	R
Actual grams	*Predicted grams*	Error	MAPE
2898	3251.16	0.121865	12%
994	891.03	0.103588	
3977	3132.10	0.212447	
3040	2800.77	0.078693	
3523	3132.10	0.110957	
3100	3299.02	0.064201	
3670	3314.44	0.096883	
3097	3480.52	0.123837	
3040	2992.69	0.015564	
3239	2992.69	0.076046	

Regression Results

PREDICTIVE MODEL

How is a regression a predictive modeling technique?

Regression Results

PREDICTIVE MODEL

How is a regression a predictive modeling technique?

- Once we have a final validated model, given the regression equation, for values of X we can predict a “Y” value

Regression Results

PREDICTIVE MODEL

How is a regression a predictive modeling technique?

- Once we have a final validated model, given the regression equation, for values of X we can predict a “Y” value

- We calculated fitted values or predicted values for the actual data values of X in our dataset

- We can use the same calculation for other (future) values of X’s

Regression Results

PREDICTIVE MODEL

For example, we have a mother with 10 years of education, Race = Black (1), expected Gestation period = 40 weeks, and Smoking = No (0),

Regression Results

PREDICTIVE MODEL

For example, we have a mother with 10 years of education, Race = Black (1), expected Gestation period = 40 weeks, and Smoking = No (0),

we can calculate the expected birthweight as:

Birthweight = _- 2834 + 156.51*Gestation + 9.57* Years Of Education -168.9 *Race_
- 174.8 *Smoking

= -2834 + 156.51 * 40 + 9.57 *10 – 168.9 * 1 – 174.8 * 0

= 3352.76 gms

Regression Results

PREDICTIVE MODEL

For example, we have a mother with 10 years of education, Race = Black (1), expected Gestation period = 40 weeks, and Smoking = No (0),

we can calculate the expected birthweight as:

$$\text{Birthweight} = -2834 + 156.51 * \mathit{Gestation} + 9.57 * \mathit{Years\ Of\ Education} - 168.9 * \mathit{Race} - 174.8 * \mathit{Smoking}$$

$$= -2834 + 156.51 * 40 + 9.57 * 10 - 168.9 * 1 - 174.8 * 0$$

$$= 3352.76 \text{ gms}$$

This is the predicted weight of the baby

Regression Results

PREDICTIVE MODEL

Why should we believe the predicted value?

Regression Results

PREDICTIVE MODEL

Why should we believe the predicted value?

If we have a good model (High R2, good fit, low MAPE), we can be confident about our predictions

Regression Results

PREDICTIVE MODEL

Why should we believe the predicted value?

If we have a good model (High R2, good fit, low MAPE), we can be confident about our predictions

In this example, our R2 is only 52%, and MAPE is 12%. This is not a great model

Regression Results

PREDICTIVE MODEL

Why should we believe the predicted value?

If we have a good model (High R2, good fit, low MAPE), we can be confident about our predictions

In this example, our R2 is only 52%, and MAPE is 12%. This is not a great model

What next?

Coming Up

Regression Analysis

Regression Assumptions

THANK YOU