

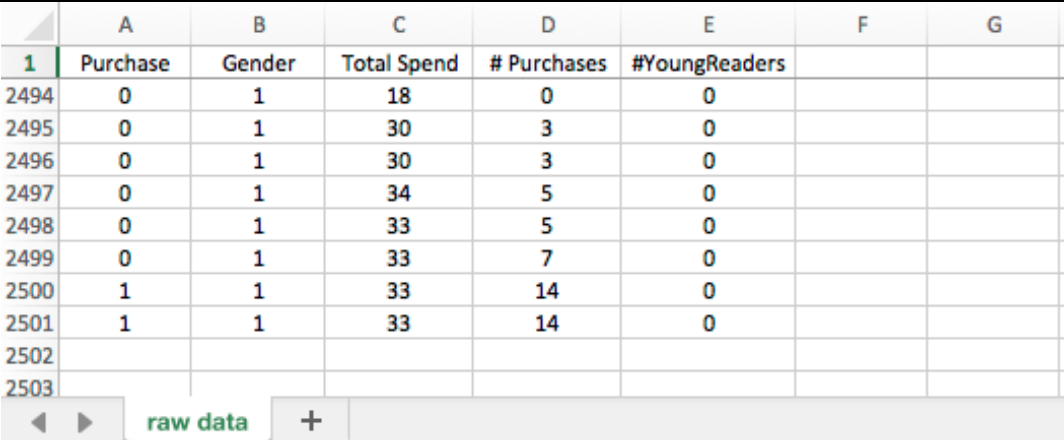
## Testing a Model Using a Holdout Sample

### Excel Step-by-Step How-to for Windows and Excel for Mac 2016 (v.16) or later

**Instructions:** Use this guide to test a regression model for a data set using a holdout sample. The process begins with creating a holdout sample, followed by creating a regression model from the data remaining after the holdout is removed. Finally, the regression model is tested against the holdout sample to verify that the model performs consistently.

**Data requirement:** One dependent variable and at least two independent variables, quantitative data

**Sample data:** Consumer data

Step	Windows Instructions + Screen Shot
	Create a Holdout
1. Organize your data so that each row represents an entry and the columns are attributes of that entry.	

2. Add random numbers immediately to the right of your data.

**=RAND()** *Note: this function requires no parameters*

Add the RAND function in the first blank cell to the right of your first row of data to generate a random number between 0 and 1. Copy this form down so the RAND function appears in all rows with data. Place the column head "RANDOM" in row 1.

	A	B	C	D	E	F
1	Purchase	Gender	Total Spend	# Purchases	#YoungReaders	RANDOM
2	1	1	113	13	5	=RAND()
3	1	1	268	11	0	0.4440328
4	1	1	308	12	1	0.16430375
5	1	1	249	3	2	0.90988658

You will notice that every time you complete an entry in the spreadsheet, the numbers in the RANDOM column will change. To avoid this, you want to replace the RAND function in each cell with the numbers in the cells. Select the RANDOM column then right click to Copy. Right click again and select Paste Special from the pop-up menu.

	A	B	C	D	E	F
1	Purchase	Gender	Total Spend	# Purchases	#YoungReaders	RA
2	1	1	113	13	5	0.7
3	1	1	268	11	0	0.4
4	1	1	308	12	1	0.5
5	1	1	249	3	2	0.5
6	1	1	44	43	25	0.1
7	1	1	250	11	2	0.8
8	1	1	140	40	22	0.3
9	1	1	280	17	9	0.1
10	1	0	90	23	11	0.4
11	1	0	103	13	5	0.9

Cut

Copy

Paste

**Paste Special.**

Insert Copied

Delete

Clear Contents

Format Cells...

Column Width

Hide

Select Values from the Paste Special menu.

	A	B	C	D	E	F
1	Purchase	Gender	Total Spend	# Purchases	#YoungReaders	RANDOM
2	1	1	113	13	5	0.6145644
3	1	1	268	11	0	0.22631056
4	1	1	308	12	1	0.60983527
5	1	1	249	3	2	0.90670842
6	1	1	44	43	25	0.16982801
7	1	1	250	11	2	0.58783187
8	1	1	140	40	22	0.62959035
9	1	1	280	17	9	0.65156647

#### Paste

☐ All

☐ Formulas

☒ Values

☐ Formats

☐ All using

☐ All exce

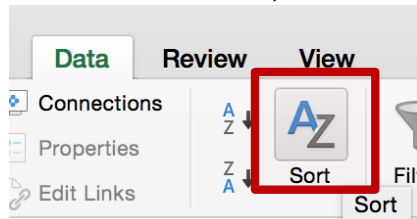
☐ Column

☐ Formula

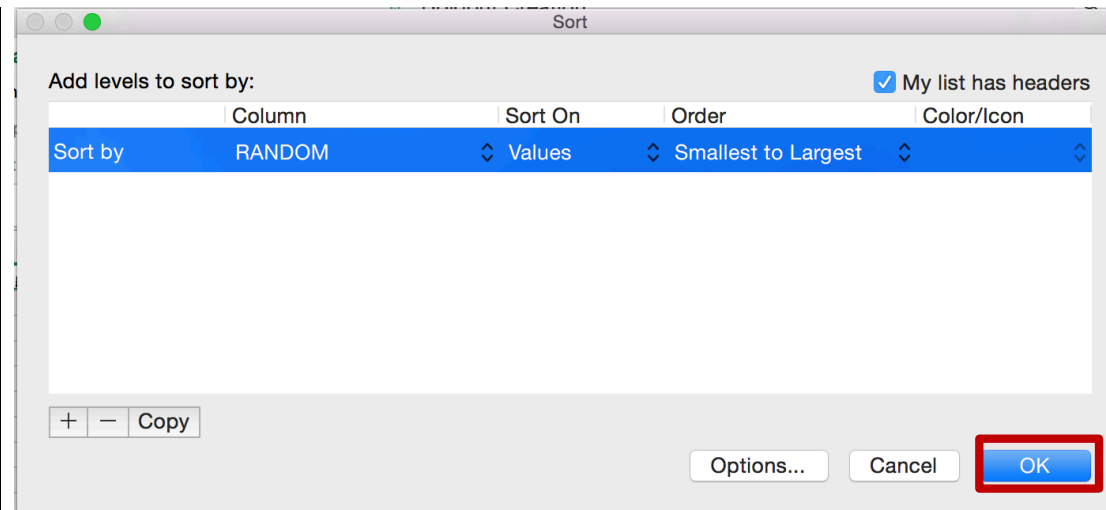
You should now have a random number to the right of your data in each row.

- Sort the rows by random number.

Select a cell in the RANDOM Column. From the Data menu, select Sort. Using the arrows, sort by Column RANDOM, click OK



E	F	G
#YoungReaders	RANDOM	
0	0.39194567	
0	0.27039337	



4. Remove a holdout sample from the data set.

Select the data from rows you want to use for a holdout sample. For example, if you had a 2500 row sample and wanted a 20% holdout, you would scroll down to row 2001 and select the last 500 rows of data (rows 2002 to 2501). Don't select the RANDOM column since it's not actually part of your original data set.

	A	B	C	D	E	F
1	Purchase	Gender	Total Spend	# Purchases	#YoungReaders	RANDOM
2498	0	1	33	5	0	0.92925508
2499	0	1	33	7	0	0.56056582
2500	1	1	33	14	0	0.75200245
2501	1	1	33	14	0	0.31887203
2502						

With the rows selected, right click and Cut the data, then click the + to add a new tab.

1	Purchase	Gender	Total Spend	# Purchases	#YoungReaders
2498	0	1	33	5	0
2499	0	1	33	7	0
2500	1	1	33	14	0
2501	1	1	33	14	0
2502					

raw data +

Right click in cell A1 of the new sheet and Paste the data into the sheet. Right click in the tab for the new sheet and select Rename. You will want to call this tab something like “Holdout” so you don’t get confused later about what these data are.

	A	B	C	D	E
497	0	1	33	5	0
498	0	1	33	7	0
499	1	1	33	14	0
500	1	1	33	14	0
501					
502					

raw data Shee Insert Sheet Delete Rename

You will now have a randomly selected set of observations from the original sample to use as your holdout sample. You can now use the observations remaining in the first tab to build your model or models.

5. Run a regression on the model-building (non-holdout) data

See the how-to guide for running a multiple regression if you need help running a regression.

Depending on your project needs, you may wish to develop several models and run several regressions, after which you compare models' effectiveness by testing each against the holdout.

	A	B	C	D	E	F	G
1	SUMMARY OUTPUT						
2							
3	<i>Regression Statistics</i>						
4	Multiple R	0.497691218					
5	R Square	0.247696548					
6	Adjusted R Square	0.24618817					
7	Standard Error	0.375794741					
8	Observations	2000					
9							
10	<i>ANOVA</i>						
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>	
12	Regression	4	92.7622335	23.1905584	164.21386	1.249E-121	
13	Residual	1995	281.7372665	0.14122169			
14	Total	1999	374.4995				
15							
16		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
17	Intercept	0.131110486	0.026508816	4.94592014	8.2128E-07	0.07912262	0.183098351
18	Gender	-0.06080402	0.017743296	-3.4268728	0.00062296	-0.0956014	-0.026006687
19	Total Spend	0.000255039	8.85462E-05	2.88028886	0.00401551	8.1386E-05	0.000428691
20	# Purchases	-0.00120176	0.00183629	-0.6544499	0.51289741	-0.004803	0.002399487
21	#YoungReaders	0.049258198	0.003463204	14.2233045	8.4271E-44	0.04246632	0.056050073
22							

*Note:* If you are following along with the supplied data set, your regression parameters will likely be slightly different than those shown here. The random sample selection yields different results each time, and as a result you will split the sample slightly differently than the split shown here.

6. Paste the raw data and a transposed copy of the coefficients for the independent variables in a new sheet.

Create a new sheet and rename it something like “analysis.” If you are testing several models, the tab name should also include a reference that helps you know which model it represents. Copy all the data remaining in the Raw Data tab and paste it into the Analysis tab.

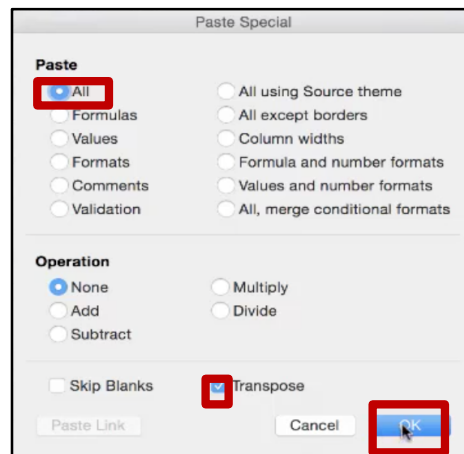
	A	B	C	D	E
1	Purchase	Gender	Total Spend	# Purchases	#YoungReaders
2	1	1	113	13	5
3	1	1	268	11	0
4	1	1	308	12	1
5	1	1	249	3	2
6	1	1	44	43	25

◀ ▶ raw data regression1 **analysis1** holdout

Now go back to the regression tab. Select and copy the variable names and coefficients from the regression report.

	<i>Coefficients</i>
Intercept	0.131110486
Gender	-0.06080402
Total Spend	0.000255039
# Purchases	-0.00120176
#YoungReaders	0.049258198

Right click in a cell of the analysis sheet that's near the top and a few columns to the right of your data, and Paste Special the attributes and the coefficient values from the regression sheet into your new sheet. A pop-up window will appear. In the paste section select All, and in the Operation section select None. Click to activate the Transpose box. Click the OK button to paste the cells into your new sheet.



Pasting with Transpose will change the orientation of your regression data so that the columns become rows. This will make your results easier to read as you continue through your analysis.

Intercept	Gender	Total Spend	# Purchases	#YoungReaders
0.13111049	-0.060804	0.00025504	-0.0012018	0.049258198



7. Apply coefficients to the sample to obtain a predicted value for each observation in the sample.

Now you'll work in your Analysis tab to make a determination based on your coefficients. In the first empty cell to the right of your coefficients, add a column head Predicted Y.

The predicted Y value will be the intercept plus the sum product of the coefficient values of all four attributes. To enter this function, begin by typing an equal sign in the first free cell in the Predicted Y column. After the equal sign coefficient value under the "Intercept" label to insert the number after the equal sign. Add a plus sign and then type "sumproduct" Under the regression data, highlight the coefficient values for the remaining attributes and hit the F4 key to lock the data and add a comma. Under the original data, highlight the four attribute's corresponding values listed in the first numerical row. *Do not* hit the F4 key to lock this data. Hit the enter key to complete the equation and reveal the predicted Y value.

SUMPRO... ✖ ✔ <i>fx</i> = $\$H\$2 + \text{SUMPRODUCT}(\$I\$2:\$L\$2, B2:E2)$												
	A	B	C	D	E	F	G	H	I	J	K	L
1	Purchase	Gender	Total Spend	# Purchases	#YoungRead	PredictedY		Intercept	Gender	Total Spend	# Purchases	#YoungReaders
2	1	0	360	58	33	<i>=<math>\\$H\\$2 + \text{SUMPRODUCT}(\\$I\\$2:\\$L\\$2, B2:E2)</math></i>		-0.060804	0.00025504	-0.0012018	0.049258198	

Double-click on the small box on the bottom right corner of the cell to insert the formula into the entire Predicted Y column.

8. Evaluate outcomes of decisions based on predicted values.

You can now investigate how well the model predicts. In the example shown, the dependent variable is a binary purchase result, so the model will be used to predict purchases.

Using the percentile function in Excel, calculate the 100<sup>th</sup>, 75<sup>th</sup>, 50<sup>th</sup> and 25<sup>th</sup> percentiles for the data set – the 100<sup>th</sup> percentile is the max value with the 75<sup>th</sup> percentile being the number that 75% of observations are less than. In empty cells to the right of the data, type the numbers 1, .75, .5 and .25. In the cells adjacent to each percentile, enter the corresponding percentile function.

Begin by typing **=percentile** in the cell to the right of the percentile cell with the value of 1. Complete the function by selecting the entire Predicted Y column and hit the F4 key to lock the data. Then select the adjacent percentile cell in Percentile column. Hit the enter key to reveal the predicted Y for the percentile.

	F	G	H	I	J	K
ad	Predicted Y		Intercept	Gender	Total Spend	# Purchases
5	0.329793951		0.13111049	-0.060804	0.00025504	-0.0012018
0	0.12543747					
1	0.183695455					
2	0.228722209		Percentile	PredictedY		
25	1.261307463		1	=PERCENTILE(\$F\$2:\$F\$2001,H6)		
2	0.219363171		0.75	0.25898832		
2	1.141621858		0.5	0.18396638		
9	0.564611161		0.25	0.13446894		
11	0.668263676					

Select this first predicted y result, and double-click the small box on the bottom right corner to insert the formula into the remaining three predicted y cells.

	A	B	C	D	E	F	G	H	I	J
1	Purchase	Gender	Total Spend	# Purchases	#YoungRead	Predicted Y		Intercept	Gender	Tot
2	1	1	113	13	5	0.329793951		0.13111049	-0.060804	0.0
3	1	1	268	11	0	0.12543747				
4	1	1	308	12	1	0.183695455				
5	1	1	249	3	2	0.228722209				
6	1	1	44	43	25	1.261307463				
7	1	1	250	11	2	0.219363171				
8	1	1	140	40	22	1.141621858				
9	1	1	280	17	9	0.564611161				
10	1	0	90	23	11	0.668263676				

In the cell right adjacent to the percentile, you will use SUMIF to sum the purchase column to determine how many purchases were made by the corresponding percentiles. Begin by typing =SUMIF into the blank cell. Select the entire Predicted Y column, excluding the title cell, and hit F4 to lock the data. Then type "<=" including the quotation marks. Select the adjacent percentile cell and hit F4 to lock the data. Then highlight the entire purchases column, excluding the title cell, and hit the enter key. Click on the small box on the bottom right corner of the cell to insert the formula into the remaining three cells.

## 9. Plan a course of action

You should now have a set of outcomes that represents the influence of different attributes. You may need to do additional calculations to establish a connection between your outcomes and specific goals or key performance indicators.

Percentile	PredictedY	#<=	#between	MAIL	Purchase	Revenue	Costs	Profit
1	1.77874289	499	304	500	304	2432	1740	692
0.75	0.25898832	195	90	500	90	720	2275	-1555
0.5	0.18396638	105	66	500	66	528	2335	-1807
0.25	0.13446894	39	39	500	39	312	2402.5	-2090.5

This table shows predicted Y being used to estimate the profitability of targeting different percentiles.

10. Test your planned response with the holdout sample.

As the table shows, targeting the top 25% of customers (as determined by coefficients for the attributes tested in the model) should be profitable. Targeting any quartile below this top 25%, however, is likely to generate a loss.

You can now test how well the model works with the holdout sample. To test your model with the holdout, you apply the regression model to the holdout data points that you randomly removed. You would apply the same treatment as with the larger sample---say, sorting them into quartiles and estimating the profitability of targeting these quartiles. Hopefully your model performs similarly.

Percentile	PredictedY	#<=	#between	MAIL	Purchase	Revenue	Costs	Profit
1	1.48715031	124	70	500	70	560	2325	-1765
0.75	0.27249767	54	23	500	23	184	2442.5	-2258.5
0.5	0.1908309	31	19	500	19	152	2452.5	-2300.5
0.25	0.13676834	12	12	500	12	96	2470	-2374

In this case, the top quartile of the holdout is profitable just as it was with the main model-building sample. So, the holdout test lends support to the validity of conclusions based upon the model-building sample.