Multiple Regression

Extra Material

Michael Black

This document refers to the Excel document "rdchem.xlsx" *on eCampus*
In this document, we want to answer the following questions:

1. Does a company's sales and spending on research and development (R&D) affect its profits?
2. Does a company's sales and spending on research and development (R&D) affect its profit margins?

We will walk through answering #1, and you will get to practice answering #2.

How to perform multiple regression using Data Analysis ToolPak

Open the associated Excel document, and use the sheet entitled "Example". We are interested in the following model:

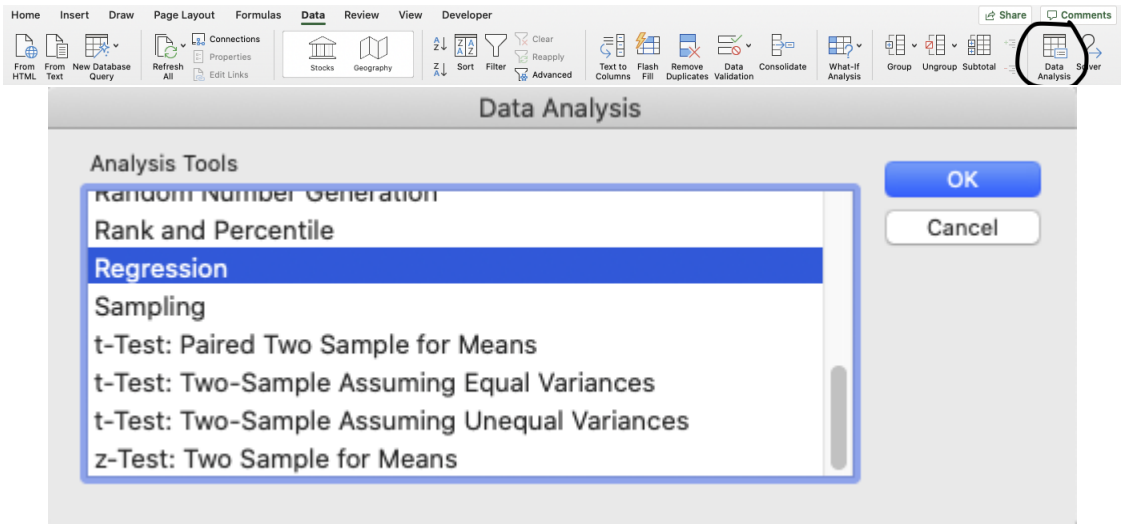
$$profits_i = \beta_0 + \beta_1 sales_i + \beta_2 rd_i + \varepsilon_i$$

where $profits_i$ is the profits of firm i, rd_i is the amount firm i spent on research and development, and $sales_i$ is the gross sales in USD for firm i. Let's estimate our model:

1. Open the dataset:

	A	B	C	D
1		profits	sales	rd
2	1	186.899994	4570.2002	430.600006
3	2	467	2830	59
4	3	107.400002	596.799988	23.5
5	4	-4.3000002	133.600006	3.5
6	5	8	42	1.70000005
7	6	47.2999992	390	8.39999962
8	7	0.89999998	93.9000015	2.5
9	8	77.4000015	907.900024	39.9000015
10	9	2563	19773	1136
11	10	4154	39709	1428
12	11	93.6999969	2936.5	45.2999992
13	12	355.399994	2513.80005	65.1999969
14	13	45.2999992	1124.80005	20.2999992
15	14	4.0999999	921.599976	15.6000004
16	15	132.399994	2432.6001	74
17	16	329	6754	147.5
18	17	289.899994	1066.30005	29.2000008
19	18	163.199997	3199.8999	92.1999969
20	19	32.5999985	150	3.70000005
21	20	83.8000031	509.700012	19.3999996
22	21	271.200012	1452.69995	74.4000015
23	22	809	8995	612
24	23	215	1212.30005	45.2999992
25	24	154.800003	906.599976	11
26	25	116	2592	66
27	26	23	201.5	10.3999996
28	27	60.5	2617.80005	48.9000015
29	28	18.3999996	502.200012	6.0999999
30	29	313	2824	178.199997
31	30	6.4000001	292.200012	3
32	31	626	7621	191
33	32	105.800003	1631.5	26

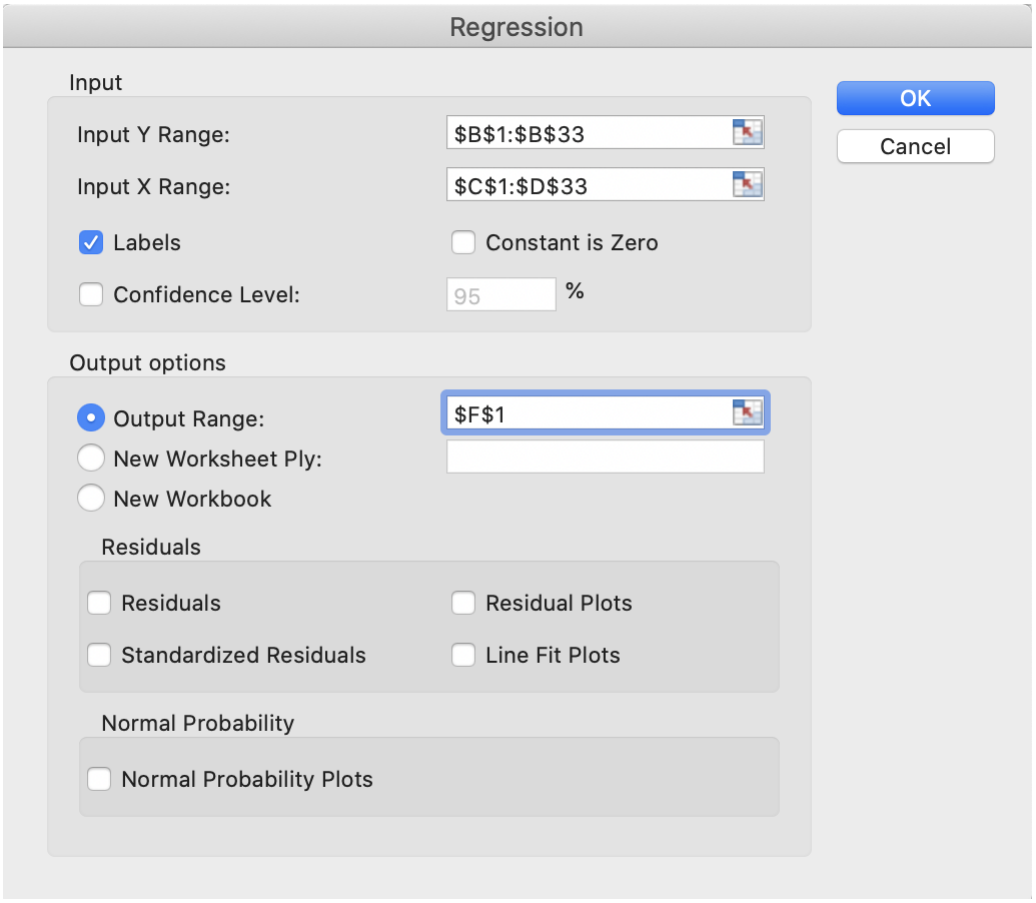
2. Open the Data Analysis ToolPak, click on "Regression":



3. Select the data according to the model. Our model is:

$$profits_i = \beta_0 + \beta_1 sales_i + \beta_2 rd_i + \varepsilon_i$$

so our y (dependent) variable is $profits$, and our x (independent) variables are $sales$, and rd. Be sure to select the first row (column headings) as well, and click the "Labels" button. Finally, place the output in the same sheet. I told Excel to start the output (results) at F1, but you can choose to put it anywhere. When you are ready, press "OK":



4. That's it! You have performed a regression.

How to interpret a multiple regression

Recall once more the model we are estimating:

$$profits_i = \beta_0 + \beta_1 sales_i + \beta_2 rd_i + \varepsilon_i$$

The regression output for this example is as follows:

F	G	H	I	J	K	L	M	N
SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.984111849							
R Square	0.968476132							
Adjusted R Square	0.966302072							
Standard Error	152.5338412							
Observations	32							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	2	20729070.86	10364535.4	445.468938	1.69931E-22			
Residual	29	674730.6085	23266.5727					
Total	31	21403801.47						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-37.55628659	30.24957292	-1.2415477	0.22435198	-99.42360979	24.3110366	-99.42361	24.3110366
sales	0.086716568	0.011517031	7.52942007	2.6725E-08	0.063161594	0.11027154	0.06316159	0.11027154
rd	0.512720438	0.269197498	1.90462557	0.06679071	-0.037850264	1.06329114	-0.0378503	1.06329114

Now, let's move through the results as you should with every regression:

1. Look at the $\bar{R}^2$. Remember we are looking at the *adjusted*-R^2 because we have a multivariate regression. For a univariate regression, we would just use the R^2. For this regression, the $\bar{R}^2$ is **0.966**, meaning that **sales and R&D explain 96.6% of the variation of profits**, given the data we have. That's pretty good. No red flags.

2. Ensure the model is jointly significant. That is, we want to make sure that we have at least one significant slope coefficient in the model. Otherwise, profits could not be explained by sales and R&D. To ensure the model is jointly significant, we need to **run an F-test**. The hypothesis test is specifically:

$$\begin{aligned} H_0 &: \hat{\beta}_1 = \hat{\beta}_2 = 0 \\ H_1 &: \hat{\beta}_1 \quad or \quad \hat{\beta}_2 \neq 0 \end{aligned}$$

   In other words:

H_0: ALL SLOPES COEFFICIENTS ARE INSIGNIFICANT
H_1: AT LEAST ONE OF THE SLOPE COEFFICIENTS IS SIGNIFICANT

   The Excel regression gives us the F-statistic. We could calculate the F-critical value ourselves, or just look at the p-value associated with the F-test. In this case, **the p-value associated with the F-test is 1.699**$\times 10^{-22}$. This number is much lower than $\alpha = 0.05$, so we **reject the null hypothesis**, meaning that for this model, we have joint significance. That is, AT LEAST ONE OF THE SLOPE COEFFICIENTS IS SIGNIFICANT. Good, this means our model is not totally useless.

3. Next, we want to determine which parameters are significant. We do this by performing a t-test **for every coefficient**:

$$\begin{aligned} H_0 &: \hat{\beta}_0 = 0 \\ H_1 &: \hat{\beta}_0 \neq 0 \end{aligned}$$

$$H_0 \quad : \quad \hat{\beta}_1 = 0$$
$$H_1 \quad : \quad \hat{\beta}_1 \neq 0$$

$$H_0 \quad : \quad \hat{\beta}_2 = 0$$
$$H_1 \quad : \quad \hat{\beta}_2 \neq 0$$

We are interested in a two-tailed test. That is, we are interested in whether each coefficient is significantly different from zero. It could be significantly less than or greater than zero. We don't care. We care that it is simply different from zero. Remember the rules we use to reject or fail to reject:

- If the absolute value of the t-stat > t-critical value, we **reject the null hypothesis**. Otherwise, we **fail to reject the null hypothesis.**
- If the p-value $< \alpha$, we **reject the null hypothesis**. Otherwise, we **fail to reject the null hypothesis.** The value for α is subjectively chosen, but is usually 0.05.
- If the number 0 is not the confidence interval, we **reject the null hypothesis**. Otherwise, we **fail to reject the null hypothesis.**

We could look directly at the p-value or confidence intervals and make our hypothesis test decision immediately. Just follow the rules above! If we want to use the t-stat to make our decision, we have to calculate the t-critical value:

```
=TINV(alpha, degrees of freedom)
```

We set $\alpha = 0.05$, and the degrees of freedom is the number of observations minus the number of explanatory variables, minus one: $n - k - 1$. We could also just look at the following cell in the regression output:

ANOVA					
	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	20729070.86	10364535	445.46894	1.69931E-22
Residual	29	674730.6085	23266.573		
Total	31	21403801.47			

So our t-critical value would be calculated as:

```
=TINV(0.05, 29)
```

So which variables are significant (significantly different from zero)? In this case, **only $\hat{\beta}_1$ is significantly different from zero**:

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	-37.55628659	30.24957292	-1.2415477	0.224352	-99.42360979	24.311037	-99.42361	24.311037
sales	0.086716568	0.011517031	7.5294201	2.673E-08	0.063161594	0.1102715	0.0631616	0.1102715
rd	0.512720438	0.269197498	1.9046256	0.0667907	-0.037850264	1.0632911	-0.0378503	1.0632911

The insignificant variables may be interpreted as being zero. There is a chance these coefficients may be different from zero, but statistically we can't tell them apart from zero. That means that our initial model:

$$profits_i = \beta_0 + \beta_1 sales_i + \beta_2 rd_i + \varepsilon_i$$

Can be written as:

$$profits_i = 0 + (0.08)sales_i + (0)rd_i + \varepsilon_i$$
$$\Rightarrow profits_i = (0.08)sales_i + \varepsilon_i$$

4. Finally, we can *interpret* the estimated model. We are interested in the effect of our explanatory variables on our dependent variable. So we are interested in how the dependent variable changes for a small change in the explanatory variables:

$$\frac{\partial profits_i}{\partial sales_i} = ?$$

$$\frac{\partial profits_i}{\partial rd_i} = ?$$

Given our model and regression results, you should be able to see that:

$$\frac{\partial profits_i}{\partial sales_i} = 0.08$$

$$\frac{\partial profits_i}{\partial rd_i} = 0$$

Profits, sales, and R&D expenditures are measured in thousands of USD. Our model suggests, then, that:

- A \$1,000 increase in sales is associated with a \$80 increase in profits, holding R&D constant.
- A \$1,000 increase in R&D is associated with a \$0 increase in profits, holding sales constant.

Why? All linear regressions follow the same pattern for interpretation:

- A [**unit**] increase in [**explanatory variable**] is associated with a [$\hat{\beta}_j$] [**increase/decrease**] in [**dependent variable**], holding [**other explanatory variables**] constant.

The unit for all of sales is thousands of USD: \$1,000. What is 0.08 thousand dollars? \$80.

Practice

Try estimating a model yourself! Use the sheet called "Practice". Estimate the following model:

$$profmarg_i = \beta_0 + \beta_1 sales_i + \beta_2 rd_i + \varepsilon_i$$

where the explanatory variables are the same, but now the dependent variable is $profmarg$: the profit margin of firm i. Answers on next page.

[Spoiler]: Practice answers

SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.061181888
R Square	0.003743223
Adjusted R Square	-0.064964141
Standard Error	7.473476589
Observations	32

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	2	6.08580192	3.04290096	0.05448067	0.947073497
Residual	29	1619.73272	55.8528523		
Total	31	1625.81852			

	Coefficients	*tandard Erro*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	9.84245564	1.4820939	6.64091233	2.791E-07	6.81123327	12.873678	6.81123327	12.873678
sales	-0.000180809	0.00056428	-0.3204232	0.75094501	-0.001334898	0.00097328	-0.0013349	0.00097328
rd	0.004341113	0.01318947	0.32913463	0.74441922	-0.022634391	0.03131662	-0.0226344	0.03131662

Remember the model is:

$$profmarg_i = \beta_0 + \beta_1 sales_i + \beta_2 rd_i + \varepsilon_i$$

- Terrible $\bar{R}^2$: our explanatory variables are describing almost none of the variation in profit margins.
- Failure of joint significance: we fail to reject the null of the F-test, meaning none of the slope coefficients are significant. We could stop our interpretations here.
- Sure enough, none of the individual coefficients are significant. That means that our model can be written as:

$$\begin{aligned} profmarg_i &= 9.84 + (0)sales_i + (0)rd_i + \varepsilon_i \\ profmarg_i &= 9.84 + \varepsilon_i \end{aligned}$$

  ...meaning according to our model and data, profit margins are, on average, 9.84%[1], but sales and R&D expenditures have no effect on profit margins.

[1] This is because the profit margin variable is measured in percentage points.