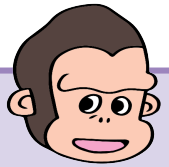
CONTENTS

第 3 章

相関

売上を上げているのは どんな要素を持った人？

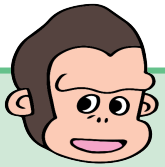
第 4 章



散布図

入社時面接の点数と売上の 関係を図にして理解しよう！

第 5 章



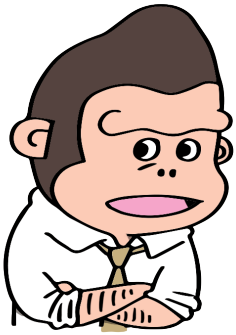
回帰分析

将来の売上予測にもっとも影響を与える要因を探せ！

この本の登場人物

山田君

IT系企業に新設されたデータ分析チームに配属された、1年目の社員。不器用ではあるが、自分で考えてコツコツと挑戦することは得意。好きな言葉は「人間万事塞翁が馬」。



ゴリラ部長

山田君の上司。部下を育てることに情熱を持つ。昔はゴリゴリの体育会系だったが、伝わりやすい、そして自分でできるようになるコミュニケーションを心がけている。ラーメンやハンバーガーなど、B級グルメの大ファンで趣味は食べ歩き。

はじめまして

みなさん、今日からよろしく！　最近ネットニュースやテレビで「データ」や「統計」という言葉を見ない日はないよね。社内にある宝物のデータを分析して意味合いを見つけて、弊社の問題を解決するということが我々のチームのミッションになります。頑張りましょう。そして、今期もっとも取り組みたい問いがこちら。

そう言うと、ゴリラ部長はホワイトボードに次の文字を書いた。

「活躍する営業は採用時にどんな要素を持っているか？　を可視化する！」

「具体的なゴール：面接終了後にわかっているデータ（例：面接の点数、適性検査のスコア、起業経験の有無など）から、なるべく活躍する可能性が高い人を採用する意思決定に役立つモデルを作る」

弊社の競争力の源泉は、優秀な営業パーソンを多く採用し営業組織のレベルを上げること。もちろん、時代の変化によって求められるスキルは変わっていく。けれども役員陣と検討した結果、

活躍する営業の要素を定義して、採用の段階からそういった人を見抜く

ということが、分析チームが全社の売上に貢献できる期待値が一番高そうだという結論になりました。そのために、まずはみんなに分析に慣れてもらい、その後、ゴールを目指して実際の分析を進めていってもらいたい。それでは、新しく配属されたみなさん、自己紹介をお願いします。

（何人かの自己紹介のあと山田君の番が来る）

山田と申します！　大学時代はサッカーばかりの4年間を過ごしており、正直データ分析どころか統計すら挫折した経験があります。ただ、試合のデータから相手にどうやって勝つかを考えたり、コツコツとトレーニングを積むことで能力を改善してきた経験を生かして、一生懸命がんばりたいと思います。どうかよろしくお願いいたします！

（みんなから拍手）

よろしく！

データ分析で一番大切なこと

ところで山田君。早速だけど、エクセルの基本的な操作は学んできたかな？

はい。とはいっても内定者研修での事前課題+αでやってきた、合計や平均を計算するという程度ですが…。

それで十分。それでは、最初の仕事をお願いするね。ここに第一営業部と第二営業部、全体では50名ずついるチームだが、それぞれから抽出してきた10名の売上データがある。これらのデータがどうなっているか、気になる点を教えてほしい。

わかりました。

えーっと、わかったと言っているけど、具体的にどんなことをすればいいかイメージできてる？

えっ。

山田君。データ分析では、やるべきことを明確にしないで分析を開始すると、時間をすごく浪費することになる。

はい。

まだ始めたばかりなので難しいと思うけど、このあたりの心構えに関して少し話をします。まずデータ分析において私がもっとも大切だと思っていることを伝えたいんだけど、ちなみになんだと思う？

えっと。うーんと（こういう時の部長、まじゴリラっぽくて威圧感すごいな）。データ分析なので、数字のセンスとかですかね？

いや、分析にセンスなんていらない。大事なのは明確な目的。

も、く、て、き？

そう。データ分析をする時には、目的を明確にすることが重要なんだ。データ分析を開始してたくさんデータに触れていると、こんなこともわかる、あんなこともわかると面白くなり、無闇に時間が過ぎてしまうことがよくある。私も始めたての頃はこれによくハマって、翌朝まで分析して、面白かったけど結局なにも進んでいなかったということがよくあった。データ分析をする時は、何を目的にして、そのためにどんなデータが出ればいいのか、ポストイットに書いてPCに貼り付けておくとよいよ。先ほど山田君が簡単に「わかりました」と言ったのでちょっと注意が必要かなと思って聞いてみたけど、目的の整理や、分析でやるべきことが言語化できていなかったよね。

その通りです。

そんなにシュンとならなくて大丈夫。毎回データ分析を始める人には同じように接していて、いつも同じような反応になるから。正直な話、私はもっとひどかった。具体的に次のアクションがイメージできていない場合は、まず考えて、順序を言語化してみる。そして難しかったら聞くようにしてください。

ありがとうございます。それで今回の分析の目的は？

今回は、山田君にチームの作業に慣れてもらうことが第一。それから、第一営業部と第二営業部のどちらがいいチームなのかを数量的に明らかにできたら嬉しいな。あと、最初のうちは次の作業や方法がわからなくなったら30分程度自分で考え、検索などで調べても行き詰まったら早めに報告してください。

わかりました！

山田君は、メモにしっかりとこう書いた。

・データ分析をする時は目的をしっかり意識する（できればポストイットやメモに書いて、分析中はPCの画面隅に貼っておく）

・今回のデータ分析の目的

1） 今のチームの仕事に慣れること

2） 第一営業部と第二営業部のどちらがいいチームなのかを数量的に明らかにすること

（席に戻ると、部長からエクセルのデータが送られてきていた）

よし、それじゃあ最初に目的を考えよう。部長は仕事に慣れろと言っていたので、まずは「自分で分析→部長にチェックしてもらう→フィードバックをもらう→追加で分析する」という一通りのサイクルを回してみるのがよさそうかな。どうせ完璧にはできないだろうし、なるべく早くやってみて、部長の意見を早めにもらうことを心がけよう。あとは「いいチーム」の定義が難しいな…。ちょっとよくわからないから、まずはデータを見てみよう。

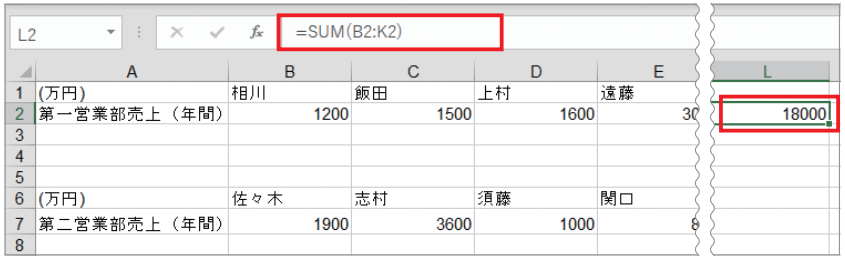
（そこにはそれぞれの営業の名前と売上だけが入っていた）

	A	B	C	D	K
1	(万円)	相川	飯田	上村	近藤
2	第一営業部売上（年間）	1200	1500		1800
3					
4					
5					
6	(万円)	佐々木	志村	須藤	戸口
7	第二営業部売上（年間）	1900	3600		3700
8					
9					

データとしてあるのは売上だけか…。ということはいったん

- **売上が高い営業さん＝活躍している**
- **売上が高いチーム＝いいチーム**

と定義して考えてみよう。まずはエクセルで、第一営業部の売上のデータを合計していこう。sum関数を使って…。



	A	B	C	D	E	L
1	(万円)	相川	飯田	上村	遠藤	
2	第一営業部売上（年間）	1200	1500	1600	30	18000
3						
4						
5						
6	(万円)	佐々木	志村	須藤	関口	
7	第二営業部売上（年間）	1900	3600	1000	8	
8						

エクセル関数

sum関数

複数のセルの合計を計算する際に、1つ1つのセルを足していってもよいのですが、sum（始まりのセル：終わりのセル）で選択すると一気にその範囲の合計値を出してくれます。またsum（）のかっこの中を「,」でつなげば、1つずつのセルを足していくことも可能です。

次は、売上の平均を出してみよう。このプロセスをまとめてできるのがaverage関数だよな。よし！　では平均の計算を関数を使ってやってみよう。

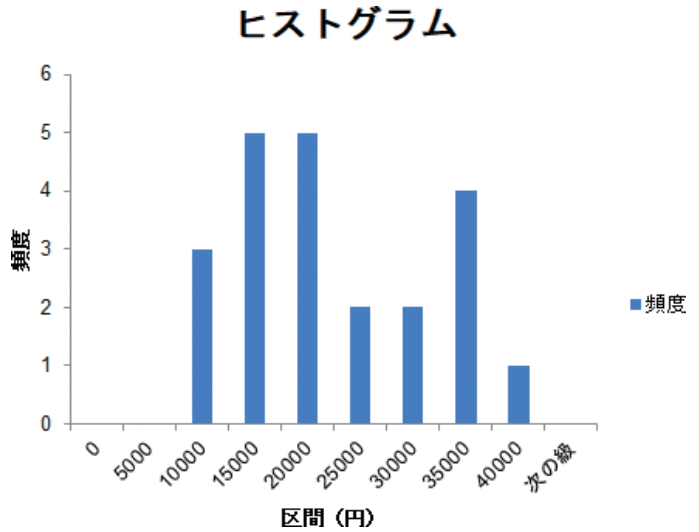
ヒストグラムとは？

寸胴のスープをきちんと混ぜて均等に掬えば、丼のスープの味をなるべく同じにすることができる。それと同じように、10人の社員のデータが、全体をきちんとかき混ぜて抽出したデータならば第一営業部と第二営業部の傾向をある程度推測できるはず。部長から第一営業部と第二営業部のデータをもらってきたけど、これを使ってどう考えたらいいのだろう。数字だけ見ていても、部分と全体が似ているのか似ていないのかよくわからないな…。そういえば部長がヒストグラムとおっしゃっていたから、まずはヒストグラムに関して調べてみよう。

ヒストグラム

ヒストグラム（英語: histogram）とは、縦軸に度数、横軸に階級をとった統計グラフの一種で、データの分布状況を視覚的に認識するために主に統計学や数学、画像処理等で用いられる。柱状図、柱状グラフ、度数分布図ともいう。（Wikipedia より）

なんか、わかるようでわからないな。ヒストグラムの例を見てみるとこんな感じか。



このグラフでは、ある会社の社員が1 ヶ月に自由に使えるお金がどれくらいあるか、またその範囲で使える人は何人いるかという関係が、それぞれX軸とY軸に表されている。X軸（横軸）に1 ヶ月あたりの使える金額が5000円刻みで設定されていて、Y軸（縦軸）にはX軸の範囲内で使える人が何人いるかが表されている。例えば5001円～ 10000円使える人は全部で3人いるということが読み取れるのか。金額範囲別の該当者が何人いるかということを視覚的に整理できるんだな。

今回の第一営業部の売上データを見てみよう。

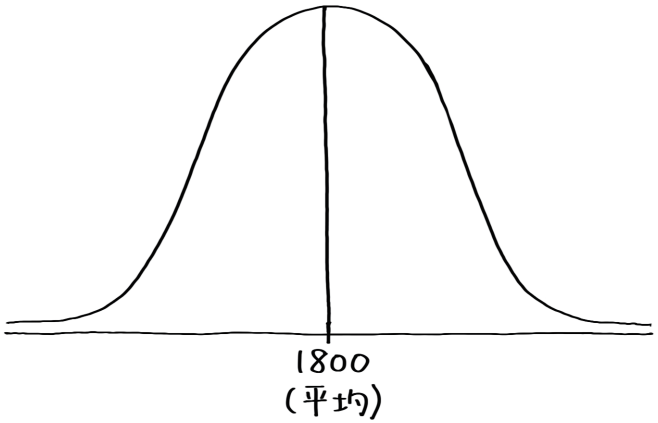
	A	B	C	D	E	F	G	H	I	J	K
1	（万円）	相川	飯田	上村	遠藤	大森	柿沼	北村	栗原	毛塚	近藤
2	第一営業部売上（年間）	1200	1500	1600	3000	2000	1900	1700	2100	1200	1800
3	売上範囲	0	500	1000	1500	2000	2500	3000			
4											
5											

その通り。そして47.72%つまり.4772を探すと？

Zが2.00のところにありました！

Z	0	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
0.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
0.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
0.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
0.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
0.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
0.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
0.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
0.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
0.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441
1.6	.4452	.4463	.4474	.4484	.4495	.4505	.4515	.4525	.4535	.4545
1.7	.4554	.4564	.4573	.4582	.4591	.4599	.4608	.4616	.4625	.4633
1.8	.4641	.4649	.4656	.4664	.4671	.4678	.4686	.4693	.4699	.4706
1.9	.4713	.4719	.4726	.4732	.4738	.4744	.4750	.4756	.4761	.4767
2.0	.4772	.4778	.4783	.4788	.4793	.4798	.4803	.4808	.4812	.4817
2.1	.4821	.4826	.4830	.4834	.4838	.4842	.4846	.4850	.4854	.4857
2.2	.4861	.4864	.4868	.4871	.4875	.4878	.4881	.4884	.4887	.4890
2.3	.4893	.4896	.4898	.4901	.4904	.4906	.4909	.4911	.4913	.4916
2.4	.4918	.4920	.4922	.4925	.4927	.4929	.4931	.4932	.4934	.4936
2.5	.4938	.4940	.4941	.4943	.4945	.4946	.4948	.4949	.4951	.4952
2.6	.4953	.4955	.4956	.4957	.4959	.4960	.4961	.4962	.4963	.4964
2.7	.4965	.4966	.4967	.4968	.4969	.4970	.4971	.4972	.4973	.4974
2.8	.4974	.4975	.4976	.4977	.4977	.4978	.4979	.4979	.4980	.4981
2.9	.4981	.4982	.4982	.4983	.4984	.4984	.4985	.4985	.4986	.4986
3.0	.4987	.4987	.4987	.4988	.4988	.4989	.4989	.4989	.4990	.4990
3.1	.4990	.4991	.4991	.4991	.4992	.4992	.4992	.4992	.4993	.4993
3.2	.4993	.4993	.4994	.4994	.4994	.4994	.4994	.4995	.4995	.4995
3.3	.4995	.4995	.4995	.4996	.4996	.4996	.4996	.4996	.4996	.4997
3.4	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4997	.4998
3.5	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998	.4998
3.6	.4998	.4998	.4999	.4999	.4999	.4999	.4999	.4999	.4999	.4999
3.7	.4999	.4999	.4999	.4999	.49991	.49992	.49992	.49992	.49992	.49992

それじゃあ、Z＝2.00がどういうことを説明しよう。それには正規分布の図を書いた方がよいね。



正規分布の問題を解く場合は、最初に必ずこの図を書くとよいです。そして、正規分布の問題を解く時にとても大事なことは、この真ん中の点がゼロじゃないということなんだ。ここは平均。

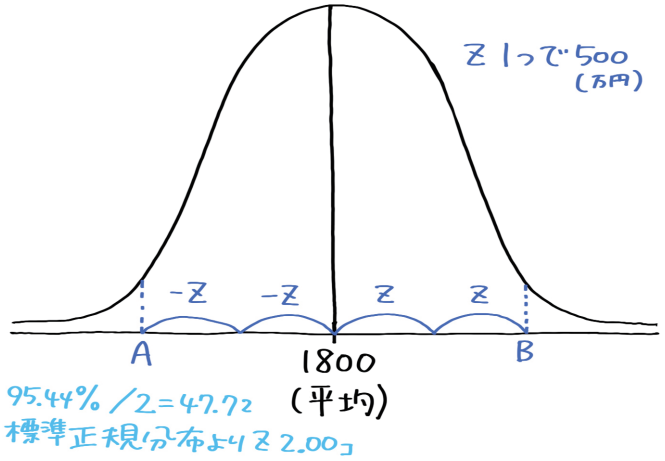
でも、時折0と書いてあるようなことがないでしょうか？

それは、平均から標準偏差で何個分離れているかを表した0であることが多いね。標準正規分布表はそうなっている。平均から標準偏差0個分離れているというのはつまり？

平均ということですね。

その通り。標準正規分布表で片側47.72%になるのは、「Z＝2.00」の時だった。つまり、平均からプラス方向にZ2個分進んだ時という意味になる。そして標準正規分布表は左右対称なので、マイナス方向にZ2個分進んだ時でもある。今回Z1個はいくつになるか

と言うと、Zは1標準偏差分なので問題で定義した値の500万円になる。



だから

A（平均からマイナス方向）：平均1800－500（Z1個つまり1標準偏差分）＊2＝800（万円）
B（平均からプラス方向）：平均1800＋500（Z1個つまり1標準偏差分）＊2＝2800（万円）

となって、答えは

第一営業部の売上平均は95.44％の確率で800万円～2800万円の間になる

ということになる。
逆に言うと、第一営業部の平均売上が2800万円以上になる可能性は2.28%（50%−47.72%）しかないということになるね。

社長がハッパかけて平均売上2800万以上いったらボーナスだ！とか言っても、可能性で考えるとすごく低いってことなんですね。

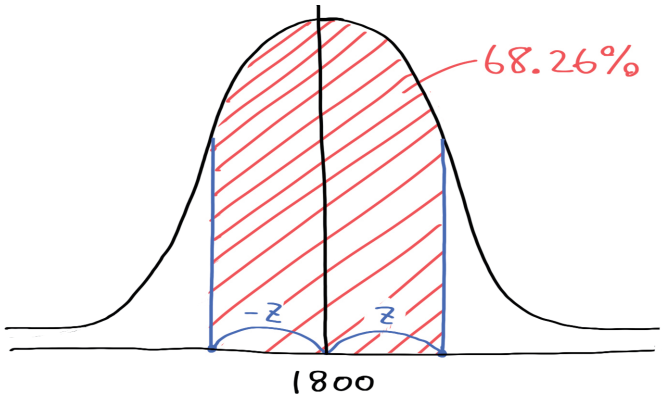
そうだね。そういうだいたいの基準感が持てるようになると、ビジネスだと強い。大外しがなくなるからね。

わかりました！

では山田君、残りの問題を1人でやってみてください。

問題❷　第一営業部の売上平均は、68.26％の確率でC ～ D万円の間になる？

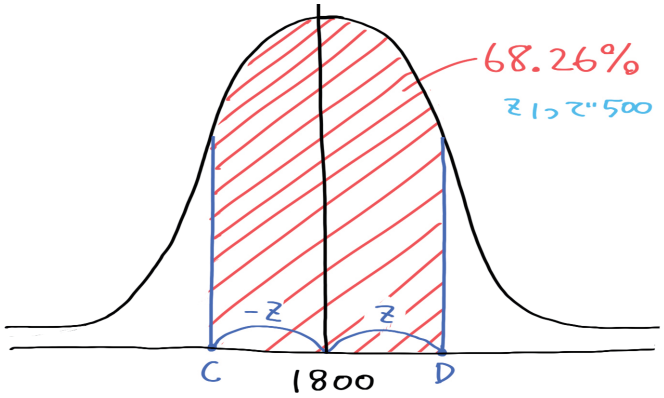
部長が言っていた通り、まずは絵を書いてみよう。



まず真ん中は平均！　それで、今回は68.26%になる時か。

標準正規分布表には68.26%は存在していない。なぜなら左右対称で半分の50%しか示さないものだから。そこで68.26%を2で割

るんだよな。すると34.13％になる。.3413は正規分布表のZ＝1.00のところにある。だから真ん中である平均から標準偏差±1つずつ進んだところが、求めるべきCとDになる。



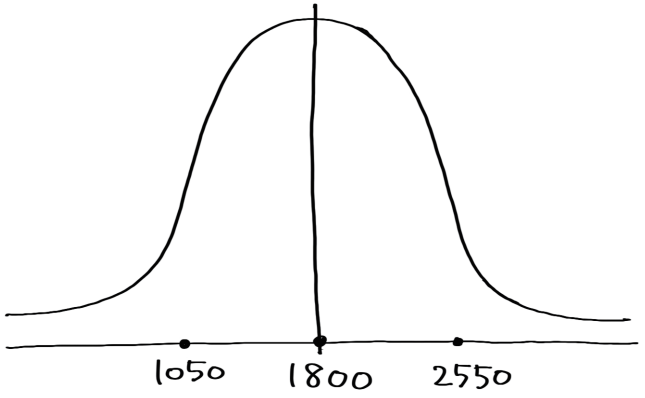
つまり

C：平均1800－500（Z1個つまり1標準偏差分）＊1＝1300（万円）
D：平均1800＋500（Z1個つまり1標準偏差分）＊1＝2300（万円）

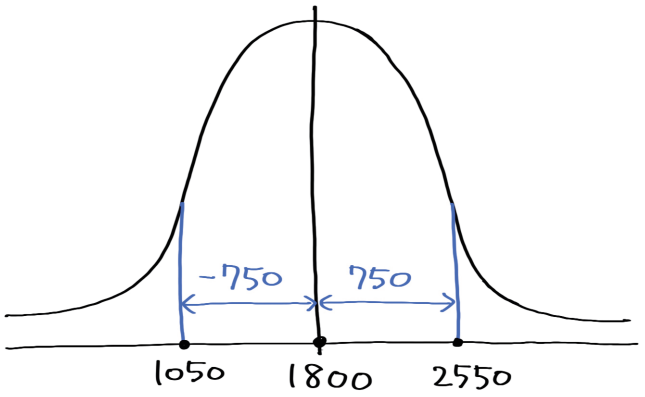
となって、答えは1300万円～ 2300万円になる。

問題❸　第一営業部の売上平均が1050万円から2550万円になる可能性は何％？

まずは正規分布の図を書く。こんな感じで。



何％というのは標準正規分布表の中にあるから、それとこの図を結びつけられればよいんだろうな。平均の1800から2550はいくつ離れているかと言うと、単純に引き算で2550−1800＝750万円か。750万円ってことはZで言うと何個分になるかと言うと、750 ／ 500で1.5個分。つまり正規分布表のZ＝1.50個のところだから43.32%。1050万円も1050−1800で、同じく750万円。今度はマイナス側に動いている。これもZ＝1.50個分動いたところだから43.32%。



$$\frac{X-Ave}{Z}=\frac{2550-1800}{500}=1.5$$

相関関係には、

① 因果関係があるケース

以外に

② 疑似相関があるケース
③ 逆の因果関係があるケース
④ 単なる偶然のケース

の3つのパターンがあり得て、それぞれがどういうものなのかを知っておく必要があります。

なるほど。知りたいです！！

うーんと山田君…私ハンバーガー大好きなんだけど、山田君も好き？

は、ハンバーガーですか。唐突ですね。もちろん大好きです。大学時代にハンバーガーとチーズバーガーを合わせて10個食べたことあります。

それはすごいね。じゃあちょっとランチタイムには早いけど、今から一緒にハンバーガーを食べに行こうか！

え、わかりました。

シェイクと電力

ここのお店のハンバーガーは美味しいよ。関東には10店舗くらい展開しているけど、こだわっていて。ハンバーガー以外に、私がとっても気に入っているのがシェイクなんだ。

お一部長、甘いものお好きなんですか？

もちろん。特にこのお店が出しているバニラとチョコとイチゴを混ぜた特注シェイクは、めちゃくちゃ最高。とんでもないカロリーだから、たまにしか飲めないんだけどね。

私も学生時代は、練習の後は甘いものが最高の癒しでした。シェイク最高ですね。注文しましょうか。

（シェイクを受け取る2人）

う、うまい！！　いやーよく冷えていてしかもこれは美味しいですね。甘すぎないけどしっかり味がして。

そうだろう。特に考えすぎて脳が疲れている時とか最高なんだよね。

脳にエネルギーを送って、もっと考えないと。

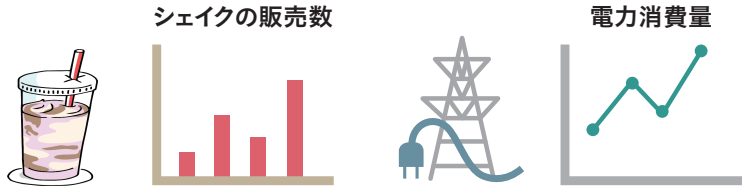
さて山田君。脳にエネルギーが来たところで問題。

（えっまだハンバーガー頼んでないけど）はい。

山田君が電力会社の社員だったと仮定します。

仮定。はいわかりました。

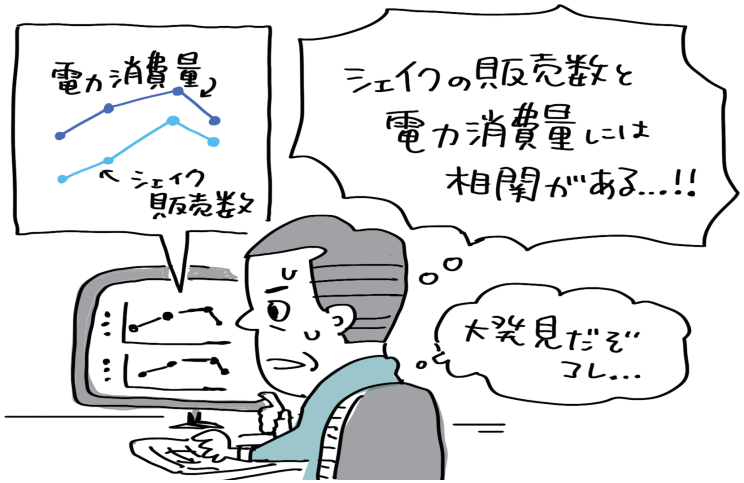
そこで、このお店のシェイクの売上と、関東地方の電力消費量との間に強い正の相関があることがわかった。つまり、シェイクが売れれば売れるほど電力消費量が上がったという事実がわかったと仮定する。



データを調べるとシェイクの販売数と電力消費量に相関があった！

なるほど。そこまでは理解しました。

そこで、山田君の電力会社の上司がこんなことを言ってきたらどうする？



「電力の供給が限界にきているので、電力消費を抑えなければいけない。そのために、ハンバーガーチェーンでシェイクを販売するのをやめさせてこい！」

うーん、それはおかしいだろって思いますね。

それはなぜかな？　2つの間には強い正の相関があるんだよ。

相関はあるけど、因果はないパターンなのかなと思います。それ以上に、ただ暑かったからシェイクも売れたし、エアコンの使用量が高まって電力も消費されただけなのかなと。つまり気温が下がらない限り、たとえシェイクの販売をやめたとしても電力消費の問題は解決されない。

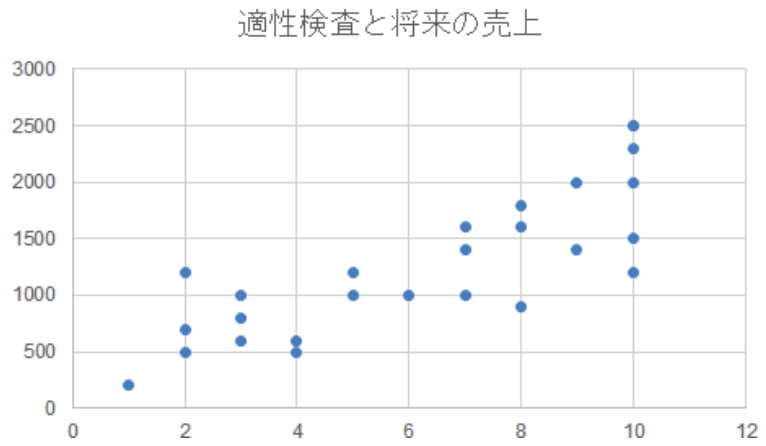
山田君、なかなか鋭いね。そうやってちゃんと言葉にして説明ができる人は、なかなか多くない。そうなんだよね。今回のハンバーガーチェーン店のケースでは、気温とシェイクの売上、気温と電力消費量の間に正の相関があっただけなんじゃないか。だから

ということは、空欄の理由は「適性検査」と「IQテスト」の相関はすでに書かれていて、総当たり表が斜めの軸を基準として左右対称だからでしょうか？

素晴らしい、大正解！

入社時のデータと将来の売上をグラフにしよう

それでは最後に、「適性検査」と「将来の売上」の関係を図にしたものを見てみよう。



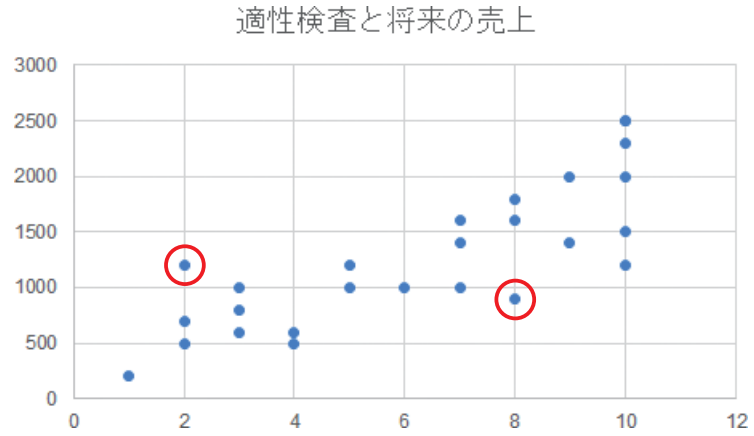
山田君、この図のY軸が「将来の売上」で、X軸が「適性検査」の点数です。この図は何の相関があると言えるかな？

えーっと。右肩上がりなので、正の相関があると言えます。

正解！　先ほど相関係数を求めたけれど、この図のように表現してみると、0.81の正の相関がどれくらい点がまとまっているのかをイメージできるよね。慣れてくると、相関係数から点のおおよその散らばりがイメージできるようになる。そうすると反対に、散らばりの形から分析のまちがいに気がつくことができるようにもなります。

なるほど！　相関係数と点の散らばりを結びつけて考えられるようになるんですね。

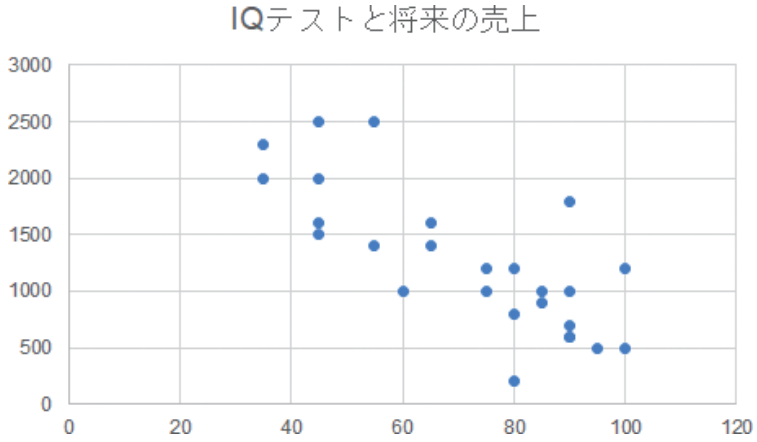
この図をよく見ると、主観ではあるけれど売上900万円くらいのところに適性検査の点数の高い人（8）がいたり、1200万円くらいのところに適性検査の点数の低い人（2）がいて、この2人は全体のトレンドから少し外れていることがわかるかな？



そうですね。右肩上がりの線からは少し外れていますね。

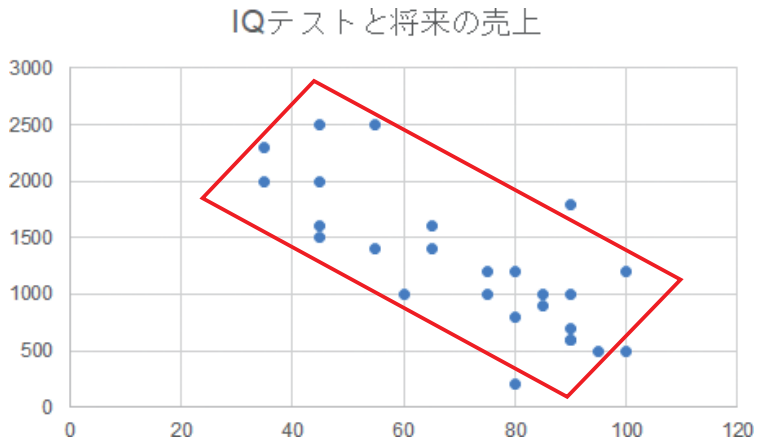
実際の運用では、こういったデータにまちがいがないかとか、なぜこうなったのかを調査するなどして、分析の精度を高めたりもします。だから、データは相関係数のような数値だけを見るんじゃなくて、ちゃんと図にして見てみるのが大事です！

次に、「IQテスト」と「将来の売上」を図にしたものが下記になります。こちらは右肩下がりなので負の相関になっているね。この中で全体のトレンドから外れている人はどれになるかな？

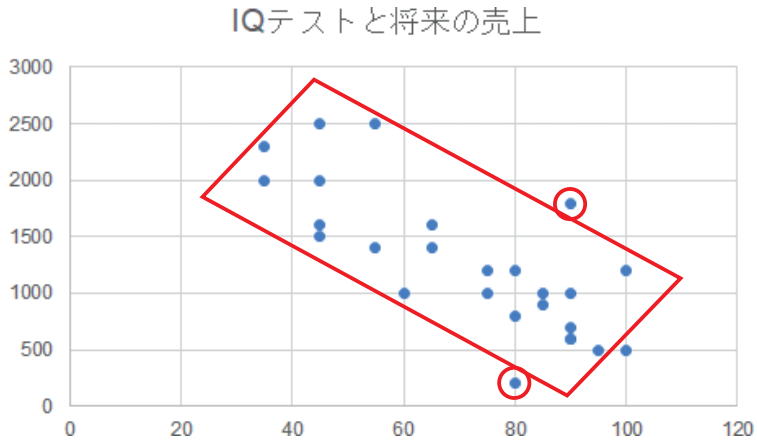


部長、すいません。初歩的な質問なのですが、トレンドって何でしょうか？

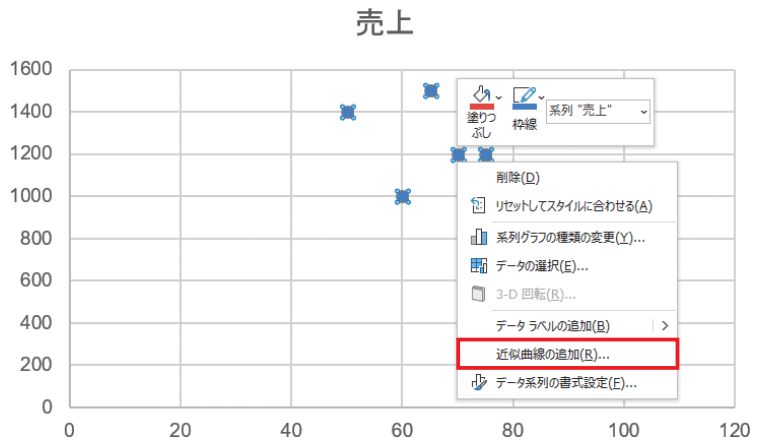
確かに、前提を確認せずに進めていた。グッドクエスチョンだね。トレンドというのは、データ全体の大きな傾向のことを指すよ。この図に入れた赤い四角のように、全体の大きな傾向を表している部分のことだね。



先ほどの「トレンドから外れている人」というのは、私の感覚だとこの図で赤く囲んだ部分の外側にある点がそれに該当する。実際の業務では、この点に当たる人がどんな人なのか？　なぜトレンドの枠から外れるのか？　といったことを考えていくと、数値だけでは見えてこない採用時のヒントなどがわかることもあります。

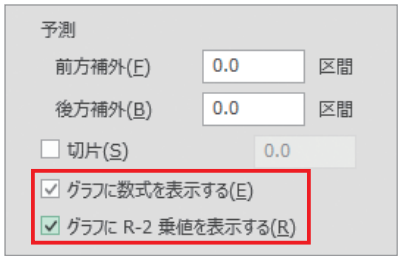


が出てきたら、そちらをクリックしてください。

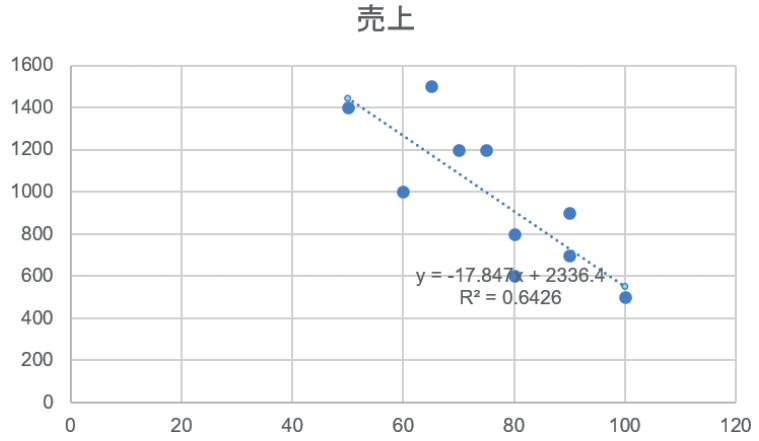


すると次のような表示が出てくるので、画面右側の設定画面で「線形近似」を選んでください。

そして「グラフに数式を表示する」と「グラフにR-2乗値を表示する」のチェックボックスにチェックを入れてください。そうするとどうなるかな？



部長！　点の中に破線の直線が表示されて、2つの数式が出てきました！



そう。この直線のことを、近似曲線と言うよ。

散布図として描かれた結果を理解しよう

そこで1つ目の問題。

問題❶　売上と面接の点数の間には相関があるか？　あるとすればなんの相関か？

これは右肩下がりの関係なので、負の相関があると言えると思います。

正解。山田君、昨日の相関の復習もちゃんとできているね。グレイト！　なんとなく面接の点数と売上には正の相関がありそうという直感があっても、こうしてデータを見るとそうなっていないこともある。では次の問題。

問題❷　この近似曲線を、エクセルはどうやって引いているでしょうか？

どうやって？　え、どうやってだろう。うーんと、点のそばを通るようなまっすぐの線を引いているようには思いますね。なんとなくすべての点の傾向をうまく表している。以前の言葉で言うと、トレンドを上手に表しているみたいな感じですかね。

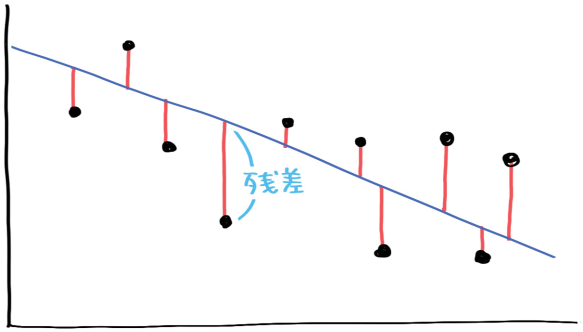
答えがわからない時は、そうやって頭に考えていることを声に出して説明していくのはよいね。山田君の思考がよくわかるし、頭の整理にもなる。よし、これはちょっと難しいので説明しよう。

エクセルで「近似曲線の追加」から「線形近似」を選択した場合、10個の点のあるX軸とY軸の世界に、エクセルが無数の直線を引いていることをイメージしてみてください。

その時、ある1つの直線とこの10個の点との距離の合計をSとします。1つ目の直線に対してS1、2つ目の直線に対してS2としていった時に、エクセルはこのSの大きさが一番小さくなる直線を残して表示します。この時のそれぞれの直線とグラフ上の1点との距離を、残差といいます。この残差という言葉を使うと、「線形近似」では

残差の合計が最小になる直線を引いている

と言うことができるよ。



残差が最小？？

山田君、ちょっと難しい言葉が出てきてもビビることはないよ。大雑把に考えてみると、このX軸とY軸の世界では、10個の点の集まりは直線上にすべて並んでいるわけじゃないよね。

はい。

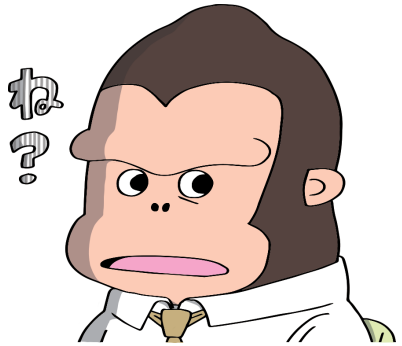
OK。文章を使った回帰分析の説明は、なかなかややこしいからね。ちなみに山田君は、回帰分析に近いことはすでにもうやっている。

えっ。いつですか？　やった記憶ないです。

回帰分析の定義に「モデルを当てはめること」「もっとも基本的なモデルはY＝aX＋b」って書いてあるよね。このY＝aX＋bという数式、どこかで見た記憶はないかな？

受験の時の数学で見たような？？

今は、受験時代の話をしているわけじゃないよね。



し、失礼しました。今までやってきた分析の中でってことですよね。うーんと。あ、そうか。先日の散布図のグラフの中に

$$Y = -17.847X + 2336.4$$

という数式が出てきました！

そうだね。実はそれが、「連続尺度の従属変数（目的変数）Yと独立変数（説明変数）Xの間にモデルを当てはめる」ということなんだ。それからこれは趣味の問題なんだけど、私はこの従属変数と独立変数という言葉がイメージしづらいので、あまり使いたくない。ここからは、目的変数と説明変数という呼び名で説明するよ。

回帰分析というのは、「目的変数」と「説明変数」の間に成り立つ数式を導く分析のことと言える。予測したい目的である「目的変数」と、この目的変数を説明してくれそうな「説明変数」という2つの変数を使って数式を作るんだ。

「説明変数」としてどういう値が「目的変数」に関係しそうかは、分析者の主観に頼ることになる。このあたりは現場の経験で、「この数値が高い人は売上が高いように予想されるな」といった感覚を生かしながら、有効な「説明変数」を探していく。予測に生きるかどうかわからなかったら、まずはいったん幅広い数値を使って分析し、検証してみるというスタンスで説明変数を集めていくとよいね。次の図は、説明変数と目的変数の関係を表したものです。

回帰分析

目的変数
（従属変数）

- 目的となる変数（例：売上など）
- この値に関係しそうな説明変数を集めてくる
- 1つのみ

説明変数
（独立変数）

- 目的変数を説明しそうな変数
- 1つ〜複数も可能

□$_1$ × 係数1 ＋ □$_2$ × 係数2 ＋ □$_3$ × 係数3 ＝ □ の式にする分析

図の下にある式では、青い四角の部分に説明変数が入る。説明変数は、1つ～複数入れることができる。それに係数をかけたものの組み合わせが「説明変数2＊係数2、説明変数3＊係数3…、説明変数n＊係数n」になり、それが右側の赤い四角の目的変数に結びつく、という式になっている。係数はあとで再度説明するけど、説明変数が働いた時に、目的変数がいくつ動くかを表している数になります。

この時、説明変数が1つの場合は単回帰分析、2つ以上の場合は重回帰分析と呼ぶよ。「単」には単騎とか単音など「1つ」という意味があるのでイメージしやすいかな。定義には「次元」という言葉が出ていたけど、ここでは変数が「1つ」なら「1次元」、「2つ」なら「2次元」と考えてくれればいい。

例えばこの式では、面接の点数Xが何点かわかれば、将来の売上予測がYの数値として求められる数式になっているよね。将来の売上Yを目的変数として、説明変数である面接の点数Xを代入するとYが求められる式になっている。

$$Y = -17.847X + 2336.4$$

なるほど。ということは、この近似曲線を求めるというのが回帰分析なんでしょうか？

大枠の理解としては正しいね。単回帰分析であればエクセル上でグラフにして見られるけど、重回帰分析になって説明変数が増えてくると表現しづらくなる。それに、それぞれの数値がどれくらい信じられるのか、という点についてはもう少し説明する必要がある。山田君、回帰分析をやってみたいかな？

はい。もちろんです！

エクセルで単回帰分析をしよう

それでは、まずはこの前の散布図のデータを使って回帰分析をしてみよう。今回は説明変数が「面接の点数」1つなので、「何」回帰分析になるかな？

	名前	面接の点数	売上
3	A	90	900
4	B	60	1000
5	C	100	500
6	D	90	700
7	E	80	800
8	F	65	1500
9	G	70	1200
10	H	50	1400
11	I	75	1200
12	J	80	600

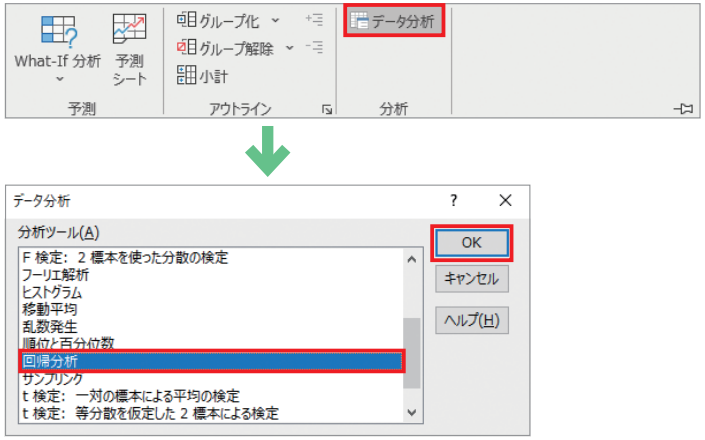
説明変数　目的変数

はい。単回帰分析ですね。

正解。今回は、面接の点数を説明変数、売上を目的変数、つまり予測したい数値に設定して、その関係を分析したい。相関を求めた時に使った、データ分析アドインを開けるかな（P.135）？

「データ」タブをクリックして、右上のアイコンの「データ分析」をクリックですね。

その通り。その中にある「回帰分析」を選択して、「OK」をクリックしてください。



できました。

それでは、最初に「ラベル」というチェックボックスにチェックを入れよう。当たり前だけど、エクセルはコンピューターつまり計算機の上で動く。計算機って、基本的には数値のデータしか計算することができない。だけど、今回のラベルにあるような「売上」や「面接の点数」といった情報は、分析結果を人間が読み取る時にはあると便利な情報だよね。「ラベル」にチェックを入れると、人間がわかりやすくなるように、データの1行目の言葉を「ラベル」として認識してくれる。

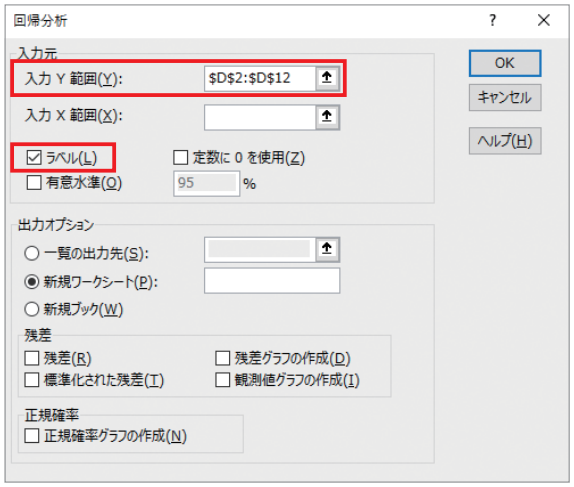
なるほど。

次に、「入力Y範囲」、「入力X範囲」という表示が見えるかな？先ほどから出ている数式にもYとX ってあったけど、このYにはどの数値を入れたらいいと思うかな？

Yは目的変数なので、今回の予想したい数値である売上ですかね？

正解。では、「入力Y範囲」に売上の数値が入ったセル範囲を入れてください。

はい。今回は「ラベル」にチェックを入れているので、「売上」という文字データが入ったD2セルから、Jさんの売上の入ったD12セルまでを選択しました。

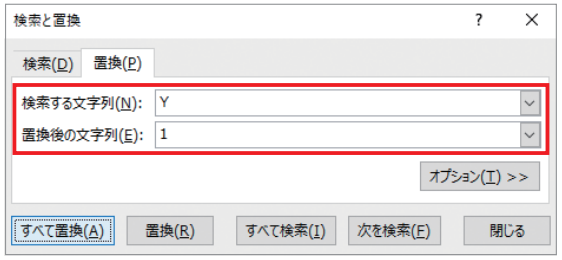


次に「入力X範囲」に数値を入れよう。Y範囲には目的変数が入ったので、「入力X範囲」には何が入る？

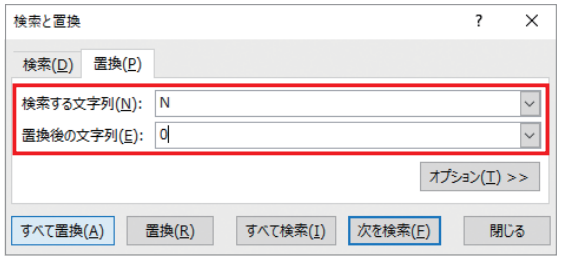
Xは目的変数を説明する可能性がある変数ということで、説明変数が入ります。今回は、「面接の点数」が説明変数になりますね。

その通り。では、「入力X範囲」に面接の点数が入ったセル範囲を入れてください。

	A	B	C
1	氏名	性別	起業経験
2	綾野太郎	M	N
3	井上大輔	M	N
4	遠藤聡子	F	N
5	岡島恵子	F	Y
6	岡本博宣	M	Y
7	河野照子	M	N
8	花田大治	M	Y
9	花本恵子	M	Y
10	皆川俊樹	M	N
11	柿崎二郎	M	Y
12	柿沼雅美	F	N
13	関根加代	M	N
14	原田大作	F	N
15	佐々木康夫	M	Y



それから「N」は「0」ですね。



最終的に、次のようなデータが得られました。

	A	B	C
1	氏名	性別	起業経験
2	綾野太郎	M	0
3	井上大輔	M	0
4	遠藤聡子	F	0
5	岡島恵子	F	1
6	岡本博宣	M	1
7	河野照子	M	0
8	花田大治	M	1
9	花本恵子	M	1
10	皆川俊樹	M	0
11	柿崎二郎	M	1
12	柿沼雅美	F	0
13	関根加代	M	0
14	原田大作	F	0
15	佐々木康夫	M	1
16	佐藤順子	F	0
17	斎藤なつみ	M	1
18	三木義晴	M	1
19	山口健斗	M	1

すばらしい。

最後に売上。これは営業部のデータですね。これは以前もらっているものがあるので、エクセルに合わせますね。できました！

	A	B	C	D	E	F	G	H	I	J	K	L
1	氏名	ストレスに対する弱さ	論理重視	仕事重視	回避的	楽観的	競争的	懐疑的	受容性	起業経験	面接の点数	売上(万円)
2	岡島恵子	5	1	2	4	8	3	10	3	0	70	¥2,000
3	山田仁	6	2	6	3	7	4	3	10	0	70	¥3,100
4	赤城武則	7	2	8	8	7	8	3	1	0	40	¥2,300
5	中島由紀子	6	2	9	2	8	1	8	4	1	50	¥4,500
6	花本恵子	3	3	7	4	8	7	3	3	1	90	¥3,100
7	須藤学	3	3	8	8	9	1	5	4	0	50	¥1,800
8	中西輝	4	3	8	4	7	6	7	3	1	70	¥3,600
9	津吹剛	1	3	9	4	8	4	10	5	1	80	¥4,400
10	土屋大志	4	3	2	6	8	9	2	5	0	40	¥1,600
11	花田大治	3	4	5	4	8	7	3	3	1	70	¥2,800
12	皆川俊樹	5	4	2	2	8	8	7	3	0	80	¥3,500
13	原田大作	2	4	7	6	7	7	2	7	0	60	¥2,500
14	水野弥一	3	4	8	8	9	1	5	4	0	60	¥2,000
15	牧口浩二	1	4	7	9	9	4	4	3	1	80	¥4,000
16	河野照子	3	5	5	3	8	5	8	5	0	60	¥2,000
17	柿沼雅美	7	5	7	4	8	10	8	5	1	80	¥3,000
18	三木義晴	2	5	9	9	6	6	10	4	1	60	¥2,500
19	山口健斗	2	5	3	9	6	6	10	4	0	40	¥1,600
20	小山田敏	4	5	10	5	7	7	7	6	1	70	¥6,000
21	仁川修二	9	5	3	6	9	3	7	7	0	40	¥1,800
22	竹原あんな	1	5	3	6	7	4	5	7	0	60	¥1,600
23	湯川秀樹	4	5	4	6	8	9	2	5	0	40	¥1,300
24	木次正光	4	5	3	3	7	10	7	5	0	50	¥1,900
25	井上大輔	7	6	3	5	8	3	2	2	0	40	¥1,500

エクセルで重回帰分析をしよう

では山田君。今から行うデータ分析を、目的変数と説明変数という言葉を使って説明してください。

はい。今回の分析のゴールは繰り返しになりますが、次のようなものになります。

> 「具体的なゴール：面接終了後にわかっているデータ（例：面接の点数、適性検査のスコア、起業経験の有無など）から、なるべく活躍する可能性が高い人を採用する意思決定に役立つモデルを作る」

この中で、「活躍する可能性が高い」という表現が、データ分析の目的を定義するにはやや曖昧です。

そこで今回は、数値データを持っていて、なおかつ分析できるものに変えるために、「売上が高い」ことを「活躍する」と定義します（作者注：もちろん売上が活躍を表すすべてではありませんが、エクセルでできる分析によっていったんなんらかの知見を得るという意味では十分であるし、これ以上の分析を定義し考えつづけるコストを考えると、ここでいったん分析をする方がコストパフォーマンスの観点でよいと判断します）。

今回の分析対象者の人数は第一営業部全体の50名。それぞれの人の売上を目的変数として、適性検査の項目であるストレスに対する弱さ、論理重視、仕事重視、回避的、楽観的、競争的、懐疑的、受容性、起業経験のあるない、面接の点数を説明変数とした重回帰分析を行います。

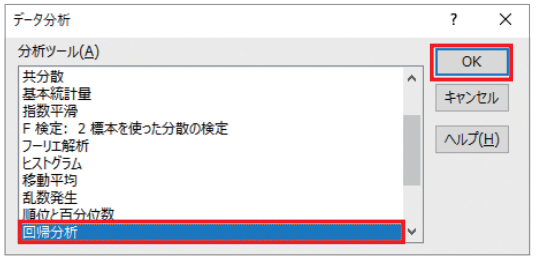
素晴らしい。それでは、エクセルを使って分析していきましょう。結果が出てきたら、この分析全体の評価と、どういうモデルを組むかということを山田君が説明してください。

わかりました。こちらがデータになります。

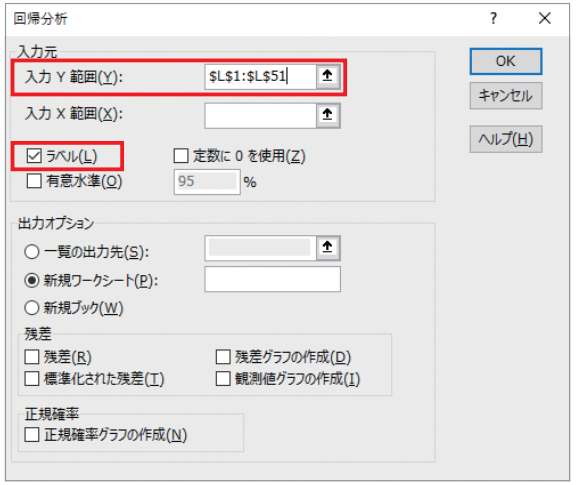
	A	B	C	D	E	F	G	H	I	J	K	L
1	氏名	ストレスに対する弱さ	論理重視	仕事重視	回避的	楽観的	競争的	懐疑的	受容性	起業経験	面接の点数	売上(万円)
2	岡島恵子	5	1	2	4	8	3	10	3	0	70	¥2,000
3	山田仁	6	2	6	3	7	4	3	10	0	70	¥3,100
4	赤城武則	7	2	8	8	7	8	3	1	0	40	¥2,300
5	中島由紀子	6	2	9	2	8	1	8	4	1	50	¥4,500
6	花本恵子	3	3	7	4	8	7	3	3	1	90	¥3,100
7	須藤学	3	3	8	8	9	1	5	4	0	50	¥1,800
8	中西輝	4	3	8	4	7	6	7	3	1	70	¥3,600
9	津吹剛	1	3	9	4	8	4	10	5	1	80	¥4,400
10	土屋大志	4	3	2	6	8	9	2	5	0	40	¥1,600
11	花田大治	3	4	5	4	8	7	3	3	1	70	¥2,800
12	笹川俊樹	5	4	2	2	8	8	7	3	0	80	¥3,500
13	原田大作	2	4	7	6	7	7	2	7	0	60	¥2,500
14	水野弥一	3	4	8	8	9	1	5	4	0	60	¥2,000
15	牧口浩二	1	4	7	9	9	4	4	3	1	80	¥4,000
16	河野照子	3	5	5	3	8	5	8	5	0	60	¥2,000
17	椎沼雅美	7	5	7	4	8	10	8	5	1	80	¥3,000

まず、エクセルの「データ」タブから「データ分析」をクリックします（P.135）。

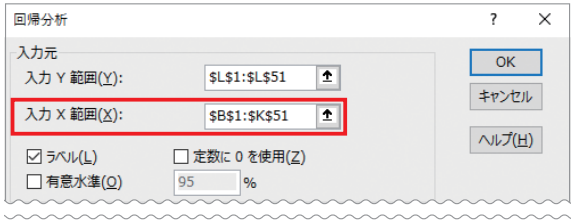
そして、出てきたデータ分析の選択肢の中から「回帰分析」を選択し、「OK」をクリックします。



次に入力Y範囲ですが、Yには求めたいものつまり目的変数を入れるので、今回は売上であるL列を選択します。分析結果がわかりやすくなるように、ラベルのチェックを忘れずにですよね。ラベルをチェックしているので、1行目の「売上」という言葉を選択しても大丈夫です。こちらはラベルとしてエクセルが認識してくれるので、数値データじゃなくても分析が可能、なおかつ分析結果が見やすくなります。



次に入力X範囲ですが、こちらは説明変数です。今回はストレスに対する弱さ、論理重視、仕事重視、回避的、楽観的、競争的、懐疑的、受容性、起業経験のあるない、面接の点数の10項目を説明変数として利用します。



今回は説明変数のラベルとしてはBのストレスに対する弱さからKの面接の点数までを選択し、データはラベルを含めて50名分なので、1から51つまりB1からK51までを選択します。そして「OK」をクリックすると…新しいタブが表示されて、結果が得られました！

	A	B	C	D	E	F	G	H	I	J
1	概要									
2										
3	回帰統計									
4	重相関 R	0.7892								
5	重決定 R2	0.6228								
6	補正 R2	0.5260								
7	標準誤差	793.6148								
8	観測数	50.0000								
9										
10	分散分析表									
11		自由度	変動	分散	観測された分散比	有意 F				
12	回帰	10	40550647.8	4055064.78	6.438405196	9.14193E-06				
13	残差	39	24563152.2	629824.415						
14	合計	49	65113800							
15										
16		係数	標準誤差	t	P-値	下限 95%	上限 95%	下限 95.0%	上限 95.0%	
17	切片	1398.10	1359.88	1.03	0.31	-1352.53	4148.72	-1352.53	4148.72	
18	ストレスに対する弱さ	8.91	47.01	0.19	0.85	-86.18	104.01	-86.18	104.01	
19	論理重視	-45.91	53.74	-0.85	0.40	-154.60	62.78	-154.60	62.78	
20	仕事重視	124.95	51.51	2.43	0.02	20.76	229.13	20.76	229.13	
21	回避的	-56.15	58.26	-0.96	0.34	-173.98	61.68	-173.98	61.68	
22	楽観的	-47.76	124.76	-0.38	0.70	-300.10	204.59	-300.10	204.59	
23	競争的	-73.91	45.60	-1.62	0.11	-166.15	18.33	-166.15	18.33	
24	懐疑的	-53.64	45.14	-1.19	0.24	-144.95	37.67	-144.95	37.67	
25	受容性	72.06	63.76	1.13	0.27	-56.90	201.02	-56.90	201.02	
26	起業経験	802.33	262.45	3.06	0.00	271.48	1333.17	271.48	1333.17	
27	面接の点数	23.74	7.44	3.19	0.00	8.69	38.79	8.69	38.79	
28										

よいですね。この分析結果が得られた際に、最初に見るべきポイントはどこでしょうか？

重決定R2の値を見ます。今回の結果は0.622です。

その意味はどういうことになりますか？

今回の10個の説明変数は、目的変数の動きを予想した時に62.2%の動きを捉えていると言える、ということになると思います。

そうだね。今回の分析はいい分析と言えるのでしょうか？

決定係数が50%以上なので、曖昧な言い方かもしれませんが「ある程度の結果」という感じでしょうか？