

## Week4 確認テスト 解答・解説

以下に、世帯の消費の実態を調べたデータ（ダミーデータ）がある。  
このデータを使って分析してみよう。

※課題に必要な分析の Excel でのやり方については補講を参照。

【問題 1】 利用データ：dummydata\_A.xlsx

目的変数を食費、説明変数を年間収入とした回帰分析による予測モデルの決定係数（ $R^2$ ）として正しいものを、次の(1)～(4)のうちから一つ選びなさい。（計算値は小数第五位で四捨五入した。）

- (1) 0.3934
- (2) 0.6475
- (3) 0.3812
- (4) 0.7504

【解答 1】

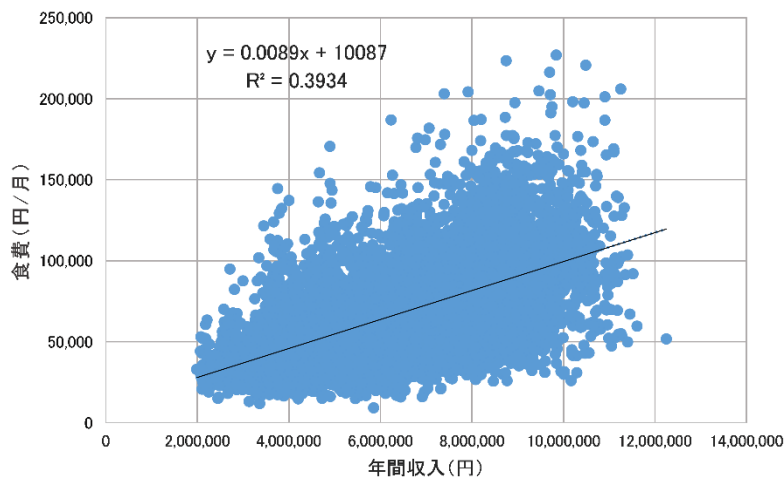
(1)

【解説 1】

食費と年間収入の散布図を描画し、近似式から回帰式を求める。

Excel のグラフツールを使って食費と年間収入の散布図を描画し、近似式と決定係数（ $R^2$ ）を表示させると以下のようなグラフができる。

$R^2=0.3934$  であるので、回答は(1)となる。



【問題 2】 利用データ：dummydata\_A.xlsx

問題 1 で求めた回帰式から、年間収入 600 万円の食費の予測値として、正しいものを次の (1)～(4)のうちから一つ選びなさい。(予測値は、Excel の散布図、近似曲線のオプションに表示される回帰式(傾きの値は小数第五位を四捨五入した値)を用いて算出、計算値は小数第一位で四捨五入した。)

- (1) 59,845 円
- (2) 62,147 円
- (3) 63,487 円
- (4) 58,463 円

【解答 2】

(3)

【解説 2】

解説 1 のグラフに表示される式は、 $y = 0.0089x + 10087$  である。

この式に  $x=600$  万円を代入すると、 $y=63487$  円となる。

【問題 3】 利用データ：dummydata\_A.xlsx

光熱・水道費と年間収入の散布図を描き、目的変数を光熱・水道費、説明変数を年間収入とした回帰分析による予測モデルを作成した際、モデルの評価として、最も適切なものを次の(1)～(4)のうちから一つ選びなさい。

- (1) データに欠損値が多いため、欠損値データを除いてモデルを作成すべきである
- (2) オーバーフィッティングが疑われる
- (3) モデルの当てはまり指標となる決定係数 ( $R^2$ ) は約 0.185 のため、モデルの精度は低いと考えられる
- (4) 決定係数 ( $R^2$ ) は 1 以上となるはずなので、データの不備等を点検し、モデルを作成しなおした方がよい

【解答 3】

(3)

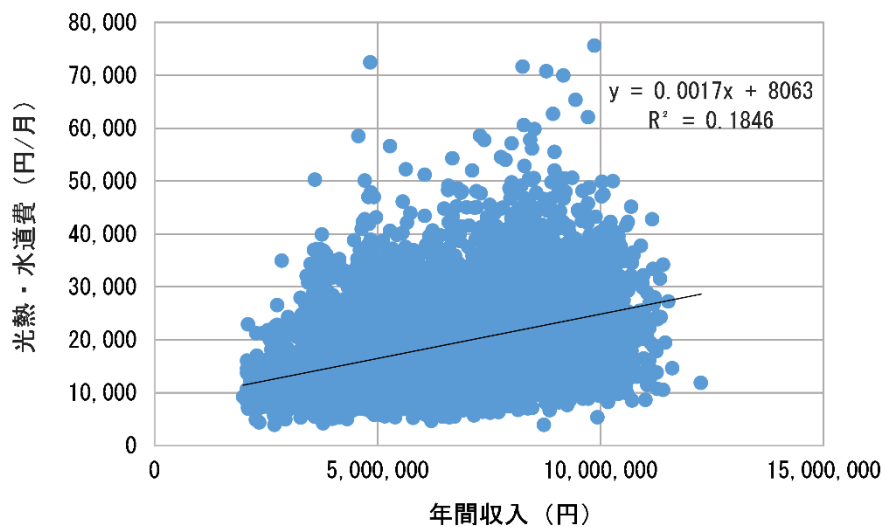
【解説 3】

ダミーデータ A を見ると、欠損値は確認できないため、(3)は適切でない。

当該モデルは線形モデルのため、オーバーフィッティングは生じない。このため(2)は適切でない。

決定係数 ( $R^2$ ) は 0～1 の値をとる。よって(4)は適切でない。

光熱・水道費と年間収入の散布図を描画し、近似式と決定係数 ( $R^2$ ) を表示すると以下となる。



決定係数  $R^2=0.1846$  であり、モデルの精度は低いと考えられる。よって(3)が回答となる。

**【問題 4】**

データの分析結果の報告として、最も適切でないものを(1)～(4)のうちから一つ選びなさい。

- (1) データの比較結果を報告する際、各変数の単位を統一した
- (2) ヒストグラムを描いたところ、データの分布がL字型の形状を示したため、データのボリュームゾーンを示す代表値として平均値のみを報告した
- (3) 毎年行っている調査であったが、報告書に、用語の定義、データの取得方法と出所を記載した
- (4) 各店舗の販売効率を報告する際、店舗ごとの売上額ではなく、各店舗の売場面積あたりの売上額を比較して報告した

**【解答 4】**

(2)

**【解説 4】**

- (1)データの比較を行う場合、単位を統一することは重要である。単位が異なると、比較条件が異なってしまうため、多寡の判断ができない。
- (3)毎年行っている調査であっても、結果解釈の前提条件となる用語の定義やデータの取得方法、出所を明記する必要がある。
- (4)店舗の売上額は、店舗面積の影響を受ける。よって各店舗の販売効率を比較したい場合は、売上面積あたりの売上額で比較するのが妥当。
- (2)ヒストグラムがL字型の場合、ボリュームゾーンを示す代表値は中央値や最頻値が適切である。平均値は中央値や最頻値よりも大幅に大きくなる。このため(2)が回答となる。

**【問題 5】**

グラフの種類と使途の組み合わせで、最も適切でないものを次の(1)～(4)のうちから一つ選びなさい。

- (1) ヒストグラム：棒の長さで、量の大きさを比較する
- (2) 折れ線グラフ：線の傾きの推移から、時系列や連続的な変化を見る
- (3) 帯グラフ：全体に占める構成比を比較する
- (4) 散布図：2種類のデータの相関をみる

**【解答 5】**

(1)

**【解説 5】**

ヒストグラムは一見棒グラフに似ているが、各階級の高さ×幅の面積が度数を表す。例えば、階級の幅が2倍になったときには、長方形の横の長さが2倍になり、縦の長さは2分の1になる。選択肢(1)の「棒の長さで、量の大きさを比較する」は棒グラフを示す記述である。よって回答は(1)となる。

※ このダミーデータは、独立行政法人 統計センターが提供している一般用マイクロデータを加工し作成したものです。