ロジスティック回帰

ロジスティック回帰は，目的変数がカテゴリカルであったり比率 (0,1) で合ったりする場合に用いる代表的な回帰手法である．

変数の尺度とダミー変数

データ分析において用いるデータは，必ずしも数量として意味をもつ連続的なものではありません．統計学の初歩に戻ってみると，Stevens, S.S. による尺度水準の4分類がありました．

スティーブンスの尺度水準

- **名義尺度** ... 他と区別し分類するための尺度．カテゴリー． 例： 性別，クラス，国籍，学籍番号
- **順序尺度** ... 順序や大小に意味を持つが，間隔に意味がない尺度． 例： 順位，Likert scale，級位，診断ステージ
- **間隔尺度** ... 順序だけでなく，間隔にも意味を持つ尺度． 例： 気温，西暦，点数
- **比率尺度** ... さらに0が原点となり，比率にも意味を持つ尺度． 例： 身長，速度，給料，時間

ロジスティック回帰はこのうち，名義尺度や順序尺度に対して回帰分析を行いたい時に用いるものです．ただ一般に，単にロジスティック回帰といった場合は名義尺度，つまりカテゴリカルデータの分析に用いるものを指すことが多いです．

では，何故名義尺度や順序尺度には線形回帰ではなく，ロジスティック回帰を使うのでしょうか．まだロジスティック回帰がどんな回帰をするものなのかは分かりませんが，とりあえず例を見てみます．

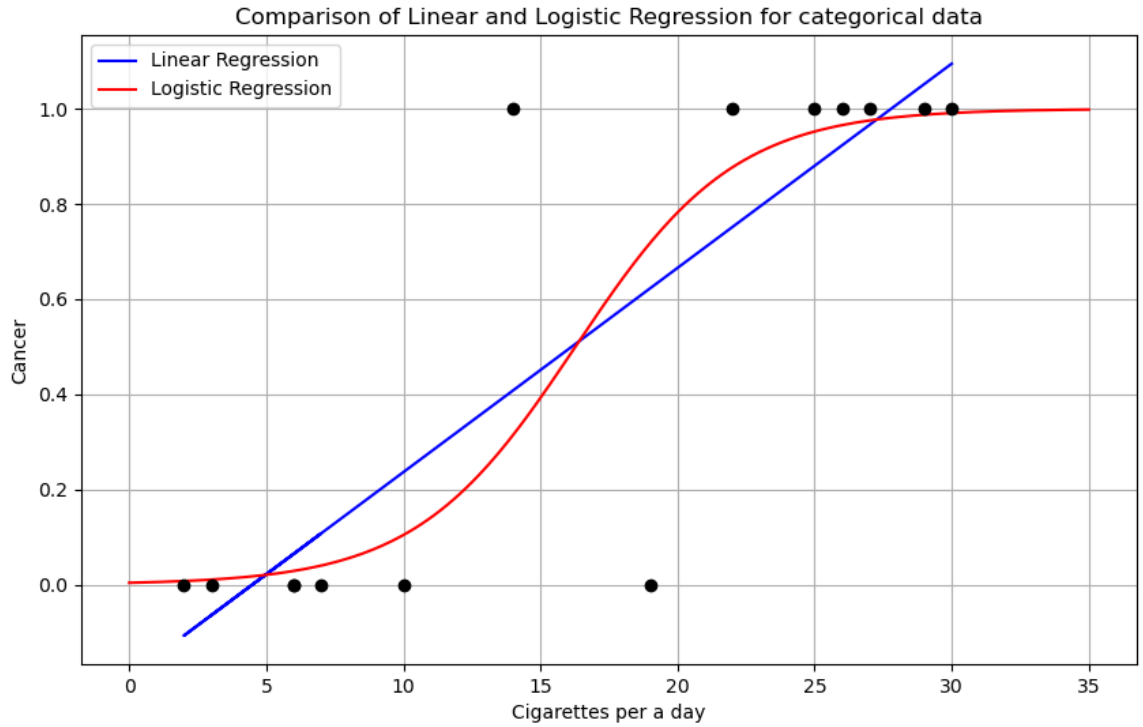
ロジスティック回帰の例

下の図は，1日に吸うたばこの本数によって発がん率 (True or False) に影響があるかをplotした仮想図です．癌の有無は，2値データとしてしか観測できませんので，1と0というダミー変数によって表しています．

▶ Pythonコード

```
# データの定義
x = np.array([3, 7, 2, 6, 6, 10, 14, 19, 22, 25,26,27,29, 30]).reshape(-1, 1)
y = np.array([0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 1,1,1,1])

# 線形単回帰モデル
lin_reg = LinearRegression()
lin_reg.fit(x, y)
y_pred_lin = lin_reg.predict(x)

# ロジスティック回帰モデル
```

```
log_reg = LogisticRegression()
log_reg.fit(x, y)
x_test = np.linspace(0, 35, 300).reshape(-1, 1)
y_pred_log = log_reg.predict_proba(x_test)[:, 1]

# プロット
plt.figure(figsize=(10, 6))

# 元データのプロット
plt.scatter(x, y, color='black', zorder=5)

# 線形回帰のプロット
plt.plot(x, y_pred_lin, label='Linear Regression', color='blue')

# ロジスティック回帰のプロット
plt.plot(x_test, y_pred_log, label='Logistic Regression', color='red')

# 設定
plt.xlabel('Cigarettes per a day')
plt.ylabel('Cancer')
plt.title('Comparison of Linear and Logistic Regression for categorical data')
plt.legend()
plt.grid(True)

plt.savefig('../figures/logistic1.png')
```



まず，青の線形回帰モデルより，ロジスティック回帰モデルの方がよく合っていそうではあると思います．

しかし重要なのは，回帰の端のあたりの挙動です．線形回帰は，(0,1)の範囲を飛び出ていることが分かるでしょうか．つまり，こいつに予測をさせると「喫煙本数が27本を超えると，発がん率は100%を超える」や

「毎日マイナス本のたばこを吸うと，癌率もマイナスになる」等の意味の分からない予言が出てしまいます．

この問題こそが，名義尺度や順序尺度に線形回帰が使えない理由です．対するロジスティック回帰は，[0,1]の間に収まった関数形をしています．これなら，発がん率が120%なんてことにはなりません．

とりあえず，最低限はここまで理解すればロジスティック回帰の目的はOKです．

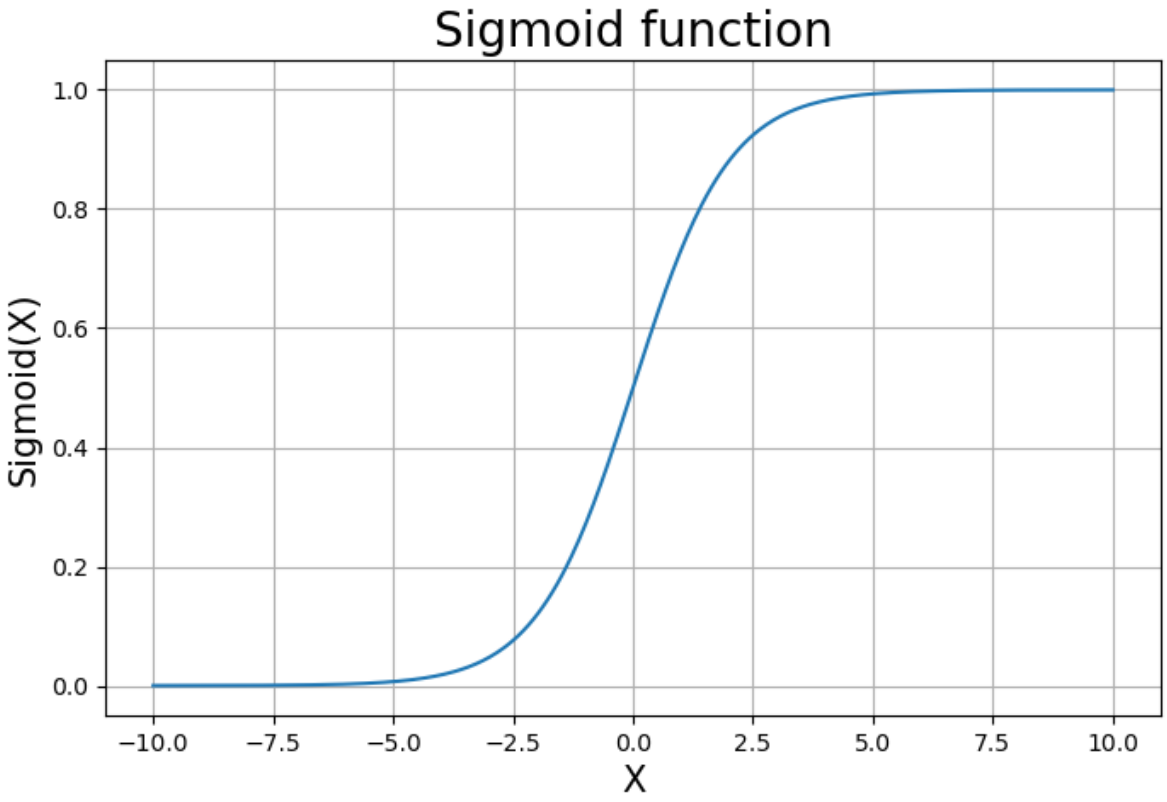
ロジスティック回帰モデル

では，実際にロジスティック回帰を扱うための関数やモデルの導出を行っていきます．

まず，線形回帰の問題点はダミーである目的変数 (y) の値が [0,1] を飛び出してしまうことにありました．それ以外の部分では，先ほどの図をよく見ればたとえば中心 (y=0.5) が一緒だったりするのが分かると思います．

ということで，線形回帰の線を両端で0,1に漸近するような形にゆがめてあげたい気持ちが芽生えます．そこで，$(-\infty, \infty)$ の値を取る変数を $[0,1]$ の範囲に収まるよう変換する**シグモイド関数**あるいは**ロジスティック関数**

$$ \begin{align} f(y) = \frac{1}{1+\exp(-y)} \end{align} $$



を使って回帰線を変換してしまうことを考えます．こうすることで，予測値が[0,1]の範囲に収まります．

これが **ロジスティック回帰**という名前の由来です．回帰モデルをロジスティック変換しているからです．

では，あとは回帰の問題を考えていきます．問題設定によってちょっと式が変わってくるので，とりあえず先ほどのたばこの例のように，目的変数が2値変数

$$ y_i = \begin{cases} 0 (癌)\\ 1 (癌ではない) \end{cases} $$

である場合を考えます．この場合，癌である確率Pを$p(y=1)$とし，n人のデータが独立である $p(y_i=1)=p_i$ と考えると y_iの従う分布はベルヌーイ分布 $B(1,p_i)$ に従うため，尤度Lは

$$ \begin{align} L = \prod_{i=1}^n f(y_i;p_i)= \prod_{i=1}^n p_i^{y_i}(1-p_i)^{1-y_i} \end{align} $$

となります．例によって積は嫌なので対数を取ると

$$ \begin{align} \log(L)= \log(\prod_{i=1}^n p_i^{y_i}(1-p_i)^{1-y_i})\ = \sum_{i=1}^n {y_i \log p_i + (1-y_i) \log(1-p_i) } \end{align} $$

と対数尤度関数が求まります．

ここで，確率p_iについては

線形回帰の$y_i = \bm{\beta}^\top\bm{x_i}$ がロジスティック変換

$$ f(y) = \frac{1}{1+\exp(-y)} $$

をかけられているので，

$$ \begin{align} p_i = \frac{1}{1+\exp(-\bm{\beta}^\top\bm{x_i})} \end{align} $$

となっています．これを$\bm{\beta}^\top\bm{x_i}$について解くと...

▶ 解法

$$ \begin{align} p_i = \frac{1}{1+\exp(-\bm{\beta}^\top\bm{x_i})} = \frac{\exp (\bm{\beta^\top x_i})}{1 + \exp (\bm{\beta^\top x_i})} \end{align} $$

とし，

$$ \begin{align} 1 - p_i = \frac{1 + \exp (\bm{\beta^\top x_i}) - \exp (\bm{\beta^\top x_i})}{1 + \exp (\bm{\beta^\top x_i})} = \frac{1}{1 + \exp (\bm{\beta^\top x_i})} \end{align} $$

も同様に出しておきます．

ここで式6,7の比を取ると

$$ \begin{align} \frac{p_i}{1-p_i} = \frac{\exp (\bm{\beta^\top x_i})(1+\exp (\bm{\beta^\top x_i}))}{1+\exp (\bm{\beta^\top x_i})} = \exp (\bm{\beta^\top x_i}) \end{align} $$

と，$\exp (\bm{\beta^\top x_i})$ であることがわかります．よって両辺の対数を取ると

$$ \begin{align} \log \frac{p_i}{1-p_i} = \bm{\beta^\top x_i} \end{align} $$

が得られます．この式が**ロジスティック回帰モデル**です．

これを先程の対数尤度関数

$$ \begin{align} \log(L)= \sum_{i=1}^n {y_i \log p_i + (1-y_i) \log(1-p_i) } \nonumber \end{align} $$ に代入すると

▶ 解法

まず， $$ \begin{align} \log(L)= \sum_{i=1}^n {y_i \log p_i + (1-y_i) \log(1-p_i) } \end{align} $$ の $\log p_i, \log (1-p_i)$ をそれぞれ

$$ \begin{align} \log p_i = \log(\frac{\exp (\bm{\beta^\top x_i})}{1 + \exp (\bm{\beta^\top x_i})})=\bm{\beta^\top x_i}-\log(1+\bm{\beta^\top x_i})\ \log(1-p_i) = \log(\frac{1}{1+\bm{\beta^\top x_i}})=-\log(1+\exp(\bm{\beta^\top x_i})) \end{align} $$

と表します．すると

$$ \begin{align} \log(L)= \sum_{i=1}^n {y_i \log p_i + (1-y_i) \log(1-p_i) }\ =\sum_{i=1}^n {y_i \bm{\beta^\top x_i} - y_i \log (1+\bm{\beta^\top x_i}) -\log (1+\bm{\beta^\top x_i}) + y_i\log (1+\bm{\beta^\top x_i}) } \end{align} $$

となります．これを整理すると

$$ \begin{align} \log(L) = \sum_{i=1}^n {y_i \bm{\beta^\top x_i} -\log(1+\exp (\bm{\beta^\top x_i})) } \end{align} $$

が得られます．これが，ロジスティック回帰で解く尤度方程式です．ちなみに，線形回帰の場合には最尤法と最小二乗法の２つの求め方がありましたが，ロジスティック回帰では最小二乗法が使えないためこのように最尤法で求めます．

【ホーム】