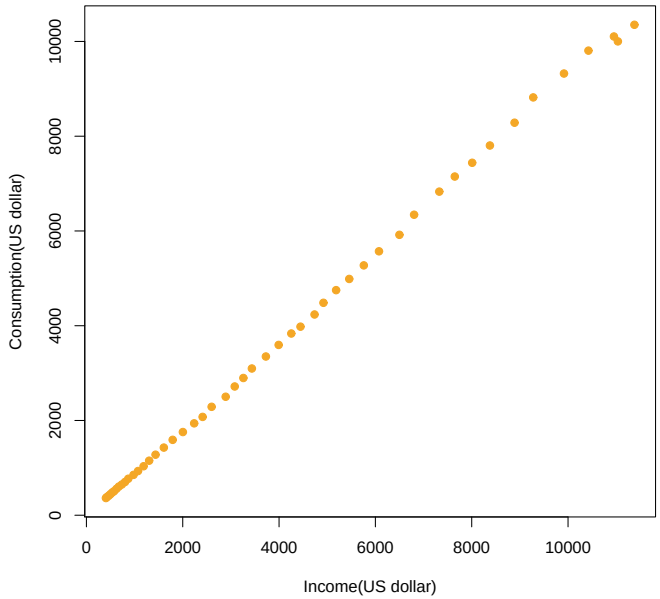
計量経済分析特論
「決定係数 R^2」

原　尚幸

新潟大・経済

推定したモデルのあてはまり

- OLSE : データに最もよくあてはまる推定量
- OLSE が実際にあてはまりが良いか悪いかは別問題
- はまりがよければ説明力の高いモデル
- あてはまりが悪ければ説明力の低いモデル

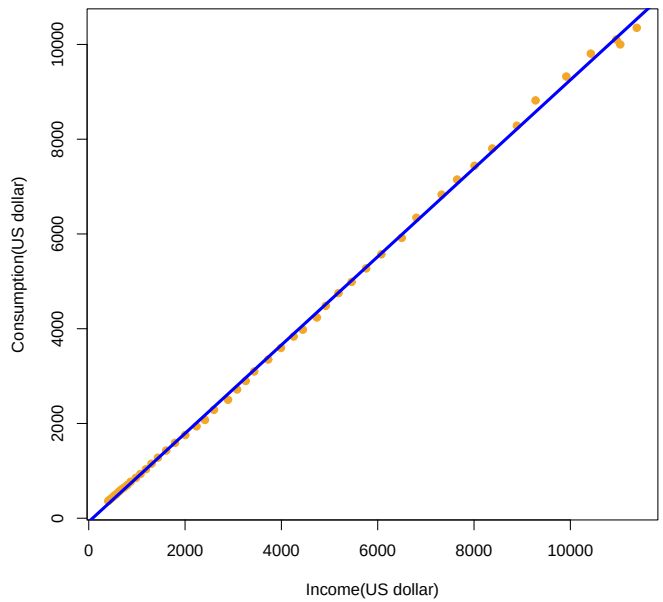


問題

- OLSE で推定したモデルは, 実際どのくらいデータにあてはまっているのか？
- あてはまりのよさを定量的に測る指標はないか？

推定したモデルのあてはまり

- OLSE : データに最もよくあてはまる
  推定量
- OLSE が実際にあてはまりが良いか悪いかは別問題
- はまりがよければ説明力の高いモデル
- あてはまりが悪ければ説明力の低いモデル

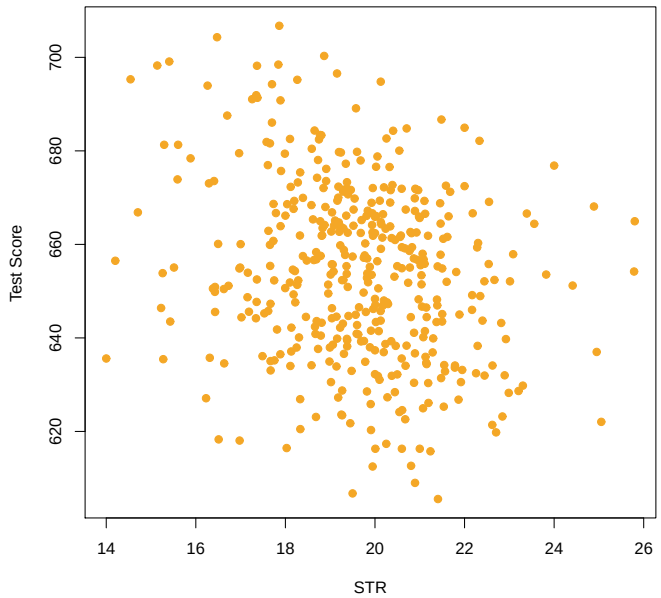


問題

- OLSE で推定したモデルは, 実際どのくらいデータにあてはまっているのか？
- あてはまりのよさを定量的に測る指標はないか？

推定したモデルのあてはまり

- OLSE : データに最もよくあてはまる推定量
- OLSE が実際にあてはまりが良いか悪いかは別問題
- はまりがよければ説明力の高いモデル
- あてはまりが悪ければ説明力の低いモデル

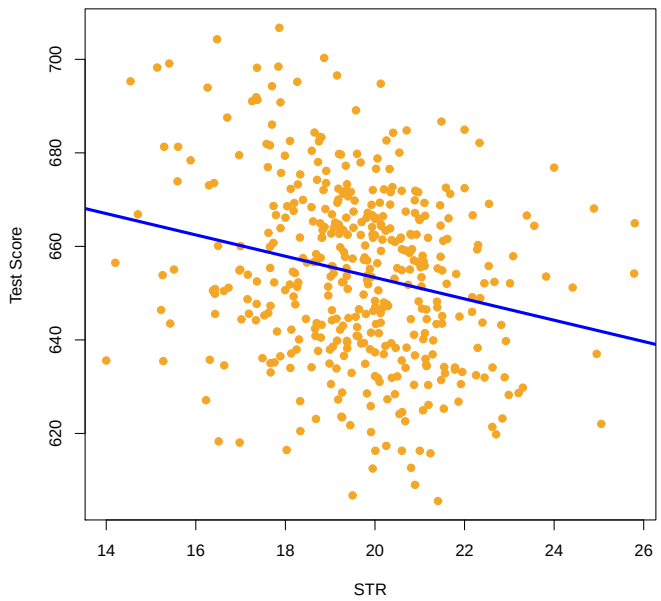


問題

- OLSE で推定したモデルは, 実際どのくらいデータにあてはまっているのか？
- あてはまりのよさを定量的に測る指標はないか？

推定したモデルのあてはまり

- OLSE : データに最もよくあてはまる
  推定量
- OLSE が実際にあてはまりが良いか悪いかは別問題
- はまりがよければ説明力の高いモデル
- あてはまりが悪ければ説明力の低いモデル



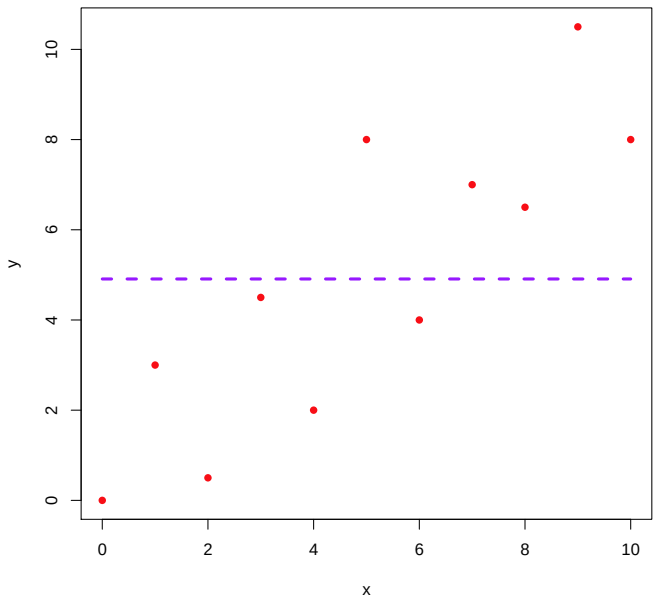
問題

- OLSE で推定したモデルは, 実際どのくらいデータにあてはまっているのか？
- あてはまりのよさを定量的に測る指標はないか？

総変動 (TSS : total sum of squares)

$$TSS := \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Y_i の平均からの乖離の二乗和
- データの分散 $\Leftrightarrow$ データの本来のばらつき



- 紫の点線：Y_i の標本平均

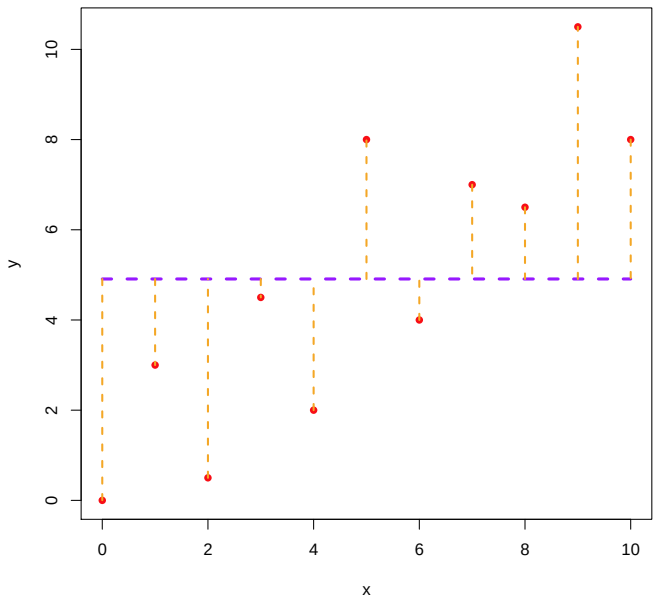
$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- 橙の点線：Y_i の平均偏差 $Y_i - \bar{Y}$
- TSS = 橙の点線の 2 乗和

総変動 (TSS : total sum of squares)

$$TSS := \sum_{i=1}^{n} (Y_i - \bar{Y})^2$$

- Y_i の平均からの乖離の二乗和
- データの分散 $\Leftrightarrow$ データの本来のばらつき



- 紫の点線：Y_i の標本平均

$$\bar{Y} = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

- 橙の点線：Y_i の平均偏差

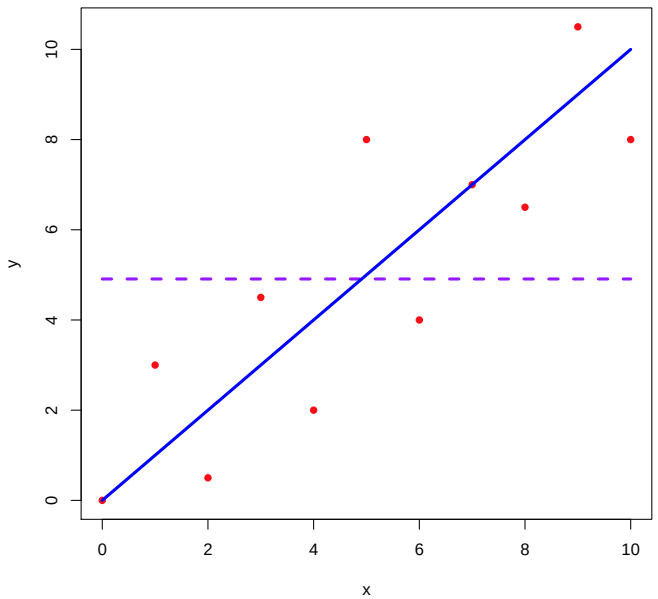
$$Y_i - \bar{Y}$$

- $TSS =$ 橙の点線の 2 乗和

回帰変動 (ESS : explained sum of squares)

$$ESS := \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- あてはめ値の平均からの変動 $\Rightarrow$ モデルの変動
- OLS 推定したモデルで説明できた変動の二乗和

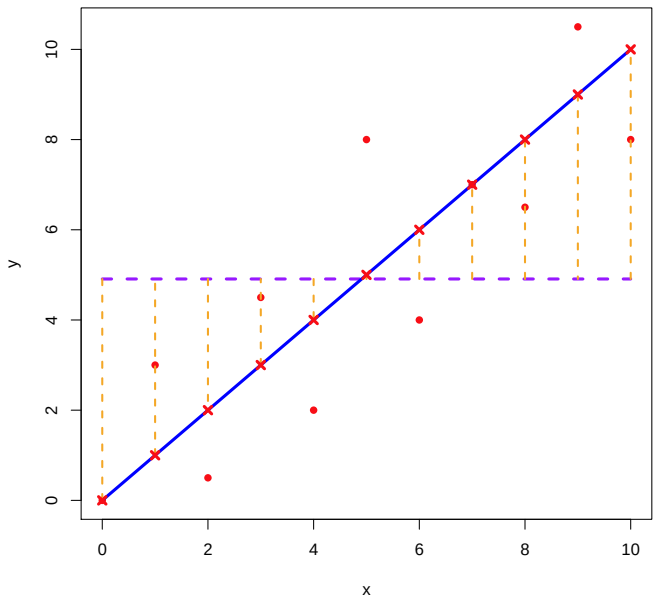


- 青の直線：OLSE による推定直線
- 橙の点線：あてはめ値の平均偏差 $\hat{Y}_i - \bar{Y}$
- ESS = 橙の点線の 2 乗和
- モデルが捉えた変動

回帰変動 (ESS : explained sum of squares)

$$ESS := \sum_{i=1}^{n} (\hat{Y}_i - \bar{Y})^2$$

- あてはめ値の平均からの変動 $\Rightarrow$ モデルの変動
- OLS 推定したモデルで説明できた変動の二乗和

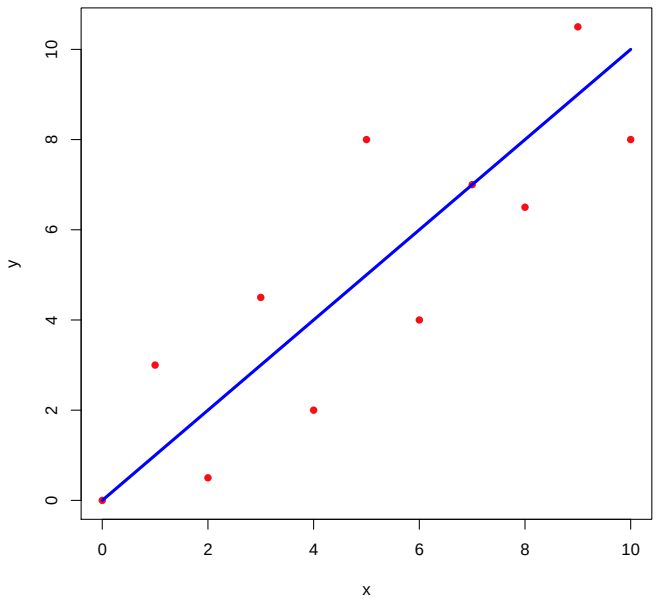


- 青の直線：OLSE による推定直線
- 橙の点線：あてはめ値の平均偏差
  $\hat{Y}_i - \bar{Y}$
- $ESS =$ 橙の点線の 2 乗和
- モデルが捉えた変動

残差変動 (RSS : residual sum of squares)

$$RSS = S_e^2 := \sum_{i=1}^{n} e_i^2$$

- Y_i の $\hat{Y}_i$ からの変動 $\Leftrightarrow$ 残差二乗和
- モデルをあてはめてもなお説明できなかった変動の二乗和

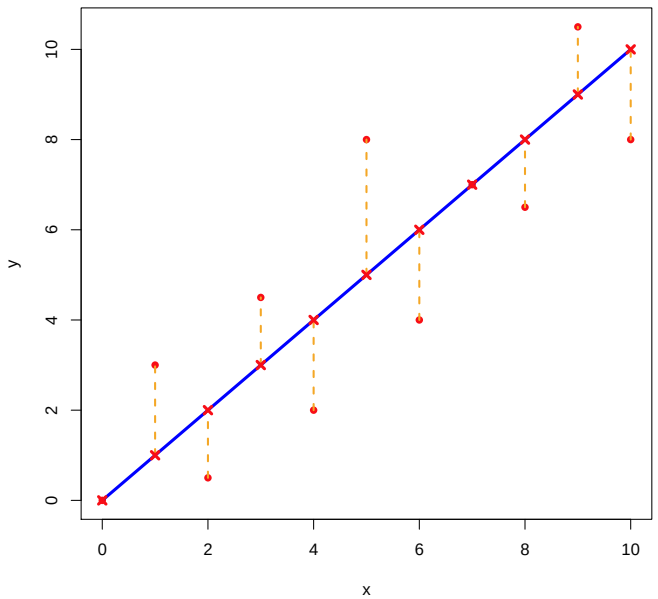


- 青の直線：OLSE による推定直線
- 橙の点線：残差
  $Y_i - \hat{Y}_i$
- $RSS =$ 残差 2 乗和

残差変動 (RSS : residual sum of squares)

$$RSS = S_e^2 := \sum_{i=1}^{n} e_i^2$$

- Y_i の $\hat{Y}_i$ からの変動 $\Leftrightarrow$ 残差二乗和
- モデルをあてはめてもなお説明できなかった変動の二乗和



- 青の直線：OLSE による推定直線
- 橙の点線：残差
  $$Y_i - \hat{Y}_i$$
- $RSS =$ 残差 2 乗和

三つの変動の関係

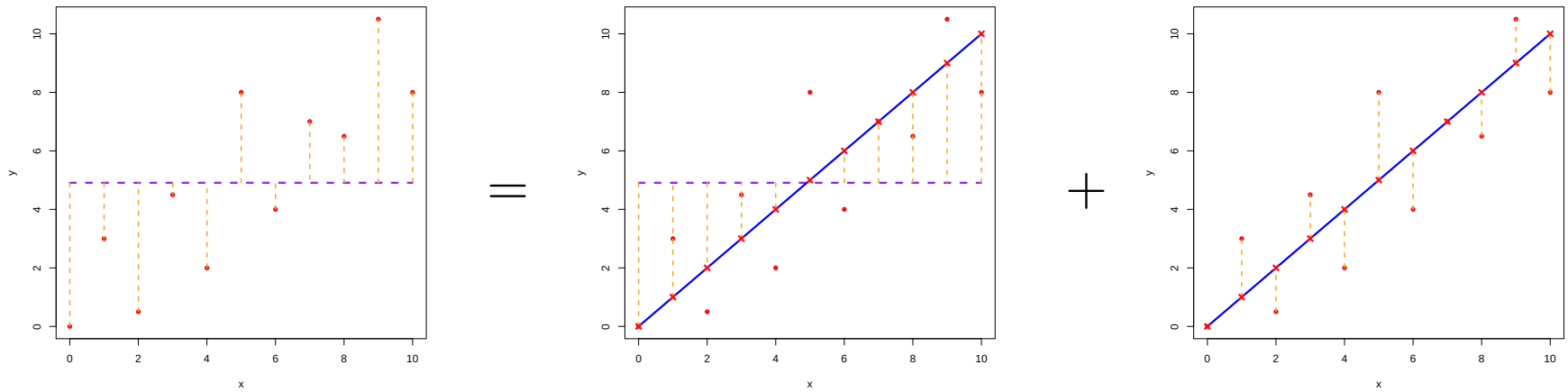
- TSS, ESS, RSS には以下の関係がある

$$TSS = RSS + ESS$$

$$\Updownarrow$$

全変動 $=$ モデルで説明できた変動

$+$ モデルで説明できなかった変動



TSS $=$ ESS $+$ RSS

決定係数 R^2

全変動に対する, 回帰で説明できた変動の割合

$$R^2 := \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

を決定係数 R^2 という

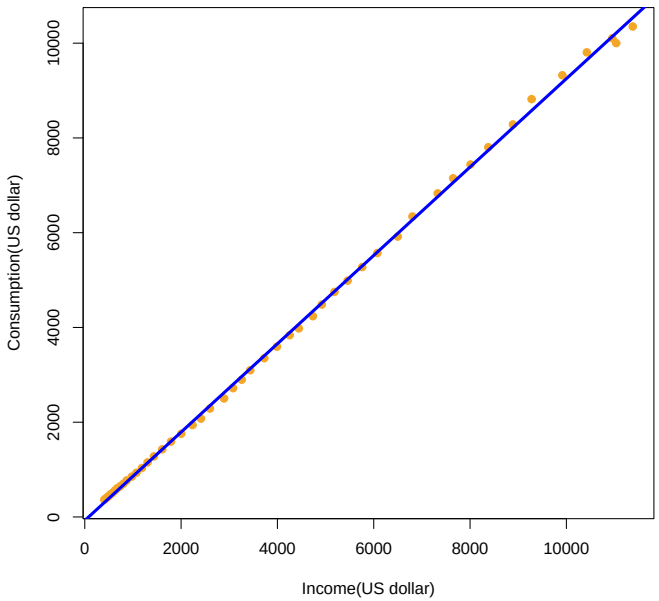
- データが本来持つ変動をモデルでどれだけ説明できたか
- あてはまりのよさを測る指標

決定係数 R^2 の性質

$$R^2 := \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

1. $0 \leq R^2 \leq 1$
2. R^2 が大 $\rightarrow$ 当てはまりがいい
   R^2 が小 $\rightarrow$ 当てはまりが悪い
3. $R^2 = 1 \Leftrightarrow RSS = S_e^2 = 0 \Leftrightarrow e_1 = \cdots = e_n = 0 \Leftrightarrow Y_i = \hat{Y}_i$
4. $R^2 = 0 \Leftrightarrow ESS = 0 \Leftrightarrow b_1 = \cdots = b_{K-1} = 0$
   $\rightarrow$ 説明変数の影響を受けていない
5. K (説明変数の数) を増やすと R^2 は増大する
   - 特に $K = n$ のとき, $S_e^2 = 0 \Leftrightarrow R^2 = 1$

アメリカの消費関数



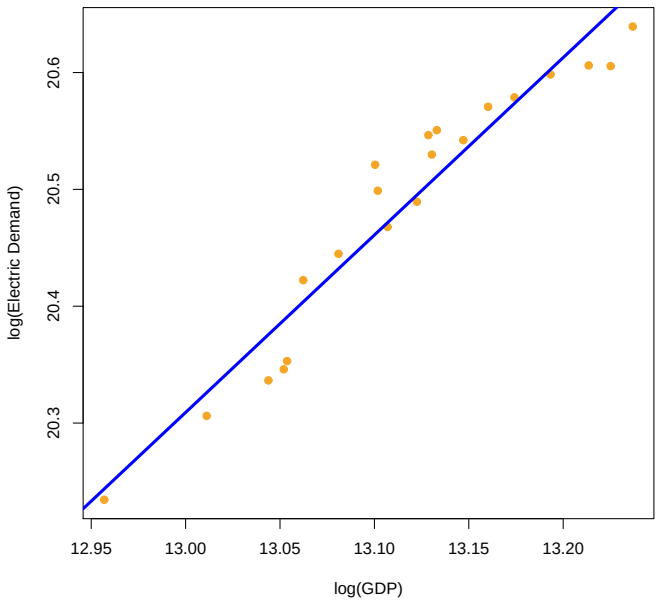
- 1962 年〜2010 年の米国における可処分所得と消費の関係

- モデル

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

- Y_i：消費支出, X_i：可処分所得
- $R^2 = 0.9993$
- データの変動の 99.93% をモデルで説明できている
- 非常にあてはまりのよいモデル

電力需要関数



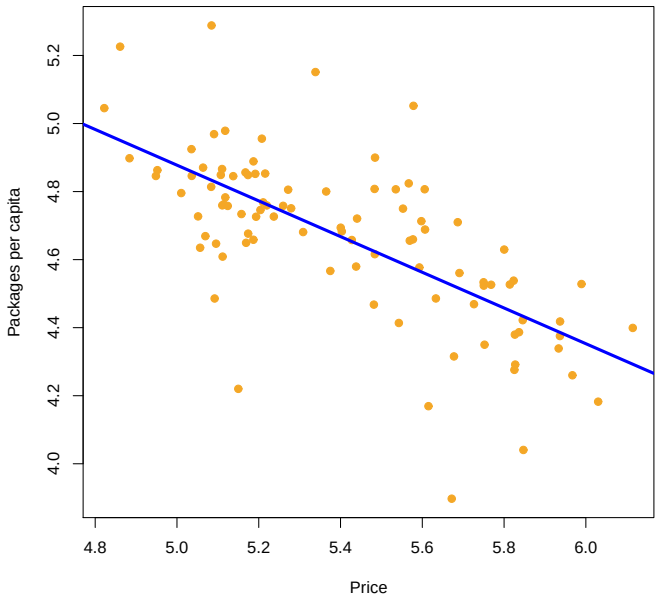
- 1980 年〜2009 年の日本における対数電力消費量と対数 GDP

- モデル

$$\log Y_i = \beta_0 + \beta_1 \log X_i + \epsilon_i$$

- Y_i：電力需要, X_i：GDP
- $R^2 = 0.9331$
- データの変動の 93.31% をモデルで説明できている
- 非常にあてはまりのよいモデル

たばこの価格と需要量の関係



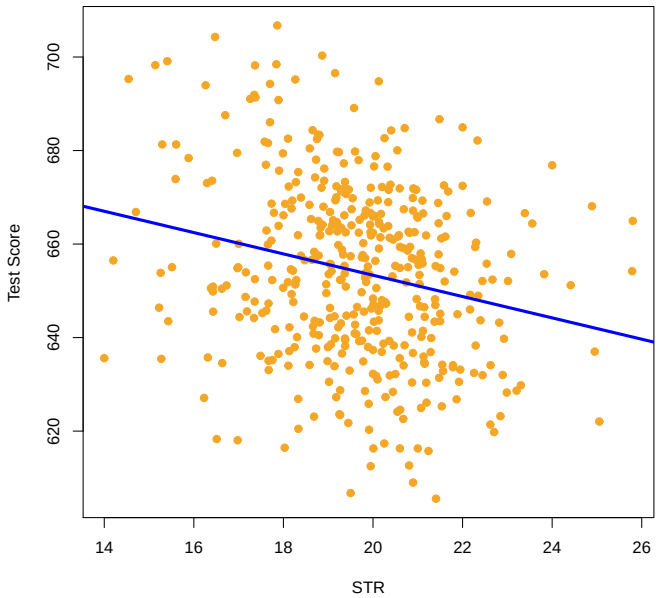
- 米国 48 州のたばこの平均価格と一人当たりの売上げ箱数 (1985 年と 1995 年)

- たばこ需要の弾性値モデル

$$\log Y_i = \beta_0 + \beta_1 \log X_i + \epsilon_i$$

- Y_i：たばこ需要, X_i：たばこ価格
- $R^2 = 0.4707$
- データの変動の 47.07% をモデルで説明できている
- 必ずしもいいモデルとは言えない

クラスの大きさと教育効果の関係



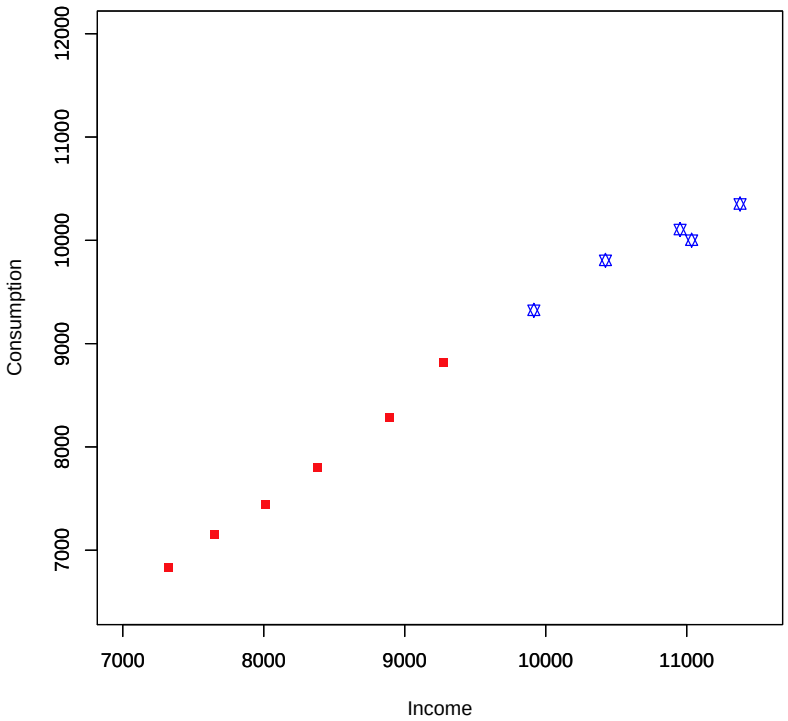
- カリフォルニア州の 420 校のテスト スコアと生徒 / 教師比 (STR)

- テスト スコアのモデル

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, i = 1, \ldots, n$$

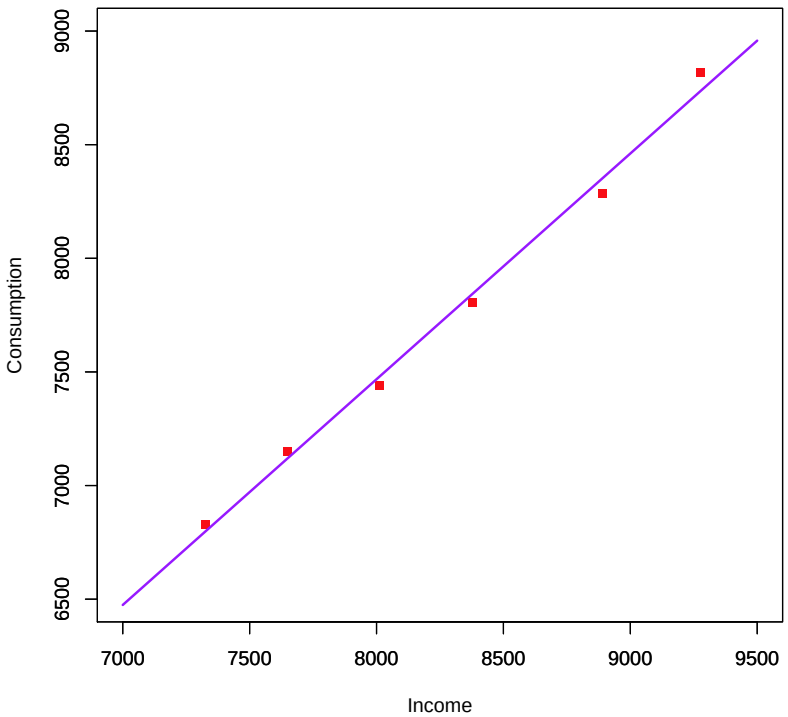
- Y_i：テスト スコア, X_i：生徒／教師比
- $R^2 = 0.05124$
- データの変動の 5.124% しかモデルで説明できていない
- ほとんどあてはまっていない
- モデルの再考の余地がある
  - 説明変数を増やす etc

あてはまりのよさとモデルのよさ



- データ：2000 年～2010 年までの
  アメリカの消費と所得
- 2000 年～2005 年までのデータ (赤) を用いて重回帰モデルを OLS 推定
- 推定したモデルを用いて, 2006 年以降の消費 (青) を予測して実測データと比較

説明変数の数と R^2

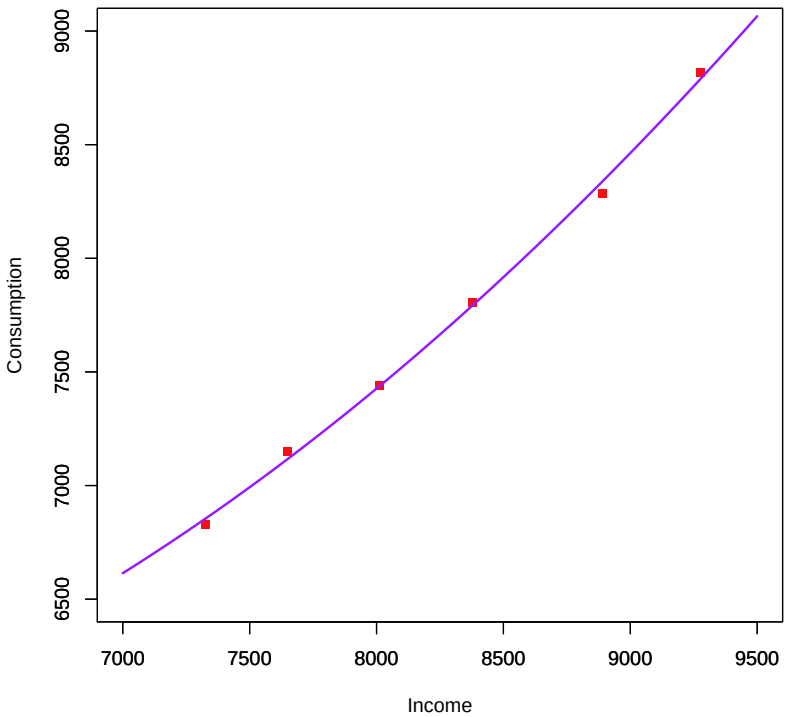


- 単回帰モデル

$$Y_i = \beta_0 + \beta_1 X_{i1} + \epsilon_i$$

- Y_i：消費支出, X_i：可処分所得
- $R^2 = 0.994$

説明変数の数と R^2

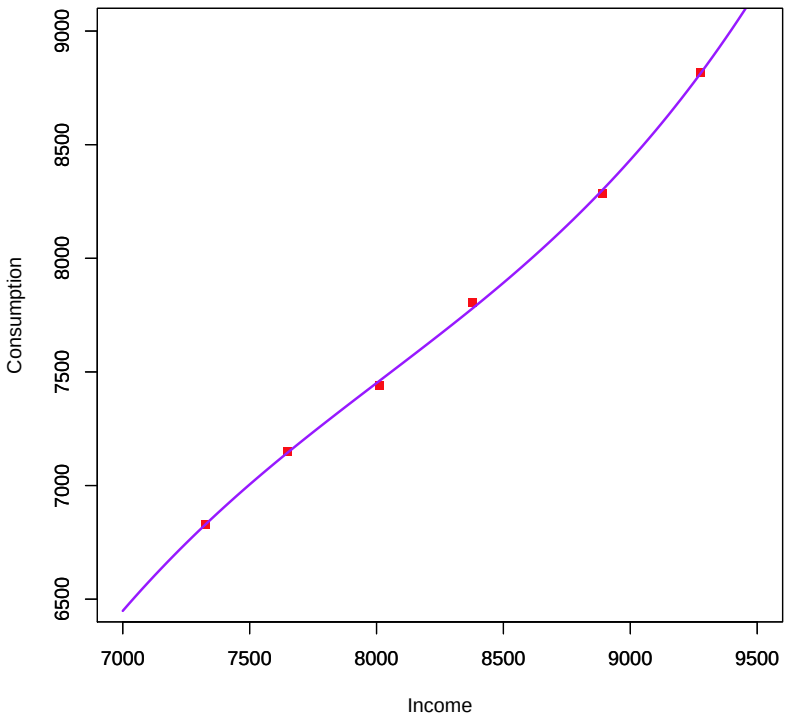


- 重回帰モデル

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \epsilon_i$$

- 2 次式のモデル
  - 1 番目の説明変数 X_{i1}
  - 2 番目の説明変数 $X_{i2} = X_{i1}^2$
- $R^2 = 0.998$
- 確かに R^2 は増加

説明変数の数と R^2

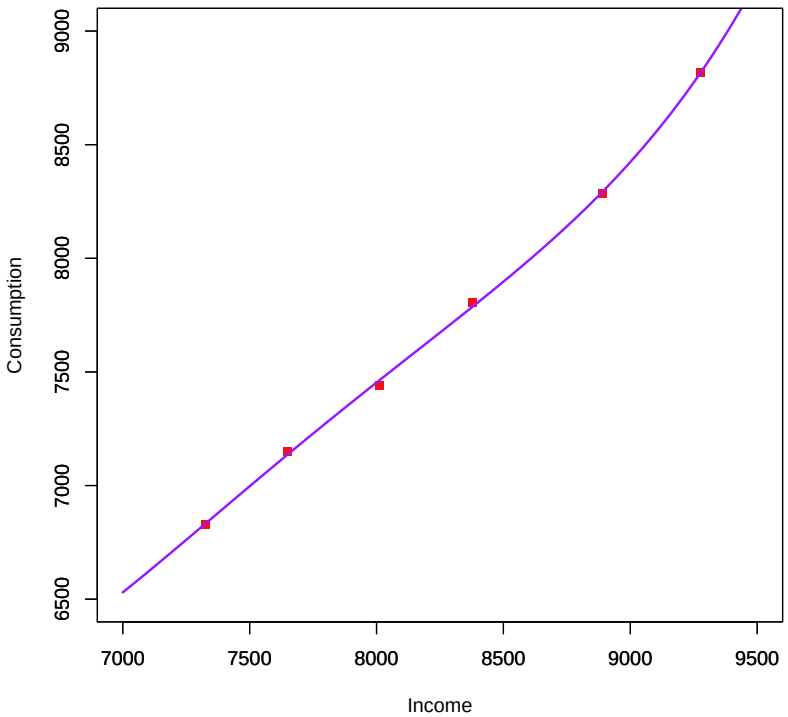


- 重回帰モデル

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i1}^3 + \epsilon_i$$

- 3 次式のモデル
  - 3 番目の説明変数 $X_{i2} = X_{i1}^3$
- $R^2 = 0.9995$
- R^2 はさらに増加

説明変数の数と R^2

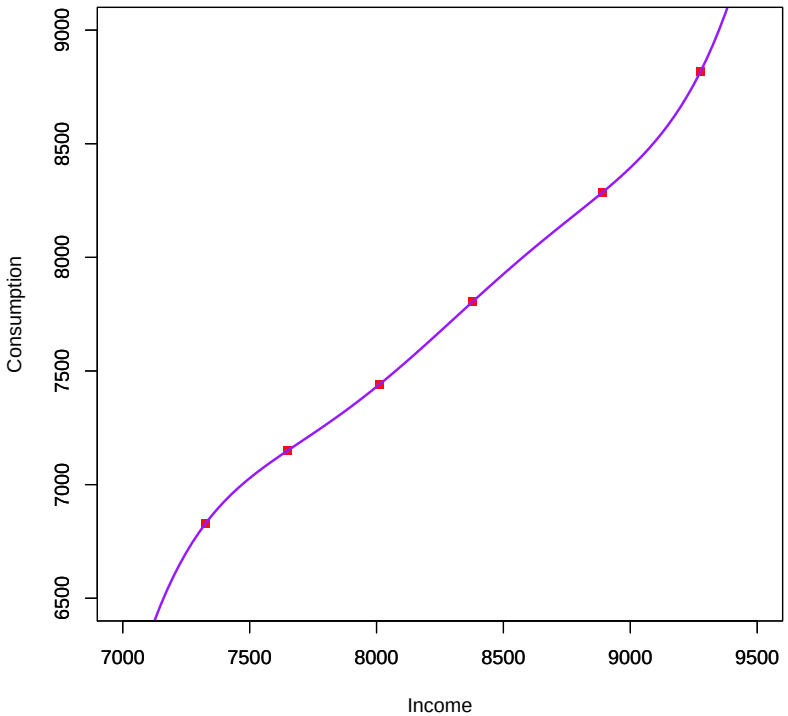


- 重回帰モデル

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i1}^3 + \beta_4 X_{i1}^4 + \epsilon_i$$

- 4 次式のモデル
  - 4 番目の説明変数 $X_{i2} = X_{i1}^4$
- $R^2 = 0.9996$
- R^2 はさらに増加

説明変数の数と R^2

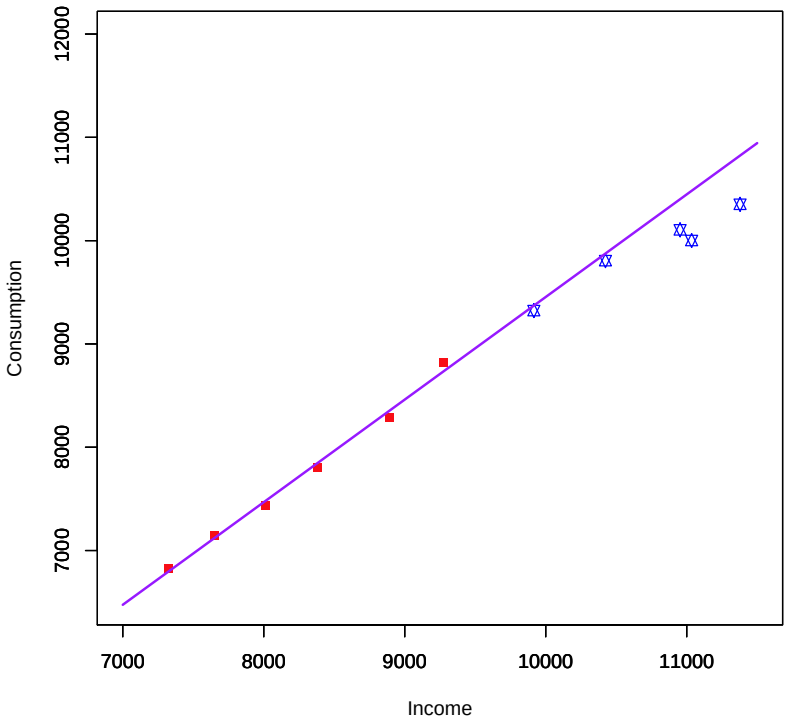


- 重回帰モデル

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i1}^3 + \beta_4 X_{i1}^4 + \beta_5 X_{i1}^5 + \epsilon_i$$

- 5 次式のモデル
  - 5 番目の説明変数 $X_{i2} = X_{i1}^5$
- $R^2 = 1$
- 残差はすべてゼロ
- OLSE はすべてのデータを通る 5 次曲線
- あてはまりのよいモデルはすぐれたモデルと言えるか？

あてはまりのよさと予測精度



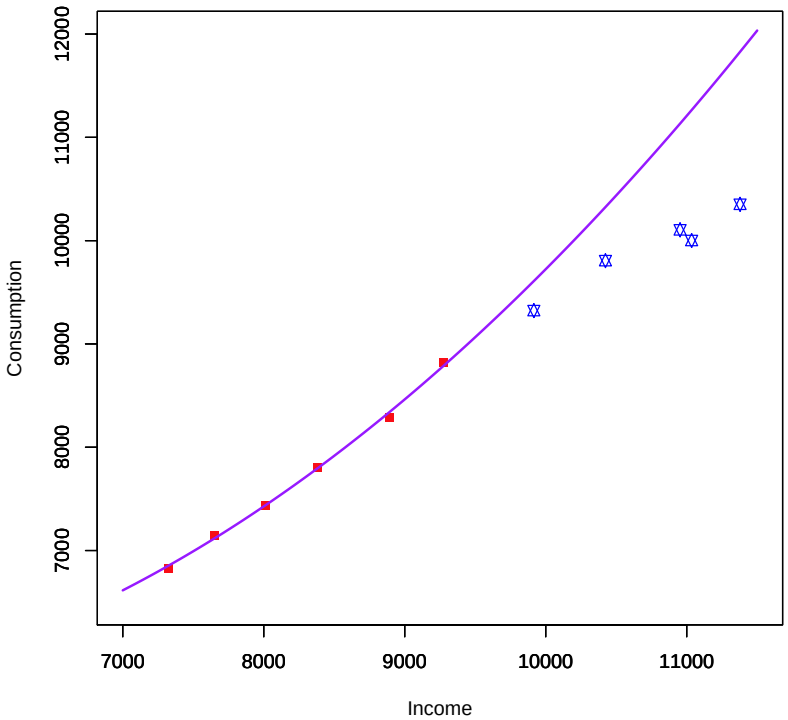
- 単回帰モデルの OLSE のあてはめ値

$$\hat{Y}_i = b_0 + b_1 X_{i1}$$

で予測

- まあまあの精度で予測できている

あてはまりのよさと予測精度

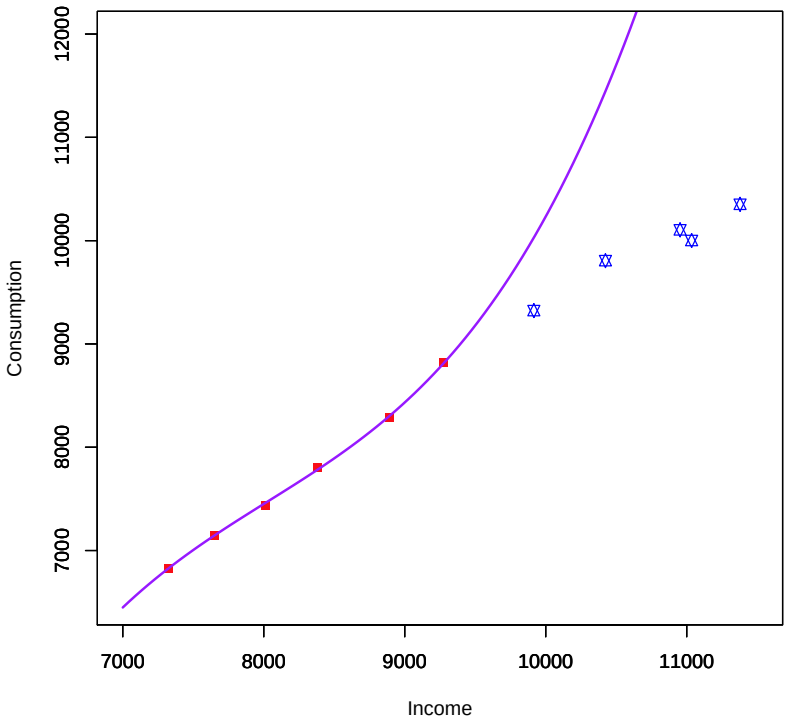


- 2 次式のモデルの OLSE のあてはめ値

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i1}^2$$

  で予測
- 実測値と予測値が離れている
- 予測の精度が低下

あてはまりのよさと予測精度



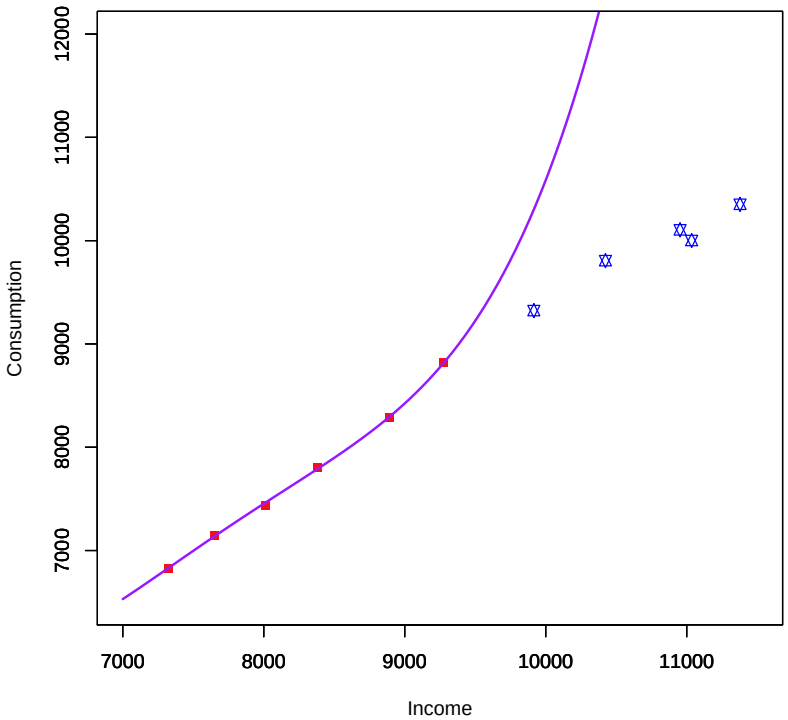
- 3 次式のモデルの OLSE のあてはめ値

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i1}^2 + b_3 X_{i1}^3$$

で予測

- 予測の精度がさらに低下

あてはまりのよさと予測精度



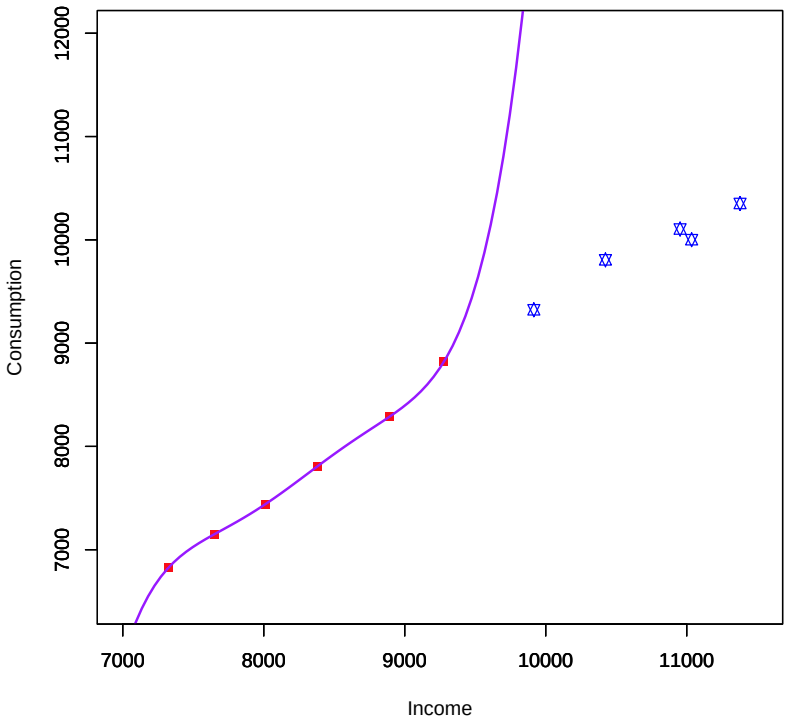
- 4 次式のモデルの OLSE のあてはめ値

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i1}^2 + b_3 X_{i1}^3 + b_4 X_{i1}^4$$

で予測

- 予測の精度がさらに低下

あてはまりのよさと予測精度



- 5 次式のモデルの OLSE のあてはめ値

$$\hat{Y}_i = b_0 + b_1 X_{i1} + b_2 X_{i1}^2 + b_3 X_{i1}^3 + b_4 X_{i1}^4 + b_5 X_{i1}^5$$

  で予測
- まったく使いものにならない
- 推定に用いたデータの範囲内におけるあてはまりはよいが, 予測精度は極端に悪い
- あてはまりがよくても, 説明変数の数が多いモデルは, 必ずしもよいモデルではない

あてはまりのよさとモデルのよさ

- 一般に重回帰モデルを用いて実証する場合, 複数のモデルの候補 (説明変数の組の候補) から最もよいモデルを選ぶ必要がある
- しかし R^2 は説明変数を増やしさえすれば大きくなるので, 最適なモデルの選択の指標としては不適切
- 実際に説明変数を増やすと, R^2 は増加するが, 予測の精度は極端に悪くなることがある (汎化能力が低下する)
- 説明変数の数が等しいモデル間のモデルの好ましさの指標としては有用なこともある