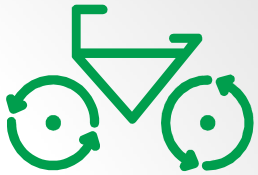
Bike Sharing Demandを用いた分析演習

2015年1月

濱岡豊研究会13期

佐藤龍治　　益田航暉　　金悠佳

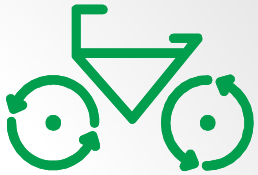
研究の要約



- 本研究ではワシントンにおける自転車の貸し出し制度の利用傾向に関する分析を行った。この際、分析には、当初kaggle.comから提示されていた「日時」、「気温」、「天気」、「季節」といった変数に加え、変数同士を組み合わせたものも用いて分析をすることで最適モデルの探索を試みた。

分析の結果、季節ダミーと体感気温を組み合わせた分析において残差は最も小さくなったが、依然として改良の余地を残すことになった。また、一日を1時間ごとに分けて利用者数を考えることはモデルへの影響力が大きいことが判明した。

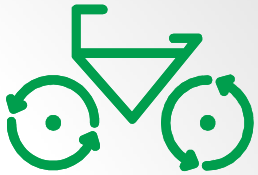
目次



研究目的
単純集計
分析１：基礎的分析
分析２：季節・風速に着目した分析
分析３：時間帯の区分・体感気温に着目した分析
分析４：回帰診断の結果に着目した分析
考察・まとめ
モデルの提出
課題

研究目的

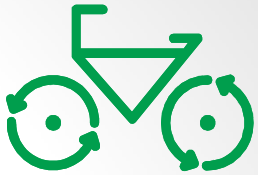
本演習の目的



- 本研究ではKaggle.com(http://www.kaggle.com/)に提供されているデータを使用し、データ分析や予測モデルの作成のノウハウを学習することを目的とする。

- Kaggle.comとは、企業や研究者がデータをサイト内で提供し、統計家やデータ分析家がその最適モデルを競い合うサイトである。会員登録をすればデータは入手・分析可能であり、企業が不特定多数の人に分析やモデルの作成を委託するというクラウドソーシングを行う場をKaggleが提供している。

研究目的

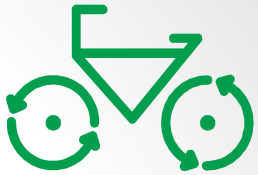


- Kaggle.comは設定した課題は、 ワシントンD.Cが運営する自転車レンタルのデータを推定用データ(1-20日)と予測用データ(20日-月末)に分け、予測用データにおいて利用傾向をより説明できるモデルを推定用データで作成することである。

- よって今回の研究では、複数のモデルを考えることによってより適合性が高いモデルを構築し、bike sharingの今後の利用を予測できるモデルを探索することを研究目的とする。

単純集計

元より設定された変数



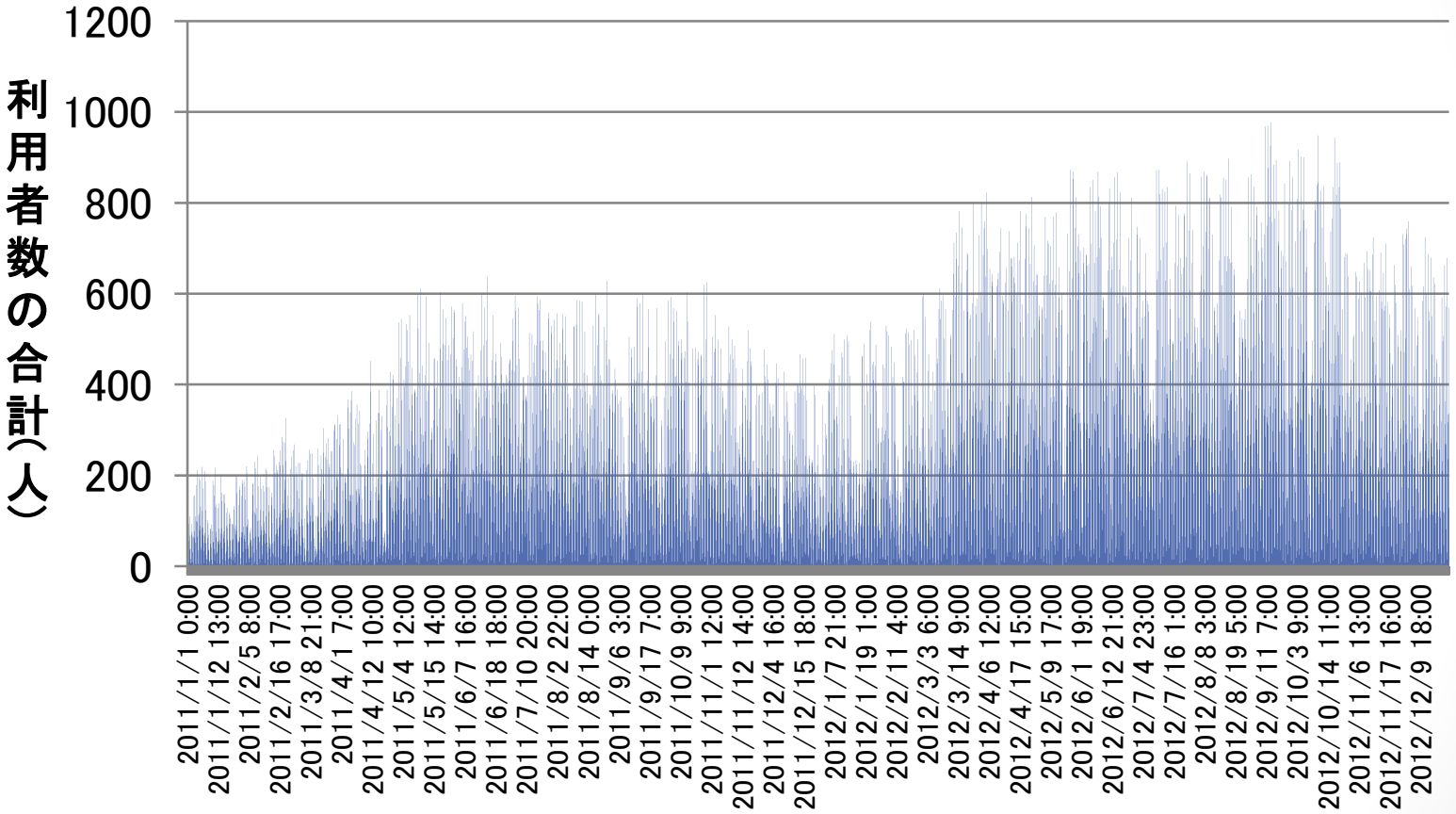
被説明変数

count: Bike Sharingの一時間当たりの利用者数の合計である。
また、casualとregisteredを合計するとcountと一致。

説明変数	意味	説明
datetime	日時	期間は2011-2012年 1時間単位
season	季節	1-3月を1、4-6月を2、7-9月を3、10-12月を4と割り当て
holiday	祝日	祝日かどうか
workingday	出勤日	出勤日であるかどうか
weather	天気	1:晴れ 2:霧or曇り 3:弱い雨or弱い雪 4:豪雨or豪雪
temp	気温	摂氏何度であるか
atemp	体感気温	体感気温が摂氏何度であるか
humidity	湿度	湿度は何%であるか
windspeed	風速	風速は何mであるか
casual	非登録利用者	事前にWebにて登録をしていない利用者(旅行者？)クレジットカード払いである
registered	登録利用者	事前にWebにて登録をしている(居住者？)クレジットカード払いではない

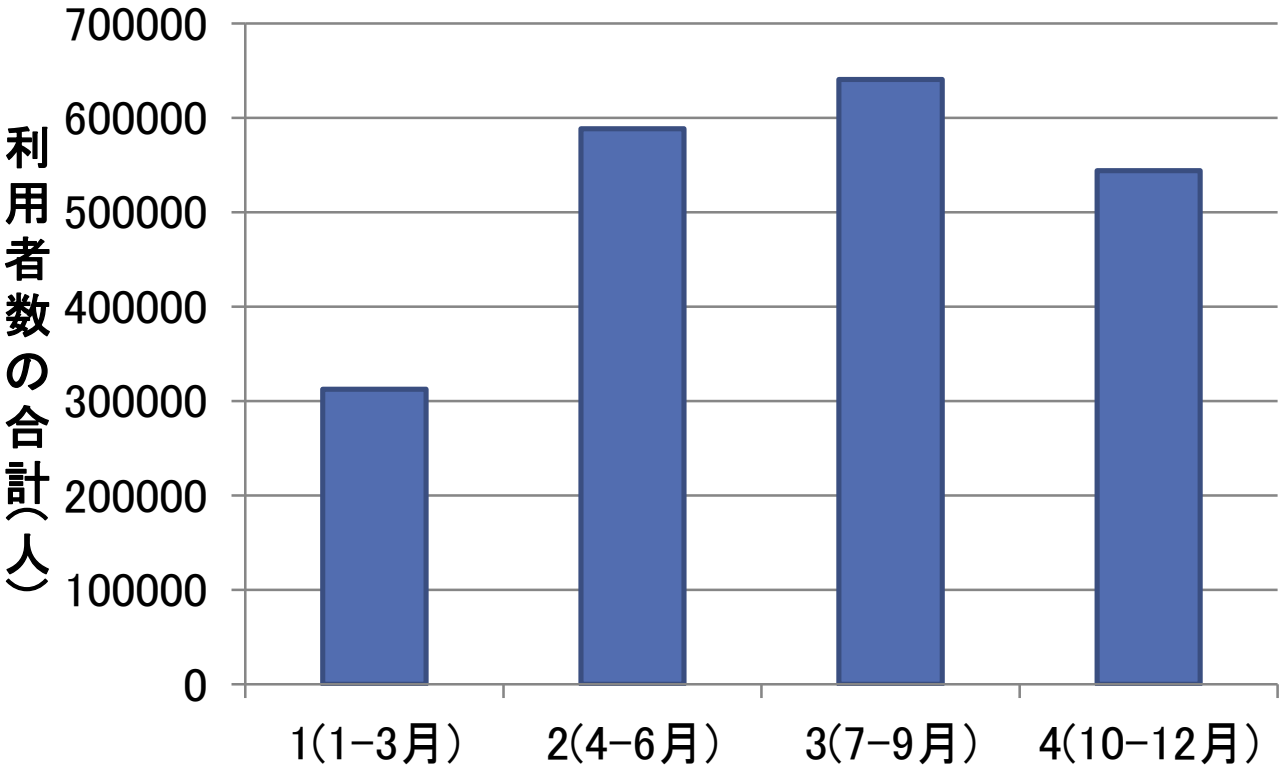
図表１　既存の変数一覧

日時と利用者数の合計のグラフ



図表２　日時・利用者のグラフ

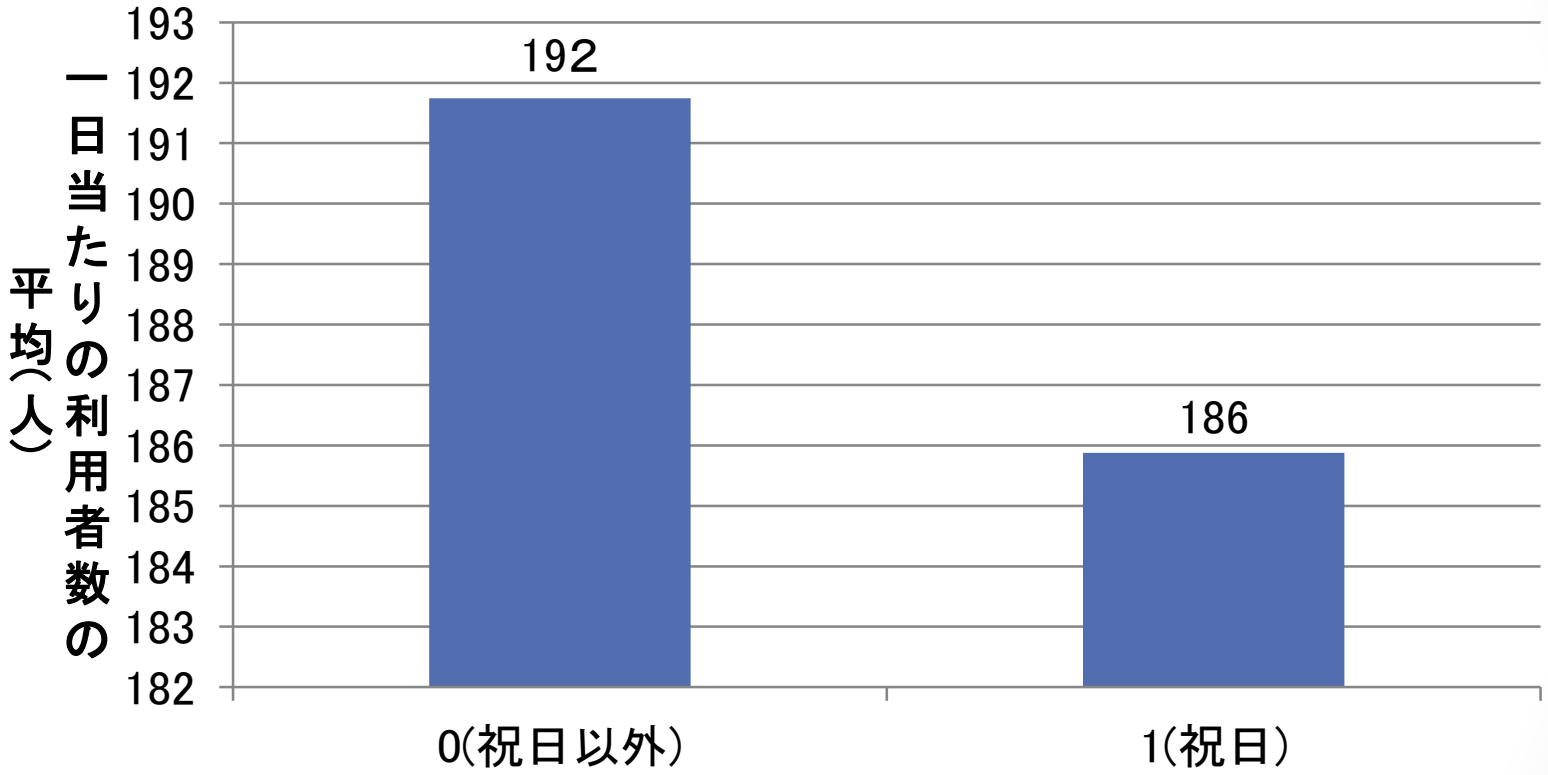
季節と利用者数の合計のグラフ



季節を３か月ごとに分類した単純集計である。夏の７～９月に最も利用者数は多く、冬の１～３月に最も利用者数は少なくなっている。

図表３　季節・利用者のグラフ

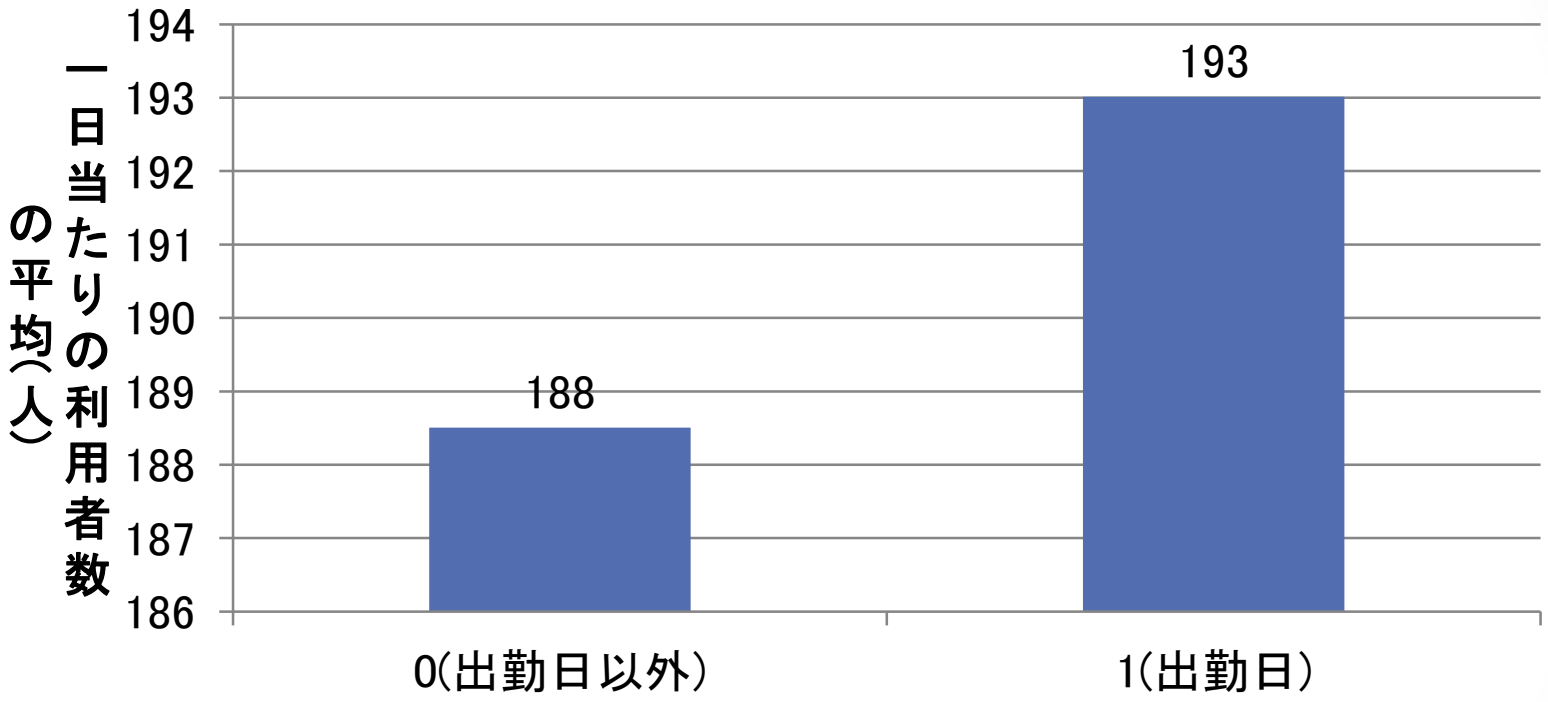
祝日と利用者数の平均のグラフ



ダミー変数の形式をとっている。一日当たりの利用者数の平均がほぼ同じであることが分かる。

図表４　祝日と利用者数のグラフ

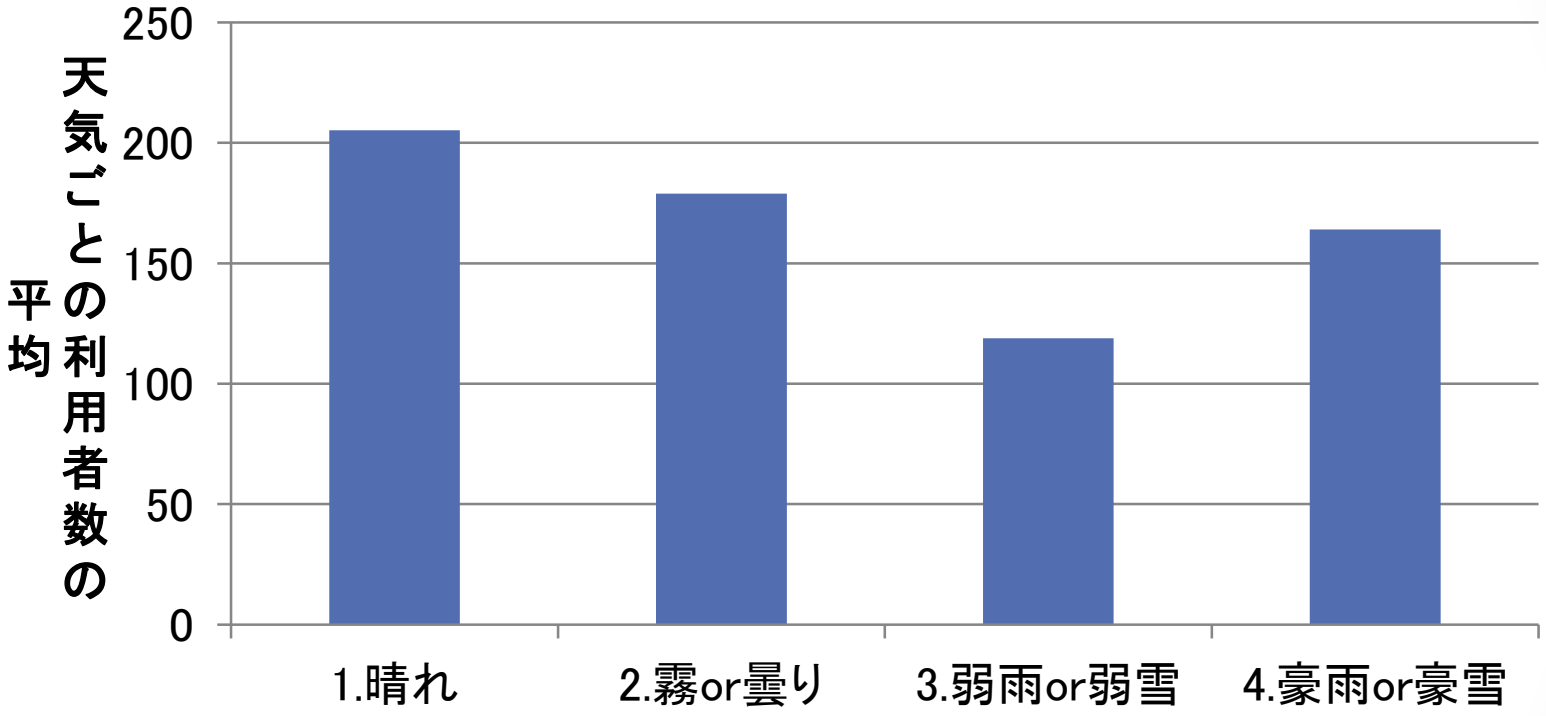
出勤日と利用者数の平均のグラフ



ダミー変数の形式をとっている。出勤日とそれ以外の日で、利用者数の違いがほとんどないことが分かる。

図表５　出勤日と利用者数のグラフ
縦軸：利用者数（人）、横軸：出勤日

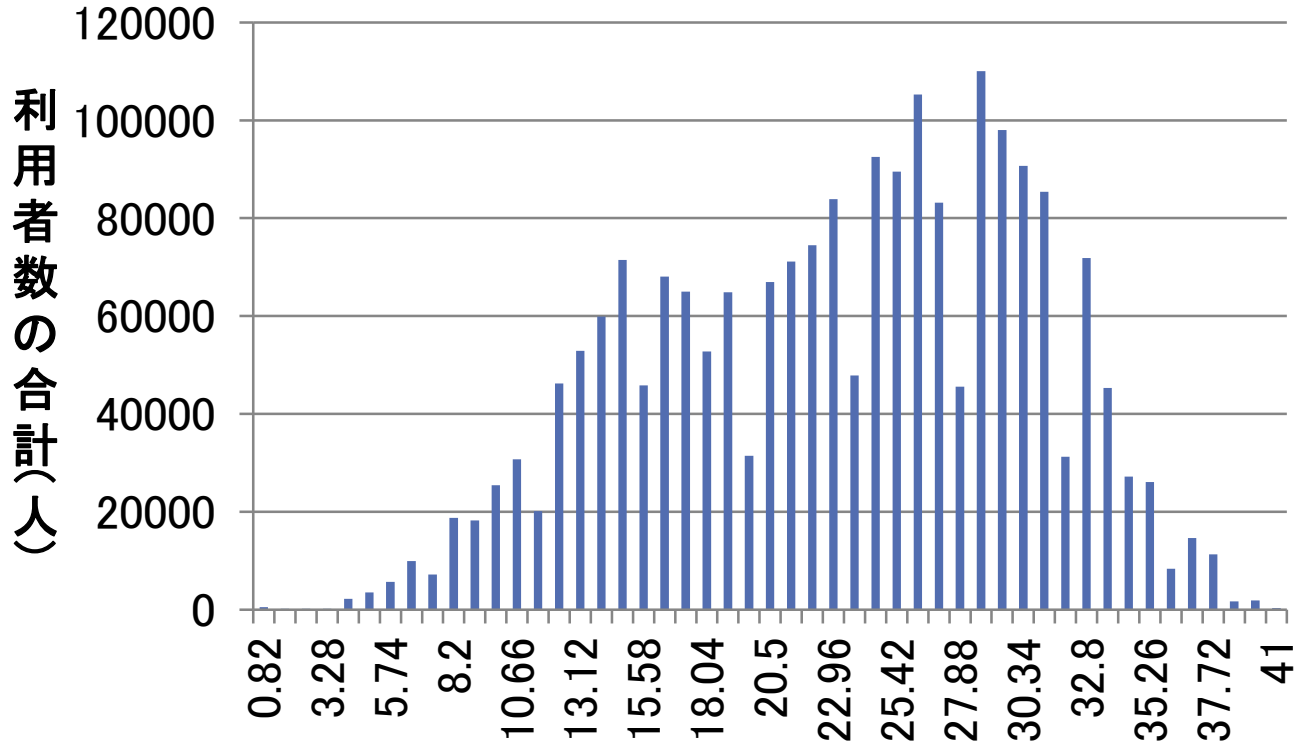
天気と利用者数の平均のグラフ



縦軸：天気ごとの利用者数の平均（人）、横軸：天気
一般に天気が良いほど利用者数が多い傾向が見られる。

図表６　天気と利用者数のグラフ）

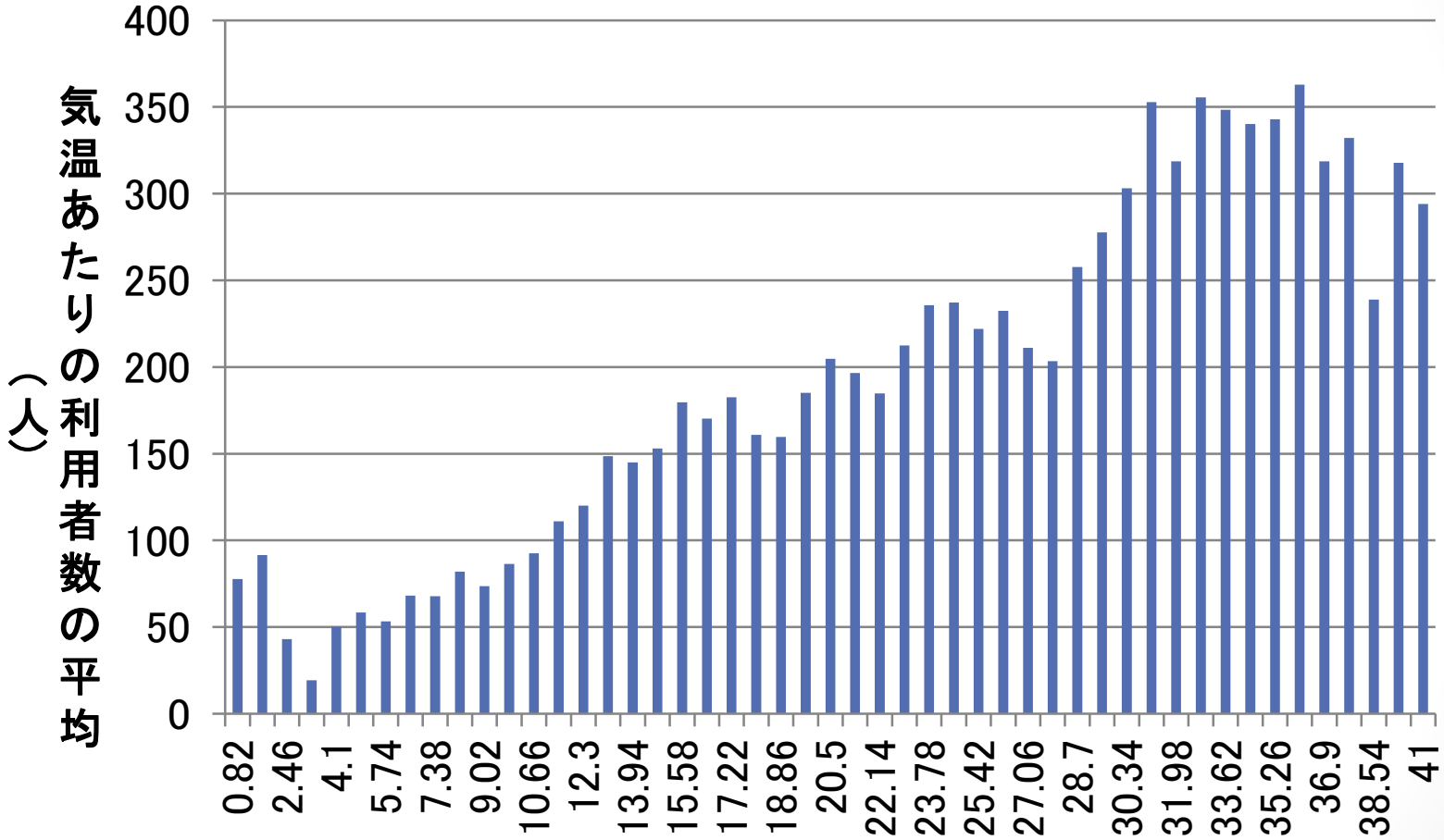
気温と利用者数のグラフ



およそ２５～３０℃あたりで利用者数は一番多くなっている。
一方で２７．８８あたりの一点で利用者数は落ち込んでいる。

図表７　気温と利用者数の合計のグラフ

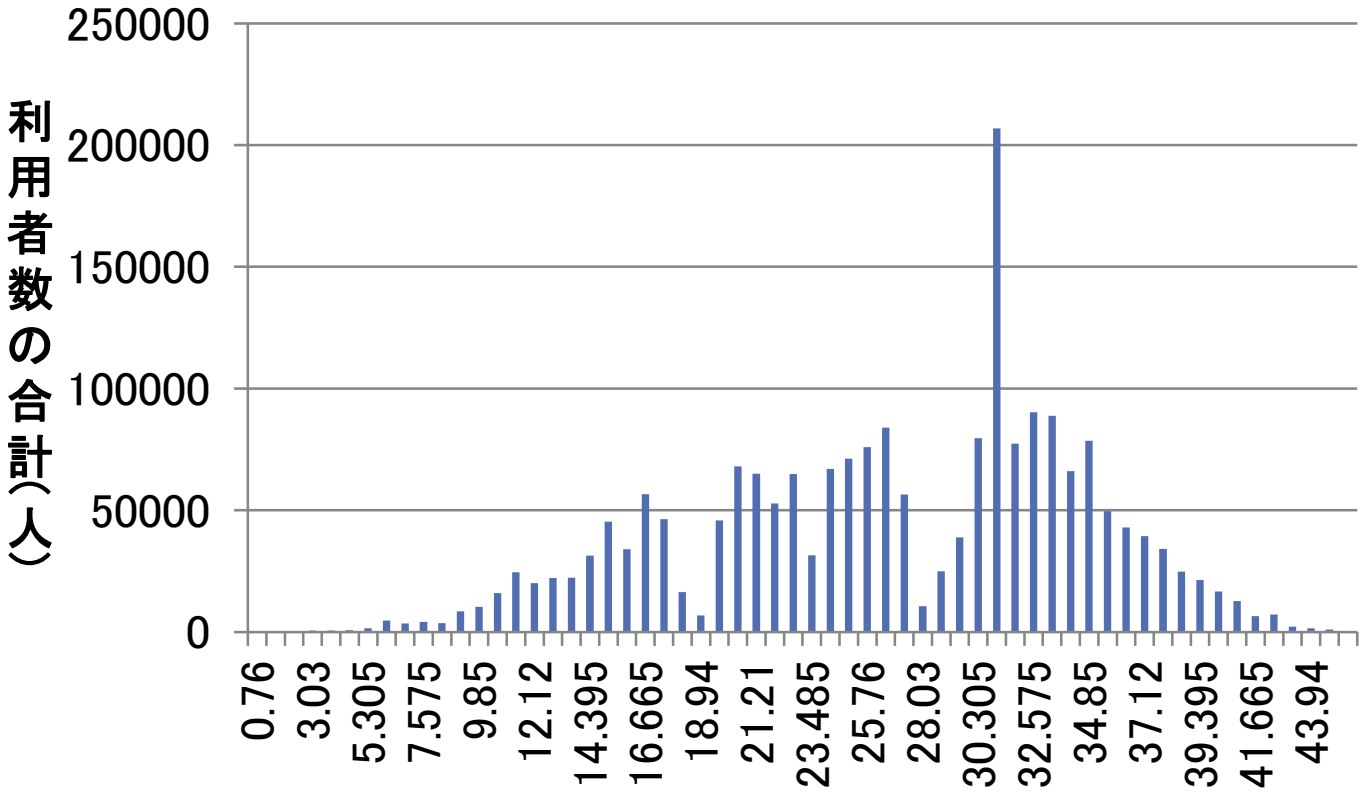
気温と利用者数の平均のグラフ



一般に気温が高くなるほど利用者数の平均の値も増加していることが分かる。

図表８　気温と利用者数の平均のグラフ

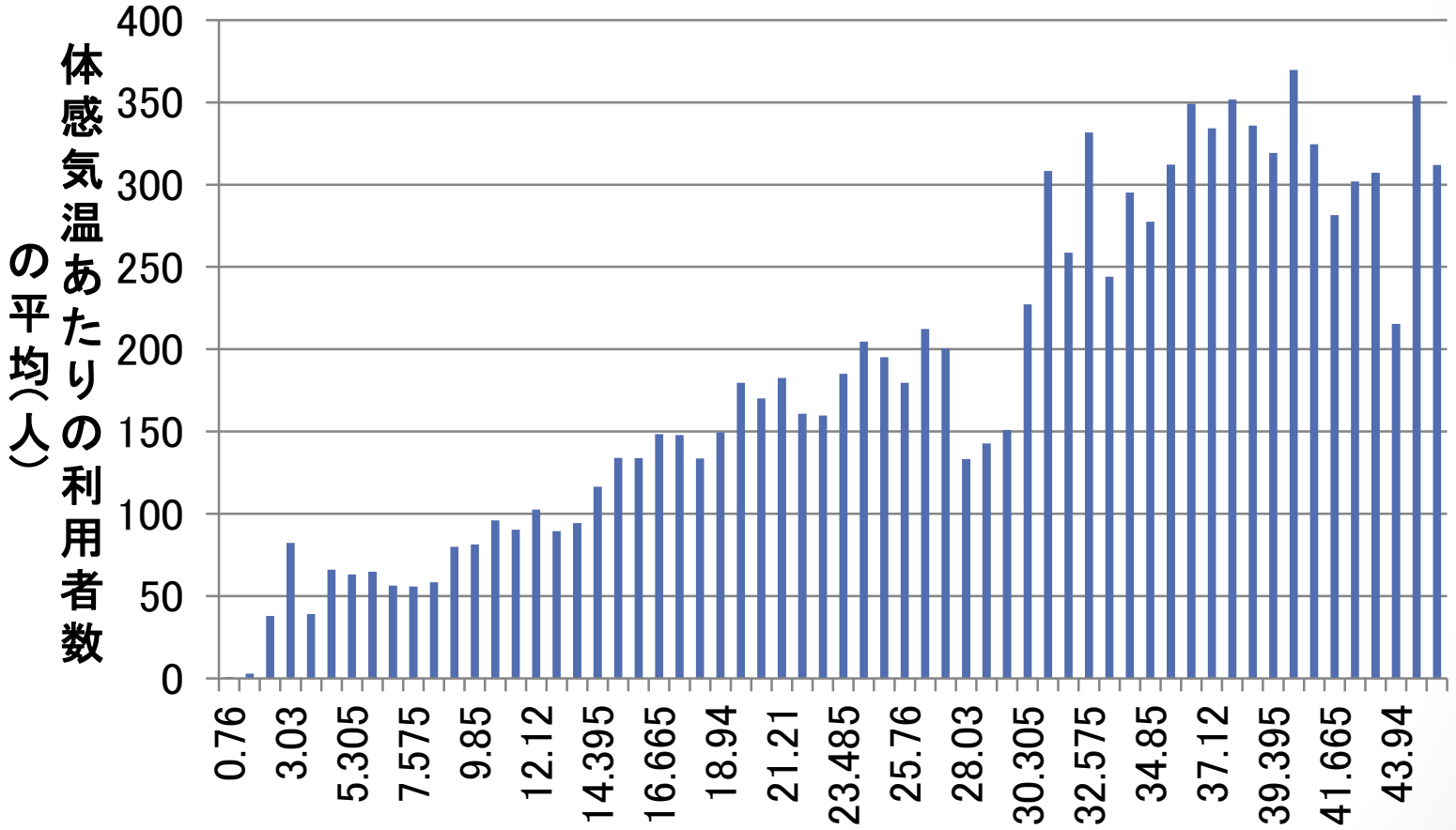
体感気温と利用者の合計のグラフ



縦軸に利用者数、横軸に体感温度をとる。
３０～３２（℃）の一点で突出して利用者数の合計が多いところがある。

図表９　体感気温と利用者数の合計のグラフ

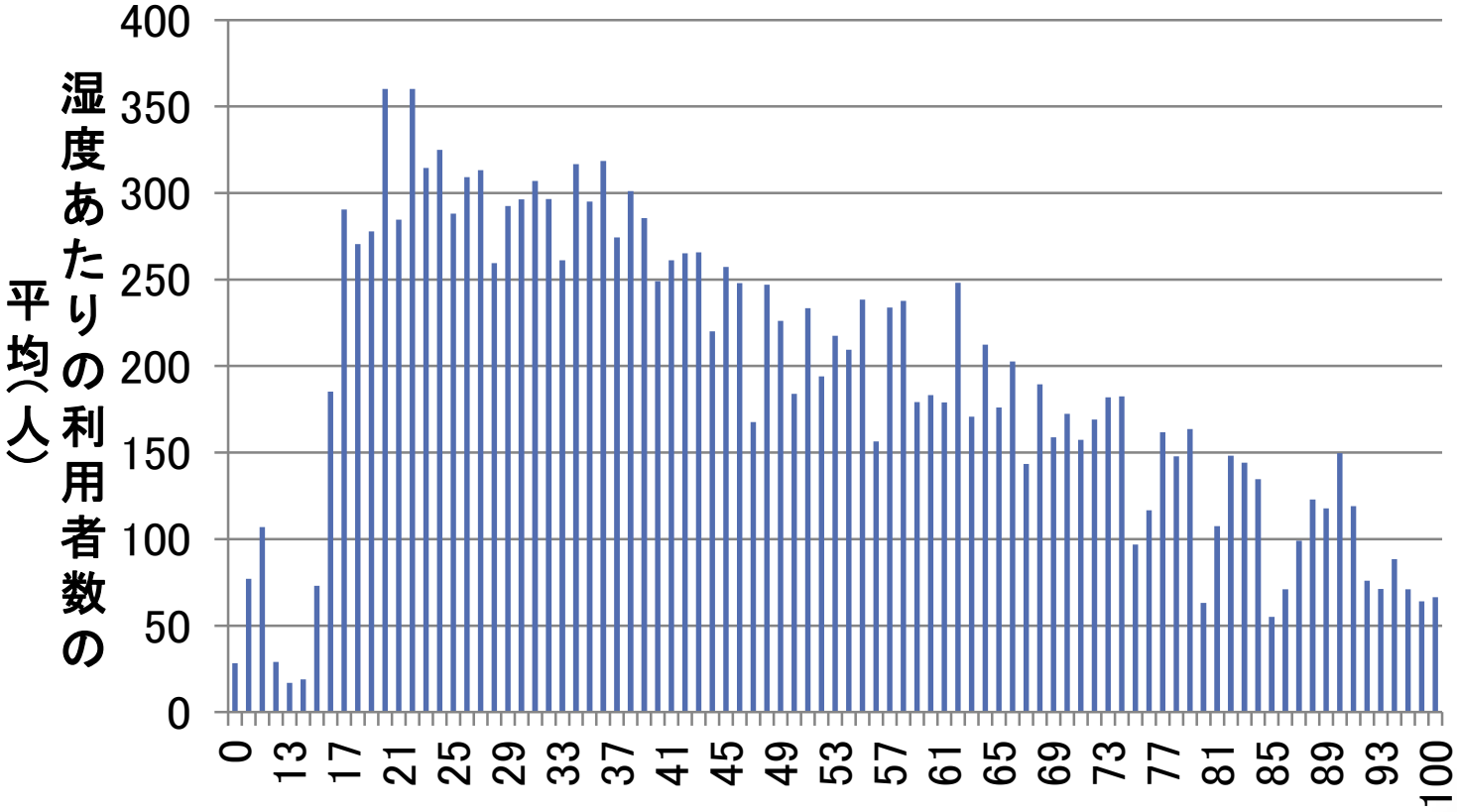
体感気温と利用者数の平均のグラフ



体感気温が上がるほど、利用者数の平均の値も増加することが分かる。

図表１０　体感気温と利用者数の平均のグラフ

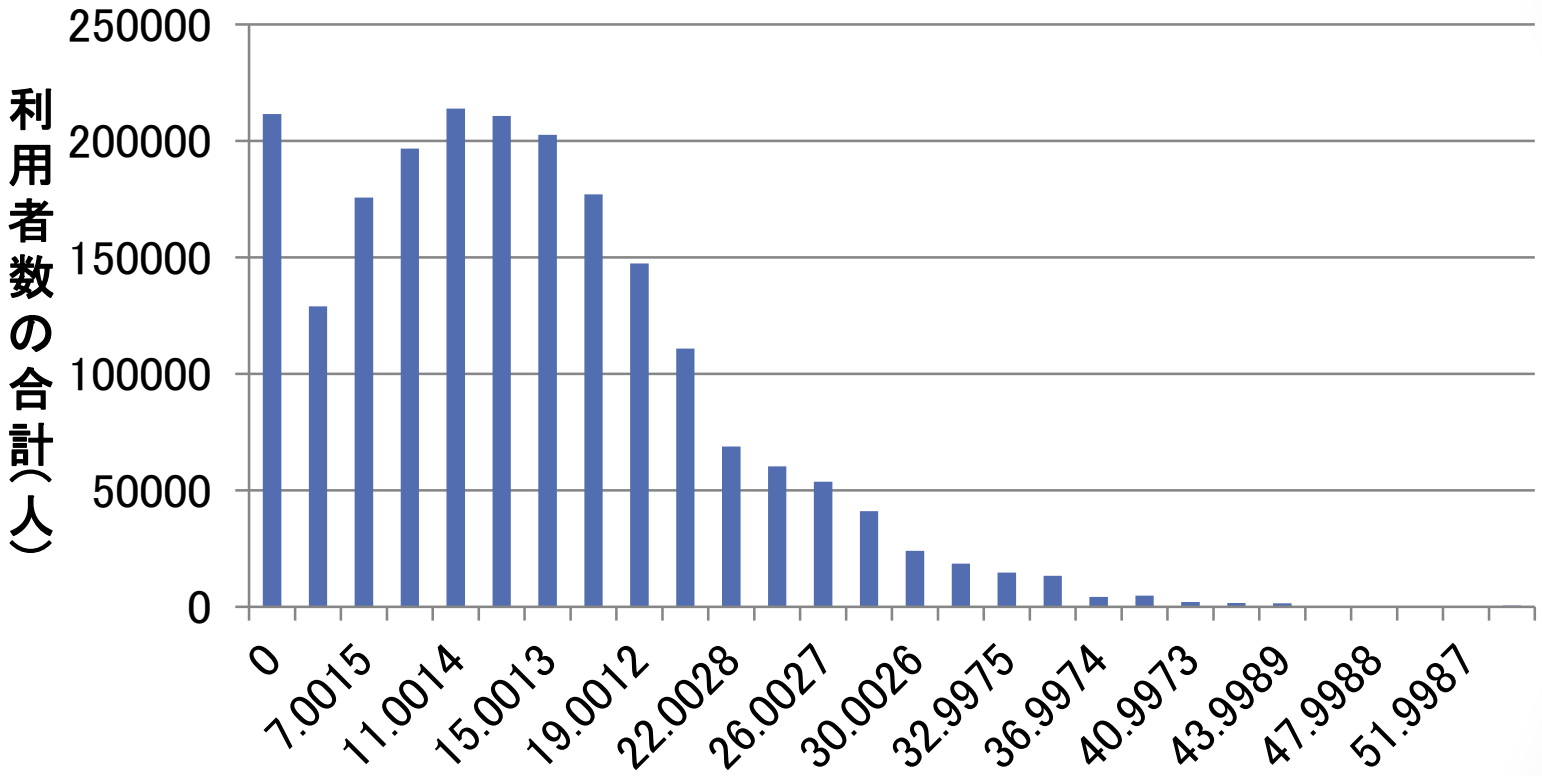
湿度と利用者数の平均のグラフ



湿度20%前後で平均は最大となり、その後減少していることが分かる。

図表１１　湿度と利用者数のグラフ

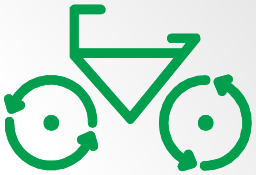
風速と利用者数の合計のグラフ



右下がりの傾向が見られ、風が強くなるほど利用者数が減少している。

図表１２　風速と利用者数の合計のグラフ

登録利用者と非登録利用者について

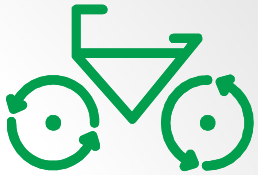


- 調査期間中におけるBike sharingの全利用者は
  208万5476人

- そのうち、
  登録利用者が169万3361人
  非登録利用者が39万2135人

- 登録者が約8割を占めているため、基本的には
  利用者の属性を考慮せず、全利用者(count)を用いる

単純集計の考察



- 2011年よりも2012年の方が利用者が多い
  →年々認知度が上がっているのではないか

- 祝日、出勤日という要素による利用者数の変化は小さい

- 体感気温の集計において、31.06度時の利用者数の合計が突出している
  →利用者数の平均を見ると値は突出していない。31.06度を記録した日が多いためと考えられる。

分析１：基礎的分析

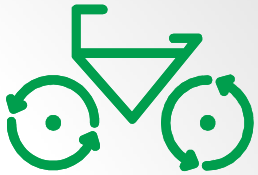
変数の追加・変更１

祝日と出勤日の変数は元々ダミー変数であり、
前回までに季節ダミーと天気ダミーは作成済みのため、他の変数についても適宜ダミー化。

①年ダミー

データは2011年-2012年の２年間において採られたものである。年ごとに利用者の動向は異なると考えられるため、年号の違いを反映させる目的で設定。
→2011年のデータに対しyear1というダミー変数を作成。

変数の追加・変更１



②月ダミー

月ごとに利用者の動向が異なると考えられるため、月の違いの影響力を反映させるために設定。

→２月を除き、１月から順にmonth1，month3-month12の計11個のダミー変数を作成。

③曜日ダミー

曜日ごとに利用者の動向が異なると考えられるため、曜日の違いの影響力を反映させるために設定。

→土曜日を除き、日曜から順にday1-day6の計６個のダミー変数を作成。

変数の追加・変更１

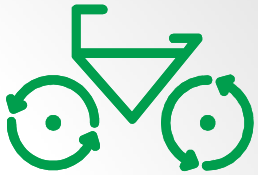
④1時間ダミー

1時間ごとに利用者の動向が異なると考えられるため、時間帯の違いの影響力を反映させるために設定。

→２３時を除き、０時から順にtime0-time22の計２３個のダミー変数を作成。

元の変数を使ったモデルと変数追加後のモデルの適合度がどのように変化するかを確認する。

分析１



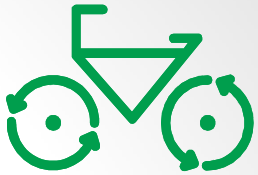
- 分析はポアソン回帰分析を用いる。

- モデルの適合度の比較は、分析結果の中から Residual devianceの値を確認する。

　Residual deviance＝逸脱残差(誤差)のことであり、この値が小さいほど予測値(モデル)が観測値(結果)と誤差が小さいことを表す。

→Residual devianceの値が小さいほど良いモデル

分析１



①元々設定されている変数のみ使用

glm(formula = 利用者数 ~ 季節 + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance: 1312614 AIC: 1382279

この数値よりもResidual devianceの値がより小さくなるようなモデルを探していく。

※分析１における以後のモデルは全て①から変更している

分析１

①の推定結果表

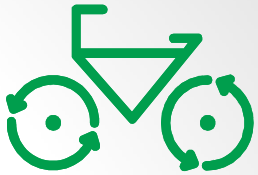
変数名	係数	標準誤差	z値	P値
季節	1.534e-01	7.200e-04	212.989	< 2e-16 ***
祝日ダミー	-3.536e-02	4.364e-03	-8.103	5.36e-16 ***
出勤日ダミー	1.869e-03	1.544e-03	1.210	0.226
天気	1.266e-02	1.289e-03	9.820	< 2e-16 ***
気温	-2.335e-03	5.425e-04	-4.304	1.68e-05 ***
体感気温	3.949e-02	5.044e-04	78.278	< 2e-16 ***
湿度	-1.559e-02	4.314e-05	-361.342	< 2e-16 ***
風速	5.375e-03	8.920e-05	60.254	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 1800567 on 10885 degrees of freedom
Residual deviance: 1312614 on 10877 degrees of freedom
AIC: 1382279 Number of Fisher Scoring iterations: 5

図表１３　分析１の推定結果表

変数部分を追加・変更することにより、適合度の高いモデルを探索する。

分析１



②年ダミーを追加して分析

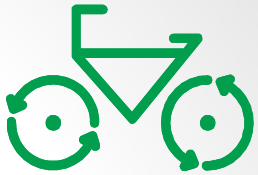
glm(formula = 利用者数 ~ 年ダミー + 季節 + 祝日ダミー + 出勤日ダミー + 天気 + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance: 1217956　AIC: 1287623

①よりもResidual devianceの値が下がっており、②の方が①よりもよいモデルであることが分かる。

この後も②と同様に分析を行う。

分析１



③月ダミー(11個)を追加して分析

glm(formula = 利用者数 ~ 季節 +月ダミー(11個) + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance: 1245939　AIC: 1315624

④曜日ダミー(６個)を追加して分析

glm(formula = 利用者数 ~ 季節 +曜日ダミー(6個) + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

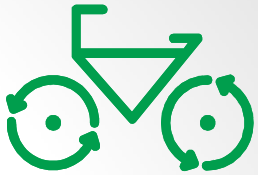
→Residual deviance: 1311743　AIC: 1381418

⑤1時間ダミー(２３個)を追加して分析

glm(formula = 利用者数 ~ 季節 +1時間ダミー(23個) + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance:　490245　AIC: 559956

分析１



⑥季節→季節ダミー(３個)に変更して分析

glm(formula = 利用者数 ~ 季節ダミー(3個) + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance: 1285343　AIC: 1355012

⑦天気→天気ダミー(３個)に変更して分析

glm(formula = 利用者数 ~ 季節 + 祝日ダミー + 出勤日ダミー +天気ダミー(3個) + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance: 1307020　AIC: 1376689

⑧作成した変数をまとめて追加して分析

glm(formula = 利用者数 ~ 年ダミー + 季節ダミー(3個) + 月ダミー(11個) + 曜日ダミー(6個)　+ 時間帯ダミー(23個) + 祝日ダミー + 出勤日ダミー +天気ダミー(3個) + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

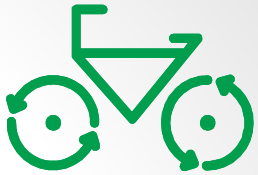
→Residual deviance:　344264　AIC: 414011

分析１　結果

	変数の追加・変更	Residual deviance	AIC
①	なし	1312614	1382279
②	年ダミー	1217956	1287623
③	月ダミー（１１個）	1245939	1315624
④	曜日ダミー（６個）	1311743	1381418
⑤	1時間ダミー（２３個）	490245	559956
⑥	季節ダミー（３個）	1285343	1355012
⑦	天気ダミー（３個）	1307020	1376689
⑧	①～⑦のダミー全て	344264	414011

図表１４　分析１結果一覧

分析１から分かること



- 1時間ダミー(２３個)を追加すると、他の変数追加時と比べてResidual devianceが大きく下がった。

  →時間帯が持つ影響力が大きいと考えられる。

- 曜日ダミーや季節ダミーを追加したとき、他の変数追加時と比べるとResidual devianceの減少幅が小さい。

  →影響力が小さいor変数の作成方法に改善点がある可能性。

分析２：季節・風速に着目した分析

変数の追加・変更２

１．季節Ⅱダミーの追加

ワシントンD.C.の気温に注目し、暑さのピークが6-8月、寒さのピークが12-2月であることに注目

→以上を踏まえて季節Ⅱダミーを作成

3-5月、6-8月、9-11月の３種類のダミー変数

２．風速１５mダミーの追加

気象庁のHPを確認したところ、15m以上の風を「強風」と認定している。

→１５mが一つの基準になるのではないか。

分析２

⑨　⑥から季節→季節Ⅱダミー(３個)に変更して分析

glm(formula = 利用者数 ~ 季節Ⅱダミー(3個) + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance: 1297002　AIC: 1366671

※⑥季節→季節ダミー(３個)に変更して分析

→Residual deviance: 1285343　AIC: 1355012

季節ダミーの方が季節Ⅱダミーよりも当てはまりが良いという結果になった。

分析２

⑩　①から風速→風速１５ｍダミーに変更して分析

glm(formula = 利用者数 ~ 季節ダミー(3個) + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速15mダミー, family = "poisson", data = train)

→ Residual deviance: 1313883　AIC: 1383548

※①元々設定されている変数のみ使用

→Residual deviance: 1312614 AIC: 1382279

当てはまりが悪くなったため、今度は風速１０ｍダミーを作成し再分析を行った。(気象庁のHPでは、１０ｍの風から表現が作られていることを参照)

glm(formula = 利用者数 ~ 季節ダミー(3個) + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速10mダミー, family = "poisson", data = train)

→Residual deviance: 1310557　AIC: 1380222

①よりモデルの当てはまりが改善した。

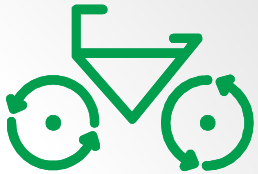
分析２　結果

	変数の追加・変更	Residual deviance	AIC
⑨	⑥から季節→季節Ⅱダミーの変更	1297002	1366671
⑩	①から風速→風速15mダミーへの変更	1313883	1383543
⑩-2	①から風速→風速10mダミーへの変更	1310557	1380222

図表１５　分析２結果一覧

※ ⑥季節ダミー追加時　Residual deviance：1285343　AIC:1355012
→分析⑨は分析⑥よりもモデルは悪化
①変数追加なし　Residual deviance：1312614　AIC:1382279
よって、分析⑩-2は分析①よりモデルの適合度が上昇

分析２から分かること



- ワシントンD.C.の気温の推移に対応した季節Ⅱダミーよりも季節ダミーの方が当てはまりがいいということが分かった。

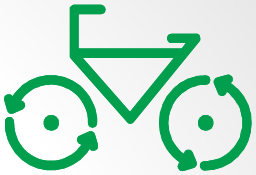
　　→分布図で確認すると冬の利用者数は夏よりも少ない。気温よりも別の要素の方が利用者数への影響力が高い可能性。

- 風速に関するダミーを作成したところ、１０ｍで区分するとモデルがわずかに改善した。

　　→ただし１５ｍの方でもResidual devianceは大きく変化はしていないため、影響力は小さい？

分析３：時間帯の区分・体感気温に着目した分析

変数の追加３



- ⑪　３時間ダミーの追加

  →1日を数時間単位で区分分けし、ダミー化

  1時間ダミーとの比較をすることが目的。

  →⑧の分析と比較

  0時から順に3時間ごとにダミー変数にする

- ⑫体感気温31.06度ダミー

  →単純集計のグラフより、体感気温が31.06度の時に利用者が急増している。よって影響力があると考えダミー変数を設定

分析３

⑪　①に３時間ダミー(7個)を追加して分析

glm(formula = 利用者数 ~ 年ダミー + 季節ダミー(3個) + 月ダミー(11個) + 曜日ダミー(6個) + 3時間ダミー(7個) + 祝日ダミー + 出勤日ダミー +天気ダミー(3個) + 気温 + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance: 658826 AIC: 728503

⑫　⑧から体感気温→体感気温31.06度ダミーに変更して分析

glm(formula = 利用者数 ~ 年ダミー + 季節ダミー(3個) + 月ダミー(11個) + 曜日ダミー(6個) + 1時間ダミー(23個) + 祝日ダミー + 出勤日ダミー +天気ダミー(3個) + 気温 + 体感気温31.06度ダミー + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance: 343700 AIC: 413447

※⑧作成した変数をまとめて追加して分析

→Residual deviance: 344264 AIC: 414011

→⑫においてモデルの当てはまりは改善した。

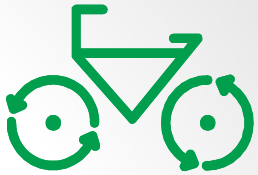
分析３　結果

	変数の追加・変更	Residual deviance	AIC
⑪	①に３時間ダミー（７個）の追加	658826	728503
⑫	⑧から体感気温→体感気温31.06度ダミーへの変更	343700	413447

図表１６　分析３結果一覧

※⑧（①～⑦で作成したダミー追加）Residual deviance:344264　AIC:414011
→分析⑫は分析⑧よりもモデルの適合度が上がった。

分析３からわかること



⑫体感気温→体感気温31.06度ダミーに変更して分析

→Residual deviance: 343700 AIC: 413447

となり、モデルはわずかに改善した。

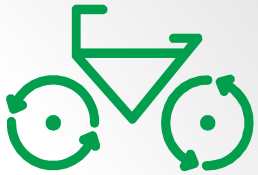
しかし作成した他の変数を含めずにKaggle側が元々設定した変数のみ使用した式で「体感気温→体感気温31.06度ダミーに変更」をしてみるとモデルは悪化。

※glm(formula = 利用者数 ~ 季節 + 祝日ダミー + 出勤日ダミー +天気 + 気温 + 体感気温 + 湿度 + 風速, family = “poisson”, data = train)において体感気温→体感気温31.06度ダミーに変更

→変数間の交互作用を分析に取り入れるべきだと考える

分析４：回帰診断の結果に着目した分析

分析４ 回帰診断



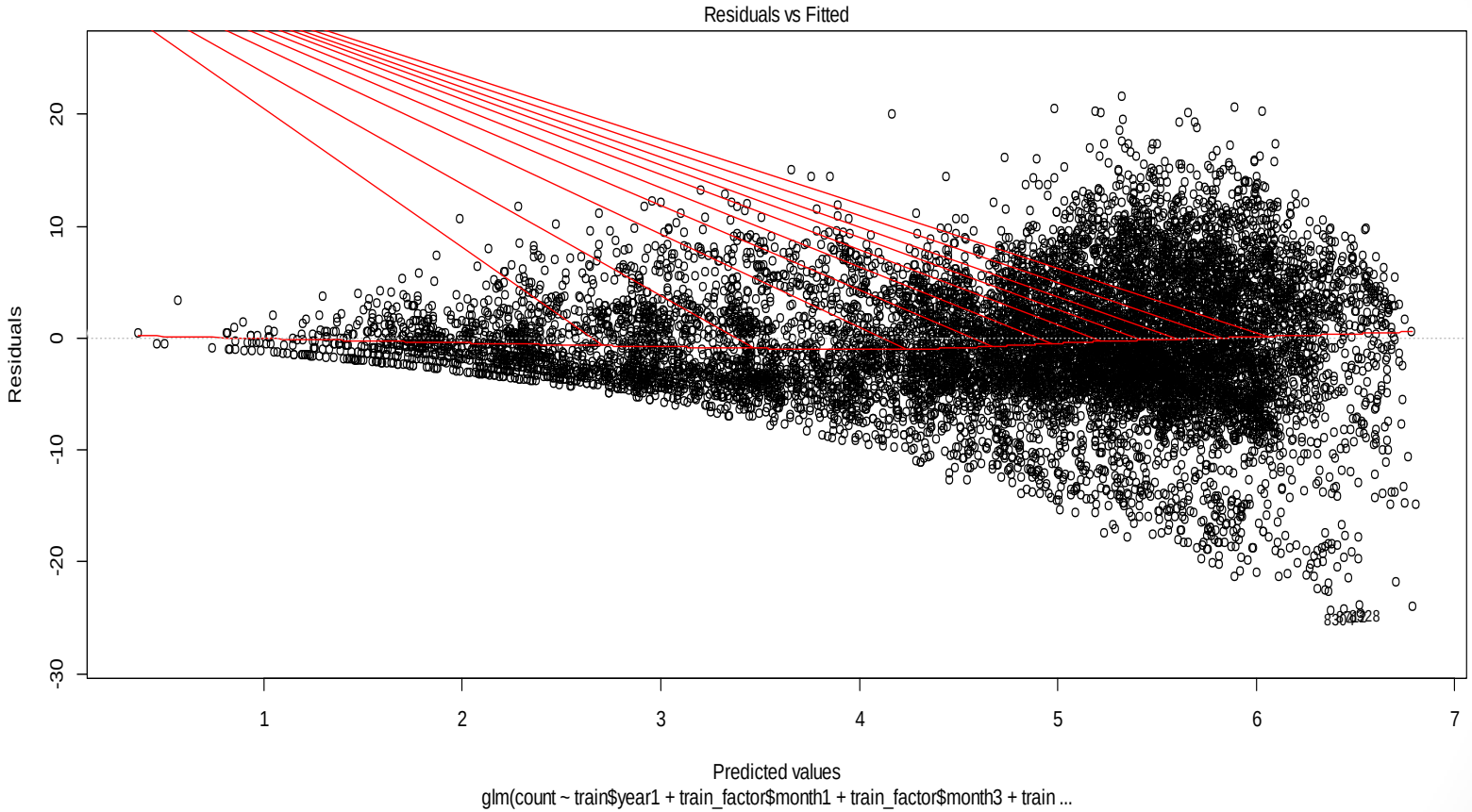
⑫　⑧から体感気温→体感気温31.06度ダミーに変更して分析

→Residual deviance: 343700 AIC: 413447

が今のところ一番良いモデルであるため、このモデルに対し回帰診断を行う。

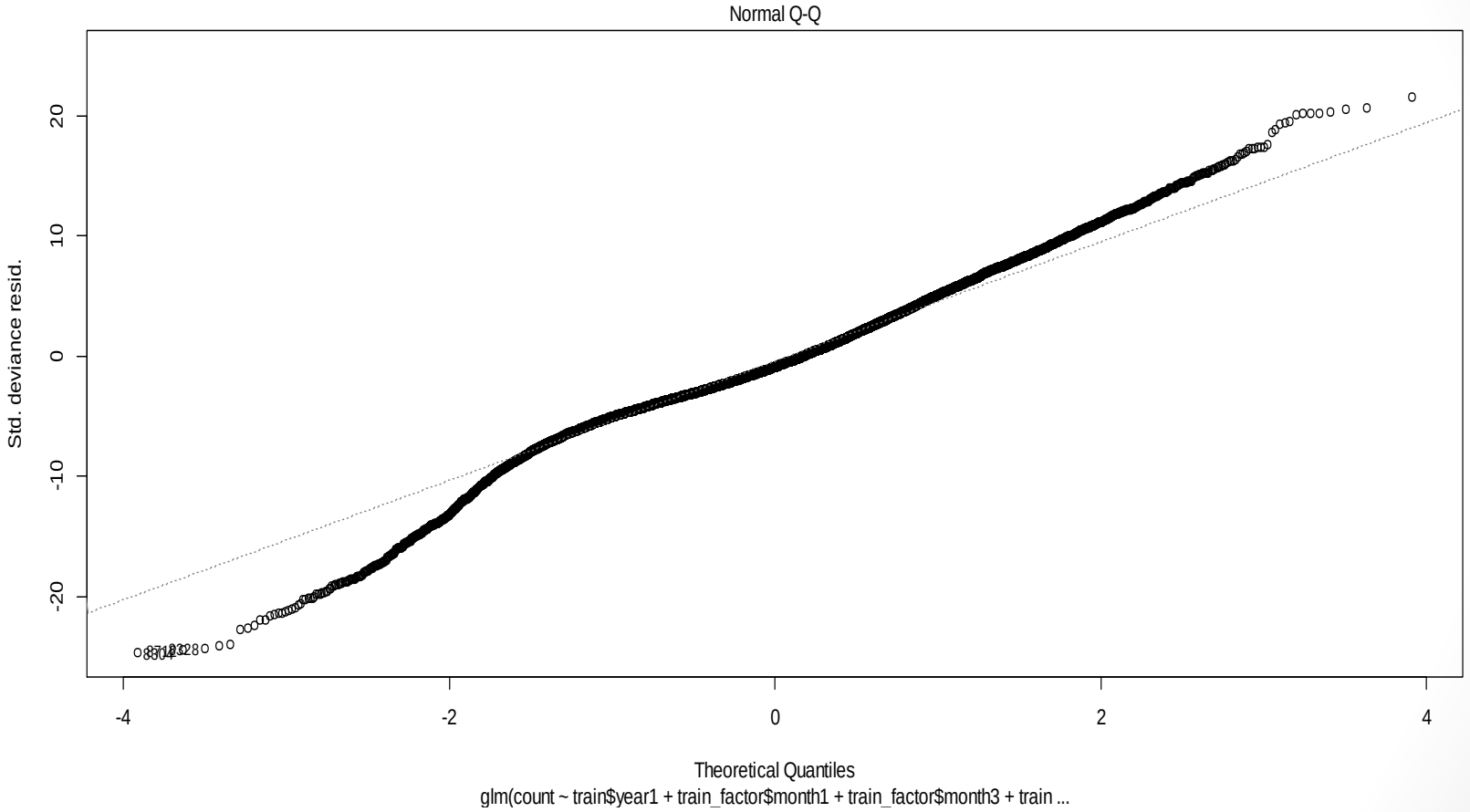
目的：残差が大きいサンプルを見つけることにより、現状のモデルが対応できていない部分を明らかにし、モデルの改善を目指す

回帰診断



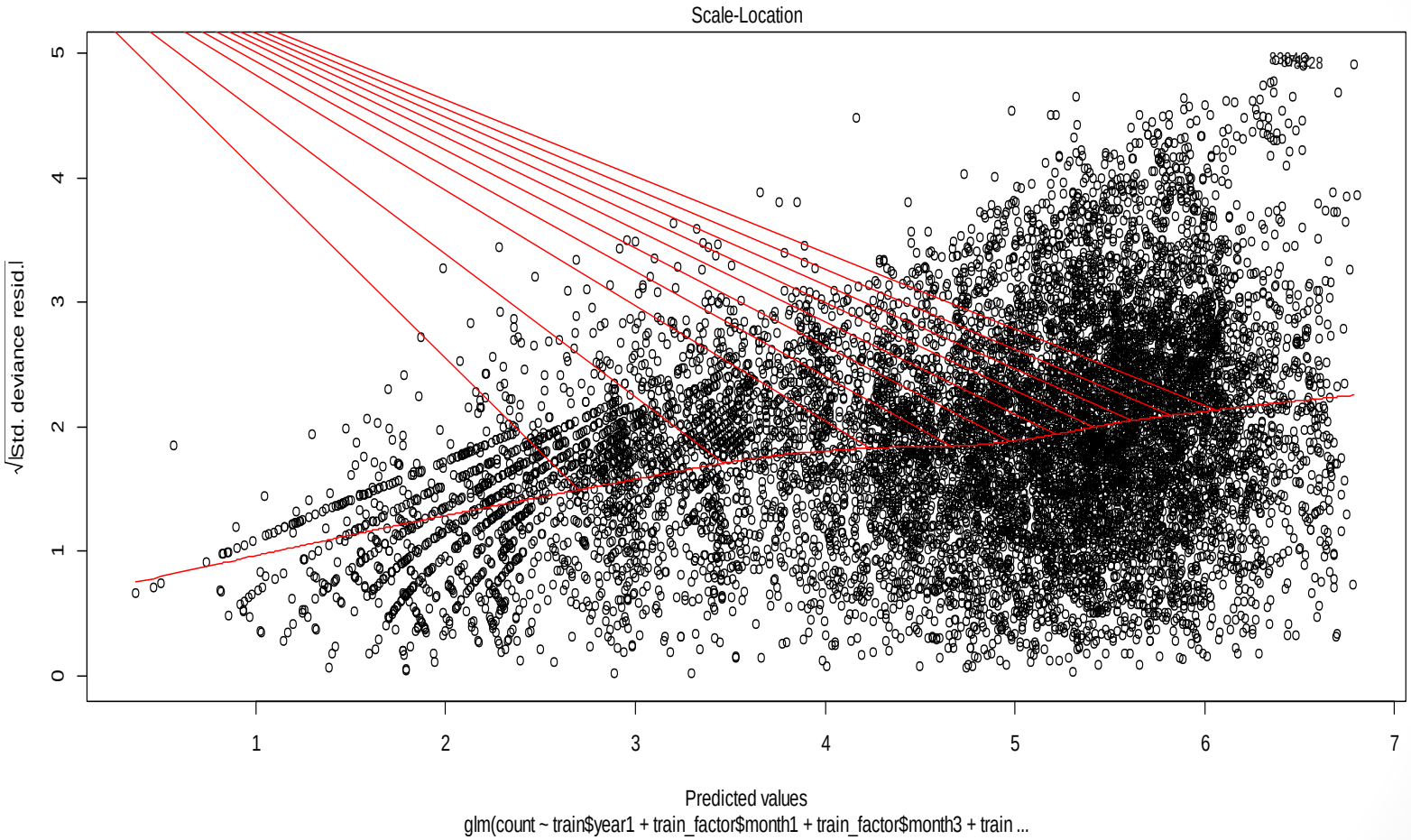
図表１７　残差とフィット値のプロット
横軸が予測値、縦軸が残差
残差の全体像を概観するために使用

回帰診断



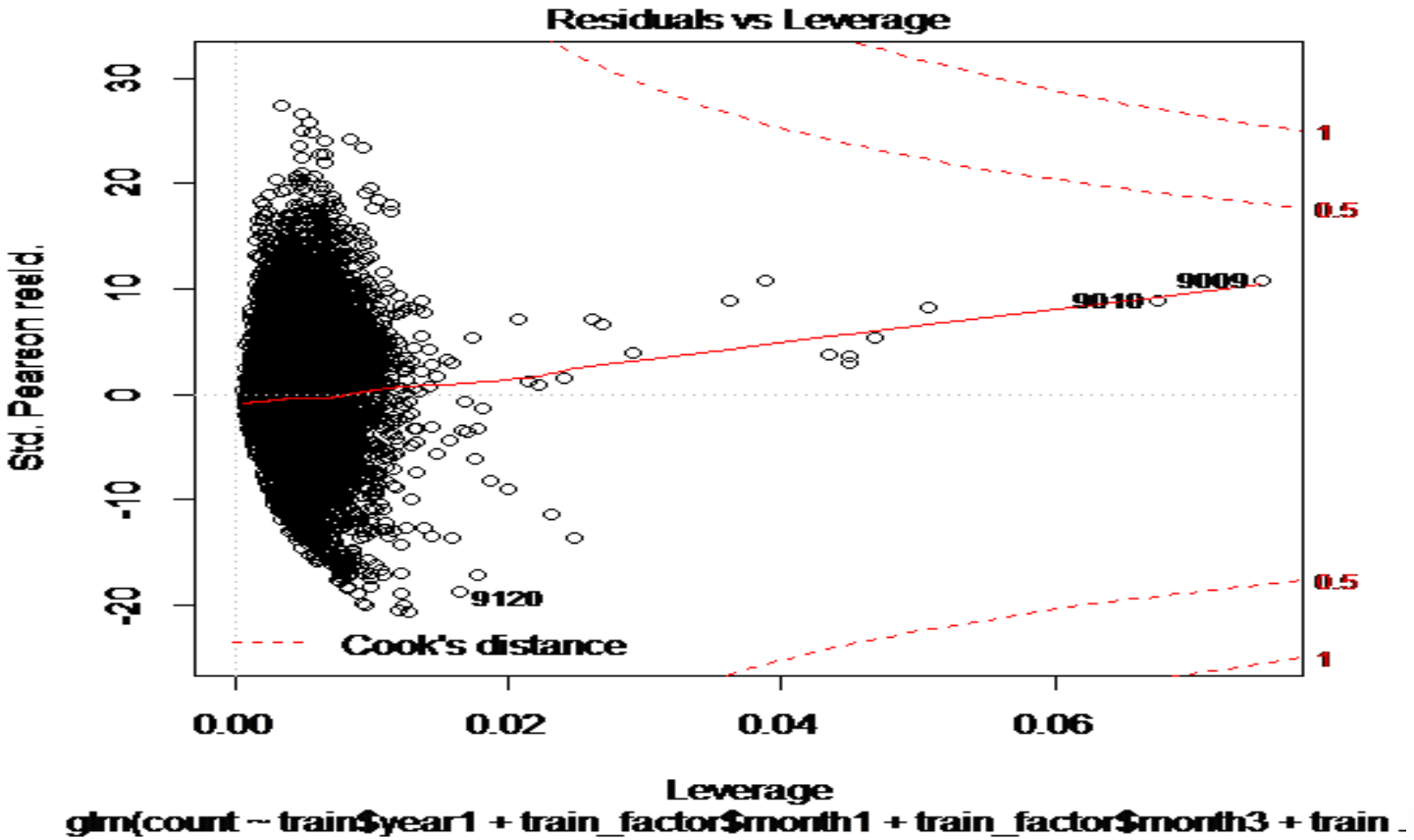
図表１８　残差の正規Q-Qプロット
データの正規性を考察するために使用
データが正規分布に従っている場合は、直線上に並ぶ。

回帰診断



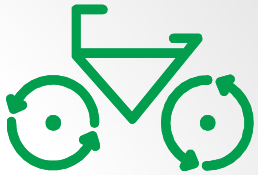
図表１９　残差の平方根プロット
残差の変動状況を考察するために使用
標準化した残差の絶対値の平方根を縦軸にし、予測値を横軸にした散布図。

回帰診断



図表２０　残差と影響力プロット（梃子値とクック距離）
１つのデータがモデルの当てはまりへの影響力を測るために使用する。
クックの距離が0.5を超えると影響力あり、1を超えると特異に大きい。
横軸は梃子値で、縦軸は標準化した残差。点線でクックの距離0.5を示している。

回帰診断より



- 1番目の図より

  8308番＝2012/7/7　12:00:00 の残差が大きい

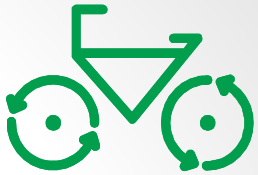
- 4番目の図より

  9009番・9010番＝2012/8/17　16:00:00・17:00:00の影響が大きい

- 全体より

  残差の影響が残っている(誤差がある)ことが視覚的にわかる。モデル改良の余地がある

変数の追加・変更４



- 回帰診断の結果より、７・８月に関する分析に改良の余地があるのではないかと考えた。

　　→気温や体感気温が高い月であるため、気温や体感気温と交互作用がありそうな変数を結びつけて分析をしてみる

⑬気温と7・8・9月ダミー(季節ダミーの内の1つ)を結びつけて一つの変数を作成して分析

⑭体感気温と7・8・9月ダミー(季節ダミーの内の1つ)を結びつけて一つの変数を作成して分析

分析４

- ⑬　⑫に気温と7・8・9月ダミー(季節ダミーの内の1つ)を結びつけた一つの変数を追加して分析

glm(formula = 利用者数 ~ 年ダミー + 季節ダミー(2個) + 月ダミー(11個) + 曜日ダミー(6個) + 1時間ダミー(23個) + 祝日ダミー + 出勤日ダミー +天気ダミー(3個) + 気温×7・8・9月ダミー + 体感気温 + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance:　340125　AIC: 409874

- ⑭　⑫に体感気温と7・8・9月ダミー(季節ダミーの内の1つ)を結びつけて一つの変数を追加して分析

glm(formula = 利用者数 ~ 年ダミー + 季節ダミー(2個) + 月ダミー(11個) + 曜日ダミー(6個) + 1時間ダミー(23個) + 祝日ダミー + 出勤日ダミー +天気ダミー(3個) + 気温 + 体感気温×7・8・9月ダミー + 湿度 + 風速, family = "poisson", data = train)

→Residual deviance:　338990　AIC: 409632

※⑫体感気温→体感気温31.06度ダミーに変更して分析

→Residual deviance:　343700 AIC: 413447

→今までで一番良いモデルよりも、残差は小さくなる結果となった。

分析４　結果

	変数の追加・変更	Residual deviance	AIC
⑬	⑫に気温＋7・8・9月ダミーの追加	340125	409874
⑭	⑫に体感気温＋7・8・9月ダミーの追加	338990	408739

図表２１　分析④結果一覧

※ 分析⑫（⑧から体感気温→体感気温31.06度ダミーに変更して分析）
Residual deviance：344264　AIC:414011
→分析⑭は分析⑫よりもモデルの適合度が改善した。

分析４

図表２２　分析⑭の推定結果表

	係数	標準誤差	z 値	P値
2011年	-4.818e-01	1.453e-03	-331.649	< 2e-16 ***
1－3月	-4.518e-01	4.123e-03	-109.571	< 2e-16 ***
4－6月	-6.902e-02	4.515e-03	-15.288	< 2e-16 ***
7－9月	8.616e-01	1.216e-02	70.870	< 2e-16 ***
1月	-1.436e-01	4.811e-03	-29.859	< 2e-16 ***
3月	1.537e-01	4.341e-03	35.409	< 2e-16 ***
4月	-8.518e-02	3.735e-03	-22.805	< 2e-16 ***
5月	3.251e-02	3.241e-03	10.032	< 2e-16 ***
6月	NA	NA	NA	NA
7月	-2.841e-02	3.805e-03	-7.465	8.36e-14 ***
8月	-1.094e-02	3.497e-03	-3.128	0.001762 **
9月	NA	NA	NA	NA
10月	7.941e-02	3.737e-03	21.249	< 2e-16 ***
11月	1.723e-01	5.962e-03	28.898	< 2e-16 ***
12月	NA	NA	NA	NA
日曜日	-9.242e-02	2.603e-03	-35.500	< 2e-16 ***
月曜日	-6.064e-02	2.664e-03	-22.762	< 2e-16 ***
火曜日	-4.123e-02	2.591e-03	-15.911	< 2e-16 ***
水曜日	-3.965e-02	2.587e-03	-15.324	< 2e-16 ***
木曜日	-2.090e-02	2.550e-03	-8.194	2.54e-16 ***
金曜日	-4.765e-03	2.571e-03	-1.854	0.063776 .
0時	-4.864e-01	8.024e-03	-60.621	< 2e-16 ***
1時	-9.552e-01	9.466e-03	-100.903	< 2e-16 ***
2時	-1.351e+00	1.105e-02	-122.279	< 2e-16 ***
3時	-2.022e+00	1.487e-02	-135.982	< 2e-16 ***
4時	-2.608e+00	1.944e-02	-134.172	< 2e-16 ***
5時	-1.470e+00	1.169e-02	-125.753	< 2e-16 ***
6時	-1.003e-01	7.325e-03	-13.689	< 2e-16 ***
7時	9.236e-01	5.918e-03	156.084	< 2e-16 ***
8時	1.428e+00	5.536e-03	257.996	< 2e-16 ***
9時	9.096e-01	5.875e-03	154.827	< 2e-16 ***
10時	6.449e-01	6.111e-03	105.530	< 2e-16 ***

分析4

12時	9.716e-01	5.843e-03	166.285	< 2e-16 ***
13時	9.535e-01	5.874e-03	162.311	< 2e-16 ***
14時	8.872e-01	5.947e－03	149.184	< 2e-16 ***
15時	9.303e-01	5.920e-03	157.148	< 2e-16 ***
16時	1.149e+00	5.763e-03	199.432	< 2e-16 ***
17時	1.566e+00	5.536e-03	282.801	< 2e-16 ***
18時	1.494e+00	5.538e-03	269.817	< 2e-16***
19時	1.190e+00	5.658e-03	210.290	< 2e-16 ***
20時	8.859e-01	5.863e-03	151.104	< 2e-16 ***
21時	6.235e-01	6.103e-03	102.171	< 2e-16 ***
22時	3.792e-01	6.399e-03	59.266	< 2e-16 ***
天気 スコア1	2.951e-01	7.826e-02	3.771	0.000163 ***
天気 スコア2	2.277e-01	7.826e-02	2.910	0.003620 **
天気 スコア3	-1.794e-01	7.830e-02	-2.292	0.021929 *
祝日	-3.730e-03	4.524e-03	-0.825	0.409574
出勤日	NA	NA	NA	NA
気温	-1.079e-02	5.678e-04	-19.011	< 2e-16 ***
体感気温	3.481e-02	5.135e-04	67.796	< 2e-16 ***
湿度	-2.647e-03	5.318e-05	-49.767	< 2e-16 ***
風速	-1.203e-03	9.189e-05	-13.091	< 2e-16 ***
体感気温×7－9月	-2.876e-02	3.953e-04	-72.759	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 1800567 on 10885 degrees of freedom
Residual deviance: 338990 on 10835 degrees of freedom
AIC: 408739
Number of Fisher Scoring iterations: 5

分析４

- 分析⑭において、適合度は高いが一部の変数(季節ダミー)がNAと表示。

→他の変数との関係に原因があると考えられるため、今回は分析⑭からNAと表示された変数4つ(6月・9月・12月・出勤日)を取り除いたモデル分析⑭-2を扱うこととする。

分析４

図表２３　分析⑭－２の推定結果表

	係数	標準誤差	z 値	P値
2011年	-4.818e-01	1.453e-03	-331.649	< 2e-16 ***
1－3月	-4.518e-01	4.123e-03	-109.571	< 2e-16 ***
4－6月	-6.902e-02	4.515e-03	-15.288	< 2e-16 ***
7－9月	8.616e-01	1.216e-02	70.870	< 2e-16 ***
1月	-1.436e-01	4.811e-03	-29.859	< 2e-16 ***
3月	1.537e-01	4.341e-03	35.409	< 2e-16 ***
4月	-8.518e-02	3.735e-03	-22.805	< 2e-16 ***
5月	3.251e-02	3.241e-03	10.032	< 2e-16 ***
6月	-2.841e-02	3.805e-03	-7.465	8.36e-14 ***
7月	-1.094e-02	3.497e-03	-3.128	0.001762 **
8月	7.941e-02	3.737e-03	21.249	< 2e-16 ***
11月	1.723e-01	5.962e-03	28.898	< 2e-16 ***
日曜日	-9.242e-02	2.603e-03	-35.500	< 2e-16 ***
月曜日	-6.064e-02	2.664e-03	-22.762	< 2e-16 ***
火曜日	-4.123e-02	2.591e-03	-15.911	< 2e-16 ***
水曜日	-3.965e-02	2.587e-03	-15.324	< 2e-16 ***
木曜日	-2.090e-02	2.550e-03	-8.194	2.54e-16 ***
金曜日	-4.765e-03	2.571e-03	-1.854	0.063776 .
0時	-4.864e-01	8.024e-03	-60.621	< 2e-16 ***
1時	-9.552e-01	9.466e-03	-100.903	< 2e-16 ***
2時	-1.351e+00	1.105e-02	-122.279	< 2e-16 ***
3時	-2.022e+00	1.487e-02	-135.982	< 2e-16 ***
4時	-2.608e+00	1.944e-02 -	134.172	< 2e-16 ***
5時	-1.470e+00	1.169e-02	-125.753	< 2e-16 ***
6時	-1.003e-01	7.325e-03	-13.689	< 2e-16 ***
7時	9.236e-01	5.918e-03	156.084	< 2e-16 ***
8時	1.428e+00	5.536e-03	257.996	< 2e-16 ***
9時	9.096e-01	5.875e-03	154.827	< 2e-16 ***
10時	6.449e-01	6.111e-03	105.530	< 2e-16 ***

分析4

11時	7.953e-01	5.965e-03	133.327	< 2e-16 ***
12時	9.716e-01	5.843e-03	166.285	< 2e-16 ***
13時	9.535e-01	5.874e-03	162.311	< 2e-16 ***
14時	8.872e-01	5.947e-03	149.184	< 2e-16 ***
15時	9.303e-01	5.920e-03	157.148	< 2e-16 ***
16時	1.149e+00	5.763e-03	199.432	< 2e-16 ***
17時	1.566e+00	5.536e-03	282.801	< 2e-16 ***
18時	1.494e+00	5.538e-03	269.817	< 2e-16 ***
19時	1.190e+00	5.658e-03	210.290	< 2e-16 ***
20時	8.859e-01	5.863e-03	151.104	< 2e-16 ***
21時	6.235e-01	6.103e-03	102.171	< 2e-16 ***
22時	3.792e-01	6.399e-03	59.266	< 2e-16 ***
天気 スコア1	2.951e-01	7.826e-02	3.771	0.000163 ***
天気 スコア2	2.277e-01	7.826e-02	2.910	0.003620 **
天気 スコア3	-1.794e-01	7.830e-02	-2.292	0.021929 *
祝日	-3.730e-03	4.524e-03	-0.825	0.409574
気温	-1.079e-02	5.678e-04	-19.011	< 2e-16 ***
体感気温	3.481e-02	5.135e-04	67.796	< 2e-16 ***
湿度	-2.647e-03	5.318e-05	-49.767	< 2e-16 ***
風速	-1.203e-03	9.189e-05	-13.091	< 2e-16 ***
体感気温×7ー9月	-2.876e-02	3.953e-04	-72.759	< 2e-16 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for poisson family taken to be 1)
Null deviance: 1800567 on 10885 degrees of freedom
Residual deviance: 338990 on 10835 degrees of freedom
AIC: 408739
Number of Fisher Scoring iterations: 5

考察・まとめ

考察

- 利用者数への影響度が高い変数として、年・月・時や気温・体感気温が挙げられる。

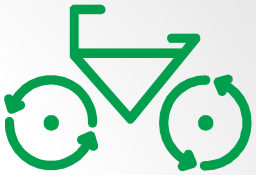
  その中でも「hour」の持つ影響力が大きい。

  →上記の変数を組み入れると、モデルの残差の変化が大きいため。交互作用にしてみるとさらにモデルの当てはまりが上昇した。

- 逆にそのほかの変数の影響力は小さい。(モデルと適合度にもあまり変化をもたらさない)

- 気温と体感気温という２つの変数は性質が異なっており、別々に考える必要性。

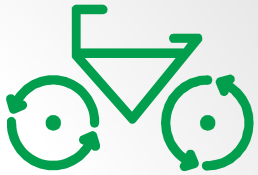
まとめ



分析①と、今回の分析において適合度が高い上位３つのモデルを取り上げ、それぞれどのような変数を用いているかを以下に示す。

変数名	分析①	分析⑧	分析⑬	分析⑭
季節	1.534e-01			
祝日	-3.536e-02	-2.691e-02	-1.283e-02	-3.730e-03
出勤日	1.869e-03	NA	NA	NA
天気	1.266e-02			
気温	-2.335e-03	4.970e-03	1.195e-02	-1.079e-02
体感気温	3.949e-02	1.544e-02	1.383e-02	3.481e-02
湿度	-1.559e-02	-2.350e-03	-2.630e-03	-2.647e-03
風速	5.375e-03	-1.554e-03	-1.650e-03	-1.203e-03
2011年		-4.767e-01	-4.816e-01	-4.818e-01
1－3月		-4.607e-01	-4.567e-01	-4.518e-01
4－6月		-5.474e-03	-8.115e-02	-6.902e-02
7－9月		2.935e-02	7.621e-01	8.616e-01
1月		-1.612e-01	-1.514e-01	-1.436e-01
3月		1.816e-01	1.563e-01	1.537e-01
4月		-1.104e-01	-7.129e-02	-8.518e-02
5月		1.987e-02	4.395e-02	3.251e-02
6月		NA	NA	NA
7月		-1.612e-01	-3.265e-02	-2.841e-02
8月		-1.034e-01 *	-8.780e-03	-1.094e-02
9月		NA	NA	NA
10月		1.192e-01	7.944e-02	7.941e-02
11月		1.974e-01	1.771e-01	1.723e-01
12月		NA	NA	NA
日曜日		-9.152e-02	-9.283e-02	-9.242e-02
月曜日		-5.243e-02	-5.243e-02	-6.064e-02
火曜日		-4.334e-02	-4.334e-02	-4.123e-02
水曜日		-3.747e-02	-3.747e-02	-3.965e-02
木曜日		-1.903e-02	-1.903e-02	-2.090e-02
金曜日		2.147e-03	2.147e-03	-4.765e-03

まとめ

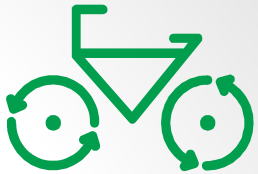


0時		-4.865e-01	-4.865e-01	-4.864e-01
1時		-9.552e-01	-9.552e-01	-9.552e-01
2時		-1.350e+00	-1.350e+00	-1.351e+00
3時		-2.020e+00	-2.020e+00	-2.022e+00
4時		-2.605e+00	-2.605e+00	-2.608e+00
5時		-1.467e+00	-1.467e+00	-1.470e+00
6時		-9.843e-02	-9.843e-02	-1.003e-01
7時		9.200e-01	9.200e-01	9.236e-01
8時		1.420e+00	1.420e+00	1.428e+00
9時		8.977e-01	8.977e-01	9.096e-01
10時		6.285e-01	6.285e-01	6.449e-01
11時		7.781e-01	7.781e-01	7.953e-01
12時		9.548e-01	9.548e-01	9.716e-01
13時		9.362e-01	9.362e-01	9.535e-01
14時		8.703e-01	8.703e-01	8.872e-01
15時		9.162e-01	9.162e-01	9.303e-01
16時		1.137e+00	1.137e+00	1.149e+00
17時		1.552e+00	1.552e+00	1.566e+00
18時		1.484e+00	1.484e+00	1.494e+00
19時		1.182e+00	1.182e+00	1.190e+00
20時		8.812e-01	8.812e-01	8.859e-01
21時		6.213e-01	6.213e-01	6.235e-01
22時		3.771e-01	3.771e-01	3.792e-01
天気 スコア1		3.069e-01	3.069e-01	2.951e-01
天気 スコア2		2.420e-01	2.420e-01	2.277e-01
天気 スコア3		-1.646e-01	-1.646e-01	-1.794e-01
気温×7－9月			-2.962e-02	
体感気温×7－9月				-2.876e-02
Residual Deviance	1313614	344264	340125	338990
AIC	1382279	414011	409874	408739

図表２４　原型モデル、適合度上位３つのモデルが用いた変数

モデルの提出

Kaggle.comへの提出

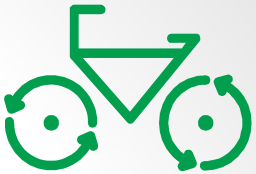


- 今回、推定用データ(1-20日まで)において作成したモデルを予測用データ(21日-月末)に当てはめ、利用者数を予測し、Kaggle.comへの提出を行った。

- Kaggle.comに提出することでモデルの適合度がスコアという形で示され、他者との比較を行うことができる。スコアが0に近くなるほど良いモデルであるとされ、順位も高くなる。

2015年1月17日現在、1960チームがBike Sharingのコンテストに参加しておりその中で1600チームのスコアが1を下回っている。

Kaggle.comへの提出



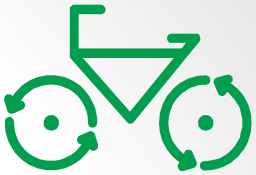
Kaggleから与えられた変数のみのモデル(分析①)よりも、今回作成したモデル(分析⑮)の両方を提出し、スコアの変化を確認した。

→分析①のスコア：3.06462
　分析⑮のスコア：3.06786

Kaggleから与えられた変数のみのモデル(分析①)よりも、今回作成したモデル(分析⑮)の方がスコアが低いという結果に。

→残差などを考えるとモデルは改善しているため、スコアが悪化するとは考えにくい。他の面から問題点を探す必要性がある。

モデルの提出における考察



推定用データにおけるモデルの適合度が上がっているにもかかわらず、予測用データにおけるモデルの適合度は下がってしまった。

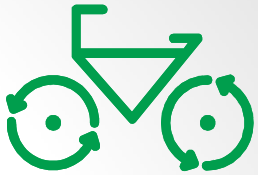
→過学習の可能性

サンプル数に比べてサンプルの説明変数の個数が多い場合や，予測モデルに複雑過ぎる関数を想定した場合，サンプルに対してはよくあてはまるモデルは構築されるが，未知のサンプルに対する予測の精度が極めて悪化することがある。今回の分析において、分析①の説明変数が7個であることに対し、

分析⑭では50個ほどの説明変数を使用しているため、予測の精度が悪化した可能性がある。

課題

課題



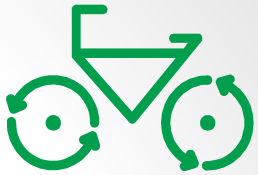
- 時間のさらなる区分分けができる可能性。
  →３時間ダミーについても、とりあえず0時から順に3時間ごとに区切ったが、より適切な分け方があるのではないか。

- datetimeという変数が使用不能
  →年・月等を抜き出すと使うことはできるが、そのまま被説明変数に組み込むとRがフリーズしてしまう。(関連するダミー変数の影響度から、datetimeもモデルへの影響が大きいと考えられるため、使用する方法を考える必要がある)

- 気温・体感気温・湿度・風速といったデータの扱い方に留意点がある。それぞれ「特定の値」でしか記録されていないため、結果に誤差を生じさせる原因になり得る。
  →「特定の値」が設定されていて、実際の値から一番近い「特定の値」を記録している？

- 過学習の可能性があるため、説明変数の数の調整や変数の作り方を見直すことで快勝することができるのではないか。

終わりに



- 本研究において、データの提供もとである
  Kaggle.comならびに
  Capital Bike Shareに感謝を申し上げます。

参考文献

- Capital Bike Share(https://www.capitalbikeshare.com/)
- Kaggle.com(http://www.kaggle.com/)
- 気象庁 知識・解説 天気予報で用いる用語「風」
  (http://www.jma.go.jp/jma/kishou/know/yougo_hp/kaze.html)
- 港区自転車シェアリング(http://docomo-cycle.jp/minato/)

- OR事典Wiki 過学習
  (http://www.orsj.or.jp/~wiki/wiki/index.php/%E9%81%8E%E5%AD%A6%E7%BF%92)
- コトバンク-自転車シェアリング
- (https://kotobank.jp/word/自転車シェアリング-190957)
- 川瀬あゆ美,宮川卓也,渡邉由比香「Titanicの乗客データを用いた分析」(慶應義塾大学商学部濱岡豊研究会2013年度個人/グループ研究)
  (http://news.fbc.keio.ac.jp/~hamaoka/GRAD_12/3_Titanic.pdf)
- トム・バンダービルト『公共交通の切り札は自転車』ニューズウィーク日本版,2013年3月1日(http://www.newsweekjapan.jp/stories/world/2013/03/post-2860_1.php)
- climatemps.com(http://www.washington-dc.climatemps.com/)
- Brandon Harris『A simple model for Kaggle Bike Sharing.』
  (http://brandonharris.io/kaggle-bike-sharing/)