STP 530
Midterm Review

Yi Zheng, Ph.D.
yi.isabel.zheng@asu.edu

ARIZONA STATE UNIVERSITY
SCHOOL OF MATHEMATICAL AND STATISTICAL SCIENCES

October 5, 2023

Example: Analyzing the bike sharing dataset

Data source: The UC Irvine Machine Learning Repository

`https://archive.ics.uci.edu/dataset/275/bike+sharing+dataset`

Pictures above: Bike sharing in New York City

Bike Sharing Dataset

Donated on 12/19/2013

This dataset contains the hourly and daily count of rental bikes between years 2011 and 2012 in Capital bikeshare system with the corresponding weather and seasonal information.

Dataset Characteristics	Subject Area	Associated Tasks
Univariate	Social	Regression
Attribute Type	**# Instances**	**# Attributes**
Integer, Real	17389	16

Information

Additional Information

Bike sharing systems are new generation of traditional bike rentals where whole process from membership, rental and return back has become automatic. Through these systems, user is able to easily rent a bike from a particular position and return back at another position. Currently, there are about over 500 bike-sharing programs around the world which is composed ...

SHOW MORE

Has Missing Values?

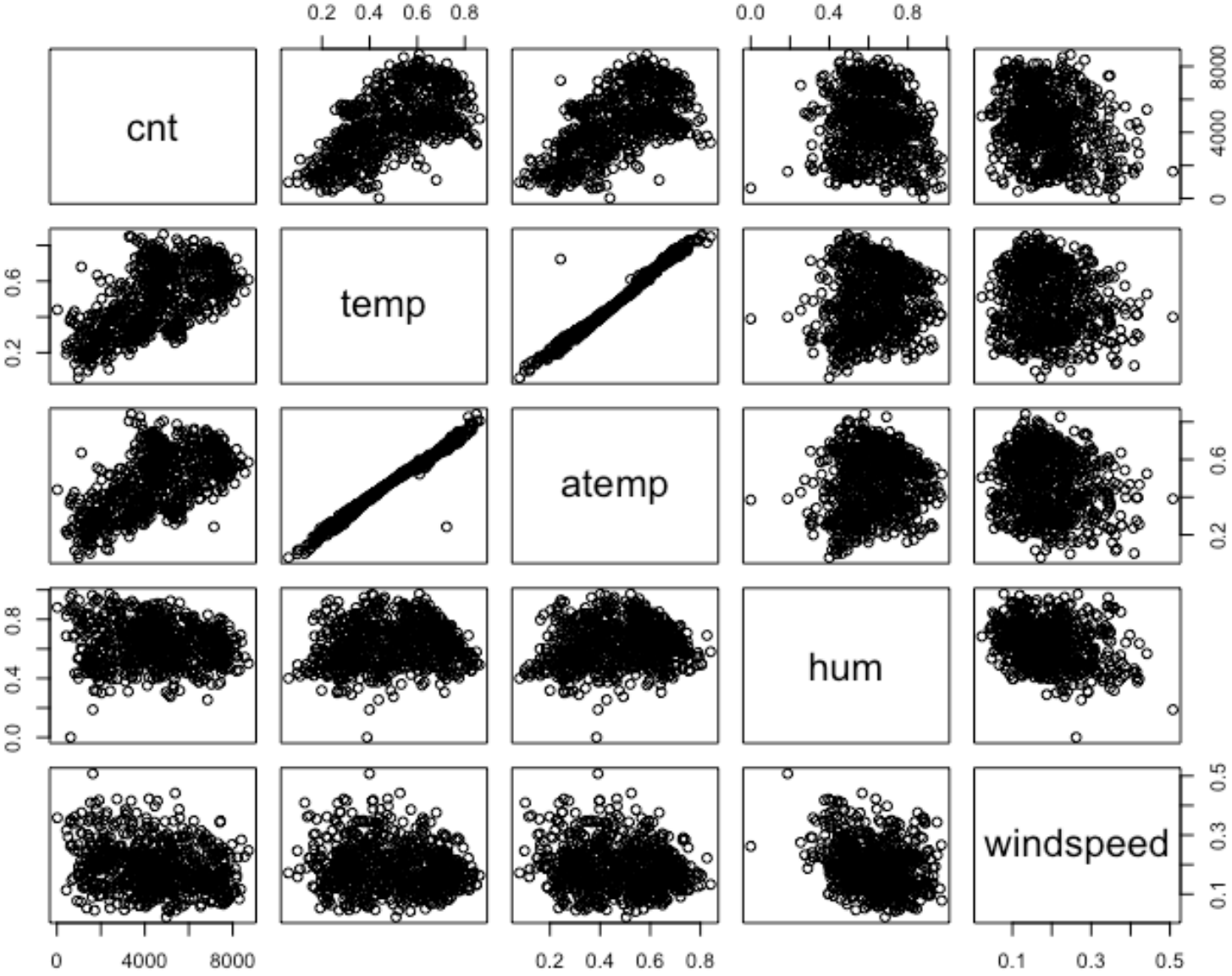
No

Attribute Information

Additional Information

Both hour.csv and day.csv have the following fields, except hr which is not available in day.csv

- instant: record index
- dteday : date
- season : season (1:winter, 2:spring, 3:summer, 4:fall)
- yr : year (0: 2011, 1:2012)
- mnth : month (1 to 12)
- hr : hour (0 to 23)
- holiday : weather day is holiday or not (extracted from http://dchr.dc.gov/page/holiday-schedule)
- weekday : day of the week
- workingday : if day is neither weekend nor holiday is 1, otherwise is 0.
+ weathersit :
  - 1: Clear, Few clouds, Partly cloudy, Partly cloudy
  - 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  - 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  - 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
- temp : Normalized temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-8, t_max=+39 (only in hourly scale)
- atemp: Normalized feeling temperature in Celsius. The values are derived via (t-t_min)/(t_max-t_min), t_min=-16, t_max=+50 (only in hourly scale)
- hum: Normalized humidity. The values are divided to 100 (max)
- windspeed: Normalized wind speed. The values are divided to 67 (max)
- casual: count of casual users
- registered: count of registered users
- cnt: count of total rental bikes including both casual and registered


cnt
temp
atemp
hum
windspeed
0.2 0.4 0.6 0.8
0.0 0.4 0.8
0 4000 8000
0.2 0.6
0.0 0.4 0.8
0.2 0.6
0.1 0.3 0.5
0 4000 8000
0.2 0.4 0.6 0.8
0.1 0.3 0.5

```
Call:
lm(formula = cnt ~ atemp + atemp2 + windspeed + hum, data = reduced.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3152.4  -966.9  -121.2  1063.9  3259.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    380.7      427.2   0.891    0.373    
atemp        28227.1     1748.7  16.141  < 2e-16 ***
atemp2      -22178.4     1848.8 -11.996  < 2e-16 ***
windspeed    -4752.6      640.5  -7.420 3.28e-13 ***
hum          -4428.2      361.0 -12.267  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1282 on 724 degrees of freedom
Multiple R-squared:  0.562,	Adjusted R-squared:  0.5596 
F-statistic: 232.2 on 4 and 724 DF,  p-value: < 2.2e-16
```

Note: We skip the diagnostics here.

1. Write out the fitted model with the estimated coefficient values.
2. Interpret (1) the intercept, and (2) the slope of humidity in the context of the problem.
3. Report and interpret the R^2 value.
4. Construct the 90% confidence interval for the slope coefficient of humidity.

Construct and interpret a 95% confidence interval and prediction interval of the predicted outcome when atemp = .5, hum = .5, windspeed = .3.

```
> predict(m, newdata=data.frame(atemp = .5, atemp2 = .25, hum = .5, windspeed = .3),
+         interval="confidence", level=.95)
       fit      lwr      upr
1 5254.175 5043.527 5464.823
> predict(m, newdata=data.frame(atemp = .5, atemp2 = .25, hum = .5, windspeed = .3),
+         interval="predict", level=.95)
       fit      lwr      upr
1 5254.175 2669.701 7838.649
```

The F-test of global model utility.

```
Call:
lm(formula = cnt ~ atemp + atemp2 + windspeed + hum, data = reduced.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3152.4  -966.9  -121.2  1063.9  3259.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    380.7      427.2   0.891    0.373    
atemp        28227.1     1748.7  16.141  < 2e-16 ***
atemp2      -22178.4     1848.8 -11.996  < 2e-16 ***
windspeed    -4752.6      640.5  -7.420 3.28e-13 ***
hum          -4428.2      361.0 -12.267  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1282 on 724 degrees of freedom
Multiple R-squared:  0.562,	Adjusted R-squared:  0.5596 
F-statistic: 232.2 on 4 and 724 DF,  p-value: < 2.2e-16
```

The F-test of extra sum of squares of humidity.

```
> anova(m3)
Analysis of Variance Table

Response: cnt
           Df     Sum Sq    Mean Sq F value    Pr(>F)
atemp       1 1098537076 1098537076 668.207 < 2.2e-16 ***
atemp2      1  144497807  144497807  87.894 < 2.2e-16 ***
windspeed   1   36775803   36775803  22.370 2.704e-06 ***
hum         1  247407931  247407931 150.491 < 2.2e-16 ***
Residuals 724 1190260914    1644007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The F-test of the linear testing approach

```
> # The linear testing approach: whether humidity should be kept in the model when
> # when they other terms are in.
>
> full.model <- lm(cnt ~ atemp + atemp2 + windspeed + hum, data=reduced.data)
> reduced.model <- lm(cnt ~ atemp + atemp2 + windspeed, data=reduced.data)
> anova(reduced.model, full.model)
Analysis of Variance Table

Model 1: cnt ~ atemp + atemp2 + windspeed
Model 2: cnt ~ atemp + atemp2 + windspeed + hum
  Res.Df        RSS Df Sum of Sq      F    Pr(>F)
1    725 1437668845
2    724 1190260914  1 247407931 150.49 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The effect size for the model comparison: Partial R-squared

```
> anova(full.model)
Analysis of Variance Table

Response: cnt
           Df     Sum Sq    Mean Sq F value    Pr(>F)
atemp       1 1098537076 1098537076 668.207 < 2.2e-16 ***
atemp2      1  144497807  144497807  87.894 < 2.2e-16 ***
windspeed   1   36775803   36775803  22.370 2.704e-06 ***
hum         1  247407931  247407931 150.491 < 2.2e-16 ***
Residuals 724 1190260914    1644007
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> anova(reduced.model)
Analysis of Variance Table

Response: cnt
           Df     Sum Sq    Mean Sq F value    Pr(>F)
atemp       1 1098537076 1098537076 553.980 < 2.2e-16 ***
atemp2      1  144497807  144497807  72.869 < 2.2e-16 ***
windspeed   1   36775803   36775803  18.546 1.887e-05 ***
Residuals 725 1437668845    1982992
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The effect size for the model comparison: Compare the R-squares

```
Call:
lm(formula = cnt ~ atemp + atemp2 + windspeed + hum, data = reduced.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-3152.4  -966.9  -121.2  1063.9  3259.1 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)    380.7      427.2   0.891    0.373    
atemp        28227.1     1748.7  16.141  < 2e-16 ***
atemp2      -22178.4     1848.8 -11.996  < 2e-16 ***
windspeed    -4752.6      640.5  -7.420 3.28e-13 ***
hum          -4428.2      361.0 -12.267  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1282 on 724 degrees of freedom
Multiple R-squared:  0.562,	Adjusted R-squared:  0.5596 
F-statistic: 232.2 on 4 and 724 DF,  p-value: < 2.2e-16
```

```
Call:
lm(formula = cnt ~ atemp + atemp2 + windspeed, data = reduced.data)

Residuals:
    Min      1Q  Median      3Q     Max 
-4165.3 -1042.6  -124.4  1119.3  3384.4 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  -1532.6      436.8  -3.509 0.000478 ***
atemp        22745.0     1856.8  12.249  < 2e-16 ***
atemp2      -16668.3     1969.6  -8.463  < 2e-16 ***
windspeed    -2948.3      684.6  -4.306 1.89e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1408 on 725 degrees of freedom
Multiple R-squared:  0.471,	Adjusted R-squared:  0.4688 
F-statistic: 215.1 on 3 and 725 DF,  p-value: < 2.2e-16
```