2021/01/20 Ex.2

Marco Scarpelli

05 giugno, 2024

Dataset exploration

```
##   bike_count mean_temp mean_wind        day
## 1       5484      23.7       1.6 No Holiday
## 2       2682      -6.9       0.8 No Holiday
## 3       5424      26.7       1.1 No Holiday
## 4       5852      13.4       1.6 No Holiday
## 5       3515       6.3       2.2 No Holiday
## 6       5114      12.4       1.0 No Holiday
```

```
## [1] 50  4
```

Point a

```
## 
## Call:
## lm(formula = bike_count ~ dummy + mean_temp:dummy + mean_wind:dummy + 
##     mean_temp + mean_wind, data = df)
## 
## Residuals:
##     Min      1Q  Median      3Q     Max 
## -3790.5  -799.5   115.5   880.1  2837.4 
## 
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)    
## (Intercept)      4084.39    1211.00   3.373 0.001561 ** 
## dummy            -798.28    1446.13  -0.552 0.583733    
## mean_temp         118.61      32.01   3.706 0.000586 ***
## mean_wind        -225.00     626.91  -0.359 0.721383    
## dummy:mean_temp   -31.18      40.56  -0.769 0.446198    
## dummy:mean_wind  -216.73     767.85  -0.282 0.779072    
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 1441 on 44 degrees of freedom
## Multiple R-squared:  0.4418, Adjusted R-squared:  0.3784 
## F-statistic: 6.966 on 5 and 44 DF,  p-value: 7.208e-05
```

The parameters are:

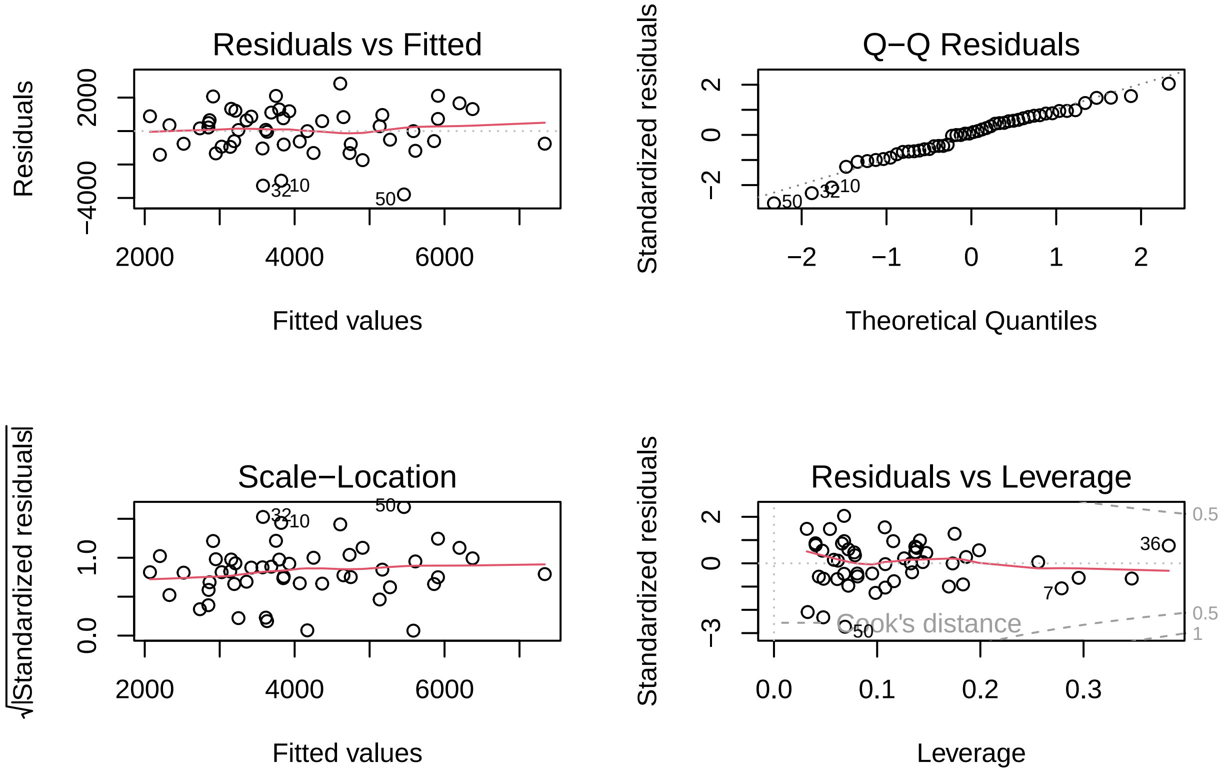
```
##     (Intercept)           dummy       mean_temp       mean_wind dummy:mean_temp 
##      4084.38789      -798.28039       118.60701      -224.99991       -31.17918 
## dummy:mean_wind
##      -216.72851
```

```
## [1] 2075365
```

Point b

Assumptions on the model

We assume homoscedastic residuals:



```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fm)
## W = 0.96913, p-value = 0.2133
```

We can see that the Gaussianity test succeds; furthermore, the plots show that all points are within Cook's distance, the Q-Q plot follows the line closely enough and the residuals exhibit no clear patttern.

New model

From the summary above, it seems that only `mean_wind` should be removed from the model, and it also seems that holiday information is not signficant (but we should verify that by removing only one thing at a time. However, there is a debate on whether the exam text is asking to remove both `mean_wind` and `mean_temp`. Here, I will remove both.

Weather info

We will remove weather information completely from the model and run `anova` on the old model and new model.

```
##
## Call:
## lm(formula = bike_count ~ dummy, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3315.3 -1133.9  -256.8  1289.2  3816.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   4850.8      411.7  11.784    9e-16 ***
## dummy        -1221.5      514.6  -2.374   0.0217 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1746 on 48 degrees of freedom
## Multiple R-squared:  0.1051, Adjusted R-squared:  0.08642
## F-statistic: 5.635 on 1 and 48 DF,  p-value: 0.02165
```

ANOVA:

```
## Analysis of Variance Table
##
## Model 1: bike_count ~ dummy + mean_temp:dummy + mean_wind:dummy + mean_temp +
##     mean_wind
## Model 2: bike_count ~ dummy
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1     44  91316071
## 2     48 146411890 -4 -55095819 6.6369 0.0002863 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The two models are different and the RSS of the second model is way higher, so we keep the first model. This means that weather info has some play in the prediction.

Holiday info

We will remove all holiday info.

```
##
## Call:
## lm(formula = bike_count ~ mean_temp + mean_wind, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3636.5 -1046.7   207.6   793.8  3251.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3220.46     720.24   4.471 4.90e-05 ***
## mean_temp      91.69      21.48   4.269 9.45e-05 ***
## mean_wind    -109.91     389.53  -0.282    0.779
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1584 on 47 degrees of freedom
## Multiple R-squared:  0.2794, Adjusted R-squared:  0.2488
## F-statistic: 9.112 on 2 and 47 DF,  p-value: 0.0004525
```

ANOVA:

```
## Analysis of Variance Table
##
## Model 1: bike_count ~ dummy + mean_temp:dummy + mean_wind:dummy + mean_temp +
##     mean_wind
## Model 2: bike_count ~ mean_temp + mean_wind
##   Res.Df       RSS Df Sum of Sq      F   Pr(>F)
## 1     44  91316071
## 2     47 117888169 -3 -26572098 4.2679 0.009922 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Again, the complete model was better. The holiday has some play in the prediction.

Point c

From what we already stated in point B, we will first remove `mean_wind` since it seems to have low significance.

```
##
## Call:
## lm(formula = bike_count ~ dummy + mean_temp:dummy + mean_temp,
##     data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3775.5  -878.9   109.0   933.1  2947.9
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      3682.96     459.80   8.010 2.86e-10 ***
## dummy           -1080.41     602.09  -1.794 0.079318 .
## mean_temp         117.37      31.51   3.724 0.000533 ***
## dummy:mean_temp   -31.99      39.98  -0.800 0.427757
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1427 on 46 degrees of freedom
## Multiple R-squared:  0.4276, Adjusted R-squared:  0.3903
## F-statistic: 11.45 on 3 and 46 DF,  p-value: 9.872e-06
```

We can remove the interaction between the dummy and the mean temperature:

```
##
## Call:
## lm(formula = bike_count ~ dummy + mean_temp, data = df)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3675.1  -798.9    43.6   949.1  2824.7
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3880.73     386.22  10.048 2.74e-13 ***
## dummy       -1423.79     420.67  -3.385  0.00145 **
## mean_temp      97.49      19.32   5.047 7.17e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1421 on 47 degrees of freedom
## Multiple R-squared:  0.4196, Adjusted R-squared:  0.3949
## F-statistic: 16.99 on 2 and 47 DF,  p-value: 2.798e-06
```

Let us check with ANOVA:

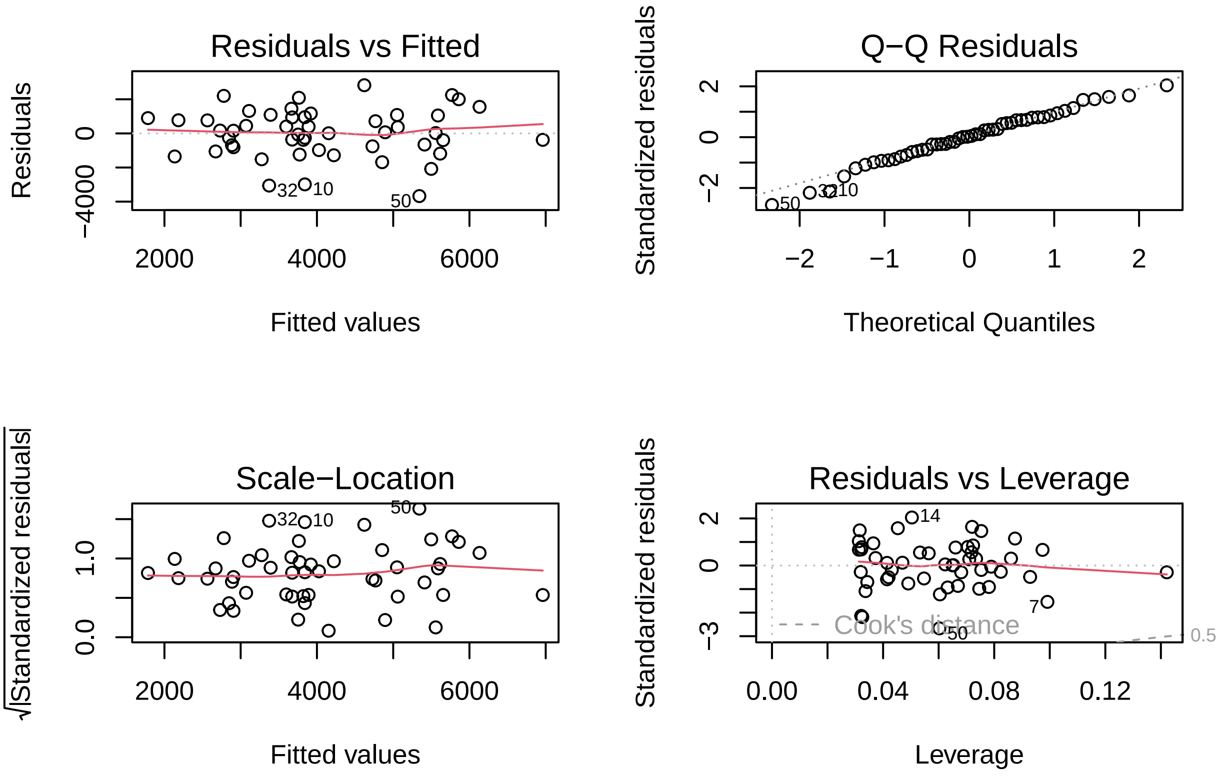
```
## Analysis of Variance Table
##
## Model 1: bike_count ~ dummy + mean_temp:dummy + mean_wind:dummy + mean_temp +
##     mean_wind
## Model 2: bike_count ~ dummy + mean_temp
##   Res.Df      RSS Df Sum of Sq      F Pr(>F)
## 1     44 91316071
## 2     47 94946679 -3  -3630608 0.5831 0.6292
```

The parameters are:

```
## (Intercept)       dummy   mean_temp
##  3880.73485 -1423.79191    97.49175
```

```
## [1] 2157879
```

Diagnostics:



```
##
##  Shapiro-Wilk normality test
##
## data:  residuals(fm4)
## W = 0.98129, p-value = 0.6076
```

Point D

Confidence intervals:

```
##            1
## fit 4075.718
## lwr 3334.344
## upr 4817.093
```

Prediction:

```
##            1
## fit 4075.718
## lwr 1121.848
## upr 7029.589
```