Statistics and Data Analysis
Regression Analysis (2)

Instructor: Ling-Chieh Kung
Department of Information Management
National Taiwan University

1. The "Bike_Month" sheet contains the monthly number public bike rentals in a city and other related information. A simple linear regression model with only *instant* and *cnt* is

$$cnt = 69033 + 5453 instant.$$

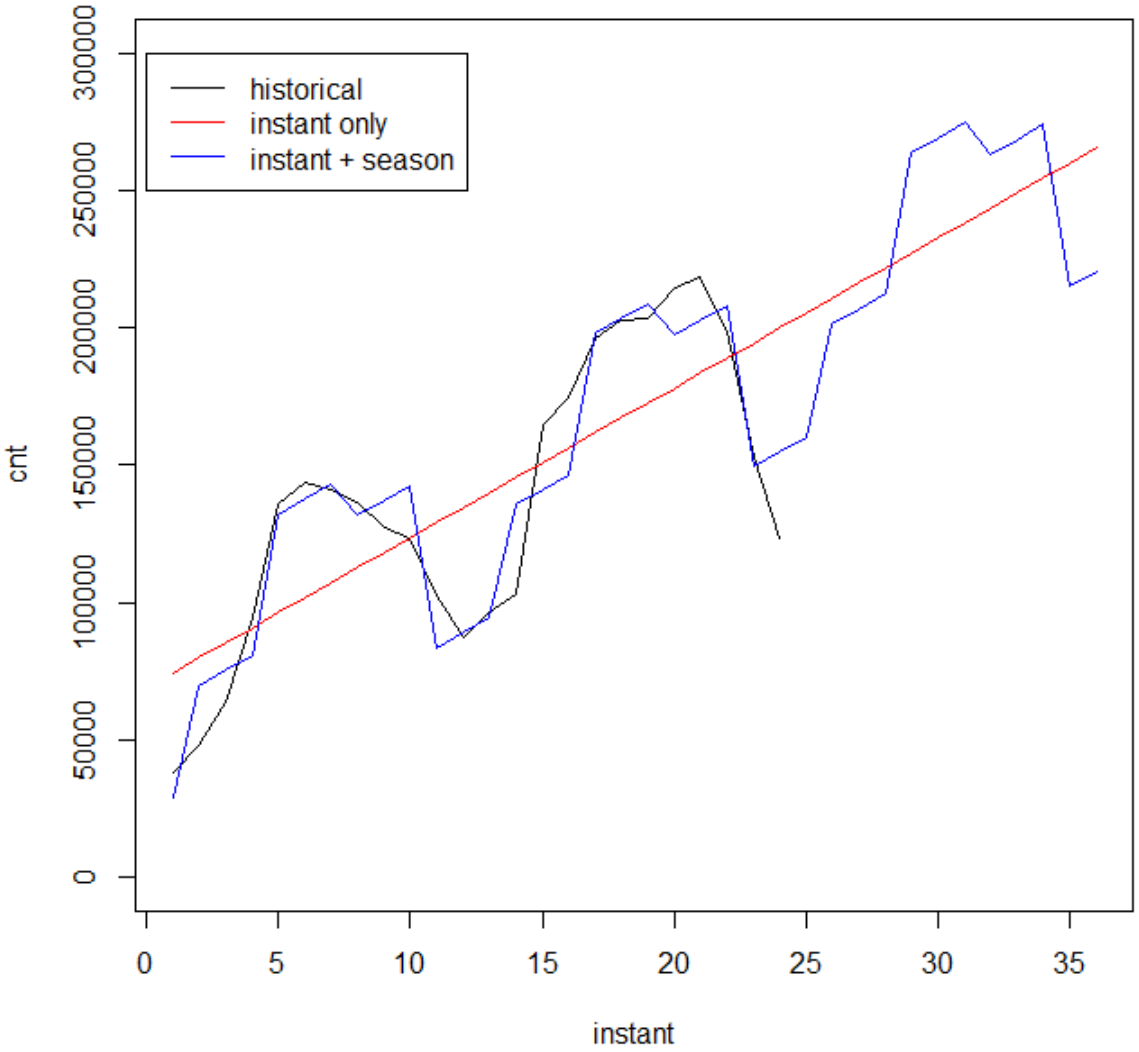
The regression model captures the trend but not the seasonal effect. Let's try to add the seasonal effect into the model.

(a) Construct a multiple linear regression model for *season*, *instant*, and *cnt*. Try to interpret the model. Is there anything weird?

(b) Construct three new columns *spring*, *summer*, and *fall* as the indicator variables for the season to be 1, 2, and 3. Construct a regression model for *cnt* with *instant* and the three indicator variables. Validate and interpret the model.

(c) Predict monthly rentals for the next January, February, and March with the old (*instant* only) and new models (*instant* plus the three indicator variables). Compare the outcomes.

(d) Let's visualize the difference.


historical
instant only
instant + season
cnt
0
50000
100000
150000
200000
250000
300000
0
5
10
15
20
25
30
35
instant

(e) Change the reference level of *season* to 3 (i.e., fall). Construct a regression model for *cnt* with *instant* and the three indicator variables. Validate and interpret the model.

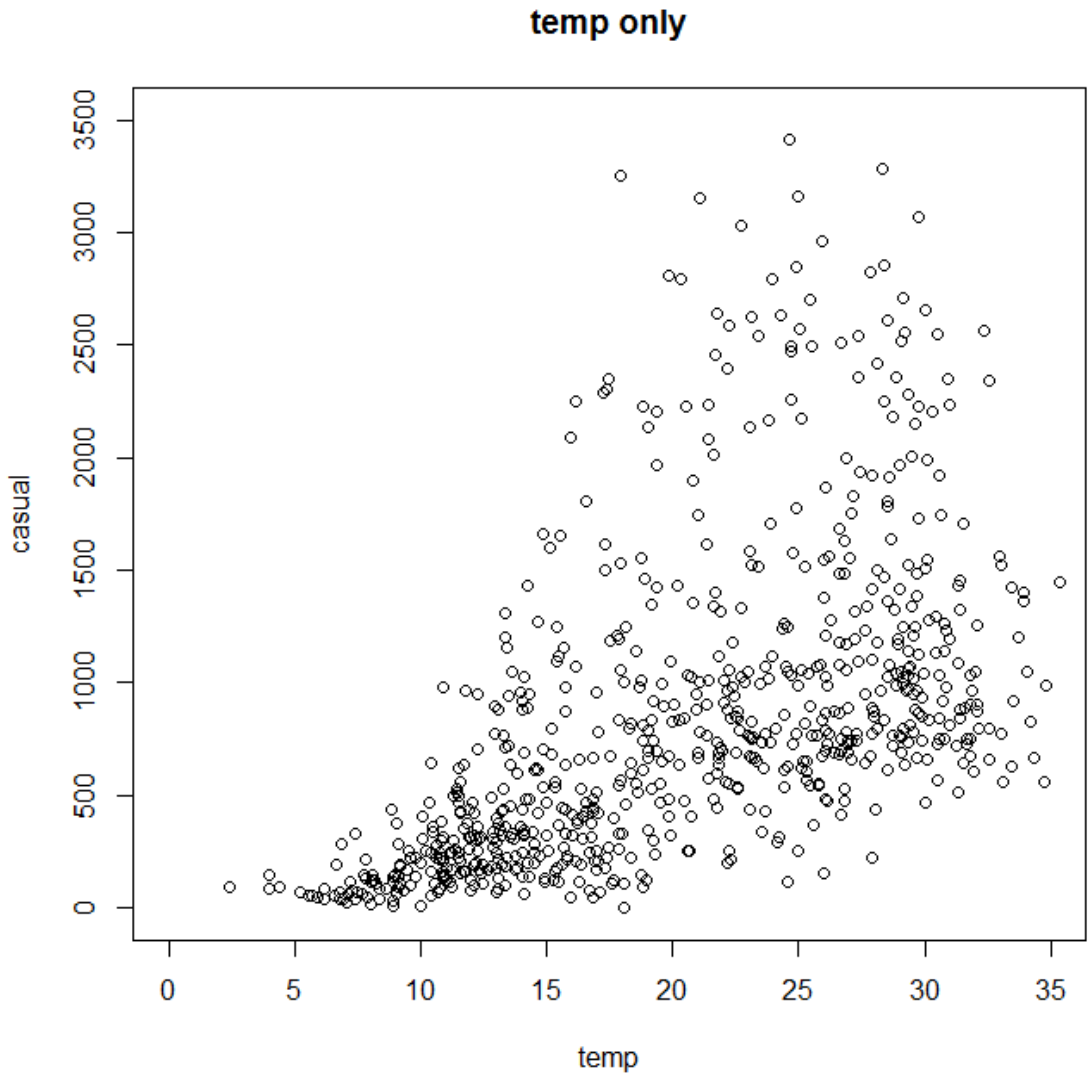
(f) Predict monthly rentals for the next January, February, and March with the two models with *instant* and three indicator variables. Compare the outcomes.

2. The "Bike_Day" sheet in "SDA-Fa15_cp11_data.xlsx" contains the daily number public bike rentals in a city and other related information. For the variable *weathersit*, 1 means sunny, 2 means cloudy, and 3 means rainy or snowy.

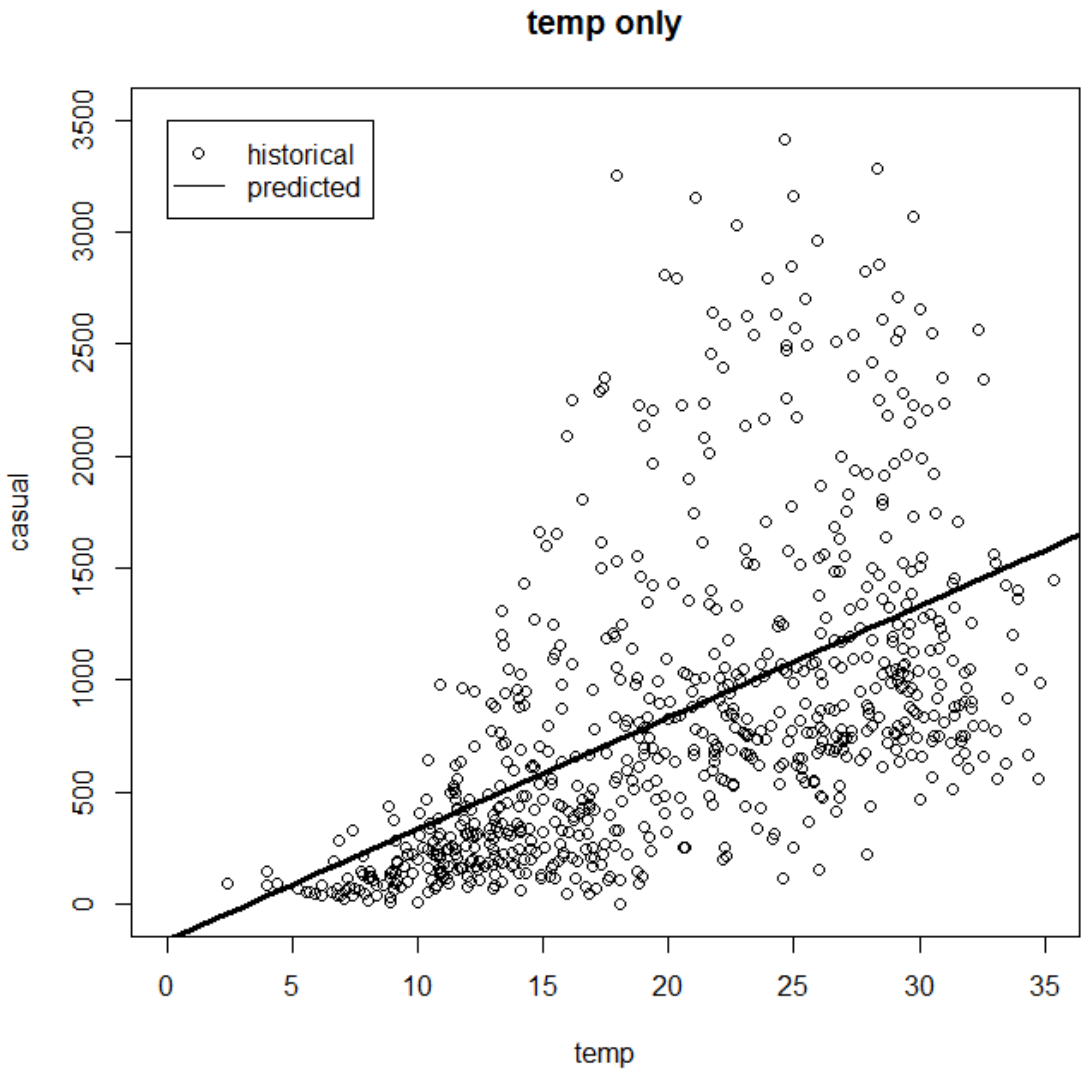
   (a) If we construct a regression model with *instant*, *weathersit*, and *cnt*, what is wrong?

   (b) Create indicator variables for *weathersit* by choosing sunny as the reference level. Construct a regression model with *instant*, the indicator variables for *weathersit*, and *cnt*. Validate and interpret the model.

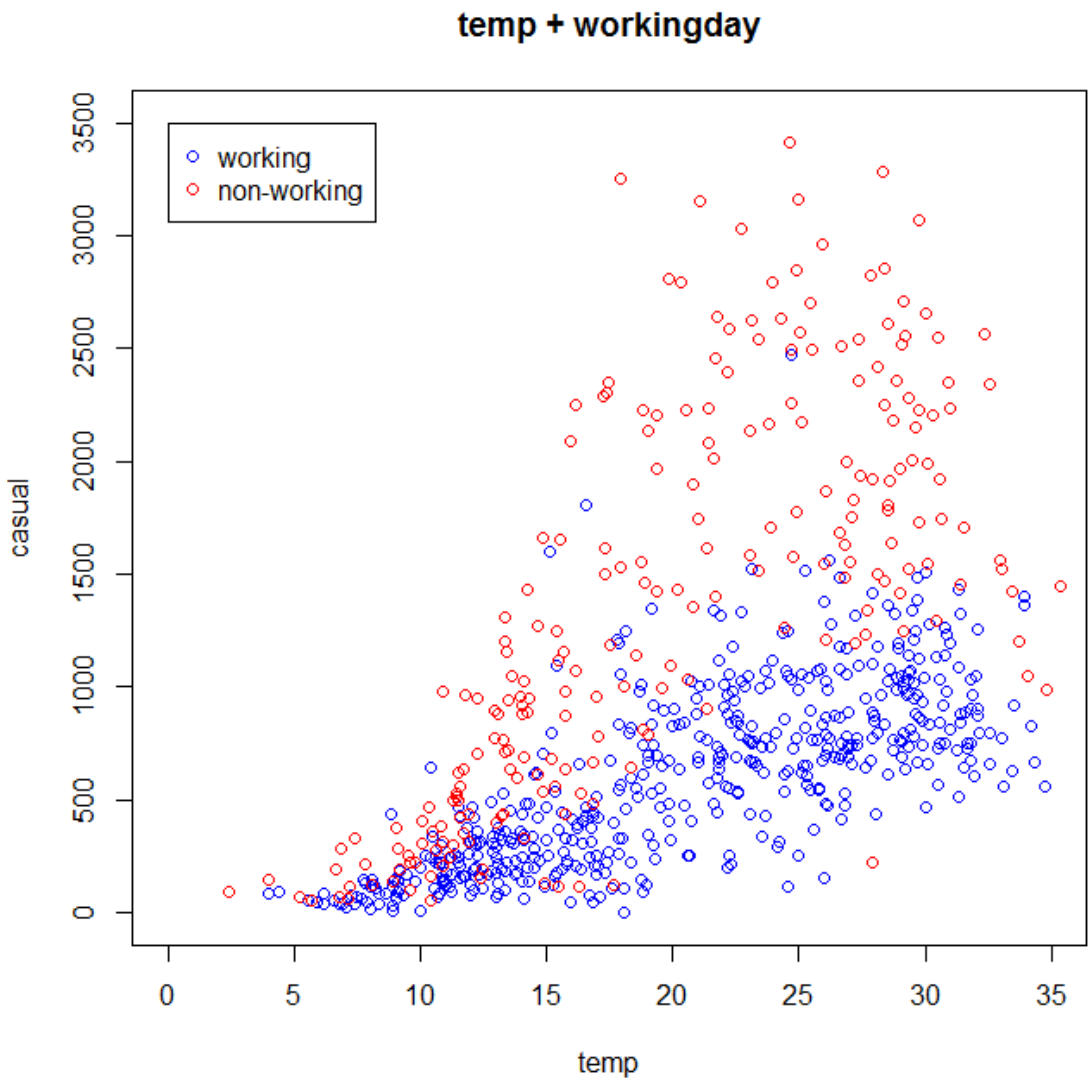
3. Consider the daily rental data. The scatter plot here suggests that *temp* affects *casual*.



(a) Construct a regression model with *temp* and *casual*. Validate and interpret the model. Construct the scatter plot for *temp* and *casual*. Add the linear trend line into the plot.



(b) It is known that *casual* has different patterns for working and non-working days. We suspect that weather condition affects *casual* in a different way for different values of *workingday* (1 for working and 0 for not). The scatter plot here suggests this.

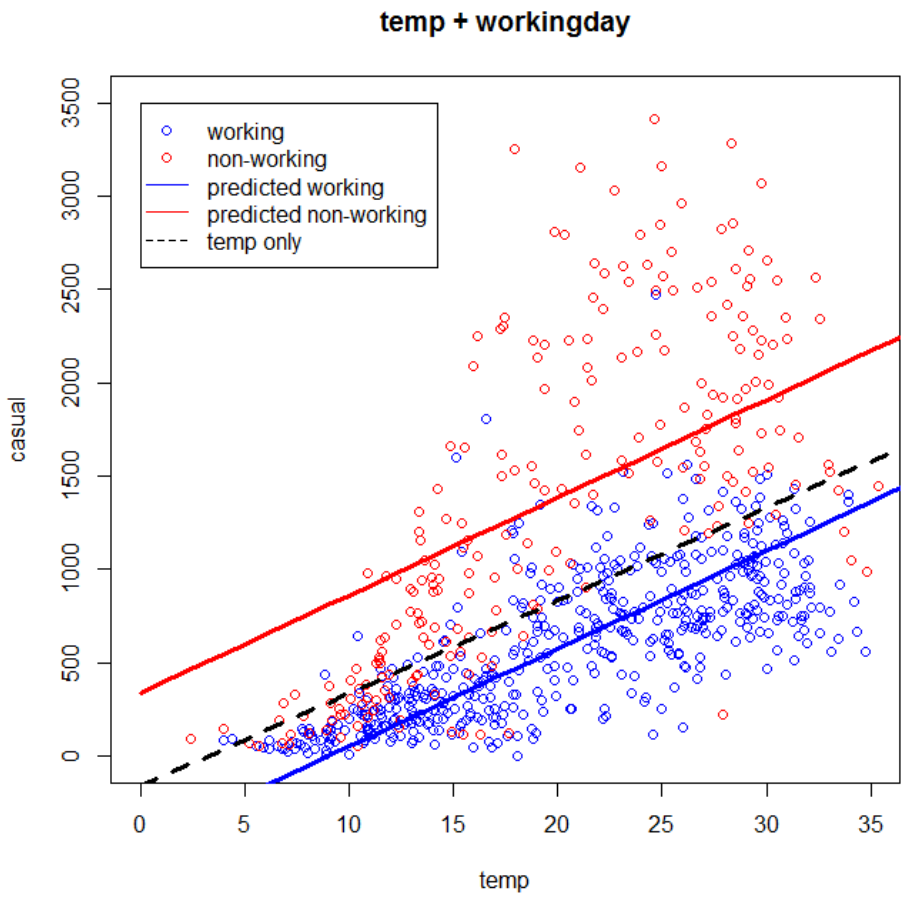


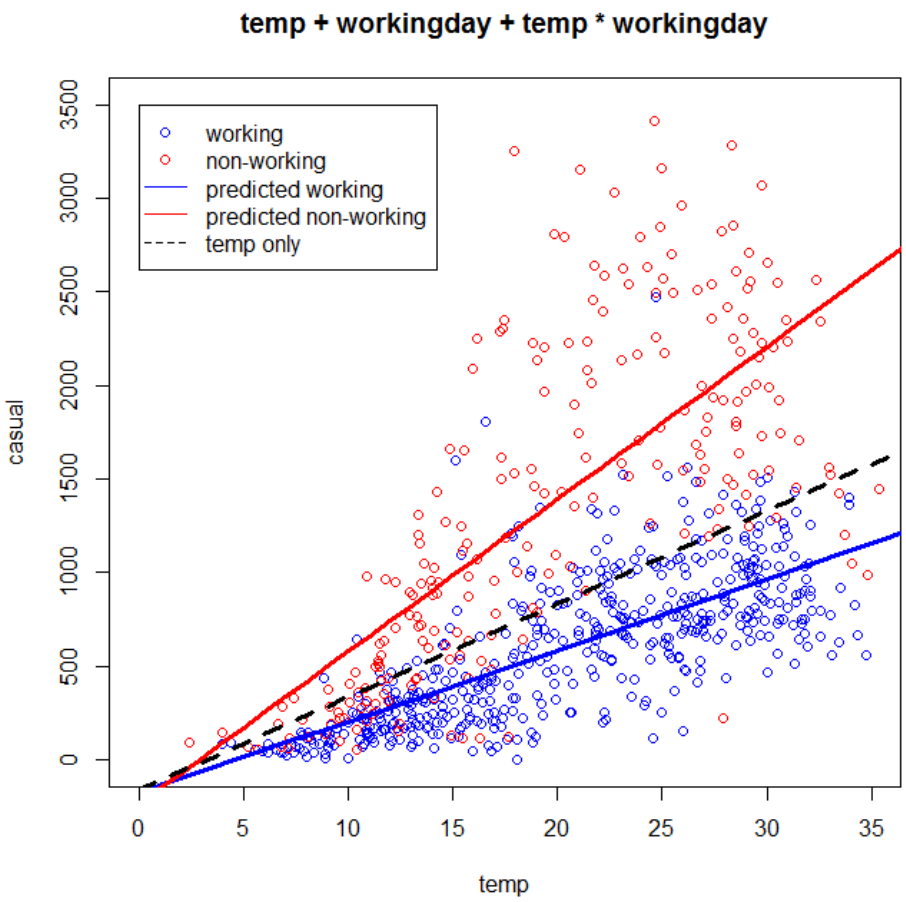
(c) Construct a regression model with *temp*, *workingday*, and *casual* with no interaction. Validate and interpret the model.

(d) Construct a regression model with *temp*, *workingday*, *temp* $\times$ *workingday*, and *casual*. Validate and interpret the model.

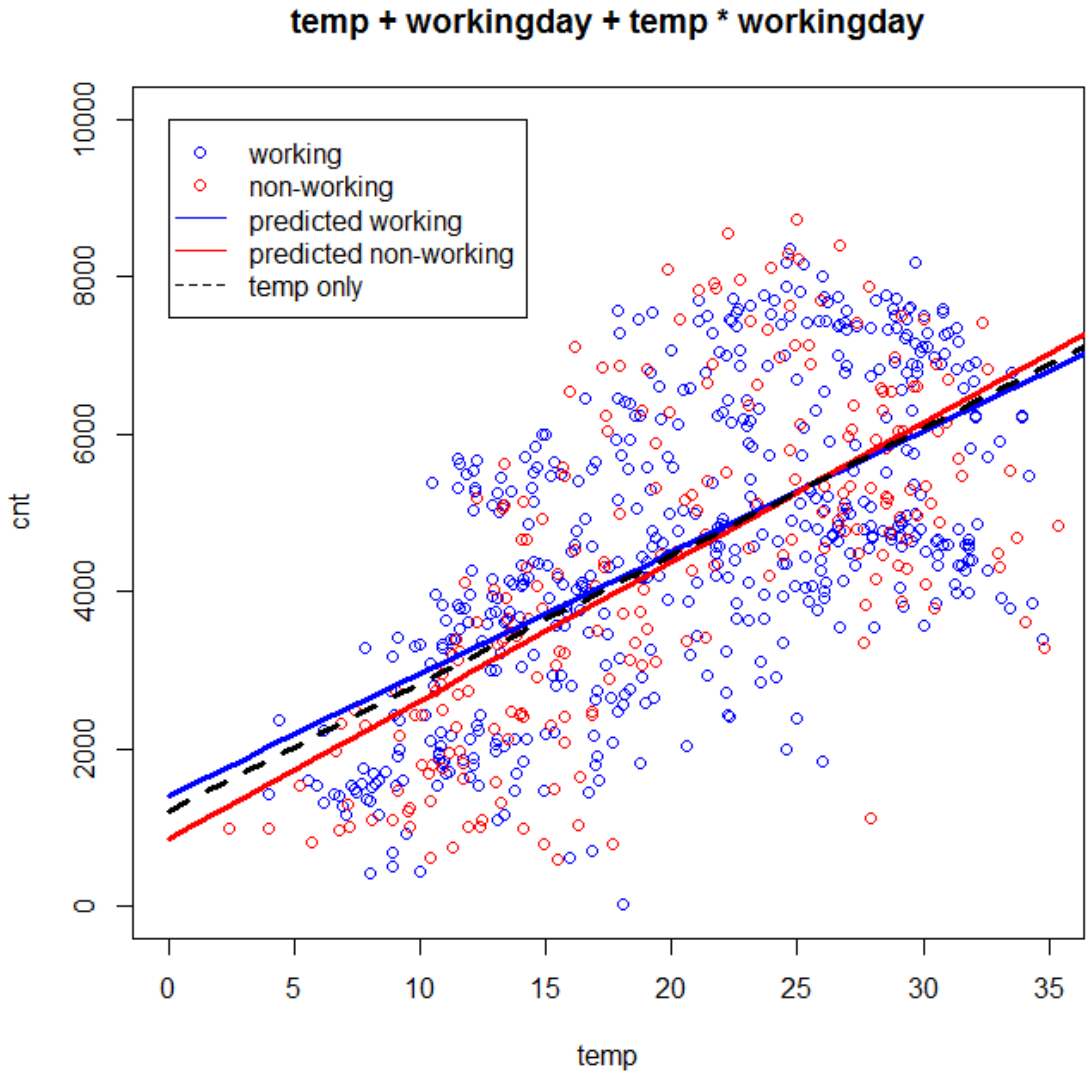
(e) Compare the above two models. Which one is better?

(f) Let's visualize the difference among the three models.


temp + workingday
working
non-working
predicted working
predicted non-working
temp only
casual
temp
0
500
1000
1500
2000
2500
3000
3500
0
5
10
15
20
25
30
35


temp + workingday + temp * workingday
working
non-working
predicted working
predicted non-working
temp only
casual
temp
0
500
1000
1500
2000
2500
3000
3500
0
5
10
15
20
25
30
35

4. Construct a multiple linear regression model with *temp*, *workingday*, *temp* × *workingday*, and *cnt*. Validate and interpret the model.



5. To predict or explain *cnt*, is it good to include *casual* or *registered* as independent variables?