Homework 8

Categorical Variables

due Oct 29 by 330pm

Question 1

Load in the `EX7.BIKE` dataset using the `data` command. This is the DC bicycle demand dataset with additional predictors besides temperature, humidity, and windspeed. There is also information about whether the day was a holiday or a working day, the day of the week, and whether it was rainy.

```
#Load up EX7.BIKE
data("EX7.BIKE")
```

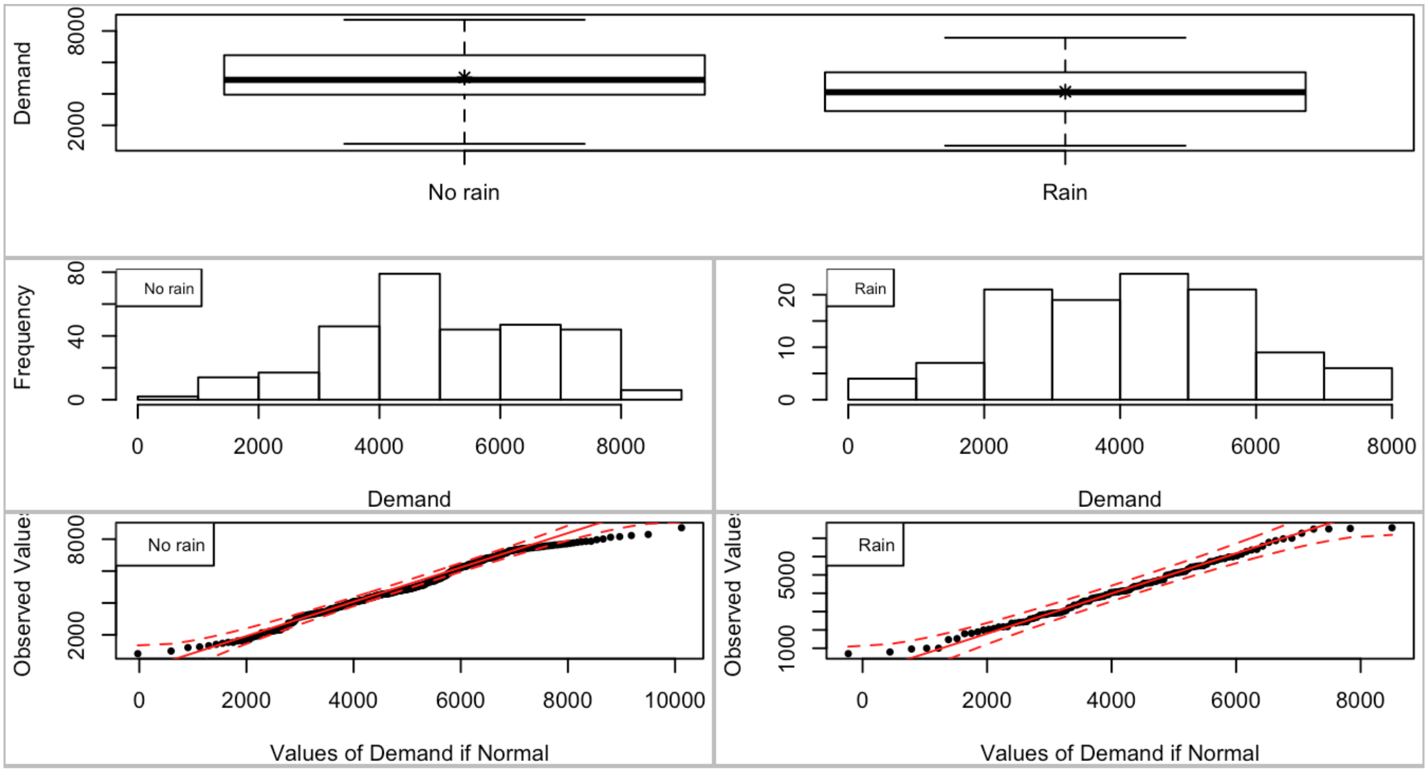
a. If we represent the categorical variable `Workingday` (yes vs. no) by an indicator variable, how should it be numerically defined? Which level is the reference level?

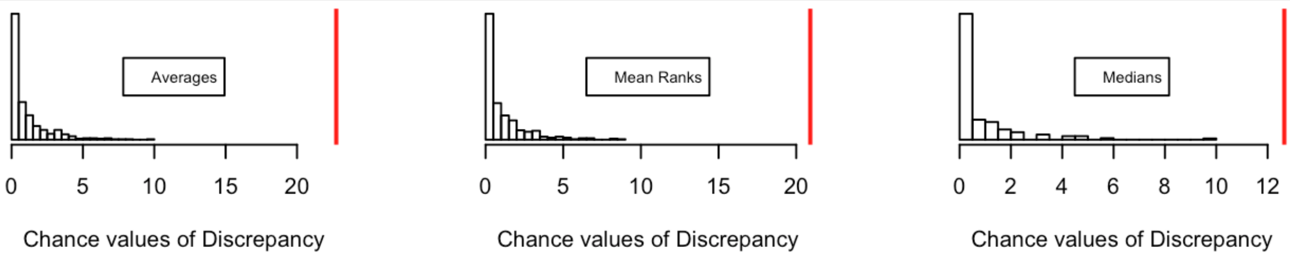
Response: 0: no; 1: yes. 0/no is reference level.

---

b. What are the overall average demands on rainy vs. non-rainy days? What is the difference in average demands, and is this difference statistically significant? Note: you can use either `associate` or fit a regression predicting `Demand` with `Weather` as the sole predictor to answer this question.

```
#Code obtaining averages, difference in averages, and statistical significance
associate(Demand~Weather, data=EX7.BIKE)
```


Demand
No rain
Rain
Frequency
Demand
Observed Values
Values of Demand if Normal


Averages
Mean Ranks
Medians
Chance values of Discrepancy

```
## Association between Weather (categorical) and  Demand (numerical)
##  using 410 complete cases
##
## Sample Sizesx
## No rain    Rain
##     299     111
##
## Permutation procedure:
##                      No rain Rain Discrepancy Estimated p-value
## Averages (ANOVA)        5046 4137       22.75                 0
## Mean Ranks (Kruskal)   205.3  206        20.9                 0
## Medians                 4891 4105       12.65                 0
## With 500 permutations, we are 95% confident that
##  the p-value of ANOVA (means) is between 0 and 0.007
##  the p-value of Kruskal-Wallis (ranks) is between 0 and 0.007
##  the p-value of median test is between 0 and 0.007
## Note:  If 0.05 is in a range, change permutations= to a larger number
##
##
##
## Advice: If it makes sense to compare means (i.e., no extreme outliers and the
## distributions aren't too skewed), use the ANOVA.  If there there are
## some obvious extreme outliers but the distributions are roughly symmetric, use
## Rank test.  Otherwise, use the Median test or rerun the test using, e.g., log1
0(y)
## instead of y
M <- lm(Demand~Weather, data=EX7.BIKE)
summary(M)
##
## Call:
## lm(formula = Demand ~ Weather, data = EX7.BIKE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4223.8 -1200.8  -101.5  1359.4  3668.2
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  5045.84      99.13   50.90  < 2e-16 ***
## WeatherRain  -908.69     190.52   -4.77 2.57e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1714 on 408 degrees of freedom
## Multiple R-squared:  0.05281,    Adjusted R-squared:  0.05049
## F-statistic: 22.75 on 1 and 408 DF,  p-value: 2.574e-06
```

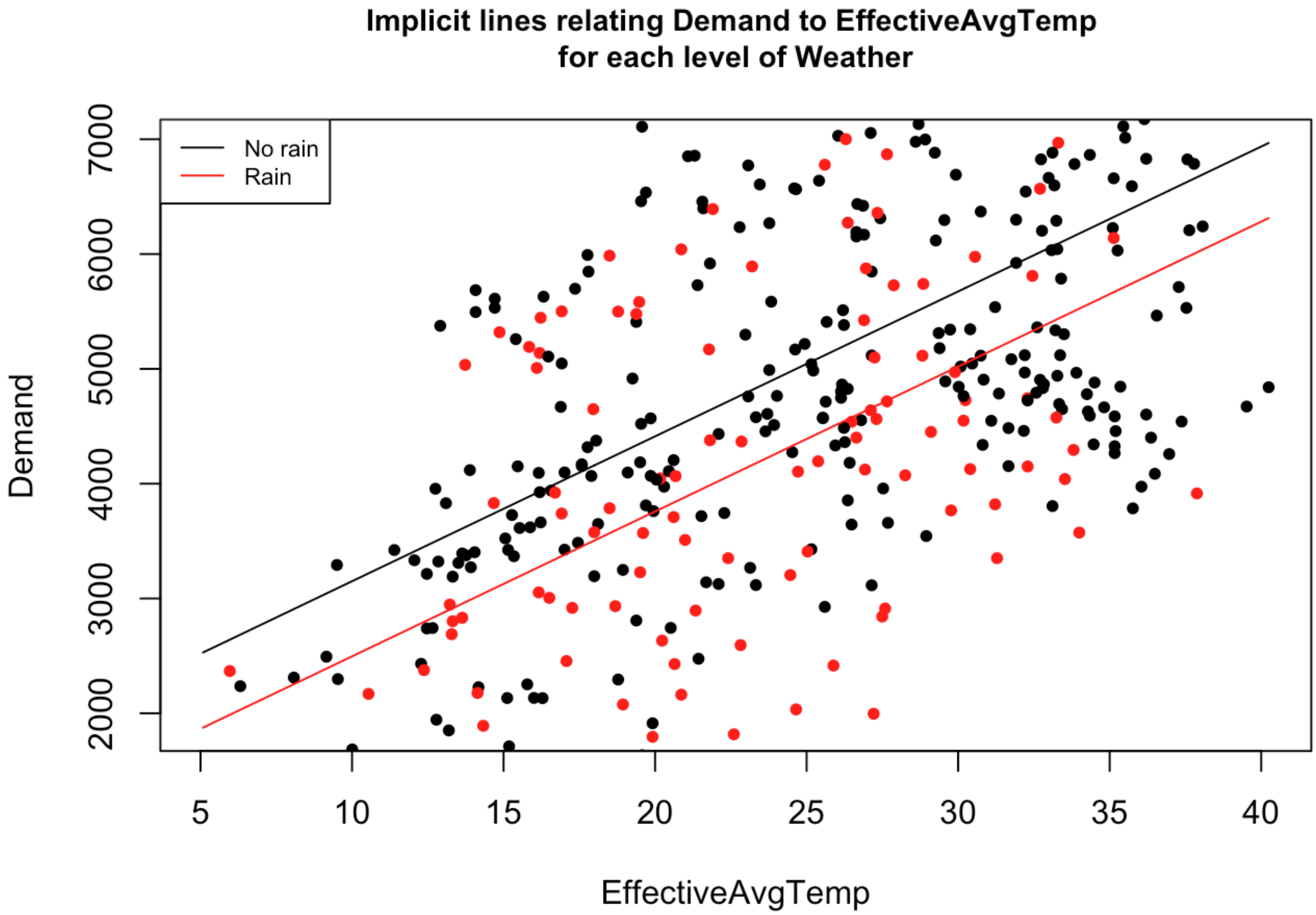
Response: Average demand for rainy weather: 4137 Average demand for non-rainy weather: 5046 Difference in average demands: 908.69 Yes, this difference is statistically significant. The p-value is 0.

c. If we really want to compare the average demands between rainy and non-rainy days, we really should take into account the effective average temperature since it tends to be cooler on days that rain (and cooler days typically have less demand anyway). Fit a regression predicting `Demand` from `EffectiveAvgTemp` and `Weather` (no interactions). Also run `summary` and `visualize.model`

```
#Model predicting Demand from EffectiveAvgTemp and Weather (no interactions)
N <- lm(Demand~Weather+EffectiveAvgTemp, data=EX7.BIKE)
#summary of model
summary(N)
```

```
##
## Call:
## lm(formula = Demand ~ Weather + EffectiveAvgTemp, data = EX7.BIKE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3197.7 -1039.8  -199.2  1139.8  3128.8
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      1887.988    241.363   7.822 4.51e-14 ***
## WeatherRain      -654.361    158.121  -4.138 4.25e-05 ***
## EffectiveAvgTemp  126.214      9.077  13.905  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1413 on 407 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3547
## F-statistic: 113.4 on 2 and 407 DF,  p-value: < 2.2e-16
```

```
#visualize model
visualize.model(N)
```




```
##
## Effect test for Weather has p-value
```

c1. You see that the implicit regression equations for rainy and non-rainy days are parallel to each other. Is this always the case when a model is fit without interactions?

Response: Yes, these models are always parallel because the regression lines for categorical variables (without interaction) have same slope.

---

c2. Among days with the same effective temperature, how much smaller is the average demand when it rains compared to when it does not rain. Is the difference in averages statistically significant? Explain.

Response: The difference in averages is smaller (654.361), but more statistically significant (p-value is 0.0000425). This means that demand is less by 654.361 on rainy days.

---

c3. Nothing to answer here, but do take note. Without accounting for the effective average temperature, we saw the difference in overall average demands was around 900 bikes. After accounting for effective average temperature, the difference in average demands is much smaller. We see that a bunch of the variation in demands on rainy vs. non-rainy days can really be attributed to variation in temperatures between these days!

This illustrates how important it is to make sure the two "individuals" you are comparing are identical as possible before looking at the difference in the average values of y. Blindly comparing rain/no rain days we get a difference in averages of around 900 bikes. However, variation in demand is attributable to many other factors. In DC, it rains more often during the winter (when it is colder) vs. the summer (when it is warmer). Intuitively, we know people are more likely to be bike riding in nicer weather. Thus, some of the difference in average demands between rainy/non-rainy days is "due" to differences in the temperature. By putting the effective average temperature into the model, we are able account for this effect since we are comparing the average demands on days with the same effective average temperature.

c4. Write out the implicit regression equations relating `Demand` to `EffectiveAvgTemp` for rainy and for non-rainy days. Round coefficients to the nearest integer.

Response: Demand = 1887.988 - 654.361(WeatherRain) + 126.214(EffectiveAvgTemp)

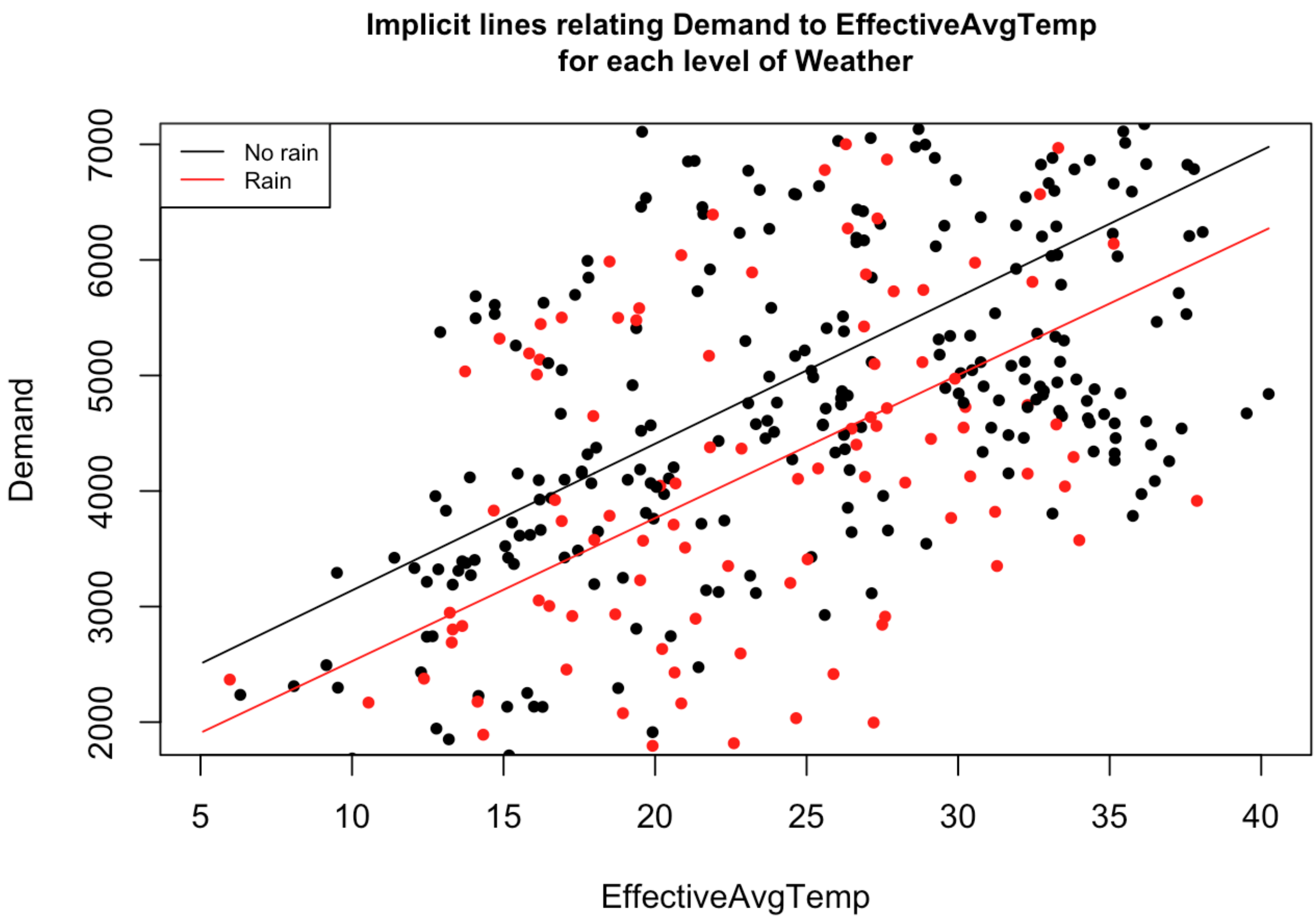
---

c5. An interaction between `Weather` and `EffectiveAvgTemp` would allow the possibility that the strength of the relationship between `Demand` and `EffectiveAvgTemp` may be different for rainy/non-rainy days. Fit the model with the interaction, obtain the summary, and visualize the model. Report the difference in the slopes of the lines relating `Demand` to `EffectiveAvgTemp`, then comment on whether the difference is statistically significant. Do we need to include this interaction effect in the model? Explain.

```
#Model predicting Demand from EffectiveAvgTemp and Weather (with interaction)
P <- lm(Demand~Weather*EffectiveAvgTemp, data=EX7.BIKE)
#summary of model
summary(P)
```

```
##
## Call:
## lm(formula = Demand ~ Weather * EffectiveAvgTemp, data = EX7.BIKE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3193  -1044   -198   1140   3126
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                1871.011    270.397   6.919 1.77e-11 ***
## WeatherRain                -582.550    537.035  -1.085    0.279
## EffectiveAvgTemp            126.892     10.301  12.319  < 2e-16 ***
## WeatherRain:EffectiveAvgTemp   -3.062     21.882  -0.140    0.889
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1415 on 406 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3531
## F-statistic: 75.43 on 3 and 406 DF,  p-value: < 2.2e-16
```

```
#visualize model
visualize.model(P)
```




```
##
## Effect test for interaction with Weather has p-value 0.8888
```

Response: Difference in slopes is 3.062 and the p-value is 0.889. This is not statistically significant. We do not need to include interaction in the model.

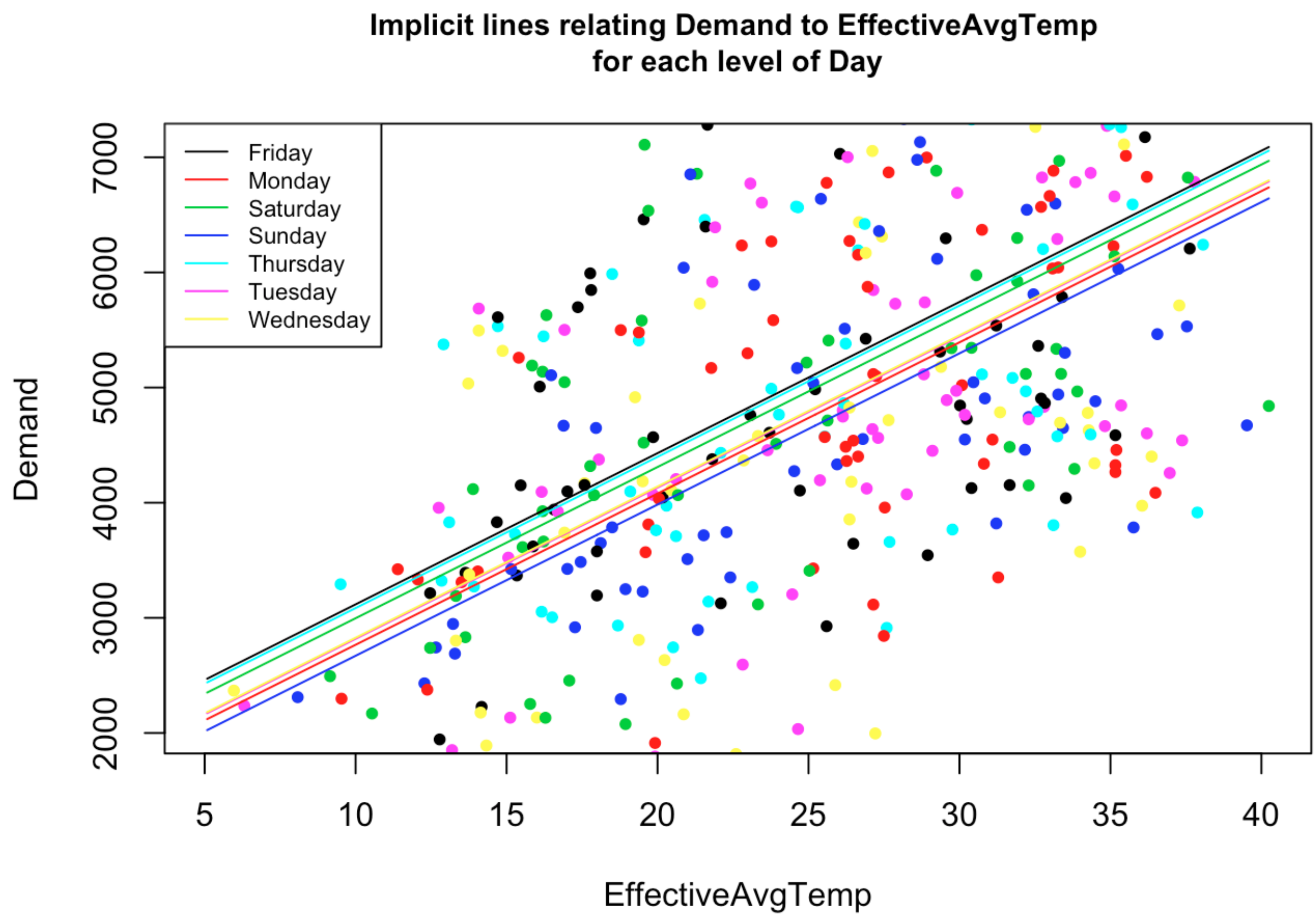
---

d. Demand for bikes may flucuate over the course of a week, so let us consider the variable `Day`. Fit a model predicting `Demand` from `EffectiveAvgTemp` and `Day` (no interactions). Run `summary` and `visualize.model`.

```
#Model predicting Demand from EffectiveAvgTemp and Day (no interaction)
R <- lm(Demand~Day+EffectiveAvgTemp, data=EX7.BIKE)
#summary of model
summary(R)
```

```
##
## Call:
## lm(formula = Demand ~ Day + EffectiveAvgTemp, data = EX7.BIKE)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -3682  -1086   -150   1115   3184
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1803.746    294.211   6.131 2.09e-09 ***
## DayMonday         -350.861    271.574  -1.292   0.1971
## DaySaturday       -120.265    261.049  -0.461   0.6453
## DaySunday         -446.406    265.780  -1.680   0.0938 .
## DayThursday        -32.291    267.956  -0.121   0.9041
## DayTuesday        -299.009    273.060  -1.095   0.2742
## DayWednesday      -290.268    271.583  -1.069   0.2858
## EffectiveAvgTemp  131.329      9.255  14.190  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1443 on 402 degrees of freedom
## Multiple R-squared:  0.3389, Adjusted R-squared:  0.3274
## F-statistic: 29.44 on 7 and 402 DF,  p-value: < 2.2e-16
```

```
#visualize model
visualize.model(R)
```




```
##
## Effect test for Day has p-value 0.5583
```

d1. `Day` **has 7 levels. Behind the scenes, how many indicator variables must be added to the regression to represent** `Day` **and which level is not represented by an indicator variable (i.e., which is the reference)?**

Response: There would be L-1, or 6, levels. The reference variable is the first alphabetically - that is Friday.

---

d2. Interpret the coefficient of `DayThursday` **, which you should have found to be -32.**

Response: On days with the same temperature, the demand will be less on a Thursday by 32.29 than on a Friday.

---

d3. For days with the same effective average temperature, which day of the week has the highest average demand and which day of the week has the lowest average demand.

Response: Friday would have the highest demand. Lowest demand would be on a Sunday.

---

d4. When comparing days with the same effective average temperature, is there any statistically

significant difference between the average demands among the days of the week? Explain.

```
#Code using drop1 that checks the significance of Day
drop1(R, test="F")
```

```
## Single term deletions
##
## Model:
## Demand ~ Day + EffectiveAvgTemp
##                  Df Sum of Sq        RSS    AIC  F value Pr(>F)
## <none>                         836723803 5972.8
## Day               6  10183783  846907586 5965.8   0.8155 0.5583
## EffectiveAvgTemp  1 419110708 1255834511 6137.3 201.3598 <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Response: No, the p-value being 0.5583 and greater than 5%, this difference is not statistically significant.

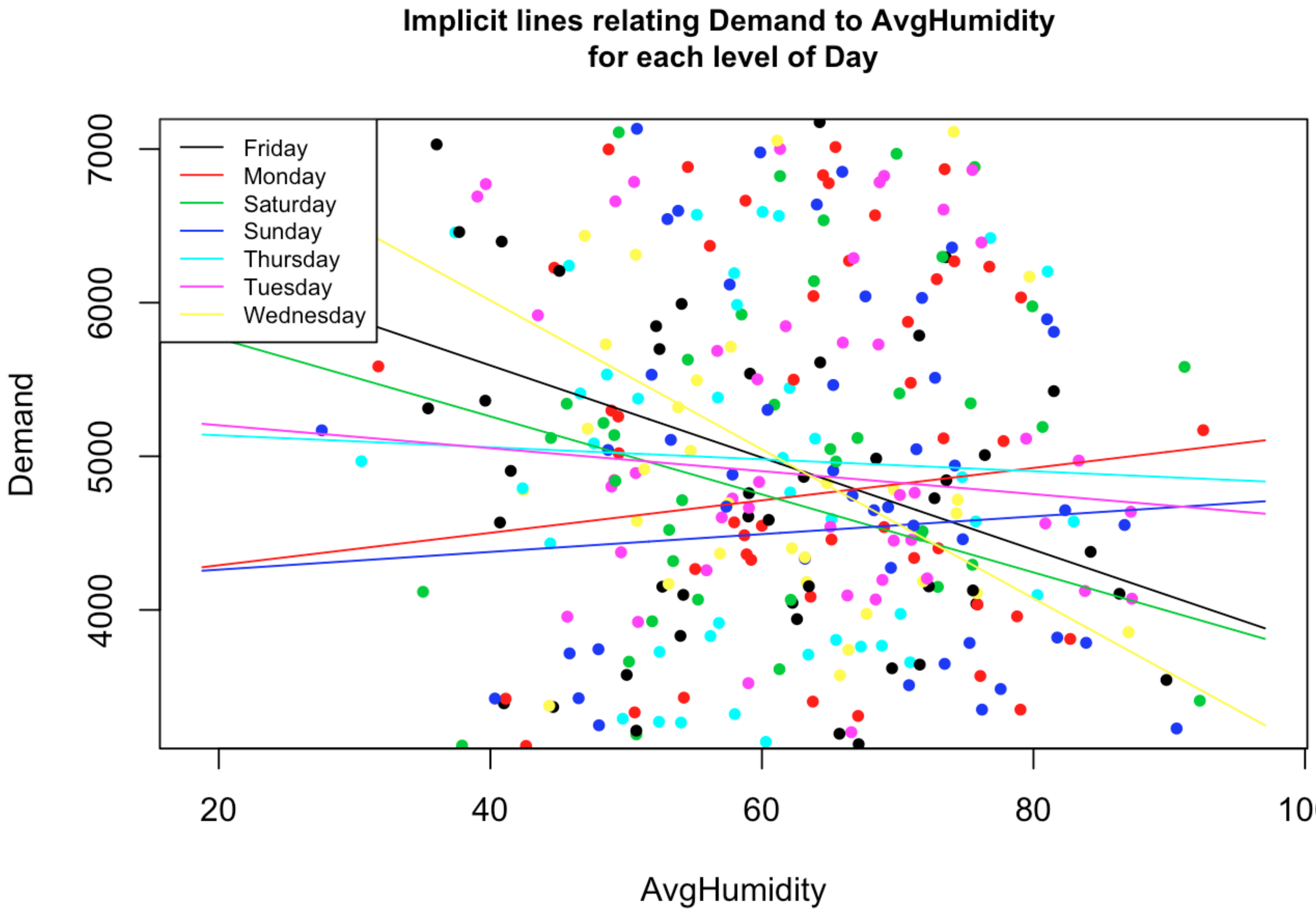
---

e. Maybe the difference in average demands for some days of the week is relatively small for certain values of the humidity and relatively large for other certain other humidities. Fit the model predicting `Demand` from `Day` and `AvgHumidity`, including the interaction. Visualize the model and perform the effect test.

```
#Model predicting Demand from AvgHumidity and Day (with interaction)
S <- lm(Demand~Day*AvgHumidity, data=EX7.BIKE)
#summary of model
summary(S)
```

```
##
## Call:
## lm(formula = Demand ~ Day * AvgHumidity, data = EX7.BIKE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4162.9 -1162.7   -95.5  1304.3  3764.1
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              6791.531   1059.124   6.412 4.08e-10 ***
## DayMonday               -2712.041   1601.124  -1.694   0.0911 .
## DaySaturday              -516.600   1379.252  -0.375   0.7082
## DaySunday               -2644.924   1504.155  -1.758   0.0794 .
## DayThursday             -1576.729   1545.627  -1.020   0.3083
## DayTuesday              -1439.909   1582.876  -0.910   0.3635
## DayWednesday             1162.029   1550.080   0.750   0.4539
## AvgHumidity               -30.000     17.098  -1.755   0.0801 .
## DayMonday:AvgHumidity      40.559     25.332   1.601   0.1102
## DaySaturday:AvgHumidity     4.614     22.092   0.209   0.8347
## DaySunday:AvgHumidity      35.776     23.772   1.505   0.1331
## DayThursday:AvgHumidity    26.098     25.163   1.037   0.3003
## DayTuesday:AvgHumidity     22.527     25.039   0.900   0.3688
## DayWednesday:AvgHumidity  -18.477     24.165  -0.765   0.4450
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1748 on 396 degrees of freedom
## Multiple R-squared:  0.04414,    Adjusted R-squared:  0.01276
## F-statistic: 1.407 on 13 and 396 DF,  p-value: 0.153
```

```
#visualize model
visualize.model(S)
```




```
##
## Effect test for interaction with Day has p-value 0.1692
```

```
#Run drop1
drop1(S, test="F")
```

```
## Single term deletions
##
## Model:
## Demand ~ Day * AvgHumidity
##                 Df Sum of Sq        RSS    AIC F value Pr(>F)
## <none>                       1209757874 6136.0
## Day:AvgHumidity  6  27915194 1237673068 6133.3   1.523 0.1692
```

e1. Name one day of the week where the relationship between `Demand` and `AvgHumidity` is positive and another day of the week where the relationship is negative.

Response:

Positive: Wednesday Negative: Monday

---

e2. Is the variation in average demands between days relatively large or relatively small when the

humidity is about 20 when compared to the variation when the humidity is around 70?

Response: Relatively small. These numbers are very small as compared to the numbers on variation without interaction.

---

e3. Are the differences in the strengths of the relationships between demand and humidity statistically significant? In other words, is the difference in average demands between days of the week larger for some values of humidity and smaller for others, from a statistical point of view?

Response: No. All p-values are greater than 5%

---

f. In this problem, interactions do not play a very important role. Fit a model predicting `Demand` from all variables in the data frame (using the ~. shortcut). When comparing two working-day Mondays with the same temperature, humidity, and windspeed, is there a statistically significant difference in average demands on rainy days vs. non-rainy days? Explain. What about holiday Fridays?

```
#Model predicting Demand from all variables (no interactions)
T <- lm(Demand~., data=EX7.BIKE)
V <- lm(Demand~Day*Workingday, data=EX7.BIKE)
#summary of model
summary(T)
```

```
##
## Call:
## lm(formula = Demand ~ ., data = EX7.BIKE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3037.1  -920.9  -214.7  1106.0  3052.3
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)      4204.351    564.622   7.446 6.06e-13 ***
## DayMonday        -318.499    263.678  -1.208  0.22780
## DaySaturday      -119.357    246.831  -0.484  0.62897
## DaySunday        -417.292    253.104  -1.649  0.10000
## DayThursday       -86.783    271.618  -0.320  0.74951
## DayTuesday       -169.033    265.110  -0.638  0.52410
## DayWednesday     -185.730    261.375  -0.711  0.47776
## Workingdayyes     -29.868    196.042  -0.152  0.87898
## Holidayyes       -384.740    414.309  -0.929  0.35365
## WeatherRain      -148.795    179.687  -0.828  0.40813
## AvgTemp          -368.805    122.887  -3.001  0.00286 **
## EffectiveAvgTemp  469.141    114.228   4.107 4.87e-05 ***
## AvgHumidity       -33.818      6.224  -5.433 9.68e-08 ***
## AvgWindspeed      -63.976     14.675  -4.360 1.66e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1346 on 396 degrees of freedom
## Multiple R-squared:  0.4334, Adjusted R-squared:  0.4148
## F-statistic:  23.3 on 13 and 396 DF,  p-value: < 2.2e-16
```

```
summary(V)
```

```
##
## Call:
## lm(formula = Demand ~ Day * Workingday, data = EX7.BIKE)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4559.1 -1198.7    -3.6  1336.3  3951.8
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 4652.74     366.93  12.680   <2e-16 ***
## DayMonday                  -1223.74    1797.57  -0.681   0.4964
## DaySaturday                 -310.51     610.61  -0.509   0.6114
## DaySunday                   -533.96     691.89  -0.772   0.4407
## DayThursday                  342.20     443.36   0.772   0.4407
## DayTuesday                  -447.74    1797.57  -0.249   0.8034
## DayWednesday                 611.32     472.35   1.294   0.1963
## Workingdayyes                545.55     475.09   1.148   0.2515
## DayMonday:Workingdayyes      794.23    1838.11   0.432   0.6659
## DaySaturday:Workingdayyes    -75.16     722.73  -0.104   0.9172
## DaySunday:Workingdayyes      -84.45     793.31  -0.106   0.9153
## DayThursday:Workingdayyes   -631.61     794.81  -0.795   0.4273
## DayTuesday:Workingdayyes     142.80    1838.39   0.078   0.9381
## DayWednesday:Workingdayyes -1756.90     679.45  -2.586   0.0101 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1760 on 396 degrees of freedom
## Multiple R-squared:  0.03109,    Adjusted R-squared:  -0.0007159
## F-statistic: 0.9775 on 13 and 396 DF,  p-value: 0.4728
```

Response: No, neither working day Mondays nor holiday Fridays hold statistically significant differences. The p-values are too large.

Question 2

In marketing analytics, the relationship between the sales of a product and the amount of money spent on advertising is often studied. Below is the output of a model predicting (thousands of dollars) based on the amount of internet advertising (thousands of dollars), the in which the product would be placed (Kitchen, Bed, Office), and the of the product (Small, Large).

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.24097    0.07193  45.059   <2e-16 ***
Ads          0.96943    0.04335  22.365   <2e-16 ***
RoomKitchen -0.12294    0.05554  -2.213   0.0277 *
RoomOffice  -0.15772    0.08736  -1.806   0.0721 .
SizeSmall    0.01089    0.05593   0.195   0.8457
```

a. Interpret the coefficient of `Ads` , which you make take to be 0.969.

Response: If all other variables were to be identical, for every thousand dollar increase in ad spending, the sales of the product would be expected to increase by $969.

---

b. Interpret the coefficient of `RoomKitchen` , which you make take to be -0.123.

Response: If all other variables were to be identical, one could expect that a kitchen item would sell for $123 less than a bedroom item.

---

c. What is the implicit regression equation relating `Sales` to `Ads` for large products that go in the bedroom?

Response: Sales = 3.24097 + 0.96943(Ads)

---

d. What is the implicit regression equation relating `Sales` to `Ads` for small products that go in the office?

Response: Sales = 3.24097 + 0.96943(Ads) + 0.01089(1) - 0.15772(1)

---

e. Among all 6 type of products (Bed Small, Bed Large, Kitchen Small, Kitchen Large, etc.), which one has the highest average sales (assuming each product has had an equal amount of advertising)?

Response: Bed Small

Question 3

A kaggle.com competition we looked at in HW3 tried to predict the “hazard score” of a property. You can think of the hazard score as a number that represents the condition of the property as determined by the inspection. Some inspection hazards are major and contribute more to the total score, while some are minor and contribute less. The total score for a property is the sum of the individual hazards. Imagine trying to predict the hazard score from the following model:

$$Hazard = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 B + \beta_4 C + \beta_5 D + \beta_5 (x_1 B) + \beta_6 (x_1 C) + \beta_7 (x_1 D)$$

In other words, we are predicting Hazard from two quantitative variables x_1 and x_2 along with a categorical variable x_8 (with levels A, B, C, and D), and the interaction of x_1 and x_8. In the above model, B is an indicator variable that equals 1 if x_8 has level B and 0 otherwise, etc. The following is the relevant output.

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.76    0.260536  10.609  < 2e-16 ***
x1           0.04    0.023769   1.736 0.082631 .
x8B         -0.05    0.261112  -0.175 0.860913
x8C          2.51    0.344397   7.289 3.17e-13 ***
x8D         -0.26    0.317947  -0.807 0.419922
x2           0.07    0.002788  23.421  < 2e-16 ***
x1:x8B      -0.01    0.024031  -0.226 0.821000
x1:x8C       0.11    0.030469   3.736 0.000187 ***
x1:x8D       0.01    0.029260   0.492 0.622422
```

a. What is the implicit regression equation predicting Hazard from x_1 and x_2 when x_8 has level A (the reference level). Your answer should be Hazard = a + b*x1* + *c*x2 for some values of a, b, and c.

Response:

$$Hazard = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

---

b. What is the implicit regression equation predicting Hazard from x_1 and x_2 when x_8 has level D. Your answer should be Hazard = a + b*x1* + *c*x2 for some values of a, b, and c.

Response:

$$Hazard = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_5 (1) + \beta_7 (x_1 * 1)$$