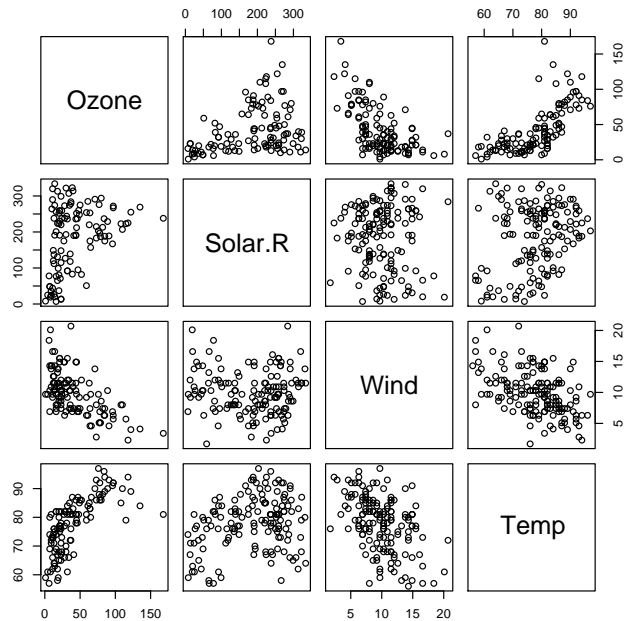
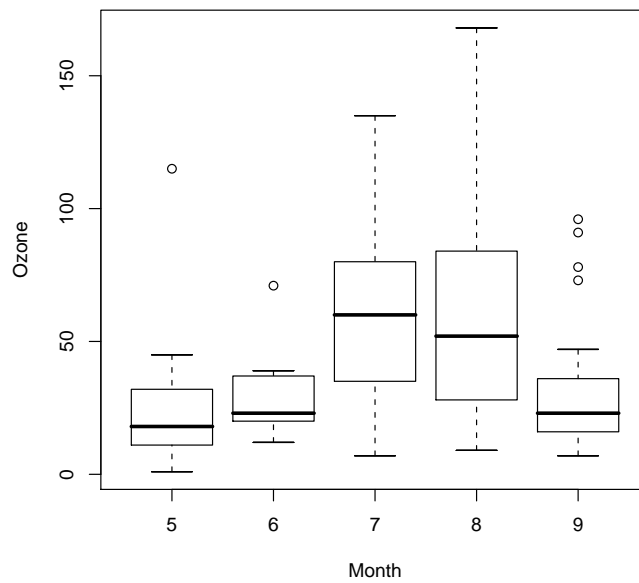


## Chapter 1, Problem Set 1

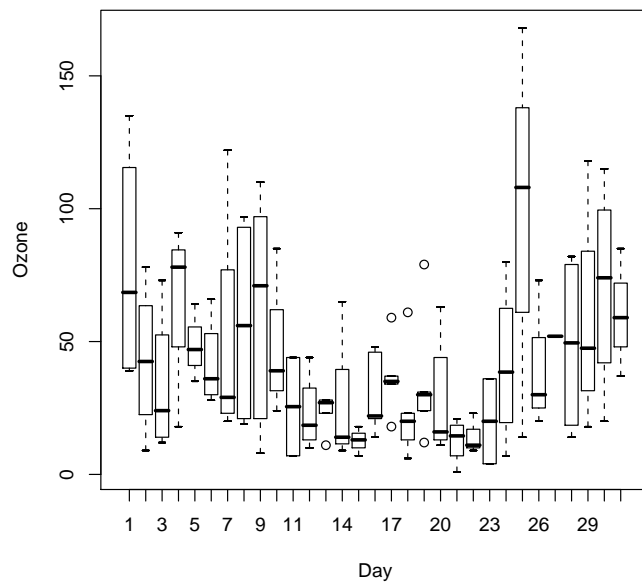
- Ozone & Solar.R: There is a positive relationship between Ozone and Solar.R. The variation increases as Ozone & Solar.R increases.  
Ozone & Wind: There is a negative relationship between Ozone and Wind. The relationship may be nonlinear. The relationship seems to flatten toward the lower right side of the plot.  
Ozone & Temp: There is a positive relationship between Ozone and Temp. The relationship may be nonlinear but the points on the upper right are pretty sparse. It is hard to tell.  
Solar.R & Wind: There seems to be little noticeable relationship between these two variables.  
Solar.R & Temp: There appears to be a positive nonlinear relationship between Solar.R & Temp, though there is a great amount of variation in this relationship. The “hole” in the middle of the plot is curious.  
Wind & Temp: There is a negative, roughly linear, relationship between Wind & Temp.



2. The distribution of Ozone varies across months. The distributions of ozone are similar in May, June, & September, and in July & August. In May, June, and September, the distributions of ozone are relatively tighter, and the medians are 18, 23, and 23, respectively. Conversely, the median ozone levels in July and August are 60 and 52, respectively. On the average, ozone concentrations are higher in July and August and the fluctuations are greater too.

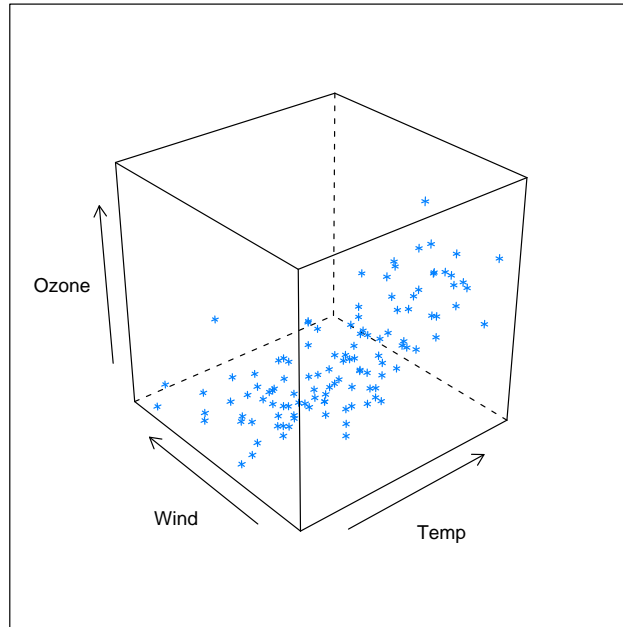


Ozone varies across day as well. In general, the median and spreads are largest towards the beginning and end of each month. Conversely, the median and spread of ozone are relatively small in the middle of the months (i.e., between days 12 and 21). This is a curious pattern. It is difficult to relate the pattern to what was learned from the scatterplot matrix.

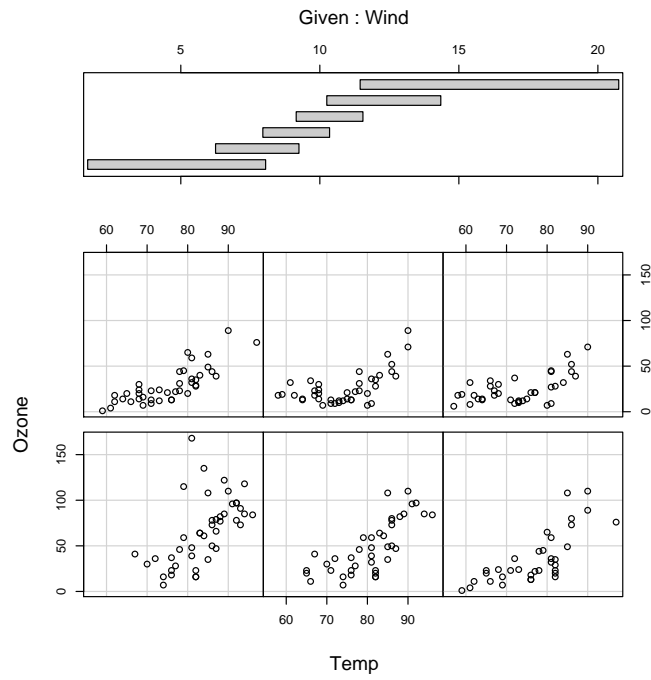


3. Month and day of the month are categorical variables. To make a scatter plot of ozone versus month or ozone versus day of the month, one would have to assume that an equal interval, quantitative value can be sensibly assigned to each categorical value.

4. From the cloud plot, it is difficult to make out any patterns from this plot of Ozone against Wind and Temp.



5. Wind speed is held constant by examining the relationship of ozone versus temperature for various subsets of the data according to slices of wind speed. For each of the six ranges of wind speed, a scatter plot of ozone versus temperature is reported.



6. On average, ozone tends to increase as temperature increases with wind speed held constant.

7. For relatively faster wind speeds (i.e., greater than roughly 10 miles per hour), ozone levels tends to increase less rapidly as temperature increases. A smaller percentage of ozone levels are between 50 and 100 ppb in the top three scatter plots. This suggests an interaction effect between wind speed and temperature.
8. From the cross-tabulation of the number of missing ozone values versus month, we see that data are missing for 21 of the 30 days in June. No other month is missing more than 5 ozone values. In the month of September, only one ozone recording is missing. If temperature and/or wind exhibit different relationships with ozone in the month of June, then the conditioning plot may not fully reveal the conditional distribution of ozone on temperature, holding wind constant.

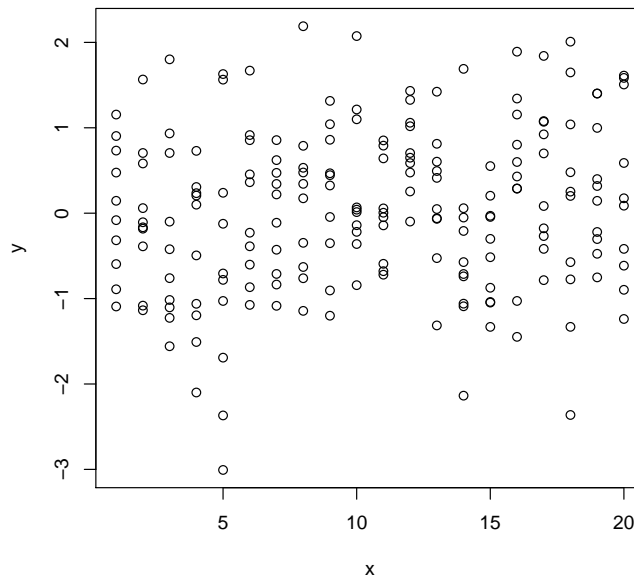
Month	# of Missing Ozone Values
5	5
6	21
7	5
8	5
9	1

9.  $Ozone_i = \beta_0 + \beta_1(Wind_i) + \beta_2(Temp_i) + \beta_3(Wind_i \cdot Temp_i) + \epsilon_i$   
 $i = 1, \dots, N, \epsilon \sim N(0, \sigma^2)$  One assumes that the disturbances are normally distributed with an expectation of 0 and variance  $\sigma^2$ , and are independent of each other and of wind speed and temperature. Also, ozone is a linear function of wind speed and temperature, both in their original units. Each of these assumptions is difficult to justify. For example, with ozone bounded at 0.0, the normality assumption is likely to be wrong and the constant variance assumption highly suspect. The earlier plots suggest as much. Insofar as ozone can remain in the atmosphere, the disturbances are likely to be related from one day to the next. Are wind and solar radiation the only relevant variables? If not, the disturbances are likely to be related to the predictors. For instance, do the data properly capture temperature inversions, which trap ozone near the earth surface? Probably not.
10.  $\widehat{Ozone}_i = -248.52 + 4.076(Temp_i) + 14.34(Wind_i) - 0.22(Wind_i \cdot Temp_i)$   
 The linear model suggests that, *ceteris paribus*, ozone increases with respect to both temperature and wind speed. There is also an interaction effect in which the impact of temperature is reduced with higher wind speeds. This is pretty much what was seen in the conditioning plots. Given all of the problems with the model, little has been learned beyond what was apparent from the conditioning plots.

## Chapter 1, Problem Set 2

1. 

```
> x=rep(1:20, times=10)
> y=rnorm(200)
```
2. The scatter plot of  $y$  against  $x$  shows no apparent relationship between the two variables. The values of  $y$  appear to be centered roughly around  $y=0$ , and the variance in  $y$  across different values of  $x$  is fairly constant.



3. From the `lm()` output, we confirm the virtually nonexistent correlation between  $y$  and  $x$ , as evidenced by the small estimated coefficient (0.001602) and  $R^2$  value (0.00008159). The p-value of the  $x$  coefficient is also well above 0.10. From the `glm()` output, we obtain the same coefficients, standard errors, t-statistics, and p-values. Rather than obtain the value of  $R^2$ , we are provided with the null and residual deviance, and  $AIC$  statistic. The null and residual deviance are close in magnitude, suggesting that conditioning  $y$  on  $x$  contributes very little. Also in both sets of output, summary statistics about the residuals are reported. In this model, the residuals exhibit a median of 0.027 and range from -3.048 to 3.432.
4. 

```
xy.lm.f = lm(y~as.factor(x))
summary(xy.lm.f)
xy.glm.f = glm(y~as.factor(x), family=gaussian)
```

`summary(xy.glm.f)`

5. When  $x$  is treated as a factor, both `lm()` and `glm()` compute the average offset relative to the value of  $x$  captured by the intercept in this case,  $x = 1$ . This is essentially a one-way Analysis of Variance (ANOVA). In comparison to the linear model of  $y$  on  $x$ , the unadjusted fit is much improved. In (3), the residual deviance was 193.13; now, it is 176.77 (the null deviance does not change). However, the *AIC* value is higher in the latter model (566.58 versus 584.88), indicating the apparent improvement is an artifact of model complexity. And although the adjusted  $R^2$  was not high in (3), now it is less than 0. In this situation, the absorption of degrees of freedom helps for the adjusted  $R^2$  and *AIC*. One would conclude properly that there is no systematic relationship. But the other output might well suggest incorrectly that there is a systematic relationship.

### Chapter 1, Problem Set 3

1. Because all of the covariates are vectors of random values from a standard normal distribution, we expect to find the following: a) on average, the coefficients should be around 0, b) by chance, a few t-statistics will be large (say, around 2), c) the F-test may suggest that all of the coefficients are not different from 0, d) the  $R^2$  will be close to 0.50 because there is close to one coefficient for every 2 observations, but e) this high ratio of coefficients to observations will result in a low adjusted  $R^2$ , thereby suggesting an overly complex model in a context when there should be nothing systematic in the relationship between the predictors and the response.
2. (Details will vary) The coefficient for  $X_{28}$  is significant at the 0.1 level, and the coefficients for  $X_2$  and  $X_{30}$  are statistically significant at the 0.05 level. But with 49 covariates, one would expect roughly  $0.05 \times 49 = 2.45$ , or between 2 and 3 coefficients, to be statistically significant when the null hypothesis is true. The  $R^2$  is 0.46, but the F-test suggests that the null hypothesis should not be rejected. The average coefficient value is -0.011, which is very close to 0.
3. On average, the coefficients are close to 0, which is consistent with the way the data were generated. There is not much evidence of a systematic relationship between the predictors and the response.
4. (Details will vary) The final stepwise model, which includes 14 predictors, is less complex than the full model with 49 predictors. The multiple  $R^2$  is roughly 0.372, which is 9 percent less in the stepwise model, but the adjusted R-squared is higher than in the full model (-0.070 versus 0.259). All of the p-values are less than 0.20, and all but three p-values are less than 0.10. The F-tests null hypothesis is rejected (p-value  $< 0.001$ ). In contradiction to initial expectations, there now seems to some systematic relationship between the predictors and the response.
5. The model suffering from overfitting. The available statistics do not, by and large, compensate for the data snooping implicit in the stepwise algorithm.
6. With more observations per predictor, the stepwise algorithm is less able to capitalize on chance patterns in the data. There will be less overfitting. Then, conventional regression results and stepwise regression results will be more alike.
7. The dangers of data snooping can materialize when it is a fitting algorithm that is doing the snooping. Serious overfitting can result. One may falsely conclude that there are systematic relationships between the predictors and the response. More generally, model selection procedures can lead to overfitting.