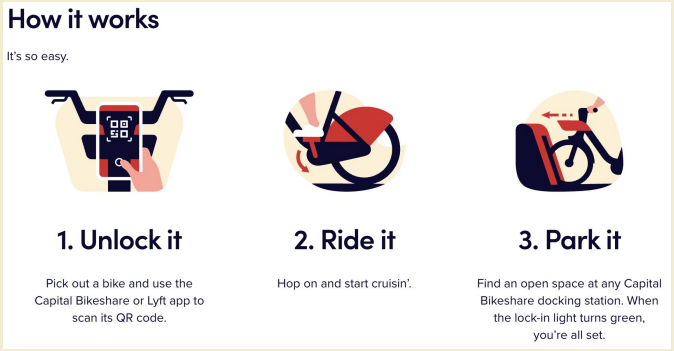
Uncovering the Stats Behind the Spokes

Uma Datar, Sarah Lozina,
Amelia Klaus

Introduction

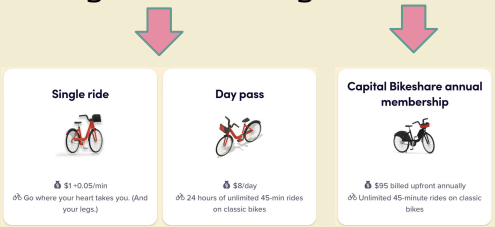
- 2 year historical log of daily data from the Capital Bikeshare system in Washington D.C.

- We want to analyze what factors contribute to the most bikes being rented out



The Variables

- Our Y variable is the **count of total rental bikes**, including casual and registered users
- Our X (independent) variables are:
  - Dteday (**Date**)
  - **Season** (1: winter, 2: spring, 3: summer, 4: fall)
  - **Year** (0: 2011, 1: 2012)
  - **Holiday** (0: not a holiday, 1: holiday)
  - **Weekday** (0: Sunday, 1: Monday, 2: Tuesday, 3: Wednesday, 4: Thursday, 5: Friday, 6: Saturday)
  - **Workingday** (0: otherwise, 1: weekend or holiday)
  - **Weathersit** (1: Clear, few clouds, partly cloudy, 2: mist and cloudy, 3: light snow or rain, 4: heavy rain or ice or thunderstorm or snow and fog)
  - **Temp** (Normalized temperature in Celsius)
  - **Atemp** (Normalized feels like temperature in Celsius)
  - **Hum** (Normalized humidity)
  - **Windspeed** (Normalized wind speed)



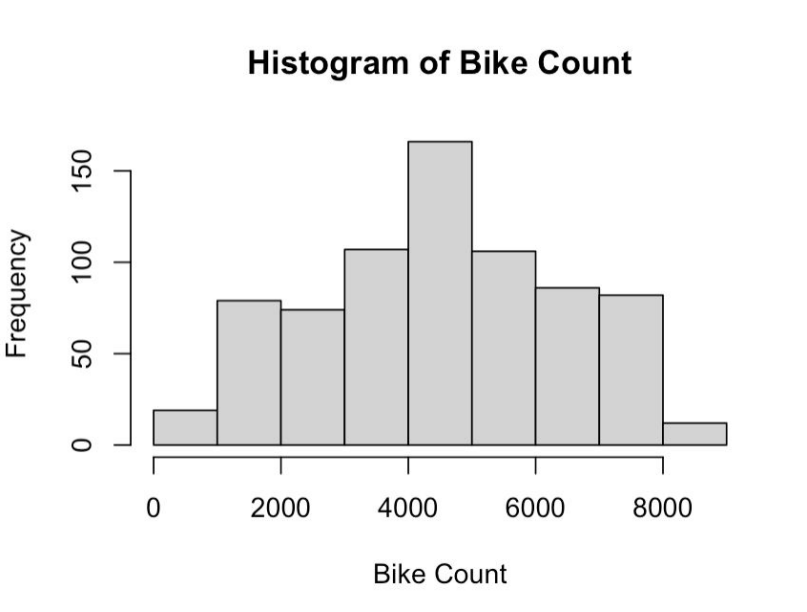
Collect Sample Data

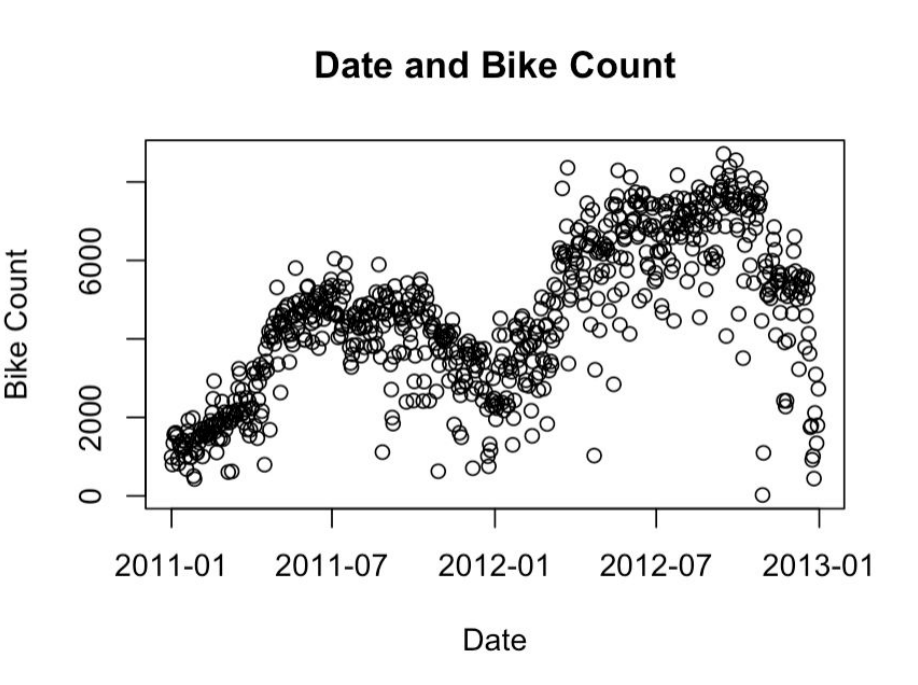
instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
1	2011-01-01	1	0	1	0	6	0	2	0.3441670	0.3636250	0.805833	0.1604460	331	654	985
2	2011-01-02	1	0	1	0	0	0	2	0.3634780	0.3537390	0.696087	0.2485390	131	670	801
3	2011-01-03	1	0	1	0	1	1	1	0.1963640	0.1894050	0.437273	0.2483090	120	1229	1349
4	2011-01-04	1	0	1	0	2	1	1	0.2000000	0.2121220	0.590435	0.1602960	108	1454	1562
5	2011-01-05	1	0	1	0	3	1	1	0.2269570	0.2292700	0.436957	0.1869000	82	1518	1600
6	2011-01-06	1	0	1	0	4	1	1	0.2043480	0.2332090	0.518261	0.0895652	88	1518	1606
7	2011-01-07	1	0	1	0	5	1	2	0.1965220	0.2088390	0.498696	0.1687260	148	1362	1510
8	2011-01-08	1	0	1	0	6	0	2	0.1650000	0.1622540	0.535833	0.2668040	68	891	959
9	2011-01-09	1	0	1	0	0	0	1	0.1383330	0.1161750	0.434167	0.3619500	54	768	822
10	2011-01-10	1	0	1	0	1	1	1	0.1508330	0.1508880	0.482917	0.2232670	41	1280	1321
11	2011-01-11	1	0	1	0	2	1	2	0.1690910	0.1914640	0.686364	0.1221320	43	1220	1263
12	2011-01-12	1	0	1	0	3	1	1	0.1727270	0.1604730	0.599545	0.3046270	25	1137	1162
13	2011-01-13	1	0	1	0	4	1	1	0.1650000	0.1508830	0.470417	0.3010000	38	1368	1406
14	2011-01-14	1	0	1	0	5	1	1	0.1608700	0.1884130	0.537826	0.1265480	54	1367	1421
15	2011-01-15	1	0	1	0	6	0	2	0.2333330	0.2481120	0.498750	0.1579630	222	1026	1248

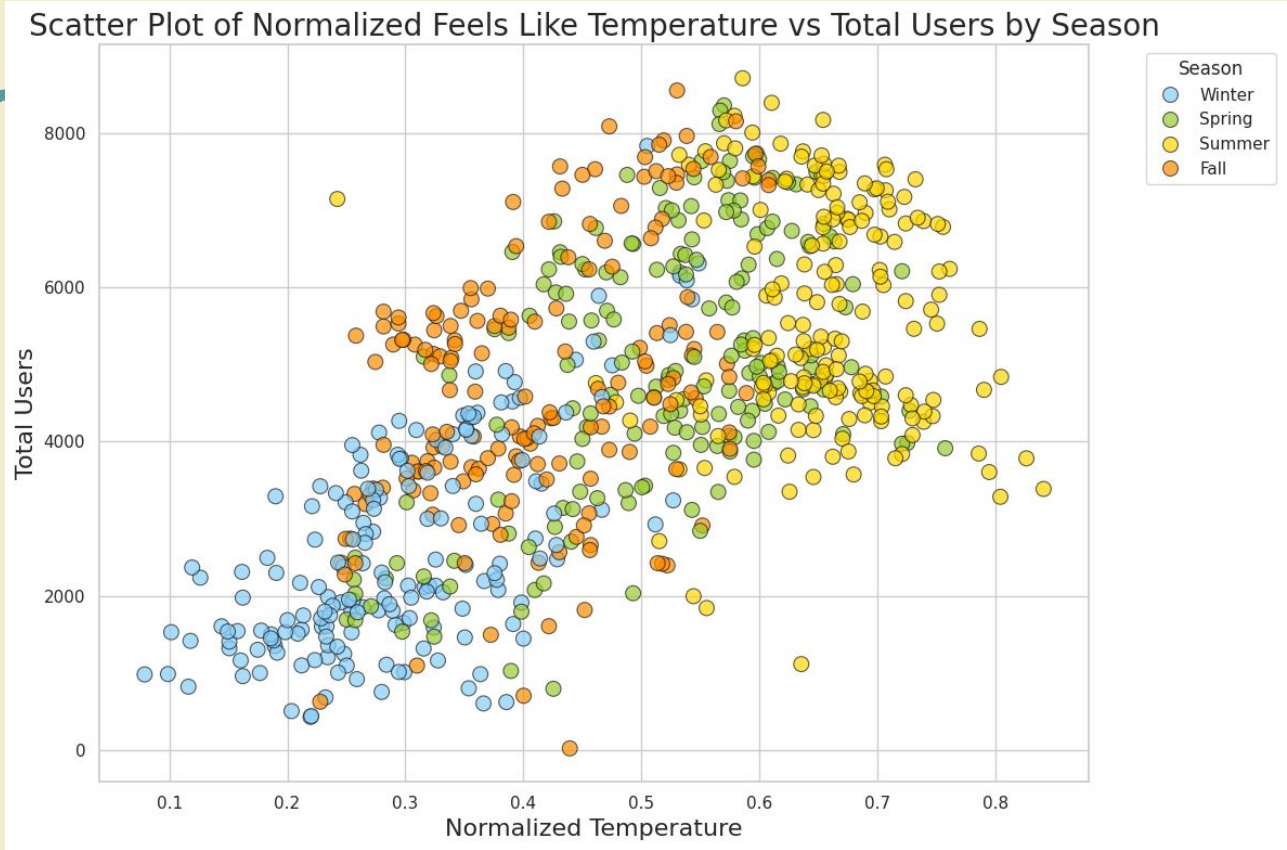
For example: On Saturday January 1st, 2011 it was misty and cloudy with 985 total users!

Preliminary Analysis

```
> range(day$cnt)
[1]    22 8714
> mean(day$cnt)
[1] 4504.349
```






Scatter Plot of Normalized Feels Like Temperature vs Total Users by Season
Season
Winter
Spring
Summer
Fall
Total Users
0
2000
4000
6000
8000
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
Normalized Temperature

First Order Linear Model

Model 1: Full first order model with all terms included.

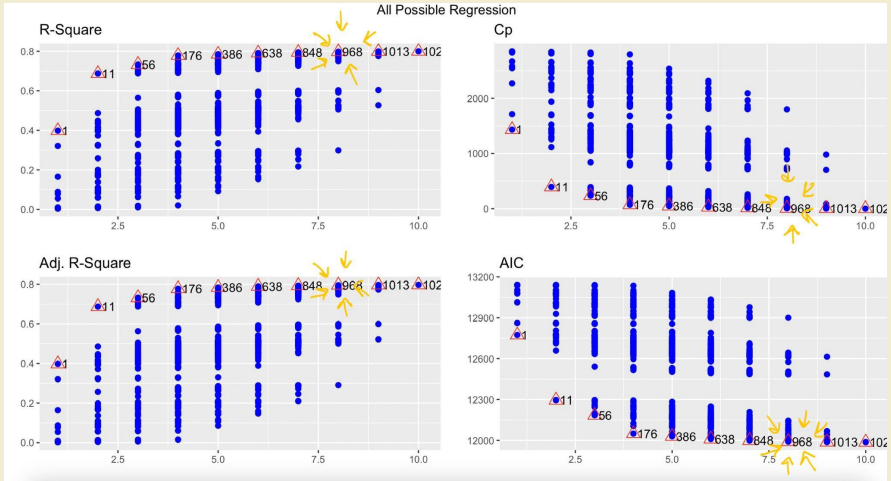
Model 2: First order model with multicollinearity resolved (temp removed)

Model 3: First order model from all-possible-regression selection (removed month and working day).

Model 1
Global F-Test
Adj R^2 = 0.7972
Standard Error = 872.4

```
> summary(bikemodel1)

Call:
lm(formula = cnt ~ season + yr + mnth + holiday + weekday + workingday +
    weathersit + temp + atemp + hum + windspeed, data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-4143.6  -442.3    49.5   546.2  2946.4

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1469.00     240.22   6.115 1.58e-09 ***
season        509.78      54.76   9.310  < 2e-16 ***
yr           2040.70      65.19  31.306  < 2e-16 ***
mnth          -38.98      17.08  -2.282  0.02276 *
holiday      -518.99     201.04  -2.582  0.01003 *
weekday        69.06      16.30   4.237 2.56e-05 ***
workingday    120.36      72.01   1.671  0.09507 .
weathersit   -610.99      78.36  -7.797 2.23e-14 ***
temp         2028.92    1403.67   1.445  0.14877
atemp        3573.27    1589.39   2.248  0.02487 *
hum         -1018.86     314.00  -3.245  0.00123 **
windspeed   -2557.57     456.28  -5.605 2.96e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 872.4 on 719 degrees of freedom
Multiple R-squared:  0.8002,    Adjusted R-squared:  0.7972
F-statistic: 261.9 on 11 and 719 DF,  p-value: < 2.2e-16
```

```
> vif(bikemodel1)
    season         yr       mnth    holiday    weekday workingday weathersit       temp      atemp        hum  windspeed
  3.548413   1.020253   3.333672   1.083126   1.024076   1.076392   1.748741  63.321299  64.343361   1.918309   1.199259
```



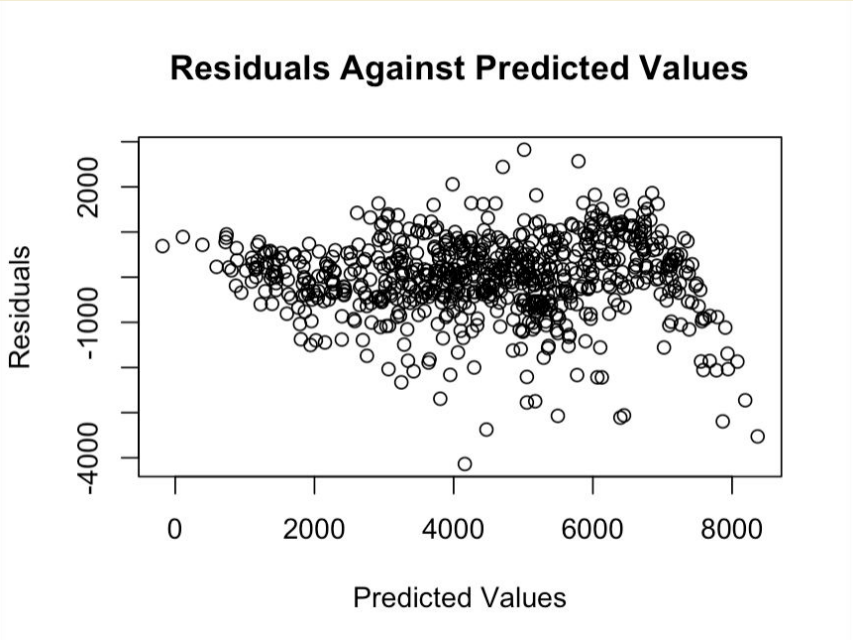
season yr holiday weekday weathersit atemp hum windspeed

1 2 3 4 5 6 7 8

Final First Order Linear Model

$\hat{Y}$ = 1467.57 + 407.23(season) + 2038.35(yr) – 614.59(holiday) + 69.21(weekday) – 592.12(weathersit) + 5931.29(atemp) – 1132.22(hum) – 2449.82(windspeed)

```
Call:
lm(formula = cnt ~ season + yr + holiday + weekday + weathersit +
    atemp + hum + windspeed, data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-4138.0  -425.5    73.3   536.4  2823.2

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1467.57     228.77   6.415 2.55e-10 ***
season        407.23      31.91  12.762  < 2e-16 ***
yr           2038.35      65.50  31.121  < 2e-16 ***
holiday      -614.59     195.40  -3.145 0.001728 **
weekday        69.21      16.35   4.234 2.60e-05 ***
weathersit   -592.12      78.49  -7.544 1.37e-13 ***
atemp        5931.29     219.28  27.048  < 2e-16 ***
hum         -1132.22     313.36  -3.613 0.000323 ***
windspeed   -2449.82     452.25  -5.417 8.27e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 876.8 on 722 degrees of freedom
Multiple R-squared:  0.7974,    Adjusted R-squared:  0.7952
F-statistic: 355.2 on 8 and 722 DF,  p-value: < 2.2e-16
```



Second Order Linear Model

```
> summary(bikemodel6)

Call:
lm(formula = cnt ~ season + yr + holiday + weekday + weathersit +
    atemp + hum + windspeed + hum * atemp, data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-4124.1  -442.0    76.7   547.3  2804.9

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1974.16     480.20   4.111 4.39e-05 ***
season        407.93      31.91  12.785  < 2e-16 ***
yr           2041.19      65.52  31.154  < 2e-16 ***
holiday      -625.51     195.56  -3.199  0.00144 **
weekday        70.24      16.36   4.292 2.01e-05 ***
weathersit   -588.48      78.52  -7.494 1.96e-13 ***
atemp        4782.13     982.57   4.867 1.39e-06 ***
hum         -1997.51     786.30  -2.540  0.01128 *
windspeed   -2491.53     453.45  -5.495 5.43e-08 ***
atemp:hum    1922.16    1602.09   1.200  0.23062
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 876.5 on 721 degrees of freedom
Multiple R-squared:  0.7978,    Adjusted R-squared:  0.7953
F-statistic: 316.1 on 9 and 721 DF,  p-value: < 2.2e-16
```

Playing around with second order terms

Model 4: Adding only curvilinear terms with no interactions.

Model 5: Model with curvilinear terms and one interaction.

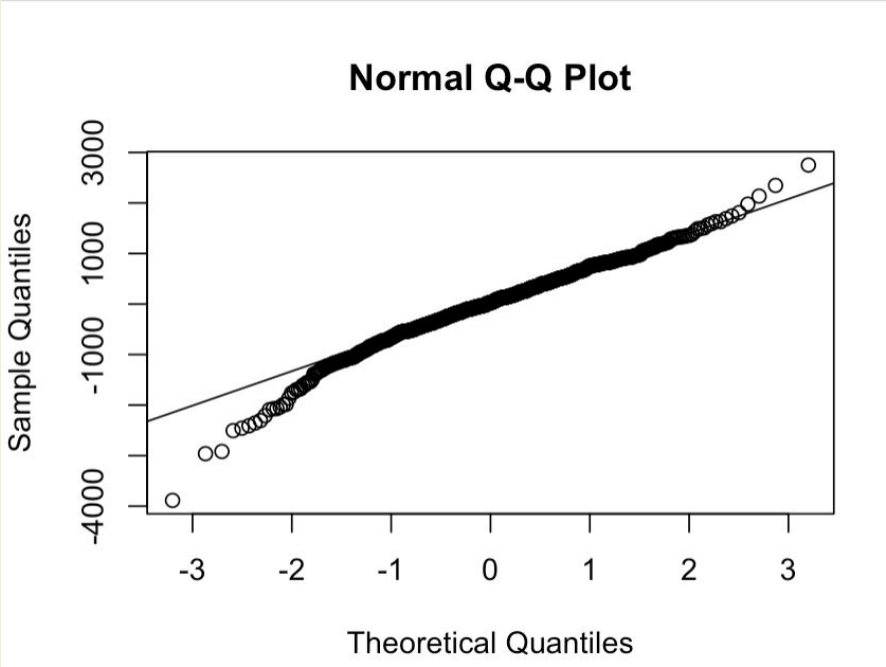
Model 6: Created model with just an interaction term.

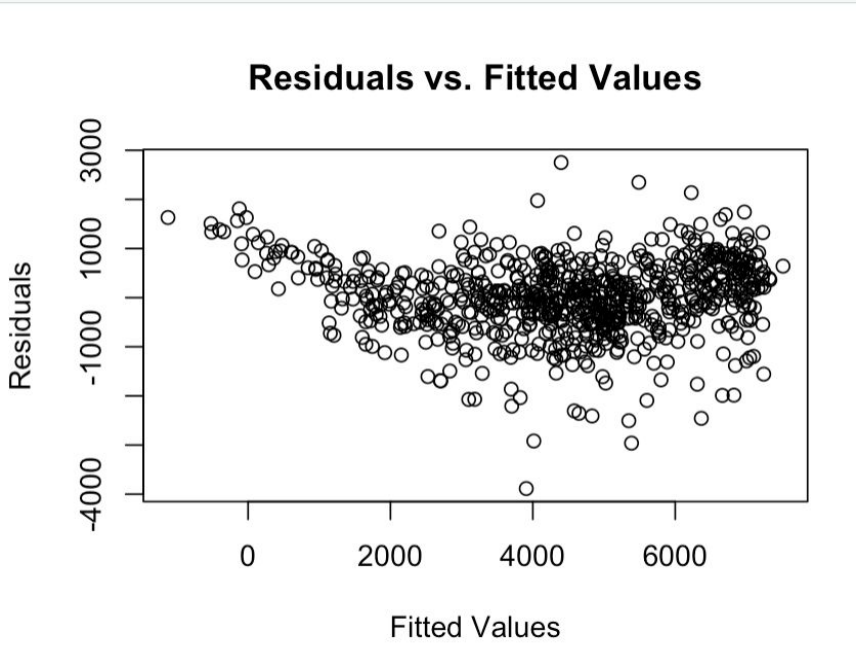
Final Second Order Linear Model

$\hat{Y}$ = -3302.5 + 328.34(season) + 1925.87(yr) – 619.45(holiday) + 67.14(weekday) – 447.46(weathersit) + 20445.12(atemp) – 16785.60(atemp^2) + 4241.07(hum) – 6462.92(hum^2) + 3968.87(windspeed) – 9531.22(windspeed^2) + 3861.96(atemp*hum) – 5815.16(atemp*windspeed)

```
> summary(bikemodel7)

Call:
lm(formula = cnt ~ season + yr + holiday + weekday + weathersit +
    atemp + I(atemp^2) + hum + I(hum^2) + windspeed + I(windspeed^2) +
    hum * atemp + atemp * windspeed, data = day)

Residuals:
    Min      1Q  Median      3Q     Max
-3886.1  -425.2    27.3   494.4  2750.5

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      -3302.57     722.32  -4.572 5.69e-06 ***
season             328.34      28.69  11.445  < 2e-16 ***
yr                1925.87      58.23  33.073  < 2e-16 ***
holiday           -619.45     172.04  -3.601 0.000339 ***
weekday             67.14      14.37   4.672 3.57e-06 ***
weathersit        -447.46      73.85  -6.059 2.21e-09 ***
atemp            20445.12    1621.76  12.607  < 2e-16 ***
I(atemp^2)      -16785.60    1193.16 -14.068  < 2e-16 ***
hum               4241.07    1474.33   2.877 0.004139 **
I(hum^2)         -6462.92    1134.53  -5.697 1.78e-08 ***
windspeed         3968.87    2207.10   1.798 0.072562 .
I(windspeed^2)   -9531.22    3571.38  -2.669 0.007785 **
atemp:hum         3861.96    1438.46   2.685 0.007425 **
atemp:windspeed  -5816.16    2708.82  -2.147 0.032119 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 769.2 on 717 degrees of freedom
Multiple R-squared:  0.8451,    Adjusted R-squared:  0.8423
F-statistic:   301 on 13 and 717 DF,  p-value: < 2.2e-16
```

Residual Analysis





Nested F-Test

```
Analysis of Variance Table

Model 1: cnt ~ season + yr + holiday + weekday + weathersit + atemp +
    I(atemp^2) + hum + I(hum^2) + windspeed + I(windspeed^2) +
    hum * atemp + atemp * windspeed
Model 2: cnt ~ season + yr + holiday + weekday + weathersit + atemp +
    hum + windspeed
  Res.Df       RSS Df  Sum of Sq      F    Pr(>F)
1    717 424252905
2    722 555025689 -5 -130772784 44.202 < 2.2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Complete Model:** Second order model with curvilinear and interaction terms.
- **Reduced Model:** Final first order model

H_0: $\beta_9 = \beta_{10} = \beta_{11} = \beta_{12} = \beta_{13} = 0$ (the true value of all coefficients is equal to 0)

H_A: at least one β_i is nonzero i = 9, 10, 11, 12, 13

P-value < 2.2×10^{-16}

Reject H_0 and conclude at least one coefficient is nonzero, so we choose the complete model

Predictions

Prediction: More users on Uma (September 10th) and Amelia's (August 3rd) birthdays than Sarah's (December 19th) in 2012

- Assigned arbitrary values for the weather on those dates
- Created a dataframe for each birthday

```
> predict(bikemodel7, SarahBday, interval = "prediction")
       fit      lwr      upr
1 4261.296 2726.861 5795.732
> predict(bikemodel7, UmaBday, interval = "prediction")
       fit      lwr      upr
1 5926.128 4387.098 7465.157
> predict(bikemodel7, AmeliaBday, interval = "prediction")
       fit      lwr      upr
1 6381.848 4840.852 7922.844
```

Conclusions + Limitations

Conclusions

- The second order model performed better than the first order model
- This model is useful for predicting the number of rented out bikes for a given day

Limitations + Future Directions

- Lot of variables, did not create a complete second order model
- Residual shape could be improved, but pretty homoscedastic

Coefficient of Variation:

```
> (100 * 769.2)/mean(day$cnt)
[1] 17.07683
```

This is not < 10%, therefore the prediction intervals for count of users generated by the model may be deemed too wide to be of practical use. This model did have the smallest s however.

THANK YOU