# Multiple Linear Regression and Inferences on Regression

Adam J Sullivan, PhD

1/29/2018

# Multiple Regression

# Multiple Regression

- We have been discussing simple models so far.
- This works well when you have:
  - Randomized Data to test between specific groups (Treatment vs Control)
- In most situations we need look at more than just one relationship.
- Think of this as needing more information to tell the entire story.

# Motivating Example

- Health disparities are very real and exist across individuals and populations.
- Before developing methods of remedying these disparities we need to be able to identify where there are disparities.In this homework we will consider a study by (Asch & Armstrong, 2007).
- This paper considers 222 patients with localized prostate cancer.

# Motivating Example

- The table below partitions patients by race, hospital and whether or not the patient received a prostatectomy.

| | Race | Prostatectomy | No Prostatectomy |
|---|---|---|---|
| **University Hospital** | White | 54 | 37 |
| | Black | 7 | 5 |
| **VA Hospital** | White | 11 | 29 |
| | Black | 22 | 57 |

# Loading the Data

You can load this data into R with the code below:

```
phil_disp <- read.table("https://drive.google.com/uc?export=download&id=0B8CsRLdwqzbzOXlIRl9VcjNJRFU", h
```

# The Data

This dataset contains the following variables:

| Variable | Description |
|---|---|
| **hospital** | 0 - University Hospital |
| | 1 - VA Hospital |
| **race** | 0 - White |
| | 1 - Black |
| **surgery** | 0 - No prostatectomy |
| | 1 - Had Prostatectomy |
| **number** | Count of people in Category |

# Consider Prostatectomy by Race

```
library(broom)
prost_race <- glm(surgery ~ race, weight=number, data= phil_disp,
                  family="binomial")
tidy(prost_race, exponentiate=T, conf.int=T)[,-c(3:4)]
```

```
##          term  estimate     p.value  conf.low conf.high
## 1 (Intercept) 0.9848485 0.930377767 0.6985457 1.3880778
## 2        race 0.4749380 0.008953745 0.2694239 0.8250258
```

# Consider Prostatectomy by Race

- What can we conclude?
- What kind of policy might we want to invoke based on this discovery?

# Consider Prostatectomy by Hospital

```
prost_hosp <- glm(surgery ~ hospital, weight=number, data= phil_disp,
                  family="binomial")
tidy(prost_hosp, exponentiate =T, conf.int=T)[,-c(3:4)]
```

```
##          term  estimate      p.value  conf.low conf.high
## 1 (Intercept) 1.4523810 6.270112e-02 0.9838382 2.1646297
## 2    hospital 0.2642013 3.409565e-06 0.1492365 0.4598822
```

# Consider Prostatectomy by Hospital

- What can we conclude?

# Multiple Regression of Prostatectomy

```
prost <- glm(surgery ~ hospital + race, weight=number, data= phil_disp,
             family="binomial")
tidy(prost, exponentiate=T, conf.int=T)[,-c(3:4)]
```

```
##          term  estimate   p.value  conf.low conf.high
## 1 (Intercept) 1.4526892 0.0681969 0.9758192 2.1830747
## 2    hospital 0.2644648 0.0001241 0.1313651 0.5145046
## 3        race 0.9981802 0.9959191 0.5006556 2.0381436
```

# Multiple Regression of Prostatectomy

- What can We conclude?
- What happened here?
- Does this change our policy suggestion from before?

# Benefits of Multiple Regression

- Multiple Regression helps us tell a more complete story.
- Multiple regression controls for confounding.

# Confounding

- Associated with both the Exposure and the Outcome
- Even if the Exposure and Outcome are not related, unmeasured confounding can show that they are.

# What Do We Do with Confounding?

- We must add all confounders into our model.
- Without adjusting for confounders are results may be highly biased.
- Without adjusting for confounding we may make incorrect policies that do not fix the problem.

# Multiple Linear Regression with appearances

- First start with univariate models
- Then perform the multiple model

# Multivariate Models

```
library(broom)
library(fivethirtyeight)
mod3 <- lm(appearances~publisher + year, data=comic_characters)
tidy3 <- tidy(mod3, conf.int=T)[,-c(3:4)]
tidy3
```

```
##              term    estimate      p.value     conf.low    conf.high
## 1     (Intercept) 1265.202320 9.811075e-78 1132.8767591 1397.5278806
## 2 publisherMarvel   -9.539045 1.242355e-11  -12.2971767   -6.7809141
## 3            year   -0.623927 5.927831e-75   -0.6904228   -0.5574312
```

# Interpreting Multiple Coefficients

- The intercept is when all coefficients are zero.
- Each other coefficient is interpreted in context to another.

# Interpreting Multiple Coefficients: Our Example

- Intercept: DC average appearances at year 0.
- Publisher Coefficient: If we consider 2 characters in the same year, the character from Marvel will have on average 9.54 less appearances than the character from DC.
- Year Coefficient: If we consider 2 characters from the same publisher, an increase in 1 year will lead to on average 0.62 less appearances.

# Further Example: Bike Sharing Data

- We have hourly data spanning 2 years
- This dataset has the first 19 days of each month.
- Goal is to find the total count of bike rented.

# Further Example: Bike Sharing Data

| Data | Fields |
|---|---|
| **datetime** | hourly date + timestamp |
| **season** | 1 = spring, 2 = summer, 3 = fall, 4 = winter |
| **holiday** | whether the day is considered a holiday |
| **workingday** | whether the day is neither a weekend nor holiday |

# Further Example: Bike Sharing Data

| Data | Fields |
| --- | --- |
| **weather** | 1: Clear, Few clouds, Partly cloudy, Partly cloudy |
| | 2: Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist |
| | 3: Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds |
| | 4: Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog |
| **temp** | temperature in Celsius |

# Further Example: Bike Sharing Data

| Data | Fields |
|---|---|
| **atemp** | "feels like" temperature in Celsius |
| **humidity** | relative humidity |
| **windspeed** | wind speed |
| **casual** | number of non-registered user rentals initiated |
| **registered** | number of registered user rentals initiated |
| **count** | number of total rentals |

# Univariate Regressions

```
library(readr)
library(tidyverse)
bikes <- read_csv("bike_sharing.csv") %>%
        mutate(season = as.factor(season)) %>%
        mutate(weather=as.factor(weather))
```

# Univariate Regressions

```
mod1 <- lm(count~season, data=bikes)
mod2 <- lm(count~holiday, data=bikes)
mod3 <- lm(count~workingday, data=bikes)
mod4 <- lm(count~weather, data=bikes)
mod5 <- lm(count~temp, data=bikes)
mod6 <- lm(count~atemp, data=bikes)
mod7 <- lm(count~humidity, data=bikes)
mod8 <- lm(count~windspeed, data=bikes)
mod9 <- lm(count~casual, data=bikes)
mod10 <- lm(count~registered, data=bikes)
```

# Univariate Regressions

```
library(broom)
tidy1 <- tidy( mod1, conf.int=T )[-1, -c(3:4) ]
tidy2 <- tidy(mod2, conf.int=T )[-1, -c(3:4) ]
tidy3 <- tidy(mod3 , conf.int=T)[-1, -c(3:4) ]
tidy4 <- tidy(mod4 , conf.int=T)[-1, -c(3:4) ]
tidy5 <- tidy(mod5, conf.int=T)[-1, -c(3:4) ]
tidy6 <- tidy(mod6 , conf.int=T)[-1, -c(3:4) ]
tidy7 <- tidy(mod7 , conf.int=T)[-1, -c(3:4) ]
tidy8 <- tidy(mod8 , conf.int=T)[-1, -c(3:4) ]
tidy9 <- tidy(mod9, conf.int=T)[-1, -c(3:4) ]
tidy10 <- tidy(mod10, conf.int=T)[-1, -c(3:4) ]
bind_rows(tidy1, tidy2, tidy3, tidy4, tidy5, tidy6, tidy7, tidy8, tidy9, tidy10)
```

# Univariate Regressions

```
##          term   estimate       p.value    conf.low  conf.high
## 2     season2  98.908111  9.756471e-94   89.559922 108.256300
## 3     season3 118.073863 1.063174e-131  108.725674 127.422052
## 4     season4  82.645034  2.127949e-66   73.297693  91.992376
## 21    holiday  -5.863841  5.736924e-01  -26.292923  14.565240
## 22 workingday   4.505252  2.264480e-01   -2.795435  11.805939
## 23   weather2 -26.281251  4.317735e-11  -34.087322 -18.475180
## 31   weather3 -86.390458  3.285377e-40  -99.096108 -73.684808
## 41   weather4 -41.236791  8.183717e-01 -393.221331 310.747749
## 24       temp   9.170540  0.000000e+00    8.769141   9.571940
## 25      atemp   8.331636  0.000000e+00    7.961788   8.701484
## 26   humidity  -2.987269 2.921542e-253   -3.154977  -2.819560
## 27  windspeed   2.249058  2.898407e-26    1.834340   2.663776
## 28     casual   2.503271  0.000000e+00    2.453989   2.552552
## 29 registered   1.164480  0.000000e+00    1.159087   1.169872
```

# Multivariate

```
mod.final <- lm(count~season+weather+humidity+windspeed, data=bikes)
tidy(mod.final)[-1,-c(3:4)]
glance(mod.final)
```

# Multivariate

```
##        term    estimate       p.value
## 2   season2 115.8007186 1.403611e-145
## 3   season3 148.3532069 7.517679e-227
## 4   season4 118.4943844 1.738000e-147
## 5  weather2  19.9875113  1.383456e-07
## 6  weather3   0.1237865  9.844830e-01
## 7  weather4 162.2596870  3.185115e-01
## 8  humidity  -3.4929513 3.860368e-273
## 9 windspeed   0.6328680  2.049791e-03
```

# Multivariate

```
##   r.squared adj.r.squared    sigma statistic p.value df    logLik      AIC
## 1 0.1949699     0.1943778 162.5889  329.2869       0  9 -70865.13 141750.3
##        BIC  deviance df.residual
## 1 141823.2 287534958       10877
```

# Inference on Linear Regressions

# Inference on Linear Regressions

1. Overall F Test of Model
2. Individual Coefficient Tests
3. Testing Groups of Variables

# Overall Model F test

- We can perform an overall F Test for a model.
- When we do this we test the following Hypothesis

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_1 = \text{ at least one } \beta_i \neq 0$$

# Overall Model F test: Bike Sharing

```
glance(mod.final)
```

# Overall Model F test: Bike Sharing

```
##   r.squared adj.r.squared    sigma statistic p.value df    logLik      AIC
## 1 0.1949699     0.1943778 162.5889  329.2869       0  9 -70865.13 141750.3
##        BIC  deviance df.residual
## 1 141823.2 287534958       10877
```

# Overall Model F test: Bike Sharing

- We have an F Statistic of 3329.3
- This yields a p-value of $< 0.0001$
- We can reject the null in favor of the alternative hypothesis.
- This suggests that at least one $\beta_I$ is not 0.

# Individual Coefficients $t$-test

- We can test each individual coefficients.
- The hypothesis we test is that:

$$H_0 : \beta_i = 0$$

$$H_1 = \beta_i \neq 0$$

- We do this with a t-test.

# Individual Coefficients $t$-test

- With the t-test we have that:

$$t_i = \frac{\beta_i}{se(\beta_i)}$$

- Then we can test this with the $t$-distribution.

# Individual Coefficients $t$-test

- Consider out Bike model:

$$\begin{aligned} E[count] = \beta_0 + \beta_1 season(Summer) + \beta_2 season(Fall) + \\ \beta_3 season(Winter) + \beta_4 weather(2) + \beta_5 weather(3) + \\ \beta_6 weather(4) + \beta_7 humidity + \beta_8 windspeed \end{aligned}$$

```
tidy(mod.final)
```

# Individual Coefficients $t$-Test

```
##          term    estimate    std.error    statistic       p.value
## 1 (Intercept) 298.3348913   7.36160428  40.52579846  0.000000e+00
## 2     season2 115.8007186   4.43879843  26.08830302 1.403611e-145
## 3     season3 148.3532069   4.50438417  32.93529177 7.517679e-227
## 4     season4 118.4943844   4.51125815  26.26637190 1.738000e-147
## 5    weather2  19.9875113   3.79203900   5.27091395  1.383456e-07
## 6    weather3   0.1237865   6.36457573   0.01944929  9.844830e-01
## 7    weather4 162.2596870 162.65541954   0.99756705  3.185115e-01
## 8    humidity  -3.4929513   0.09609386 -36.34936864 3.860368e-273
## 9   windspeed   0.6328680   0.20523232   3.08366623  2.049791e-03
```

# F-test for Groups of Coefficients

- Many times we want to be able to test the significance of groups of coefficients.
- We can do this with an F-test as well.
- For example we may want to test that:

$$H_0 : \beta_1 = \beta_2 = 0$$

$$H_1 : \text{ at least } 1 \ \beta_i \neq 0$$

# Groups of Coefficients Example

- Consider `Season` in our bike example.
- Only the first coefficient is significant.
- We may want to know if we the whole variable is worth having in the model.
- We will use the `anova()` function in R.

# Groups of Coefficients Example

```
mod1 <- lm(count~season+weather+humidity+windspeed, data=bikes)
mod2 <- lm(count~weather+humidity+windspeed, data=bikes)
anova(mod1, mod2)
```

# Groups of Coefficients Example

```
## Analysis of Variance Table
##
## Model 1: count ~ season + weather + humidity + windspeed
## Model 2: count ~ weather + humidity + windspeed
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1  10877 287534958
## 2  10880 320760441 -3 -33225483 418.96 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

# Groups of Coefficients Example 2

- Consider `weather` in our bike example.
- Only the first coefficient is significant.
- We may want to know if we the whole variable is worth having in the model.
- We will use the `anova()` function in R.

# Groups of Coefficients Example 2

```
mod1 <- lm(count~season+weather+humidity+windspeed, data=bikes)
mod2 <- lm(count~season+humidity+windspeed, data=bikes)
anova(mod1, mod2)
```

# Groups of Coefficients Example 2

```
## Analysis of Variance Table
##
## Model 1: count ~ season + weather + humidity + windspeed
## Model 2: count ~ season + humidity + windspeed
##   Res.Df       RSS Df Sum of Sq      F    Pr(>F)
## 1  10877 287534958
## 2  10880 288348337 -3   -813379 10.256 9.704e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```