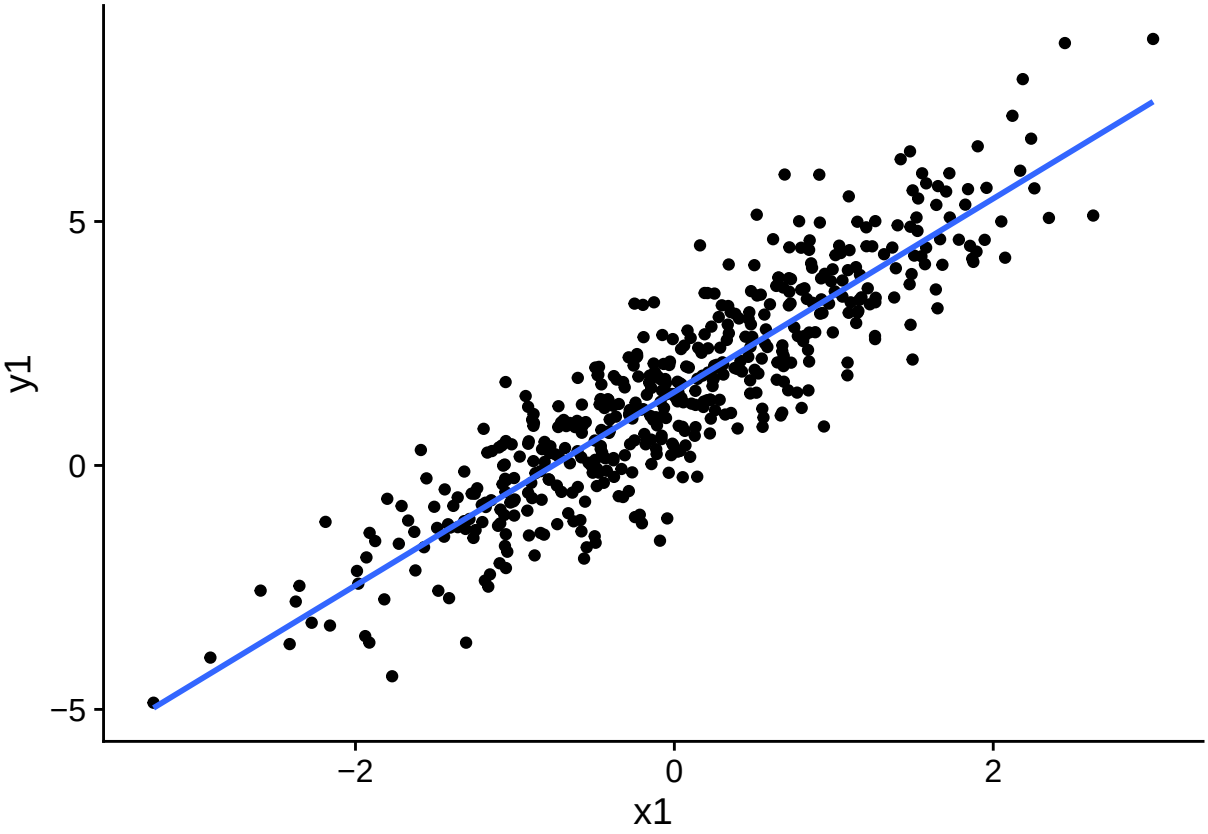
BIFX 553 Intro

1/18/2018

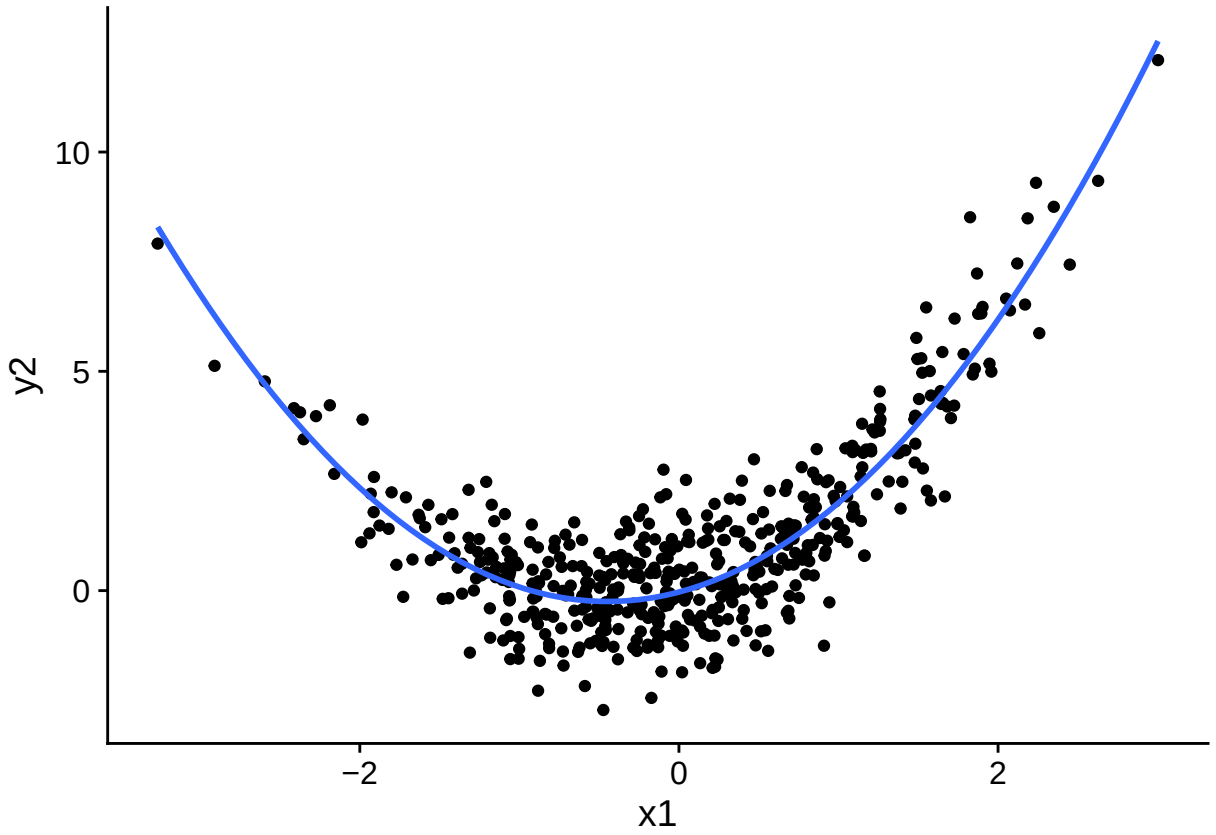
Intro to BIFX 553 and Linear Regression

What are linear regression models?

I their simplest form, a linear regression model is used to explain a relationship between two variables.



In this first figure we can see that the blue regression summarizes the relationship between x1 and y1. When x1 is equal to -2, for example, y1 is usually somewhere around -3.



As we can see here, a linear model doesn't have to be restricted to a straight line. It appears that there is a quadratic relationship between x1 and y2, and we can include that hypothesis in our model.

What makes linear models linear is that the dependent variable, y, is a linear combination of the independent variables in the model (in this case, `x` and x^2 are our independent variables). Thus we can describe this model with this equation:

$$E(y) = intercept + x\beta_1 + x^2\beta_2.$$

In algebra 1 we might write this model out as

$$y = ax^2 + bx + c.$$

We will talk about more complicated models another time. We can model very complicated relationships between many variables and their effect on an outcome, but for now we'll stick with simple linear relationships.

Mathematical representation of regression models

If we wanted to model the relationship between height and weight, we could use the observation that the slope of log weight regressed on log of height is approximately 2. As you might guess, this rule of thumb isn't very precise, but using this information we would write this model as:

$$\begin{aligned} \log(height) &= \beta_0 + \log(weight)\beta_1 + \varepsilon \\ &= \beta_0 + \log(weight) * 2 + \varepsilon \end{aligned}$$

where ϵ is the error term, accounting for the fact that everyone is a little different.

Interpretation of Intercept

We may want to estimate the intercept, β_0, to make our predictions a little better. The interpretation of β_0 is the value of the dependent variable when the independent variable (or variables) are equal to 0. In this case, the interpretation of the intercept, β_0, isn't terribly relevant, since we will never really want to predict the height of an individual who weighs only one kilogram (because the log of 1 kilogram is 0).

Interpretation of other regression coefficients

The interpretation of the other regression coefficient is a little more straight forward. β_1 represents how much $\log(height)$ changes for every unit increase in $\log(weight)$. So, on average the $\log(height)$ of an individual increases by $2\log(cm)$ for each additional $\log(kilogram)$.

Problem 1

If we collected some data and estimated $\beta_0 = -3.6$, what would be the average height of someone weighing 72 kg?

Problem 2

Suppose the lifetime cost of a kidney transplant depends on how far away the donor lives, the amount of time spent in the operating room, the number of days spent in the hospital, and the number of years the patient lives after transplant (to cover the cost of immunosuppression therapy). Write a linear model equation given this information.

Categorical Variables

The interpretation of categorical variables is a little different. For this section, we will use the following data set:

```
require(dplyr)
set.seed(238947)
n <- 200 # sample size
dat <- data_frame(cnt = rnorm(n) + 3,
                  cat = ifelse(rbinom(n, size = 1, prob = 0.5), 'A', 'B'),
                  resp1 = 2 + 2*(cat == 'B') + 0.5*cnt + rnorm(n),
                  resp2 = 2 + 0.5*cnt + 1.5*(cat == 'B')*cnt + rnorm(n),
                  resp3 = 2 + 2*(cat == 'B') + 0.5*cnt + 1.5*(cat == 'B')*cnt + rnorm(n))
```

First lets familiarize ourselves with the data.

```
summary(dat)
```

```
##       cnt             cat                resp1            resp2
##  Min.   :0.237   Length:200         Min.   :0.7538   Min.   : 0.8226
##  1st Qu.:2.396   Class :character   1st Qu.:3.5956   1st Qu.: 3.3884
##  Median :3.019   Mode  :character   Median :4.5727   Median : 4.7126
##  Mean   :3.019                      Mean   :4.5150   Mean   : 5.6186
##  3rd Qu.:3.696                      3rd Qu.:5.3246   3rd Qu.: 7.7014
##  Max.   :5.428                      Max.   :8.3765   Max.   :13.2782
```

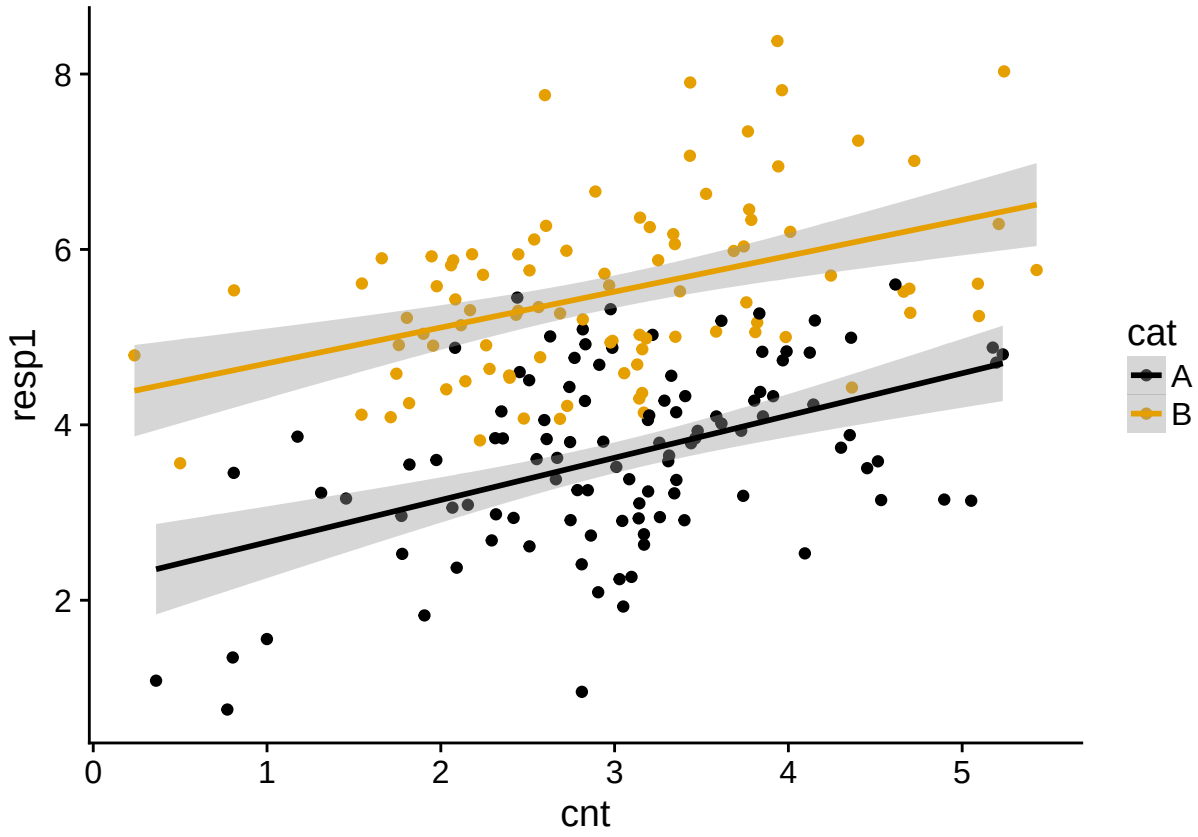


Figure 1: Response 1 - only the intercept is affected.

```
##      resp3
##  Min.   : 0.347
##  1st Qu.: 3.477
##  Median : 4.772
##  Mean   : 6.370
##  3rd Qu.: 9.375
##  Max.   :14.316
```

A categorical variable can affect the slope or the intercept of the regression line, or both.

```
require(ggplot2)
ggplot(dat, aes(cnt, resp1, color = cat)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    scale_color_manual(values = cbbPalette)
```

```
ggplot(dat, aes(cnt, resp2, color = cat)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    scale_color_manual(values = cbbPalette)
```

```
ggplot(dat, aes(cnt, resp3, color = cat)) +
    geom_point() +
    geom_smooth(method = 'lm') +
    scale_color_manual(values = cbbPalette)
```

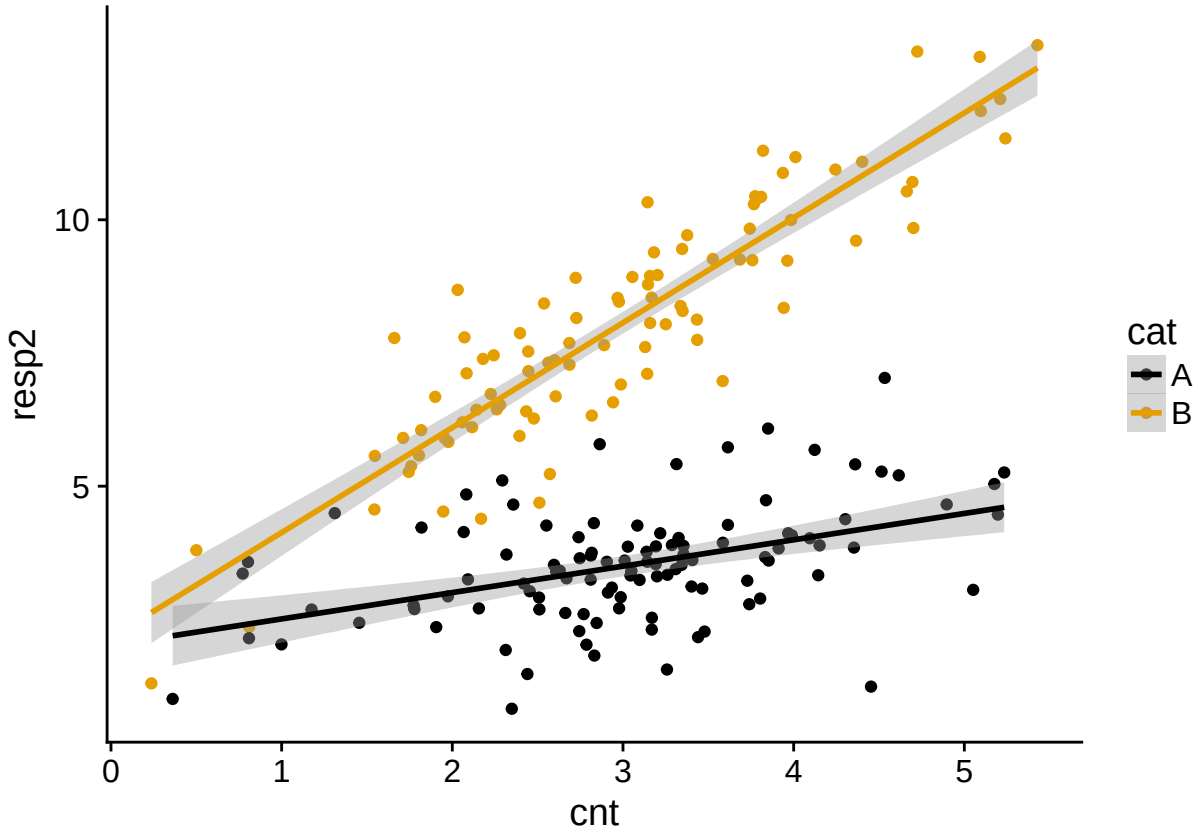


Figure 2: Response 2 - only the slope is affected.

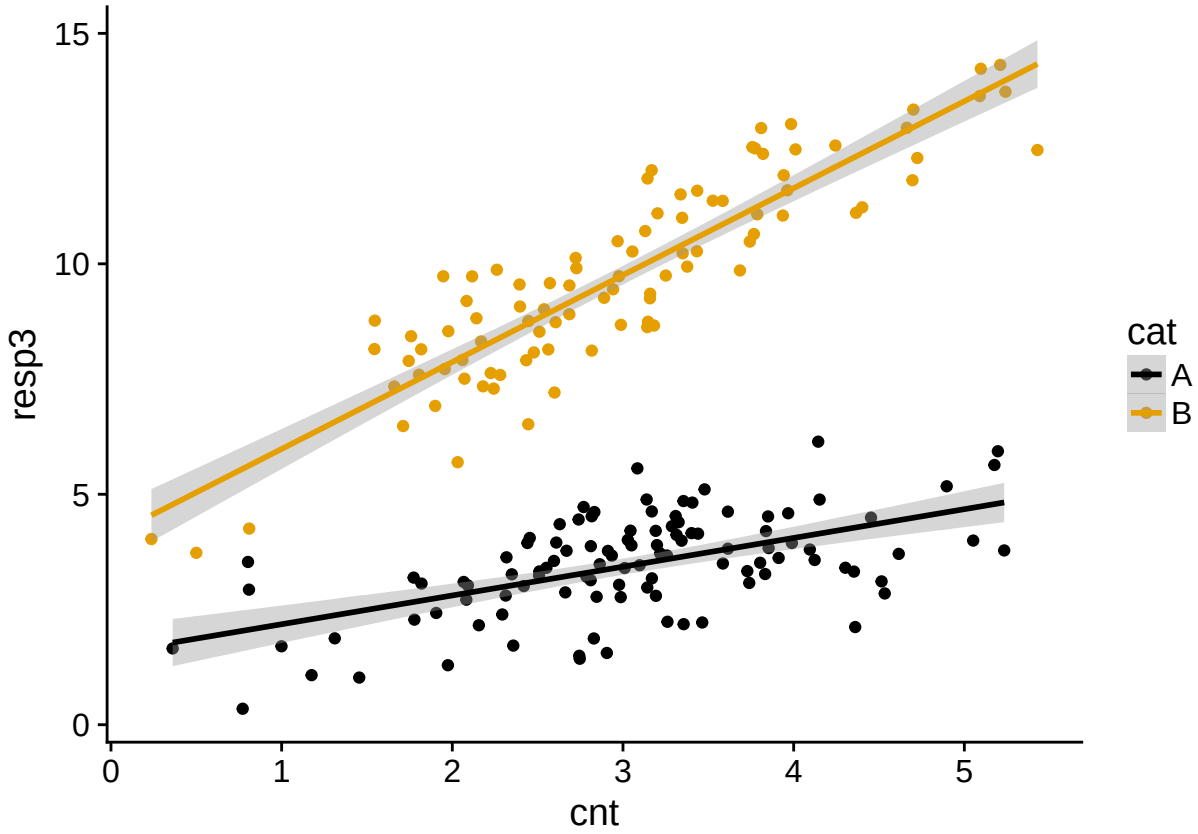


Figure 3: Response 3 - both the slope and the intercept are affected.

Interpretation of model coefficients

In this section we will use `lm` to estimate the effects of `cnt` and `cat` on the response variables in `dat`.

```
# only the intercept is affected
model1 = lm(resp1 ~ cat*cnt, data = dat)
summary(model1)
```

```
##
## Call:
## lm(formula = resp1 ~ cat * cnt, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.57841 -0.71051  0.00062  0.64640  2.47501
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.17954    0.28882   7.546 1.64e-12 ***
## catB         2.11228    0.40557   5.208 4.80e-07 ***
## cnt          0.48174    0.09014   5.344 2.51e-07 ***
## catB:cnt    -0.07286    0.12750  -0.571    0.568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9086 on 196 degrees of freedom
## Multiple R-squared:  0.568,  Adjusted R-squared:  0.5614
## F-statistic:  85.9 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
# only the slope is affected
model2 = lm(resp2 ~ cat*cnt, data = dat)
summary(model2)
```

```
##
## Call:
## lm(formula = resp2 ~ cat * cnt, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.97964 -0.57637 -0.02787  0.55858  2.77485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.01323    0.31567   6.378 1.26e-09 ***
## catB         0.14727    0.44328   0.332     0.74
## cnt          0.49494    0.09852   5.024 1.14e-06 ***
## catB:cnt     1.47423    0.13935  10.579  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.993 on 196 degrees of freedom
## Multiple R-squared:  0.8808, Adjusted R-squared:  0.879
## F-statistic: 482.8 on 3 and 196 DF,  p-value: < 2.2e-16
```

```
# both are affected
model3 = lm(resp3 ~ cat*cnt, data = dat)
```

```
summary(model3)
```

```
##
## Call:
## lm(formula = resp3 ~ cat * cnt, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.23287 -0.67559  0.03659  0.65974  2.08056
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1.5566     0.3002   5.185 5.36e-07 ***
## catB          2.5404     0.4216   6.026 8.19e-09 ***
## cnt           0.6240     0.0937   6.660 2.70e-10 ***
## catB:cnt      1.2620     0.1325   9.523  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9444 on 196 degrees of freedom
## Multiple R-squared:  0.9307, Adjusted R-squared:  0.9296
## F-statistic: 877.5 on 3 and 196 DF,  p-value: < 2.2e-16
```

Problem 3

Write out the equations for our models, based on the output from our linear regressions.

```
# E(resp_1) = 2.2 + 2.1*(cat == 'B') + 0.5*cnt - 0.1*(cat == 'B')*cnt
# E(resp_2) = 2.0 + 0.1*(cat == 'B') + 0.5*cnt + 1.5*(cat == 'B')*cnt
# E(resp_3) = 1.6 + 2.5*(cat == 'B') + 0.6*cnt + 1.3*(cat == 'B')*cnt
```

Model updates

We could drop the non-significant terms if we want to.

```
require(magrittr)
```

```
## Loading required package: magrittr
```

```
##
## Attaching package: 'magrittr'
```

```
## The following object is masked from 'package:purrr':
##
##     set_names
```

```
## The following object is masked from 'package:tidyr':
##
##     extract
```

```
update(model1, . ~ . - cat:cnt) %>%
    summary()
```

```
##
## Call:
```

```
## lm(formula = resp1 ~ cat + cnt, data = dat)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.58717 -0.71707  0.01572  0.65189  2.44014
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.29070    0.21312  10.748  < 2e-16 ***
## catB         1.89252    0.12867  14.708  < 2e-16 ***
## cnt          0.44532    0.06364   6.997 3.98e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.907 on 197 degrees of freedom
## Multiple R-squared:  0.5673, Adjusted R-squared:  0.5629
## F-statistic: 129.1 on 2 and 197 DF,  p-value: < 2.2e-16
```

```
update(model2, . ~ . - cat) %>%
    summary()
```

```
##
## Call:
## lm(formula = resp2 ~ cnt + cat:cnt, data = dat)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.9554 -0.5846 -0.0421  0.5701  2.8008
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.08791    0.22112   9.443  < 2e-16 ***
## cnt          0.47274    0.07222   6.546 5.02e-10 ***
## cnt:catB     1.51813    0.04419  34.358  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9908 on 197 degrees of freedom
## Multiple R-squared:  0.8807, Adjusted R-squared:  0.8795
## F-statistic: 727.4 on 2 and 197 DF,  p-value: < 2.2e-16
```

Howell1 data set found here.

```
require(readr) # also loaded by library(tidyverse)
require(dplyr)
require(magrittr)
howell1 <- read_delim('https://raw.githubusercontent.com/rmcelreath/rethinking/master/data/Howell1.csv'
    mutate(sex = ifelse(male, 'male', 'female'))
```

Practice

- Familiarize yourself with the data.
- What variables are important?
- Develop a model to predict weight given the provided data (hint: you may want to work only with data from adults).

Bonus Challenge Problem

- Develop a model for children.
  - What challenges did you encounter?
  - Is your model useful?

Provide an Rmd file with your analysis and comments.