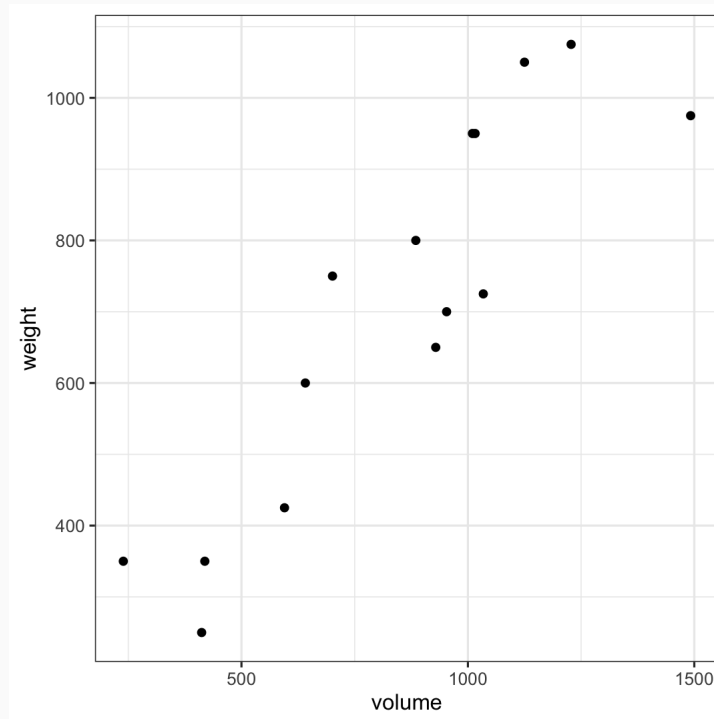


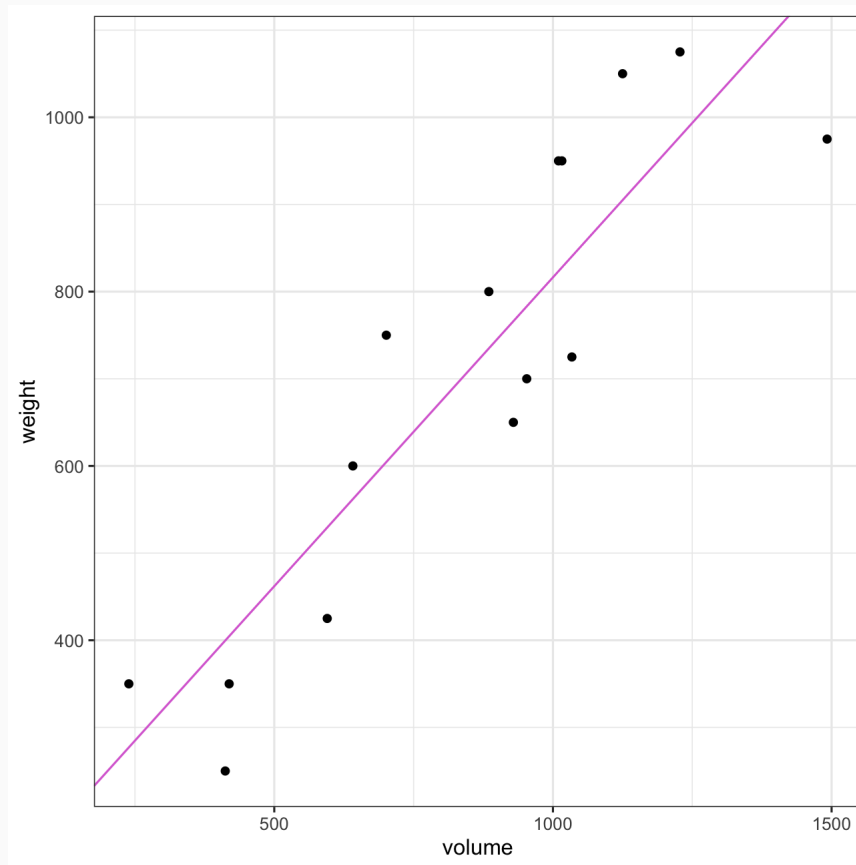
# Extending the Linear Model

## Example: shipping books

```
ggplot(books, aes(x = volume, y = weight)) +  
  geom_point()
```



```
ggplot(books, aes(x = volume, y = weight)) +  
  geom_point() +  
  geom_abline(intercept = m1$coef[1],  
              slope = m1$coef[2], col = "orchid")
```



# Fitting the linear model

```
m1 <- lm(weight ~ volume, data = books)
```

```
summary(m1)
```

```
##
## Call:
## lm(formula = weight ~ volume, data = books)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -189.97 -109.86   38.08  109.73  145.57
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.67931   88.37758   1.218   0.245
## volume       0.70864    0.09746   7.271 6.26e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 123.9 on 13 degrees of freedom
## Multiple R-squared:  0.8026,    Adjusted R-squared:  0.7875
## F-statistic: 52.87 on 1 and 13 DF,  p-value: 6.262e-06
```

# Multiple Regression

Allows us create a model to explain one *numerical* variable, the response, as a linear function of many explanatory variables that can be both *numerical* and *categorical*.

We posit the true model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \epsilon; \quad \epsilon \sim N(0, \sigma^2)$$

## Estimating $\beta_0, \beta_1$ etc.

In least-squares regression, we're still finding the estimates that minimize the sum of squared residuals.

$$e_i = y_i - \hat{y}_i$$

$$RSS = \sum_{i=1}^n e_i^2$$

And yes, they have a closed-form solution.

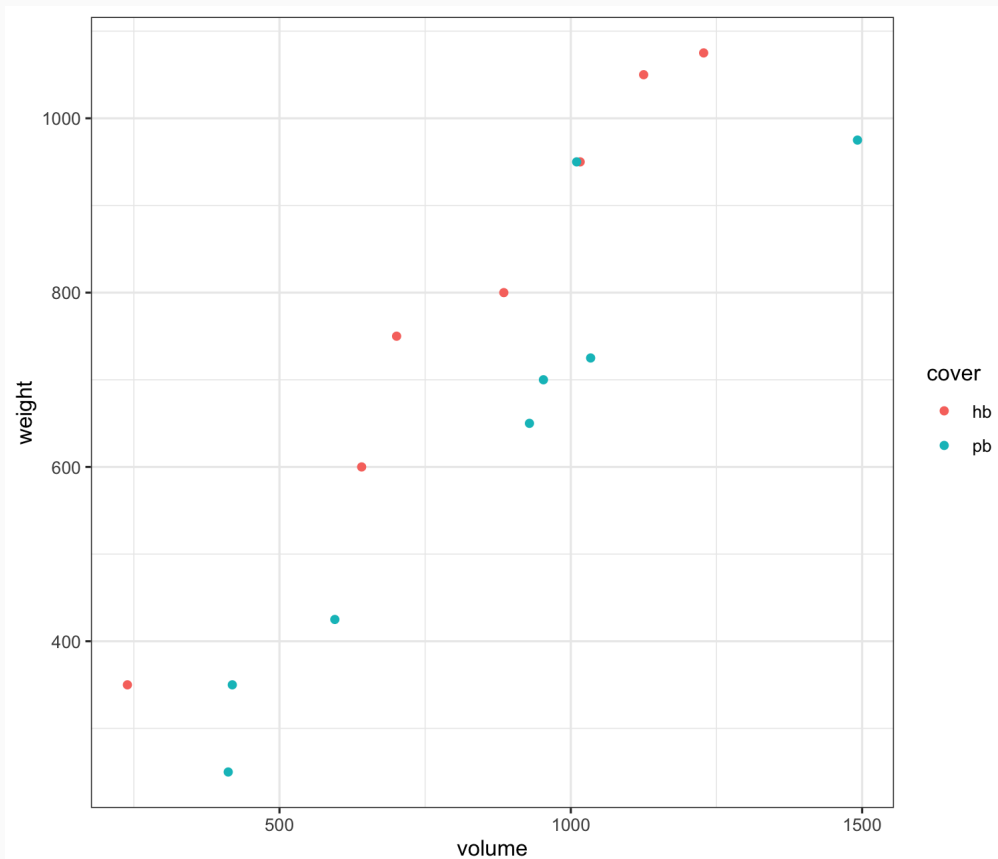
$$\hat{\beta} = (X'X)^{-1}X'Y$$

In R:

```
lm(Y ~ X1 + X2 + ... + Xp, data = mydata)
```

# Example: shipping books

```
ggplot(books, aes(x = volume,  
                  y = weight,  
                  color = cover)) +  
  geom_point()
```



# Example: shipping books

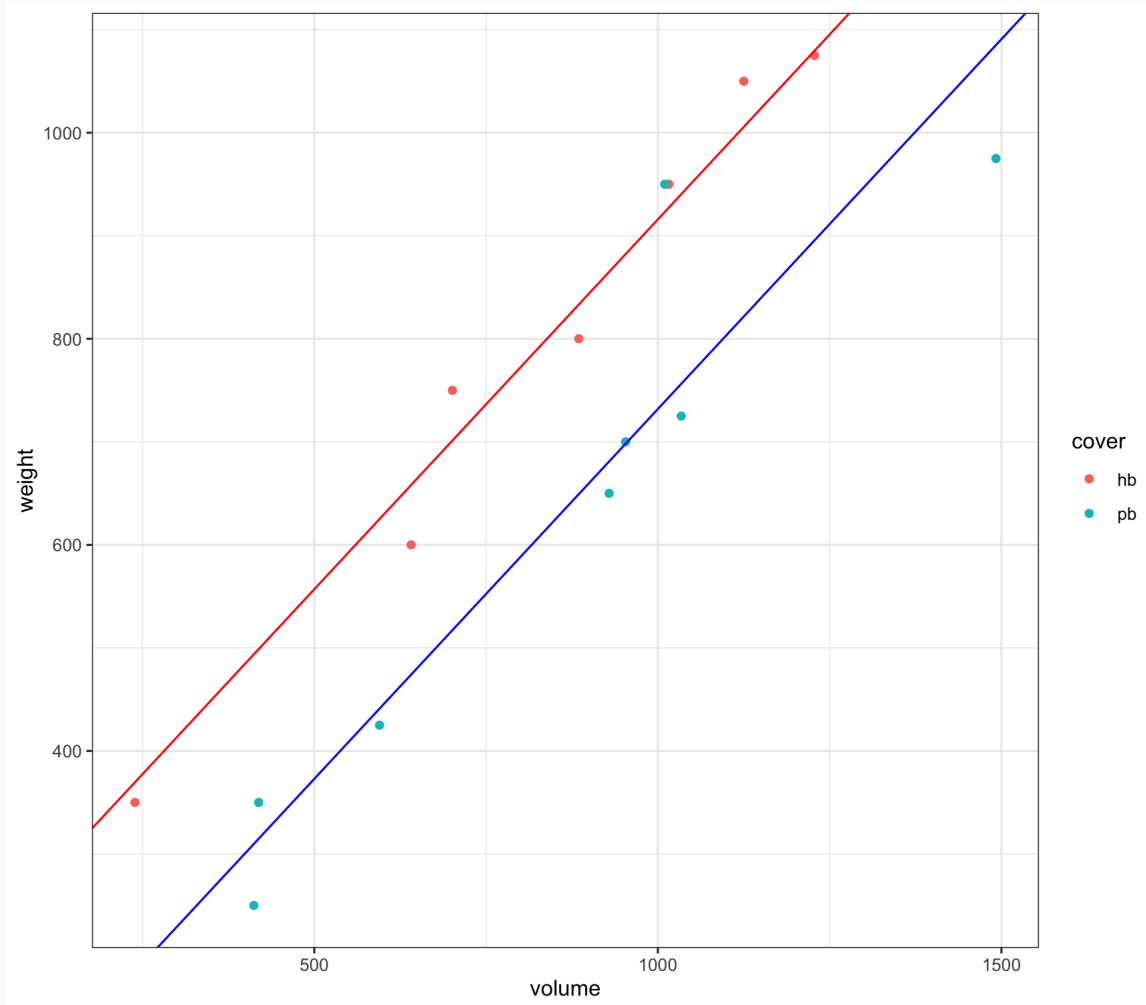
```
m2 <- lm(weight ~ volume + cover, data = books)
```

```
summary(m2)
```

```
##  
## Call:  
## lm(formula = weight ~ volume + cover, data = books)  
##  
## Residuals:  
##      Min       1Q   Median       3Q      Max   
## -110.10  -32.32  -16.10   28.93  210.95   
##  
## Coefficients:  
##              Estimate Std. Error t value Pr(>|t|)      
## (Intercept)  197.96284    59.19274   3.344 0.005841 **   
## volume        0.71795     0.06153  11.669 6.6e-08 ***   
## coverpb     -184.04727    40.49420  -4.545 0.000672 ***   
## ---  
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1  
##  
## Residual standard error: 78.2 on 12 degrees of freedom  
## Multiple R-squared:  0.9275,    Adjusted R-squared:  0.9154   
## F-statistic: 76.73 on 2 and 12 DF,  p-value: 1.455e-07
```



# Example: shipping books



# MLR slope interpretation

The slope corresponding to the dummy variable tell us:

- How much vertical separation there is between our lines
- How much **weight** is expected to increase if **cover** goes from 0 to 1 and **volume** is left unchanged.

Each  $\hat{\beta}_i$  tells you how much you expect the  $Y$  to change when you change the  $X_i$ , while **holding all other variables constant**.

## Activity

Load in the LA homes data set and fit the following model:

$$\logprice \sim \logsqft + bed + city$$

```
URL <- "http://andrewpbray.github.io/data/LA.csv"  
LA <- read.csv(URL)
```

1. What is the geometry of this model?
2. What appears to be the reference level for `city`?
3. In the context of this problem, what is suggested by the *sign* of the coefficient for `bed`? Do this make sense to you?

```
LA <- read.csv("http://andrewpbray.github.io/data/LA.csv")
LA <- mutate(LA, logprice = log(price), logsqft = log(sqft))
m1 <- lm(logprice ~ logsqft + bed + city, data = LA)
summary(m1)
```

```
##
## Call:
## lm(formula = logprice ~ logsqft + bed + city, data = LA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.26020 -0.24897 -0.01613  0.21804  1.37277
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    5.13068    0.21201  24.200  <2e-16 ***
## logsqft        1.20729    0.03036  39.769  <2e-16 ***
## bed           -0.03010    0.01284  -2.345   0.0191 *
## cityLong Beach -0.88280    0.03467 -25.464  <2e-16 ***
## citySanta Monica -0.09416    0.04022  -2.341   0.0194 *
## cityWestwood   -0.46244    0.04876  -9.484  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3704 on 1588 degrees of freedom
## Multiple R-squared:  0.8742,    Adjusted R-squared:  0.8738
## F-statistic: 2206 on 5 and 1588 DF,  p-value: < 2.2e-16
```

# Interactions

Does the relationship between `logsqft` and `logprice` change depending on the `city`?

```
m2 <- lm(logprice ~ logsqft + bed + city + logsqft:city,  
          data = LA)
```

```
summary(m2)
```

```
##
## Call:
## lm(formula = logprice ~ logsqft + bed + city + logsqft:city,
##     data = LA)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.30385 -0.23866 -0.01576  0.21562  1.36668
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    4.43151    0.38515   11.506 < 2e-16 ***
## logsqft        1.29578    0.05019   25.820 < 2e-16 ***
## bed           -0.03794    0.01296   -2.928 0.003460 **
## cityLong Beach -0.53386    0.37968   -1.406 0.159902
## citySanta Monica  1.75128    0.47010    3.725 0.000202 ***
## cityWestwood    2.43192    0.90674    2.682 0.007394 **
## logsqft:cityLong Beach -0.03663    0.04730   -0.774 0.438807
## logsqft:citySanta Monica -0.24345    0.06052   -4.022 6.03e-05 ***
## logsqft:cityWestwood -0.38773    0.12251   -3.165 0.001581 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3676 on 1585 degrees of freedom
## Multiple R-squared:  0.8763,    Adjusted R-squared:  0.8757
```

# Interactions

Does the relationship between `logsqft` and `logprice` change depending on the number of `bed`?

```
m3 <- lm(logprice ~ logsqft + bed + logsqft:bed, d
```

```
summary(m3)
```

```
##
## Call:
## lm(formula = logprice ~ logsqft + bed + logsqft:bed, data = m3)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.75668 -0.32825 -0.04576  0.31841  1.85602
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  2.803227   0.271328  10.331 < 2e-16 ***
## logsqft      1.487273   0.040007  37.175 < 2e-16 ***
## bed         -0.644164   0.067255  -9.578 < 2e-16 ***
## logsqft:bed  0.064093   0.008023   7.989 2.59e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4783 on 1590 degrees of freedom
## Multiple R-squared:  0.7899,    Adjusted R-squared: 0.7879
```