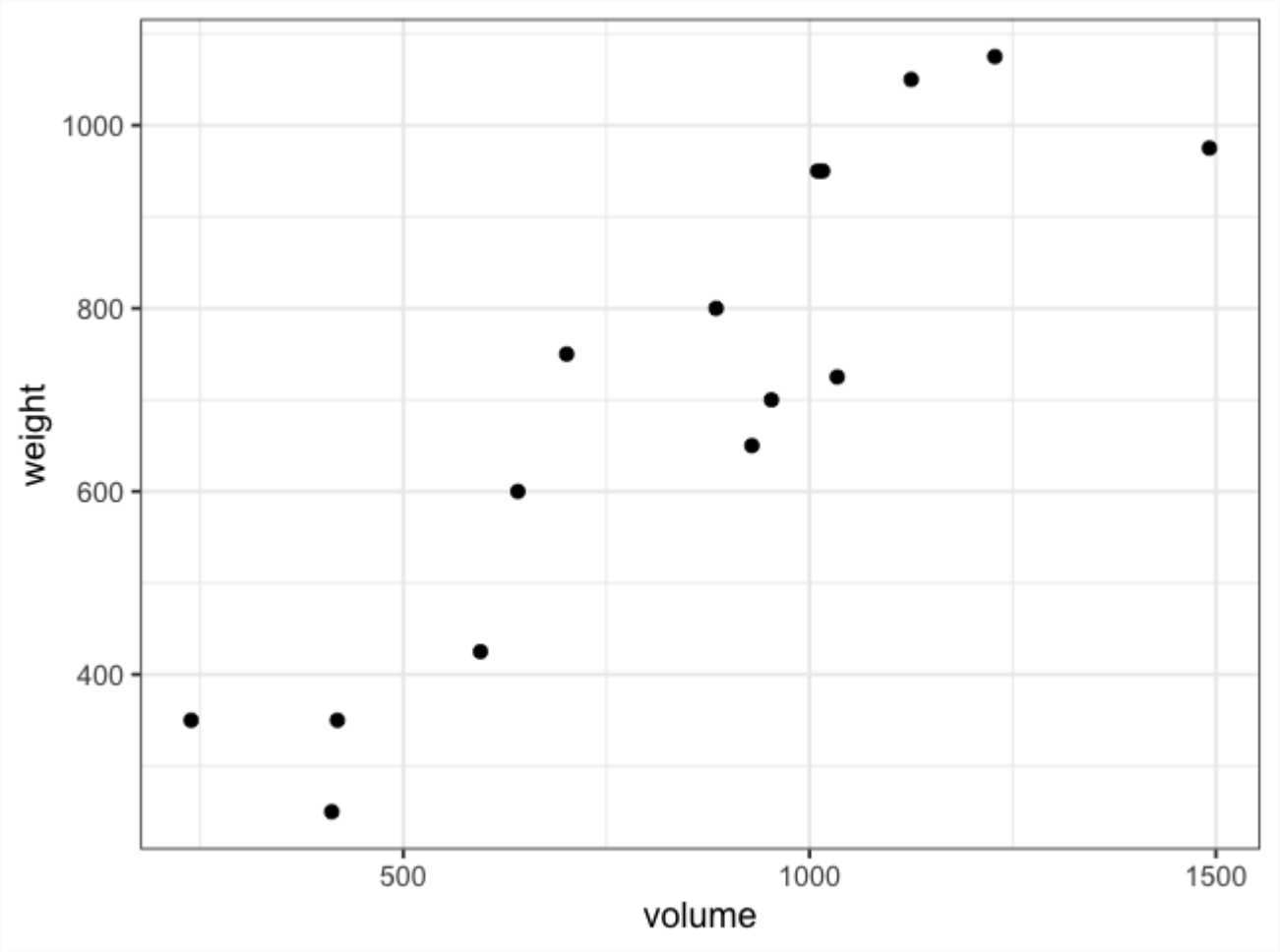
SLR to MLR

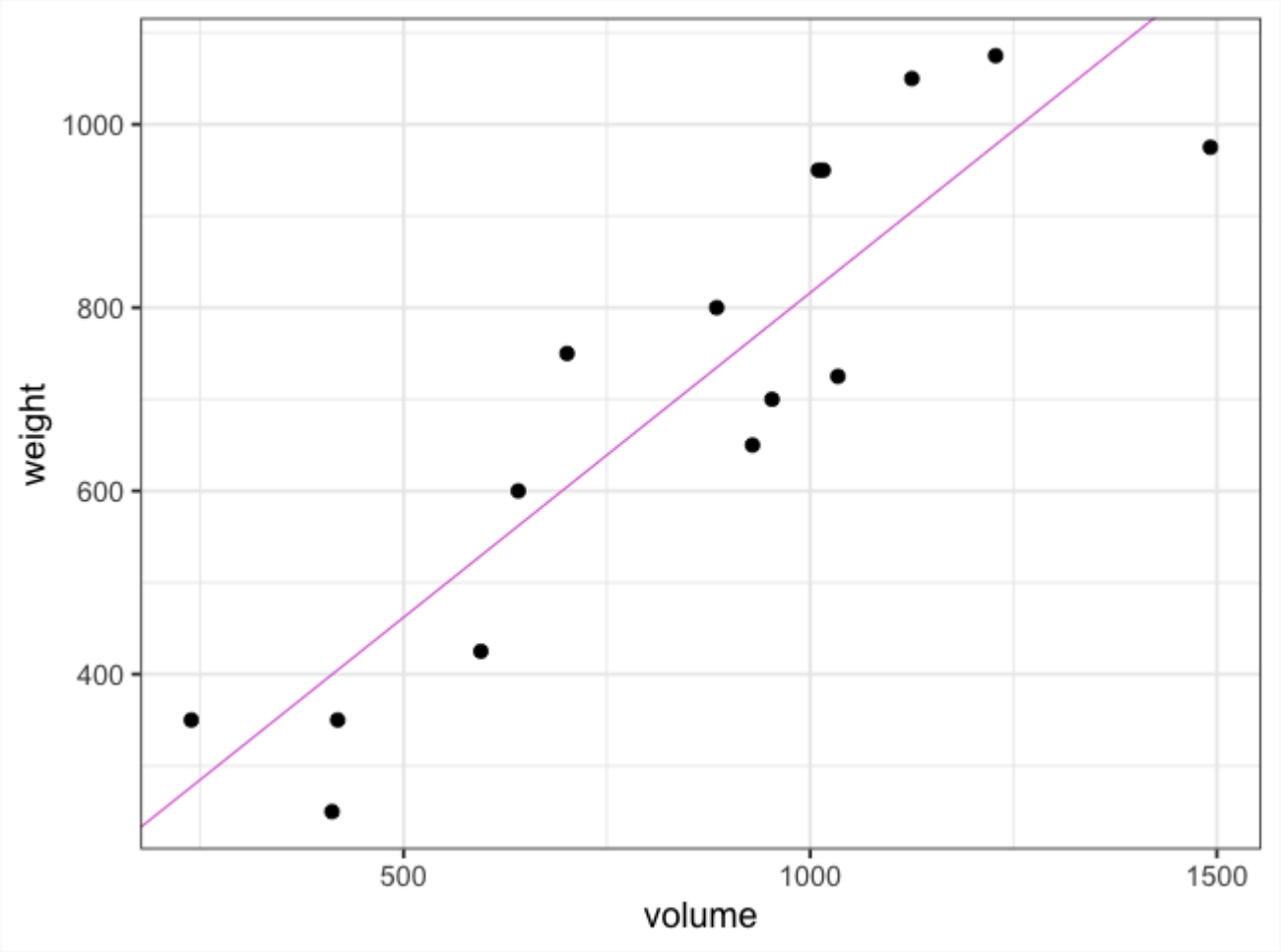
Example: shipping books

When you buy a book off of Amazon, you get a quote for how much it costs to ship. This is based on the weight of the book. If you didn't know the weight a book, what other characteristics of it could you measure to help predict weight?

Shipping books visualized



Shipping books visualized, cont.



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
## -190.0 -109.9   38.1  109.7  145.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.6793    88.3776    1.22     0.24
## volume        0.7086     0.0975    7.27  6.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124 on 13 degrees of freedom
## Multiple R-squared:  0.803,    Adjusted R-squared:  0.787
## F-statistic: 52.9 on 1 and 13 DF,  p-value: 6.26e-06
```

Q1: What is the equation for the line?

$$\hat{y} = 107.7 + 0.708x$$

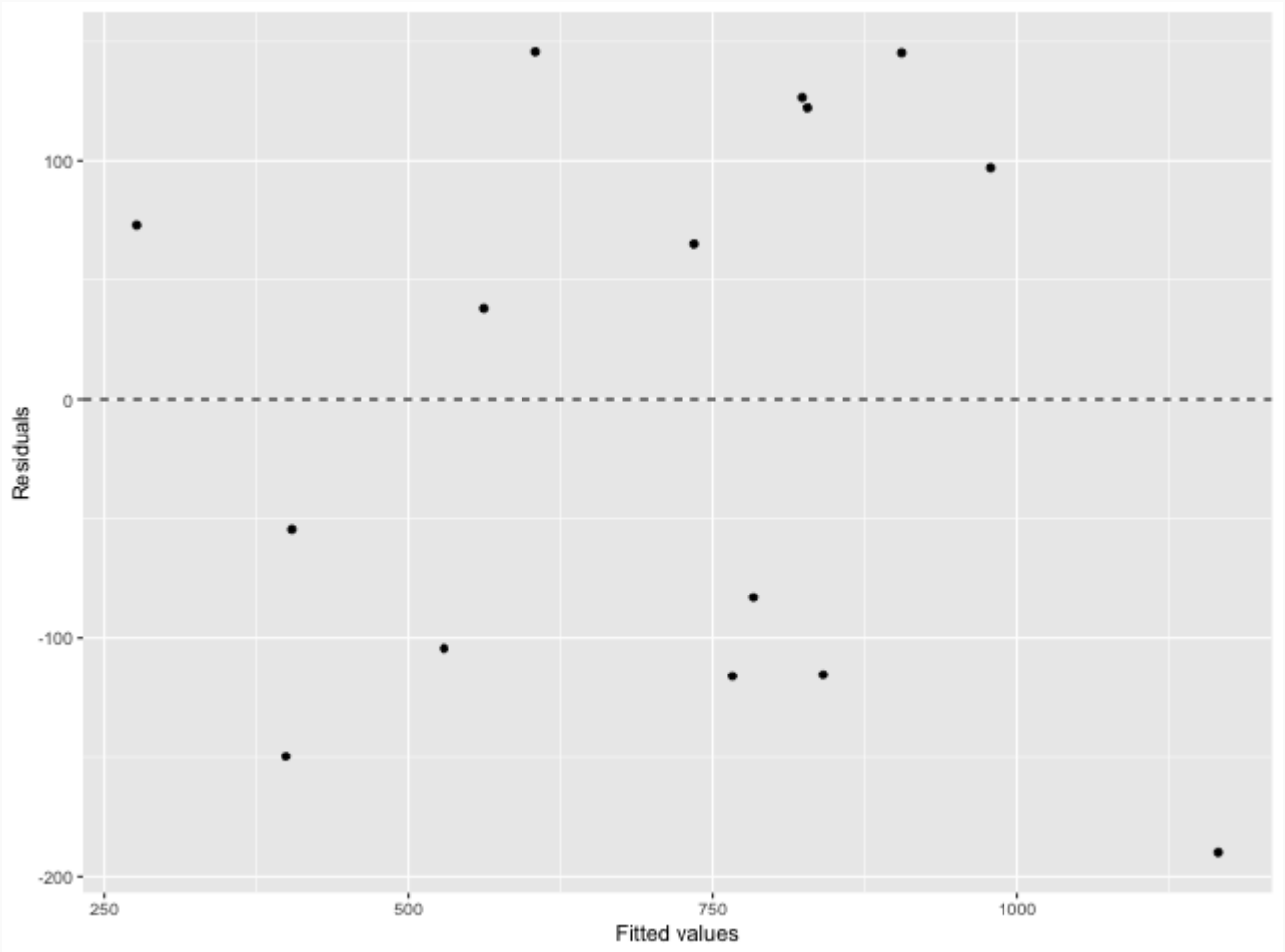
$$\widehat{weight} = 107.7 + 0.708volume$$

Q2: Does this appear to be a reasonable setting to apply linear regression?

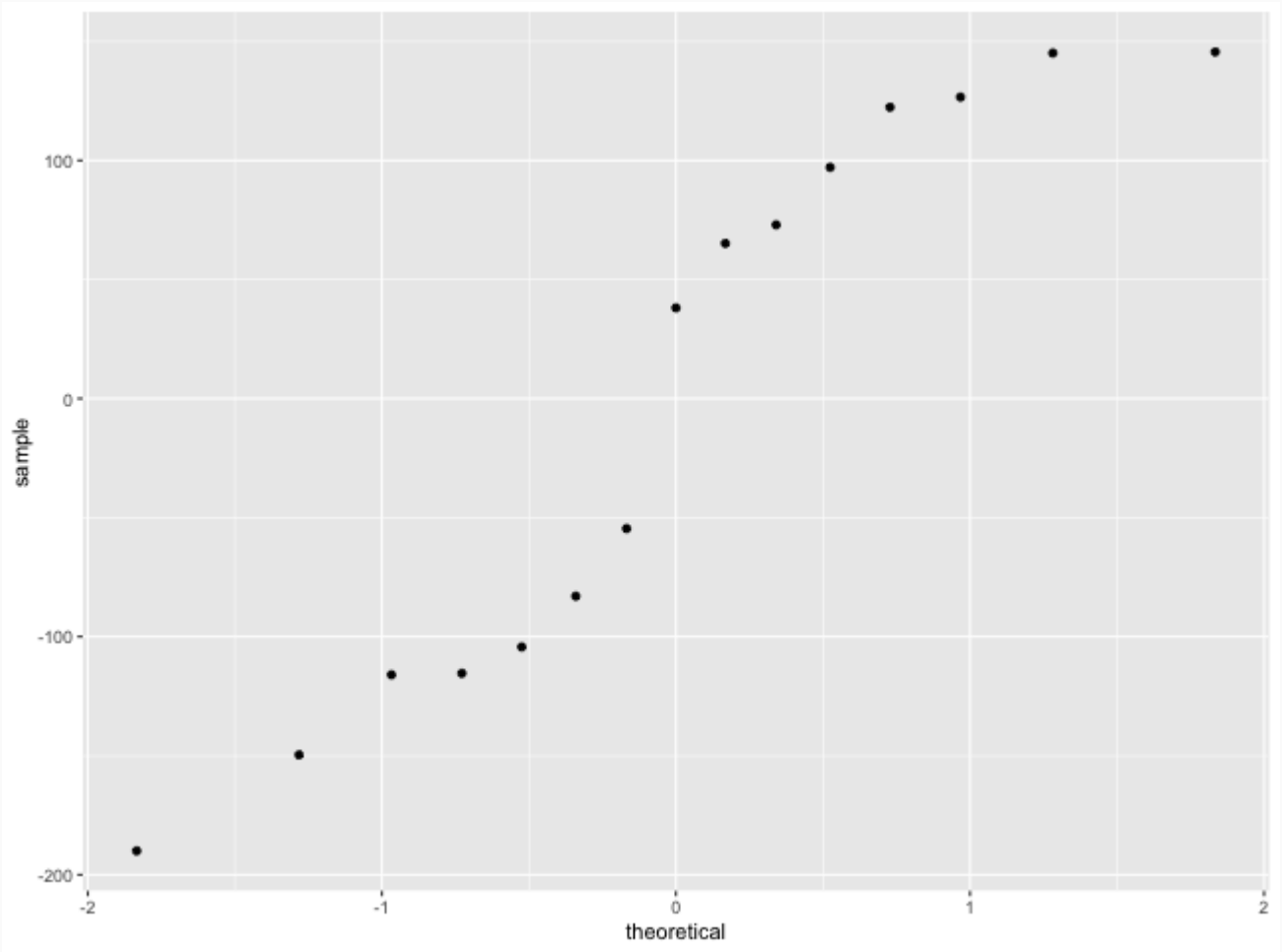
We need to check:

1. Linear trend
2. Independent observations
3. Normal residuals
4. Equal variance

Residual Plot



QQ plot



Q3: Is volume a significant predictor?

```
summary(m1)
```

```
##
## Call:
## lm(formula = weight ~ volume, data = books)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -190.0 -109.9   38.1  109.7  145.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 107.6793    88.3776    1.22     0.24
## volume        0.7086     0.0975    7.27  6.3e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 124 on 13 degrees of freedom
## Multiple R-squared:  0.803,    Adjusted R-squared:  0.787
## F-statistic: 52.9 on 1 and 13 DF,  p-value: 6.26e-06
```

Q4: How much of the variation in weight is explained by the model containing volume?

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:

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

We use the data to estimate our fitted model:

$$\hat{Y} = b_0 + b_1 X_1 + b_2 X_2 + \ldots + b_p X_p$$

Estimating β_0, β_1 etc.

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

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

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

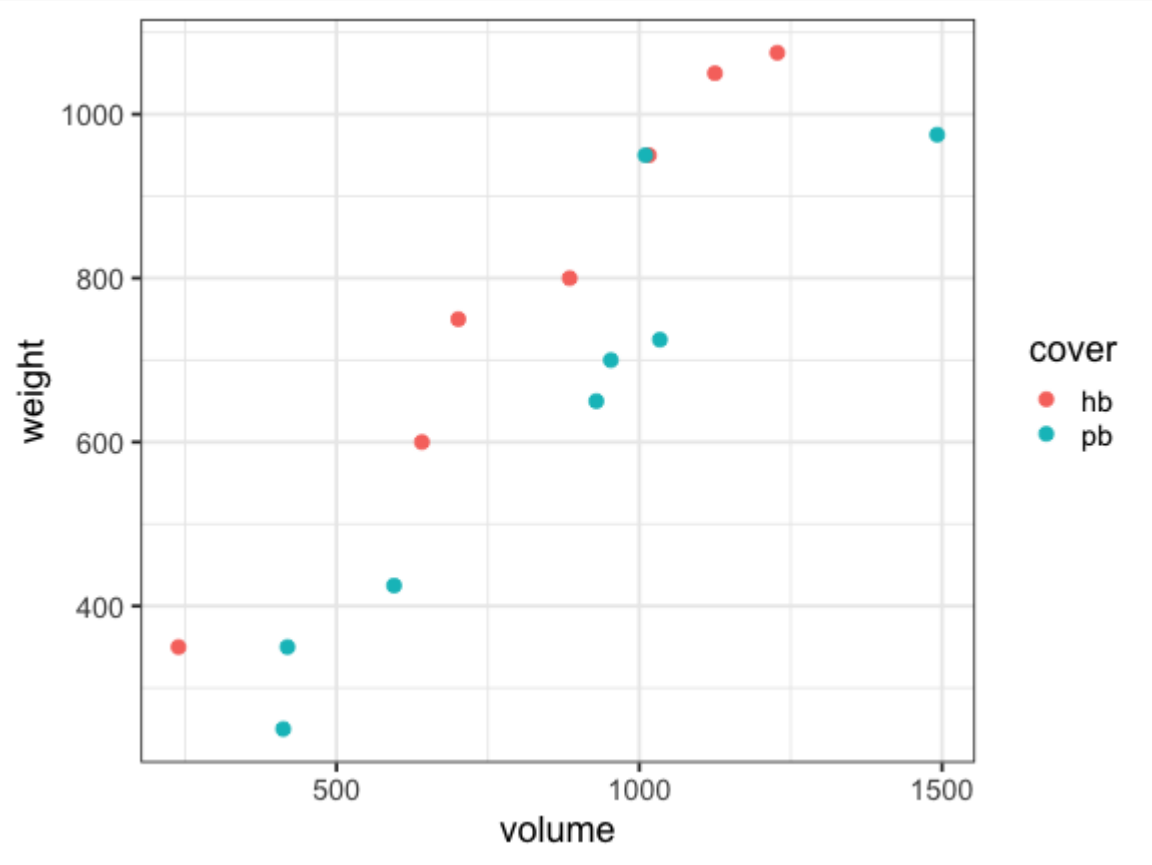
And yes, they have a closed-form solution.

$$\mathbf{b} = (X'X)^{-1}X'Y$$

In R:

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

Example: shipping books



Example: shipping books

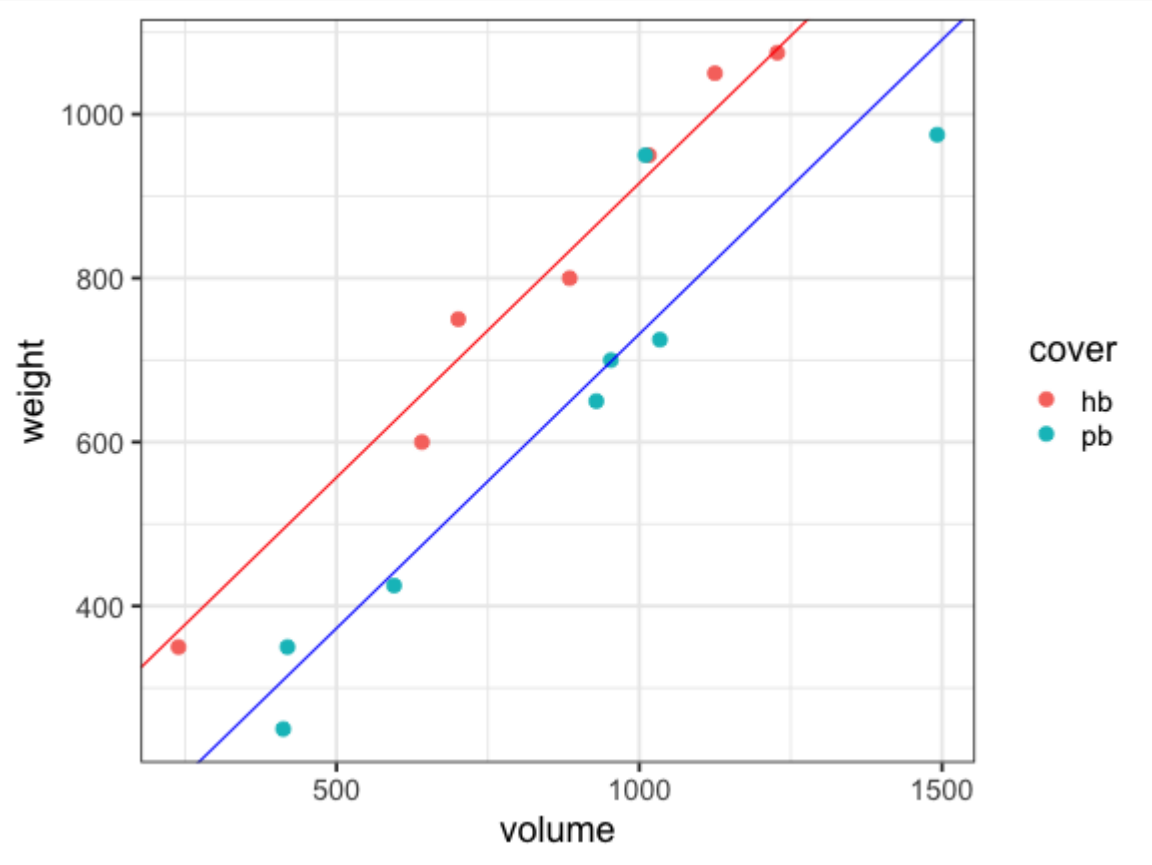
```
m2 <- lm(weight ~ volume + cover, data = books)
summary(m2)
```

```
##
## Call:
## lm(formula = weight ~ volume + cover, data = books)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -110.1  -32.3  -16.1   28.9  210.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  197.9628    59.1927    3.34  0.00584 **
## volume         0.7180     0.0615   11.67  6.6e-08 ***
## coverpb     -184.0473    40.4942   -4.55  0.00067 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0
##
```

How do we interpret these estimates?

boardwork

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 b_i tells you how much you expect the Y to change when you change the X_i, while **holding all other variables constant**.

Your turn 1

```
summary(m2)
```

```
##
## Call:
## lm(formula = weight ~ volume + cover, data = books)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -110.1  -32.3  -16.1   28.9  210.9
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  197.9628    59.1927    3.34  0.00584 **
## volume         0.7180     0.0615   11.67  6.6e-08 ***
## coverpb     -184.0473    40.4942   -4.55  0.00067 ***
## ---
## 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:  0.927,    Adjusted R-squared:  0.915
## F-statistic: 76.7 on 2 and 12 DF,  p-value: 1.45e-07
```

- Is the difference between cover types significant?

Your turn 2

```
summary(m2)$coef
```

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  197.963    59.1927    3.34 5.84e-03
## volume         0.718     0.0615   11.67 6.60e-08
## coverpb     -184.047    40.4942   -4.55 6.72e-04
```

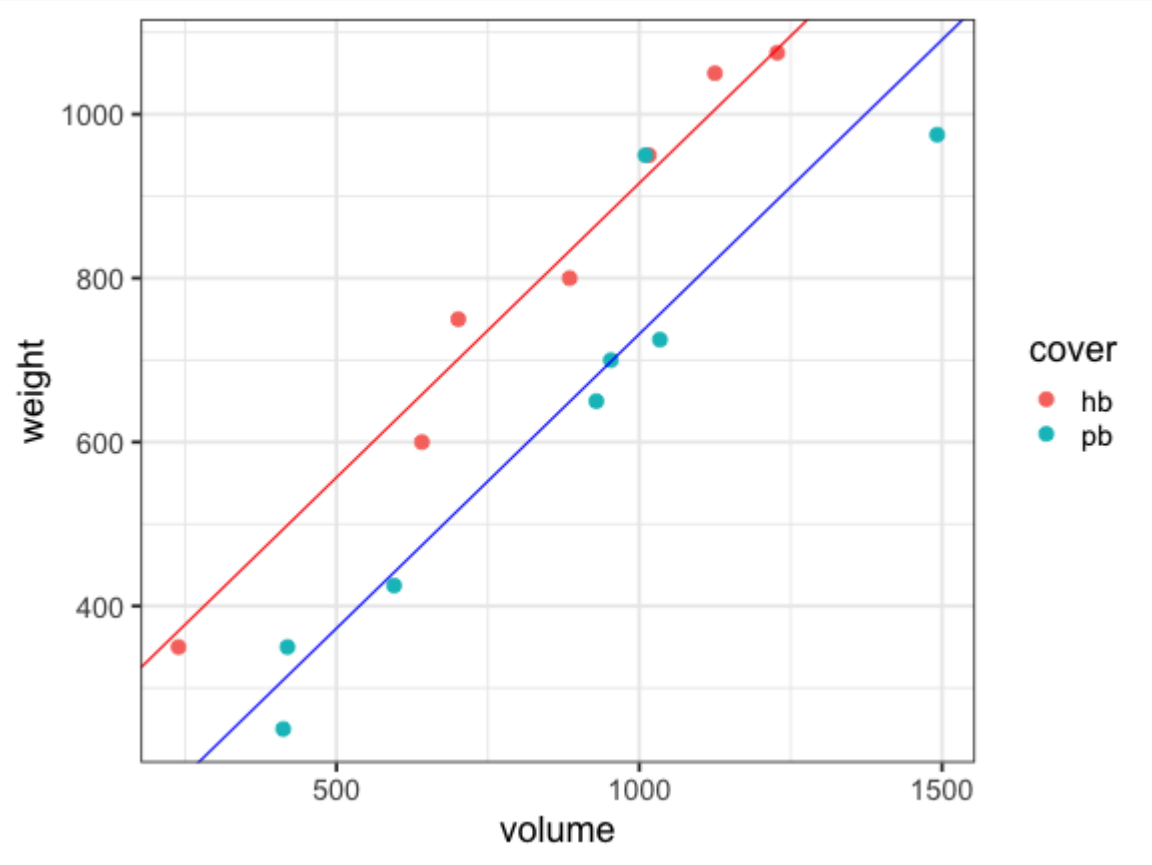
```
qt(.025, df = nrow(books) - 3)
```

```
## [1] -2.18
```

Which of the follow represents the appropriate 95% CI for `coverpb`?

1. $197 \pm 1.96 \times 59.19$
2. $-184 \pm 2.18 \times 40.5$
3. $-184 \pm -4.55 \times 40.5$

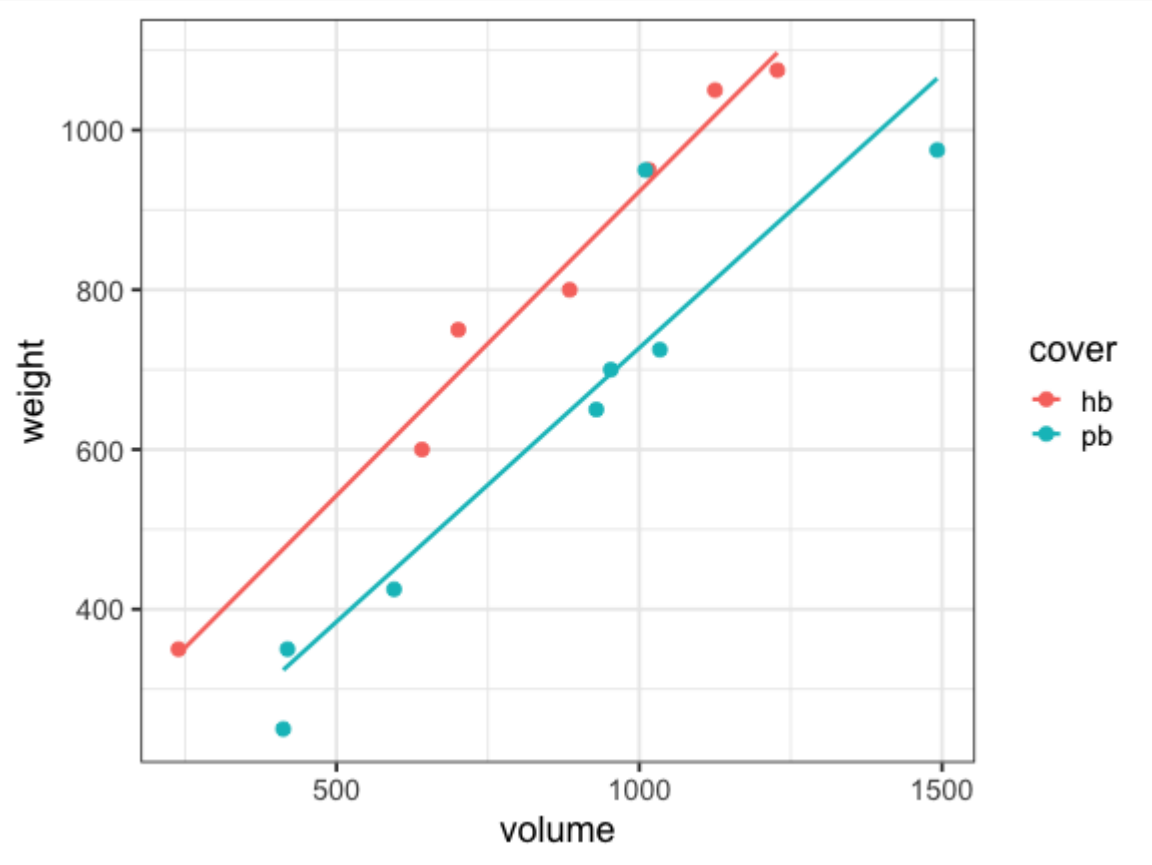
Extending the model



The two cover types have different intercepts. Do they share the same slope?

boardwork

Extending the model



Interaction terms

```
m3 <- lm(weight ~ volume + cover + volume:cover,
         data = books)
summary(m3)
```

```
##
## Call:
## lm(formula = weight ~ volume + cover + volume:cover, data = books)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##  -89.7  -32.1  -21.8   17.9  215.9
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)     161.5865    86.5192    1.87    0.089 .
## volume            0.7616     0.0972    7.84  7.9e-06 ***
## coverpb        -120.2141   115.6590   -1.04    0.321
## volume:coverpb   -0.0757     0.1280   -0.59    0.566
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 80.4 on 11 degrees of freedom
## Multiple R-squared:  0.93,    Adjusted R-squared:  0.911
## F-statistic: 48.5 on 3 and 11 DF,  p-value: 1.24e-06
```

Take home messages

- There is a statistically significant relationship between volume and weight.
- There is a statistically significant difference in weight between paperback and hardcover books, when controlling for volume.
- There is no strong evidence that the relationship between volume and weight differs betwen paperbacks and hardbacks.

This is **inference**, which required **valid models**. We'll check diagnostics next time.