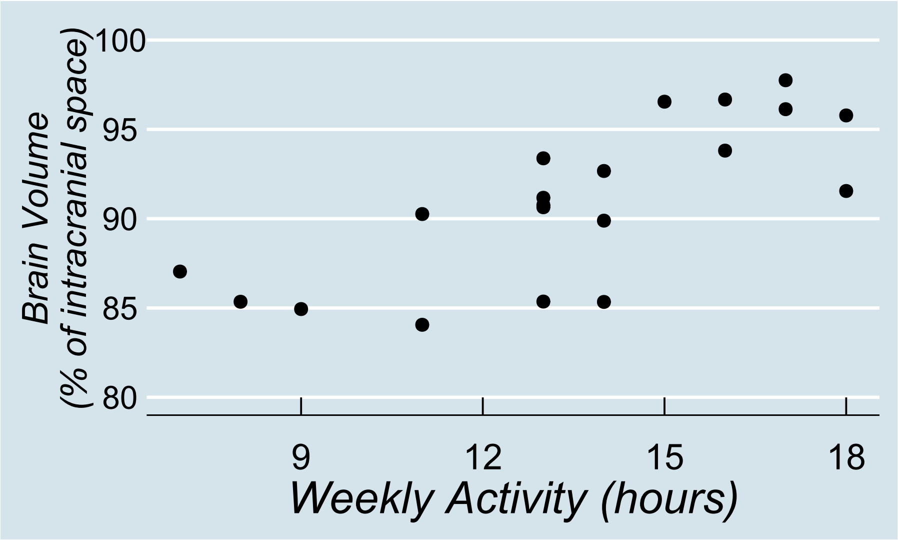
Week 7: The Linear Model

Univariate Statistics and Methodology using R

Department of Psychology
The University of Edinburgh

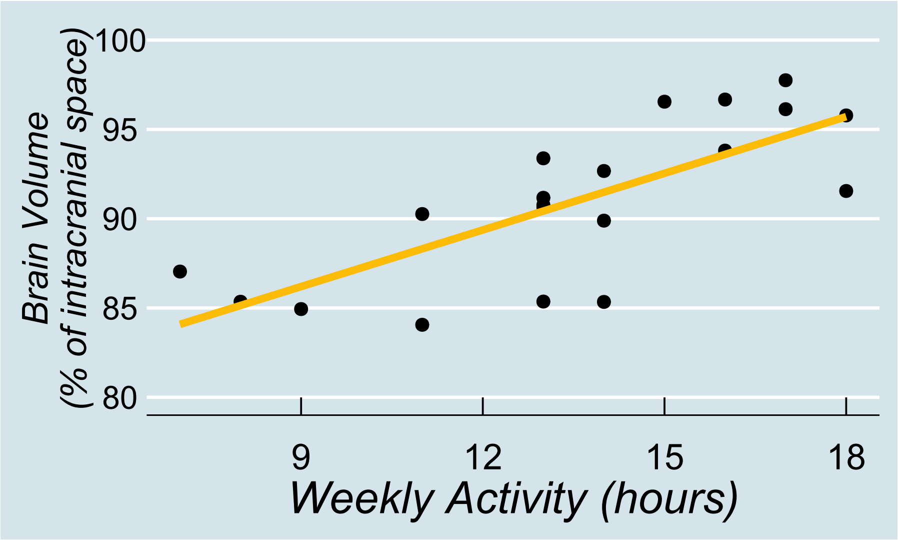
Part 1: Correlation++

Exercising our brains



$$r = 0.7488, p = 0.0001$$

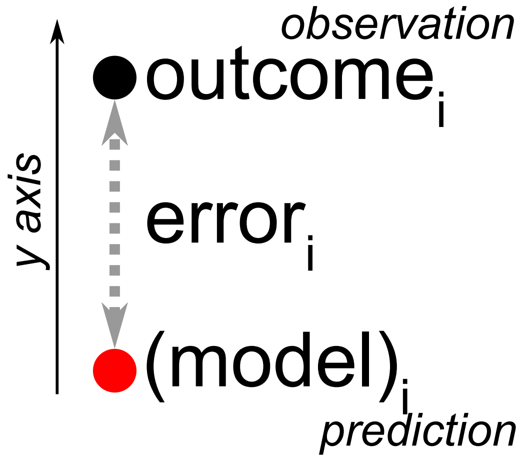
Exercising our brains (2)



"for every extra 1 hour more weekly activity, brain volume increases by 1.06 (% of intracranial space)"

The Only Equation You Will Ever Need

$$\text{outcome}_i = (\text{model})_i + \text{error}_i$$



The Only Equation You Will Ever Need

$$\text{outcome}_i = (\text{model})_i + \text{error}_i$$

- to get any further, we need to make *assumptions*
- nature of the **model**

(linear)

- nature of the **errors**

(normal)

A Linear Model

$$\text{outcome}_i = (\text{model})_i + \text{error}_i$$

$$y_i = b_0 \cdot 1 + b_1 \cdot x_i + \epsilon_i$$

so the linear model itself is...

$$\hat{y}_i = b_0 \cdot 1 + b_1 \cdot x_i$$

$$\text{y} \sim 1 + \text{x}$$

A Linear Model

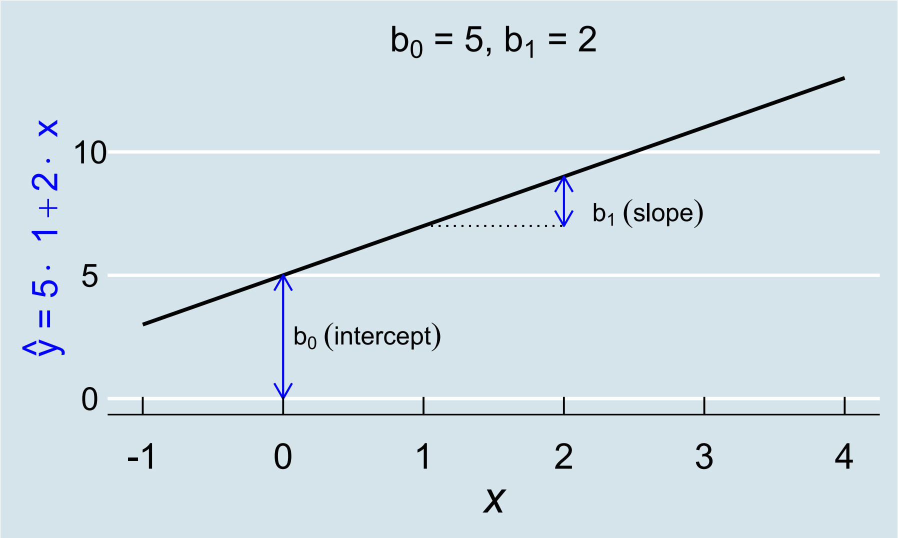
$$\text{outcome}_i = (\text{model})_i + \text{error}_i$$

$$y_i = b_0 \cdot 1 + b_1 \cdot x_i + \epsilon_i$$

so the linear model itself is...

$$\hat{y}_i = b_0 \cdot 1 + b_1 \cdot x_i$$

$$y \sim 1 + x$$



A Linear Model

$$\text{outcome}_i = (\text{model})_i + \text{error}_i$$

$$y_i = b_0 \cdot 1 + b_1 \cdot x_i + \epsilon_i$$

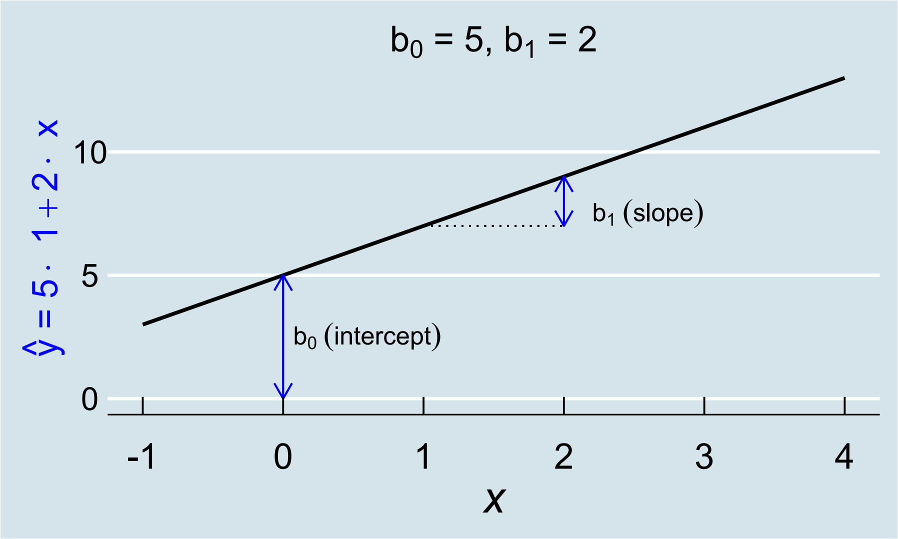
so the linear model itself is...

$$\hat{y}_i = b_0 \cdot 1 + b_1 \cdot x_i$$

$$\text{y} \sim 1 + \text{x}$$

$$\hat{y} = b_0 + b_1 \cdot x_i$$

$$\text{y} \sim \text{x}$$

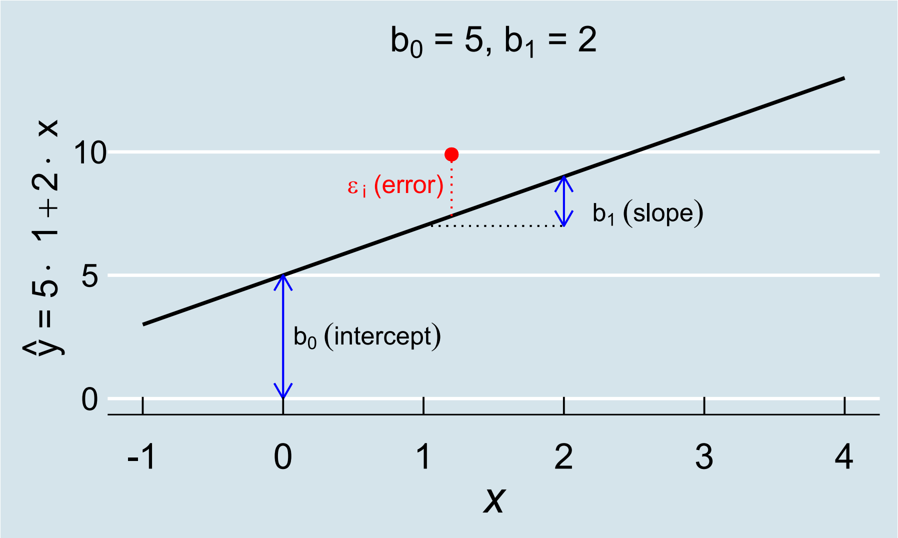


Take An Observation

$x_i = 1.2, y_i = 9.9$

$$\hat{y}_i = b_0 + b_1 \cdot x_i = 7.4$$

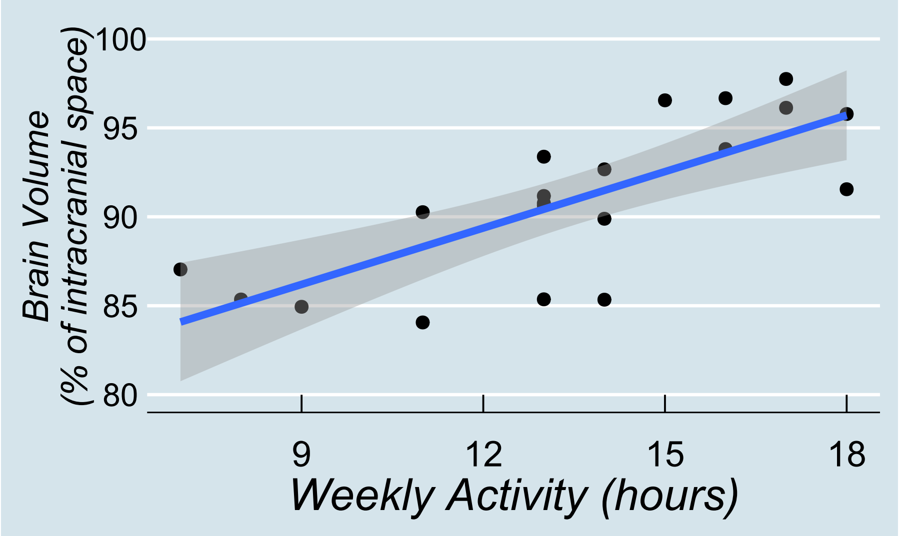
$$y_i = \hat{y}_i + \epsilon_i = 7.4 + 2.5$$



More Brain Exercises

"for every extra 1 hour more weekly activity, brain volume increases by 1.06 (% of intracranial space)"

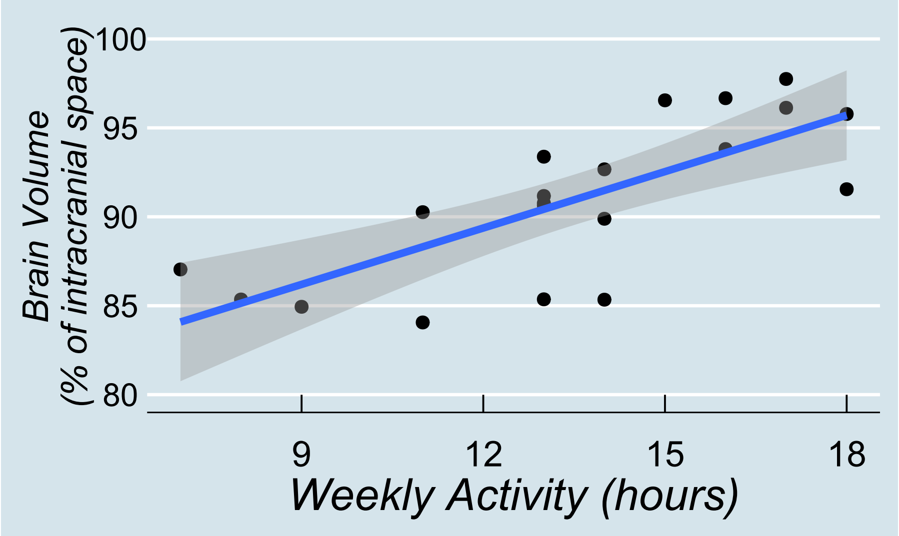
```
+ geom_smooth(method="lm")
```



More Brain Exercises

"for every extra 1 hour more weekly activity, brain volume increases by 1.06 (% of intracranial space)"

```
+ geom_smooth(method="lm")
```



but how can we evaluate our model?

Linear Models in R

```
mod <- lm(brain_vol ~ weekly_actv, data=dat)
```

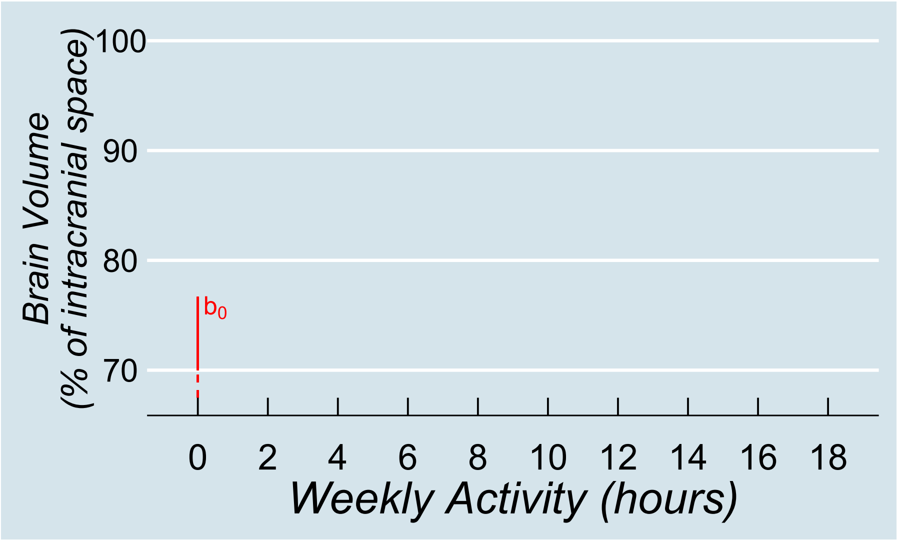
Linear Models in R

```
mod <- lm(brain_vol ~ weekly_actv, data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = brain_vol ~ weekly_actv, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.144 -1.342  0.274  2.199  4.009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.685      3.052   25.12  1.8e-15 ***
## weekly_actv    1.057      0.221    4.79  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.02 on 18 degrees of freedom
## Multiple R-squared:  0.561,    Adjusted R-squared:  0.536
## F-statistic:   23 on 1 and 18 DF,  p-value: 0.000145
```

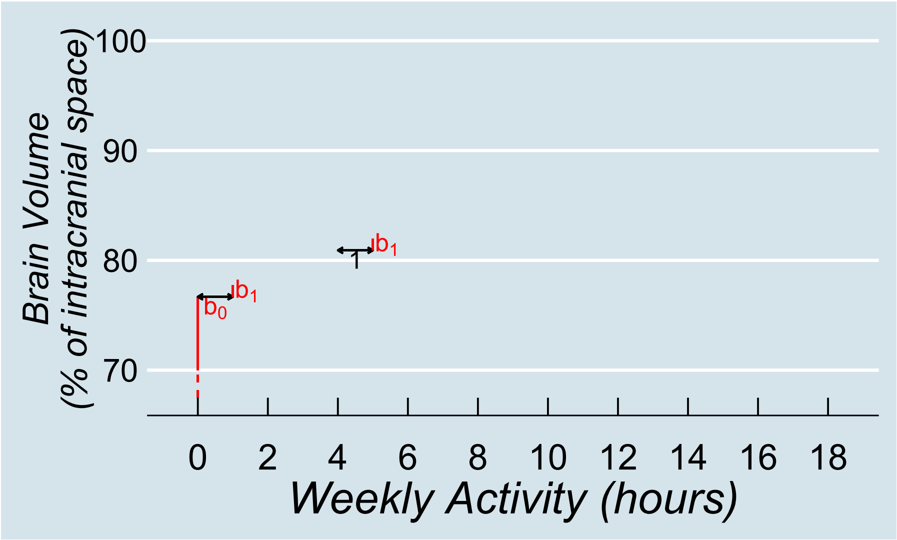
Intercept and Slope Again

$$b_0 = 76.7; \quad b_1 = 1.06$$



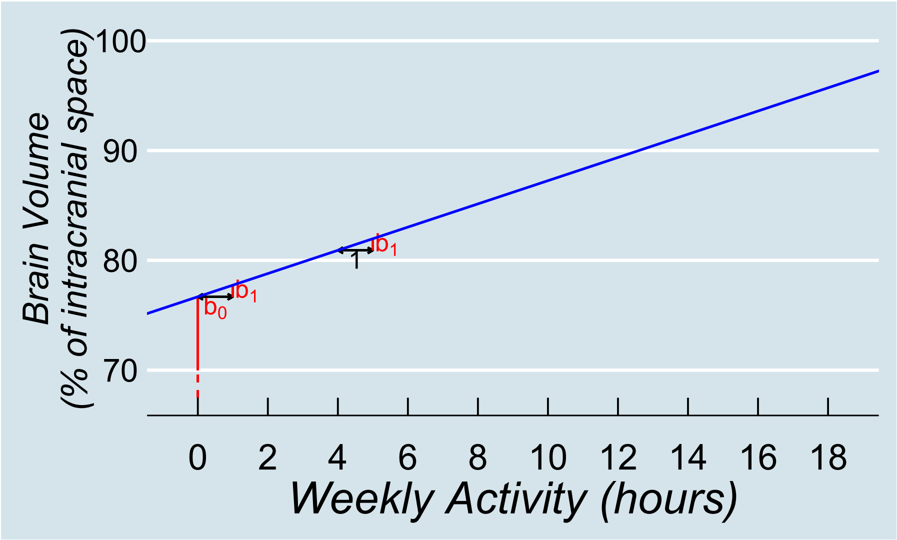
Intercept and Slope Again

$$b_0 = 76.7; \quad b_1 = 1.06$$



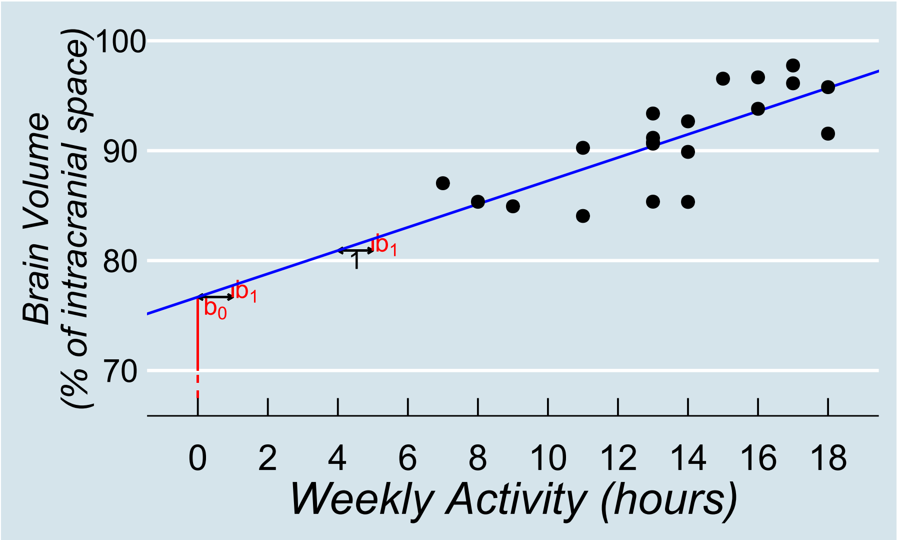
Intercept and Slope Again

$$b_0 = 76.7; \quad b_1 = 1.06$$



Intercept and Slope Again

$$b_0 = 76.7; \quad b_1 = 1.06$$



End of Part 1

Part 2

Significance

Intercept and Slope

```
mod <- lm(brain_vol ~ weekly_actv, data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = brain_vol ~ weekly_actv, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.144 -1.342  0.274  2.199  4.009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.685      3.052   25.12  1.8e-15 ***
## weekly_actv    1.057      0.221    4.79  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.02 on 18 degrees of freedom
## Multiple R-squared:  0.561,    Adjusted R-squared:  0.536
## F-statistic:   23 on 1 and 18 DF,  p-value: 0.000145
```

Are We Impressed?

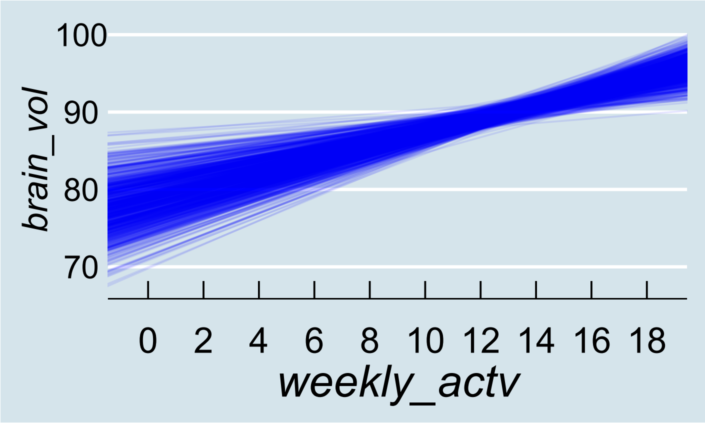
- we have an intercept of 76.7 and a slope of 1.06
- in NHST world, our pressing question is

Are We Impressed?

- we have an intercept of 76.7 and a slope of 1.06
- in NHST world, our pressing question is

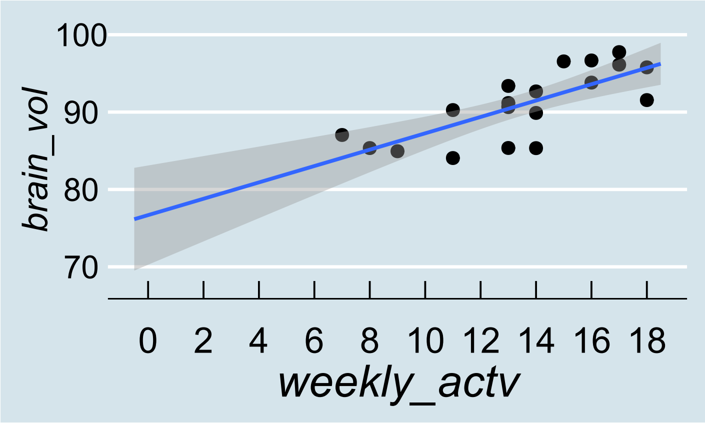
> how likely would we have been to find these parameters under the null hypothesis?

Testing Chance



- repeatedly sampling 20~datapoints from the population
  - variability in *height* of line = variability in intercept (b_0)
  - variability in *angle* of line = variability in slope (b_1)

We've Seen This Before



- shaded area represents "95% confidence interval"
  - if we repeatedly sampled 20 items from the population...
  - assumes that the 20 we have are the *best estimate* of the population

The Good Old *t*-Test

```
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.685      3.052   25.12  1.8e-15 ***
## weekly_actv    1.057      0.221    4.79  0.00015 ***
```

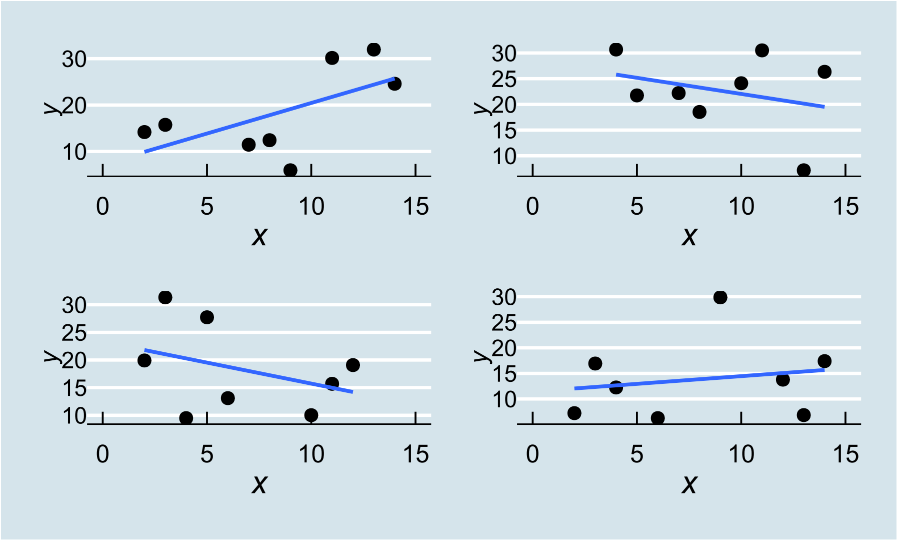
- for each model parameter we are interested in whether it is *different from zero*
- **intercept**: just like a mean
- **slope**: does the best-fit line differ from horizontal?

The Good Old *t*-Test

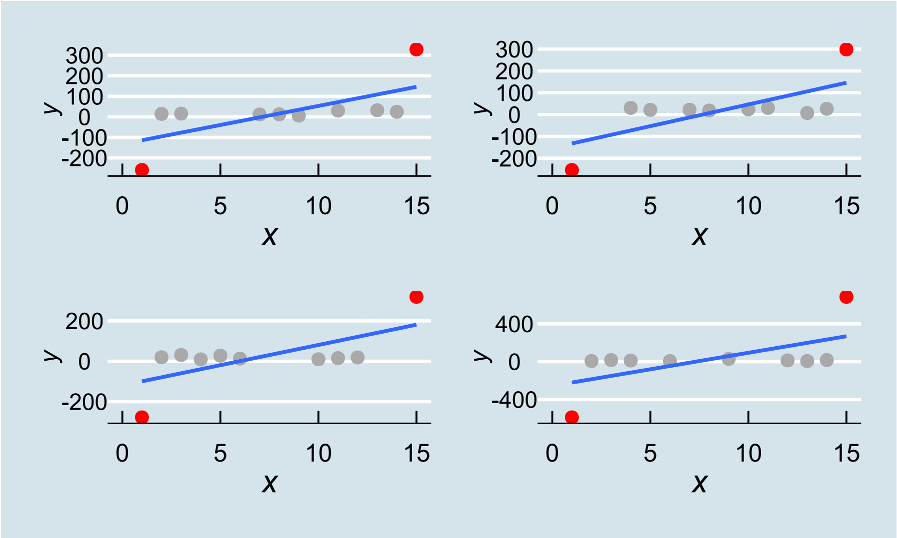
```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    76.685      3.052   25.12  1.8e-15 ***
## weekly_actv     1.057      0.221    4.79  0.00015 ***
```

- for each model parameter we are interested in whether it is *different from zero*
- intercept: just like a mean
- slope: does the best-fit line differ from horizontal?
- these are just (two-tailed) one-sample t-tests
  - standard error is the standard deviation of doing these lots of times
  - t value is $\frac{\text{Estimate}}{\text{Std. Error}}$
  - to calculate p, we need to know the *degrees of freedom*

Degrees of Freedom



Degrees of Freedom



Degrees of Freedom

- in fact we subtract 2 degrees of freedom because we "know" two things
    - intercept (b_0)
    - slope (b_1)
- the remaining degrees of freedom are the *residual* degrees of freedom

Degrees of Freedom

- in fact we subtract 2 degrees of freedom because we "know" two things
  - intercept (b_0)
  - slope (b_1)
- the remaining degrees of freedom are the *residual* degrees of freedom
- the *model* also has associated degrees of freedom
  - 2 (intercept, slope) - 1 (knowing one affects the other)

the models we have been looking at have 20 observations and 1 predictor

(1, 18) degrees of freedom

Linear Models in R

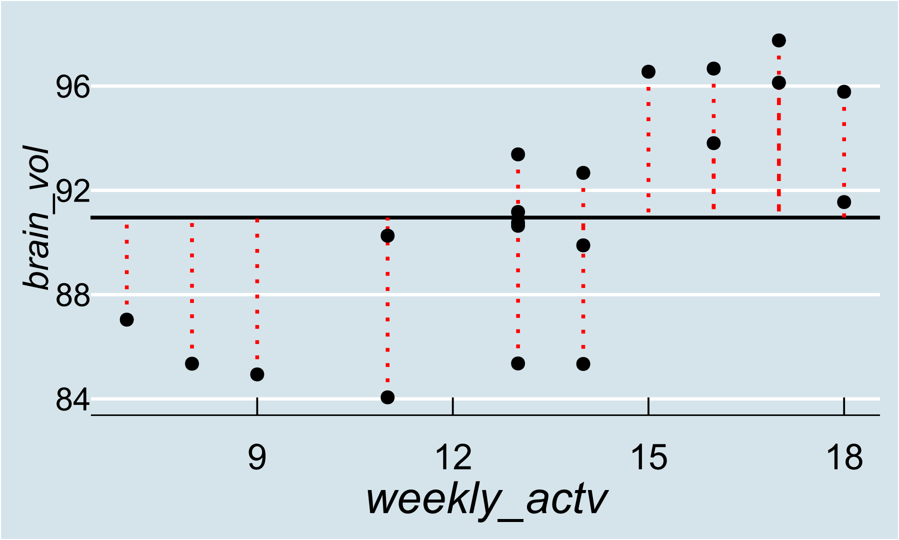
```
mod <- lm(brain_vol ~ weekly_actv, data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = brain_vol ~ weekly_actv, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.144 -1.342  0.274  2.199  4.009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.685      3.052   25.12  1.8e-15 ***
## weekly_actv    1.057      0.221    4.79  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.02 on 18 degrees of freedom
## Multiple R-squared:  0.561,    Adjusted R-squared:  0.536
## F-statistic:   23 on 1 and 18 DF,  p-value: 0.000145
```

Total Sum of Squares

$$\text{total SS} = \sum (y - \bar{y})^2$$

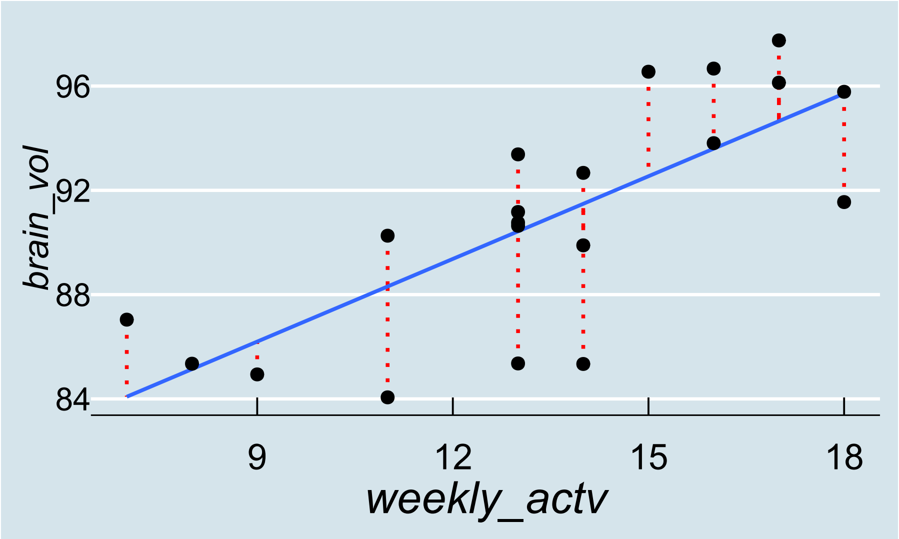
- sum of squares between observed y and mean $\bar{y}$
- represents the total amount of variance in the model
- how much does the observed data vary from a model which says "there is no effect of x" (**null model**)?



Residual Sum of Squares

$$\text{residual SS} = \sum (y - \hat{y})^2$$

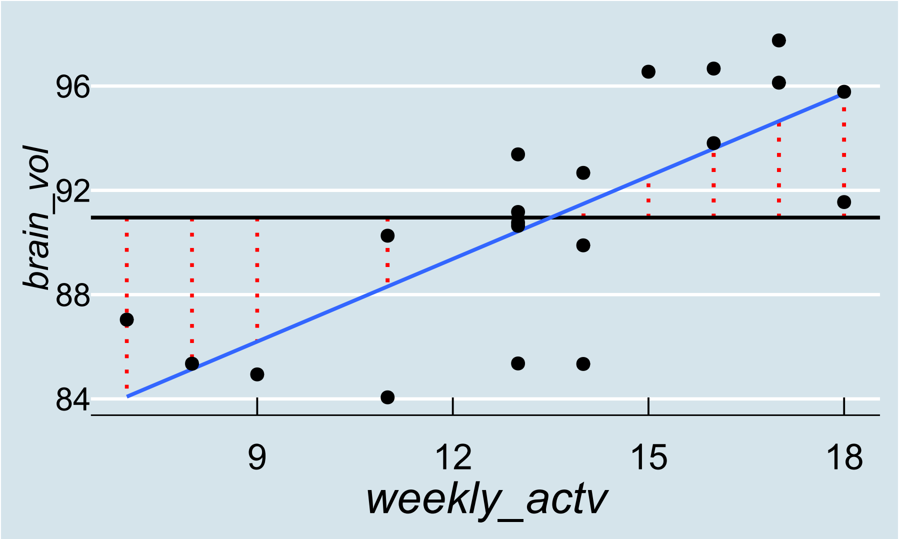
- sum of squared differences between observed y and predicted $\hat{y}$
- represents the unexplained variance in the model
- how much does the observed data vary from the existing model?



Model Sum of Squares

$$\text{model SS} = \sum (\hat{y} - \bar{y})^2$$

- sum of squared differences between predicted $\hat{y}$ and mean $\bar{y}$
- represents the additional variance explained by the current model over the null model

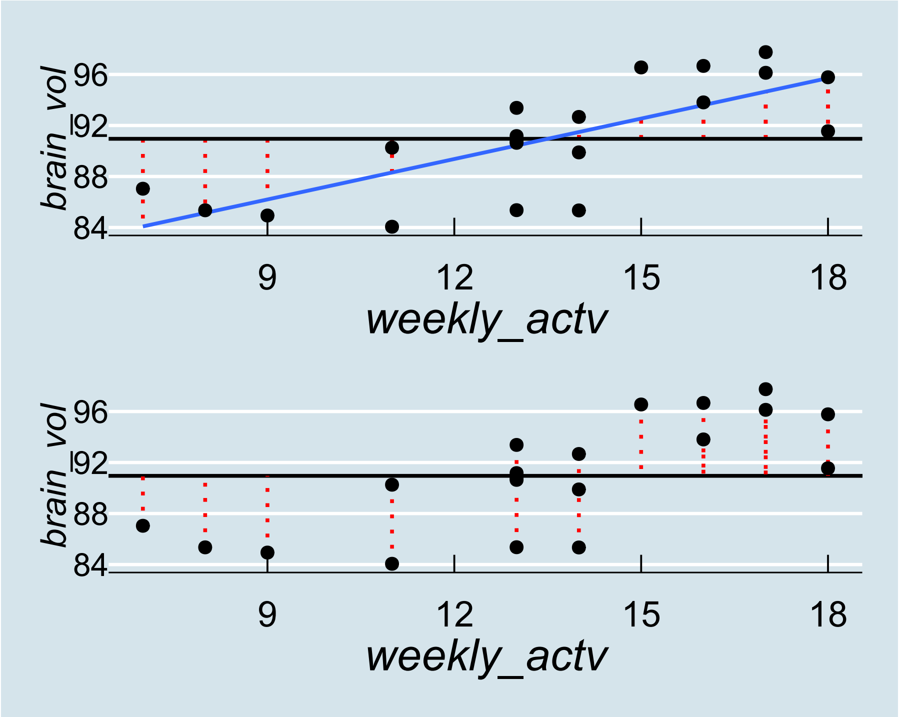


Testing the Model: R^2

$$R^2 = \frac{\text{model SS}}{\text{total SS}} = \frac{\sum(\hat{y} - \bar{y})^2}{\sum(y - \bar{y})^2}$$

"how much the model improves over the null"

- $0 \leq R^2 \leq 1$
- we want R^2 to be large
- for a single predictor, $\sqrt{R^2} = |r|$



Testing the Model: *F*

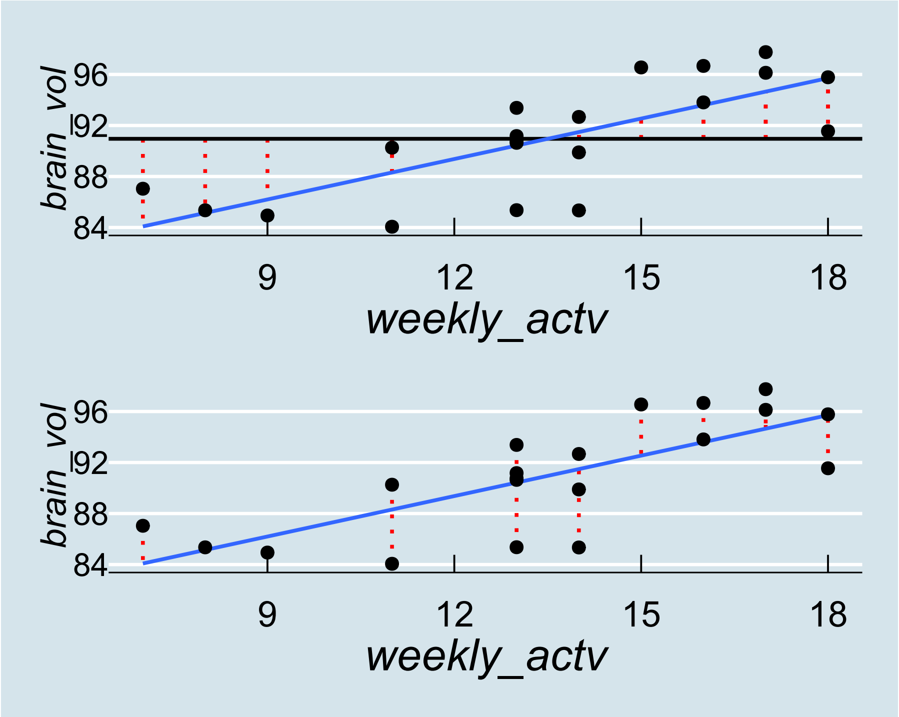
F ratio depends on mean squares

$$(\,\mathrm{MS}_x = \mathrm{SS}_x/\mathrm{df}_x\,)$$

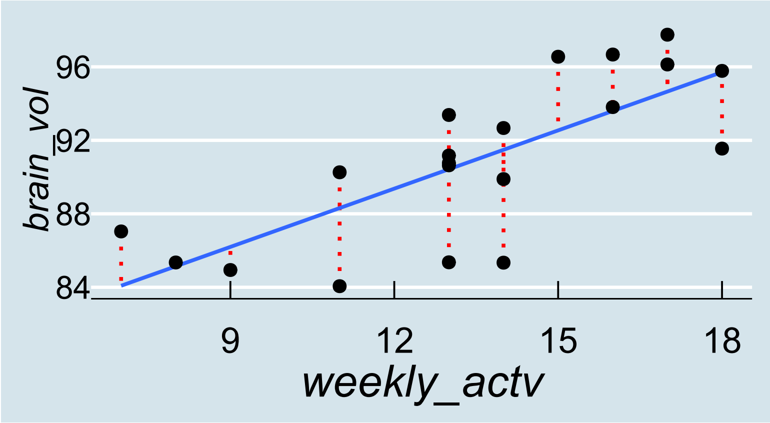
$$F = \frac{\text{model MS}}{\text{residual MS}} = \frac{\sum(\hat{y}-\bar{y})^2/\mathrm{df}_m}{\sum(y-\hat{y})^2/\mathrm{df}_r}$$

"how much the model improves over chance"

- $0 < F$
- we want F to be large
- significance of F does not always equate to a large (or theoretically sensible) effect



A Linear Model for 20 Brains



- a linear model describes the **best-fit line** through the data
- minimises the error terms ϵ or **residuals**

Two Types of Significance

```
mod <- lm(brain_vol ~ weekly_actv, data=dat)
summary(mod)
```

```
##
## Call:
## lm(formula = brain_vol ~ weekly_actv, data = dat)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.144 -1.342  0.274  2.199  4.009
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   76.685      3.052   25.12  1.8e-15 ***
## weekly_actv    1.057      0.221    4.79  0.00015 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.02 on 18 degrees of freedom
## Multiple R-squared:  0.561,	Adjusted R-squared:  0.536
## F-statistic:   23 on 1 and 18 DF,  p-value: 0.000145
```

The Good, the Bad, and the Ugly

- we can easily extend this approach
  - use more than one predictor
  - generalised linear model

- not a panacea
  - depends on *assumptions* about the data
  - depends on *decisions* about analysis

The Good, the Bad, and the Ugly

- we can easily extend this approach
  - use more than one predictor
  - generalised linear model

- not a panacea
  - depends on *assumptions* about the data
  - depends on *decisions* about analysis

- like other statistics, linear models don't tell you "about" your data
- they simply assess what is (un)likely to be due to chance
- the key to good statistics is *common sense and good interpretation*

End