# Regression

Evaluation Methods & Statistics Lecture 6

# What will be covered

- Hypothesis Testing
- Types of data
- P value revisited
- Regression
  - What it does
  - How it does it

# Our Research Question

- People may be anxious about posting tweets due to judgments from others
- This anxiety might explain why some twitter users do not post frequently
- Does posting anxiety predict the number of  twitter posts?

# The Scientific Process

1. Generate a hypothesis
2. Design experiment/study to test the hypothesis
3. Collect data from sample
4. Fit statistical model to the data
5. Assess how well this model represents the data (the model fit)

# Hypothesis testing

The Null Hypothesis (H0)

Posting anxiety **does not** predict twitter posting

The Research Hypothesis(H1)

Posting anxiety **significantly predicts** twitter posting

# Hypothesis testing

- H0 given more weight
  - Experiment run to disprove H0 => will not reject it unless evidence is sufficiently strong
  - Discount the simple before adopting something more complex (Occam's razor)

# Types of Data- Categorical

- Nominal
  - Data ascribing objects or values to distinct categories
    - Eg. High, Low
- Ordinal
  - Nominal data with explicit order/ranks
    - 1st, 2nd, 3rd
- Although we know category and order we don't know the quantity of difference between values

# Types of Data- Continuous

- Interval
  - Equal intervals in the scale
  - E.g. 5 point Likert Scale

- Ratio
  - Equal intervals with a true 0 point
    - E.g. Reaction Time
    - Distance along scale should be divisible
    - X on scale should equal 2x/2

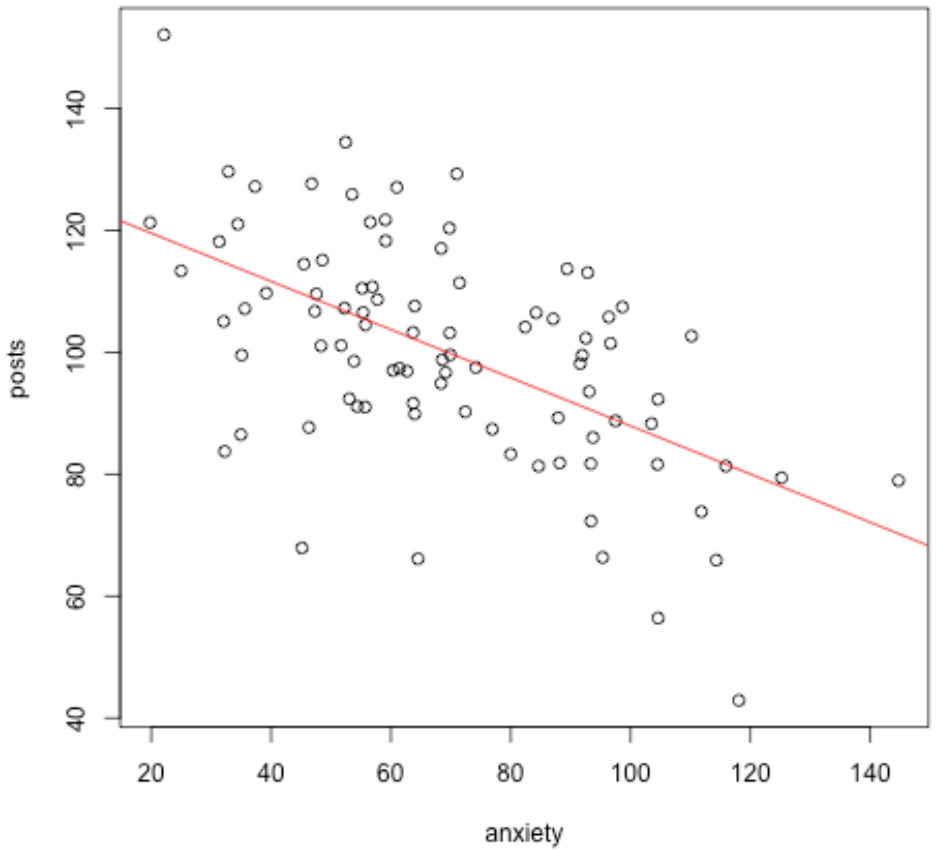
# What is Regression?

- Significance of a predictor variable **(posting anxiety)** on outcome variable **(twitter posts)**
- Predictor can be continuous or categorical
- Allows us to predict future posts based on knowledge of predictor value
- **Here we are looking at linear regression (straight line model)**

# Equation of Straight Line

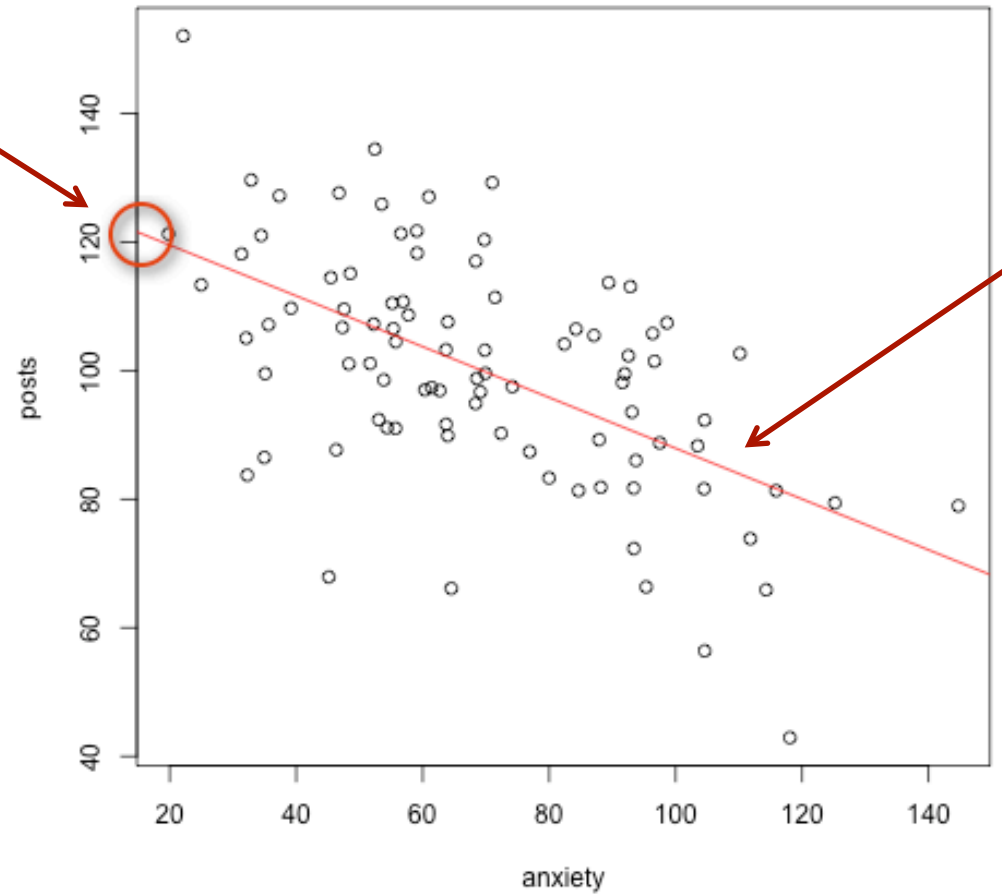
- y= mx +c --- Familiar?
- $Y_i = \beta_1 X_i + \beta_0 + \varepsilon_i$
  - $Y_i$= Outcome Variable value
  - $\beta_0$= intercept
  - $\beta_1$= gradient of regression line
  - $X_i$= Participant $_i$ score on our predictor variable
  - $\varepsilon_i$ =residual difference between score predicted for $Y_i$ and the one actually attained in the data

# Our Data



# Our Data

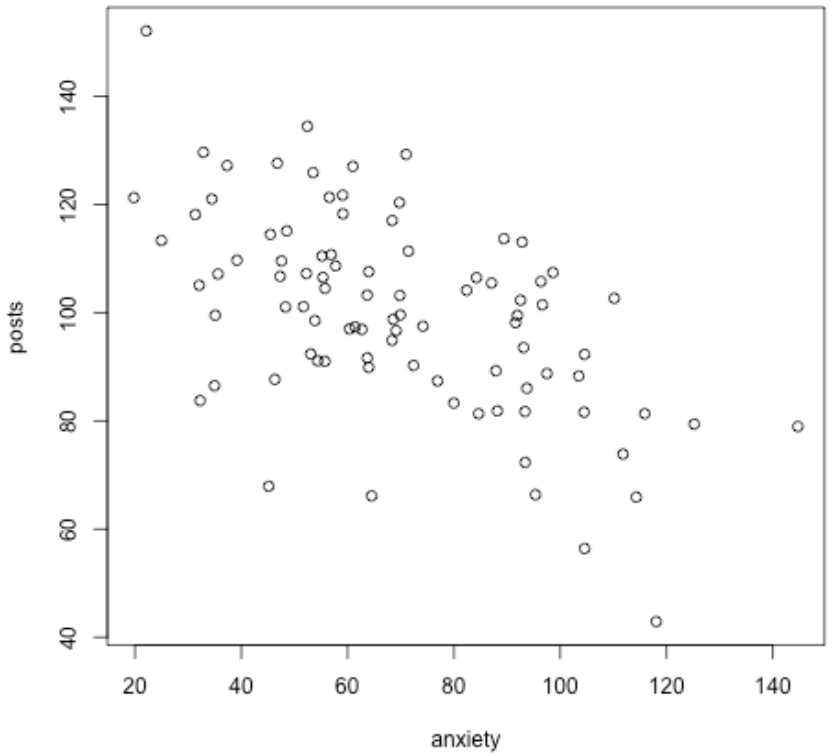
Intercept
$\beta_0$




Gradient of the line- $\beta_1$

Change in *y* with a unit change in x

# How do we fit this model?

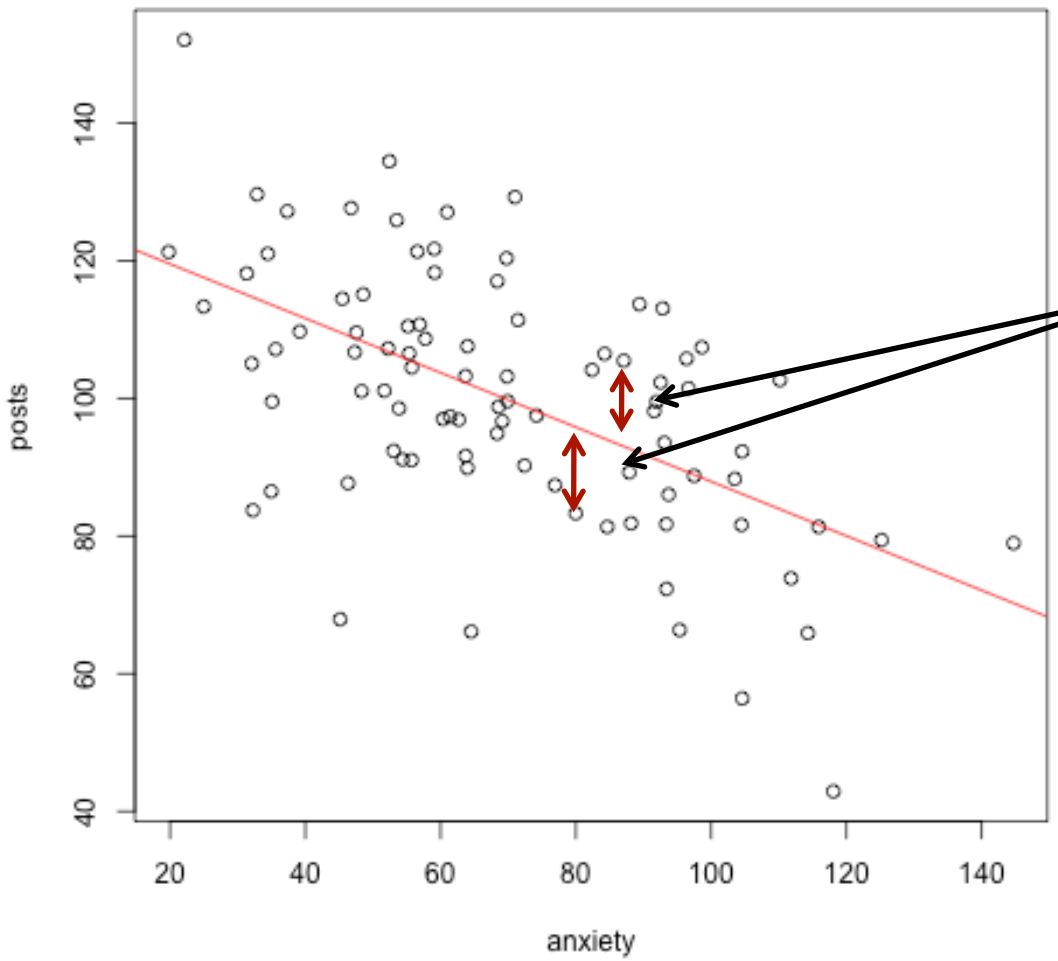



- Data looks linear
- We could just draw a line we feel is the *best fit*
- Very subjective
- Wouldn't be sure it was the best fit
- Use technique called **method of least squares**

# Method of Least Squares

- Residual difference between the line and the actual data point
- The line of best fit (regression line)= line that leads to minimum residual
- The residuals are squared and summed (SS) as some will be negative some positive.

# Our Data

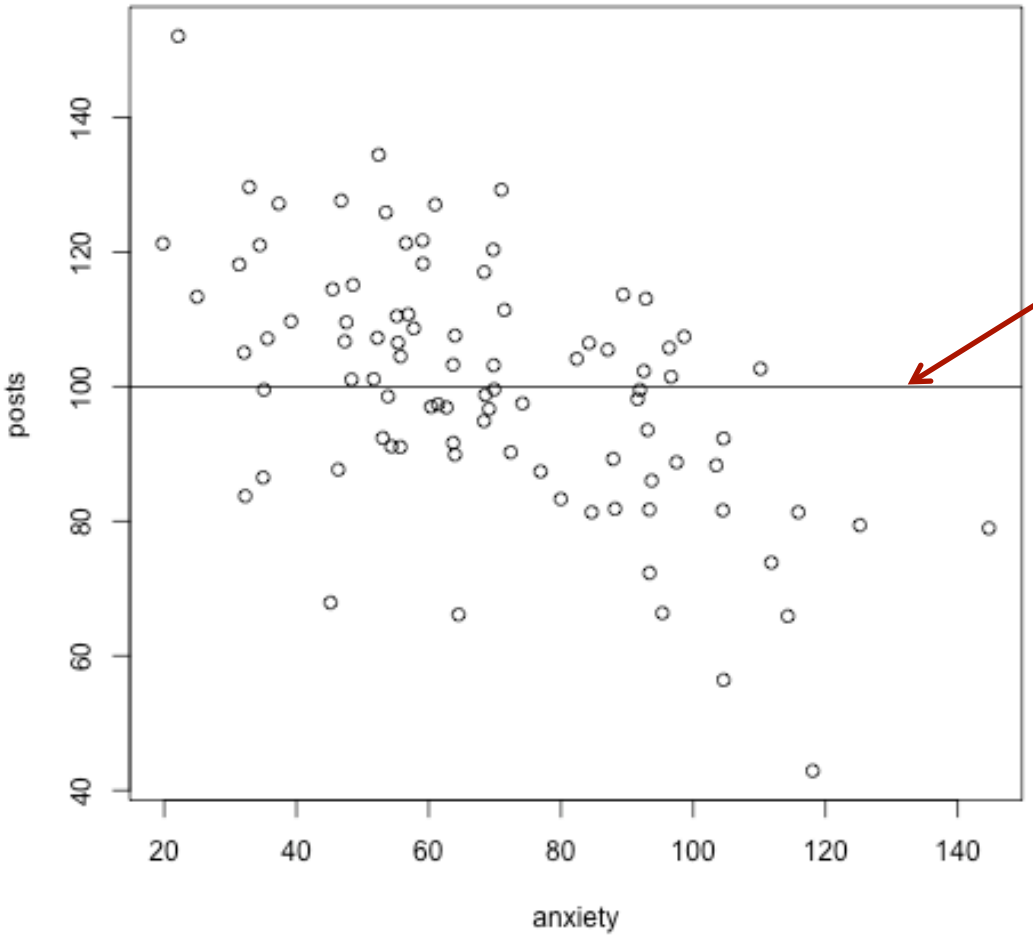



Residuals

# Is it the best model?

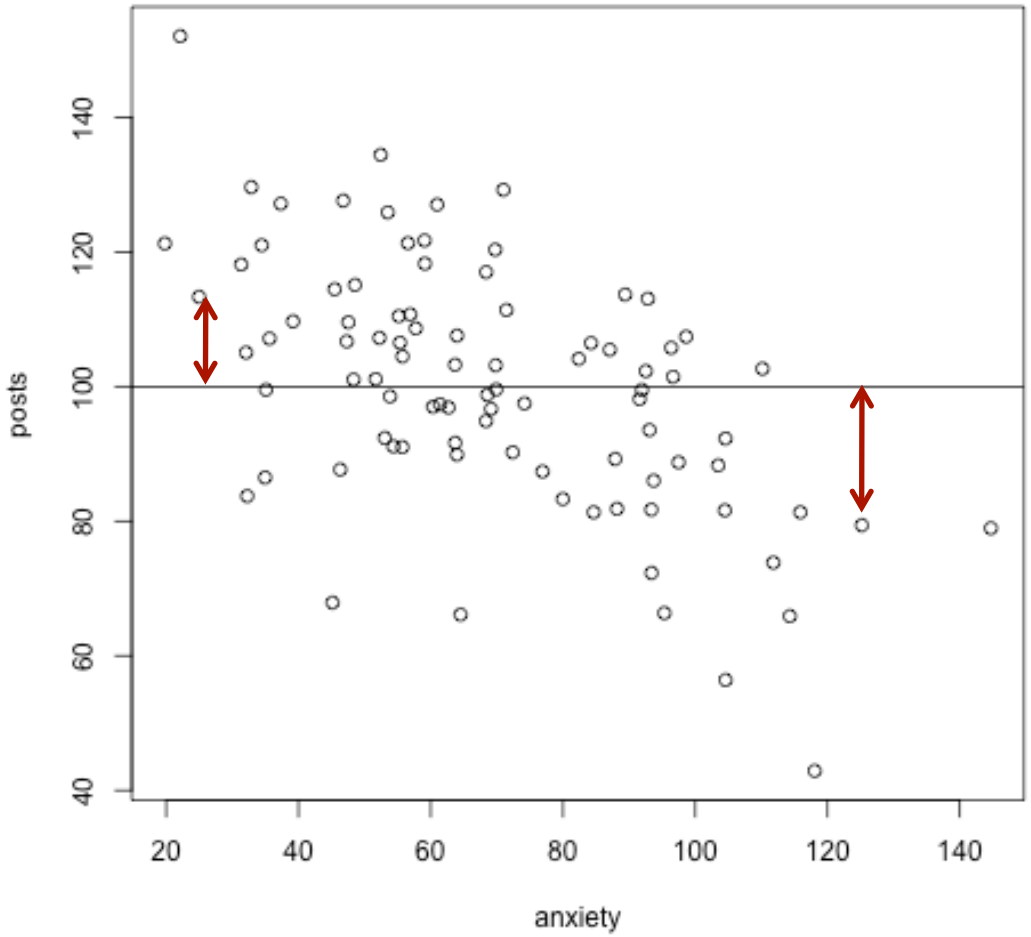
- Although it may be the line of best fit available, it might still be a poor description of the data
- We need to find out how adequate the best fit model is
- We therefore compare the best fit model to the most basic model (the mean) using Sum of Squares

# Our Data




Mean of y-
Simplest Model

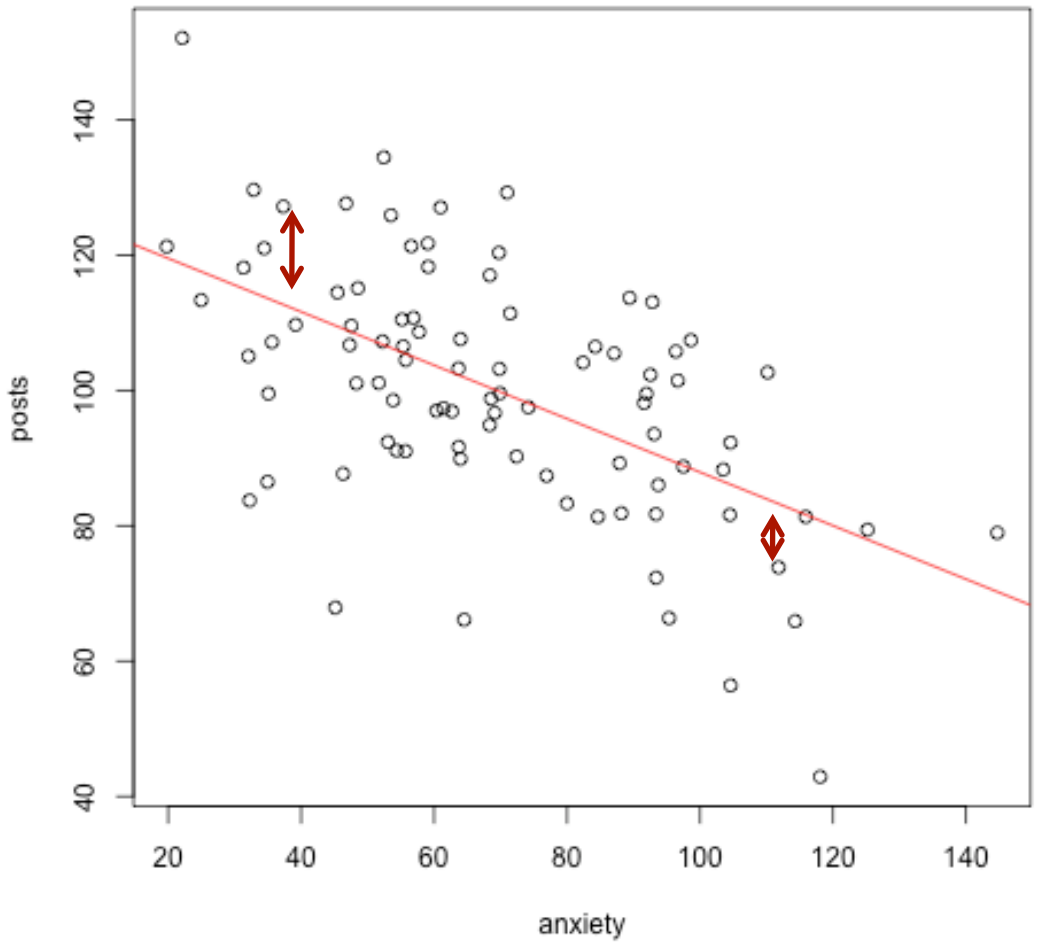
# Total Sum of Squares (SST)




SST= $\Sigma$ (Observed – Mean Y)$^2$

Total amount of differences present when simplest model applied

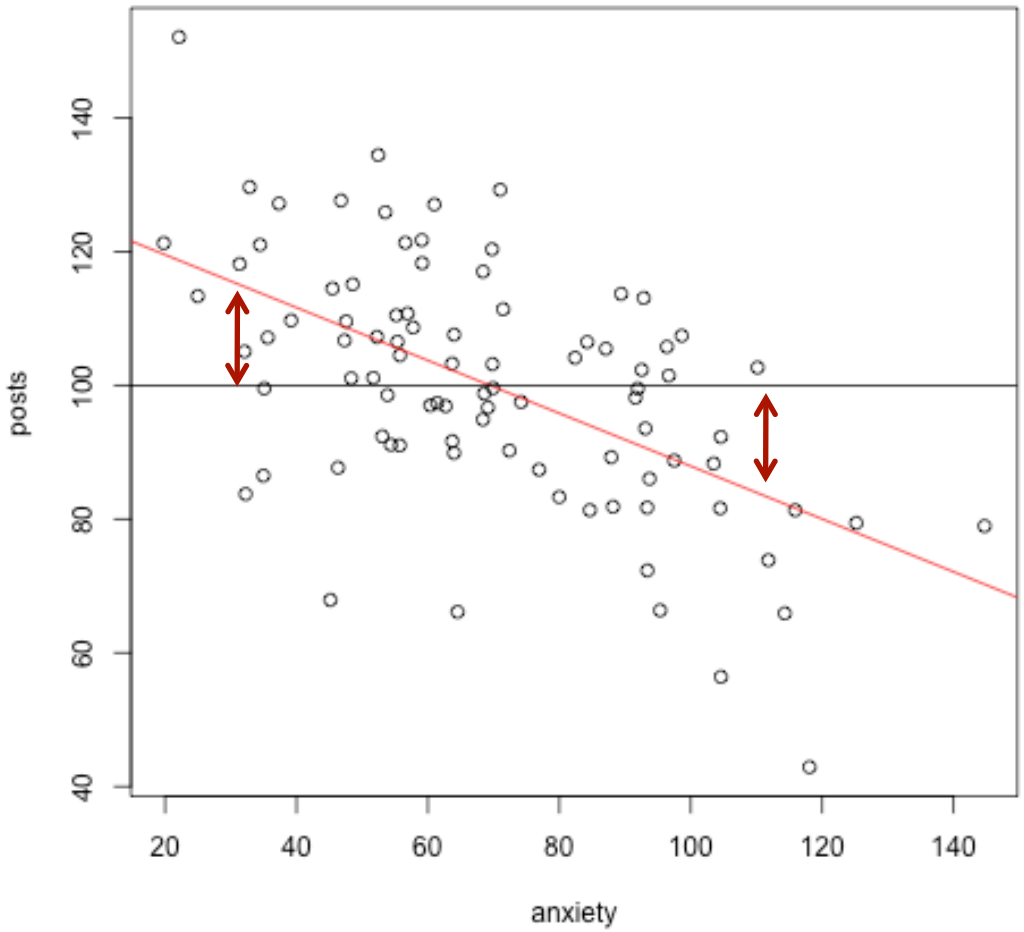
# Residual Sum of Squares (SSR)




SSR= $\Sigma$ (Observed – Model )$^2$

Total amount of differences present with best fit model applied

# Model Sum of Squares (SSM)




SSM= $\Sigma$ (Model – Mean Y)$^2$

Total amount of differences between the predicted Y from best fit and the mean of Y

Alternatively: SSM= SST- SSR

# How good is the model?

- If SSM is large then suggests model has made big improvement over just using the mean
- If SSM is small then model has made little improvement over the mean
- We can get a number to show us how much improvement ($R^2$)
- $R^2$= SSM/SST
- How much variation explained by the model as a proportion of how much there was in the first place
- $R^2$x100 = % variation explained by model

# How good is this model?

- Can also assess this through F Ratio Test
- F = Improvement due to model ($MS_M$)/difference between model and observed data ($MS_R$)
- We want F to be large ($MS_M$ to be large and $MS_R$ to be small)

# Is the predictor significant?

- Does anxiety significantly predict posts?
- Mean model assumes $\beta_1$ to be 0
  - No increase in y with unit increase in x
- We want to see whether the $\beta_1$ for regression model is significantly larger than 0
- This is done by running a t-test on these values (more next week)
- P value is the probability that the obtained t value would occur if $\beta_1$ was actually 0

# The Philosophy of Statistics

**test statistic= $s^2$ explained by model / $s^2$ unexplained by model**

we know how frequently certain test statistic values occur

We can therefore calculate the probability of obtaining that value by chance (the p value)

# The Philosophy of Statistics

- As the test stat gets larger there is less likelihood that the test stat is due to chance

- When this likelihood falls below 0.05 (Fisher's Criterion) we say that our findings are statistically significant

  - In other words we accept the H1 and reject H0

# What is a P Value?

- The probability that the results obtained occurred by chance assuming there is no effect at all
- As p value gets lower then more certain we can reject our null hypothesis
- p<= 0.05, p <= 0.01, p<= 0.001

# How do we get a p value?

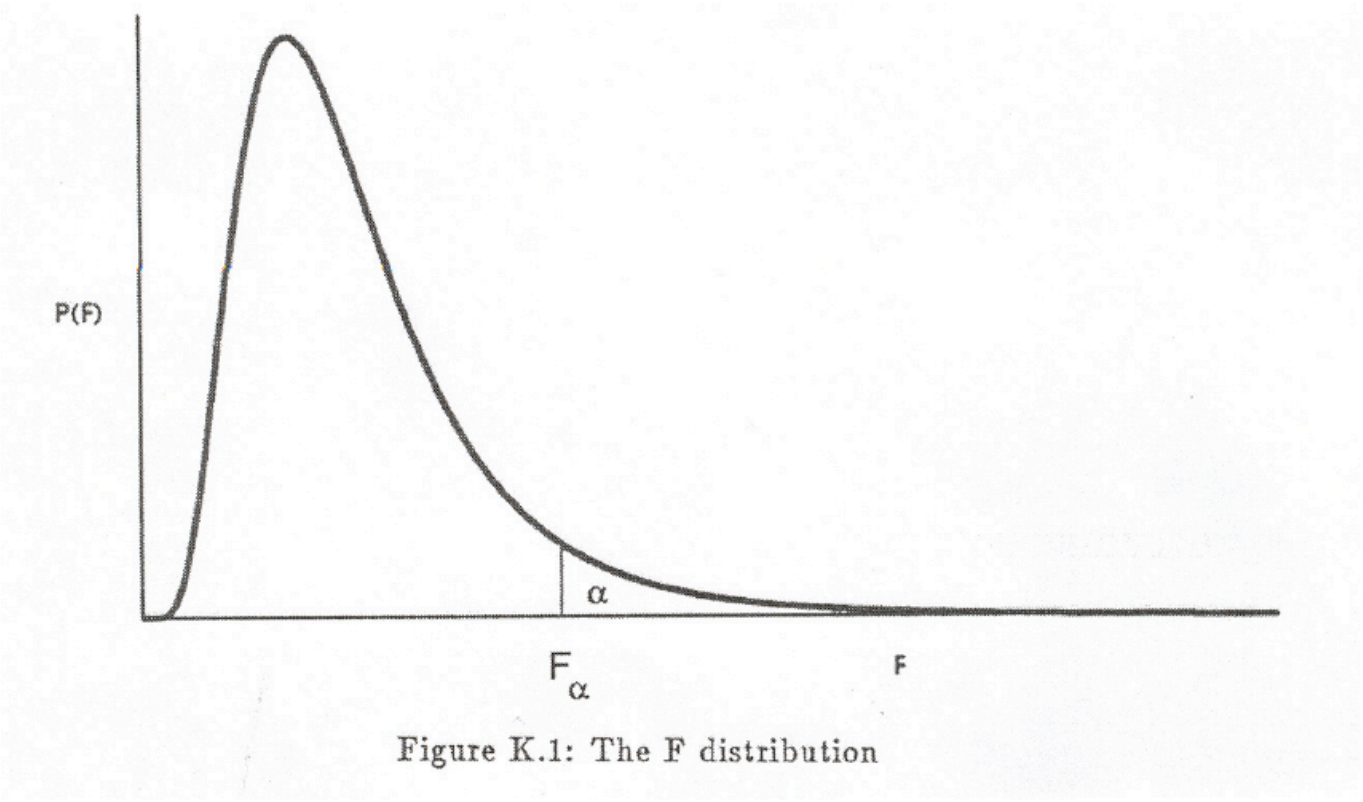



Figure K.1: The F distribution

# Reporting Statistics

- American Psychological Association Style Guide
- E.g. Correlation reporting:
  - **r(82)= -.79, p<0.001**

Degrees of freedom (n-2)

Pearson's correlation coefficient

P value

- Has instructions for each test
- Guide is available online

# Reporting Statistics

- Reporting regression
  - Report regression table

  - In text reporting of predictor findings and regression analysis results:
    - $b = -.34$, $t(225) = 6.53$, $p < .001$
    - $R^2 = .12$, $F(1, 225) = 42.64$, $p < .001$

# Readings

- Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R. Chapter 7