

Session 4: Linear models

Overview

Multiple regression continued

- Categorical predictors and interactions
- Polynomial regression

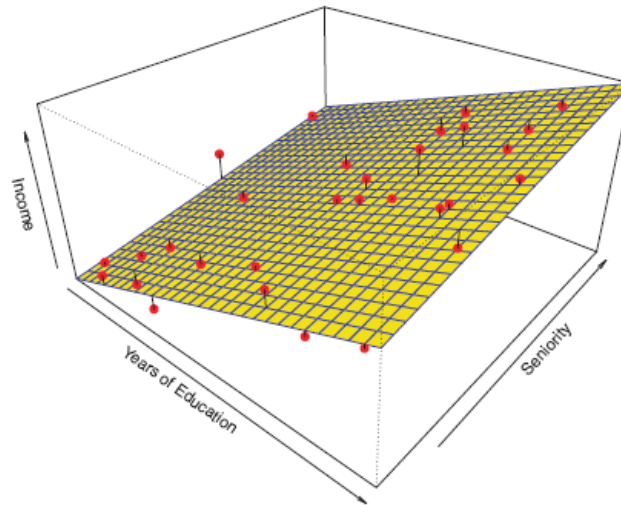
Logistic regression

Analysis of Variance

Multiple regression

In multiple regression we try to predict a quantitative response variable y using several predictor variables x_1, x_2, \dots, x_k

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x_1 + \hat{\beta}_2 \cdot x_2 + \dots + \hat{\beta}_k \cdot x_k$$



Categorical predictors

When a qualitative predictor has k levels, we need to use $k - 1$ dummy variables to code it

- e.g., we would need two dummy variables to have different intercepts for Assistant, Associate and Full Professors

$$x_{i1} = \begin{cases} 1 & \text{if Assistant Professor} \\ 0 & \text{if Full Professor} \end{cases} \quad x_{i2} = \begin{cases} 1 & \text{if Associate Professor} \\ 0 & \text{if Full Professor} \end{cases}$$

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \beta_2 + \epsilon_i & \text{if Associate Professor} \\ \beta_0 + \epsilon_i & \text{if Full Professor} \end{cases}$$

Categorical predictors

Predictors can be categorical as well as quantitative

- When a qualitative predictor has k levels, we need to use $k - 1$ dummy variables to code it

Suppose we want to predict faculty salary as a function of endowment with separate intercepts for faculty rank

```
> summary(fit_prof_rank_offset)
```

Call:

```
lm(formula = salary_tot ~ log_endowment + rank_name, data = IPED_2)
```

Residuals:

Min	1Q	Median	3Q	Max
-52464	-10844	-2703	6936	74994

Coefficients:

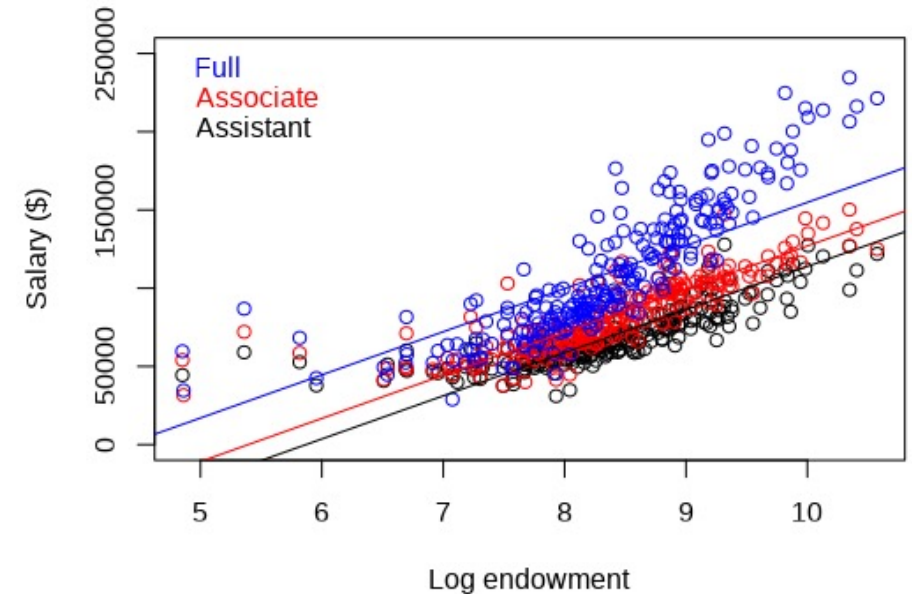
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	-120822.1	6713.9	-18.00	<0.0000000000000002 ***
log_endowment	27569.9	791.7	34.82	<0.0000000000000002 ***
rank_nameAssociate	-27855.4	1685.5	-16.53	<0.0000000000000002 ***
rank_nameAssistant	-40973.7	1685.5	-24.31	<0.0000000000000002 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18370 on 707 degrees of freedom

Multiple R-squared: 0.7192, Adjusted R-squared: 0.718

F-statistic: 603.7 on 3 and 707 DF, p-value: < 0.0000000000000022



$$\hat{y}_i = \begin{cases} \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 & \text{if assistant professor} \\ \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_3 & \text{if associate professor} \\ \hat{\beta}_0 + \hat{\beta}_1 x_{i1} & \text{if full professor} \end{cases}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i3}$$

$$= -120,822 + 27,570x_{i1} - 40,973x_{i2} - 27,855x_{i3}$$

Interaction terms

An ***interaction effect*** occurs when the response variable y is influenced by the levels of two or more predictors in a non-additive way

We can model this using an equation with an interaction term

$$y = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_3 (x_1 \cdot x_2) + \epsilon$$

An interaction term between a quantitative and categorical variable corresponds to different slopes depending for the quantitative variable depending on the value of the categorical variable

Interaction terms

If Full Professor:

$$\text{salary} \approx \beta_0 + \beta_1 \cdot \text{endowment}$$

If Assistant Professor:

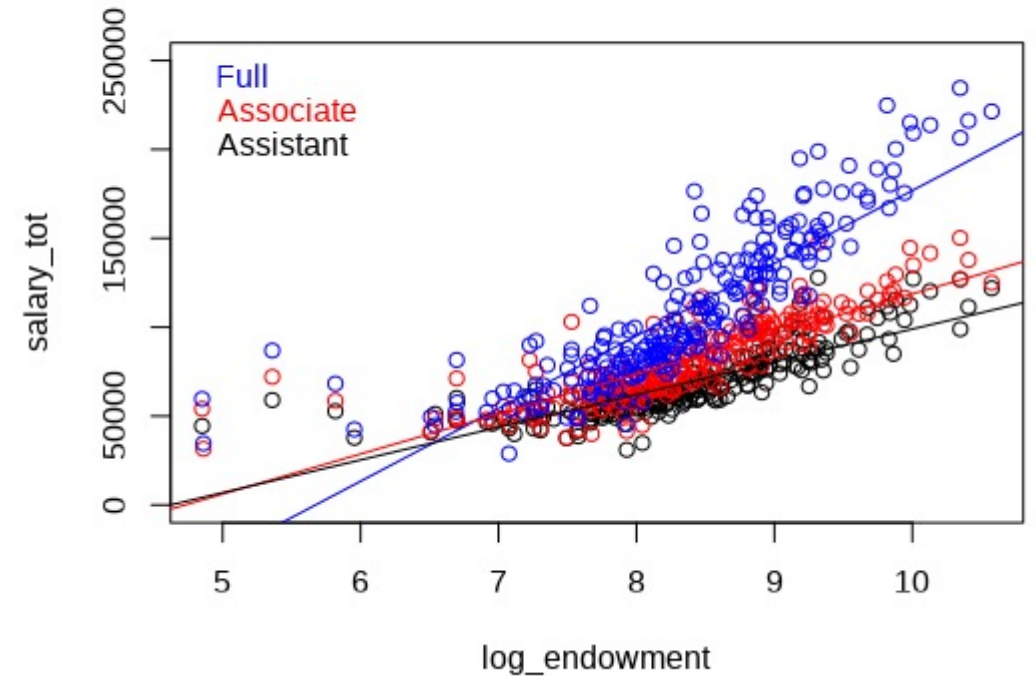
$$\text{salary} \approx (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \cdot \text{endowment}$$

Additive term if Assistant Professor

Change in slope if Assistant Professor

$$x_{i2} = \begin{cases} 1 & \text{if assistant professor} \\ 0 & \text{otherwise} \end{cases}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \hat{\beta}_3 x_{i1} \cdot x_{i2}$$



Interaction terms

$$\begin{aligned}\text{salary} \approx & \beta_0 + \beta_1 \cdot \text{endowment} \\ & + \beta_2 \cdot \text{assistant_rank_dummy} \\ & + \beta_3 \cdot (\text{assistant_rank_dummy} \cdot \text{endowment})\end{aligned}$$

Let's try it in R...

Multicollinearity

Multicollinearity occurs when our predictors (x_i 's) are correlated.

This can lead to unstable estimates of the regression coefficients

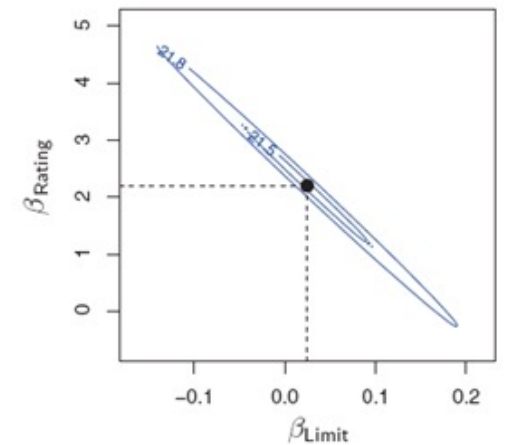
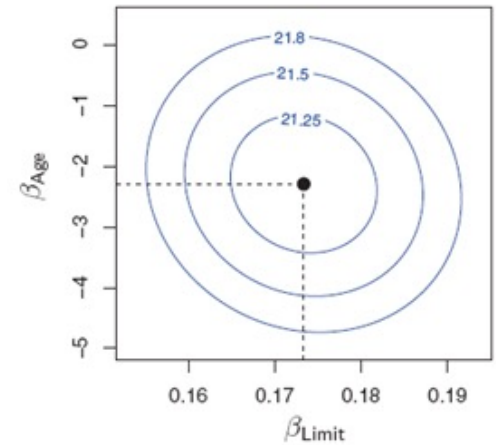
- Which will lead to large SE on the coefficients and consequently they will not appear to be statistically significant.

The **variance inflated factor** can be used to test for multicollinearity each explanatory

- Rule of thumb: $VIF > 5$ suspect for multicollinearity

`car::vif(lm_fit)`

Contours of equal SSR residuals



Non-linear relationships

Polynomial regression extends linear regression to non-linear relationships by including nonlinear transformations of predictors

$$\begin{aligned}\text{salary} = & \beta_0 + \beta_1 \cdot \text{endowment} \\ & + \beta_2 \cdot (\text{endowment})^2 + \\ & + \beta_3 \cdot (\text{endowment})^3 + \varepsilon\end{aligned}$$

Still a linear equation but non-linear in original predictors

Logistic regression

In **logistic regression** we try to predict whether a case belongs to one of two categories

- Does a case belong to category a or category b ?
- Example: can we predict if a faculty member is an Assistant or Full professor based on the salary level?

Making predictions for a categorical variable is called **classification**

- The field of Machine Learning has developed many classification methods

In logistic regression we build a conditional probability model:

- $P(\text{Class} = a \mid x)$
- $P(\text{Assistant Professor} \mid \text{salary} = \$60,000)$

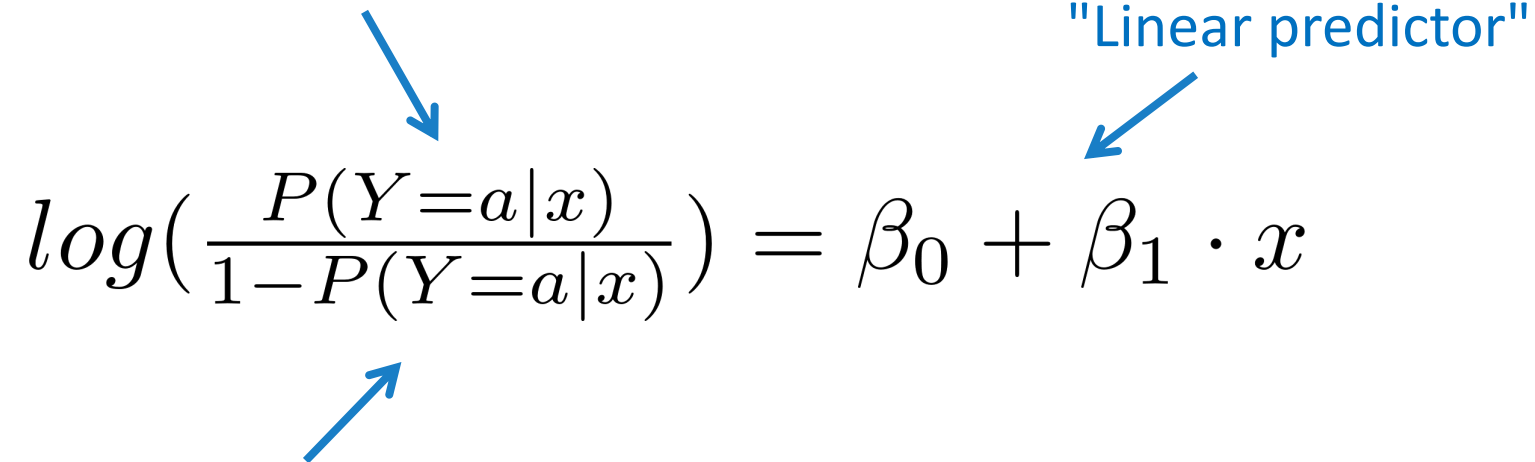
Generalized linear models

Generalized linear models use a linear combinations of predictors to predict ***a function of the mean***

If Y is a binary response variable ($Y = 0$ or 1)

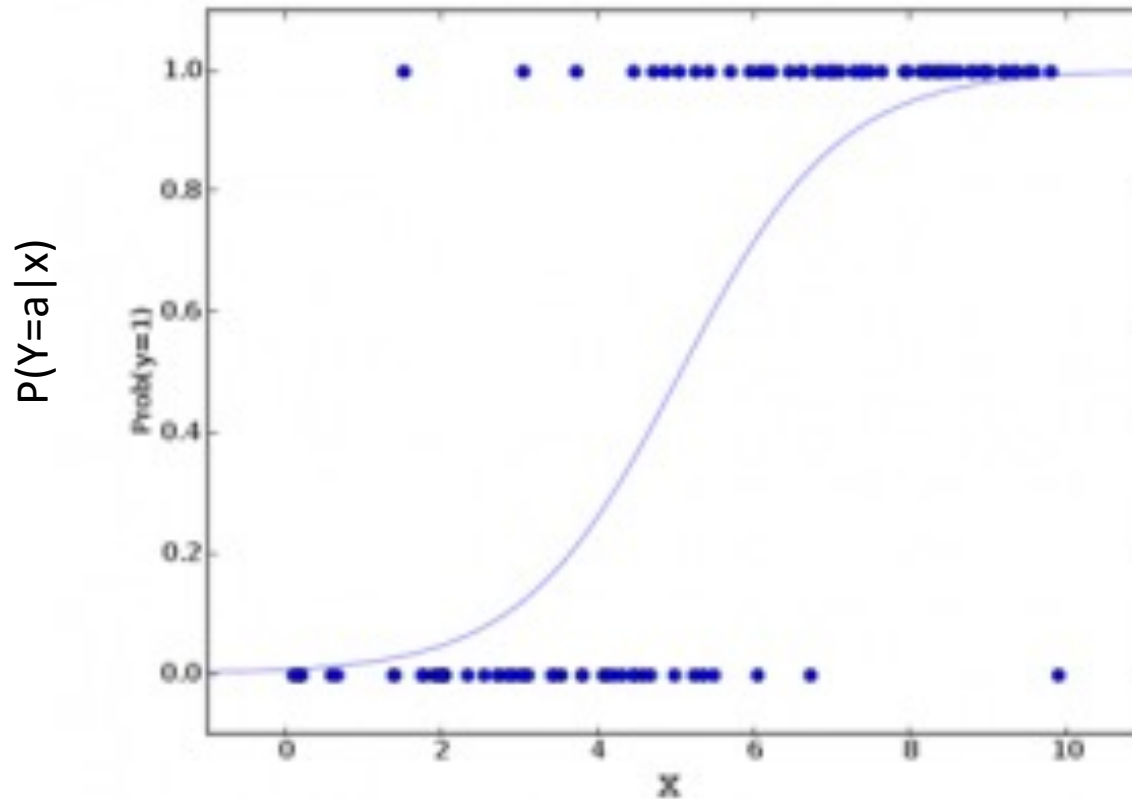
$P(Y = 1|x)$ is the mean of Y

"Linear predictor"


$$\log\left(\frac{P(Y=a|x)}{1-P(Y=a|x)}\right) = \beta_0 + \beta_1 \cdot x$$

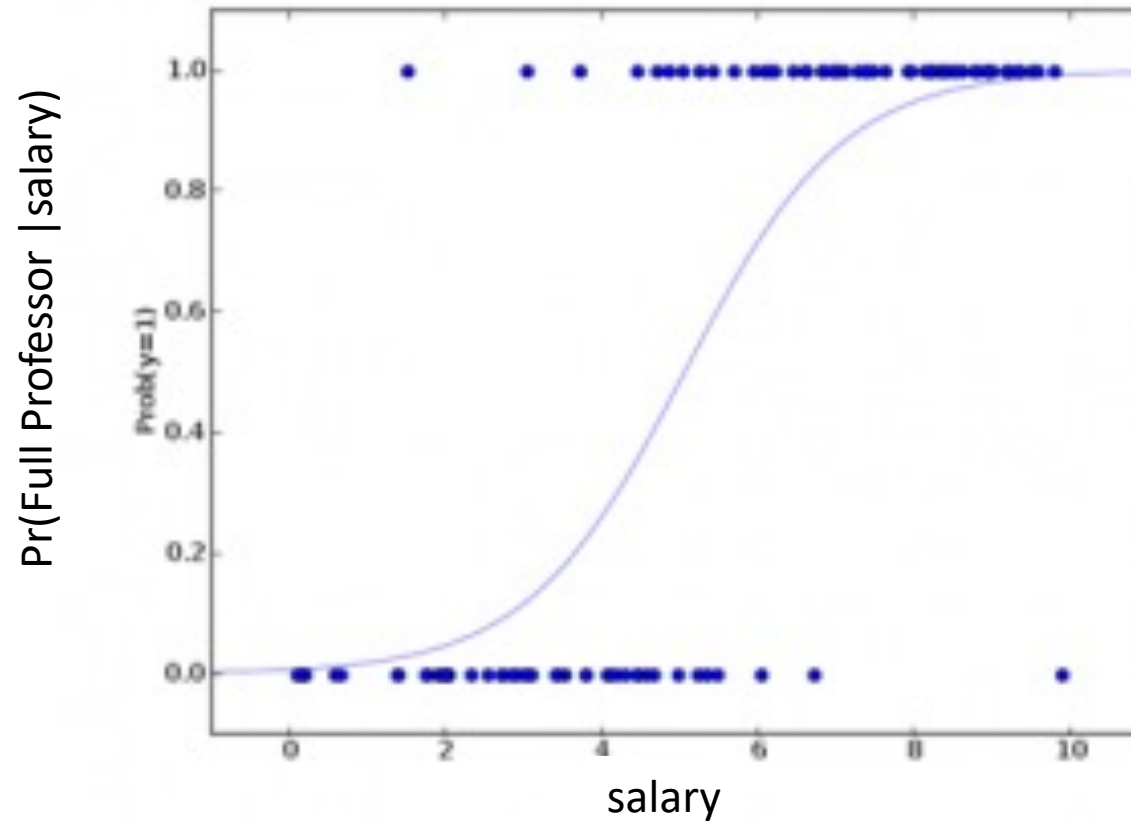
The logit function (log-odds) is a "link function" that links the mean to the linear predictor

Plotting the logistic function



$$P(Y = a|x_1) = \frac{e^{\beta_0 + \beta_1 \cdot x_1}}{1 + e^{\beta_0 + \beta_1 \cdot x_1}}$$

Plotting the logistic function



$$P(\text{Full Professor} \mid \text{salary}) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{salary}}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 \cdot \text{salary}}}$$

One-way ANOVA

A **one-way analysis of variance (ANOVA)** is a parametric hypothesis test that can be used to examine if a set of means are all the same.

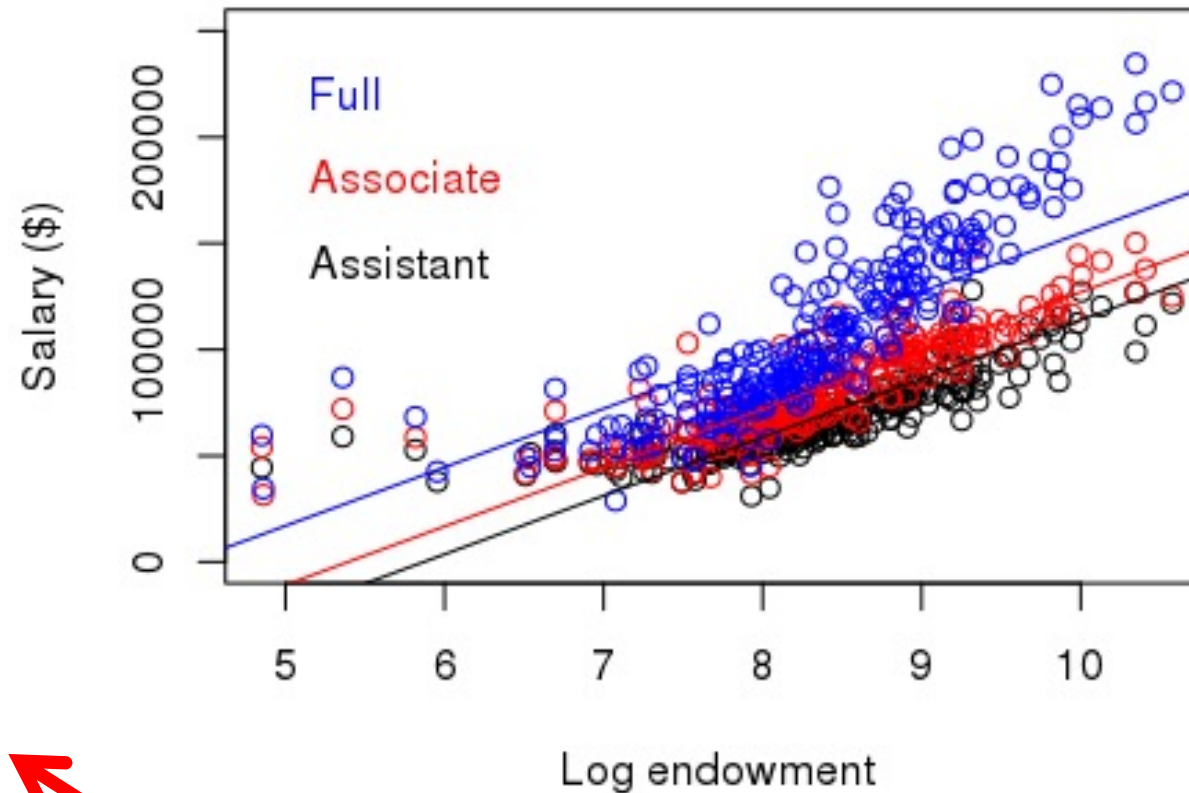
$$H_0: \mu_1 = \mu_2 = \dots = \mu_k$$

$$H_A: \mu_i \neq \mu_j \text{ for some } i, j$$

The statistic we use for a one-way ANOVA is the F-statistic

$$F = \frac{\text{between-group variability}}{\text{within-group variability}} = \frac{\frac{1}{K-1} \sum_{i=1}^K n_i (\bar{x}_i - \bar{x}_{tot})^2}{\frac{1}{N-K} \sum_{i=1}^K \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2}$$

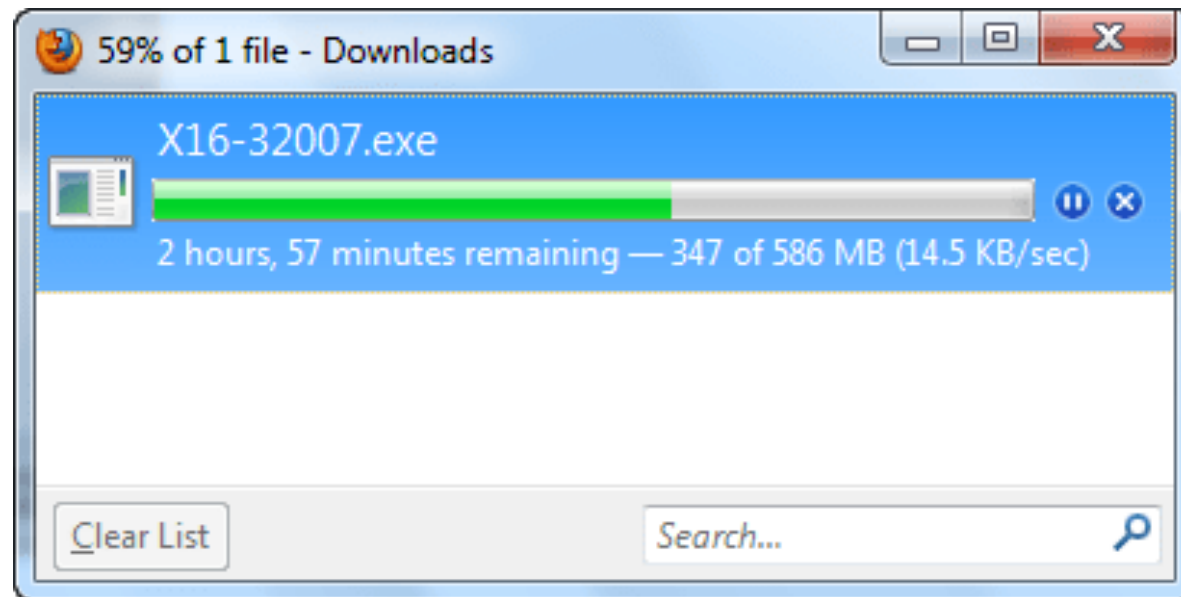
ANOVA as regression with only categorical predictors



Common slope for
log endowment

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} - \text{X} + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \text{X} + \epsilon_i & \text{if Assistant Professor} \\ \beta_0 + \beta_2 + \text{X} + \epsilon_i & \text{if Associate Professor} \\ \beta_0 + \text{X} + \epsilon_i & \text{if Full Professor} \end{cases}$$

Let's try it in R...

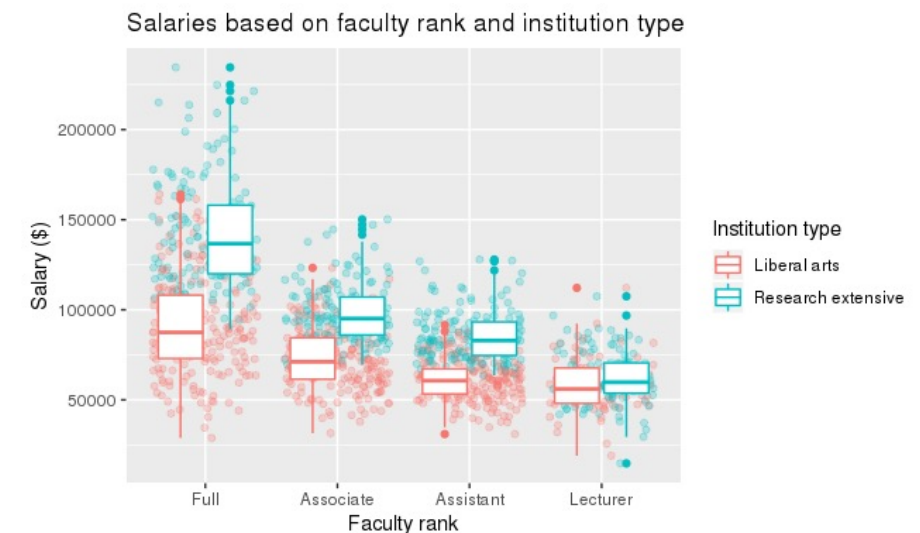


Factorial ANOVA

In a **factorial ANOVA**, we model the response variable y as a function of **more than one** categorical predictor

Example 1: Do faculty salaries depend on faculty rank, and the type of college/university

- Factors are:
 - **Rank:** Lecturer, Assistant, Associate, Full
 - **Institute:** liberal arts college, research university
 - 4 x 2 design



Factorial ANOVA

Example 2: A student at Queensland University of Technology conducted an experiment to determine what types of sandwiches ants prefer.

- Factors he looked at were:
 - **Bread:** rye, whole wheat multigrain, white
 - **Filling:** peanut butter, ham and pickle, and vegemite
 - 4 x 3 design

The student creating 4 sandwiches of all combinations of bread and filling (48 sandwiches total) and randomly left pieces in front of ant nests.

He then measured how many ants were on the sandwiches 5 minutes later.

Two-way ANOVA hypotheses

Main effect for A (bread type doesn't matter or institution type doesn't matter)

$$H_0: \alpha_1 = \alpha_2 = \dots = \alpha_j = 0$$

$$H_A: \alpha_j \neq 0 \text{ for some } j$$

Where:

Main effect for B (filling doesn't matter)

$$H_0: \beta_1 = \beta_2 = \dots = \beta_K = 0$$

$$H_A: \beta_k \neq 0 \text{ for some } k$$

α_j : is the “effect” for factor A at level j

β_k : is the “effect” for factor B at level k

Interaction effect:

$$H_0: \text{All } \gamma_{jk} = 0$$

$$H_A: \gamma_{jk} \neq 0 \text{ for some } j, k$$

γ_{jk} : is the interaction between level j of factor A, and level k of factor B.

Repeated measures ANOVA

In a **repeated measures ANOVA**, the same case/observational units are measured at each factor level.

Example: Do people prefer chocolate, butterscotch or caramel sauce?

Between subjects experiment: different people rate chocolate, butterscotch or caramel sauce.

- Run a between subjects ANOVA (as we have done before)

Within subjects experiment: each person in the experiment gives ratings for all three toppings.

- Run a repeated measures ANOVA

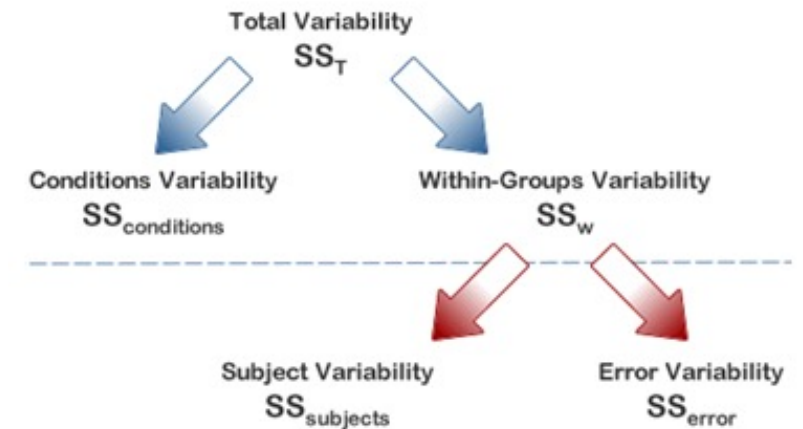
Repeated measures ANOVA

The advantages of a repeated measures ANOVA is that we can potentially reduce a lot of the variability between the cases

- This is a generalization of a paired t-test to more than two population means

To run a repeated measures ANOVA, we use a factor called ID that has a unique value for each observational unit

```
aov(reaction_time ~ condition * position + participant,  
    data = popout_log_data)
```

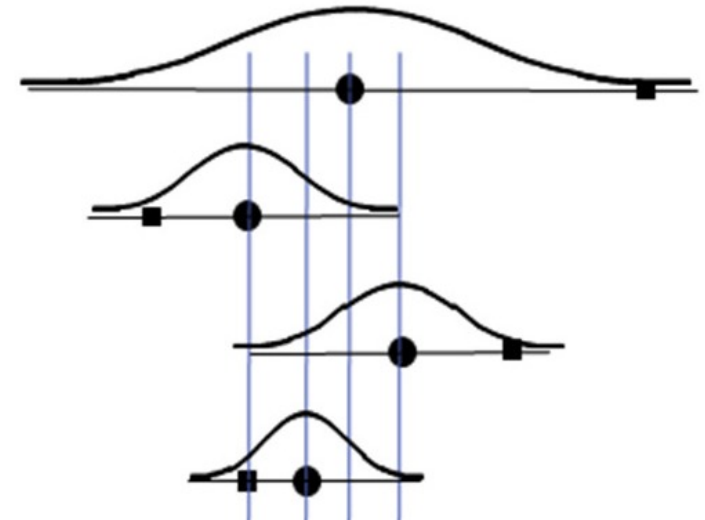


Brief mention: random effects models

In a random effects ANOVA, factor levels are viewed as being randomly generated from an underlying distribution, rather than having a fixed number of levels.

For example, we could view participants in an experiment as being a random sample from participants in a population.

- We then just estimate a mean and standard deviation for the underlying population, rather a separately ID for each participant.
 - This leads to few parameters and hence more degrees of freedom.



You can run mixed effects models in R using the [lme4](#) package