THE UNIVERSITY *of* EDINBURGH

Week 11: Some Kind of End to the Course

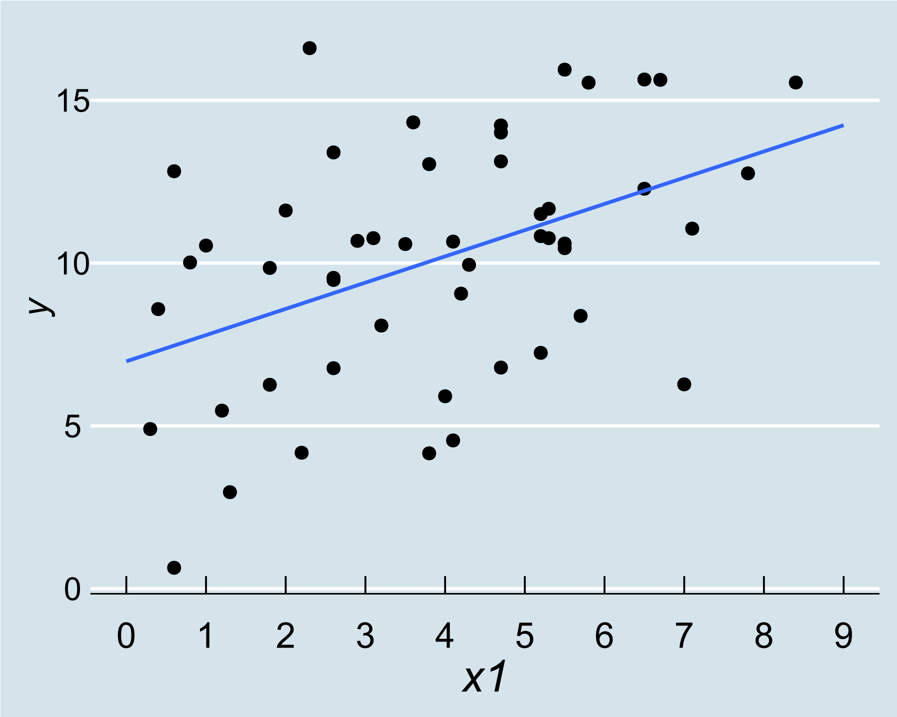
Univariate Statistics and Methodology using R

Department of Psychology
The University of Edinburgh

Part 1

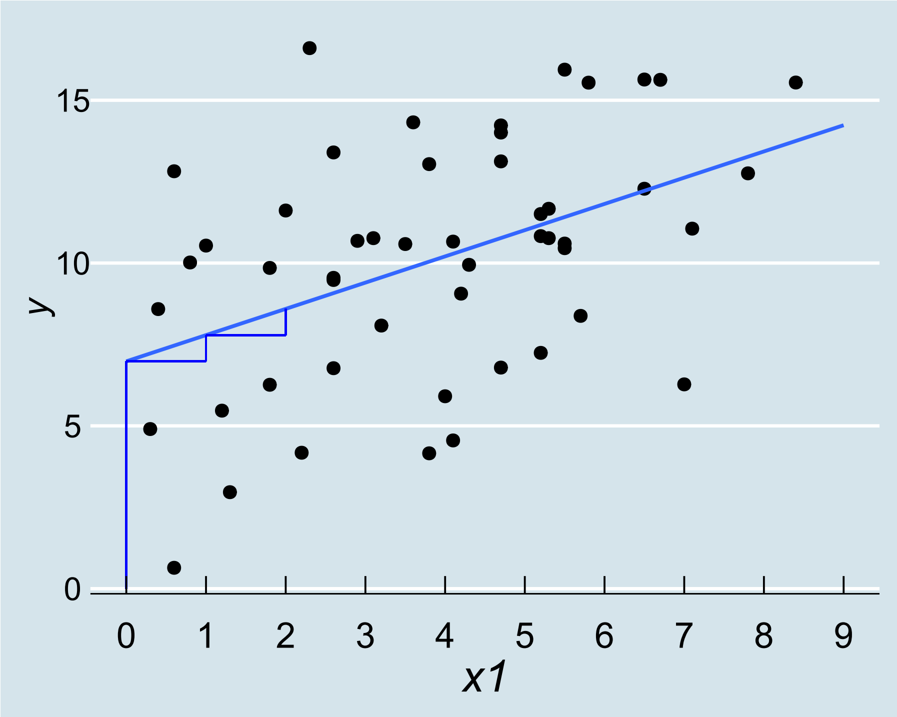
Week 7 - 10 Recap

Describing a pattern with a line




```
ggplot(df, aes(x = x1, y = y)) +
  geom_point()
```

Defining a line



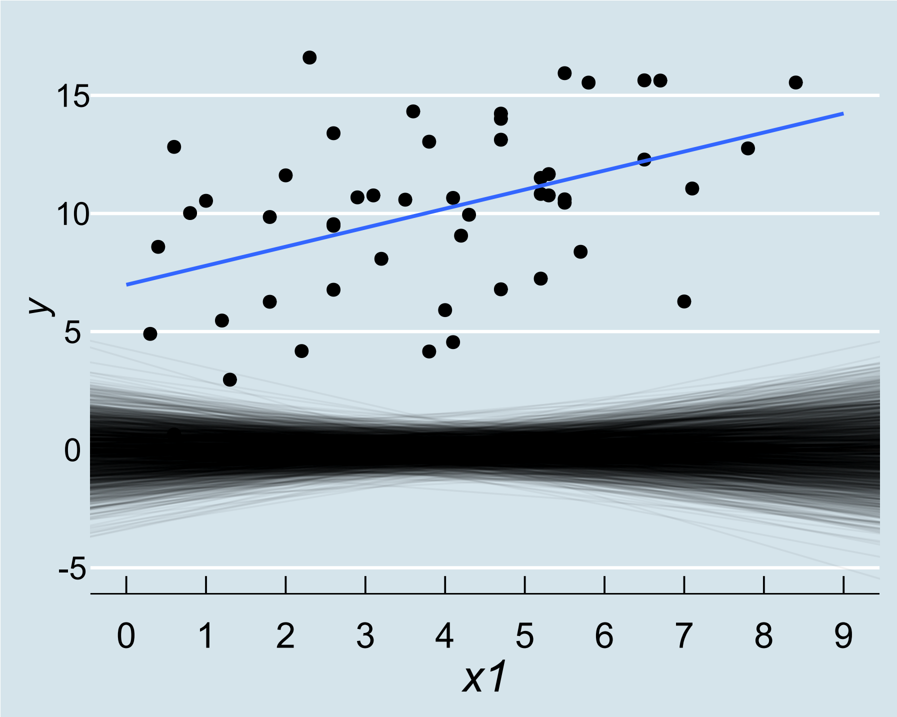
$$y_i = b_0 + b_1(x_{1i}) + \varepsilon_i$$

A line can be defined by two values:

- A starting point (Intercept)
- A slope (y across $x1$)

Fitting a linear model to some data provides coefficient estimates $\hat{b}_0$ and $\hat{b}_1$ that minimise σ_ε.

Testing the coefficients (1)

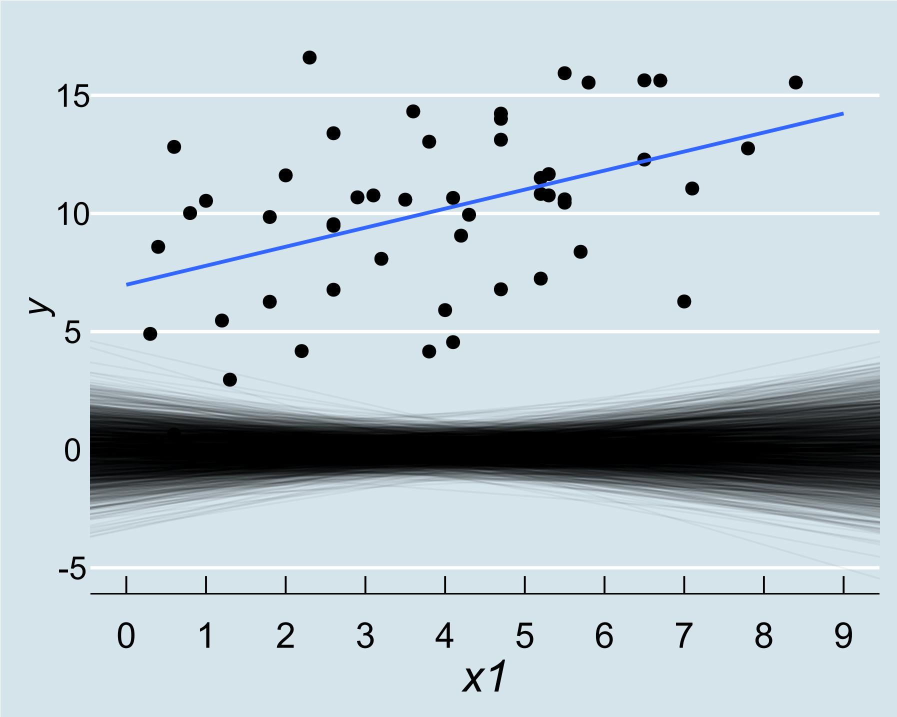


$$\hat{y} = \color{red}{\hat{b}_0} + \hat{b}_1(x_1)$$

In the "null universe" where $b_0 = 0$, when sampling this many people, what is the probability that we will find an intercept at least as extreme as the one we *have* found?

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.982      1.026    6.81  1.5e-08 ***
## x1             0.806      0.234    3.45   0.0012 **
```

Testing the coefficients (2)

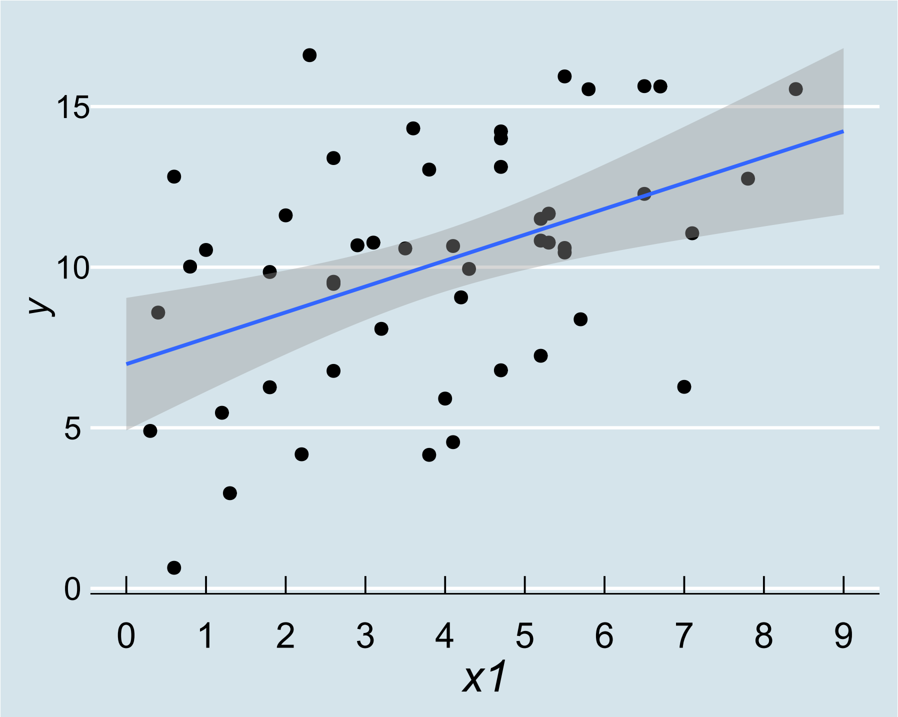


$$\hat{y} = \hat{b}_0 + \hat{b}_1(x_1)$$

In the "null universe" where $b_1 = 0$, when sampling this many people, what is the probability that we will find a relationship at least as extreme as the one we *have* found?

```
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)    6.982      1.026    6.81  1.5e-08 ***
## x1             0.806      0.234    3.45   0.0012 **
```

Coefficient Sampling Variability



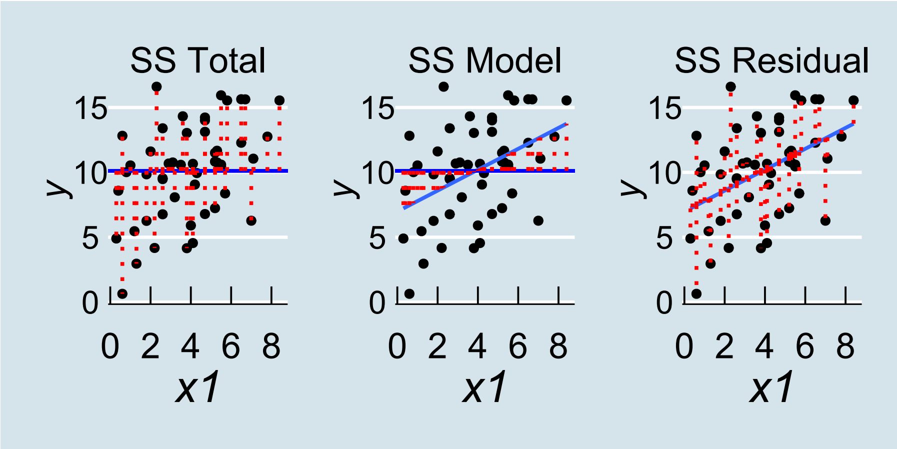
$$\hat{y} = \hat{b}_0 + \hat{b}_1(x_1)$$

Plausible range of values for b_0 and b_1:

```
confint(model)
```

```
##             Estimate 2.5 % 97.5 %
## (Intercept)   6.9816 4.919  9.044
## x1            0.8057 0.336  1.275
```

Testing the model



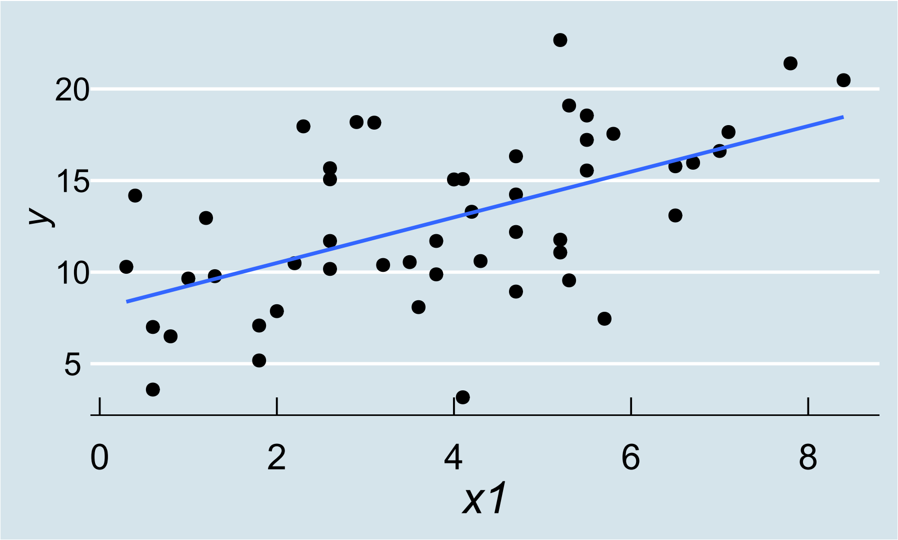
$$F_{df_{model},df_{residual}} = \frac{MS_{Model}}{MS_{Residual}} = \frac{SS_{Model}/df_{Model}}{SS_{Residual}/df_{Residual}}$$

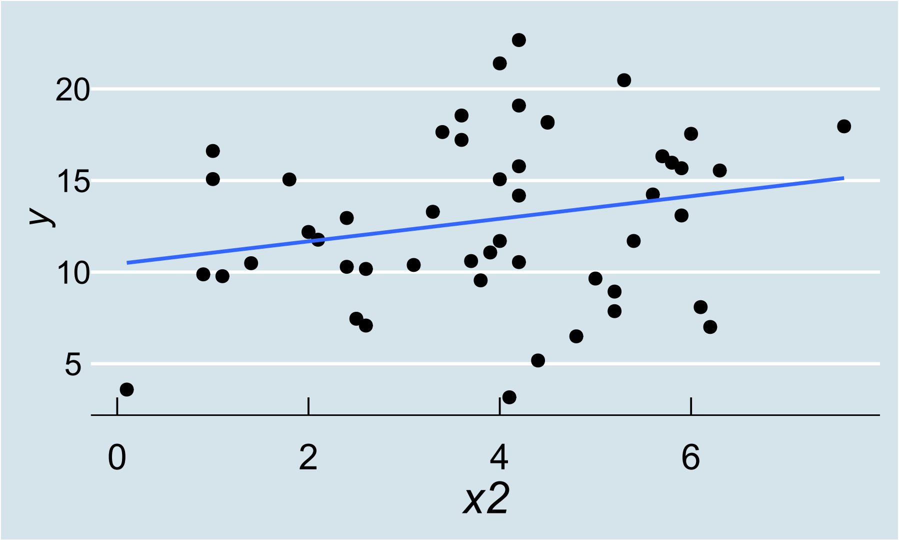
$$df_{model} = \text{nr predictors}$$

$$df_{residual} = \text{sample size} - \text{nr predictors} - 1$$

```
##
## Multiple R-squared:  0.199,	Adjusted R-squared:  0.182
## F-statistic: 11.9 on 1 and 48 DF,  p-value: 0.00118
```

More predictors





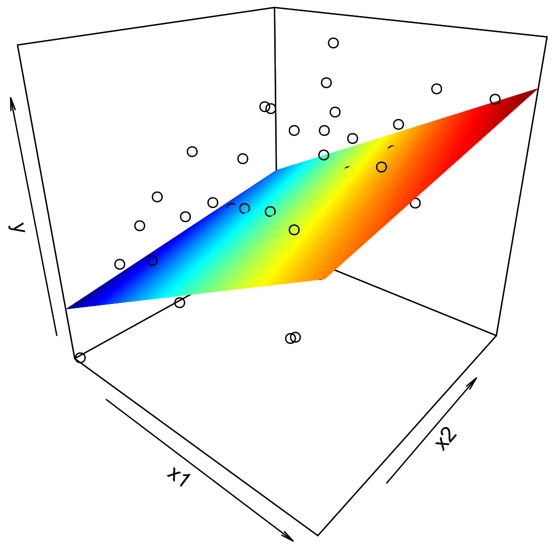
More predictors (2)

$$y_i = b_0 + b_1(x_{1i}) + b_2(x_{2i}) + \varepsilon_i$$

- A starting point (Intercept)
- A slope (across x_1)
- *Another* slope (across x_2)

Coefficient estimates $\hat{b}_0, \hat{b}_1, \hat{b}_2$ minimise σ_ε.

More predictors (3)

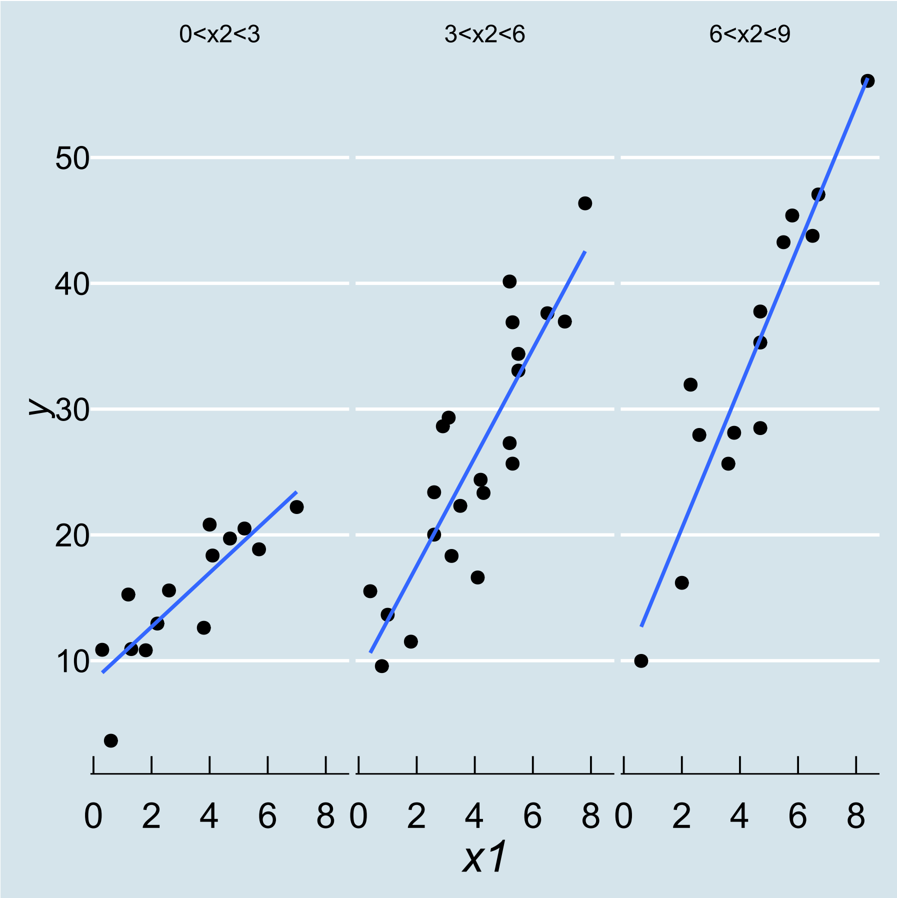


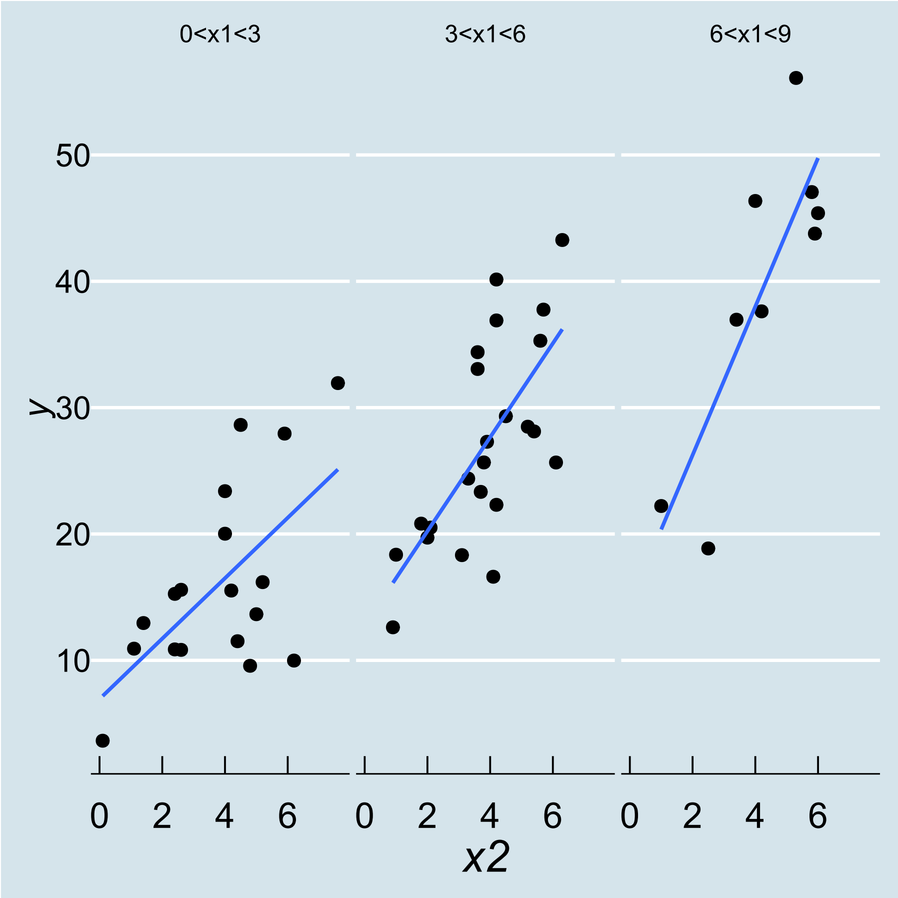
Even more predictors...

$$y_i = b_0 + b_1(x_{1i}) + b_2(x_{2i}) + \ldots + b_k(x_{ki}) + \varepsilon_i$$

- A starting point (Intercept)
- A slope (across x_1)
- A slope (across x_2)
- ...
- ...
- A slope (across x_k)

associations that depend on other things



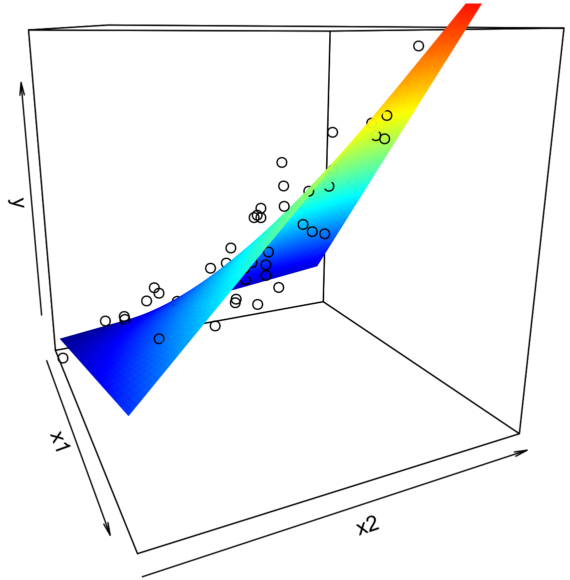


interactions

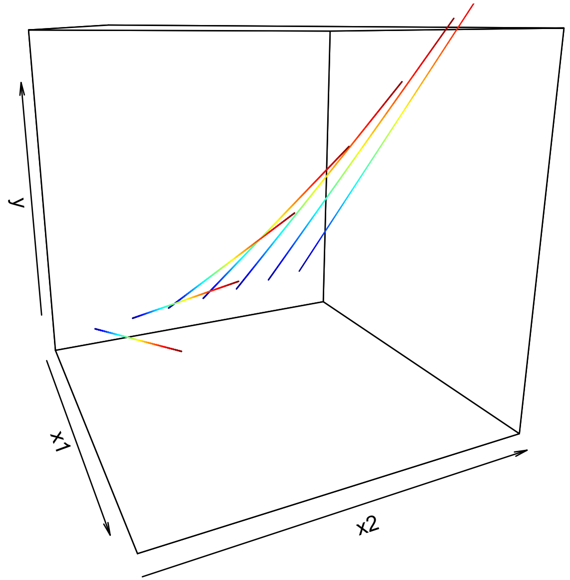
$$y_i = b_0 + b_1(x_{1i}) + b_2(x_{2i}) + b_3(x_{1i} \cdot x_{2i}) + \varepsilon_i$$

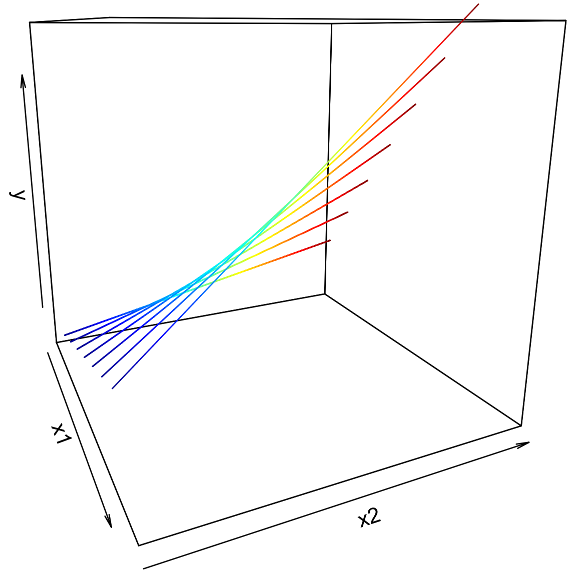
- starting point (Intercept)
- A slope (across x_1)
- A slope (across x_2)
- ...
- How slope across x_1 changes across x_2

interactions



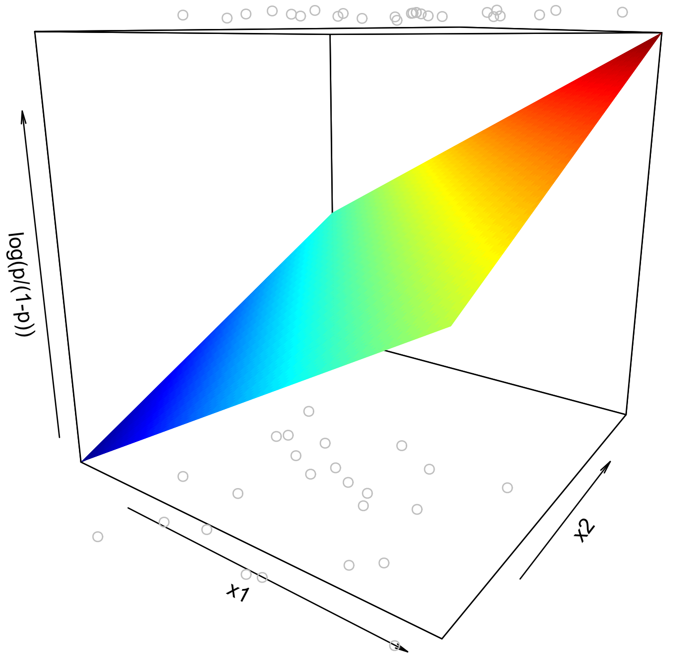
interactions (2)



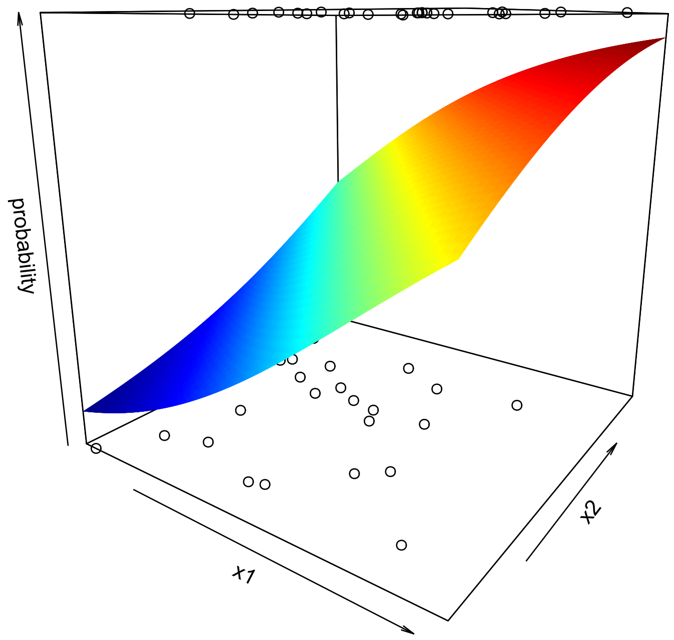


other outcomes

$$ln(\frac{p}{1-p}) = b_0 + b_1(x_{1i}) + b_2(x_{2i})$$



$$ln(\frac{p}{1-p}) \Rightarrow \frac{p}{1-p} \Rightarrow p$$



Checking Assumptions: Linear Models

required

- linearity of relationship
- for the *residuals*:
  - normality
  - homogeneity of variance
  - independence

desirable

- uncorrelated predictors
- no "bad" (overly influential) observations

Checking Assumptions: Logit Models

required

- linearity of relationship between IVs and log-odds
- for the *residuals*:
    - ~~normality~~
    - ~~homogeneity of variance~~
    - independence

desirable

- uncorrelated predictors
- no "bad" (overly influential) observations
- large samples (due to maximum likelihood fitting)

End of Part 1

Part 2

Common Tests as linear models

```
usmr <- read_csv("https://uoepsy.github.io/data/usmr2022.csv")
```

lm vs correlation

regression, continuous predictor

```
summary(lm(sleeprating ~ height, data = usmr))
```

```
##
## Call:
## lm(formula = sleeprating ~ height, data = usmr)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -65.42 -11.66   5.52  16.96  36.16
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   18.785     46.879     0.4     0.69
## height         0.279      0.278     1.0     0.32
##
## Residual standard error: 22.6 on 76 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.013,    Adjusted R-squared:  4.73e-05
## F-statistic:    1 on 1 and 76 DF,  p-value: 0.32
```

Correlation

```
cor.test(usmr$height, usmr$sleeprating)
```

```
##
##     Pearson's product-moment correlation
##
## data:  usmr$height and usmr$sleeprating
## t = 1, df = 76, p-value = 0.3
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.1112  0.3284
## sample estimates:
##    cor
## 0.1142
```

lm vs t.test

regression, intercept

```
summary(lm(height ~ 1, data = usmr))
```

```
##
## Call:
## lm(formula = height ~ 1, data = usmr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.407  -7.607  -0.107   7.523  20.893
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   168.11       1.04     162   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.22 on 78 degrees of freedom
##   (1 observation deleted due to missingness)
```

one sample t.test

```
t.test(usmr$height, mu=0)
```

```
##
##      One Sample t-test
##
## data:  usmr$height
## t = 162, df = 78, p-value <2e-16
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  166.0 170.2
## sample estimates:
## mean of x
##     168.1
```

lm vs t.test (2)

regression, binary predictor

```r
summary(lm(height ~ catdog, data = usmr))
```

```
##
## Call:
## lm(formula = height ~ catdog, data = usmr)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.655  -7.501   0.499   8.399  19.499
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   166.36       1.55  107.60   <2e-16 ***
## catdogdog       3.15       2.07    1.52     0.13
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.15 on 77 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.0291,	Adjusted R-squared:  0.0165
## F-statistic: 2.31 on 1 and 77 DF,  p-value: 0.133
```

two sample t.test

```r
t.test(height ~ catdog, data = usmr,
       var.equal = TRUE)
```

```
##
##     Two Sample t-test
##
## data:  height by catdog
## t = -1.5, df = 77, p-value = 0.1
## alternative hypothesis: true difference in means between group cat and group
## 95 percent confidence interval:
##  -7.2706  0.9796
## sample estimates:
## mean in group cat mean in group dog
##             166.4             169.5
```

lm vs Traditional ANOVA

> If you should say to a mathematical statistician that you have discovered that linear multiple regression and the analysis of variance (and covariance) are identical systems, he would mutter something like "Of course—general linear model," and you might have trouble maintaining his attention. If you should say this to a typical psychologist, you would be met with incredulity, or worse. Yet it is true, and in its truth lie possibilities for more relevant and therefore more powerful research data.
>
> Cohen (1968)

History

Multiple Regression

- introduced c. 1900 in biological and behavioural sciences
- aligned to "natural variation" in observations
- tells us that means $(\bar{y})$ are related to groups $(g_1, g_2, \ldots, g_n)$

ANOVA

- introduced c. 1920 in agricultural research
- aligned to experimentation and manipulation
- tells us that groups $(g_1, g_2, \ldots, g_n)$ have different means $(\bar{y})$

- both produce F-ratios, discussed in different language, but identical

lm vs Traditional ANOVA

regression, binary predictor

```
summary(lm(height ~ eyecolour, data = usmr))
```

```
##
## lm(formula = height ~ eyecolour, data = usmr)
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)      170.254      2.183   77.99   <2e-16 ***
## eyecolourbrown    -3.682      2.609   -1.41     0.16
## eyecolourgreen     2.717      4.125    0.66     0.51
## eyecolourgrey     -0.254      9.515   -0.03     0.98
## eyecolourhazel    -2.404      3.935   -0.61     0.54
## eyecolourother    -4.834      5.775   -0.84     0.41
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.26 on 73 degrees of freedom
##   (1 observation deleted due to missingness)
## Multiple R-squared:  0.0563,    Adjusted R-squared:  -0.00837
## F-statistic: 0.871 on 5 and 73 DF,  p-value: 0.505
```

anova

```
summary(aov(height ~ eyecolour, data = usmr))
```

```
##               Df Sum Sq Mean Sq F value Pr(>F)
## eyecolour      5    373    74.7    0.87   0.51
## Residuals     73   6261    85.8
## 1 observation deleted due to missingness
```

Why Teach LM/Regression?

- LM has less restrictive assumptions
    - especially true for unbalanced designs/missing data
- LM is far better at dealing with covariates
    - can arbitrarily mix continuous and discrete predictors
- LM is the gateway to other powerful tools
    - mixed models and factor analysis ($\longrightarrow$ MSMR)
    - structural equation models

End

Goodbye!