Aalto University
School of Science

Correlations, regression

Experimental and Statistical Methods in Biological Sciences I

Heini Saarimäki, BECS

9/10/2014

Outline for today

- Introduction to the basic concepts

- Demos and exercises
  - Correlations
  - Regression models
  - Contrasts

What we know so far

1. Preparing your data

- Factors as factors
- Missing values coded as NAs

2. Describing your data

- Plotting
- Descriptive statistics

3. Statistical comparisons

- T-tests
- Between-subjects ANOVA

What we know after today

1. Preparing your data

2. Describing your data

3. Statistical comparisons

4. Correlations

- Correlations
- Regression

0. One step back...
Testing for hypotheses

1) State the hypotheses (null and alternative)

example:
H0: there are no differences between groups
H1: there is a difference between groups

0. One step back...
Testing for hypotheses

1) State the hypotheses (null and alternative)
2) Formulate an analysis plan (how to use sample data to evaluate the null hypothesis)

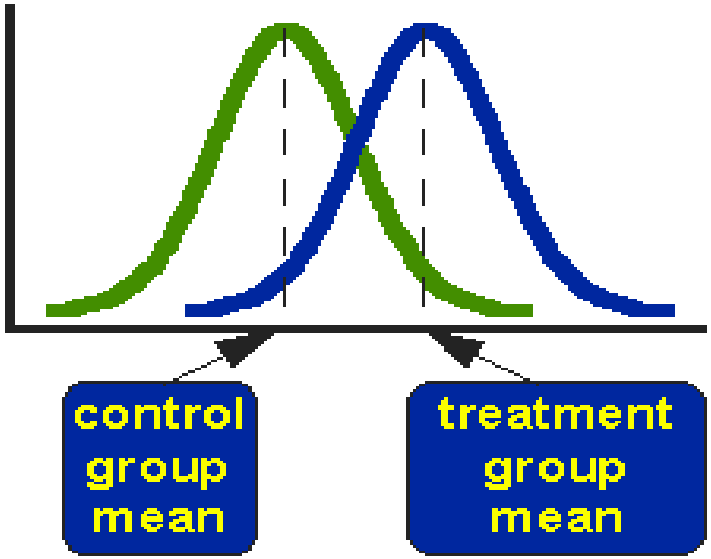
    significance level: 0.05 is acceptable
    test method: two-sample t test

0. One step back...
Testing for hypotheses

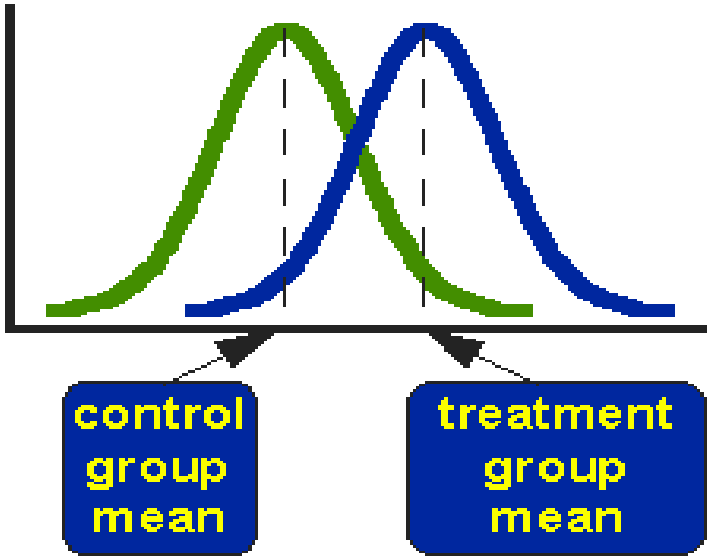
1) State the hypotheses (null and alternative)
2) Formulate an analysis plan (how to use sample data to evaluate the null hypothesis)
3) Analyze sample data (find the value of the test statistic described in the analysis plan)

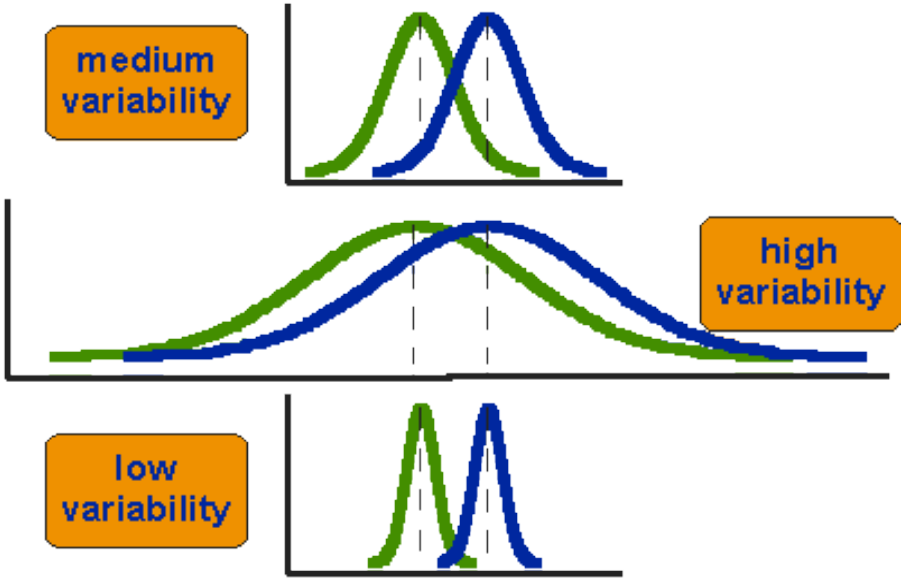
    calculate standard error, degrees of freedom
    use these to calculate test statistic (t-score)
    assess the p value (=probability of observing a sample statistic as extreme as the test statistic)

T test



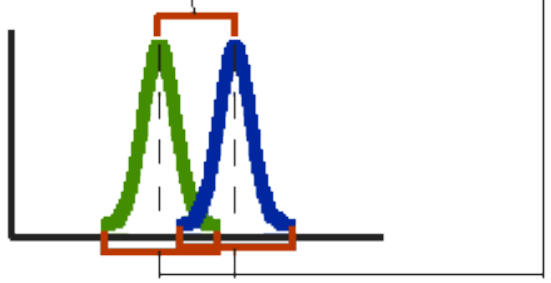
T test


control group mean
treatment group mean


medium variability
high variability
low variability

T test

$$\frac{\text{signal}}{\text{noise}} = \frac{\text{difference between group means}}{\text{variability of groups}} = \frac{\bar{X}_T - \bar{X}_C}{SE(\bar{X}_T - \bar{X}_C)} = \text{t-value}$$



0. One step back...
Testing for hypotheses

1) State the hypotheses (null and alternative)
2) Formulate an analysis plan (how to use sample data to evaluate the null hypothesis)
3) Analyze sample data (find the value of the test statistic described in the analysis plan)

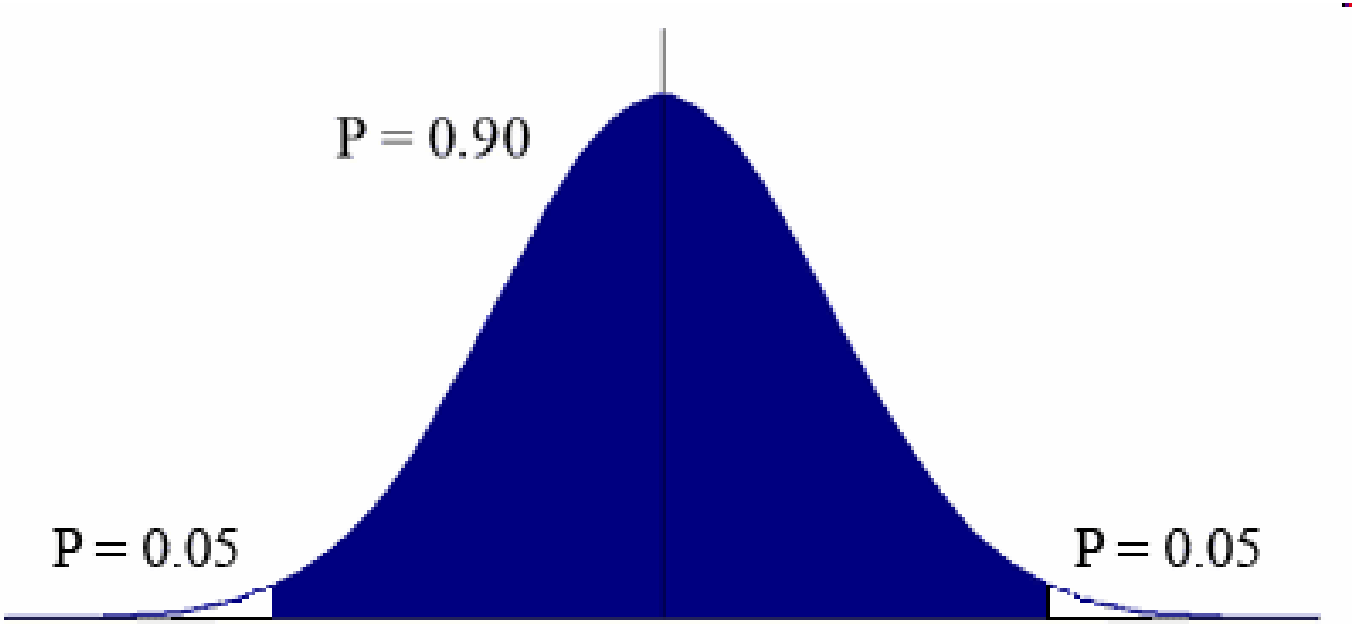
    calculate standard error, degrees of freedom
    use these to calculate test statistic (t-score)
    assess the p value (=probability of observing a sample statistic as extreme as the test statistic)

0. One step back... Testing for hypotheses

1) State the hypotheses (null and alternative)
2) Formulate an analysis plan (how to use sample data to evaluate the null hypothesis)
3) Analyze sample data (find the value of the test statistic described in the analysis plan
4) Interpret results (apply the decision rule described in the analysis plan – if the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis)

    compare p value to the significance level, reject null hypothesis when p value is less than significance level

Probability to observe this p value



0. One step back... Testing for hypotheses

1) State the hypotheses (null and alternative)
2) Formulate an analysis plan (how to use sample data to evaluate the null hypothesis)
3) Analyze sample data (find the value of the test statistic described in the analysis plan
4) Interpret results (apply the decision rule described in the analysis plan – if the value of the test statistic is unlikely, based on the null hypothesis, reject the null hypothesis)

    compare p value to the significance level, reject null hypothesis when p value is less than significance level

Data for the lecture

- If you want to play around with it, you can find the data here:
  - http://becs.aalto.fi/~heikkih3/ses
  - Use `read.table` and set `header=TRUE`:

```
data <- read.table('http://becs.aalto.fi/ ~heikkih3/ses',
    header=T)
```

- Data has been simulated

1. Getting started

- Load in data
- Prepare the data as usual
  - Missing values, factors as factors
- Describe the data
  - `summary`, `describe`, plots
- What's new: learn some new ways to look at the data
  - scatterplots, `corr.test`, `coplot`, `xyplot`

summary(data)

```
> summary(data)
  Intelligence        Height           SES             sex        
 Min.   : 86.74   Min.   :104.6   Min.   :1.400   Min.   :0.0000  
 1st Qu.: 96.62   1st Qu.:116.8   1st Qu.:2.700   1st Qu.:0.0000  
 Median : 99.91   Median :120.1   Median :3.000   Median :0.0000  
 Mean   :100.71   Mean   :120.4   Mean   :3.031   Mean   :0.4979  
 3rd Qu.:103.36   3rd Qu.:123.9   3rd Qu.:3.400   3rd Qu.:1.0000  
 Max.   :400.00   Max.   :250.0   Max.   :5.000   Max.   :1.0000  
                                                  NA's   :1
```

- Data has some obvious outliers
- Sex is not coded as a factor
- NA already coded in the data

describe(data)

```
> describe(data)
             var   n   mean    sd median trimmed  mad    min max  range  skew kurtosis   se
Intelligence   1 487 100.71 14.45  99.91  100.04 5.00  86.74 400 313.26 18.23   374.81 0.65
Height         2 487 120.43  7.89 120.06  120.19 5.34 104.60 250 145.40  9.05   147.22 0.36
SES            3 487   3.03  0.51   3.00    3.03 0.59   1.40   5   3.60  0.00     0.01 0.02
sex            4 486   0.50  0.50   0.00    0.50 0.00   0.00   1   1.00  0.01    -2.00 0.02
```

```
library(psych)

describe(data)
```

Examine trends visually and statistically

```
plot(data)
```

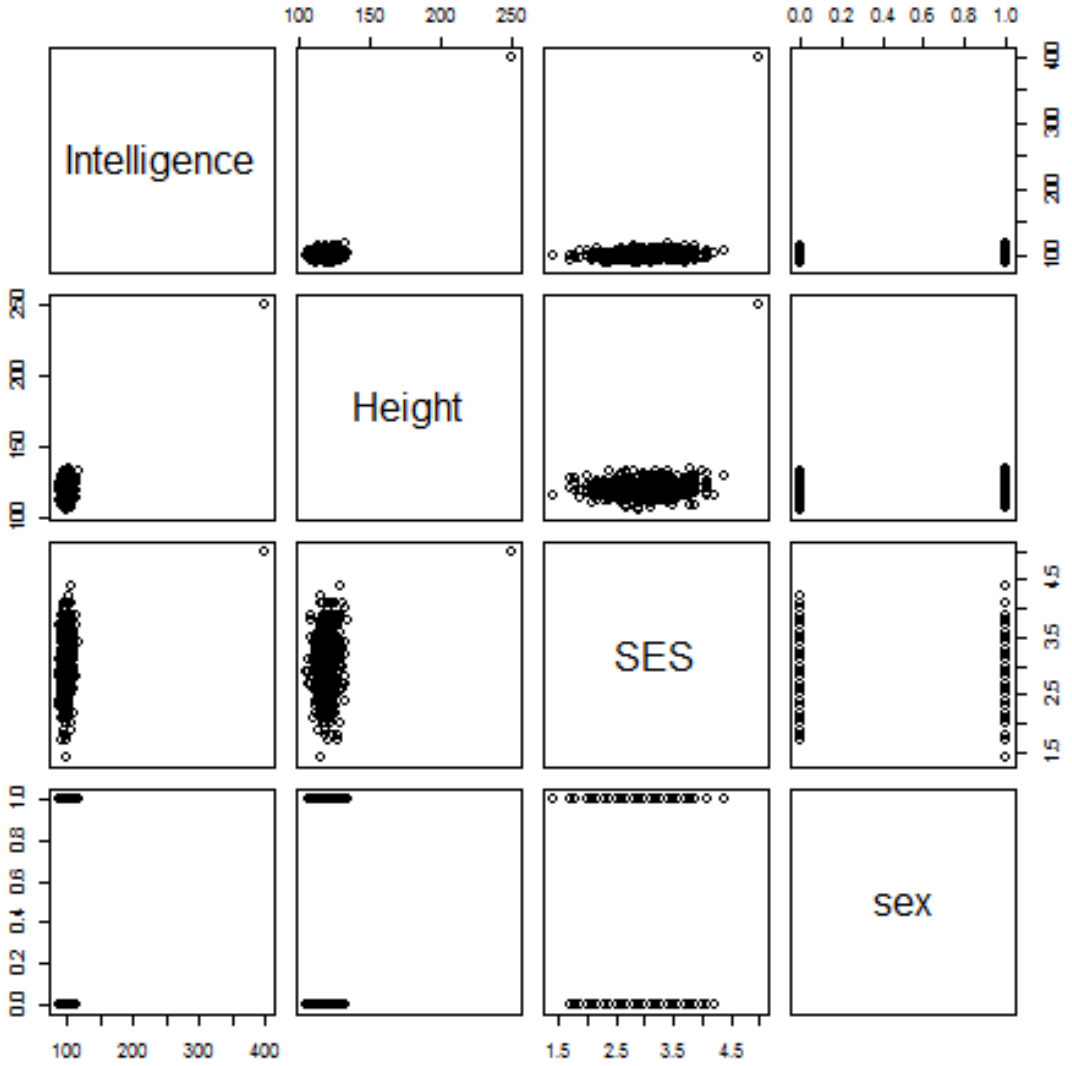
- What does our data look like? - scatterplot between all variables

```
corr.test(data)
```

- Simple correlation matrix is often an easy way to start identifying trends
- We learn more about correlations in a bit…

plot(data)

Looks like we have some serious outliers…



corr.test(data)

```
> corr.test(data)
Call:corr.test(x = data)
Correlation matrix
             Intelligence Height  SES   sex
Intelligence         1.00   0.72 0.22  0.00
Height               0.72   1.00 0.21 -0.01
SES                  0.22   0.21 1.00  0.03
sex                  0.00  -0.01 0.03  1.00
Sample Size
             Intelligence Height SES sex
Intelligence          487    487 487 486
Height                487    487 487 486
SES                   487    487 487 486
sex                   486    486 486 486
Probability values (Entries above the diagonal are adjusted for multiple tests.)
             Intelligence Height  SES sex
Intelligence         0.00   0.00 0.00   1
Height               0.00   0.00 0.00   1
SES                  0.00   0.00 0.00   1
sex                  0.96   0.91 0.58   0
```

- Correlation matrix
- Sample size
- p values

plot and corr.test

- The data is being heavily influenced by at least one outlier
- Height and intelligence are VERY strongly associated
- Sex still needs to be recoded

Preparing the data

```
data$sex <- factor(data$sex, levels=c(0,1), labels=c(”male”,
”female”)
```

```
which(Height > 200)
```

```
which(intelligence > 200)
```

```
which(is.na(data$sex))
```

- The problem is being caused by case 38

```
data2 <- data[-38,]
```

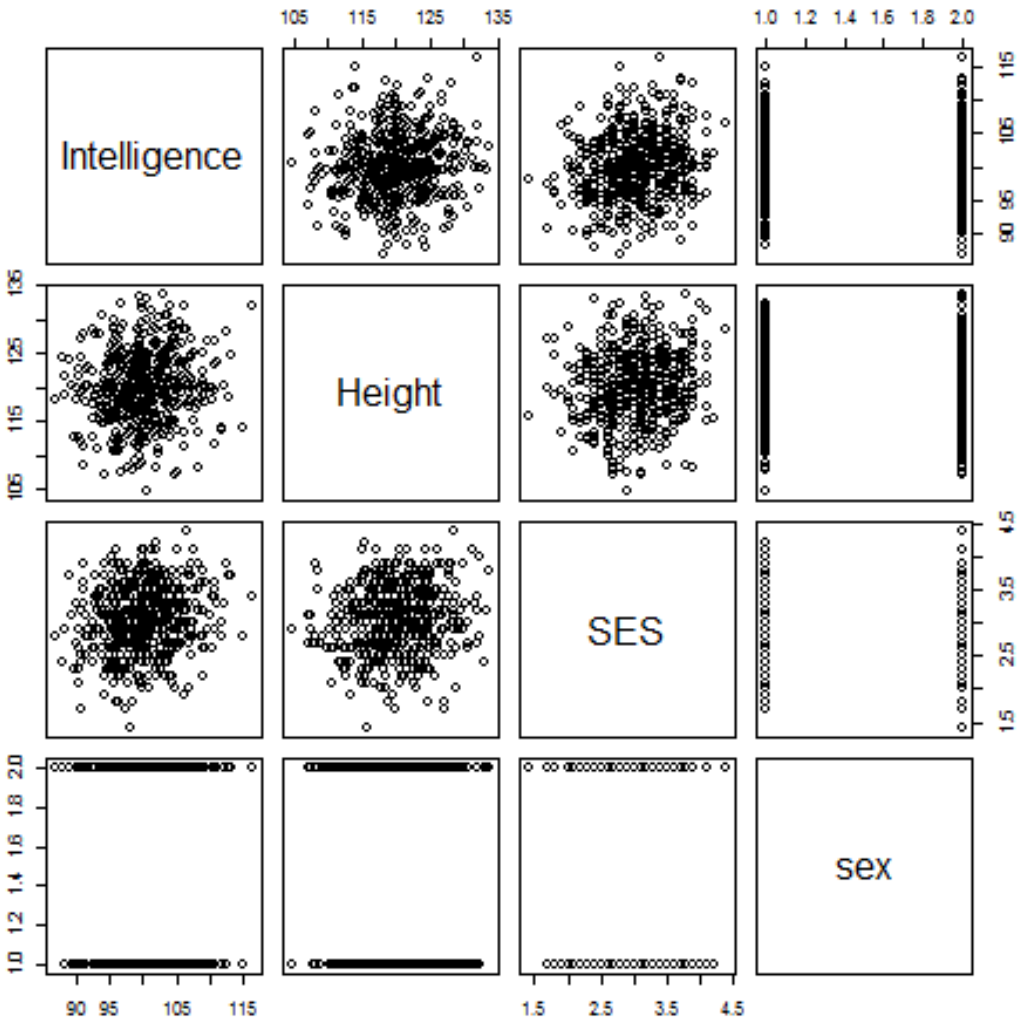
Examine data2

- `detach(data)`
- `attach(data2)`
- `plot(data2)`
- `corr.test(data2)`

- Evaluate impact of dropping the case visually
- You can if you wish rerun all analyses on data and data2, for comparison

plot(data2)

No more extreme outliers distorting our data



corr.test comparison

- data

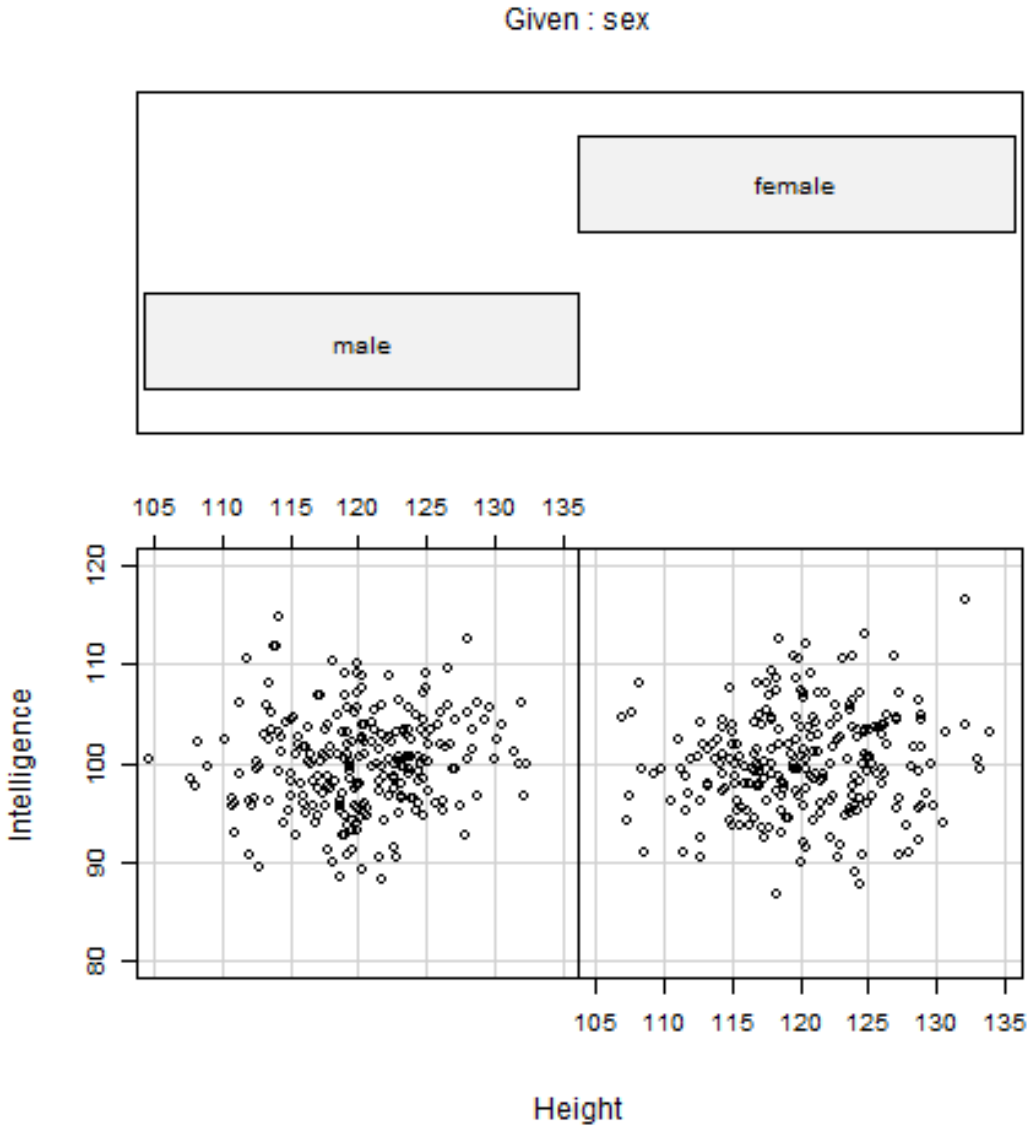
	Intelligence	Height	SES
Intelligence	1.00	0.72	0.22
Height	0.72	1.00	0.21
SES	0.22	0.21	1.00

- data2

	Intelligence	Height	SES
Intelligence	1.00	0.09	0.16
Height	0.09	1.00	0.13
SES	0.16	0.13	1.00

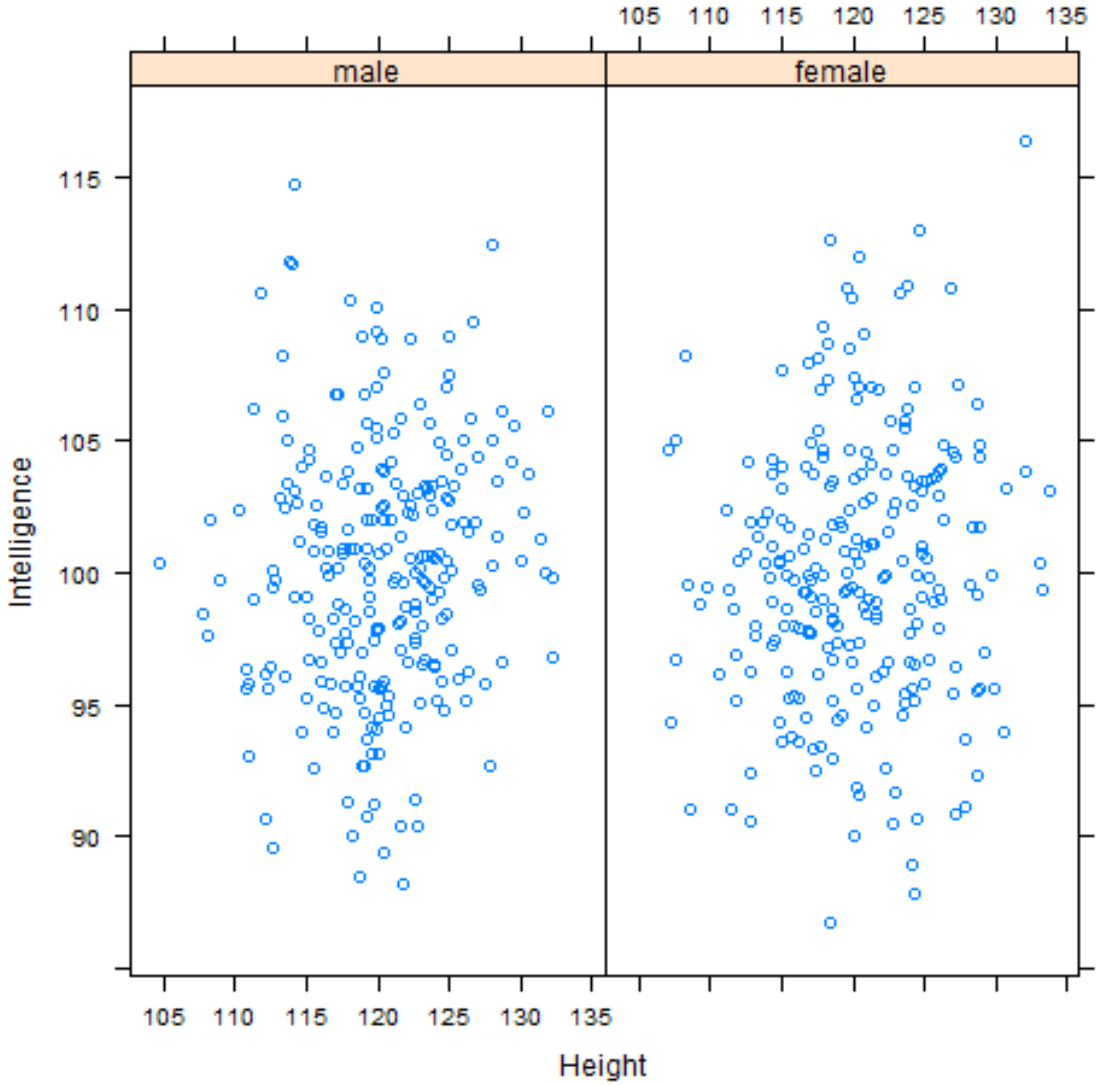
Conditioning plots

```
coplot(Intelligence
~ Height | sex,
xlim=c(105,135)
ylim=c(80,120))
```



Or ...

```
library(lattice)

xyplot(Intelligence ~
Height | sex)
```



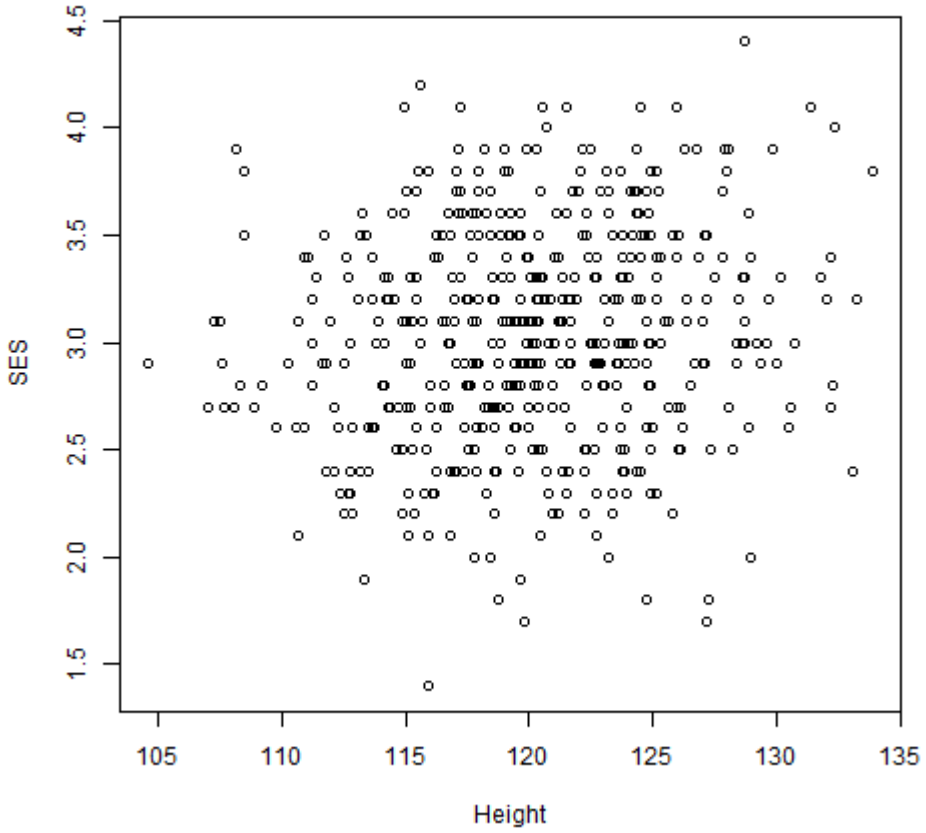
2. Correlation

- Covariance and correlations

- Different correlation coefficients:
  - Pearson
  - Spearman
  - Partial

Visualization: scatterplots

```
plot(data)

plot(SES ~ Height)
```



Pearson product-moment correlation coefficient

- Covariance and correlation between two variables

```
cov(SES, Intelligence)
```

```
cor(SES, Intelligence)
```

```
cor.test(SES, Intelligence)
```

  - Tests for null hypothesis that correlation = 0
- For the whole data frame

```
corr.test(data2)
```

- Gives Pearson's correlation coefficient by default

Pearson product-moment correlation coefficient

- Gives
  - Correlation coefficient
  - Absolute value, positive vs negative correlation
  - T test statistics and p value
- Pairwise deletion of cases is default
  - If want to select just complete cases, use ”complete”
- Adjust for multiple tests
  - Modify tests by changing ”adjust”

Pearson product-moment correlation coefficient

```
> corr.test(data)
Call:corr.test(x = data)
Correlation matrix
             Intelligence Height  SES   sex
Intelligence         1.00   0.72 0.22  0.00
Height               0.72   1.00 0.21 -0.01
SES                  0.22   0.21 1.00  0.03
sex                  0.00  -0.01 0.03  1.00
Sample Size
             Intelligence Height SES sex
Intelligence          487    487 487 486
Height                487    487 487 486
SES                   487    487 487 486
sex                   486    486 486 486
Probability values (Entries above the diagonal are adjusted for multiple tests.)
             Intelligence Height  SES sex
Intelligence         0.00   0.00 0.00   1
Height               0.00   0.00 0.00   1
SES                  0.00   0.00 0.00   1
sex                  0.96   0.91 0.58   0
```

- **p values:**
  raw values below diagonal, values above diagonal are adjusted for multiple tests

Spearman's rank correlation coefficient

- Similar, but now add ”method” as a named argument:

```
cor.test(SES, Intelligence, method=”spearman”)
corr.test(data2, method=”spearman”)
```

- Same details:
  - Correlation coefficient
  - Result of t test

Partial correlations

- `pcor.test(x,y,z,)` from `ppcor` package

```
pcor.test(SES, Intelligence, Height)
```

  - Partial correlation between SES and Intelligence while controlling for Height

- Results:
  - Coefficient, p value, t value

3. Regression models

- Concepts
- How to build up your model
- Simple model with one predictor
  - Testing for assumptions
- Another model with two predictors
  - Model comparison
- Modeling interactions and polynomials

Linear regression

- ANOVA is just a special case of a linear model
  - Where predictor is categorical
- In fact, the information stored by R in both cases is similar – just the standard output differs:
  - ANOVA:

```
summary(anova.model)          # ANOVA table
summary.lm(anova.model)       # regression table
```

  - Regression:

```
summary(reg.model)            # regression table
anova(reg.model)              # ANOVA table
```

Linear regression

- We can predict any data using the following general equation:

$$outcome_i = model + error_i$$

- Describing a straight line:

$$y_i = b_0 + b_1 X_i + \varepsilon_i$$

b_i
- Regression coefficient for the predictor, slope of the regression line

b_0
- Intercept (value of Y when X = 0)

Linear regression

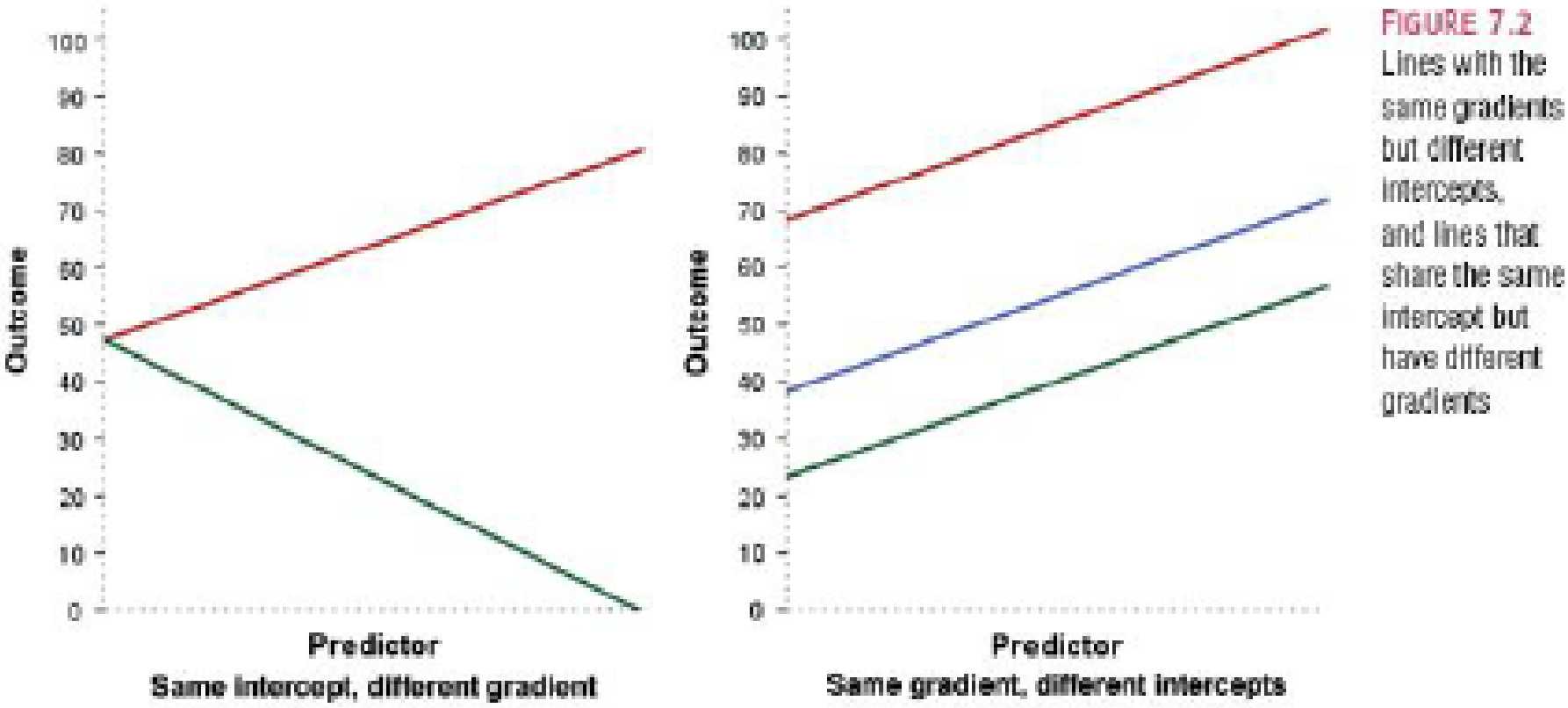


FIGURE 7.2
Lines with the same gradients but different intercepts, and lines that share the same intercept but have different gradients

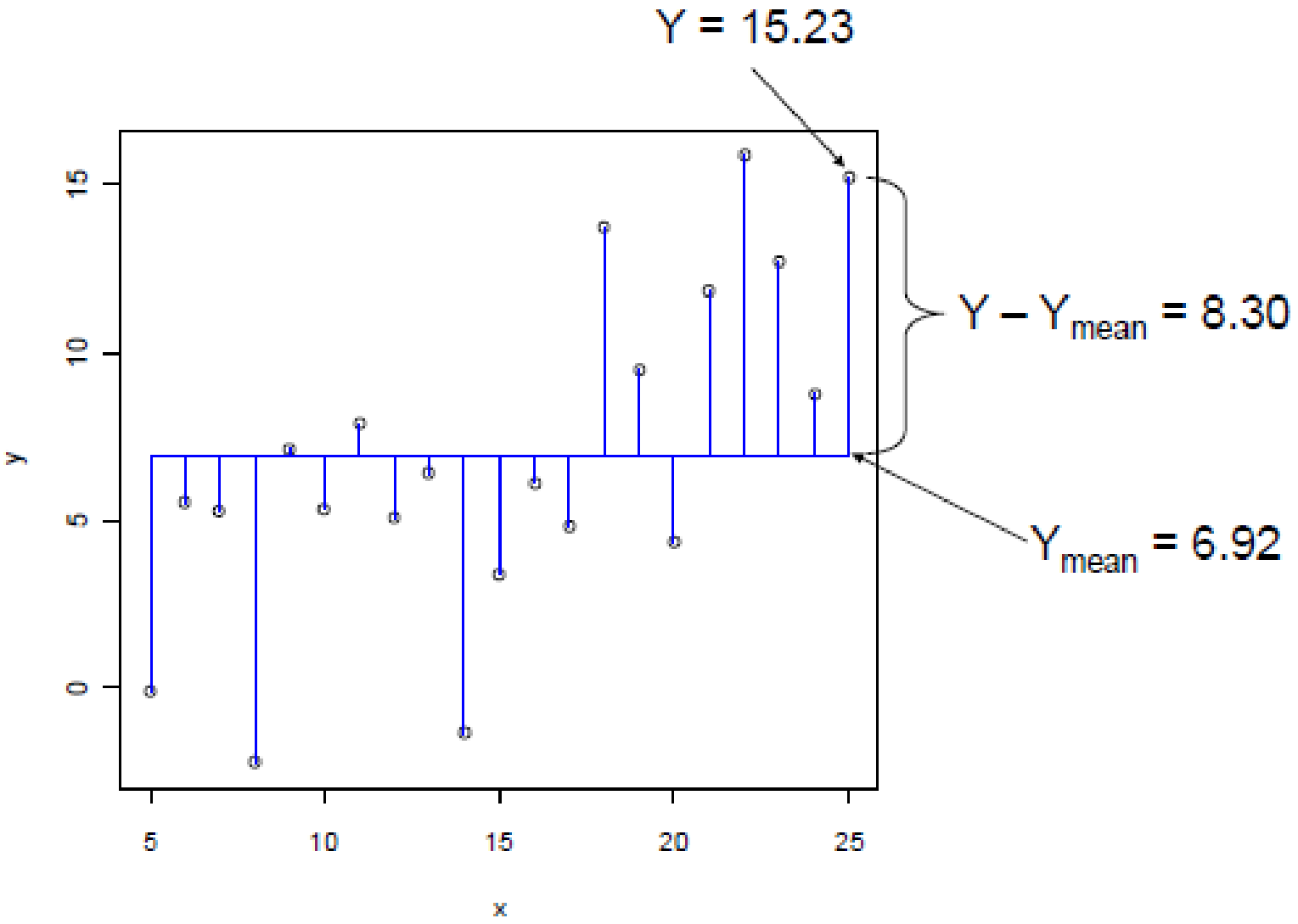
Field (2009)

Linear regression

- Idea:
  - Fit a regression line to the data
    - The regression line summarizes or models your observations
    - Predicts a value on the outcome variable (Y_pred) for each value of a single observed variable, X
- How?
  - Method of least squares – provides the regression line in which the sum of squared differences between the observed values and the values predicted by the model is as small as possible
    - $\sum(Y - Y_{pred})^2$

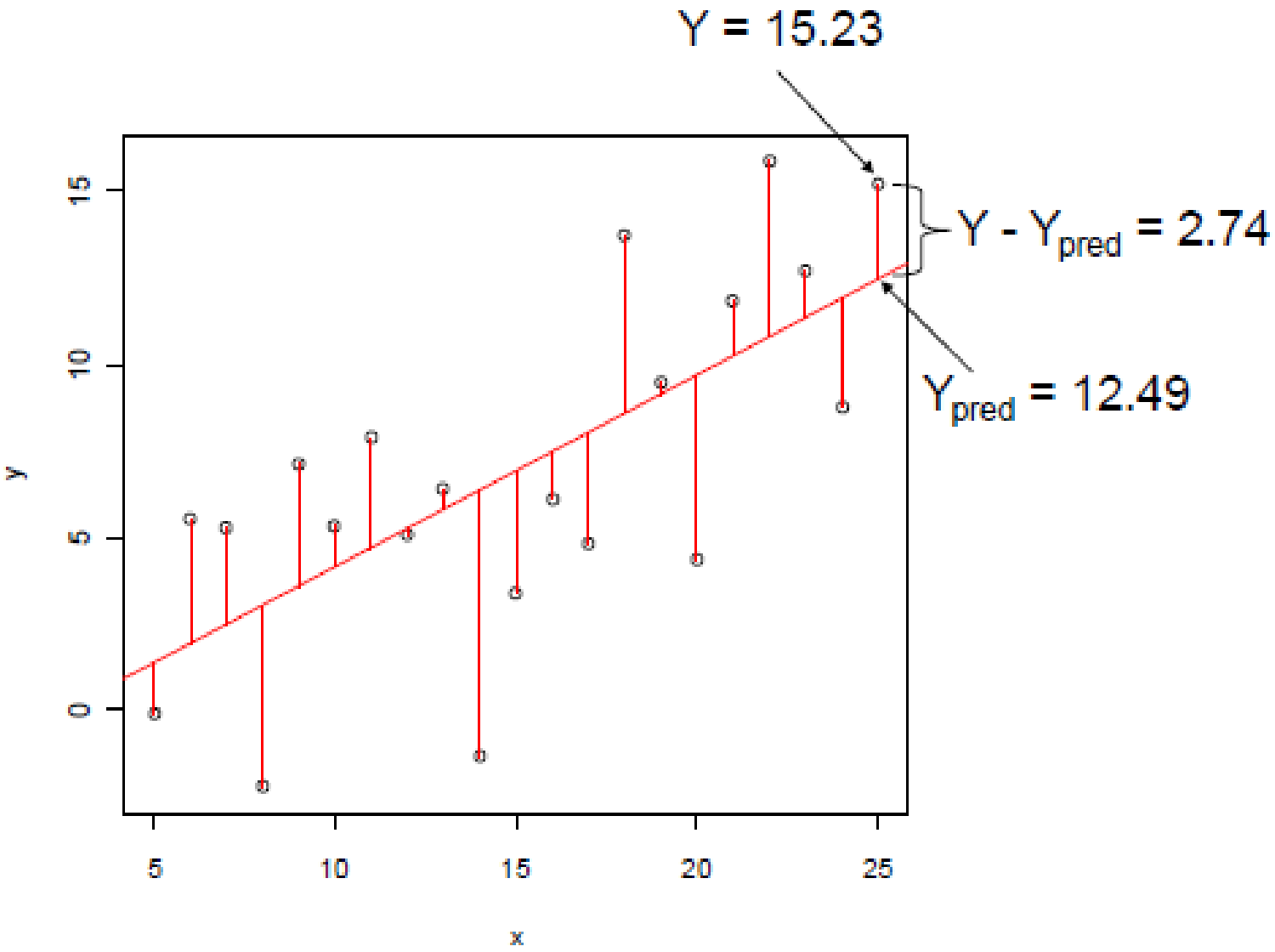
Sums of squares

- Total sums of squares (SS_total)
  - Sum of squared differences between observed values of Y and the mean of Y
  - $SS_{total} = \sum(Y - Y_{mean})^2$


Y = 15.23
Y – Y mean = 8.30
Y mean = 6.92
y
x
0
5
10
15
5
10
15
20
25

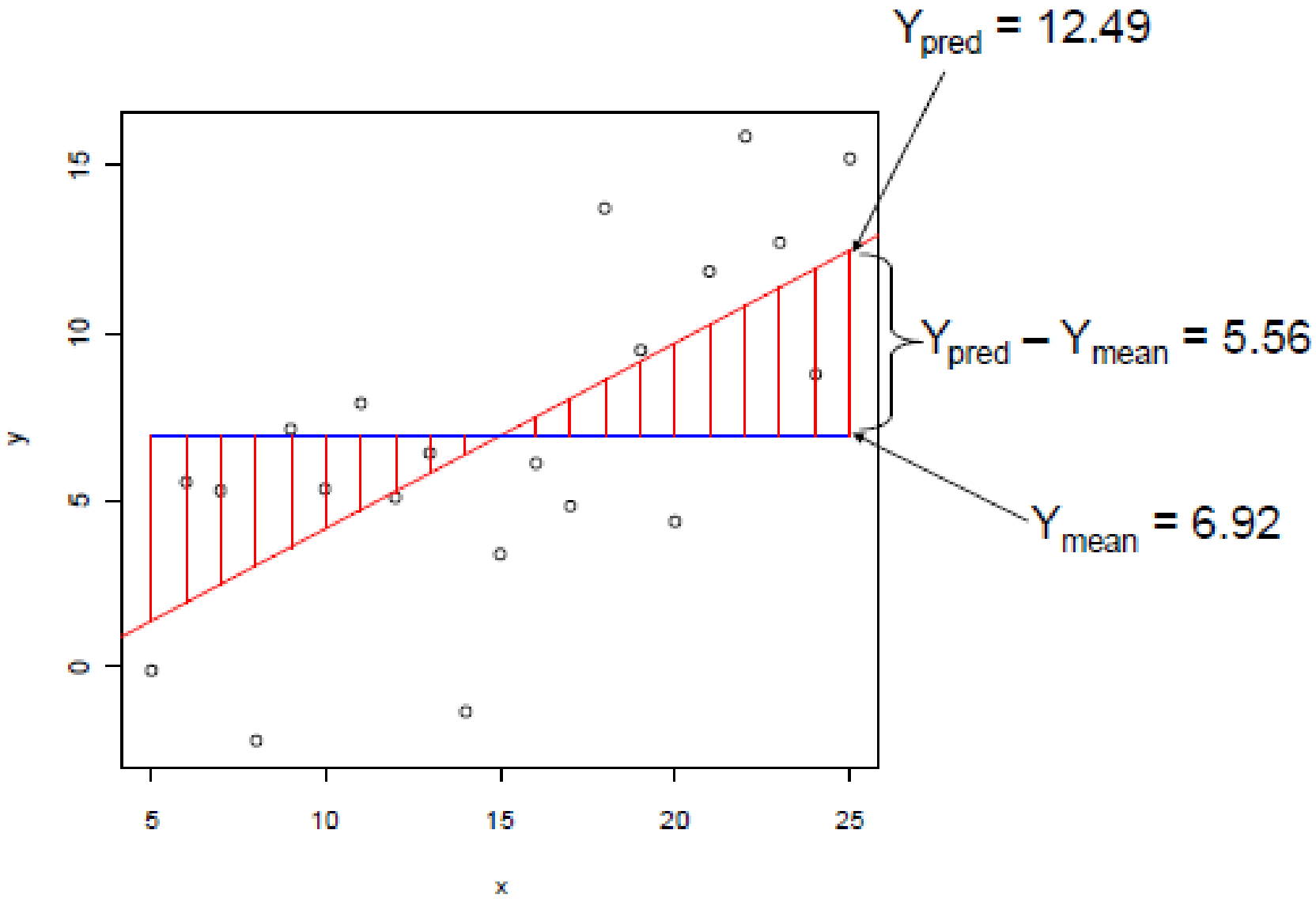
Sums of squares

- Total sums of squares (SS_total)
  - Sum of squared differences between observed values of Y and the mean of Y
  - $SS_{total} = \sum(Y - Y_{mean})^2$
- Residual sums of squares (SS_residual)
  - Sum of squared differences between observed values of Y and predicted values of Y
  - $SS_{residual} = \sum\left(Y - Y_{pred}\right)^2$


Y = 15.23
Y - Ypred = 2.74
Ypred = 12.49
y
x
0
5
10
15
5
10
15
20
25

Sums of squares

- Total sums of squares (SS_total)
  - Sum of squared differences between observed values of Y and the mean of Y
  - $SS_{total} = \sum(Y - Y_{mean})^2$
- Residual sums of squares (SS_residual)
  - Sum of squared differences between observed values of Y and predicted values of Y
  - $SS_{residual} = \sum(Y - Y_{pred})^2$
- Model sums of squares (SS_model)
  - Sum of squared differences between predicted values of Y and the mean of Y
  - $SS_{model} = \sum(Y_{pred} - Y_{mean})^2$


Y_{pred} = 12.49
$Y_{pred} - Y_{mean}$ = 5.56
Y_{mean} = 6.92
y
x
0
5
10
15
5
10
15
20
25

Sums of squares

- Total sums of squares (SS_total)
  - Sum of squared differences between observed values of Y and the mean of Y
  - $SS_{total} = \sum(Y - Y_{mean})^2$
- Residual sums of squares (SS_residual)
  - Sum of squared differences between observed values of Y and predicted values of Y
  - $SS_{residual} = \sum(Y - Y_{pred})^2$
- Model sums of squares (SS_model)
  - Sum of squared differences between predicted values of Y and the mean of Y
  - $SS_{model} = \sum(Y_{pred} - Y_{mean})^2$ = SS_total – SS_residual

Linear regression

- Testing the model: R^2
  - The proportion of variance accounted for by the regression model
  - $R^2 = \frac{SS_M}{SS_T}$
  - Indicates how much the model improves the prediction of Y over just using the mean of Y

- (in ANOVA: how much does adding a separate group mean improve the prediction of Y over just using the same mean for all groups)

Regression models

- R uses function `lm`

```
y ~ A                 # A is continuous
y ~ as.factor(A)      # A is categorical
y ~ A + B             # models A and B main effects
y ~ A + B + A:B       # models A, B, and their interaction
y ~ A*B               # shortcut for: y ~ A + B + A:B
```

- Create variables for polynomial regression

```
A_2 = a^2
A_3 = a^3
```

- Save results of `lm()`:

```
model1 <- lm()
```

Relationship with ANOVA

- Looks familiar?

- This is because ANOVA is part of the same family: it is just a special case of general linear model
  - Where predictor is categorical variable

Simple model

```
model1 <- lm(Height ~ SES)
```

- SES is our only predictor
- Data about the model stored in model1

```
summary(model1)

anova(model1)

coefficients(model1)

confint(model1, level=0.95)

plot(model1)

fitted(model1)
resid(model1)
```

summary(model1)

```
> summary(model1)

Call:
lm(formula = Height ~ SES, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3935  -3.4263  -0.1277   3.6421  13.7291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.1517     1.4414  80.583  < 2e-16 ***
SES           1.3240     0.4697   2.819  0.00502 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.219 on 484 degrees of freedom
Multiple R-squared: 0.01615,    Adjusted R-squared: 0.01412
F-statistic: 7.946 on 1 and 484 DF,  p-value: 0.005018
```

summary(model1)

- Gives data on coefficients
  - beta_0 = intercept
  - beta_1 = gradient
  - As in…

    y = beta_0 + beta_1 * x
- R^2 and adjusted R^2
- Significance of predictors
- F-ratio

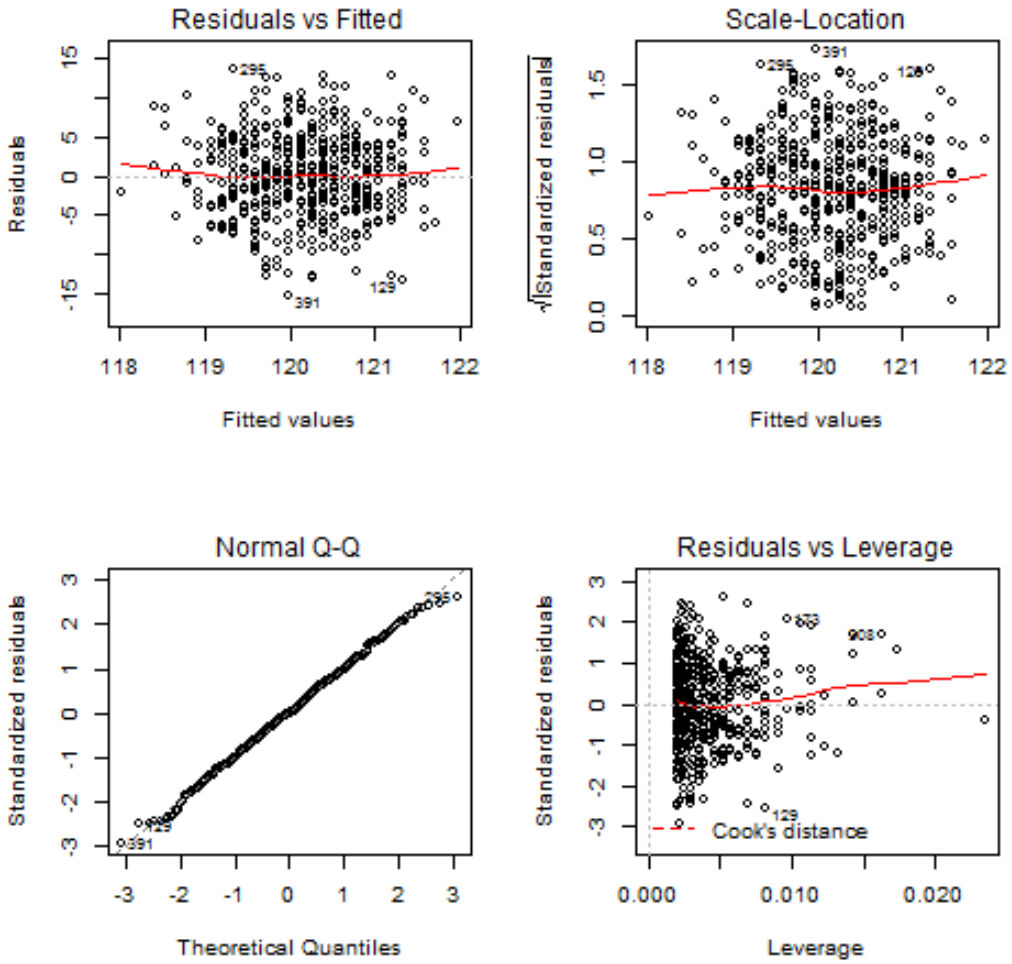
anova(model1)

```
> anova(model1)
Analysis of Variance Table

Response: Height
           Df  Sum Sq Mean Sq F value    Pr(>F)
SES         1   216.4 216.433  7.9457 0.005018 **
Residuals 484 13183.7  27.239
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Sums of squares
- DF
- `anova` is very useful when you want to COMPARE models

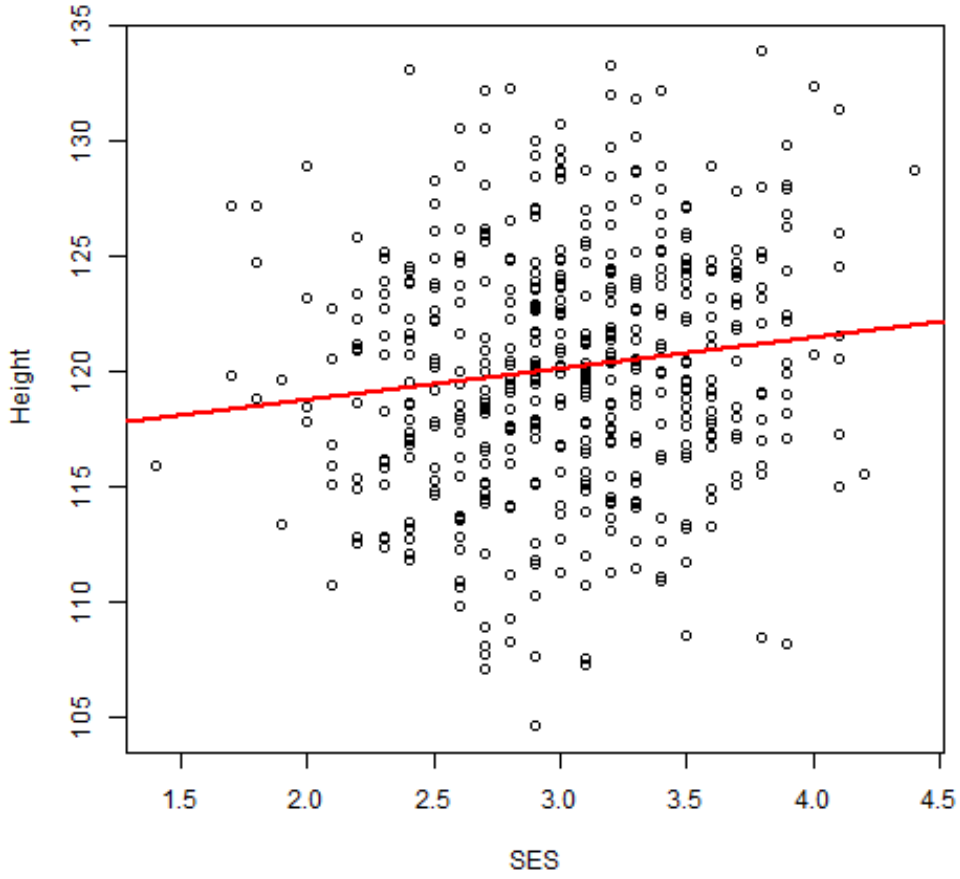
plot(model1)

```
layout(matrix(c(1,2,3,4),2,2))
plot(model1)
```



Plotting the regression line

```
plot(Height~SES)

abline(model1,
col=”red”, lwd=2)
```



Testing the assumptions

- Influential cases
- Multicollinearity
- Homoscedasticity
- Linearity

- Parsimonity!
  - More complex models are penalized in adjusted R^2

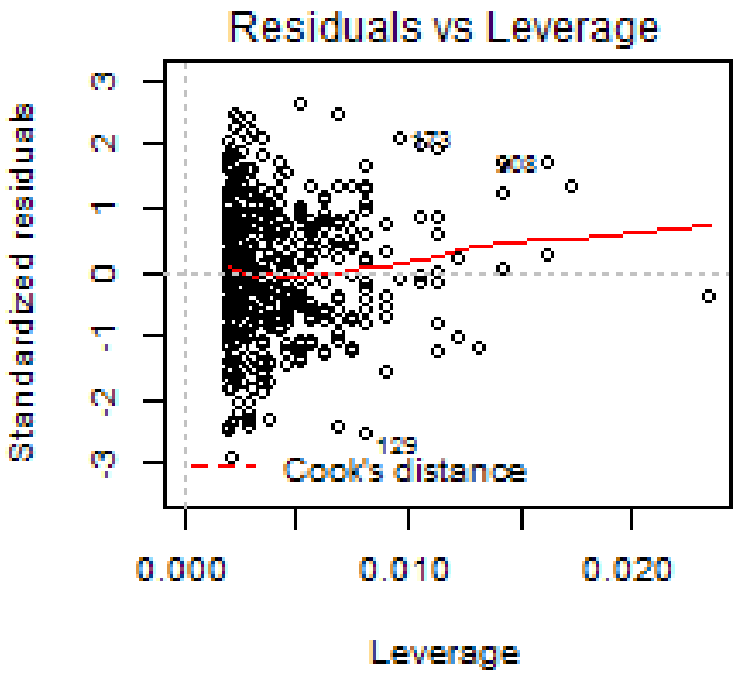
Influential cases

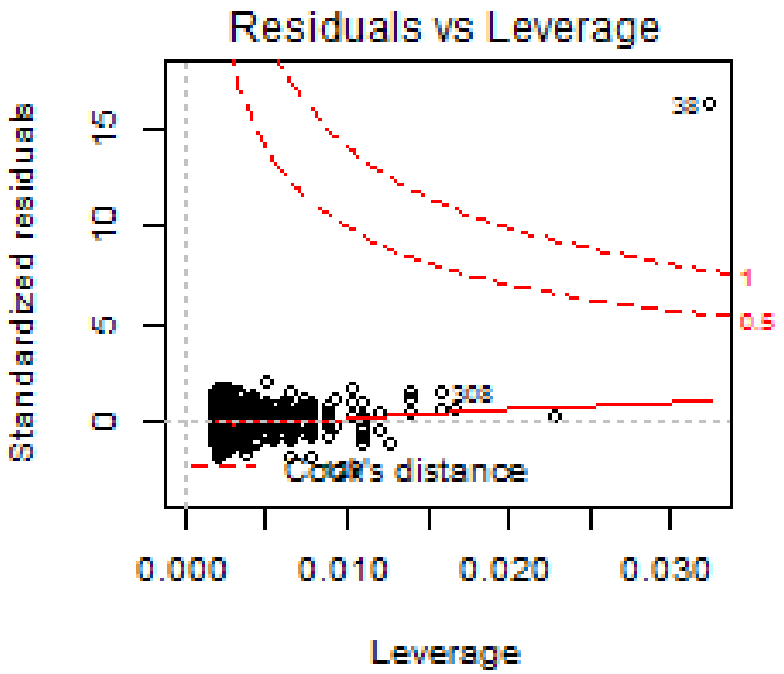
- Cook's distances

```
influence.measures(model1)
```

  - Tests which cases have a large impact on the model
  - Can also be identified graphically
  - Cases above 0.5 or 1 may be problematic, though some suggest sample size must be taken into account
  - Note case 38 in "data"

data vs data2





Multicollinearity

- Independent variables should not be highly intercorrelated
- Can be dealt with by averaging the variables, or by factor analysis

- You can use the correlation matrix to examine this.

Homoscedasticity

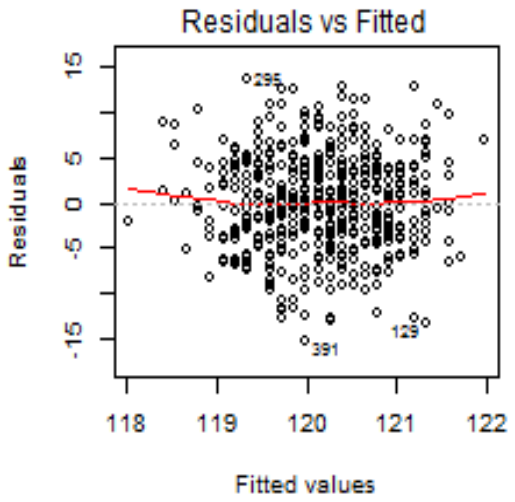
- Variance should be homogenous

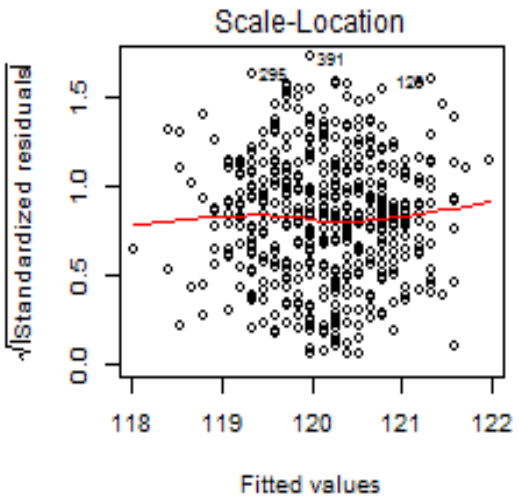
```
library(car)

ncvTest(model1)
```

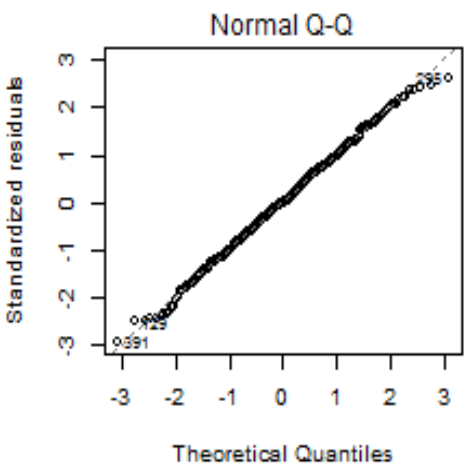
- Results should be non-significant
- Significant results indicate regression model may be biased
  - In this case, transformation of data may help
- Residuals vs. fitted plots are also useful

```
> ncvTest(model1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 0.0009664605    Df = 1     p = 0.9751994
```

Homoscedasticity







Linearity

- Can be identified from histograms
- Or take QQ-plot separately
- If the data points are distributed equally around the horizontal line, data is likely linear
- If violated, non-linear transformation may help

Model interpretation

```
> summary(model1)

Call:
lm(formula = Height ~ SES, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.3935  -3.4263  -0.1277   3.6421  13.7291

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 116.1517     1.4414  80.583  < 2e-16 ***
SES           1.3240     0.4697   2.819  0.00502 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.219 on 484 degrees of freedom
Multiple R-squared: 0.01615,    Adjusted R-squared: 0.01412
F-statistic: 7.946 on 1 and 484 DF,  p-value: 0.005018
```

Model interpretation

- SES predicts height significantly
- Look at R^2 for effect size
- What about adding another predictor?

Model 2

```
model2 <- lm(Height ~ SES+Intelligence)
```

- Here, SES is entered into the model first, followed by Intelligence
- We can test the effect of intelligence after accounting for SES
- We can test whether including intelligence makes for a better overall model

summary(model2)

```
> summary(model2)

Call:
lm(formula = Height ~ SES + Intelligence, data = data2)

Residuals:
     Min       1Q   Median       3Q      Max
-15.4323  -3.5151  -0.1559   3.8376  13.6279

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  108.43589    4.86348  22.296   <2e-16 ***
SES            1.19758    0.47501   2.521   0.0120 *
Intelligence   0.08091    0.04872   1.661   0.0974 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.21 on 483 degrees of freedom
Multiple R-squared: 0.02174,    Adjusted R-squared: 0.01769
F-statistic: 5.366 on 2 and 483 DF,  p-value: 0.004954
```

Model 2 interpretation

- Intelligence is not a significant predictor of height
- Does it improve our model?
- Use `anova(model1,model2)` and see if the extra predictor significanlty improves the model

anova(model1,model2)

```
> anova(model1,model2)
Analysis of Variance Table

Model 1: Height ~ SES
Model 2: Height ~ SES + Intelligence
  Res.Df   RSS Df Sum of Sq      F Pr(>F)
1    484 13184
2    483 13109  1    74.861 2.7583 0.0974 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Final model

- SES, but not intelligence, appears to affect height
- Despite correlating with height, intelligence's contribution to height can be explained by SES
- BUT remember the order of our predictors were determined by theory, not statistically
- So some researchers might claim the order of entry was incorrect!

Interactions and polynomials

- You may want to test interactions as well as main effects
- Polynomials help when a straight line predicts the data poorly
- Evaluating graphs and simple linear models may suggest interactions and polynomials

Testing interactions

```
model3 <- lm(Height ~ SES*Intelligence)
```

- Tests
  - Main effect of SES
  - Main effect of Intelligence
  - Interaction between SES and Intelligence
- Interactions go AFTER main effects

- Again, `summary(model3), anova(model1,model3)`

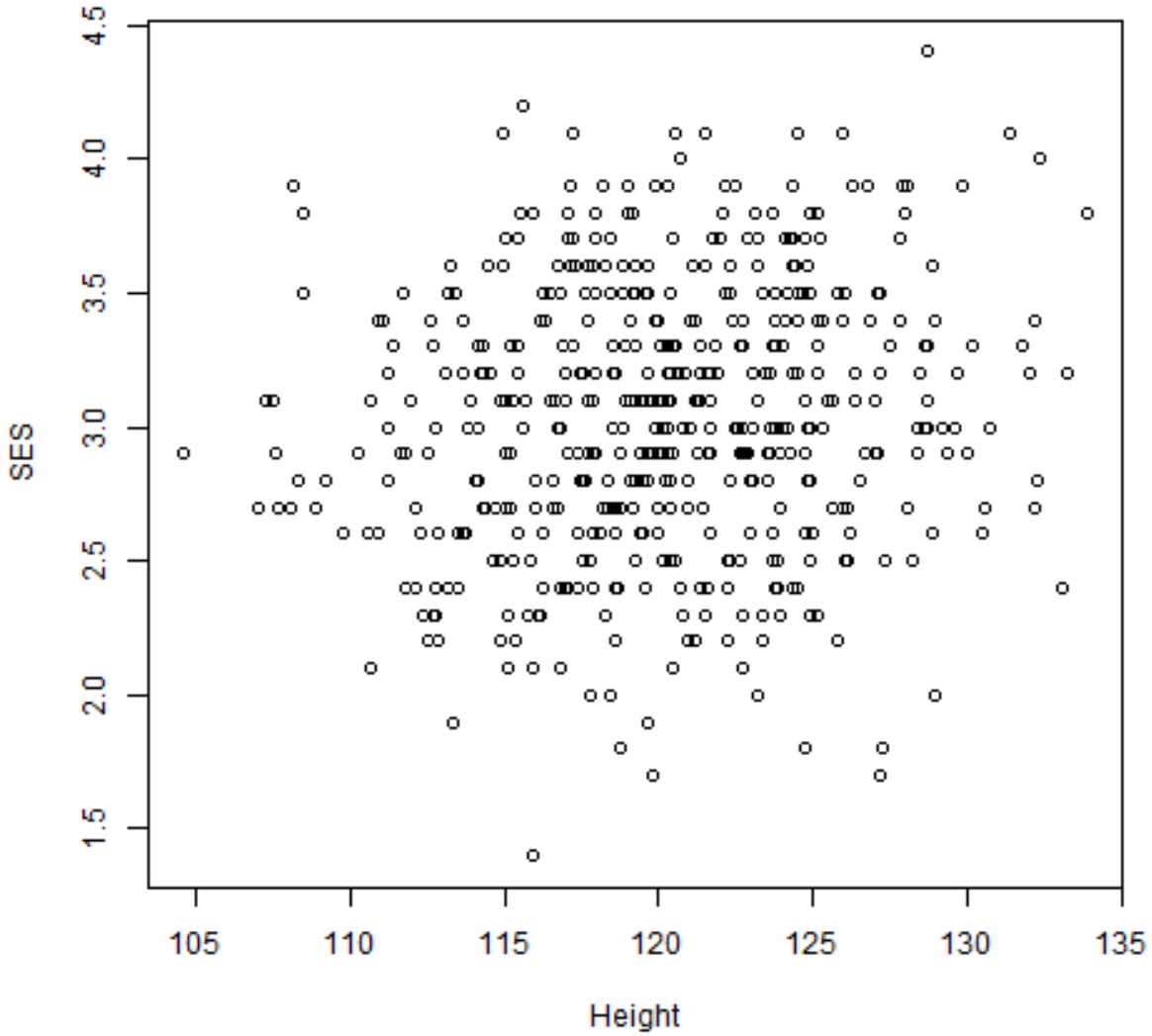
Polynomials

- Is the trend linear?
- Start by creating a new variable where you square each data point, modelling a quadratic (curved) line

```
data2$SES2 <- data2$SES^2

model4 <- lm(Height ~ SES+data2$SES2)
```

- Add higher polynomials until you reach one that is non-significant
- Here, the quadratic function is non-significant… indicating the trend is linear
- This just confirms what the plots tell us

Polynomials



4. Contrasts

- Planned comparisons between a subset of categories in categorical variables which have > 3 levels for ANOVA or linear models
- You can define *k-1* contrasts (where *k* is the number of levels in your category)

4. Contrasts

- E.g., back to our Education groups from last week (comprehensive school, secondary, higher)

- Formulate your hypothesis:
  What could meaningful group comparisons include?
  - Example 1: compare higher education to other levels of education
  - Example 2: compare secondary education to comprehensive school

Logic behind contrasts

Category	Contrast 1	Contrast 2
Comprehensive school	-0.5	-1
Secondary	-0.5	1
Higher	1	0

- Contrast 1: compare higher education to other levels of education
- Contrast 2: compare secondary education to comprehensive school
- Note that contrasts 1 and 2 are orthogonal, so they can be added to the same analysis.

Setting up contrasts: Example

Check the levels of a categorical variable:

```
> levels(education)
"1"  "2"  "3"
```

Save and check contrasts for the categorical variable

```
> contrasts(education) <- cbind(c(-0.5,-0.5,1),c(-1,1,0),)
> contrasts(education)
```

Once a contrast is saved, it will be added to anova automatically

```
> anova.model <- aov(digitsymbol_1 ~ education)
> summary.lm(anova.model)
```

Data for today

Variable	Type	Description
Performance	**continuous**	**Exam mark**
Hours	continuous	Hours of class missed
Educ	categorical	Years of education prior to course
Rating	continuous	Student rating of the course
Entry	continuous	Score on exam taken in the first week of the course
Extra	continuous	Additional work
Stress	continuous	Self-reported stress level