Statistics for CSAI II

5 - Regression

Dr. Travis J. Wiltshire |

TILBURG UNIVERSITY

Modules

1. Introduction and Probability
2. Sampling Theory
3. Revisiting Hypothesis Testing & Intro to Correlation
4. Correlation
5. *Intro to Regression*
6. More Regression Centering and Checking Assumptions
7. Multiple Regression and Assumptions
8. Interactions
9. Multiple Regression with Categories
10. Multiple Regression with Polynomials
11. Mixed Models
12. Growth Curve Analysis

PE Updates

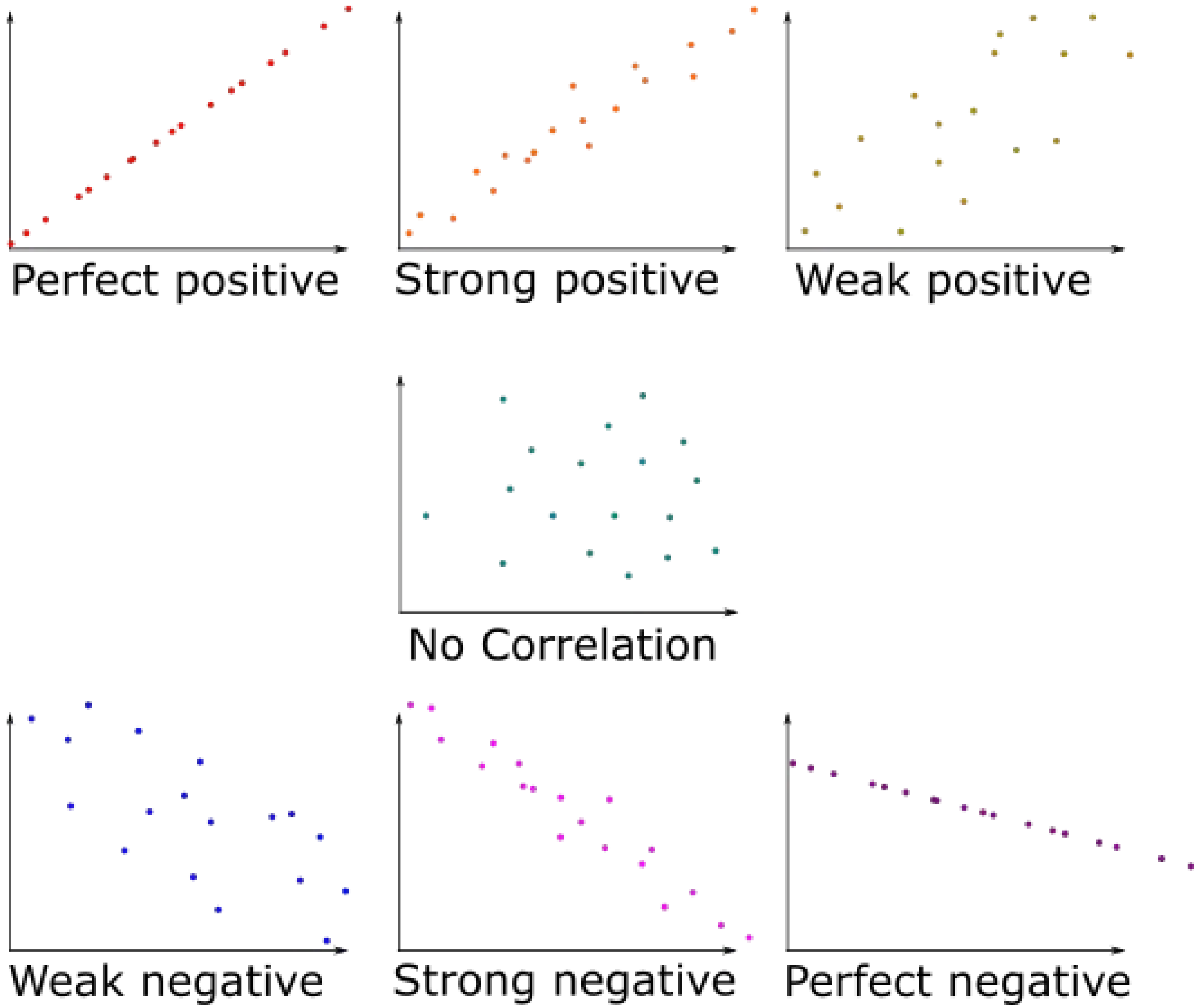
- We strongly encourage you to attend the in person practical sessions as there is opportunity for interaction and feedback, but are allowed to complete the assignments remotely (within the allotted time windows)
- You must complete a majority of the assignment and show a valid attempt at completing each task
- Files must be knitted correctly to be acceptable (no screen shots, word files, etc.). Ensuring you are set up to knit correctly is your responsibility. You can comment out code that is giving your trouble with knitting and still show it if you want us to consider it a valid attempt.

WOOCLAP

Outline

1. Linear regression with one predictor
2. How to assess the fit of a regression model?
   - Total sum of squares
   - Model sum of squares
   - Residual sum of squares
   - F
   - R^2
3. How to do regression using **R**?
4. Interpreting the output of a regression model

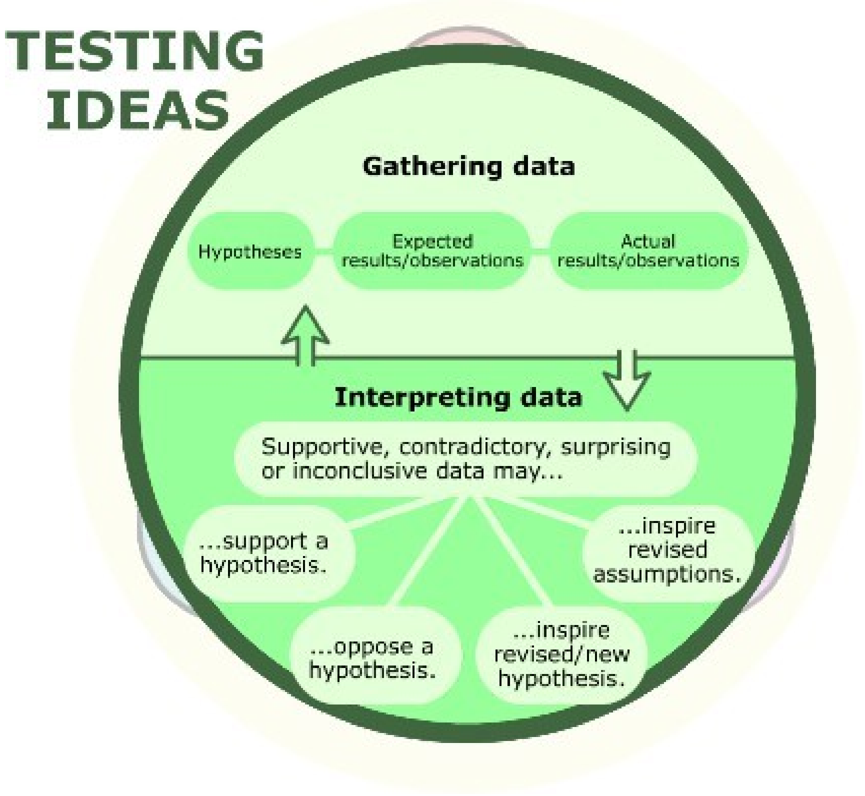
Different types of correlation relationships



Hypothesis Testing and a Model Based Approach

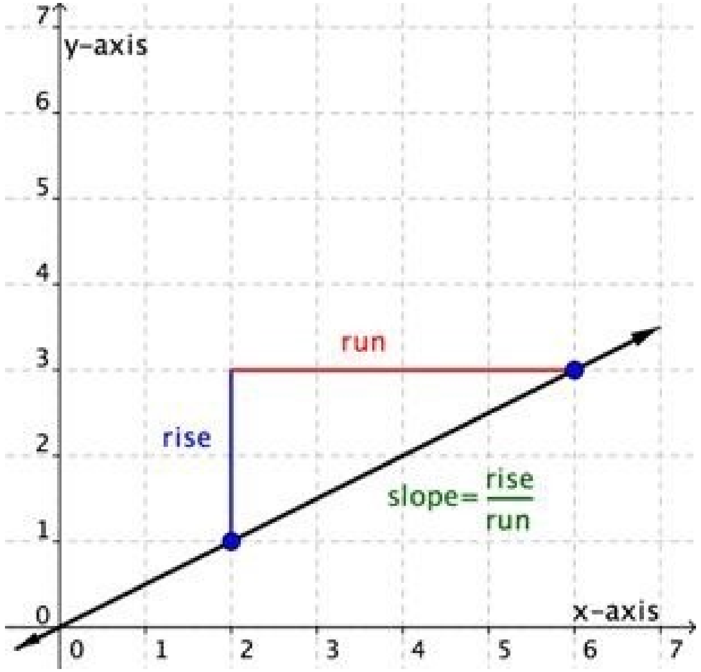
Data = Model predictions + Error

- Comparison (Relationship between variables) ⇒ Correlation
- Prediction (predicting outcomes vs predicting relationship vs using the model for prediction) ⇒ Regression
- **H_o**: no relationship between data and model


TESTING IDEAS
Gathering data
Hypotheses
Expected results/observations
Actual results/observations
Interpreting data
Supportive, contradictory, surprising or inconclusive data may...
...support a hypothesis.
...oppose a hypothesis.
...inspire revised/new hypothesis.
...inspire revised assumptions.

What is Regression?

- A way of predicting the value of one variable from another.
  - It is a hypothetical **model** of the relationship between variables.
    - Dependent variable (y) and one or more independent variables (x)
    - Typically, both are interval scale variables
  - The model used is a **linear** one.
  - Therefore, we describe the relationship using **the equation of a straight line**.



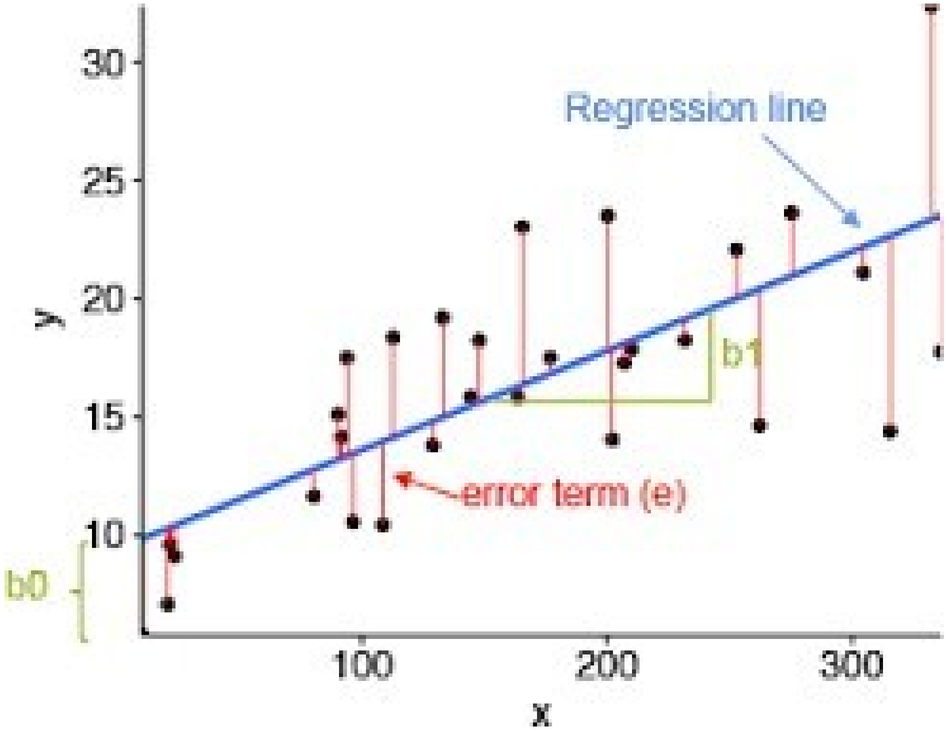
$$y = \underbrace{m}_{\text{slope}}x + \underbrace{b}_{\text{y-intercept}}$$

$$y = \underbrace{3}_{\text{slope}}x - \underbrace{5}_{\text{y-intercept}}$$

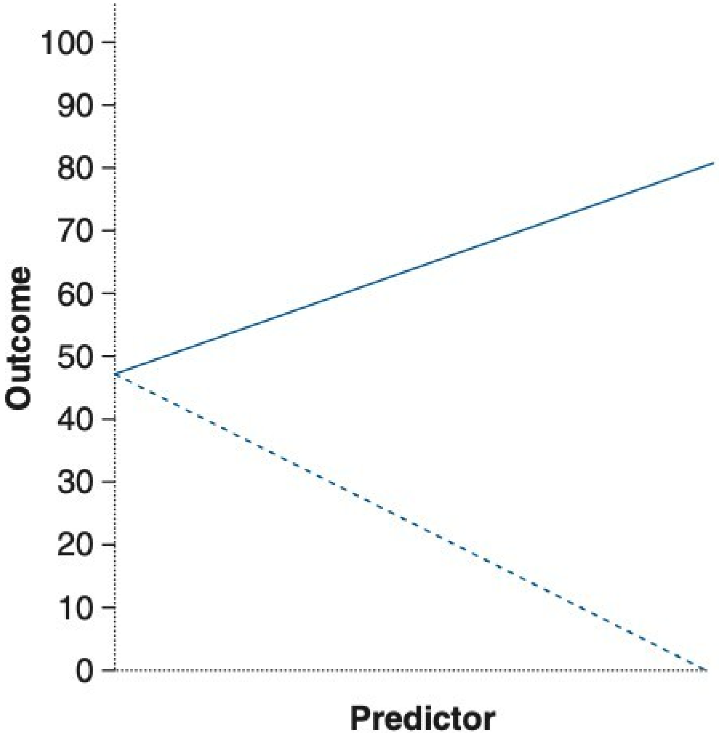
Describing a Straight Line

$$Y_i = b_0 + b_i X_i +$$

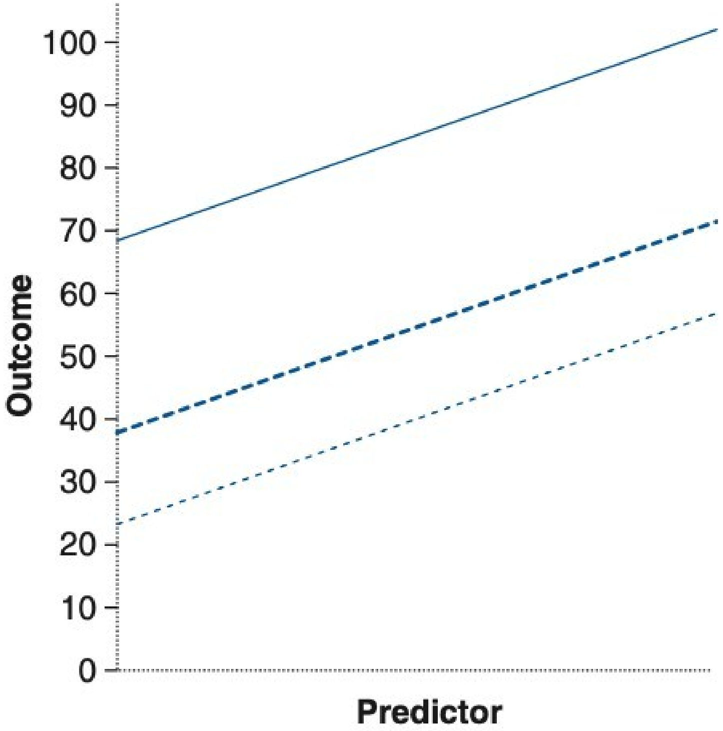
- b_i χ_i
  - Regression coefficient for the predictor
  - Gradient (**slope**) of the regression line
  - Direction/strength of relationship
- b_o
  - Intercept (value of *Y* when *X* = 0)
  - Point at which the regression line crosses the *Y*-axis (ordinate)



Gradients and Intercepts

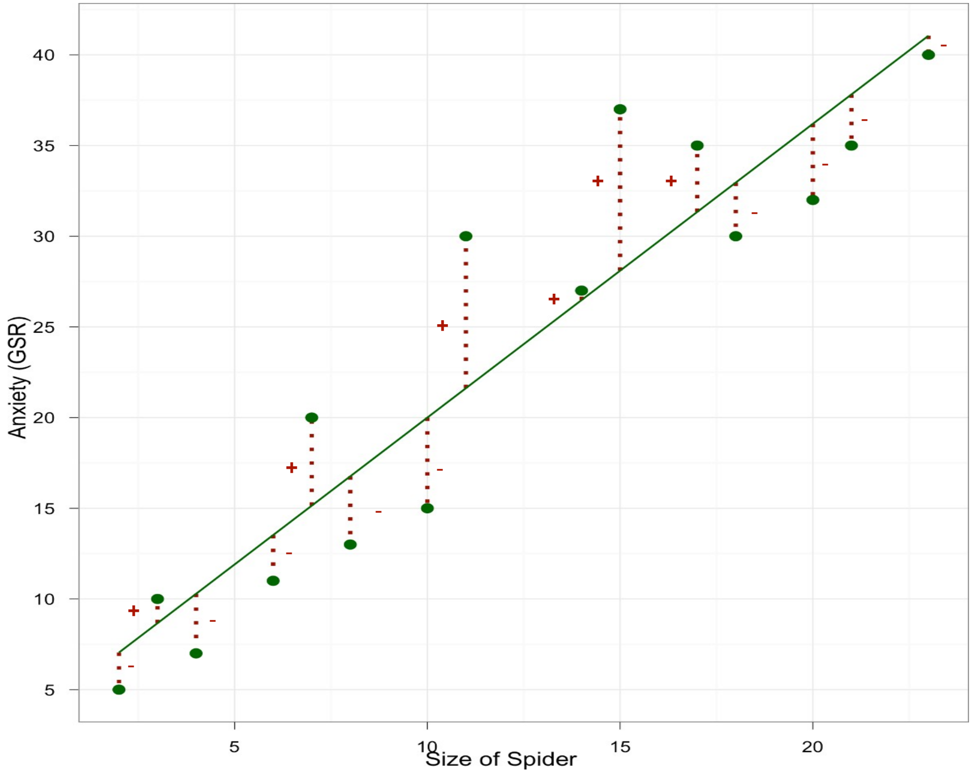


Same intercept, different gradlent



Same gradient, different intercepts

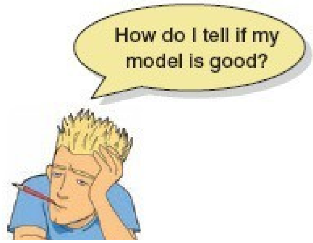
Ordinary Least Squares



This graph shows a scatterplot of some data with a line representing the general trend. The vertical lines (dotted) represent the differences (or residuals) between the line and the actual data

- Our coefficient **estimates** b_o and b_i are determined based on minimizing the difference between the observed data and the value predicted by the linear approximation
  - **Minimizing the sum of square residuals**
  - Other methods exist (e.g., loglikelihood, maximum likelihood)

How Good Is the Model?



- The regression line is only a model based on the data.
- **This model might not reflect reality.**
  - We need to test how well the model fits the observed data.

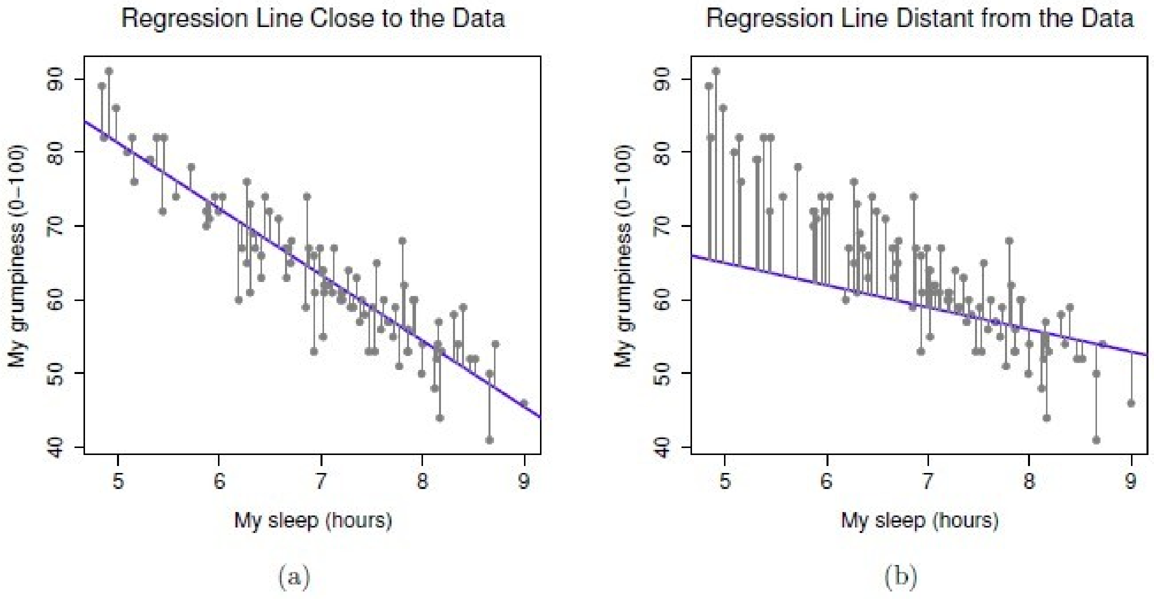


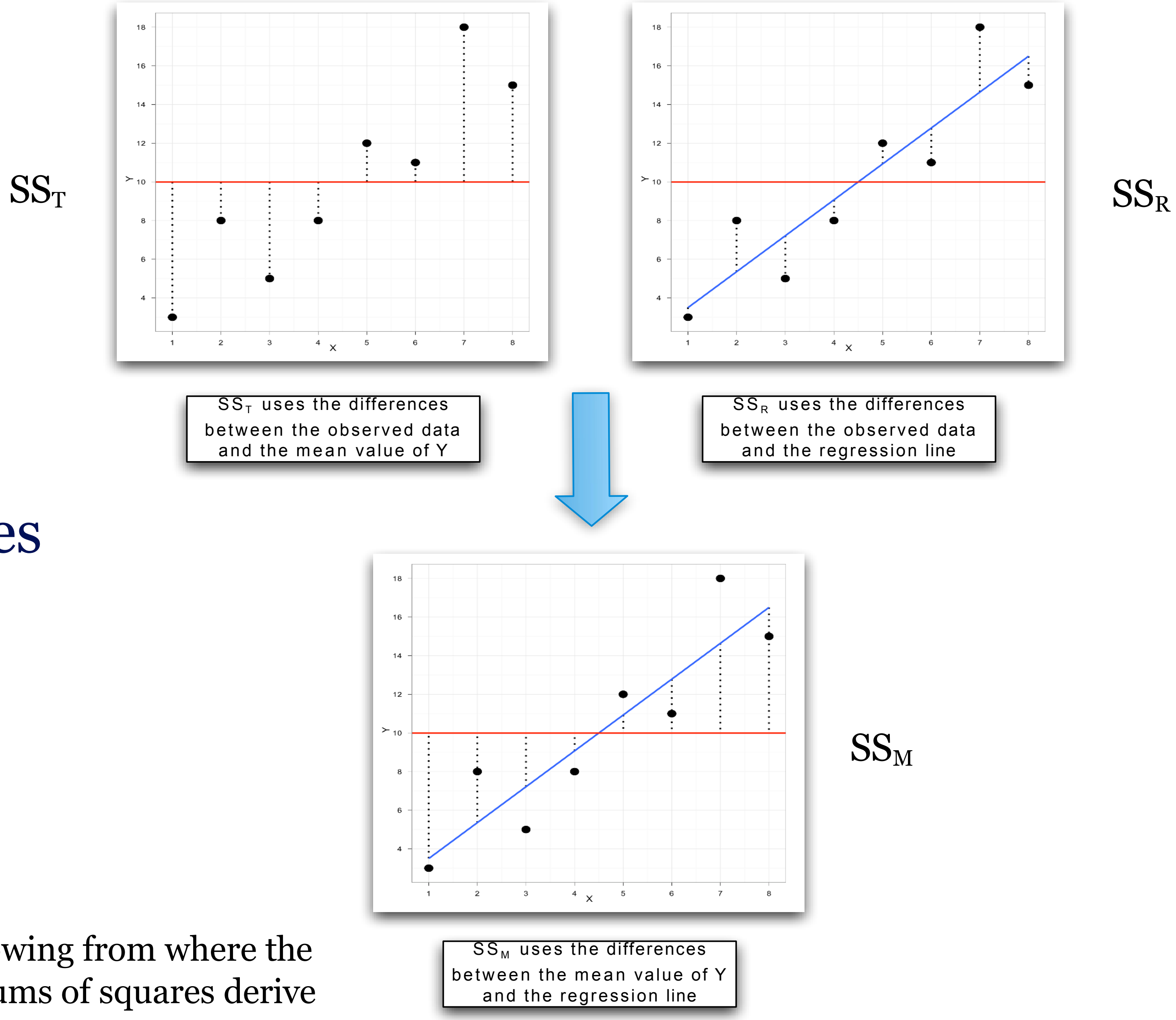
Figure 15.3: A depiction of the residuals associated with the best fitting regression line (panel a), and the residuals associated with a poor regression line (panel b). The residuals are much smaller for the good regression line. Again, this is no surprise given that the good line is the one that goes right through the middle of the data.

Sums of Squares

- SS_T : Total sum of squares

- SS_R : Residual sum of squares

- SS_M : Model sum of squares



SS_T

SS_R

SS_T uses the differences between the observed data and the mean value of Y

SS_R uses the differences between the observed data and the regression line

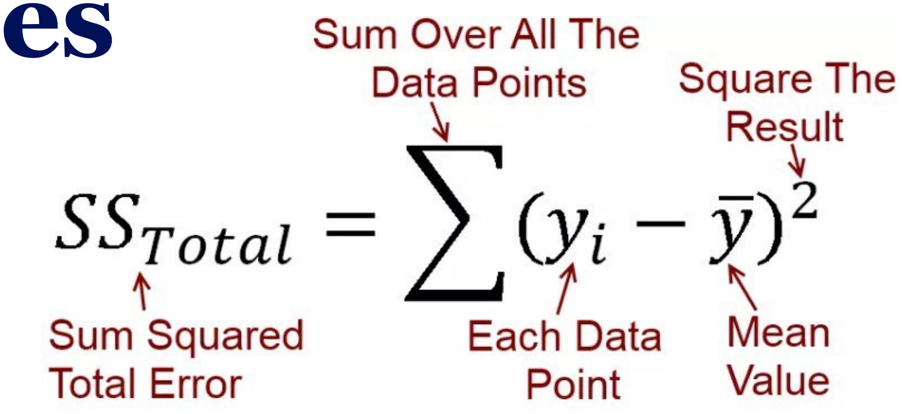
SS_M

SS_M uses the differences between the mean value of Y and the regression line

Diagram showing from where the regression sums of squares derive

Calculating Sums of Squares

- $\mathrm{SS}_{\mathrm{Total}}$
  - **Total** variability: variability between scores and the mean.
  - Inaccurracy of the worst possible model (the mean)
- $\mathrm{SS}_{\mathrm{Residual}}$
  - **Residual**/error variability: variability between the regression model and the actual data.
  - Inaccurracy of the regression model.
- $\mathrm{SS}_{\mathrm{Model}}$
  - **Model** variability: difference in variability between the model and the mean.
  - How much better does the (best) regression model perform than the (worst) mean model? Higher is better.



$$SS_{Total} = \sum (y_i - \bar{y})^2$$

Sum Over All The Data Points

Square The Result

Sum Squared Total Error

Each Data Point

Mean Value

$$\mathrm{SS}_{res} = \sum_i (Y_i - \hat{Y}_i)^2$$

Each regression model point

$$\mathrm{SS}_{mod} = \mathrm{SS}_{tot} - \mathrm{SS}_{res}$$

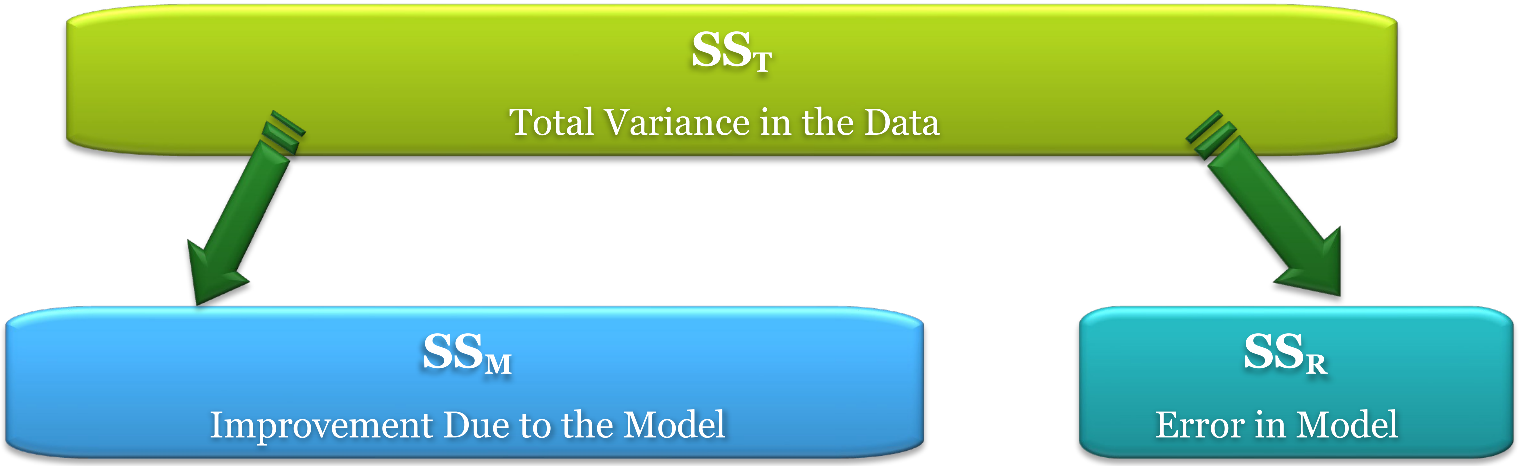
Testing the Model: *R*²

- Coefficient of determination *R*²
  - The proportion of variance in the dependent variable that is predictable from the independent variable(s)
  - The Pearson Correlation Coefficient Squared
  - If $R^2 = 1$, the model explains 100% of the variance in the data

$$\boxed{R^2 = \frac{SS_M}{SS_T}}$$

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

Testing the Model: ANOVA



- If the model results in better prediction than using the mean, then we expect SS_M to be much greater than SS_R

Testing the Model: ANOVA

- Mean squared error
  - Sums of squares are total values.
  - They can be expressed as averages.
  - These are called mean squares, MS.

$$\mathrm{MS}_{mod} = \frac{\mathrm{SS}_{mod}}{df_{mod}}$$

$$\mathrm{MS}_{res} = \frac{\mathrm{SS}_{res}}{df_{res}}$$

$$\boxed{F = \frac{\mathrm{MS_M}}{\mathrm{MS_R}}}$$

$$df_{mod} = K.$$

$$df_{res} = N - K - 1.$$

K = Number of predictor variables in model

Regression: An Example

- A record company boss was interested in predicting record sales from advertising.
- Data
  - 200 different album releases
- Outcome variable:
  - Sales (CDs and downloads) in the week after release
- Predictor variable:
  - The amount (in units of £1000) spent promoting the record before release.

"Yes sir, both your precious metal and heavy metal investments are doing well."

Regression: An Example

Testing the whole model (ANOVA)

- Null hypothesis: There is no relationship between the predictors and the outcome.

$$H_0 : Y_i = b_0 + \epsilon_i$$

$$H_1 : Y_i = \left(\sum_{k=1}^{K} b_k X_{ik} \right) + b_0 + \epsilon_i$$

Testing individual coefficients (t-test)

- Null hypothesis: The true population regression coefficient is 0.

$$\begin{aligned} H_0 : &\quad b = 0 \\ H_1 : &\quad b \neq 0 \end{aligned}$$

$$t = \frac{\hat{b}}{\mathrm{SE}(\hat{b})}$$

Regression in R

- We run a regression analysis using the *lm()* function – lm stands for ‘linear model’. This function takes the general form:

```
newmodel <- lm(outcome ~ predictors)
```

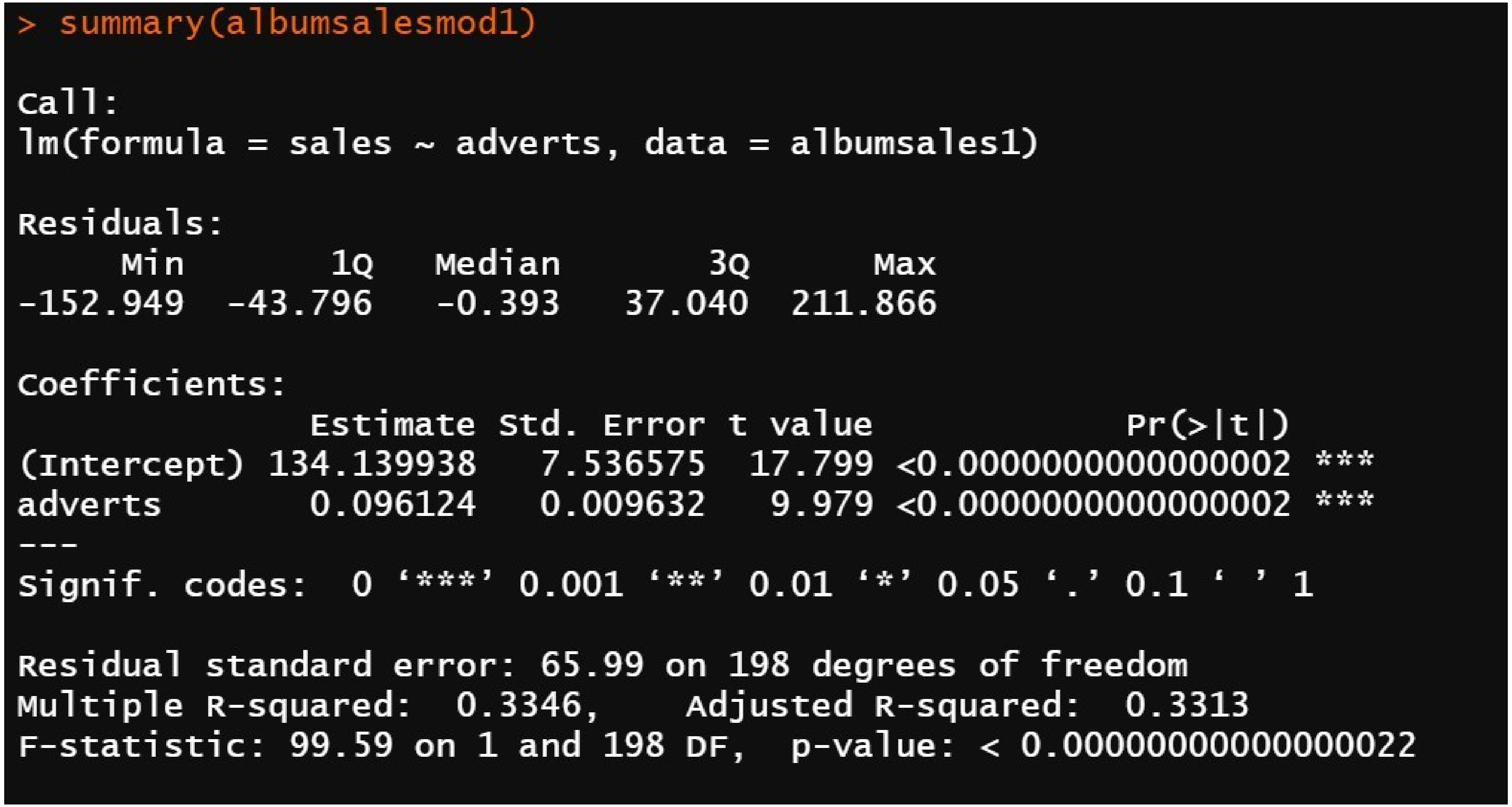
```
albumsalesmod1 <- lm(album1$sales ~ album1$adverts)
```

- or we can tell **R** what dataframe to use (using data = nameOfDataFrame), and then specify the variables without the *dataFrameName$* before them:

```
albumsalesmod1 <- lm(sales ~ adverts, data = album1)
```

Output of a Simple Regression

- We have created an object called *albumsalesmod1* that contains the results of our analysis. We can show the object by executing:



```
> summary(albumsalesmod1)

Call:
lm(formula = sales ~ adverts, data = albumsales1)

Residuals:
     Min        1Q    Median        3Q       Max 
-152.949   -43.796   -0.393    37.040   211.866 

Coefficients:
              Estimate Std. Error t value            Pr(>|t|)    
(Intercept) 134.139938   7.536575  17.799 <0.0000000000000002 ***
adverts       0.096124   0.009632   9.979 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 65.99 on 198 degrees of freedom
Multiple R-squared:  0.3346,    Adjusted R-squared:  0.3313 
F-statistic: 99.59 on 1 and 198 DF,  p-value: < 0.00000000000000022
```

Example Regression Write Up

The amounts of money spent on advertising not only explained a significant 33% of the variance in album sales (R^2 = .33, F(1, 198) = 99.59, p < .001), but it also positively predicted album sales, b = .096, 95% CI [0.08, 0.12], t(198) = 9.98, p < .001, .

Making Predictions with our Model

$$\begin{aligned}\text{Record Sales}_i &= b_0 + b_1\text{Advertising Budget}_i\\ &= 134.14 + \Big(0.09612\ \xi\text{Advertising Budget}_i\end{aligned}$$

$$\Big)$$

$$\begin{aligned}\text{Record Sales}_i &= 134.14 + \Big(0.09612\ \xi\text{Advertising Budget}_i\\ \Big)\\ &= 134.14 + \big(0.09612\ \xi\ 100\big)\\ &= 143.75\end{aligned}$$

Standardized vs unstandardized regression coefficients

- Unstandardized b
  - b coefficient is the amount by which our dependent variable changes if we change the independent variable by one unit while keeping other independent variables constant.
- Standardized Beta β
  - Measured in units of standard deviation. A beta value of 1.25 indicates that a change of one standard deviation in the independent variable results in a 1.25 standard deviations increase in the dependent variable.
  - Allows for comparing the contribution of multiple independent variables, as they are expressed on the same scale.

Standardizing in R

Standardizing using a function:

```
install.packages(“effectsize”)
library(“effectsize”)
effectsize(albumsalesmod1)
```

Standardize in our model:

```
albumsalesmod2<-lm(scale(sales)~scale(adverts),data=albumsales1)
summary(albumsalesmod2)
```

Run your own regression on exam anxiety data

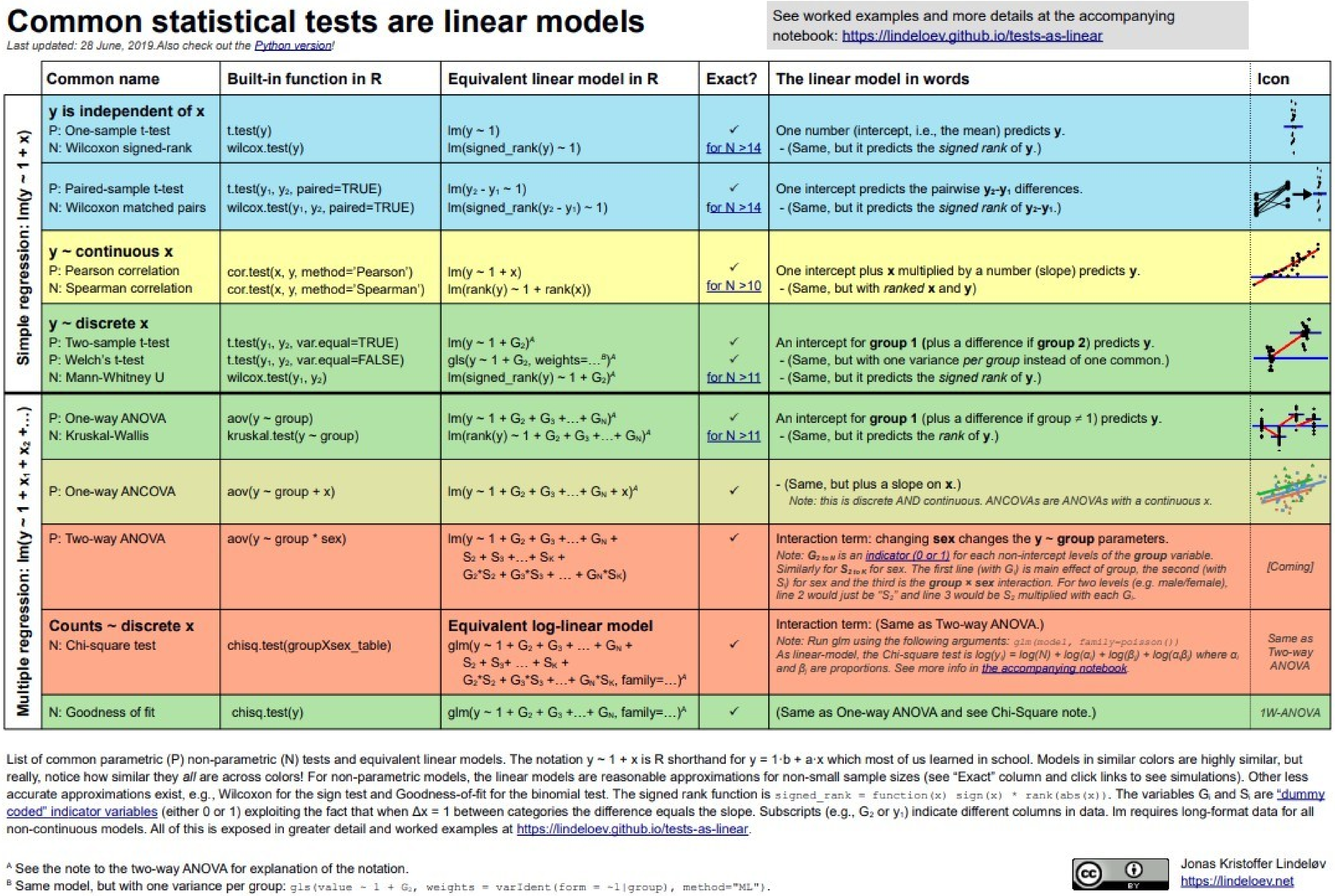
- Load the Exam Anxiety.dat file into R.
- Make a prediction about the relationship between exam anxiety and exam performance
- Run a linear regression using the lm() function
- Interpret the output
- Get the standardized beta coefficients
- Bonus: create a scatter plot with the regression line
- Write a summary report of the results

Summing Up

- Understand linear regression with one predictor
- Understand how we assess the fit of a regression model
  - Total sum of squares
  - Model sum of squares
  - Residual sum of squares
  - *F*
  - R^2
- Know how to do regression using **R**
- Interpret a regression model

Preparing for Module 6: More Regression

- Winter, B. (2013). Linear models and linear mixed effects models (page 1-21).



Common statistical tests are linear models

Last updated: 28 June, 2019.Also check out the Python version!

See worked examples and more details at the accompanying notebook: https://lindeloev.github.io/tests-as-linear

	Common name	Built-in function in R	Equivalent linear model in R	Exact?	The linear model in words	Icon

Simple regression: lm(y ~ 1 + x)	**y is independent of x**
	P: One-sample t-test
	N: Wilcoxon signed-rank
		t.test(y)
		wilcox.test(y)
			lm(y ~ 1)
			lm(signed_rank(y) ~ 1)
				✓
				for N >14
					One number (intercept, i.e., the mean) predicts **y**.
					- (Same, but it predicts the *signed rank* of **y**.)	

	P: Paired-sample t-test
	N: Wilcoxon matched pairs
		t.test(y_1, y_2, paired=TRUE)
		wilcox.test(y_1, y_2, paired=TRUE)
			lm(y_2 - y_1 ~ 1)
			lm(signed_rank(y_2 - y_1) ~ 1)
				✓
				for N >14
					One intercept predicts the pairwise **y_2-y_1** differences.
					- (Same, but it predicts the *signed rank* of **y_2-y_1**.)	

	y ~ continuous x
	P: Pearson correlation
	N: Spearman correlation
		cor.test(x, y, method='Pearson')
		cor.test(x, y, method='Spearman')
			lm(y ~ 1 + x)
			lm(rank(y) ~ 1 + rank(x))
				✓
				for N >10
					One intercept plus **x** multiplied by a number (slope) predicts **y**.
					- (Same, but with *ranked* **x** and **y**)	

	y ~ discrete x
	P: Two-sample t-test
	P: Welch's t-test
	N: Mann-Whitney U
		t.test(y_1, y_2, var.equal=TRUE)
		t.test(y_1, y_2, var.equal=FALSE)
		wilcox.test(y_1, y_2)
			lm(y ~ 1 + G_2)[A]
			gls(y ~ 1 + G_2, weights=...[B])[A]
			lm(signed_rank(y) ~ 1 + G_2)[A]
				✓
				✓
				for N >11
					An intercept for **group 1** (plus a difference if **group 2**) predicts **y**.
					- (Same, but with one variance *per group* instead of one common.)
					- (Same, but it predicts the *signed rank* of **y**.)	

Multiple regression: lm(y ~ 1 + x_1 + x_2 +...)	P: One-way ANOVA
	N: Kruskal-Wallis
		aov(y ~ group)
		kruskal.test(y ~ group)
			lm(y ~ 1 + G_2 + G_3 +...+ G_N)[A]
			lm(rank(y) ~ 1 + G_2 + G_3 +...+ G_N)[A]
				✓
				for N >11
					An intercept for **group 1** (plus a difference if group ≠ 1) predicts **y**.
					- (Same, but it predicts the *rank* of **y**.)	

	P: One-way ANCOVA	aov(y ~ group + x)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + x)[A]	✓	- (Same, but plus a slope on **x**.) *Note: this is discrete AND continuous. ANCOVAs are ANOVAs with a continuous x.*	
	P: Two-way ANOVA	aov(y ~ group * sex)	lm(y ~ 1 + G_2 + G_3 +...+ G_N + S_2 + S_3 +...+ S_K + G_2*S_2 + G_3*S_3 + ... + G_N*S_K)	✓	Interaction term: changing **sex** changes the **y ~ group** parameters. *Note: **$G_{2 \text{ to } N}$** is an indicator (0 or 1) for each non-intercept levels of the **group** variable. Similarly for **$S_{2 \text{ to } K}$** for sex. The first line (with G_i) is main effect of group, the second (with S_i) for sex and the third is the **group × sex** interaction. For two levels (e.g. male/female), line 2 would just be "S_2" and line 3 would be S_2 multiplied with each G_i.*	*[Coming]*
	Counts ~ discrete x N: Chi-square test	chisq.test(groupXsex_table)	**Equivalent log-linear model** glm(y ~ 1 + G_2 + G_3 + ... + G_N + S_2 + S_3+ ... + S_K + G_2*S_2 + G_3*S_3 +...+ G_N*S_K, family=...)[A]	✓	Interaction term: (Same as Two-way ANOVA.) *Note: Run glm using the following arguments:* `glm(model, family=poisson())` *As linear-model, the Chi-square test is $\log(y_i) = \log(N) + \log(\alpha_i) + \log(\beta_j) + \log(\alpha_i\beta_j)$ where α_i and β_j are proportions. See more info in the accompanying notebook.*	*Same as Two-way ANOVA*
	N: Goodness of fit	chisq.test(y)	glm(y ~ 1 + G_2 + G_3 +...+ G_N, family=...)[A]	✓	(Same as One-way ANOVA and see Chi-Square note.)	*1W-ANOVA*

List of common parametric (P) non-parametric (N) tests and equivalent linear models. The notation y ~ 1 + x is R shorthand for y = 1·b + a·x which most of us learned in school. Models in similar colors are highly similar, but really, notice how similar they *all* are across colors! For non-parametric models, the linear models are reasonable approximations for non-small sample sizes (see "Exact" column and click links to see simulations). Other less accurate approximations exist, e.g., Wilcoxon for the sign test and Goodness-of-fit for the binomial test. The signed rank function is `signed_rank = function(x) sign(x) * rank(abs(x))`. The variables G_i and S_i are "dummy coded" indicator variables (either 0 or 1) exploiting the fact that when Δx = 1 between categories the difference equals the slope. Subscripts (e.g., G_2 or y_1) indicate different columns in data. lm requires long-format data for all non-continuous models. All of this is exposed in greater detail and worked examples at https://lindeloev.github.io/tests-as-linear.

[A] See the note to the two-way ANOVA for explanation of the notation.
[B] Same model, but with one variance per group: `gls(value ~ 1 + G₂, weights = varIdent(form = ~1|group), method="ML")`.

Jonas Kristoffer Lindeløv
https://lindeloev.net

Thanks!

Material from Field et al Discovering Statistics with R.