Statistics for CSAI II

6 — Regression and Assumptions

Travis J. Wiltshire, Ph.D. |

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

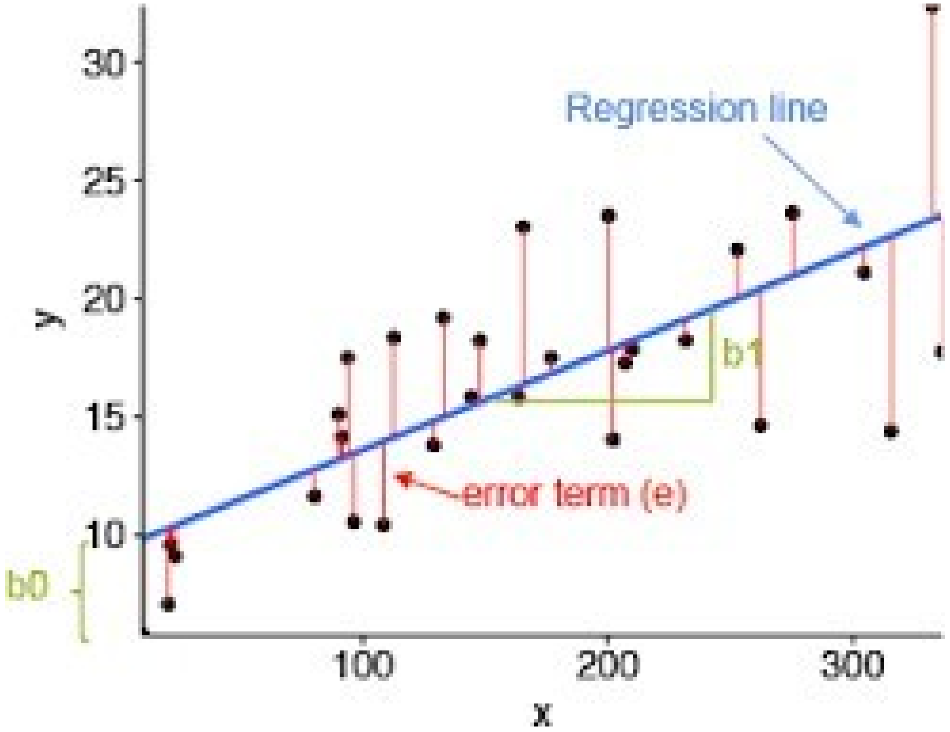
Outline

1. Understand linear regression with one predictor
   - Regression with continuous predictor
   - Regression with categorical predictor
2. Meaningful intercepts
3. Assumptions of Regression
4. Global checking of assumptions
   - Gvlma
   - check_model() from performance package

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_0
  - Intercept (value of Y when $X = 0$)
  - Point at which the regression line crosses the Y-axis (ordinate)



Output of a Simple Regression

```
> 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
```

Making Predictions with our Model

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

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

Regression in Matrix Algebra Form

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{\beta} + \underbrace{\begin{bmatrix} e_1 \\ e_2 \\ e_3 \end{bmatrix}}_{e} = X\beta + e$$

Note that the matrix-vector multiplication $X\beta$ results in

$$X\beta = \begin{bmatrix} \beta_0 + \beta_1 x_1 \\ \beta_0 + \beta_1 x_2 \\ \beta_0 + \beta_1 x_3 \end{bmatrix},$$

which is essentially just a compact way of writing the regression model.

- More details here
- And here

Regression with a Categorical Predictor

- Categorical variables: variables made up of **categories** of objects/entities
- But regression analysis requires numeric variables…
- Solution: Dummy coding
  - Adding a numeric variable that assigns value relative to reference category
  - Alternative: Effect coding (each category is compared to the overall mean; using -1 and 1)

Subject	**Sex**	**Voice.Pitch**
1	female	233 Hz
2	female	204 Hz
3	female	242 Hz
4	male	130 Hz
5	male	112 Hz
6	male	142 Hz

$$\text{pitch} \sim \text{sex} + \varepsilon$$

Dummy coding of categorical variables in R

```
> pitch = c(233,204,242,130,112,142)
> sex = as.factor(c(rep("female",3),rep("male",3)))
> my.df = data.frame(sex,pitch)
> my.df
     sex pitch
1 female   233
2 female   204
3 female   242
4   male   130
5   male   112
6   male   142
> # Check how R is treating the dummy coding
> contrasts(my.df$sex)
       male
female    0
male      1
> # Change reference category to male
> my.df$sex <- relevel(my.df$sex, ref = "male")
> contrasts(my.df$sex)
       female
male        0
female      1
```

Subject	Sex	Voice.Pitch
1	female	233 Hz
2	female	204 Hz
3	female	242 Hz
4	male	130 Hz
5	male	112 Hz
6	male	142 Hz

$$pitch \sim sex + \varepsilon$$

Regression with Categorical Predictor in R

```
> #model pitch by sex
> xmdl = lm(pitch ~ sex, my.df)
> summary(xmdl)

Call:
lm(formula = pitch ~ sex, data = my.df)

Residuals:
      1       2       3       4       5       6 
  6.667 -22.333  15.667   2.000 -16.000  14.000 

Coefficients:
            Estimate Std. Error t value  Pr(>|t|)    
(Intercept)   226.33      10.18  22.224 0.0000243 ***
sexmale       -98.33      14.40  -6.827   0.00241 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 17.64 on 4 degrees of freedom
Multiple R-squared:  0.921,	Adjusted R-squared:  0.9012 
F-statistic: 46.61 on 1 and 4 DF,  p-value: 0.002407
```

```
> # Check how R is treating the dummy coding
> contrasts(my.df$sex)
       male
female    0
male      1
> mn_female <- mean(my.df[my.df$sex=="female",]$pitch)
> mn_female
[1] 226.3333
> mn_male <- mean(my.df[my.df$sex=="male",]$pitch)
> mn_male
[1] 128
> mn_male - mn_female
[1] -98.33333
```

Run a regression on exam anxiety data

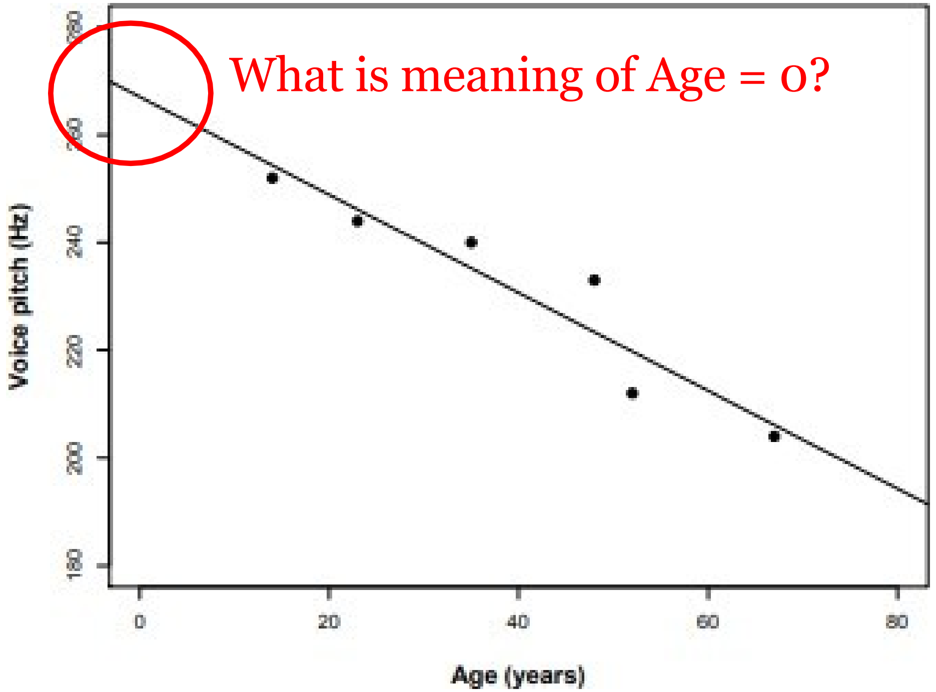
- Load the Exam Anxiety.dat file into R (or csv).
- Examine the data and make a prediction about the relationship between exam anxiety and gender
- Run a linear regression using the lm() function
- Interpret the output using summary() function

Meaningless intercepts with continuous variables

```
> age = c(14,23,35,48,52,67)
> pitch = c(252,244,240,233,212,204)
> my.df = data.frame(age,pitch)
> xmdl = lm(pitch ~ age, my.df)
> summary(xmdl)
```

```
Call:
lm(formula = pitch ~ age, data = my.df)

Residuals:
     1      2      3      4      5      6 
-2.338 -2.149  4.769  9.597 -7.763 -2.115 

Coefficients:
            Estimate Std. Error t value   Pr(>|t|)
(Intercept) 267.0765     6.8522   38.98 0.00000259 ***
age          -0.9099     0.1569   -5.80    0.00439 **
```

Subject	Age	Voice.Pitch
1	14	252 Hz
2	23	244 Hz
3	35	240 Hz
4	48	233 Hz
5	52	212 Hz
6	67	204 Hz



Using centering to make a more meaningful intercept

```
my.df$age.c = my.df$age - mean(my.df$age)
xmdl = lm(pitch ~ age.c, my.df)
summary(xmdl)
```

```
Coefficients:
             Estimate Std. Error t value    Pr(>|t|)
(Intercept) 230.8333     2.8113   82.11 0.000000132 ***
age.c        -0.9099     0.1569   -5.80     0.00439 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.886 on 4 degrees of freedom
Multiple R-squared:  0.8937,    Adjusted R-squared:  0.8672
F-statistic: 33.64 on 1 and 4 DF,  p-value: 0.004395
```

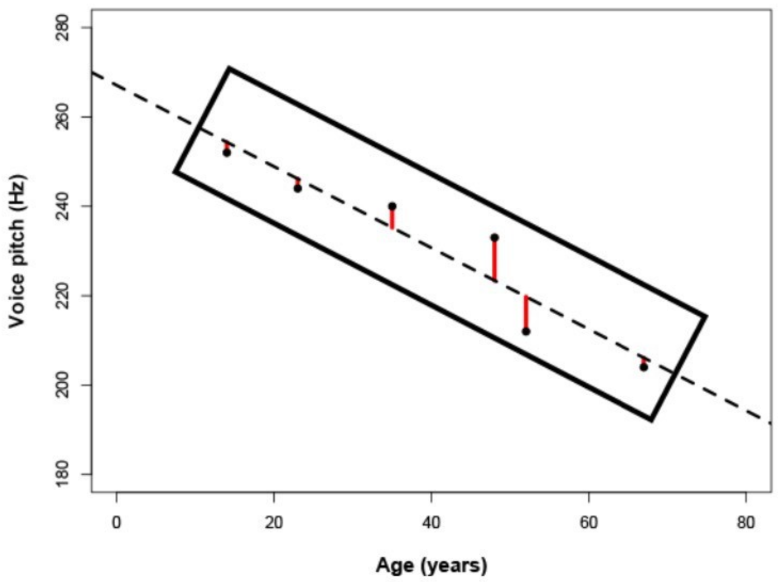
Now our intercept tells us the mean voice pitch.

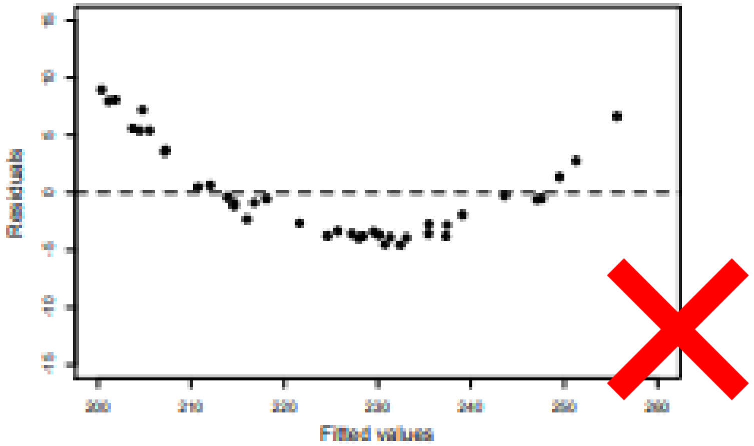
```
> mean(pitch)
[1] 230.8333
```

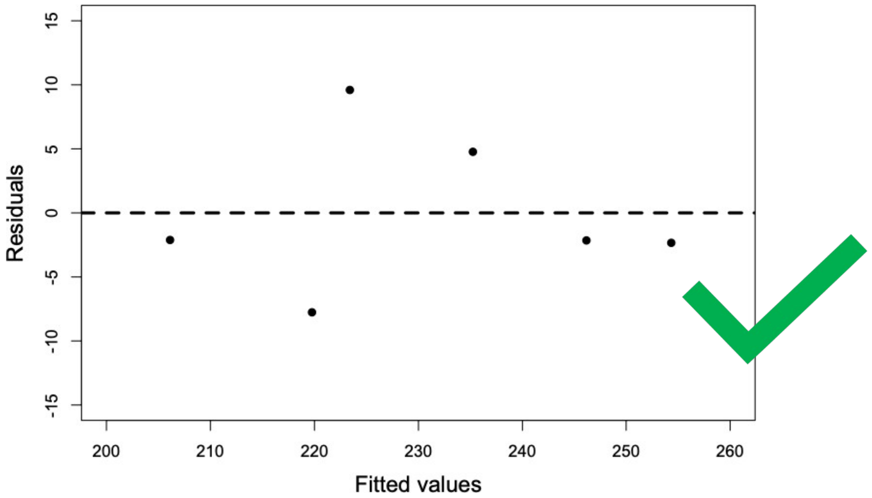
Assumptions of Regression

Assumption 1: Linearity

- The outcome (dependent variable) is the result of a linear combination of the predictors (independent variables)
  - Check by looking at residuals plot
  - If residuals plot shows a curve or some other pattern, the linearity assumption is violated



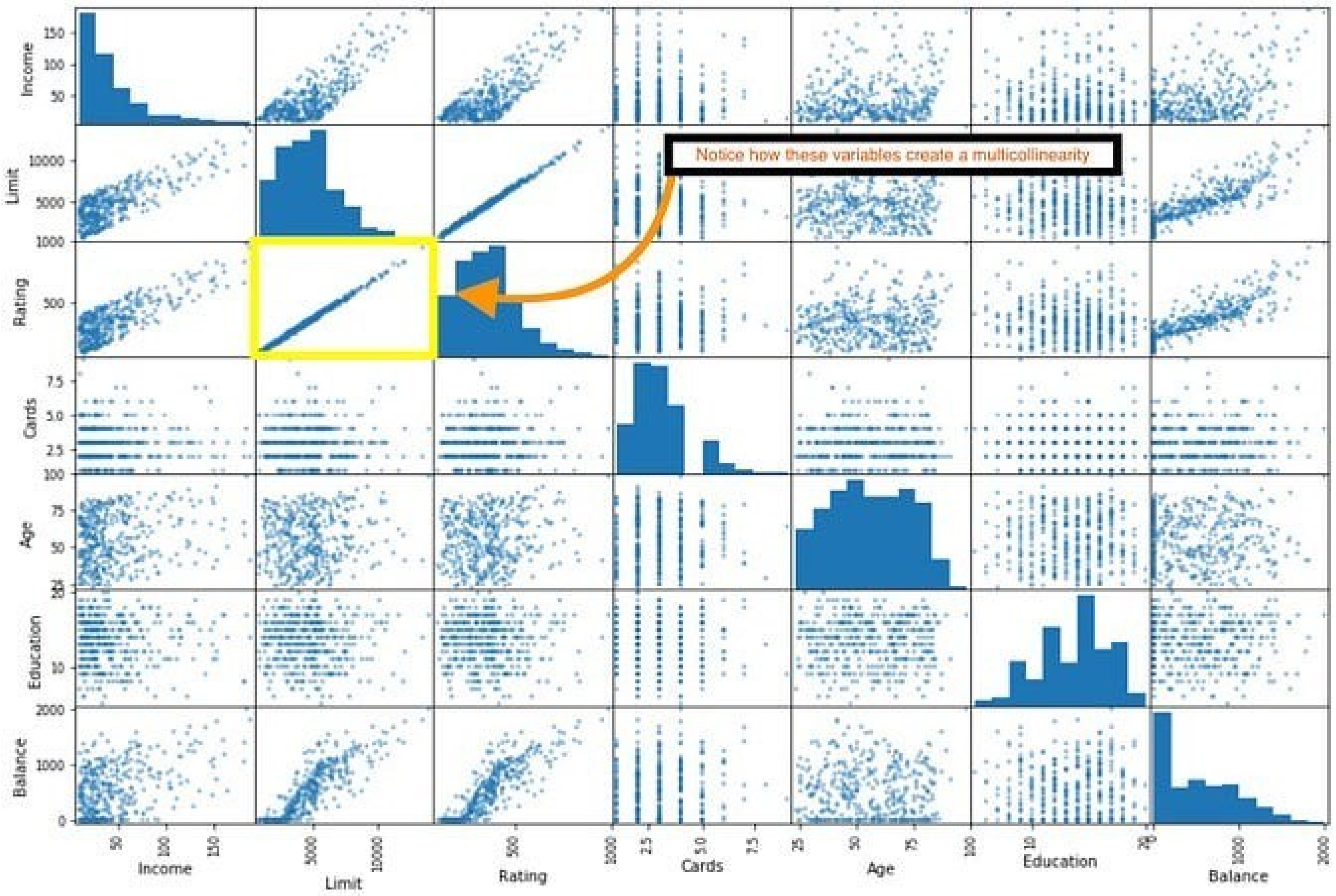





```
plot(fitted(xmdl), residuals(xmdl), pch=20)
abline(a=0,b=0, lty=2)
```

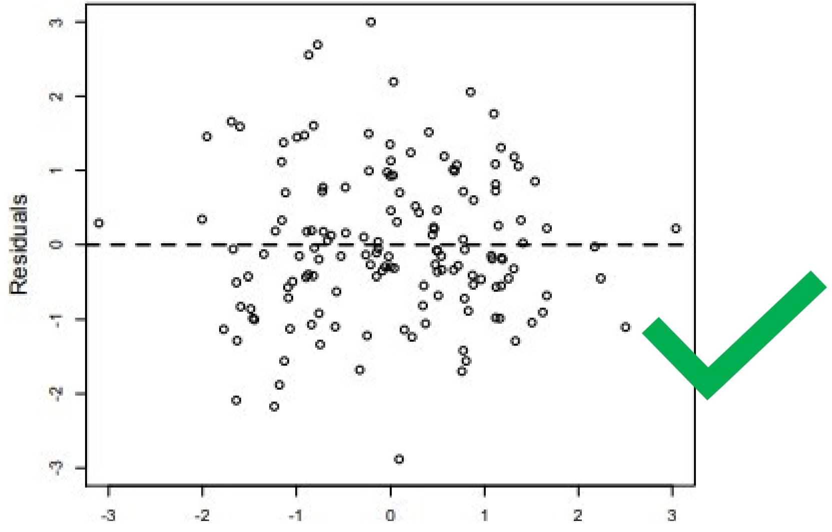
Assumption 2: Absence of collinearity

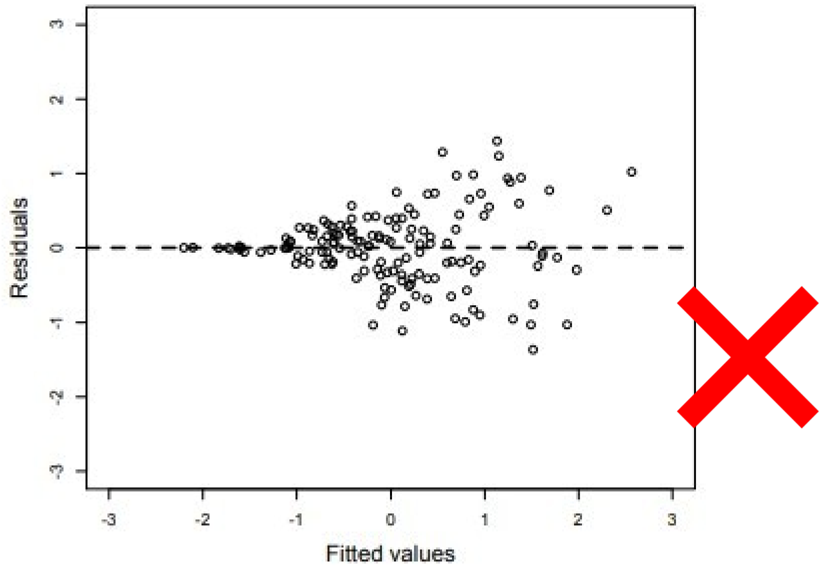
- Avoid correlated predictors to keep interpretation of the model stable



Assumption 3: Homoscedasticity

- Homoscedasticity = absence of heteroskedasticity
- The variance of the data should be approximately equal across the range of predicted values
- Check the residual plots
- Possible remedy: consider log-transforming your response data

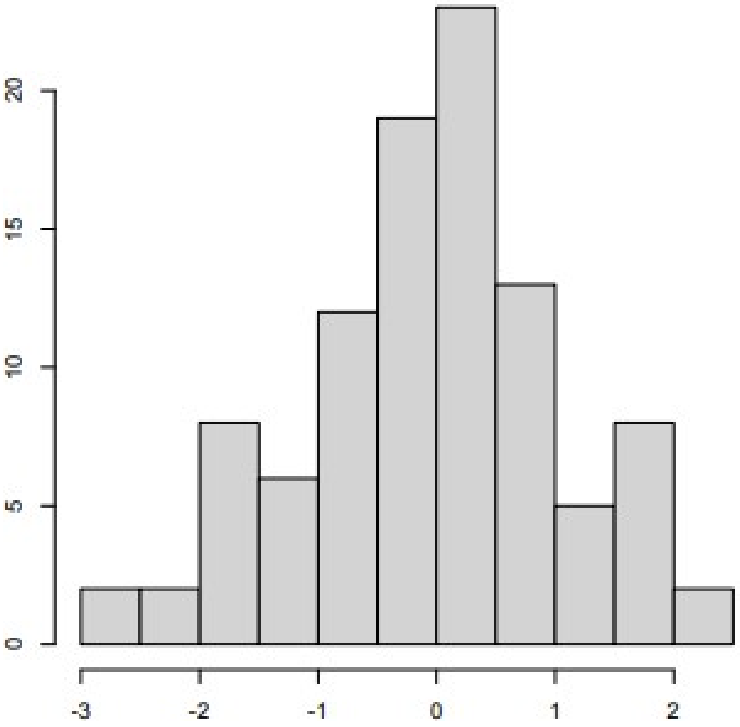


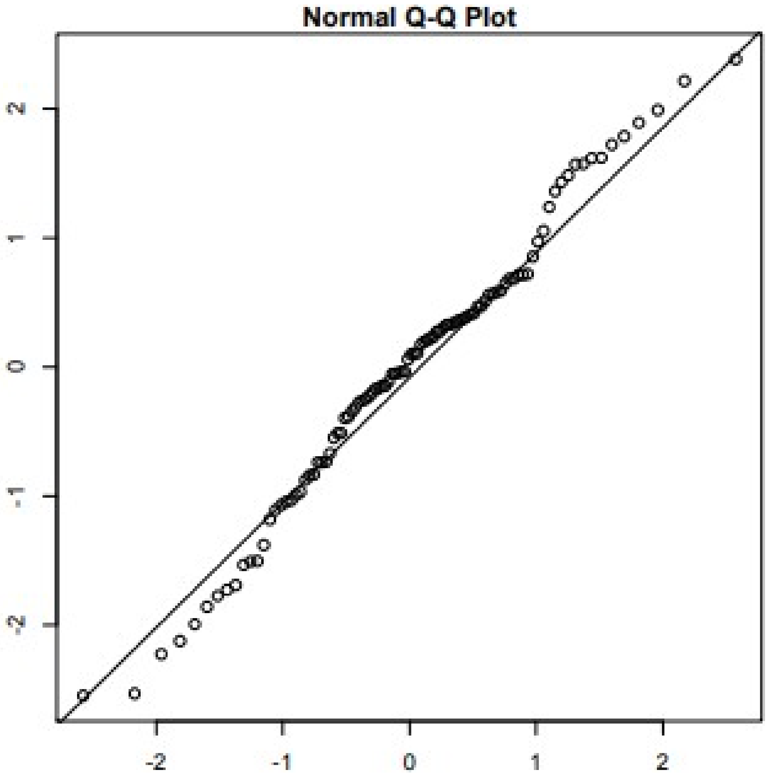


Assumption 4: Normality of Residuals

Examine the histogram or q-q plot of the residuals.

```
> hist(residuals(xmdl))
> qqnorm(residuals(xmdl))
```

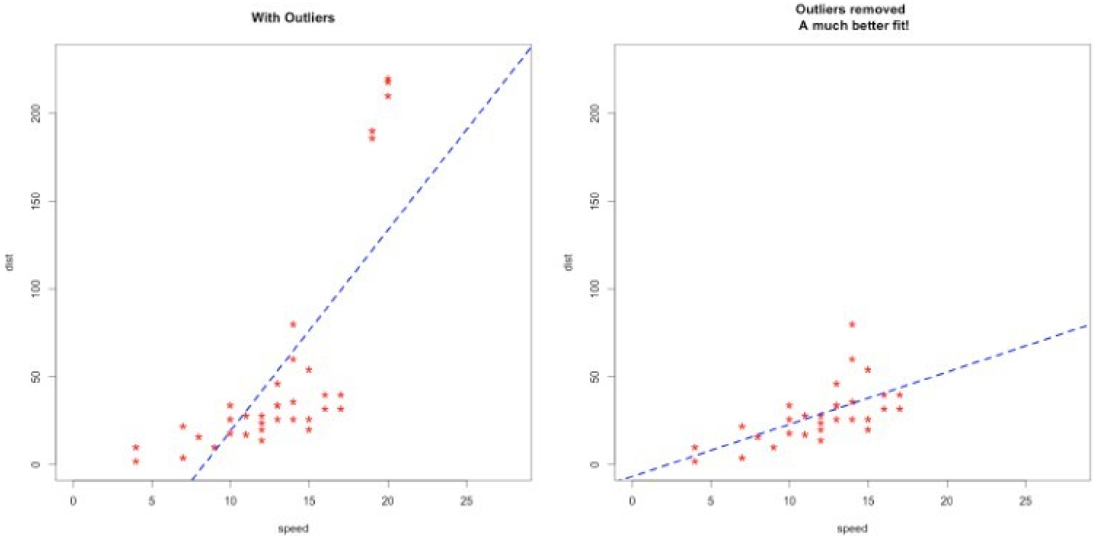




Assumption 5: absence of influential points (outliers)

- Extreme data points (outliers) can drastically change the the results
- Check with “leave-one-out” diagnostics for each data point
  - Adjustments to the coefficients if that estimate is left out
- Warning signs:
  - The adjustments change the sign of the coefficients
  - Adjustments by at least half of the absolute value of the coefficients could be concerning

```
> dfbeta(xmdl)

  (Intercept)          age
1  -3.3645662  0.06437573
2  -1.6119656  0.02736278
3   1.5481303 -0.01456709
4  -0.0259835  0.05092767
5   0.8707699 -0.06479736
6   1.8551808 -0.06622744
```



Assumption 6: Independence * IMPORTANT*

- Each observation must be independent (from a different subject)
- Dangers of violating independence:
  - Increased chance of spurious results
  - Meaningless p-values
- Part of the experimental design
- Use mixed models to resolve non-independencies

Study 1

Subject	Sex	Voice.Pitch
1	female	233 Hz
2	female	204 Hz
3	female	242 Hz
4	male	130 Hz
5	male	112 Hz
6	male	142 Hz

Study 2

Subject	Age	Voice.Pitch
1	14	252 Hz
2	23	244 Hz
3	35	240 Hz
4	48	233 Hz
5	52	212 Hz
6	67	204 Hz

Check the assumptions of exam anxiety data

- For the model you made examining the relationship between exam anxiety and gender
- Check the linearity assumption
- Check the homoscedasticity assumption
- Check the normality of the residuals
- Check for influential data points

gvlma() – Global check of the assumptions

- Global Stat ⇐ Linearity
- Skewness ⇐ Normality
- Kurtosis ⇐ Influential points
- Link Function ⇐ is your dependent variable truly continuous, or categorical?
- Heteroscedasticity ⇐Homoscedasticity

• Don't blindly trust the output: combine both methods !

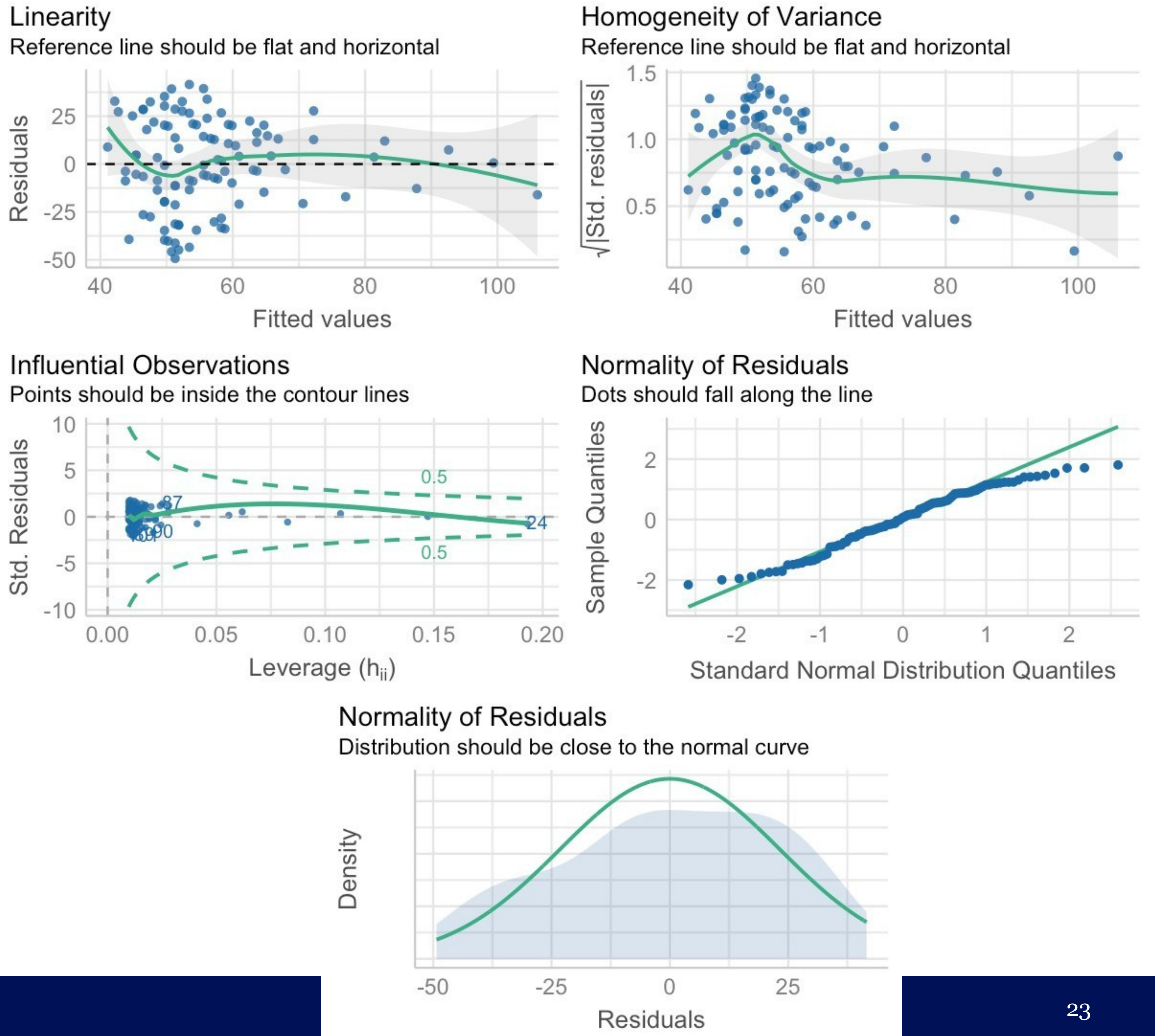
```
> # Global Check of the assumptions
> require(gvlma)
> gvlma(xmdl)

Call:
lm(formula = pitch ~ age.c, data = my.df)

Coefficients:
(Intercept)        age.c
   230.8333      -0.9099


ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
Level of Significance =  0.05

Call:
 gvlma(x = xmdl)

                    Value p-value                Decision
Global Stat        1.9167  0.7511 Assumptions acceptable.
Skewness           0.2132  0.6443 Assumptions acceptable.
Kurtosis           0.1942  0.6595 Assumptions acceptable.
Link Function      1.1268  0.2885 Assumptions acceptable.
Heteroscedasticity 0.3825  0.5363 Assumptions acceptable.
```

Original paper here.

Check the assumptions of exam anxiety data with gvlma() and check_model()

- Try gvlma() on the exam anxiety model
  - Don't forget to install and load the gvlma package
- Use check_model() from performance package on model
- Compare it with the conclusions you made from checking the graphs

check_model() from package performance



Running another model and checking the assumptions

- Load and inspect the driving.csv dataset
- Run two models (make predictions first)

    Age as predictor of errors at time2

    Gender as a predictor of errors at time2
- Interpret the output from the summary() function
- For age, generate a meaningful intercept and rerun the model with this
- Check the assumptions of both models
- Write a summary of the results for one of the models

Summing Up

- Understand linear regression with one predictor
  - Regression with continuous predictor
  - Regression with categorical predictor
- Meaningful intercepts
- Assumptions of Regression
- Global checking of assumptions
  - Gvlma
  - check_model() from performance package

Preparing for Module 7: Multiple Regression and Assumptions

- Required Reading: Field, Miles & Field – CH 7 (pp. 261-301)

Thanks! See you next week!
Questions?

TILBURG UNIVERSITY

Material from Bodo Winter tutorial