Regression analysis II & III: Multiple regression, confounding, interactions, categorical variables, assumptions, leverage

Regression analysis IV: To explain, to predict or to describe

Manuela Zucknick

Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
manuela.zucknick@medisin.uio.no

MF9130E – Introductory Course in Statistics, 10-05-2023

Outline

Aalen chapter 11.4-11.6, Kirkwood and Sterne chapters 11 and 12

- **Multiple linear regression** (briefly: multiple regression)

- More on linear regression models: **confounding**, **interactions**, **categorical covariates** with more than 2 levels, regression **assumptions**, **leverage** effect.

- To explain, to predict or to describe?: How the purpose of the analysis decides what is important.

Outline for today

08.30-10.00: Regression analysis II: multiple regression, confounding, interaction effects.

10.15-11.15: R exercise for regression II.

11.15-11.45: Discussion of the R exercise for regression II in class.

- LUNCH

12.45-13.45: Regression analysis III: Multiple regression (continued), categorical variables, assumptions, leverage effect.

14.00-14.45: R exercise for regression III.

14.45-15.15: Discussion of the R exercises for regression III in class.

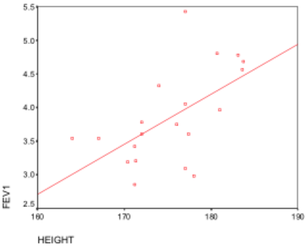
15.15-16.00: To explain, to predict or to describe?: How the purpose of the analysis decides what is important.

Yesterday: Simple linear regression

A **simple linear regression** describes the relationship between 1 independent variable (covariate, or predictor) and the dependent variable (response variable, or outcome) via a line.

Toy example: association between FEV1 and height.
Estimated regression line:

$$\text{FEV1} \approx -9.19 + 0.07 \cdot \text{height} \tag{1}$$



Relationship between simple linear regression and t-test

- There is a connection between the two approaches:
- Student's t-test (with equal variances) for the difference in the population mean between two independent groups is **equivalent** to a simple linear regression with the grouping as predictor variable.

Let us see this in a toy example:

Table 9.4 24 hour total energy expenditure (MJ/day) in groups of lean and obese women (Prentice *et al.*, 1986)

	Lean (n = 13)	Obese (n = 9)
	6.13	8.79
	7.05	9.19
	7.48	9.21
	7.48	9.68
	7.53	9.69
	7.58	9.97
	7.90	11.51
	8.08	11.85
	8.09	12.79
	8.11	
	8.40	
	10.15	
	10.88	
Mean	8.066	10.298
SD	1.238	1.398

R output for the t-test

R output for the Student's t-test (with equal variances) for the difference in `energy` between the lean and obese:

```
> t.test(energy ~ group, data=energy, var.equal=TRUE)

	Two Sample t-test

data:  energy by group
t = -3.9456, df = 20, p-value = 0.000799
alternative hypothesis: true difference in means between group Lean and group Obese is not equal to 0
95 percent confidence interval:
 -3.411451 -1.051796
sample estimates:
 mean in group Lean mean in group Obese
           8.066154           10.297778
```

R output for the simple linear regression

```
> fit <- lm(energy ~ group, data=energy)
> summary(fit)

Call:
lm(formula = energy ~ group, data = energy)

Residuals:
    Min      1Q  Median      3Q     Max
-1.9362 -0.6153 -0.4070  0.2614  2.8138

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   8.0662     0.3618  22.297 1.34e-15 ***
groupObese    2.2316     0.5656   3.946 0.000799 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.304 on 20 degrees of freedom
Multiple R-squared:  0.4377,    Adjusted R-squared:  0.4096
F-statistic: 15.57 on 1 and 20 DF,  p-value: 0.000799
```

Multiple regression

- Is an extension of the simple linear regression with one independent variable (predictor / covariate),
- Still a continuous response (dependent) variable, but several explanatory (independent) variables (multiple predictors / covariates),
- The independent variables can be continuous, dichotomous or have more than two categories,
- The **multiple linear regression model** is defined as

$$Y = b_0 + b_1 x_1 + \cdots + b_p x_p.$$

Regression coefficients

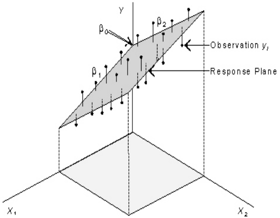
$$Y = b_0 + b_1 x_1 + \cdots + b_n x_n.$$

- $b_1, \ldots, b_n$ are called regression coefficients,
- b_i can be interpreted as the effect of one unit increase of the variable x_i when the other variables remain unchanged,
- also called **adjusted effect**,
- Not necessarily a causal effect.

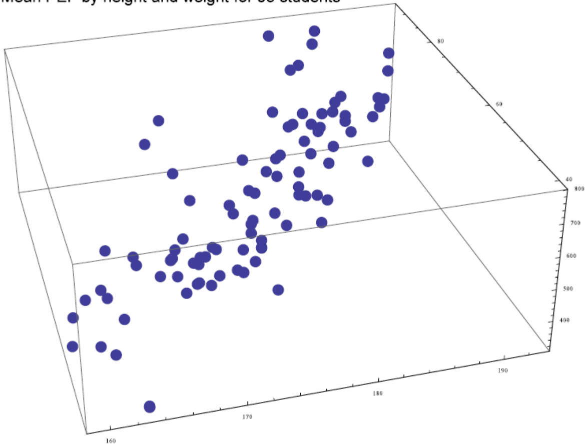
Interpretation

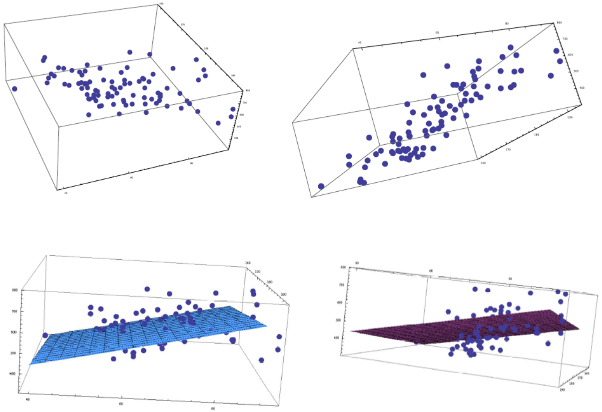
- Geometrically this corresponds to viewing data as points in a high-dimensional space.
- Beyond three dimensions we cannot picture such a space, but mathematically there is no difficulty with high-dimensional spaces.

Regression with two independent variables:



Mean PEF by height and weight for 95 students





Multiple regression via a toy example

Example: data on **systolic blood pressure**

Description	Name
Id	Id
Systolic blood pressure	SBP
Quetelet index (BMI)	QUET
Age	AGE
Smoking status	SMK

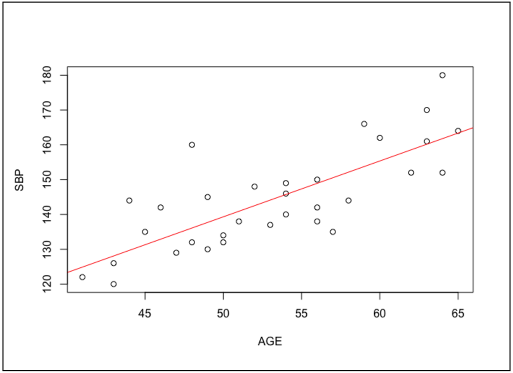
Simple linear regression: SBP vs AGE

```
> fit <- lm(SBP ~ AGE, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ AGE, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.548  -6.990  -2.481   5.765  23.892 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  59.0916    12.8163   4.611 6.98e-05 ***
AGE           1.6045     0.2387   6.721 1.89e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.245 on 30 degrees of freedom
Multiple R-squared:  0.6009,	Adjusted R-squared:  0.5876 
F-statistic: 45.18 on 1 and 30 DF,  p-value: 1.894e-07
```

- Note that $\hat{b}_0 = 59.09$ and $\hat{b}_1 = 1.61$,
- Confidence interval for b_1 $(1.12, 2.09)$ (calculate in R with `confint()`)
- $H_0 : b_1 = 0$ is rejected, as $p < 0.001$.
- SBP increases 1.6 units for **each year**.

Simple linear regression: SBP vs Age

```
> plot(SBP ~ AGE, data=bloodpressure)
> abline(reg=fit, col="red")
```



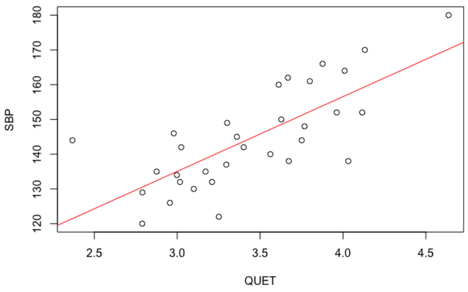
Simple linear regression: SBP vs QUET

```
> fit <- lm(SBP ~ QUET, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max
-19.231  -7.145  -1.604   7.798  22.531

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   70.576     12.322   5.728 2.99e-06 ***
QUET          21.492      3.545   6.062 1.17e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.812 on 30 degrees of freedom
Multiple R-squared:  0.5506,    Adjusted R-squared:  0.5356
F-statistic: 36.75 on 1 and 30 DF,  p-value: 1.172e-06
```

- Note that $\hat{b}_0 = 70.58$ and $\hat{b}_1 = 21.49$,
- Confidence interval for b_1 $(14.25, 28.73)$ (calculate in R with `confint()`)
- $H_0 : b_1 = 0$ is rejected, as $p < 0.001$.
- SBP increases 21.49 units for **each unit of QUET**.

Simple linear regression: SBP vs QUET

```
> plot(SBP ~ QUET, data=bloodpressure)
> abline(reg=fit, col="red")
```



Multiple regression: Combining AGE and QUET

```
> fit <- lm(SBP ~ QUET + AGE, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET + AGE, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.667  -6.793  -2.732   5.318  19.600 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  55.3234    12.5347   4.414 0.000129 ***
QUET          9.7507     5.4025   1.805 0.081489 .  
AGE           1.0452     0.3861   2.707 0.011253 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.916 on 29 degrees of freedom
Multiple R-squared:  0.6412,	Adjusted R-squared:  0.6165 
F-statistic: 25.92 on 2 and 29 DF,  p-value: 3.505e-07
```

- QUET does not have a significant effect on SBP, when adjusting for AGE,
- When AGE increases, then SBP will increase with 1.045 units,
- This is a significant increase $(p = 0.01)$, confidence interval $(0.26, 1.84)$ (calculate in R with `confint()`).

Confounding

What did we learn from the two previous models?

- Adjustment for AGE leads to a weaker relationship between SBP and QUET.
- AGE is associated with both SBP and QUET, and affects the association between them.

This implies that AGE is a **confounding variable.**

Confounders (more on this topic tomorrow)

Definition

A **confounder** is a variable that is a **common cause** of the exposure and the response (disease), and **NOT an effect** of the exposure or the disease.

- Confounding variables are important when we want to estimate (causal) effects from various exposures.
- As they cause both the exposure and the response, they are likely to cause biases.
- They can be dealt with by **adjusting in a multiple regression model**: always adjust for potential confounders by including them in the regression model!
- Multivariate regression models are thus important to include potential relevant variables.
- Be careful not to include common effects (also called *colliders*).

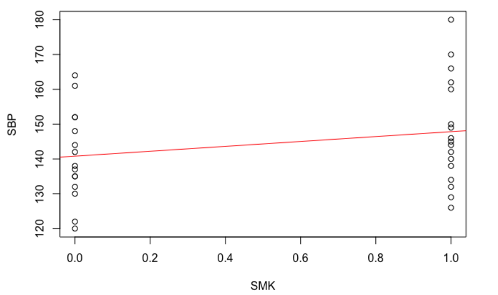
Simple linear regression: SBP vs SMK

```
> fit <- lm(SBP ~ SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ SMK, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-21.824  -9.056  -2.812  11.200  32.176 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  140.800      3.661  38.454   <2e-16 ***
SMK            7.024      5.023   1.398    0.172    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.18 on 30 degrees of freedom
Multiple R-squared:  0.06117,	Adjusted R-squared:  0.02988 
F-statistic: 1.955 on 1 and 30 DF,  p-value: 0.1723
```

- Note that $\hat{b}_0 = 140.80$ and $\hat{b}_1 = 7.02$,
- Confidence interval for b_1 $(-3.24, 17.28)$ (calculate in R with `confint()`)
- $H_0 : b_1 = 0$ is not rejected, as $p = 0.17$,
- Average difference between the two groups is 7.02.

Simple linear regression: SBP vs SMK

```
> plot(SBP ~ SMK, data=bloodpressure)
> abline(reg=fit, col="red")
```



Multiple regression: Combining AGE, QUET and SMK

```
> fit <- lm(SBP ~ QUET + AGE + SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET + AGE + SMK, data = bloodpressure)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5420  -6.1812  -0.7282   5.2908  15.7050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  45.1032    10.7649   4.190 0.000252 ***
QUET          8.5924     4.4987   1.910 0.066427 .  
AGE           1.2127     0.3238   3.745 0.000829 ***
SMK           9.9456     2.6561   3.744 0.000830 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.407 on 28 degrees of freedom
Multiple R-squared:  0.7609,	Adjusted R-squared:  0.7353 
F-statistic: 29.71 on 3 and 28 DF,  p-value: 7.602e-09
```

- Both AGE and SMK have significant effects,
- When AGE increases 1 unit, SBP increases with 1.2 units,
- Confidence interval: $(0.55, 1.88)$, $p = 0.001$,
- Smokers have 10 units higher SBP than non-smokers, confidence interval $(4.5, 15.4)$, $p = 0.001$.

Removing QUET from the model

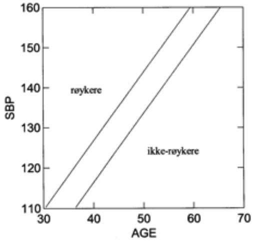
```
> fit <- lm(SBP ~ AGE + SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ AGE + SMK, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-10.639  -5.518  -1.637   4.900  19.616 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  48.0496    11.1296   4.317 0.000168 ***
AGE           1.7092     0.2018   8.471 2.47e-09 ***
SMK          10.2944     2.7681   3.719 0.000853 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.738 on 29 degrees of freedom
Multiple R-squared:  0.7298,	Adjusted R-squared:  0.7112 
F-statistic: 39.16 on 2 and 29 DF,  p-value: 5.746e-09
```

- Both AGE and SMK still have significant effects.
- Removing QUET lead to a slight decrease in the R^2: we might consider keeping it.

Closer look at the effect of AGE and SMK

$$\text{SBP} = 48.05 + 1.71 \cdot \text{AGE} + 10.29 \cdot \text{SMK}$$

- One year increase in age yields an increase of SBP 1.71 units,
- Non-smokers model: $\text{SBP} = 48.05 + 1.71 \cdot \text{AGE}$
- Smokers model: $\text{SBP} = 58.34 + 1.71 \cdot \text{AGE}$

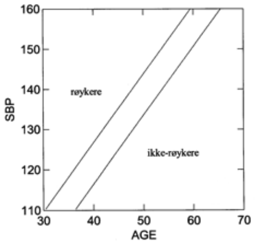


- The effect on SBP of the increase in AGE is the same regardless if one is a smoker or not. Is this realistic?
- NO $\rightarrow$ In reality, the effect of age could be larger for smokers.

Closer look at the effect of AGE and SMK

$$\text{SBP} = 48.05 + 1.71 \cdot \text{AGE} + 10.29 \cdot \text{SMK}$$

- One year increase in age yields an increase of SBP 1.71 units,
- Non-smokers model: $\text{SBP} = 48.05 + 1.71 \cdot \text{AGE}$
- Smokers model: $\text{SBP} = 58.34 + 1.71 \cdot \text{AGE}$



- The effect on SBP of the increase in AGE is the same regardless if one is a smoker or not. Is this realistic?
- NO $\rightarrow$ In reality, the effect of age could be larger for smokers.

Interaction between two explanatory variables

- If the effect of one variable might depend on another variable,
- we have to build a common model for main effects as well as interactions:

$$\text{SBP} = b_0 + b_1 \cdot \text{AGE} + b_2 \cdot \text{SMK} + b_3 \cdot \text{AGE} \cdot \text{SMK}$$

- This is easily done in R with either the "*" or ":" operators:

```
lm(SBP ~ AGE*SMK, data=bloodpressure)
```

or

```
lm(SBP ~ AGE + SMK + AGE:SMK, data=bloodpressure)
```

Interaction between two explanatory variables

```
> fit <- lm(SBP ~ AGE*SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ AGE * SMK, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-11.036  -4.961  -1.958   5.552  20.665 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  58.5743    14.8048   3.956 0.000472 ***
AGE           1.5152     0.2703   5.605 5.32e-06 ***
SMK         -12.8460    21.7153  -0.592 0.558888    
AGE:SMK       0.4349     0.4048   1.074 0.291840    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.717 on 28 degrees of freedom
Multiple R-squared:  0.7405,	Adjusted R-squared:  0.7127 
F-statistic: 26.63 on 3 and 28 DF,  p-value: 2.369e-08
```

- Note that the interaction term is not significant, so we may drop this from the model if there are no particular biological/clinical reasons for keeping it,

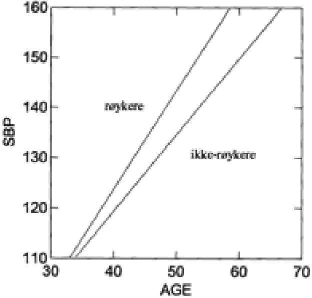
Interpretation

For the non-smokers (SMK $=0$):

$$\begin{aligned}\text{SBP} =& \hat{b}_0 + \hat{b}_1 \cdot \text{AGE} + \hat{b}_2 \cdot 0 + \hat{b}_3 \cdot \text{AGE} \cdot 0 \\ =& 58.57 + 1.52 \cdot \text{AGE}\end{aligned}$$

For the smokers (SMK $= 1$):

$$\begin{aligned}\text{SBP} =& \hat{b}_0 + \hat{b}_1 \cdot \text{AGE} + \hat{b}_2 \cdot 1 + \hat{b}_3 \cdot \text{AGE} \cdot 1 \\ =& 45.72 + 1.96 \cdot \text{AGE}\end{aligned}$$



Other possible interactions

```
> fit <- lm(SBP ~ QUET*SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET * SMK, data = bloodpressure)

Residuals:
     Min       1Q   Median       3Q      Max 
-22.3713  -5.5705  -0.6357   7.4972  17.1051 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   49.312     19.972   2.469   0.0199 *  
QUET          26.303      5.703   4.612 8.01e-05 ***
SMK           29.944     24.164   1.239   0.2256    
QUET:SMK      -6.185      6.932  -0.892   0.3799    
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 8.948 on 28 degrees of freedom
Multiple R-squared:  0.6511,	Adjusted R-squared:  0.6137 
F-statistic: 17.42 on 3 and 28 DF,  p-value: 1.408e-06
```

Other possible interactions

```
> fit <- lm(SBP ~ QUET*AGE, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET * AGE, data = bloodpressure)

Residuals:
    Min      1Q  Median      3Q     Max 
-13.385  -6.208  -2.284   6.243  21.926 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)  
(Intercept) 207.3696    86.3654   2.401   0.0232 *
QUET        -34.1170    25.2168  -1.353   0.1869  
AGE          -1.8468     1.6686  -1.107   0.2778  
QUET:AGE      0.8224     0.4625   1.778   0.0863 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 8.601 on 28 degrees of freedom
Multiple R-squared:  0.6776,    Adjusted R-squared:  0.6431 
F-statistic: 19.62 on 3 and 28 DF,  p-value: 4.742e-07
```

Model selection

- None of these interactions had significant effects, so in the light of a parsimony criterion (so to save degrees of freedom) we will skip the interactions in the final model.
- Automatic model selection is possible, but hard to use in practice.
- Models motivated by causal interpretations should be based on subject matter knowledge, not just an algorithm.

Final multiple regression model

No significant interactions, so we end up with the following model:

$$\text{SBP} = b_0 + b_1 \cdot \text{AGE} + b_2 \cdot \text{QUET} + b_3 \cdot \text{SMK}$$

```
> fit <- lm(SBP ~ QUET + AGE + SMK, data=bloodpressure)
> summary(fit)

Call:
lm(formula = SBP ~ QUET + AGE + SMK, data = bloodpressure)

Residuals:
     Min       1Q   Median       3Q      Max 
-13.5420  -6.1812  -0.7282   5.2908  15.7050 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  45.1032    10.7649   4.190 0.000252 ***
QUET          8.5924     4.4987   1.910 0.066427 .  
AGE           1.2127     0.3238   3.745 0.000829 ***
SMK           9.9456     2.6561   3.744 0.000830 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.407 on 28 degrees of freedom
Multiple R-squared:  0.7609,	Adjusted R-squared:  0.7353 
F-statistic: 29.71 on 3 and 28 DF,  p-value: 7.602e-09
```

Interaction

- Interaction means that the effect of a variable depends on a second variable,
- Not the same a confounding variable,
- Multivariate regression enables us to analyze interaction effects,
- We often need large data sets to get significant interaction effects.

- A variable Z that has an interaction effect on variable X is sometimes called an effect modifier of X.

Assumptions: residuals

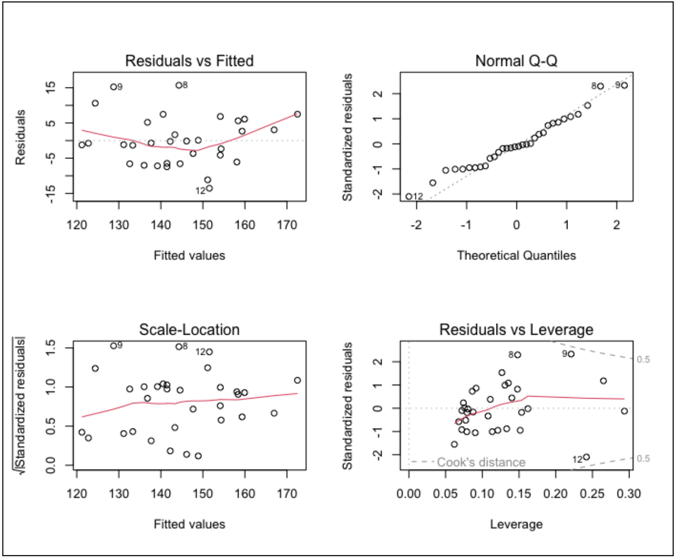
$$
\begin{aligned}
e_1 &= y_1 - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_{11} - \cdots - \hat{\beta}_p \cdot x_{p1} \\
&\vdots \\
e_n &= y_n - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_{1n} - \cdots - \hat{\beta}_p \cdot x_{pn}
\end{aligned}
$$

- Divide by empirical standard deviation to get standardized residuals,
- Standardized residuals should:
  - Be independent,
  - Be normally distributed around 0, regardless of the size of the fitted value.

Check assumptions with R

- Normality plot for residuals (Normal Q-Q plot): top-right plot on next slide
- Residual plot: Plot residuals against fitted values: top-left and bottom-left plots on next slide

Model diagnostics plots in R



Explanatory variables with more than two categories

We will go back to the birth weight data set (birth.dta).

Response variables:

BWT Birth weight

Explanatory variables:

AGE Age

LWT Mothers weight

SMK Smoking status

ETH Ethnicity, $1 =$ White, $2 =$ Black, $3 =$ Other

Categorical variables with more than two levels

- Are formally included in the analysis with dummy variables,
- In some softwares (e.g. SPSS) one has to manually construct two dummy-variables to include ethnicity.
- In R this is done automatically provided we make sure that the categorical variable is included as a `factor` variable.
- Character variables are automatically translated into factor, but not numeric variables.
- With this, R will internally create two new dummy variables under the hood:

ETH	Eth(1)	Eth(2)
White	0	0
Black	1	0
Other	0	1

Simple regression including a categorical predictor (with more than 2 levels)

```
> fit <- lm(bwt ~ as.factor(eth), data=birth)
> summary(fit)

Call:
lm(formula = bwt ~ as.factor(eth), data = birth)

Residuals:
     Min       1Q   Median       3Q      Max 
-2095.01  -503.01   -13.74   526.99  1886.26 

Coefficients:
                    Estimate Std. Error t value Pr(>|t|)    
(Intercept)          2719.69     140.04  19.420   <2e-16 ***
as.factor(eth)other    84.32     165.00   0.511   0.6099    
as.factor(eth)white   384.05     157.87   2.433   0.0159 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 714.1 on 186 degrees of freedom
Multiple R-squared:  0.05075,	Adjusted R-squared:  0.04054 
F-statistic: 4.972 on 2 and 186 DF,  p-value: 0.007879
```

Simple regression including a categorical predictor (with more than 2 levels)

```
> #Since eth is a character variable (text, not numbers), R will actually
> #automatically translate it into a factor variable:
> fit <- lm(bwt ~ eth, data=birth)
> summary(fit)

Call:
lm(formula = bwt ~ eth, data = birth)

Residuals:
     Min       1Q   Median       3Q      Max
-2095.01  -503.01   -13.74   526.99  1886.26

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  2719.69     140.04  19.420   <2e-16 ***
ethother       84.32     165.00   0.511   0.6099
ethwhite      384.05     157.87   2.433   0.0159 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 714.1 on 186 degrees of freedom
Multiple R-squared:  0.05075,	Adjusted R-squared:  0.04054
F-statistic: 4.972 on 2 and 186 DF,  p-value: 0.007879
```

Multiple regression with all available predictors: AGE, LWT, SMK and ETH

```
> fit <- lm(bwt ~ age + lwt + smk + eth, data=birth)
> summary(fit)

Call:
lm(formula = bwt ~ age + lwt + smk + eth, data = birth)

Residuals:
     Min       1Q   Median       3Q      Max
-2281.79  -447.32    22.18   472.27  1747.79

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 2330.426    337.061   6.914 7.61e-11 ***
age           -2.036      9.817  -0.207 0.835894
lwt            3.999      1.737   2.302 0.022480 *
smksmoker   -400.326    109.207  -3.666 0.000323 ***
ethother     110.929    166.953   0.664 0.507251
ethwhite     511.535    157.028   3.258 0.001339 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 681.9 on 183 degrees of freedom
Multiple R-squared:  0.1484,    Adjusted R-squared:  0.1251
F-statistic: 6.377 on 5 and 183 DF,  p-value: 1.744e-05
```

Testing if the multi-level categorical variable is significant

Once we have fitted a regression model including a multi-level categorical variable, we might want to test if there is a significant overall effect of that variable.

We do not get this from the regression output, but we can use the `anova` command to perform a so-called likelihood-ratio test, which compares the model with ETH to the model without ETH.

Remember that 'ETH' is encoded with 2 'dummy variables': R then tests the null-hypothesis that the regression coefficient for both dummy variables are equal to 0.

R output

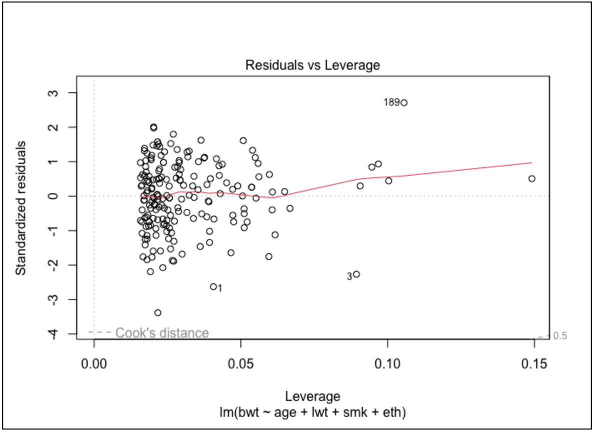
```
> fit <- lm(bwt ~ age + lwt + smk + eth, data=birth)
> fit0 <- lm(bwt ~ age + lwt + smk, data=birth)
> anova(fit0, fit)
Analysis of Variance Table

Model 1: bwt ~ age + lwt + smk
Model 2: bwt ~ age + lwt + smk + eth
  Res.Df      RSS Df Sum of Sq      F    Pr(>F)
1    185 92935223
2    183 85091158  2   7844064 8.4349 0.0003133 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Note that the p-value is 0.0003, so the variable is significant.

Robustness: leverage and influence of observations

- Sometimes a single individual can have a huge influence on the estimates in a regression model,
- This is something we want to avoid as it makes the conclusion more arbitrary,
- A single individual will typically have more influence on the final estimate if it is very untypical in terms of covariates, and also has a relatively large residual value,
- How different an individual is from the average, in terms of covariates, is quantified by the 'leverage',
- It is common to assess the influence by plotting the squared residual against the leverage for every individual,
- We can use the fourth plot of the model diagnostics plots that are generated by running `plot(fit)`.

Standardized residuals vs leverage



- Potential influence points are indicated by their ID.
- We can use Cook's distance > 1 as an indication for a potential influence point (not the case here).

Summary

Key words

- Multiple linear regression
- Confounder / collider (more tomorrow)
- Interaction effects
- Categorical covariates with more than 2 levels
- Regression assumptions / leverage effect

Statistical Science
2010, Vol. 25, No. 3, 289–310
DOI: 10.1214/10-STS330

To Explain or to Predict?

Galit Shmueli

Abstract. Statistical modeling is a powerful tool for developing and testing theories by way of causal explanation, prediction, and description. In many disciplines there is near-exclusive use of statistical modeling for causal explanation and the assumption that models with high explanatory power are inherently of high predictive power. Conflation between explanation and pre-

To Explain
To Predict
or
To Describe?

Galit Shmueli 徐茉莉

Institute of Service Science

國立清華大學 NATIONAL TSING HUA UNIVERSITY

ISBIS 2019 Satellite Conference
August 15-16, 2019
Lanai Kijang, Kuala Lumpur, Malaysia

ISBIS: International Society for Business and Industrial Statistics
An Association of the International Statistical Institute

Definitions: Describe

Descriptive modeling
statistical model for approximating a distribution or relationship

Descriptive power
goodness of fit, generalizable to population

Description: **Sailer et al. (2023)**. Caressed by music: Related preferences for velocity of touch and tempo of music?

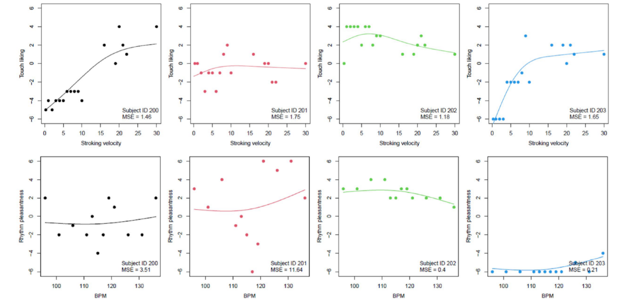


FIGURE 11
Touch (upper) and beat rating patterns (lower) with fit line for four participants. MSE=mean of the squared residuals as a measure of the goodness of the fit.

- Describe relationships between variables x and y.
- We are mainly interested in: the fitted regression curve

Definitions: Explain

Explanatory modeling
theory-based, statistical testing of causal hypotheses

Explanatory power
strength of relationship in statistical model

Explanation: **Kristiansen et al. (2021)**. Mediators Linking Maternal Weight to Birthweight and Neonatal Fat Mass in Healthy Pregnancies

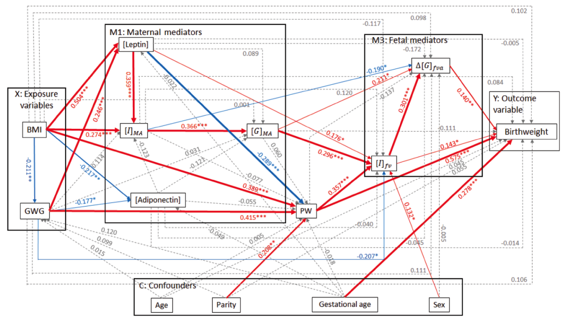


Figure 4. The path analysis with birthweight as the outcome (n = 165) presented with standardized β values. Red arrows symbolize a positive association; blue arrows symbolize an inverse association; gray dashed arrows indicate no statistically significant association. Adiponectin, maternal levels of adiponectin; Age, maternal age; BMI, maternal pregestational body mass index; Leptin, maternal levels of leptin; ΔG_{fva}, fetal venous-arterial glucose difference; G_{MA}, maternal arterial levels of glucose; GWG, gestational weight gain; I_{fv}, fetal venous levels of insulin; I_{MA}, maternal arterial levels of insulin; PW, placental weight; Sex, fetal sex. *$P < .05$; ** $P < .01$; *** $P < .001$.

- Explain/ understand the nature of a relationships between variables x and y.
- We are mainly interested in: coefficients $\hat{a}$, $\hat{b}$ and their p-values

Definitions: Predict

Predictive modeling
empirical method for predicting new observations

Predictive power
ability to accurately predict new observations

Prediction: **Maros et al. (2020)**. Machine learning workflows to estimate class probabilities for precision cancer diagnostics on DNA methylation microarray data

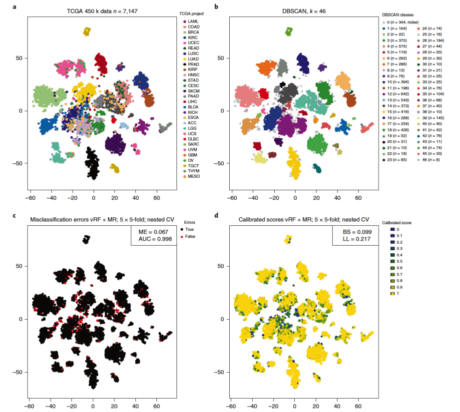


Fig. 2 | External validation on various tumor types from TCGA. a-d, We generated the external validation cohort (n = 7,147) by extracting and combining 450k DNA methylation microarray data from 30 different TCGA projects. First, an unsupervised variance filtering was performed to reduce

- Predict y from other data x
- We are mainly interested in: fitted/ predicted values $\hat{y}$

Monopolies in Different Fields

Different Scientific Goals

Different ***generalization***

Explanatory Model:
test/quantify causal effect between *constructs* for "average" unit in population

Descriptive Model:
test/quantify distribution or correlation structure for *measured* "average" unit in population

Predictive Model:
predict *values* for new/future individual units

Summary: To explain, to predict or to describe

- Description: Scatterplots with the fitted regression curves.

- Explanation: Tables of the estimated regression coefficients with their confidence intervals (or standard errors) and p-values

  Crucial that the model contains the right set of covariates (confounders, not colliders - see tomorrow) and that no strong multi-collinearity exists, normality of the residuals

- Prediction: Prediction performance on a new never seen test data set, e.g. test RSS (sum of squares of residuals) or test R^2

  We do not care about the regression coefficients, therefore inclusion of confounders, avoidance of multi-collinearity etc. not so important.

For more details see the abridged Shmueli (2019) presentation provided to the class.