Regression analysis III

1. Multiple linear regression: categorical covariates with more than 2 levels
2. Model assumptions, leverage
3. Closing: To explain, to predict or to describe

Valeria Vitelli
Oslo Centre for Biostatistics and Epidemiology
Department of Biostatistics, UiO
valeria.vitelli@medisin.uio.no

MF9130E – Introductory Course in Statistics
24.04.2024

Outline

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

1. Morning: Regression II
   - Introduction to **Multiple linear regression** (briefly: multiple regression)
   - More details on linear regression models: **confounding**, **interactions**
2. Afternoon: Regression III
   - **categorical covariates** with more than 2 levels
   - Multiple regression **assumptions**, **leverage** effect
   - To explain, to predict or to describe? How the purpose of the analysis decides what is important

Schedule for today

08.30-10.15:	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
12.45-14.00:	Regression analysis III: Multiple regression (continued), categorical variables, assumptions, leverage effect. To explain, to predict or to describe?
14.00-15.00:	R exercise for regression III
15.00-15.30:	Discussion of the R exercises for regression III in class
15.30-16.00:	Course Summary

Conclusion this morning:
Final multiple regression model

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

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

```
> 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
```

Conclusion this morning: Interaction Effects

- 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

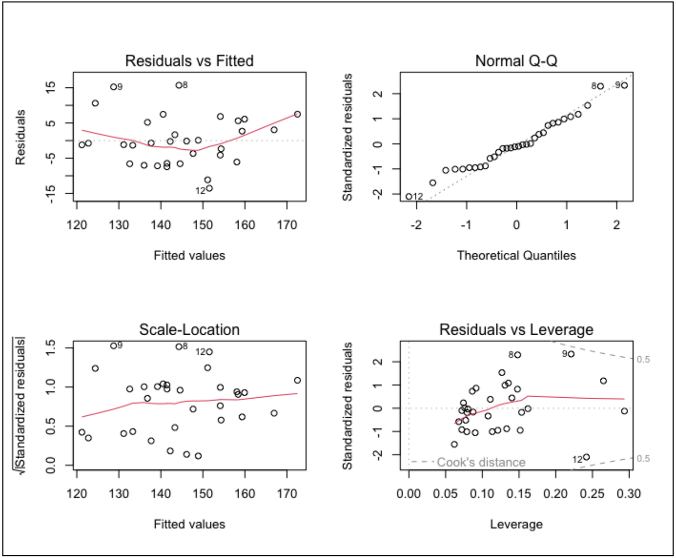
$$
\begin{aligned}
e_1 &= y_1 - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_{11} - \cdots - \hat{\beta}_p \cdot x_{p1} \\
&\vdots \\
e_n &= y_n - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_{1n} - \cdots - \hat{\beta}_p \cdot x_{pn}
\end{aligned}
$$

- 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

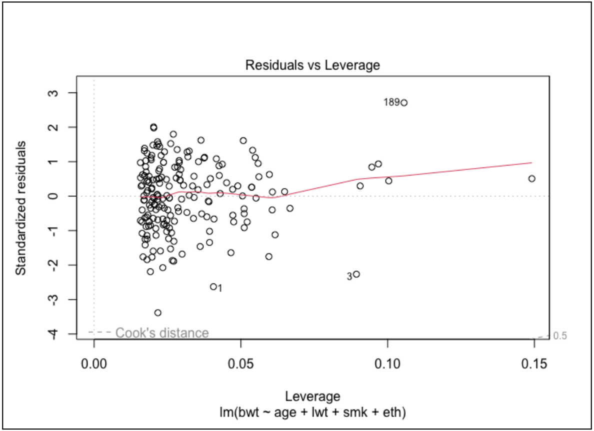
```
> fit <- lm(bwt ~ age + lwt + smk + eth, data=birth)
> fit0 <- lm(bwt ~ age + lwt + smk, data=birth)
> anova(fit0, fit)
Analysis of Variance Table

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

- Categorical covariates with more than 2 levels
- Regression assumptions
- Robustness, 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?

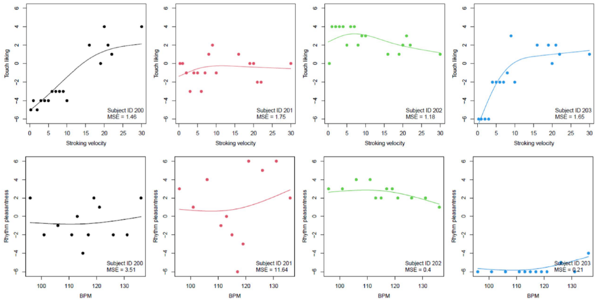


FIGURE 11

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

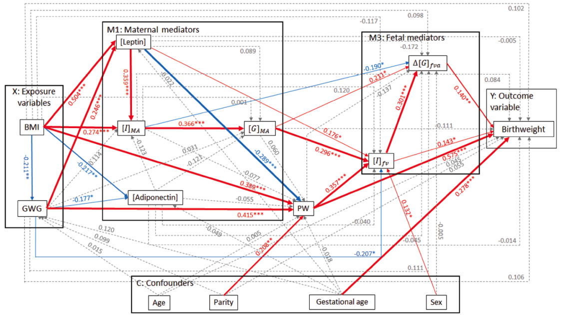


Figure 4. The path analysis with birthweight as the outcome (n = 165) presented with standardized β values. Red arrows symbolize a positive association; blue arrows symbolize an inverse association; gray dashed arrows indicate no statistically significant association. Adiponectin, maternal levels of 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

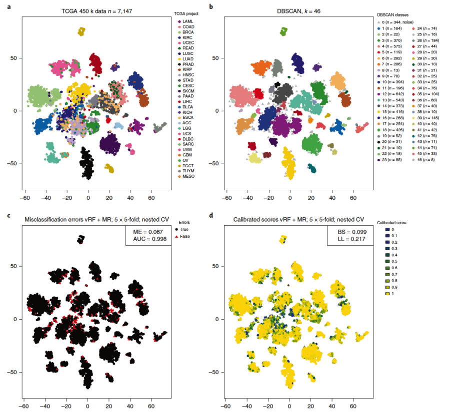


Fig. 2 | External validation on various tumor types from TCGA. a–d, We generated the external validation cohort (n = 7,147) by extracting and combining 450k DNA methylation microarray data from 30 different TCGA projects. First, an unsupervised variance filtering was performed to reduce

- 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.