Multiple Regression

Philippe Rast

PSC 204B:
General Linear Model
January, 16

UC Davis, Winter 2018

Topics

Single-level Regression:

Week 1	Linear Regression (G&H: 3,4)
Week 2	Multiple Regression
Week 3	Violation of Assumptions
Week 4	Logistic Regression and GLM (G&H: 5, 6)
Week 5	Model comparison, Over-fitting, Information Criteria (McE: 6)
Week 6	Regression inference via simulations (G&H: 7–10)

Multilevel Regression:

Week 7	Multilevel Linear Models (G&H: 11–13)
Week 8	Multilevel Generalized Models (G&H: 14, 15)
Week 9	Bayesian Inference (G&H: 18 / McE: 1, 2, 3)
Week 10	Fitting Models in Stan and brms (G&H: 16, 17 / McE: 11)

Overview

Tuesday:

1. Goodness of fit
2. Suppression
3. Polynomial Models

Thursday:

4. Linear Transformations
5. Nonlinear Transformations

Multiple Regression: Example and Data

Table: Data from the Longitudinal Aging Study Amsterdam

Person	1	2	3	4	5	6	7	8	9	10
x_1: Age	56	57	58	61	63	72	73	75	76	83
x_2: Education†	9	13	11	9	12	8	10	8	11	12
y: PS‡	25	34	31	19	38	21	23	16	18	17

†Years in school ‡Processing speed.

```
lm(formula = PS ~ age + Education)

            Estimate Std. Error t value Pr(>|t|)
(Intercept)  35.8551    14.6944   2.440   0.0448 *
age          -0.5108     0.1594  -3.204   0.0150 *
Education     2.2109     0.8581   2.576   0.0367 *
---
Residual standard error: 4.523 on 7 degrees of freedom
Multiple R-squared:  0.7296,Adjusted R-squared:  0.6524
```

Goodness of Fit: Variance Explained (R^2)

- How good is our prediction of y given x_1 and x_2?
- Analogous to the simple linear regression, variance explained is defined as:

$$R^2 = \frac{\hat{\mathbf{y}}^{*\prime}\hat{\mathbf{y}}^*}{\mathbf{y}^{*\prime}\mathbf{y}^*} = 1 - \frac{\hat{\boldsymbol{\epsilon}}'\hat{\boldsymbol{\epsilon}}}{\mathbf{y}^{*\prime}\mathbf{y}^*}.$$

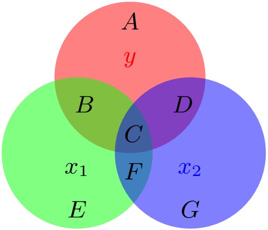
- If we use our values from the example we obtain

$$R^2 = \frac{386.42}{529.60} = 1 - \frac{143.18}{529.60} = 0.7296 \approx 73\%.$$

- Age and education explain 73% of the variance in processing speed.

Goodness of Fit: Variance Decomposition

- How much variance do age and education explain by themselves?
- Problem: Strictly speaking “by-themselves” does not exist in multiple regression.
- Venn-diagram (cf. Cohen et al., 2003)



- The circles represent the (unit) variances of y, x_1 and x_2.

Goodness of Fit: Variance decomposition

- Hence, the total variance explained in y by x_1 and x_2 is
$$R_T^2 = \frac{B+C+D}{A+B+C+D}.$$
- By x_1 *alone* explained variance is
$$R_{x_1}^2 = \frac{B}{A+B+C+D}.$$
- Accordingly, the variance explained of x_2 *alone* is
$$R_{x_2}^2 = \frac{D}{A+B+C+D}.$$
- The amount of variance which is explained by both, x_1 *and* x_2 is
$$R_{x_1,x_2}^2 = \frac{C}{A+B+C+D}.$$

Goodness of Fit: Variance decomposition

- Given these definitions it follows that the
  - variance explained by x_1 in a simple linear regression corresponds to

$$R_1^2 = \frac{B + C}{A + B + C + D}.$$

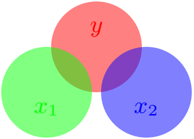
  - variance explained by x_2 in a simple linear regression corresponds to

$$R_2^2 = \frac{C + D}{A + B + C + D}.$$

Note C is the part which is explained by both x_1 and x_2. Hence, the total variance explained of the multiple regression model is smaller than the sum of variance explained from simple regression models based on x_1 and x_2 alone.

Goodness of Fit: Orthogonality

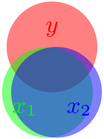
- The Venn-diagram can be used to visualize what happens when predictor variables are independent (orthogonal) from each other.



- If x_1 and x_2 are independent from each other then $C = 0$ and $F = 0$.
- In this case the total variance explained in a *multiple* regression corresponds exactly to the sum of explained variances in a simple regression.
- It holds: $R_T^2 = R_1^2 + R_2^2$ (only if $x_1 \perp x_2$).

Goodness of Fit: Multicollinearity

- The opposite is the case if we have **multicollinearity**, that is if two (or more) predictors are highly correlated among each other. In that case, a Venn-diagram might look like this:



- In this case the proportion C is so large in comparison to B and D that it is difficult to describe what unique contribution either x_1 or x_2 have on y.
- In other words: One of the predictors x_1 or x_2 is practically *redundant*.

Goodness of Fit: Commonality Analysis

- The common and unique amounts of variance can be obtained by means of commonality analysis.
- Approach for two predictor variables†
  - Step 1: Compute total variance explained (R_T^2) in a multiple regression.
  - Step 2: Compute variance explained only for x_1 (R_1^2) in a simple linear regression.
  - Step 3: Compute variance explained only for x_2 (R_2^2) in a simple linear Regression.
  - Then $B = R_T^2 - R_2^2 = 73\% - 33\% = 40\%$.
  - Then $D = R_T^2 - R_1^2 = 73\% - 47\% = 26\%$.
  - Then $C = R_1^2 + R_2^2 - R_T^2 = 47\% + 33\% - 73\% = 7\%$.

† For the approach in the case of more than two predictor variables see e.g. Seibold, D. R. & McPhee, R. D. (1979). Commonality analysis: A method for decomposing explained variance in multiple regression analysis. *Human Communication Research*, *5*, 355-365.

Suppression

- Explained variance in a multivariate model typically does not exceed the sum of the single R^2's
- This is not always the case
- Suppression:
  - Suppression variables increase predictive validity of another variable by its inclusion into a regression.
  - The omission of suppressors or confounders will lead to either an underestimation or an overestimation of the effect of X on Y,

Goodness of Fit: Three Cases of Suppression, Overview

- Three variables, the dependent y, and two predictors x_1, x_2.
- Assumption 1: x_2 is the suppressor variable.
- Assumption 2: $r_{yx_1} \geq r_{yx_2}$, that is, the criterion is more closely related to the first predictor.

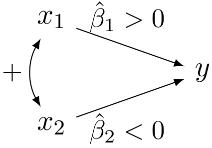
Table: Types of Suppression

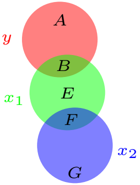
$r_{x_1x_2}$	r_{yx_2}	Type
$\neq 0$	$= 0$	Classical Suppression
> 0	$0 < r_{yx_2} < r_{yx_1} r_{x_1x_2}$	Net Suppression
< 0	> 0	Cooperative Suppression

- Indication of presence of suppression: $R_T^2 > r_{yx_1}^2 + r_{yx_2}^2$

Goodness of Fit: Three Cases of Suppression

1. Classical Suppression





- y is independent of the suppressor variable x_2 ($r_{yx_2} = 0$, right graph).
- x_1 and x_2 are connected positively (curved arrow).
- Then a negative regression weight is created for x_2 (left graph)!
- The same applies to the inverse relation.
- x_2 suppresses irrelevant variance in x_1 for the prediction of y (area F in the graph).

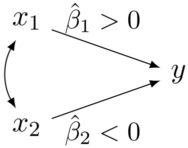
Goodness of Fit: Three Cases of Suppression

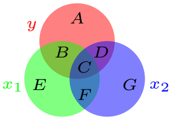
1. Classical Suppression: Example

- y: Grade in stats class
- x_1: Stats-quiz (time-limited)
- x_2: Reading speed (suppressor)
- Relations: $r_{yx_1} = .38, r_{yx_2} = 0, r_{x_1x_2} = .45.$
- Regression models
  - Model 1: $R^2_{y\cdot x_1} = r^2_{yx_1} = .144$
  - Model 2: $R^2_{y\cdot x_2} = r^2_{yx_2} = 0$
  - Model 3: $R^2_{y\cdot x_1,x_2} = \frac{r^2_{yx_1}+r^2_{yx_2}-2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1-r^2_{x_1x_2}} = \frac{r^2_{yx_1}}{1-r^2_{x_1x_2}} = .181$
- The reading speed suppresses the proportion of the variance in the stats-quiz, which is irrelevant for the grade.

Goodness of Fit: Three Cases of Suppression

2. Net suppression





- y correlates positively with x_1 and x_2, with $r_{x_1x_2} > 0$ and $0 < r_{yx_2} < r_{yx_1}r_{x_1x_2}$.
- Then the regression from y to x_2 is negative and x_2 is a suppressor variable.
- The same applies to inverse relation.
- Again, x_2 suppresses those variance parts in x_1 that are irrelevant for predicting y.

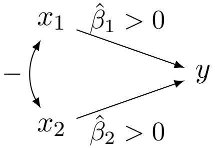
Goodness of Fit: Three Cases of Suppression

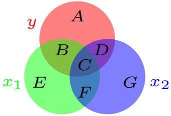
2. Net suppression: Example

- y: Degree of damage
- x_1: Severity of fire
- x_2: Number of fire fighters in action (suppressor)
- Relations: $r_{yx_1} = .65, r_{yx_2} = .25, r_{x_1x_2} = .70.$
- It holds: $0 < r_{yx_2} < r_{yx_1} r_{x_1x_2}$, because $0 < .25 < .455.$
- Regression models
  - Model 1: $R^2_{y \cdot x_1} = r^2_{yx_1} = .42$
  - Model 2: $R^2_{y \cdot x_2} = r^2_{yx_2} = .0625$
  - Model 3: $R^2_{y \cdot x_1, x_2} = \frac{r^2_{yx_1} + r^2_{yx_2} - 2 r_{yx_1} r_{yx_2} r_{x_1x_2}}{1 - r^2_{x_1x_2}} = .505$
- If we keep the severity of the fire constant, then a higher number of firefighters in action will reduce the damage.

Goodness of Fit: Three Cases of Suppression

3. Cooperative suppression





- y correlates positively with x_1 and positively with x_2.
- x_1 and x_2 correlate negatively.
- The same applies to inverse relations.
- Now, x_1 and x_2 in the other variable suppress those variance parts that are irrelevant for a y prediction.
- Both variables x_1 and x_2 are then suppressor variables.

Goodness of Fit: Three Cases of Suppression

3. Cooperative suppression: Example

- y: Grade in stats class
- x_1: Statistics knowledge of instructor
- x_2: Comprehensibility of instructor
- Relations: $r_{yx_1} = .30, r_{yx_2} = .25, r_{x_1x_2} = -.35.$
- Regression models
  - Model 1: $R^2_{y\cdot x_1} = r^2_{yx_1} = .09$
  - Model 2: $R^2_{y\cdot x_2} = r^2_{yx_2} = .0625$
  - Model 3: $R^2_{y\cdot x_1,x_2} = \frac{r^2_{yx_1}+r^2_{yx_2}-2r_{yx_1}r_{yx_2}r_{x_1x_2}}{1-r^2_{x_1x_2}} = .234$
- If the statistics knowledge of the instructor is kept constant, comprehensibility gains in importance. If the comprehensibility is kept constant, the statistical knowledge becomes more important.

Goodness of Fit: Three Cases of Suppression

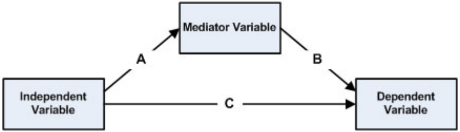
To summarize

- Suppression variables increase the predictive validity of another variable by its inclusion into a regression equation.
- Three types of suppression
  - Classical suppression
  - Net suppression
  - Cooperative suppression

- Omitting suppressor variables either reduces or artificially inflates the magnitude of a relationship between two variables.

Mediation

- Models in which a the influence of a variable on the outcome is mediated via another variable are termed *mediation* models.



- In the figure, the influence of the independent variable x_1 on outcome y is mediated by variable the mediator variable x_2.
- We can quantify the amount of mediation, i.e., whether the effect of x_1 on y is *completely* or only *partially* mediated by x_2. We now have two causal levels.
- See: http://davidakenny.net/cm/mediate.htm

Mediation: Approach

- Practical approach to test mediation

Step 1: Show that the causal variable is correlated with the outcome: Via regression analysis, show that x_1 has effect on y (estimate test path c, *total effect* of x_1).

Step 2: Show that the causal variable is correlated with the mediator: Via regression analysis, show that x_1 has effect on x_2 (estimate and test path a).

Step 3: Show that the mediator affects the outcome variable. Via regression analysis, show that x_1 and x_2 have an effect on y (estimate and test path b and c').

- For our example (slide 4) we obtain for the variables $y =$ PS, $x_1 =$ Age und $x_2 =$ Education.

Step 1: $\beta_c = -0.55^*$.

Step 2: $\beta_a = -0.02$.

Step 3: $\beta_b = 2.21^*$, $\beta_{c'} = -0.51^*$.

  - The effect of mediation is $\frac{c-c'}{c} = .072 \approx 7\%$.
  - Approximately 7% of the age effect on PS are mediated by education.

Predictors

Curvilinear Relations

- Polynomials
- "Construction" of curvilinear functions
- Interpretation of parameters
- Nonlinear transformations
- Outlook on alternative measures

Linear Model: Curvilinear Outcome

- A model is said to be linear when it is linear in *its parameters*.

e.g. The model $y = \beta_0 + \beta_1 x + \beta_2 x^2 + \epsilon$

and

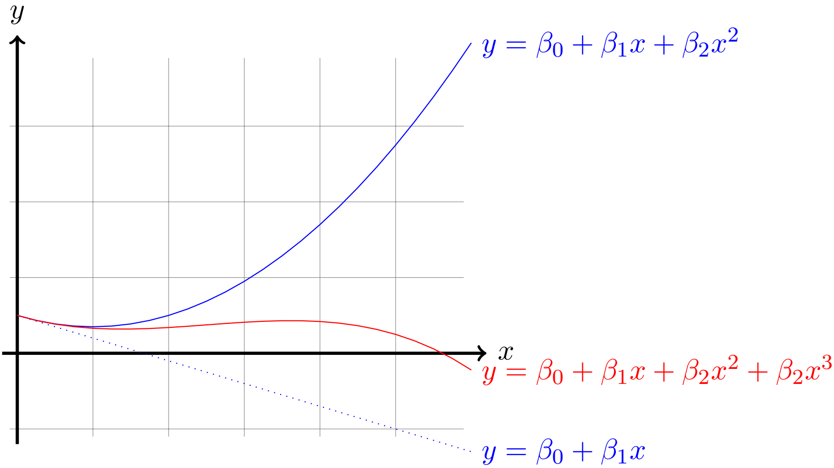
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1^2 + \beta_4 x_2^2 + \beta_5 x_1 x_2 + \epsilon$

are linear in their parameters.

i.e. Predictors are simply multiplied by regression coefficient.

- Term with highest exponent is referred to as the *leading term* or *highest order term*.
- Leading term determines the overall shape of regression function
- Outcome "looks" curved

Linear Model: Curvilinear Outcome



Linear Model: Curvilinear Outcome

Different solutions to the "problem" of non-linearity

- Polynomial regression
- Monotonic nonlinear transformation
- Nonlinear regression
- nonparametric regression

- Polynomial models can be used in those situations where the relationship between study and explanatory variables is curvilinear
- Sometimes a nonlinear relationship in a small range of explanatory variable can also be modeled by polynomials

Polynomial models in one variable

- The k^{th} order polynomial model in one variable is given by
  $y = \beta_0 + \beta_1 x + \beta_2 x^2 + ... + \beta_k x^k + \epsilon$
- If $x_j = x_j$, $j = 1, ..., k$, then the model is a multiple linear regressions model with k explanatory variables $x_1, x_2, ..., x_k$ but with onle one predictor$x$.
- Essentially, x^k is an interaction of x with k times itself.
- We can also write $\mathbf{y} = \mathbf{X}\beta + \epsilon$
- The linear regression model includes the polynomial regression model.
- Techniques for fitting linear regression model can be used for fitting the polynomial regression model.

Polynomial models in one variable

For example

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \epsilon_i$$

- In matrix notation:

  $\hat{\mathbf{y}} = \mathbf{X}\hat{\boldsymbol{\beta}}$

  with elements

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 \\ 1 & x_2 & x_2^2 \\ \vdots & \vdots & \vdots \\ 1 & x_n & x_N^2 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

  and estimator $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ for $\hat{\boldsymbol{\beta}}$

- These models can be conveniently estimated via the OLS approach

Polynomial Models: What Order?

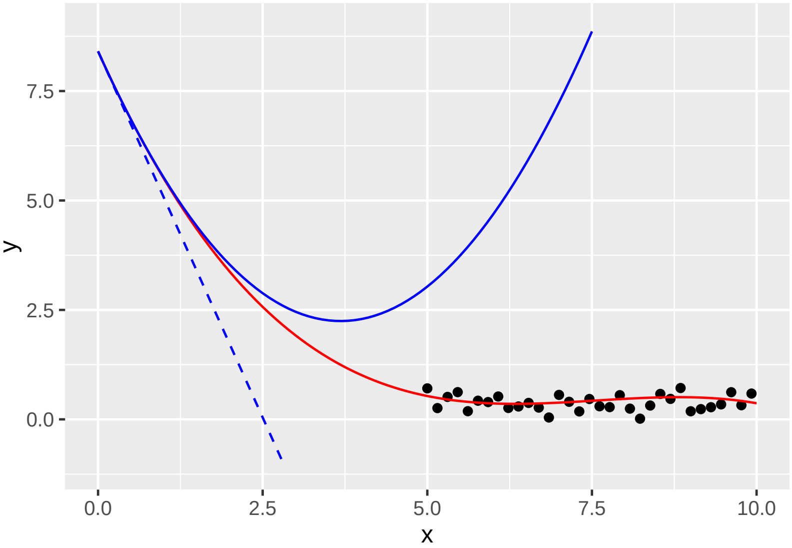
- We can fit up to a polynomial of oder $(q-1)$ for a variable whose scale contains q distinct values.
- Selection process should be ideally theoretically guided
- Polynomials of higher order are difficult to interpret
- Polynomial equations may approximate non-linear relationships

Interpretation of Parameters

- The interpretation of parameter β_0 (intercept) may not be useful.
- Given that we usually don't extrapolate from polynomial models, all parameters outside the data range may not be interpreted.
- If we wish to interpret, e.g. intercept, we need to "pull" it into the observed data range.
- Centering $x_i - \bar{X}$ shifts the y-axis into the data range and β_0 may be interpreted.
  - If we mean-center, the slope β_1 gives us the average slope
  - Centering at different places along the x-axis may be used to address substantive questions

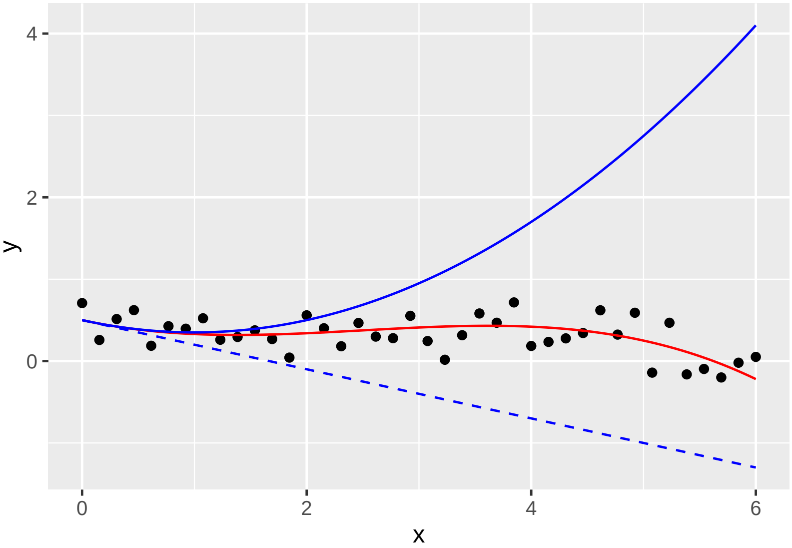
Interpretation of Parameters: Example

Intercept is outside of data range:



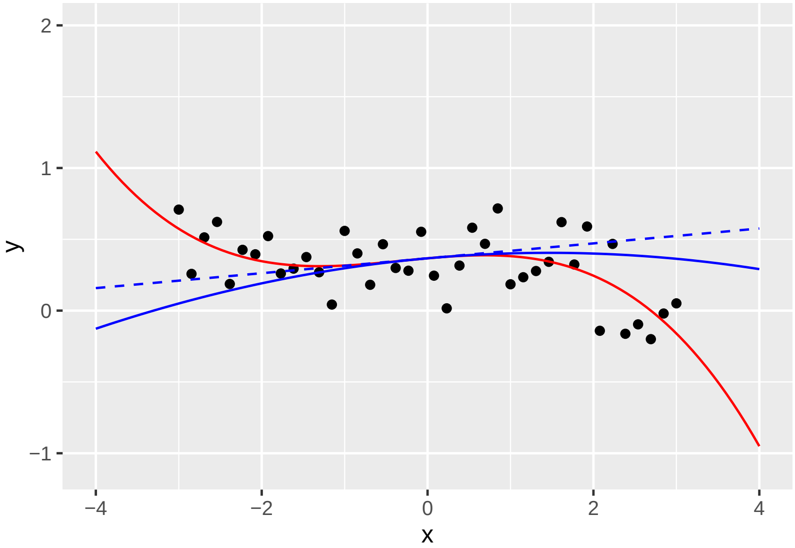
Interpretation of Parameters: Example

Intercept is at the beginning of data range



Interpretation of Parameters: Example

Intercept is at the mean $\bar{x}$



Effects of Centering

Example: Verbal learning across five trials

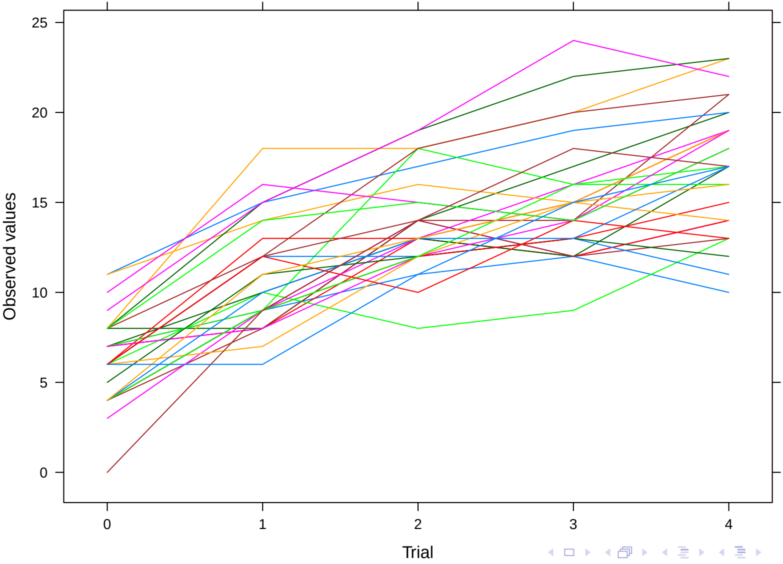


Illustration of Centering

E.g. In higher order longitudinal models ($\hat{y} = \beta_0 + \beta_1 t + \beta_2 t^2$) re-centering of time ($t$) yields different models/estimates/significance for lower order terms:

Time coded as: 0, 1, 2, 3, 4

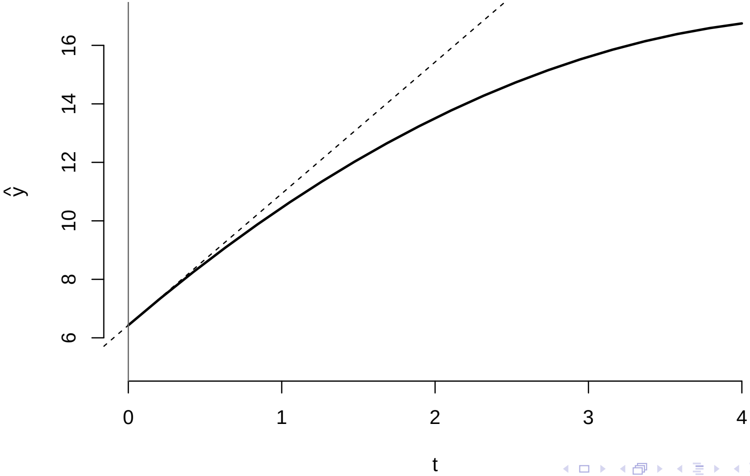
	Value	Std.Error	DF	t-value	p-value
Intercept	6.43	0.21	1342.00	30.59	0.00
Time	4.50	0.13	1342.00	35.74	0.00
Time^2	-0.48	0.03	1342.00	-15.86	0.00

Time coded as: -4, -3, -2, -1, 0

	Value	Std.Error	DF	t-value	p-value
Intercept	16.78	0.21	1342.00	79.83	0.00
Time	0.67	0.13	1342.00	5.34	0.00
Time^2	-0.48	0.03	1342.00	-15.86	0.00

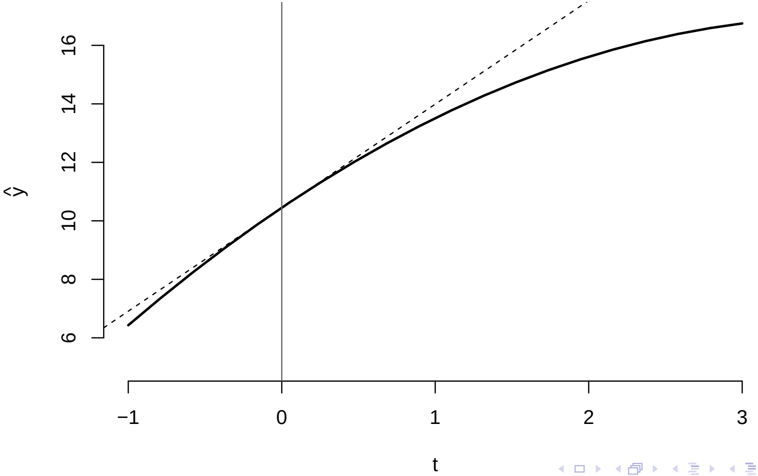
Simple slope or Instantaneous Rate of Change

Example: $\frac{\partial \hat{y}}{\partial t} = \beta_1 + 2\beta_2 t = 4.50 + 2(-0.48)t = 4.50$



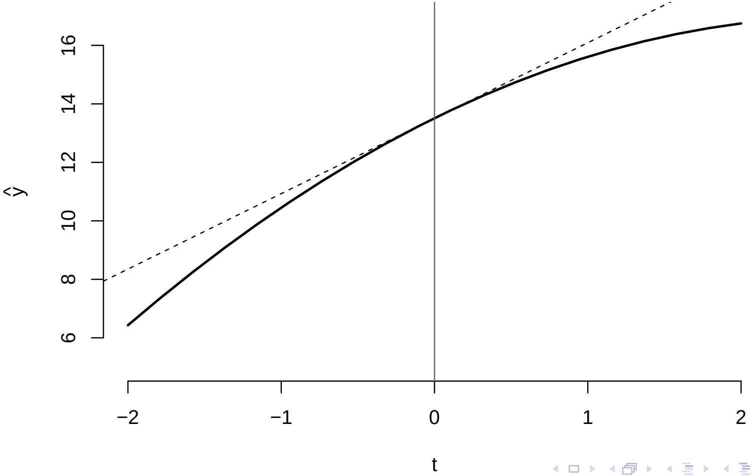
Simple slope or Instantaneous Rate of Change

Example: $\frac{\partial \hat{y}}{\partial t} = \beta_1 + 2\beta_2 t = 4.50 + 2(-0.48)t = 3.54$



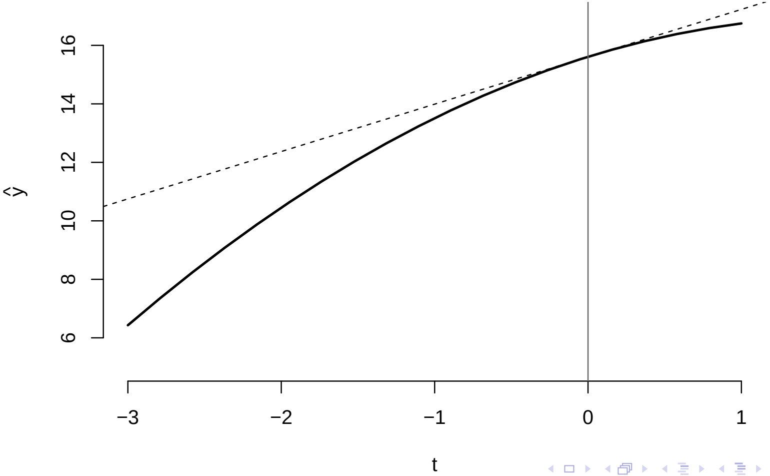
Simple slope or Instantaneous Rate of Change

Example: $\frac{\partial \hat{y}}{\partial t} = \beta_1 + 2\beta_2 t = 4.50 + 2(-0.48)t = 2.58$



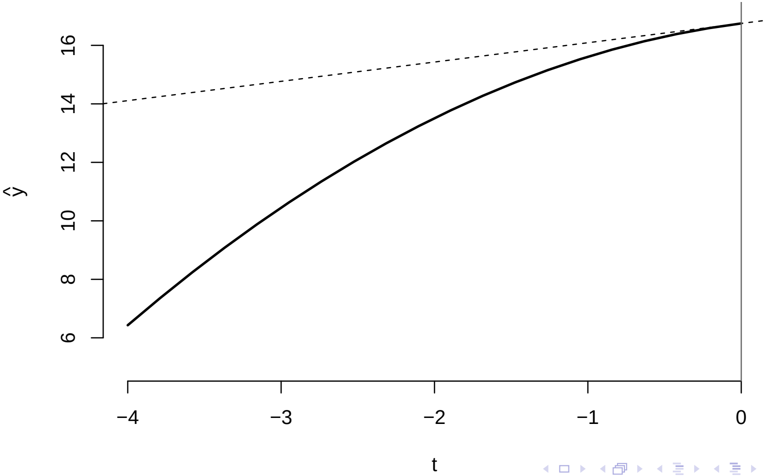
Simple slope or Instantaneous Rate of Change

Example: $\frac{\partial \hat{y}}{\partial t} = \beta_1 + 2\beta_2 t = 4.50 + 2(-0.48)t = 1.62$



Simple slope or Instantaneous Rate of Change

Example: $\frac{\partial \hat{y}}{\partial t} = \beta_1 + 2\beta_2 t = 4.50 + 2(-0.48)t = 0.66$



Simple slope or Instantaneous Rate of Change

- Simple slopes are essentially tangent lines at a particular value of x
- We can obtain these slopes by either re-centering x to that given value, *or*
- we can obtain these slopes by calculating the partial derivatives with respect of x for a given regression function – and then enter the given value of x to obtain the slope of the tangent

Multimorbidity: Example from Clinical Research

Multimorbidity: Co-occurrence of medical conditions in same individual

Investigation of multimorbidity and aging

- High prevalence
  - Over 65 year, 65% report at least one and 50% report two or more medical conditions
- Complex, multi-variate
- Highly intertwined
  - Aging-related changes in physical and cognitive capabilities and mental health
  - Known to affect psychological distress and quality of life
  - Presence of psychological distress increases with severity of multimorbidity / number of medical conditions

Aim: Understand how the effects of chronic conditions evolve over time relative to aging-related and end of life changes

Breaking Down Complexity

Identification of periods in time where multimorbidity impacts particular outcomes, such as depressive symptoms, versus periods of time where this is not the case

- Reduce complexity of the phenomenon
- Identify regions of significance
- ▷ *When* does it matter?

- Methodological counterpart
  - Johnson-Neyman (J-N) Technique / direct regression for rates of change
  - Probe moderators and identify regions of statistical significance
  - In context of a curvilinear longitudinal model with higher-order terms

J-N Technique

- Originally developed in ANCOVA framework.
  - Determine significance of difference among two groups on one variable while holding constant two other variables
- Extend to longitudinal modeling
- ▷ Exact computation of conditions and boundary values where moderator elicits statistically significant slopes

Health and Retirement Study

HRS:

- Longitudinal Study
- $N = 2526$ who died in the period between 1994 to 2006 and who were between 50 and 90 years of age at their death
  - 51% female
  - Age at death $M = 76$, $SD = 9$
  - Average time-to-death was 7.9 years ($SD = 2.58$ years).
- Measurement interval: 2 years

Variables of Interest

- **Depressive Symptoms**
  - Short version of the Center for Epidemiological Studies Depression Scale (CES-D)
  - Range: 0–8
- **Multimorbidity**
  - Comprises: Hypertension, diabetes, CVD, stroke, and cancer (present $= 0$, not present $= 1$)
  - Summed across conditions to obtain Multimorbidity Index
  - MMI: $M = 1.4$, $SD = 1.1$, ranging from 0 to 5

  MMa average multimorbidity across study
  MMc wave-specific multimorbidity, person centered
- Time scaled as **time-to-death**
- **Age–at–Death**

Statistical Model

Curvilinear model

- Polynomial model
- Interactions and quadratic effects
- Main Effects:
  - Time–to–death (TD), Age–at–death (AD), MMc, and MMa
  - 13 interaction terms
- Lower order terms are dependent on highest order interactions:
  - $AD \times MMa \times TD^2$
  - $AD \times MMc \times TD^2$
- J-N technique
- Mixed Effect model with focus on fixed effects

Results

Selected fixed effects estimates for CES-D

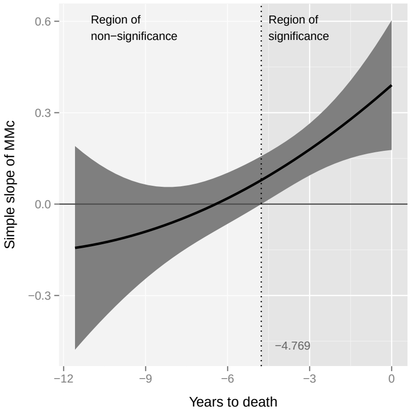
	Intercept at death			5 years prior to death		
	Estimate	S.E.	t-Value	Estimate	S.E.	t-Value
Intercept	2.14**	0.14	15.34	1.38**	0.07	19.03
TD	0.22**	0.05	3.99	0.09**	0.02	5.67
MMc	0.39**	0.11	3.59	0.07	0.04	1.65
MMa	0.15	0.19	0.80	0.79**	0.09	8.61
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
TD$\times$MMc	0.08	0.04	1.94	0.05**	0.01	3.43
TD$\times$MMa	-0.22**	0.07	-3.15	-0.04	0.02	-1.91
AD$\times$MMa	0.02	0.02	1.27	-0.02**	0.01	-2.98
AD$\times$MMc$\times$TD^2	0.00	0.00	0.55	0.00	0.00	0.55

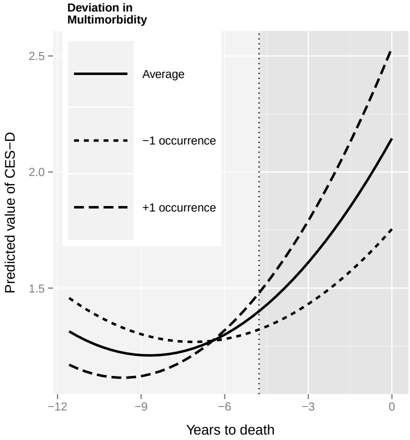
Changes in Multimorbidity: MMa

- Instantaneous rates of change of CES-D along the time-to-death axis across three conditions of average multimorbidity (MMa)

Changes in Multimorbidity: MMc

- Differing effect of MMc on CES-D across time
- One-unit change in MMc impacts changes in CES-D differently.
- From 4.8 years prior to death changes are significant.





Effect of Age at Death: Animation across Age Band

Animation

Example: Concluding Remarks

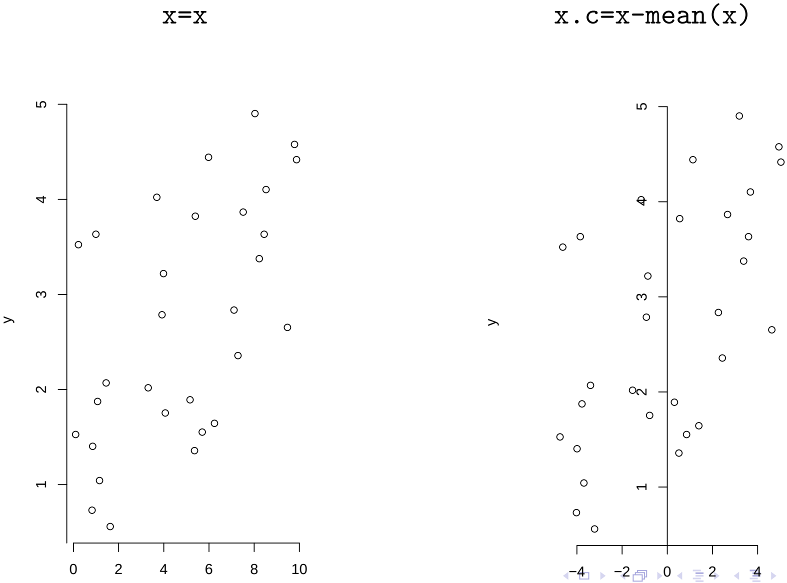
- Multimorbidity reveals different facets:
    - Increase in average MM shows opposite effect of increases within persons
    - Might explain different findings in the past in these regards
- Multivariate: Type of MM, Proximity to death and age-at-death have all different effects on well-being

- J-N technique is useful to continuously probe moderating effects
- ▷ Identify when certain effects are or are not statistically significant
- In the context of multimorbidity
- ▷ Particularly useful for interpreting the complex interactions across time and identify regions of significance

Linear Transformations

- Transformations that maintain the scale and interval
- Centering
  - Grand-mean centering: $x_i - \bar{x}$
  - Centering to specific value, eg. 5: $x_i - 5$
- Multiplication: Unit changes
  - 1 year $(\times 12)$ 12 months $(\times 4.3)$ 52 weeks...
  - Height in mm, cm, m etc.
- Standardization to a metric: z, T, etc.

Centering



Centering

```
lm(formula = y ~ x)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.60213 0.35180 4.554
x            0.23068 0.06115 3.773
---

Residual se: 1.039 on 28 df
Multiple R^2:  0.337
Adjusted R^2:  0.3133
```

```
lm(formula = y ~ x.c)

Coefficients:
Estimate Std. Error t value Pr(>|t|)
(Intercept)  2.71993 0.18965 14.342
x.c          0.23068 0.06115  3.773
---

Residual se: 1.039 on 28 df
Multiple R^2:  0.337
Adjusted R^2:  0.3133
```

Alternatively

- Using a conventional centering point
  - Center based on an understandable reference point

Rescaling

- Changing units (eg. x/10)

```
lm(formula = y ~ x)
coef.est coef.se
(Intercept) 4.73     0.39
x           0.22     0.07
---
n = 30, k = 2
residual sd = 1.15, R-Squared = 0.25

> x2 <- x/10

lm(formula = y ~ x2)
coef.est coef.se
(Intercept) 4.73     0.39
x2          2.19     0.71
---
n = 30, k = 2
residual sd = 1.15, R-Squared = 0.25
```

Standardization

So far:

- (Re-)centered
- Rescaling

Combination of both approaches

- Re-centering and rescaling:

  e.g. Standardizing

Standardizing $\hat{\boldsymbol{\beta}}$-Parameters

- The parameters in vector $\hat{\boldsymbol{\beta}}$ indicate by how many units the dependent variable changes with respect to changes in the independent variable.
- The advantage of these raw scores is that corresponding parameters from different studies can be compared across different samples.
- Example: In our data we obtained $\hat{\beta}_1 = -.51$ and $\hat{\beta}_2 = 2.21$ for the influence of age and education on PS.

  In another, independent study, we may obtain $\hat{\beta}_1 = -.66$ and $\hat{\beta}_2 = 3.01$ for *the same variables*.

  We can now compare these estimates directly: The estimates for the β-parameter in the second study are larger compared to the first study.†

†Whether they are significantly different i.e., whether their respective populations differ would have to be tested.

Standardizing $\hat{\beta}$-Parameters

- To compare the relative influence of independent variables within a sample the $\hat{\beta}$-parameters are not useful – unless they are all on the exact same metric.
- For meaningful comparisons we need to *standardize* the $\hat{\beta}$-parameters.
- We can achieve this in two ways:
  - Either by standardizing the dependent and independent variables (e.g. via z-standardization). And then running the regression.
  - Or we multiply the $\hat{\beta}$-parameter of the independent variable j with the ratio s_j/s_y (the ratio of the standard deviations).

Standardizing $\hat{\beta}$-Parameters

```
age = c(56, 57, 58, 61, 63, 72, 73, 75, 76, 83)
Education = c( 9, 13, 11,  9, 12,  8, 10,  8, 11, 12)
PS = c(25, 34, 31, 19, 38, 21, 23, 16, 18, 17)

age.z <- (age - mean(age))/sd(age)

age.z <- scale(age)
educ.z <- scale(Education)
```

Table: Data from the Longitudinal Aging Study Amsterdam (z-stand.)

Person	1	2	3	4	5	6	7	8	9	10
x_1: Age	-1.19	-1.09	-0.98	-0.67	-0.46	0.48	0.59	0.79	0.90	1.64
x_2: Education†	-0.73	1.53	0.39	-0.73	0.96	-1.30	-0.17	-1.30	0.39	0.96
y: PS‡	25	34	31	19	38	21	23	16	18	17

†Number of school years ‡Processing speed.

Standardizing $\hat{\beta}$-Parameters

```
lm(formula = PS ~ age.z + educ.z)
coef.est coef.se
(Intercept) 24.20     1.43
age.z       -4.86     1.52
educ.z       3.91     1.52
---
n = 10, k = 3
residual sd = 4.52, R-Squared = 0.73
```

- The unit change is now 1 SD.
- $\sigma_{\text{age}} = 9.51$
- Equivalent as age/9.51

```
lm(formula = PS ~ I(age/9.51) + educ.z)
coef.est coef.se
(Intercept) 58.63    10.84  [age not centered!]
I(age/9.51) -4.86     1.52
educ.z       3.91     1.52
```

Standardizing $\hat{\beta}$-Parameters

Interpretation

- After z-standardizing the different variables are on a comparable metric – with a study. One unit corresponds to on standard deviation.
- We can now interpret the z-standardized β-parameters:
  - Coefficient of age (-4.86) is larger than education (3.91)
  - Change in one age SD is associated with a decrease in PS of 4.86 seconds
  - Change in one education SD is associated with an increase in PS of 3.91 seconds

Standardizing $\hat{\beta}$-Parameters

```
age = c(56, 57, 58, 61, 63, 72, 73, 75, 76, 83)
Education = c( 9, 13, 11,  9, 12,  8, 10,  8, 11, 12)
PS = c(25, 34, 31, 19, 38, 21, 23, 16, 18, 17)

(age - mean(age))/sd(age)

age.z <- scale(age)
educ.z <- scale(Education)
ps.z <- scale(PS)
```

Table: Data from the Longitudinal Aging Study Amsterdam (z-stand.)

Person	1	2	3	4	5	6	7	8	9	10
x_1: Age	-1.19	-1.09	-0.98	-0.67	-0.46	0.48	0.59	0.79	0.90	1.64
x_2: Education[†]	-0.73	1.53	0.39	-0.73	0.96	-1.30	-0.17	-1.30	0.39	0.96
y: PS[‡]	0.10	1.28	0.88	-0.68	1.79	-0.42	-0.15	-1.07	-0.81	-0.94

[†]Number of school years [‡]Processing speed.

Standardizing $\hat{\boldsymbol{\beta}}$-Parameters

No need to estimate the intercept

```
lm(formula =
ps.z ~ -1 + age.z + educ.z)
coef.est coef.se
age.z  -0.63     0.18
educ.z  0.51     0.18
---
n = 10, k = 2
residual sd = 0.55, R-Squared = 0.73
```

```
lm(formula =
PS ~ -1 + age.c + Education)
coef.est coef.se
age.c      -0.51     0.15
Education  2.35     0.13
---
n = 10, k = 2
residual sd = 4.24, R-Squared = 0.98
```

- The z-standardized $\hat{\beta}$-parameters are denoted by the superscript z.
- The computation for the second approach uses the standard deviations: $\left({}^{SD_x}\!/_{SD_y}\right)\beta_x = \beta_x^z$

$$\hat{\boldsymbol{\beta}}^z = \frac{1}{s_y}\begin{bmatrix} \sqrt{s_{x_1}^2} & 0 \\ 0 & \sqrt{s_{x_2}^2} \end{bmatrix} \hat{\boldsymbol{\beta}} = \frac{1}{7.277}\begin{bmatrix} 9.024 & 0 \\ 0 & 1.6763 \end{bmatrix}\begin{bmatrix} -.51 \\ 2.21 \end{bmatrix} = \begin{bmatrix} -0.633 \\ 0.509 \end{bmatrix}.$$

Standardizing $\hat{\beta}$-Parameters

- Interpretation:
  - For one additional standard deviation in chronological age we predict a reduction of 0.633 standard deviations in PS (and vice versa for a reduction in age)
  - For one additional standard deviation in years of education we predict slower processing speed of .509 standard deviations.

Note: The standardized parameters are highly dependent on the standard deviation of the respective parameters in the sample. Comparing these parameters *across* different samples is typically not recommended!

Standardizing der $\hat{\boldsymbol{\beta}}$-Parameter

Some thoughts

- Standardized β-parameters are strongly affected by bounded data (which limit magnitude of variance, see Cohen & Cohen, 1983).
- Standardized regression coefficients mostly find application as effect sizes in regression analyses.
- Standardized regression coefficients can be useful to facilitate computations
    - Especially in Bayesian applications, computation can be greatly facilitated

Regression to the Mean

- If both x and y are standardized, then the regression intercept is zero and the slope is simply the correlation between x and y (for the univariate case)
- Slope of a regression of two standardized variables must always be between -1 and 1

▷ In general, the slope of a regression with one predictor is $\beta = \rho\sigma_y\sigma_x$, where ρ is the correlation between the two variables and σ_x and σ_y are standard deviations of x and y.

Regression to the Mean

```
lm(formula = y ~ x)
coef.est coef.se
(Intercept) 1.15     0.44
x           0.10     0.07
---
n = 30, k = 2
residual sd = 1.12, R-Squared = 0.07

lm(formula = scale(y) ~ scale(x))
coef.est coef.se
(Intercept) 0.00     0.18
scale(x)    0.26     0.18
---
> cor(x,y)
[1] 0.2616495
```

Regression to the Mean

- Hence, when x and y are standardized, regression slope is always less than 1.
- When x is 1 standard deviations above the mean, the predicted value of y is somewhere between 0 and 1 standard deviations above the mean.
- y is predicted to be closer to the mean (in standard-deviation units) than x

▷ Regression to the mean

Example: If a woman is 10 inches taller than the average for her sex, and the correlation of mothers' and (adult) sons' heights is 0.5, then her son's predicted height is 5 inches taller than the average for men. He is expected to be taller than average, but not so much taller–thus a "regression" (in the nonstatistical sense) to the average. (cf. Gelman & Hill, 2006)

Nonlinear Transformations

- Assumption: Linearity and additivity
- ▷ $\mathbf{y} = \mathbf{X}'\boldsymbol{\beta}$
- Sometimes these are not reasonable assumptions
- Scale is not maintained
- Parameters will obtain different values
  - $\boldsymbol{\beta}$ parameters change
  - Consequently, interpretation will be affected

Transformations of x and/or y

- Transformations may be made on
  - Predictor x
  - Dependent variable y

- Goals of transformations
  - Simplify relationship
  - ▷ Typically linearize
  - Eliminate heteroscedasticity
  - ▷ Change variance structure
  - Normalize residuals
  - ▷ Make distribution of residuals “normal”

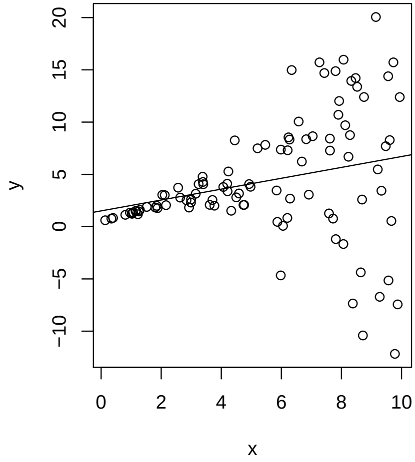
Rationale for Transforming

- Transformations should be taken only after careful consideration
- Nonlinearity may arise from outliers
- Diagnostic plots
  - Residual vs. $\hat{y}$
  - Density plots

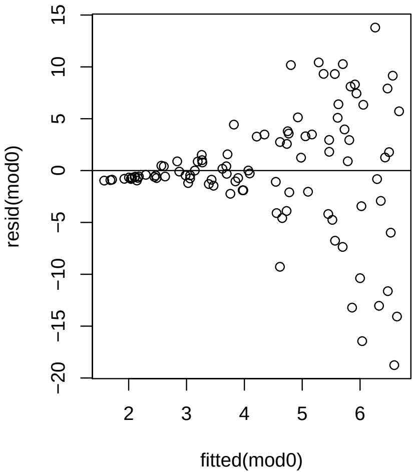
- Consequence of transformation may not always work
- Danger of fixing one problem but creating another
  - Possible introduction of outliers

Example: Heteroskedasticity

```
plot(x,y)
abline(coefficients(mod0))
```




```
plot(fitted(mod0), resid(mod0))
abline(h=0)
```



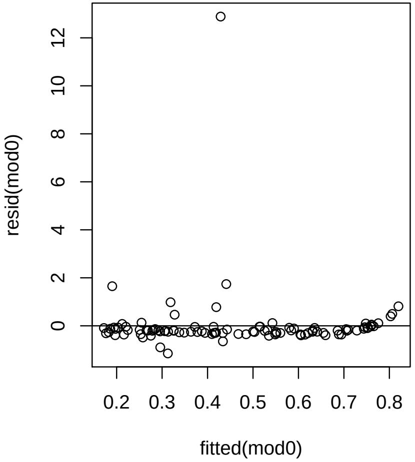
Example: Heteroskedasticity

- Fan spread
- "penalize" large values

e.g. $^1/_y$

```
mod0 <- (lm(1/y ~ x))

plot(fitted(mod0),
   resid(mod0))
abline(h=0)
```



Logarithms and Exponents

- Transformations often involve log and e
- $\log_m(x)$ logarithm of value x to the base m

e.g. $\log_{10}(1000) = 3$ since $10^3 = 1000$

- Natural logarithm $\ln_e(x)$

e.g $x = \exp(\ln_e(x)) = e^{ln_e(x)}$, hence, $\exp()$ is called the antilog of $\ln_e$

- The antilog returns the original number
  - Important for returning to original metric after transformation
- Some transformations have no solution (e.g. $\ln_e(-10)$)

Logarithmic transformations

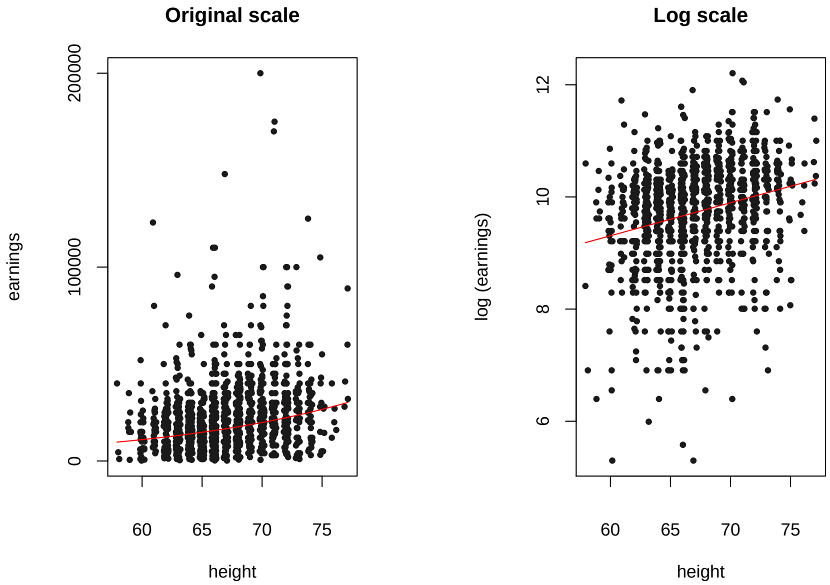
- A linear model on the logarithmic scale corresponds to a multiplicative model on the original scale

  $\log y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \epsilon_i$
- Exponentiating both sides yields

$$
\begin{aligned}
y_i &= \exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \epsilon_i) \\
&= B_0 \times B_1^{X_{i1}} \times B_2^{X_{i2}} \ldots E_i
\end{aligned}
$$

  - ▷ $B_0 = e^{\beta_0}, B_1 = e^{\beta_1}, B_2 = e^{\beta_2}, \ldots$ are exponentiated regression coefficients (all positive)
  - ▷ $E_i = e^{\epsilon_1}$ is the exponentiated error term
- On the scale of the original data y_i, the predictors $X_{i1}, X_{i2}, \ldots$ come in *multiplicatively*.

Example: Logarithmic regression



Example: Logarithmic regression

- Predicting earnings from height

```
log.earn <- log (earn) R code
earn.logmodel.1 <- lm (log.earn ~ height)

lm(formula = log.earn ~ height)
            coef.est coef.se
(Intercept)    5.74    0.45
height         0.06    0.01
n = 1192, k = 2
residual sd = 0.89, R-Squared = 0.06
```

- Estimated coefficient $\beta_1 = 0.06$ implies that a difference of 1 inch in height corresponds to an expected positive difference of 0.06 in log(earnings)

Example: Logarithmic regression

- Earnings are multiplied by exp(0.06)
- ▷ $\exp(0.06) \approx 1.06$
- Difference of 1 in the predictor corresponds to an expected positive difference of about 6% in the outcome variable
- Similarly, if β_1 were -0.06, then a positive difference of 1 inch of height would correspond to an expected negative difference of about 6% in earnings.

Natural log vs log-base-10

- ln, or natural log yields % change in y for one unit in x
- Coefficients on ln scale are directly interpretable as approximate proportional differences.
- The actual transformed value is hard to understand

$\log_{10}$

- The advantage of $\log_{10}$ is that the predicted values themselves are easier to interpret
- $\log_{10}(1,000) = 3$ or $\log_{10}(100,000) = 5$
- Coefficients are harder to interpret

```
lm(formula = log10.earn ~ height)
            coef.est coef.se
(Intercept)    2.493   0.197
height         0.026   0.003
n = 1187, k = 2
residual sd = 0.388, R-Squared = 0.06
```

Intrinsically Linear vs. Intrinsically Nonlinear

It's not always clear whether a function is linear or nonlinear.

- Intrinsically linear

- Elements can be transformed to make function linear
- How does the error ϵ enter the model?
- Multiplicative errors: Error increases with increasing y
  $y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} \exp(\epsilon)$ can be linearized with rules
  $\log(bX) = \log b + \log X$ and $\log(X^b) = b \log(X)$ to
  $\log y = \log \beta_0 + \beta_1 \log x_1 + \beta_2 \log x_2 + \epsilon$
- This model can now be solved via OLS

Intrinsically Linear vs. Intrinsically Nonlinear

- Intrinsically *non*linear

- Elements can *not* be transformed to make function linear
- How does the error ϵ enter the model?
- Additive errors: Error variance is constant
  $y = \beta_0 x_1^{\beta_1} x_2^{\beta_2} + \epsilon$
- Problem: $\log(a+b) = \log(a(1+b/a)) = \log a + \log(1+b/a)$
  $\log y = \log(\beta_0 x_1^{\beta_1} x_2^{\beta_2}(1 + \beta_0 x_1^{\beta_1} x_2^{\beta_2} \frac{1}{\epsilon}))$
- Error is now multiplied!
- Intrinsically nonlinear
- Nonlinear regression instead of OLS

Linearizing Based on Theory

Some models are used to linearize certain type of data

- Learning data tend to be nonlinear with a bend and an asymptote
- e.g. Michaelis Menten (diminishing returns model)
- Typical approach is to take data and linearize according to that model
- After linearizion, add error term, to keep it additive
- Implies multiplicative error terms in non linearized form

- There are a number of transformations based on models that can be linearized
- There are also a number of transformations ($\log$, e, ratio etc.) are typically used in certain fields and/or for certain data.

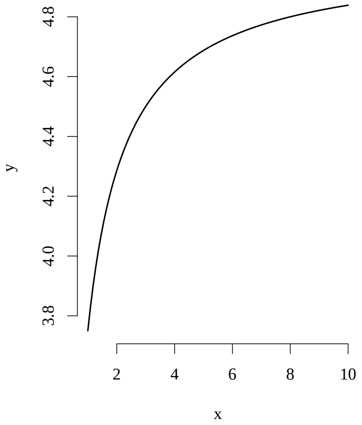
Linearizing Based on Theory

Asymptotic function (Michaelis-Menten or diminishing returns model):

$$y = \frac{ax}{1+bx}.$$

Linearize with reciprocal transformation:

$$\frac{1}{y} = \frac{1+bx}{ax}$$
$$\frac{1}{y} = \frac{1}{ax} + \frac{bx}{ax} = \frac{1}{ax} + \frac{b}{a}$$



With y=1/y; x=1/x; A=1/a: C=b/a we obtain the linear model:

$$y = AX + C$$

And now we can add error term: $y = AX + C + \epsilon$

Convenient Linearization Methods

Some methods to linearize are not founded in theory but are convenient

- logarithmic transformations
- e.g. $\mathbf{y} = [1, 10, 100, 1000, 10,000]'$ can be transformed to $\log_{10}(\mathbf{y}) = [0, 1, 2, 3, 4]'$
- Reciprocal transformations
  - Common in reaction time data and interpreted as rates
  - Reduces variance due to large numbers

Empirical Transformation Methods

Some methods are completely data driven

- Goal is to linearize and normalize data
- Typically
  - Attempt to undo all unfavorable conditions in data
  - Power transformations: $y* = y^{\lambda}$
  - ▷ Linearizing one-bend data
  - Logit and Probit transformations
  - ▷ Linearizing two-bend data
- Target may be x, y or both variables
  - Problem: Data may be bent, but homoscedastic
  - Transformation would reduce bend but introduce heteroscedasticity
  - ▷ Error variance is affected by transformations of y
  - ▷ Nonlinearity is affected by transformations of x

Empirical Transformation Methods

- How to choose λ
  - Iterative methods
  - Most common: Box-Cox transformation (in R: `boxcox()`)
- Considered to be "old school"
- Modern approach is to use generalize linear model

Nonlinear Regression

Some models can not be linearized

- Intrinsically nonlinear
- Arises with additive errors

e.g. $y = c^{(\exp(\beta x))} + \epsilon$

- OLS will not work as there is not an analytic solution to our minimization problem.
- Iterative methods using maximum likelihood
    - Starting with parameter configurations and approaching best solution step by step

Summary

- Power polynomials are very flexible and can fit almost any shape – given sufficient polynomials
  - We typically prefer interpretability over best fit and cubic terms may be the highest order polynomial we see in psychological research
- Nonlinear transformations
  - Change relative spacing on scores of a scale to either
    - address curvature and/or
    - heteroscedasticity and/or
    - non-normality of residual distribution
  - Transformations can be based on theoretical models, on convenience, or be empirically determined by the data.
- Nonlinear regression
  - Regression that is base on an intrinsically (or inherently) nonlinear model