MAT20306 - Advanced Statistics

Lecture 7: Multiple linear regression

WAGENINGEN UR
For quality of life

Biometris
Quantitative Methods brought to Life

Multiple Linear Regression: Inference and Modeling

1) Comparing models: Extra Sums-of-Squares principle
2) Test for several β 's simultaneously (full vs reduced model)
3) Estimation of mean response for given x-values, with CI; Prediction of y for given x-values, with a prediction interval PI
4) Collinearity
5) Modeling:
   1. Variable and model selection : several aspects
   2. Quadratic regression and Interaction
   3. Dummy variables

O&L Sections 12.4, 12.5, 12.6

Seasonal catch of bass, example 12.17 in O&L

y = seasonal catch of bass in a lake (per mile^2) – given in 1000 units
x_1 = number of lake shore residences (per mile^2 lake area)
x_2 = size of lake (mile^2)
x_3 = 1 for public access of lake and 0 otherwise (dummy variable)
x_4 = index for structures that offer shelter for bass.

Seasonal catch of bass, example 12.17 in O&L

y = seasonal catch of bass in a lake (per mile^2) – given in 1000 units
x_1 = number of lake shore residences (per mile^2 lake area)
x_2 = size of lake (mile^2)
x_3 = 1 for public access of lake and 0 otherwise (dummy variable)
x_4 = index for structures that offer shelter for bass.

A commission doubts the need for variables x_3 and x_4 $\rightarrow$ one construct a test for
H_0: $\beta_3 = \beta_4 = 0$ vs H_a: at least one of β_3, $\beta_4 \neq 0$.

Fit the Full Model and Reduced Model and compare the two residual sums of squares (output on p693, 694 O&L):

1. Full Model with all 4 variables: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \varepsilon$
2. Reduced Model without x_3 and x_4 : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$

Extra sums of squares

- If a regressor x enters a regression model, the *SSE* will *decrease* and the *SSR* will *increase* with the same amount.
- Increase in *SSR* = decrease in *SSE* = *extra sum of squares* due to entering x into a given model.

- Starting model with *residences & size* only:

$$SSE_1 = 23.425$$

- Now add *structure & access*:

$$SSE_2 = 2.276$$

- Extra SS of *structure & access* (after *residences & size*) = 23.425 – 2.276=**21.149**

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.913	2	1.456	1.057	.369[a]
	Residual	23.425	17	1.378		
	Total	26.338	19			

a. Predictors: (Constant), Size, Residence
b. Dependent Variable: Catch

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.062	4	6.016	39.652	.000[a]
	Residual	2.276	15	.152		
	Total	26.338	19			

a. Predictors: (Constant), Structure, Access, Residence, Size
b. Dependent Variable: Catch

- Generally, the extra sum of squares depends on the order of model terms, e.g. SS of x_1 first and x_2 after x_1 is generally not the same as SS for x_2 and SS for x_1 after x_2.

2. F–test for subset of β's

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	24.062	4	6.016	39.652	.000[a]
	Residual	2.276	15	.152		
	Total	26.338	19			

a. Predictors: (Constant), Structure, Access, Residence, Size
b. Dependent Variable: Catch

— ANOVA table F-test looks at all regressors together. How many in this case?

Remember we want to test: H_0: $\beta_3 = \beta_4 = 0$ versus at least one β_3, or $\beta_4 \neq 0$,

For a regression model with $k = 4$ explanatory variables, we compare two models:

1 Full Model (FM) with x_1, x_2, x_3 and x_4 → $\mathrm{SSE}_{\mathrm{FM}}$

2 Reduced Model (RM) with x_1, x_2 only (model under H_0) → $\mathrm{SSE}_{\mathrm{RM}}$

Test statistic: $$F = \frac{\Delta SSE / \Delta dfE}{MSE_{FM}} = \frac{(\mathrm{SSE}_{\mathrm{RM}} - \mathrm{SSE}_{\mathrm{FM}})/(dfE_{RM} - dfE_{FM})}{\mathrm{MSE}_{\mathrm{FM}}}$$

Under H_0: $F \sim \mathrm{F}(\mathrm{df}_1, \mathrm{df}_2)$ with $df_1 = \Delta dfE$, and, $df_2 = dfE_{FM}$

Number of bass: full vs reduced model

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	2.913	2	1.456	1.057	.369[a]
	Residual	23.425	17	1.378		
	Total	26.338	19			

a. Predictors: (Constant), Size, Residence
b. Dependent Variable: Catch

<u>Full model</u>:
$SSE_{FM} = 2.276$
$dfE_{FM} = 15$

<u>Reduced model</u>:
$SSE_{RM} = 23.425$
$dfE_{RM} = 17$

1) H_0: $\beta_3 = \beta_4 = 0$ versus at least one β_3, or $\beta_4 \neq 0$

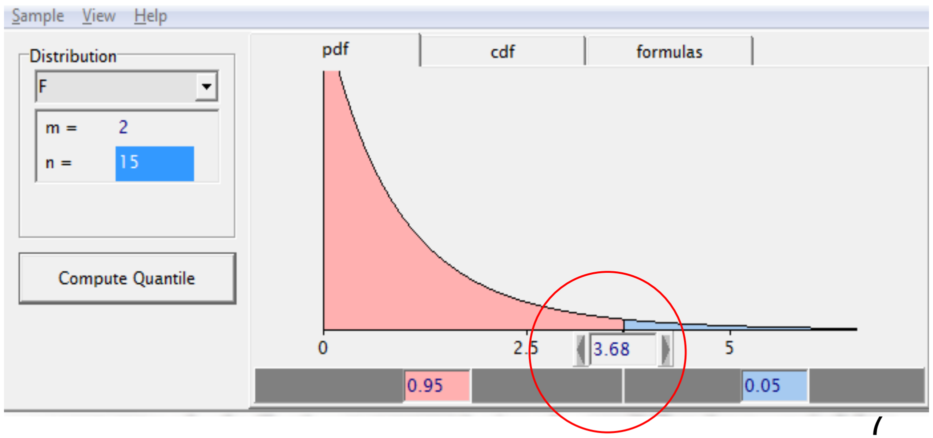
2/3) Test statistic: $F = \frac{\Delta SSE/\Delta dfE}{MSE\_FM}$; under H_0, F ~ F(2, 15)

4/5) RR: F > F(2,15, 0.05) = 3.68

6) F = $\frac{(23.425-2.276)/(17-15)}{0.152} = 69.7$

7) 69.7 > 3.68, so

8) H_0 is rejected, H_a is proven.
We cannot omit the two variables.



3. Prediction using multiple linear regression

- Similar to simple regression, we can be interested in the mean (expected) response μ_y at specific values x_1*, x_2*, …, x_k* of the regressors:

$$\mu_y = \beta_0 + \beta_1 x_1^* + \cdots + \beta_k x_k^*$$

- The estimated mean response at x_1*, x_2*, …, x_k* is obtained by replacing the β's by their LSE:

$$\hat{\mu}_y = b_0 + b_1 x_1^* + \cdots + b_k x_k^*$$

- We will read the corresponding standard error (of the estimated mean response) se($\hat{\mu}_y$) from SPSS output.
- The (1- α) CI for mean response has limits: $\hat{\mu}_y \pm t_{dfE}(\alpha/2) \cdot se(\hat{\mu}_y)$

- This is a confidence interval for μ_y and not a prediction interval for y !

Prediction: catch

- Want to construct a 0.95 Confidence Interval for the expected catch of lakes where the number of residences is 55, with an area of 1.5 square miles, with public access, and with structure index equal to 52.

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	-2.784	.816		-3.413	.004
	Residence	.027	.009	.401	2.931	.010
	Size	.504	.221	.323	2.281	.038
	Access	.743	.202	.317	3.676	.002
	Structure	.051	.005	.867	11.258	.000

a. Dependent Variable: Catch

- $\hat{\mu}_y = -2.784 + 0.027x_1 + 0.504x_2 + 0.743x_3 + 0.051x_4$
- $\hat{\mu}_y = -2.784 + 0.027 \times 55 + 0.504 \times 1.5 + 0.743 \times 1 + 0.051 \times 52 = 2.8$
- $CI(\hat{\mu}_y) = \hat{\mu}_y \pm t_{15}(0.025) \times se\left(\hat{\mu}_y\right) = 2.8 \pm 2.131 \times 0.12 = (2.58, 3.1)$

From SPSS

Prediction: catch in SPSS

- 0.95 confidence interval for the expected catch when the number of residences is 55, a lake of 1.5 square miles, with public access, and with a structure index equal to 52.

Lake	Catch	Residence	Size	Access	Structure	PRE_1	RES_1	SEP_1	LMCI_1	UMCI_1
1	3.6000	92.2000	.2100	0	81	3.93365	-.33365	.20164	3.50387	4.36343
2	.8000	86.7000	.3000	0	26	1.01949	-.21949	.20046	.59222	1.44676
3	2.5000	80.2000	.3100	0	52	2.17972	.32028	.14103	1.87911	2.48033
4	2.9000	87.2000	.4000	0	64	3.02615	-.12615	.16187	2.68114	3.37116
5	1.4000	64.9000	.4400	0	40	1.22167	.17833	.18943	.81791	1.62543
6	.9000	90.1000	.5600	0	22	1.03699	-.13699	.24511	.51455	1.55942
7	3.2000	60.7000	.7800	0	80	3.32550	-.12550	.22167	2.85302	3.79798
8	2.7000	50.9000	1.2100	0	60	2.25683	.44317	.20810	1.81328	2.70039
9	2.2000	86.1000	.3400	1	30	1.97100	.22900	.20710	1.52959	2.41242
10	5.9000	90.0000	.4000	1	90	5.17347	.72653	.24440	4.65254	5.69439
11	3.3000	80.4000	.5200	1	74	4.15859	-.85859	.17839	3.77836	4.53882
12	2.9000	75.0000	.6600	1	50	2.85729	.04271	.13777	2.56364	3.15095
13	3.6000	70.0000	.7800	1	61	3.34616	.25384	.13620	3.05585	3.63648
14	2.4000	64.6000	.9100	1	40	2.19321	.20679	.13604	1.90325	2.48317
15	.9000	50.0000	1.1000	1	22	.97736	-.07736	.21346	.52238	1.43233
16	2.0000	50.0000	1.2400	1	50	2.47947	-.47947	.16150	2.13523	2.82370
17	1.9000	51.2000	1.4700	1	37	1.96275	-.06275	.14243	1.65917	2.26632
18	3.1000	40.1000	2.2100	1	61	3.26502	-.16502	.18723	2.86596	3.66409
19	2.6000	45.0000	2.4600	1	39	2.39735	.20265	.21022	1.94927	2.84543
20	3.4000	50.0000	2.8000	1	53	3.41833	-.01833	.28556	2.80967	4.02698
.	.	55.0000	1.5000	1	52	2.84661	.	.12094	2.58883	3.10438

$x_1^*, x_2^*, x_3^*, x_4^*$ $\quad \hat{\mu}_y \quad se(\hat{\mu}_y) \quad CI_{1-\alpha}(\mu_y)$

Prediction continued

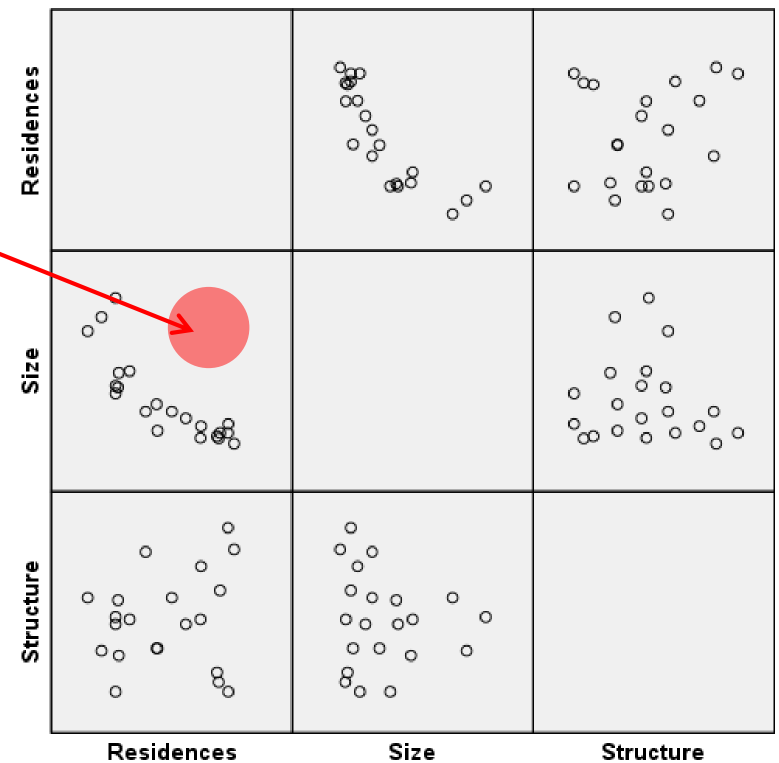
- Similar to simple regression, a prediction interval for y can be constructed as well.
- This interval contains all likely values for *y*, considering the estimated values for the β's and σ_ε and their se's.
- The (1–α) prediction interval limits: $\left(\hat{y} \pm t_{\alpha/2,dfE} \cdot se(\hat{y})\right)$
- [NB. se($\hat{y}$) (individual prediction) is $\sqrt{s_\varepsilon^2 + se(\hat{\mu}_y)^2}$ with $s_\varepsilon^2 = MSE$]
- This interval is wider than the (1–α)-confidence interval for μ_y.

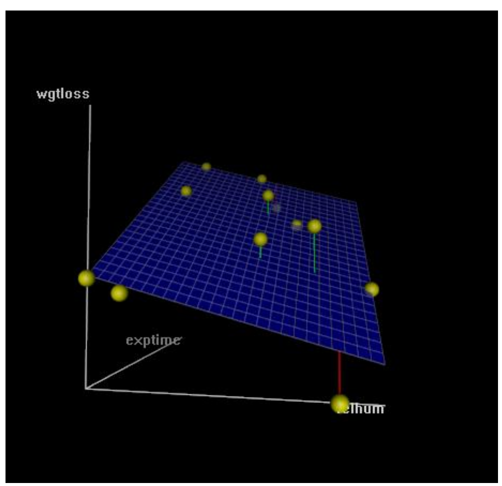
2.39735	.20265	.21022	1.94927	2.84543	1.45395	3.34075
3.41833	-.01833	.28556	2.80967	4.02698	2.38891	4.44774
2.84661	.	.12094	2.58883	3.10438	1.97731	3.71590

$CI_{1-\alpha}(\mu_y)$ $CI_{1-\alpha}(y)$

Extrapolation in regression

- Extrapolation is prediction of y for values of the explanatory variables that are outside the (multidimensional) experimental region.
- This is potentially hazardous, because often we cannot be sure that the model holds outside the experimental region.
- For example, combinations of large number of residences (x_1) and large lake sizes (x_2) are outside the experimental region (were not assessed).

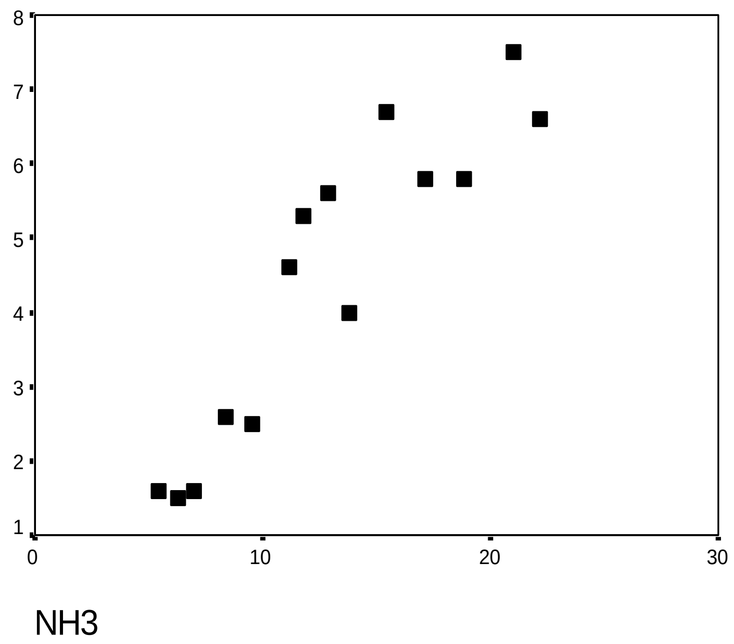


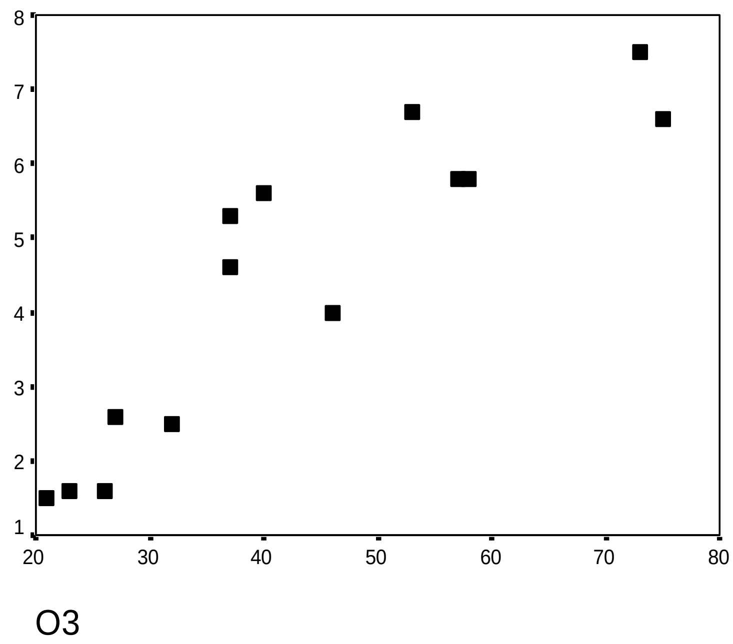


Multi (collinearity)

Plant damage example

Relationship of **damage** of plants (y) vs. NH_3 (ammonia) and O_3 (ozone) levels of surrounding air is investigated. Both NH_3 and O_3 are *observed*, not fixed.





Strong positive relationship of *damage* with NH_3 and O_3.

Plant damage continued

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	47.844	2	23.922	28.585	.000[a]
	Residual	9.206	11	.837		
	Total	57.049	13			

a. Predictors: (Constant), O3, NH3

b. Dependent Variable: DAMAGE

F–test for H_0: $\beta_1 = \beta_2 = 0$

very significant.

Coefficients[a]

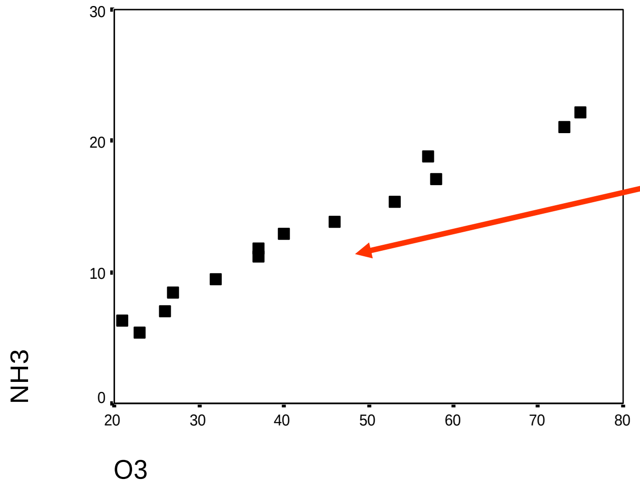
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-9.32E-02	.667		-.140	.891
	NH3	.490	.316	1.268	1.549	.150
	O3	-4.22E-02	.097	-.357	-.437	.671

a. Dependent Variable: DAMAGE

But, NH_3 and O_3 do not show a significant effect with separate t–tests: P-values are 0.15 and 0.67, and both are > 0.05.

How come?

Plant damage continued



Answer: Collinearity

that is: NH_3 and O_3 are strongly correlated.

Consequence: including NH_3 when O_3 is already in the model, does not improve the fit, and vice versa.

From graph: $NH_3 \approx 5 + 0.25 * O_3$

Multicollinearity

- (Severe) problems when there are high correlations among the explanatory variables:
  - problems with interpretation of the β's
  - and even numerical problems.

- Some x-variables may be (nearly) replaced by (linear) combinations of other x-variables: different sets of values for β's show nearly the same fit (almost the same SSE).

- Indicators of the problem are:
  - high variance inflation factors (VIF s) or low tolerances (TOL).
  - Possibly : high standard errors for (some) $\hat{\beta}$'s

Variance inflation factor

- From O&L:

$$se(\hat{\beta}_j) = s_\varepsilon \sqrt{\frac{1}{\sum_i (x_{ij} - \bar{x}_j)^2 (1 - R_j^2)}} = s_\varepsilon \sqrt{\frac{VIF_j}{\sum_i (x_{ij} - \bar{x}_j)^2}}$$

  where R^2_j is the proportion of **variation in x_j** "explained" by the other x-variables.

- So, a large VIF_j leads to a large standard error for $\hat{\beta}j$.
- The higher R^2_j the more variable x_j is related to (some of) the other x-variables.
- Variance Inflation Factor (VIF): $VIF_j = 1/(1 - R_j^2)$

$$R_j^2 = 0 \Rightarrow VIF_j = 1$$

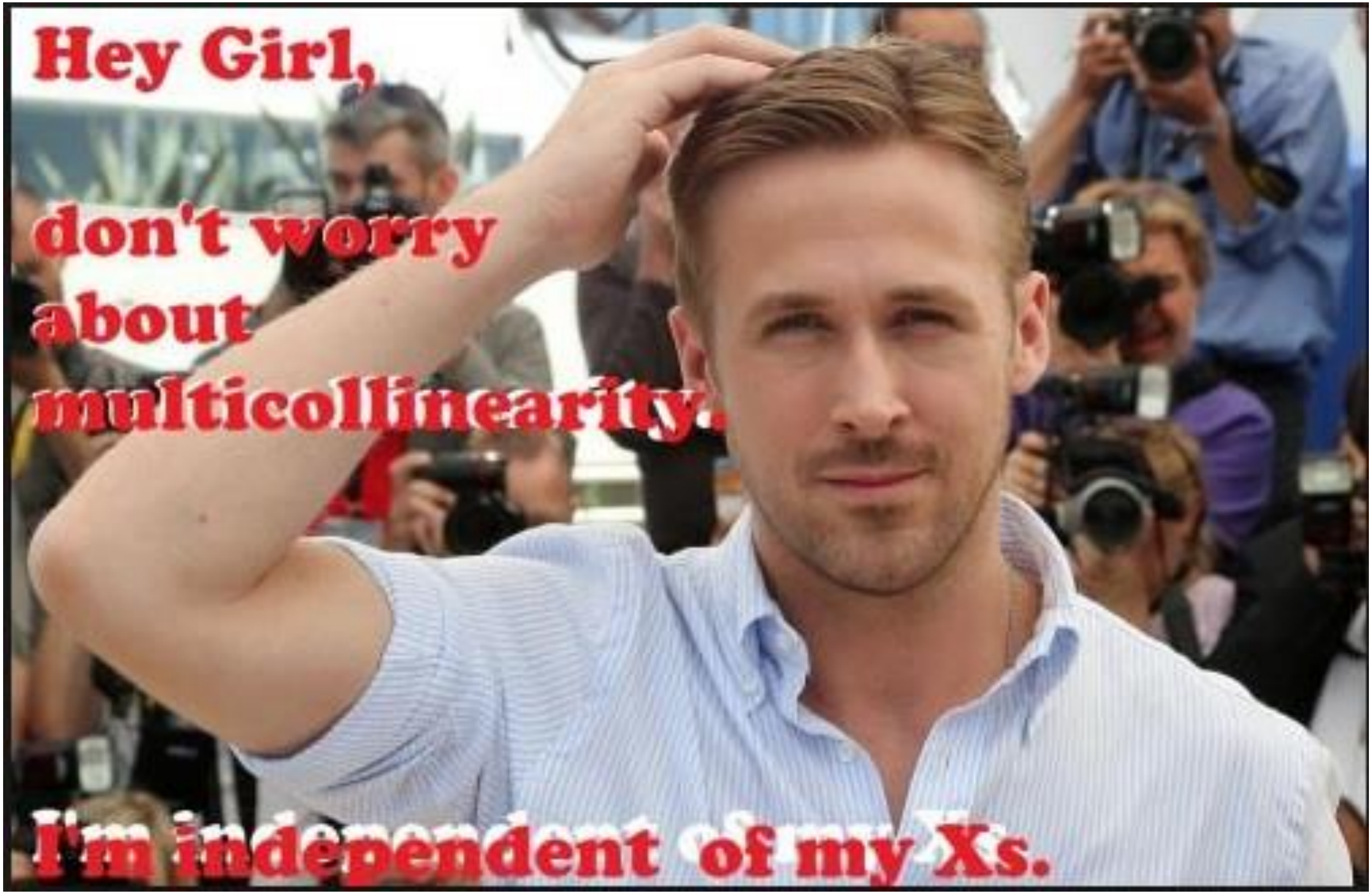
$$R_j^2 = 1 \Rightarrow VIF_j = \infty$$

- Tolerance (TOL): $TOL = 1/VIF$
- We are worried when a $VIF_j > 10$, or $TOL_j < 0.1$.
- In a designed experiment, collinearity problems can be avoided by a proper choice of the values of the x's by the researcher.

Remedy for collinearity, before and after analysis

- No “cure”, but some precautions you can take beforehand:
  - try to make a judicious choice of x-variables beforehand
  - do not put variables in the model that can be expected to be strongly related among each other: choose one of them
  - inspect correlations between x-variables
  - be careful with an observational study: x-variables may be strongly related in your sample, but not in the underlying population of interest
- After you have fitted a model:
  - Inspect standard errors, VIFs or TOLs
  - fit models with subsets of the x-variables as well and see what happens with β coefficients, their standard errors, significance of F– and t–tests.

Time to smile ☺



Topics @ Advanced Level

Multiple linear regression

GENERAL LINEAR MODEL

Confidence intervals

Hypothesis testing

Writing paragraphs

Advancing in the grammar of the language

Modeling 1: use of dummy variables

- Qualitative explanatory variables (e.g. treatment factors with *t* levels), can be represented by dummy variables.
- A dummy variable (or indicator variable) takes values 0 or 1, indicating absence or presence of e.g. a treatment.
- Consider *2* groups, A and B in which response y has expected values (means) μ_A and μ_B . To test equality: use 2-sample t-test.

Modeling 1: use of dummy variables

- Qualitative explanatory variables (e.g. treatment factors with t levels), can be represented by dummy variables.
- A dummy variable (or indicator variable) takes values 0 or 1, indicating absence or presence of e.g. a treatment.
- Consider *2* groups, A and B in which response y has expected values (means) μ_A and μ_B . To test equality: use 2-sample t-test.
- Define $x_A = 1$ for units in group A, and $x_A=0$ for units in group B, then: $\mu_y = \beta_0+\beta_1 x_A \rightarrow$ Regression model

  For units in group A: $\mu_y = \beta_0+\beta_1= \mu_A$.

  For group B: $\mu_y = \beta_0 \quad = \mu_B \rightarrow$

  H_0: $\mu_A = \mu_B$ is equivalent to testing H_0: $\beta_1=0$!

- In general, $(t-1)$ dummies are needed with their coefficients to have a model for mean response for t treatments.

Example, model with four treatments

- Example with treatments: 1 … 4 with means $\mu_1 \dots \mu_4$.
- E.g. 4 diets, each diet applied to 2 people, y = weight loss, compare the 4 population means of the diets

- Regression model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$

- This is multiple linear regression model with 3 regressors, with values:
  - if treatment 1 is used: $x_1 = 1, x_2 = x_3 = 0$
  - if treatment 2 is used: $x_2 = 1, x_1 = x_3 = 0$
  - if treatment 3 is used: $x_3 = 1, x_1 = x_2 = 0$
  - if treatment 4 is used: $x_1 = x_2 = x_3 = 0$ Treatment 4 is the ***reference***
- What do $\beta_0 = \mu_4$, β_1, β_2 and β_2 represent?

Treatment			
1	2	3	4
$\mu_1 = \beta_0 + \beta_1$	$\mu_2 = \beta_0 + \beta_2$	$\mu_3 = \beta_0 + \beta_3$	$\mu_4 = \beta_0$

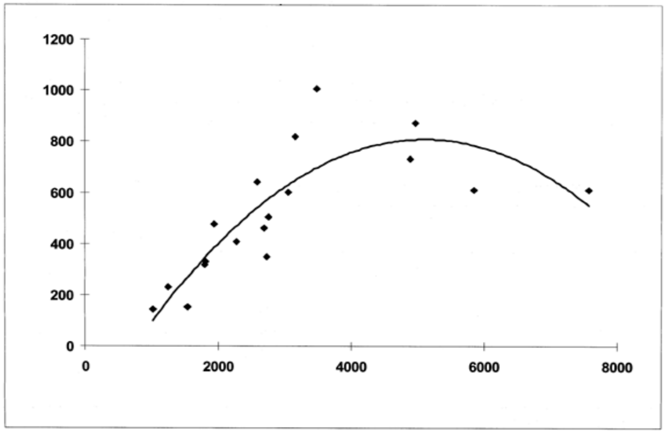
Modeling 2: Quadratic regression

- A quadratic regression model looks like this:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \varepsilon$$

- Now rename regressor x as x_1, and x^2 as x_2 . The regression model becomes

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

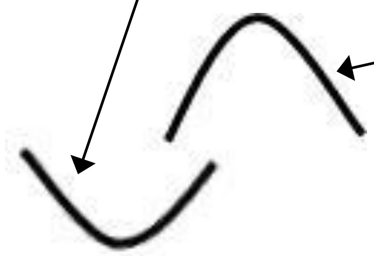


- Systematic part of model is quadratic function of x.
- Model is non-linear in variable x, but linear in coefficients β_1, β_2, β_3.
- So, it is a (general) linear model, and can be fitted with linear regression.

- Higher order terms (e.g. cubic x^3) can be added to model, result is higher order polynomial, can still be fitted by linear regression.

Quadratic regression, continued

- Interpretation of parameters:
  - β_0 is value of the curve where y-axis is cut by graph at x = 0 (intercept)
  - β_1 is the slope (or tangent) at that point
  - β_2 determines the amount of curvature and its sign indicates whether the graph is a valley (for positive β_2) or a hill (for negative β_2)



Modeling 3: Interactions

- Statistical interaction between two regressors: the **effect of one regressor on the response** depends on the level of other regressor.

- Can, in the simplest case, be modeled (and thus tested) with cross-product terms . For example with two predictors:

  $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$ where $x_3 = x_1 \times x_2$

- Two regression lines in one model:

- if $x_2 = 0 \Rightarrow \mu = \beta_0 + \beta_1 x_1$, so slope for x_1 is β_1

  if $x_2 = 1 \Rightarrow \mu = (\beta_0 + \beta_2) + (\beta_1 + \beta_3)\, x_1$, so slope is $\beta_1 + \beta_3$

There may be more ways to model interaction. This is the only one we present.

Anxiety of rats: different slopes

- We allow for a different effect of dose on anxiety for drugs A and B: the slopes of the regression lines may be different.

- Add product $x_1 x_2$ as a third variable ($x_3 = x_1 * x_2$) to model:

$$\mu_y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

- for drug A: $x_2 = 0$, $x_3 = 0$, so $\mu_y = \beta_0 \quad + \beta_1 x_1$

  for drug B: $x_2 = 1$, $x_3 = x_1$, so $\mu_y = (\beta_0 + \beta_2) + (\beta_1 + \beta_3) x_1$

- intercept for A: β_0

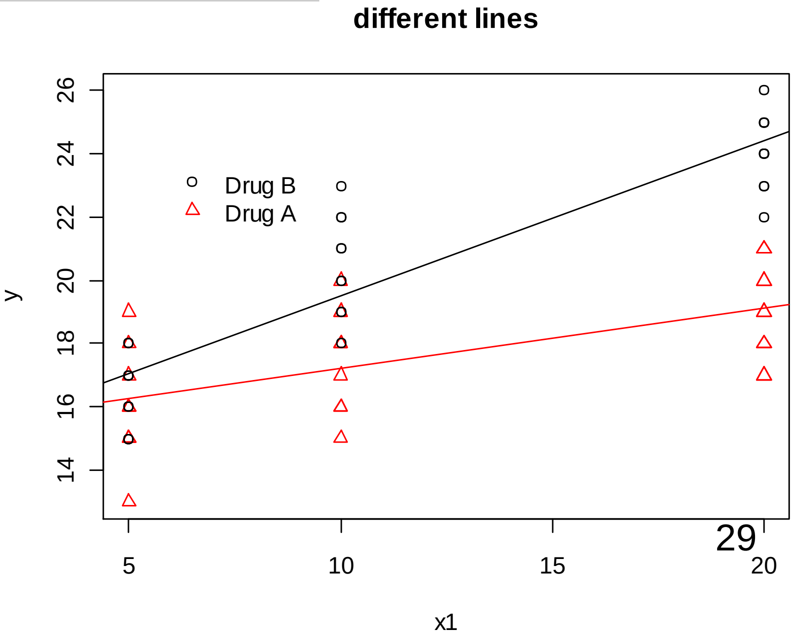
  intercept for B: $\beta_0 + \beta_2$

- slope for A: β_1

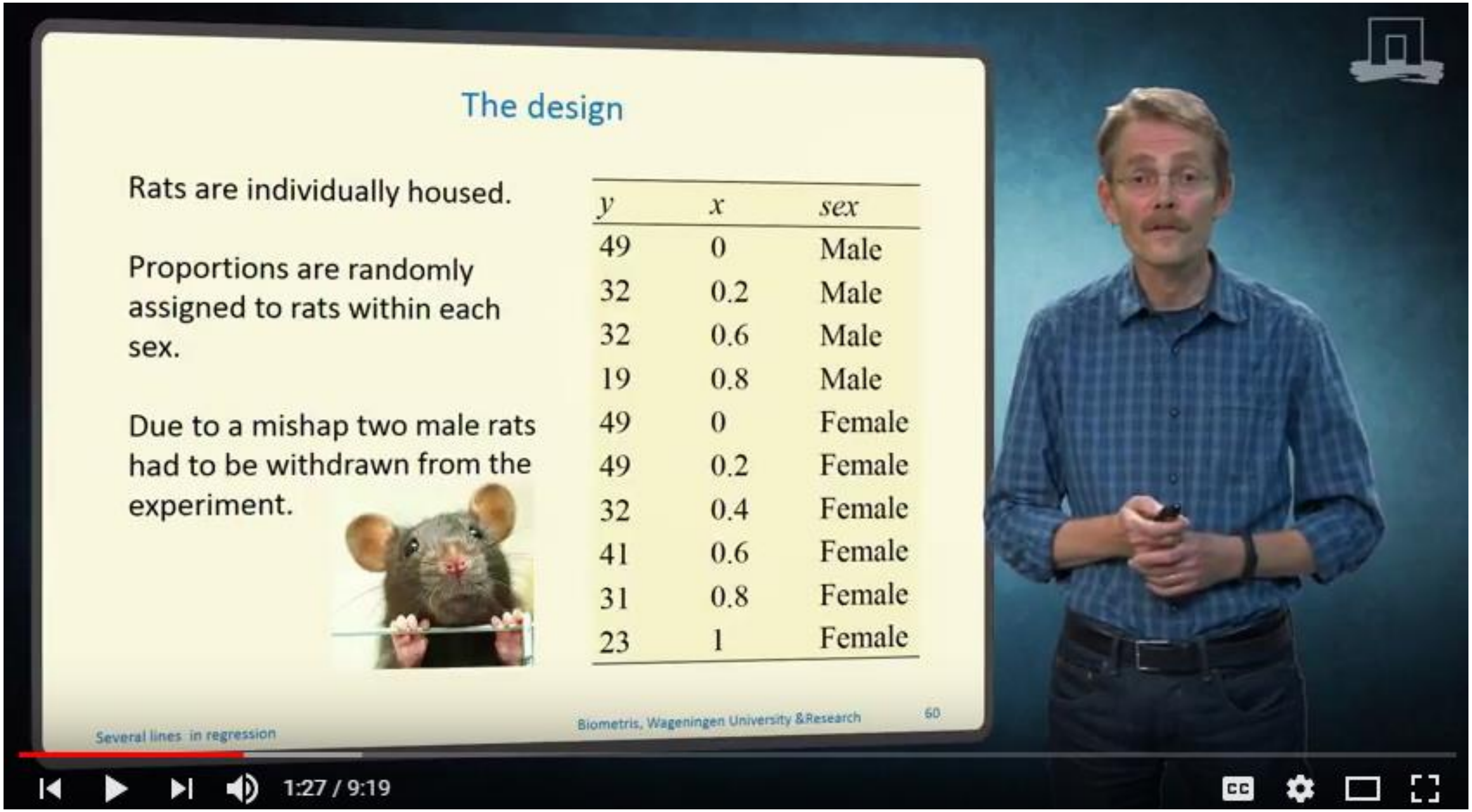
  slope for B: $\beta_1 + \beta_3$

drug A is the reference

- β_3: difference in slope between B and A
- β_2 : difference in intercept between B and A



Modeling 3: Two lines in one multiple regression



More about modeling

- Quadratic or cubic terms, interactions, such as product terms, may improve the fit of the model
- Transformation of y may improve the fit. It may either improve upon the assumptions for the error terms ε, or on the structure of μ.
- The log transformation changes the model from multiplicative to additive. Most other transformations make interpretation (more) difficult.
- Sometimes a transformation of y helps to reduce differences between variances, but at the same time violates the normality assumption. In that case a more advanced class of models from the generalized linear models (not part of this course - - - MSLS), may be more appropriate.