Violation of Assumptions

Philippe Rast

PSC 204B:
General Linear Model
January, 23

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

Statistical Inference

- So far we've been discussing model structures
- Descriptive statistics
- Typically we want to make assertions about population
- Brief refresher on statistical inference
- Many "violations" are only relevant for statistical inference

Statistical Inference

- Inferential statistics deals with precision of estimates in with respect to the population (inferring from sample to population)
- What can we conclude from sample estimates to population?
- In our example: Do age and reaction time relate negatively to each other in the population – not only in the sample?
- How precise is our estimate of β_0 and especially of β_1?
- We can't reach perfect precision – but we can try not to exceed a certain boundary, a certain probability of error (in other words "level of significance")

Statistical Validation: $\hat{\beta}_0$ and $\hat{\beta}_1$ as Stochastic Variables

- $\hat{\beta}_0$ and $\hat{\beta}_1$ are (as estimates of the population parameters β_0 and β_1) to be seen as stochastic variables.
- Explanation:

Formal:
1. Given Equation $\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ we see that $\hat{\boldsymbol{\beta}}$ depends directly from $\mathbf{y}$
2. y_i is defined as stochastic variable.
3. Hence, $\hat{\beta}_0$ and $\hat{\beta}_1$ are random variables (as in stochastic).

Descriptive:
1. For each randomly drawn sample we would obtain slightly different estimates for $\hat{\beta}_0$ and $\hat{\beta}_1$.
2. In other words: The estimates of $\hat{\beta}_0$ and $\hat{\beta}_1$ are at least partly due to a random (stochastic) component.
3. The random proportion gets smaller with increasing sample size.

Statistical Validation: Expected value of $\hat{\beta}_1$

- The expected value of $\hat{\beta}_1$ is β_1, i.e.

$$\mathsf{E}(\hat{\beta}_1) = \beta_1. \tag{1}$$

- Meaning:
  - The probability that $\hat{\beta}_1$ and β_1 coincide is relatively the largest (among all other possibilities)
  - If we estimated $\hat{\beta}_1$ an infinite number of times (i.e., in an infinite number of samples of equal size), our average value would be β_1.
- This property of an estimator is referred to as unbiasedness – or more generally *bias*.

Statistical Validation: Variance of $\hat{\beta}_1$

- The variance of $\hat{\beta}_1$ corresponds to the ratio of error variance and sum of squares of x, i.e.

$$\mathsf{V}(\hat{\beta}_1) = \frac{\sigma_\epsilon^2}{\sum_{i=1}^{N}(x_i - \bar{x})^2}. \tag{2}$$

- Meaning:
  - The smaller the error variance (σ_ϵ^2, in the numerator of Equation 2), the smaller the variance of $\hat{\beta}_1$. I.e., the smaller the error, the higher the precision of the estimate of the slope parameter.
  - The larger the sum of squares of x (in the denominator of Equation 2), the smaller the variance of $\hat{\beta}_1$. I.e., if the predictor variable has a large variation our estimate of the slope parameter becomes more precise.
- The variance of an estimator defines its *efficiency*.

Statistical Validation: Variance of $\hat{\beta}_1$, Proof

$$
\begin{aligned}
\mathsf{V}(\hat{\beta}_1) &= \mathsf{E}(\hat{\beta}_1 - \mathsf{E}\{\hat{\beta}_1\})^2 \\
&= \mathsf{E}\left(\left\{ \beta_1 + \frac{\sum_{i=1}^N x_i^* \epsilon_i}{\sum_{i=1}^N [x_i^*]^2} \right\} - \beta_1 \right)^2 \\
&= \mathsf{E}\left(\frac{\sum_{i=1}^N [x_i^*]^2 \epsilon_i^2}{\left\{ \sum_{i=1}^N [x_i^*]^2 \right\}^2} \right) \\
&= \frac{\sum_{i=1}^N [x_i^*]^2 \mathsf{E}(\epsilon_i^2)}{\left\{ \sum_{i=1}^N [x_i^*]^2 \right\}^2} \\
&= \frac{\sigma_\epsilon^2}{\sum_{i=1}^N [x_i^*]^2} = \sigma_{\hat{\beta}_1}^2 . \qquad (3)
\end{aligned}
$$

Row 2 to 3: For a random variable $e = a_1 e_1 + \ldots a_n e_n = \Sigma_{i=1}^n a_i e_i$ it holds $\mathsf{V}(e) = \Sigma_{i=1}^n a_i^2 \mathsf{V}(e_i) + \Sigma_{i,j=1}^n a_i a_j \mathsf{C}(e_i, e_j)$. Errors are defined as uncorrelated and covariance term $\Sigma_{i,j=1}^n a_i a_j \mathsf{C}(e_i, e_j)$ drops!

Statistical Validation: Distribution of $\hat{\beta}_1$

- $\hat{\beta}_1$ is nothing more than a linear combination of each single y_i
- Hence, $\hat{\beta}_1$ is normally distributed!
- More precisely:

$$\hat{\beta}_1 \sim N(\mathsf{E}\{\hat{\beta}_1\}, \mathsf{V}\{\hat{\beta}_1\}) \tag{4}$$

$$\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma_\epsilon^2}{\sum_{i=1}^{N}[x_i^*]^2}\right). \tag{5}$$

Statistical Validation: Hypothesis Test of $\hat{\beta}_1$

- An easy way to test hypotheses for $\hat{\beta}_1$ would be to construct a z-test:

$$\begin{aligned} z &= \frac{\hat{\beta}_1 - \beta_1}{\sigma_{\hat{\beta}_1}} \\ &= \frac{(\hat{\beta}_1 - \beta_1)\sqrt{\sum_{i=1}^{N}[x_i^*]^2}}{\sigma_\epsilon} \sim N(0,1). \end{aligned} \tag{6}$$

- Such a test assumes that σ_ϵ^2 is known, which is never the case in the real world.
- In fact, we only have an estimate of σ_ϵ^2, namely $\hat{\sigma}_\epsilon^2$.
- $\hat{\sigma}_\epsilon^2 = \frac{\Sigma_{i=1}^{N}(y_i - \hat{y})^2}{N-p-1}$ with p variables

Statistical Validation: Hypothesis Test of $\hat{\beta}_1$

- We can now construct a t-test.

$$\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}_\epsilon / \sqrt{\sum_{i=1}^{N} [x_i^*]^2}} \sim t_{N-1} \tag{7}$$

- We can now test the statistical significance of $\hat{\beta}_1$ via the t-distribution.

Conclusion

- Looking closer at the equation for the t-value, we notice that it is similar to the equation for the z-value. (cf. Equation 6 and 7). *But*
- we replaced σ_ϵ^2 with the sample estimate $\hat{\sigma}_\epsilon^2$.
- The significance is not tested anymore via the normal- but via the t-distribution.
- With an increasing number of degrees of freedom (i.e., with increasing sample size) the t-distribution approximates the normal distribution. Beginning with about $N > 30$ the differences become minimal.
- This makes sense intuitively, as with increasing N the estimate of the error variance becomes more precise (the correction of N on $N-1$ and $N-2$ has hardly an effect at all).
- ▷ All of these things are based on the assumptions defined under the Gauss-Markov theorem!

Coding Schemes

1. Understand the t-test in the context of a simple linear regression model.
2. Understand dummy coding and how it can be used in real data.
3. Be aware of advantages and disadvantages of dummy coding.

t-Test as Simple Linear Regression Model

- The prerequisites for the multiple regression only stated that it was necessary for predictors to be non-stochastic variables.
- Hence, it possible to use non-continuous variables, that is, categorical features, as predictors.
- The requirements for the regression analysis are not bound to the scale of measurement of the independent variable.
- The dependent variable remains unaffected by the predictor scale and is still measured at the interval level.

Definition: Statistical methods where the DV is on an interval scale and IV's are categorical are typically viewed as analyses of variance.

t-Test as Simple Linear Regression Model

- The simplest situation is the t-test that compares the mean of the dependent variables from two groups.
- The following table shows the measurements of processing speed (PS) for women (w) and men (m).

Table: Data from the Longitudinal Aging Study Amsterdam

Person	1	2	3	4	5	6	7	8	9	10
z: Gender	w	w	w	w	w	m	m	m	m	m
y: PS†	25	19	21	16	17	34	31	38	23	18

†Processing Speed

- Descriptive Stats:
  $\hat{\mu}_{y_w} = 19.6$, $\hat{\sigma}^2_{y_w} = 12.80$,
  $\hat{\mu}_{y_m} = 28.8$, $\hat{\sigma}^2_{y_m} = 66.70$.

t-Test as Simple Linear Regression Model

- A classic t-test for Table 1:
- Standard deviation of the difference among w and m (in case $n = n_m = n_w$.):

$$\hat{\sigma}_{\hat{\mu}_{y_m} - \hat{\mu}_{y_w}} = \sqrt{\frac{\hat{\sigma}^2_{y_m} + \hat{\sigma}^2_{y_w}}{n}} = \sqrt{\frac{12.8 + 66.7}{5}} = \sqrt{\frac{79.5}{5}} = 3.98.$$

- Hence, t is

$$t = \frac{\hat{\mu}_{y_m} - \hat{\mu}_{y_w}}{\hat{\sigma}_{\hat{\mu}_{y_m} - \hat{\mu}_{y_w}}} = \frac{28.8 - 19.6}{3.98} = \frac{9.2}{3.98} = 2.31.$$

- The critical t-value for $\alpha = .05$ with 8 df and tested two-tailed $t_{crit.} = 2.306$. Given that calculated t-value is larger than the critical value, the difference in the mean value of 9.2 between men and women is statistically significant.

t-Test as Simple Linear Regression Model

- This result can also be obtained by via regression analysis.
- We define:

$$x_i = \begin{cases} 0 \text{ if } z_i = w \\ 1 \text{ if } z_i = m \end{cases} \tag{8}$$

- Furthermore we define (with $N = n_w + n_m = 2n$):

$$\mathbf{y} = \begin{bmatrix} 25 \\ 19 \\ \vdots \\ 23 \\ 18 \end{bmatrix} ; \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} ; \mathbf{X} = \begin{bmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_i \\ \vdots & \vdots \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 1 \\ 1 & 1 \end{bmatrix} ; \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_i \\ \vdots \\ \epsilon_N \end{bmatrix} .$$

t-Test as Simple Linear Regression Model

```
lm(formula = ps ~ gender)
coef.est coef.se
(Intercept) 19.60     2.82
gender       9.20     3.99
---
n = 10, k = 2
residual sd = 6.30, R-Squared = 0.40
```

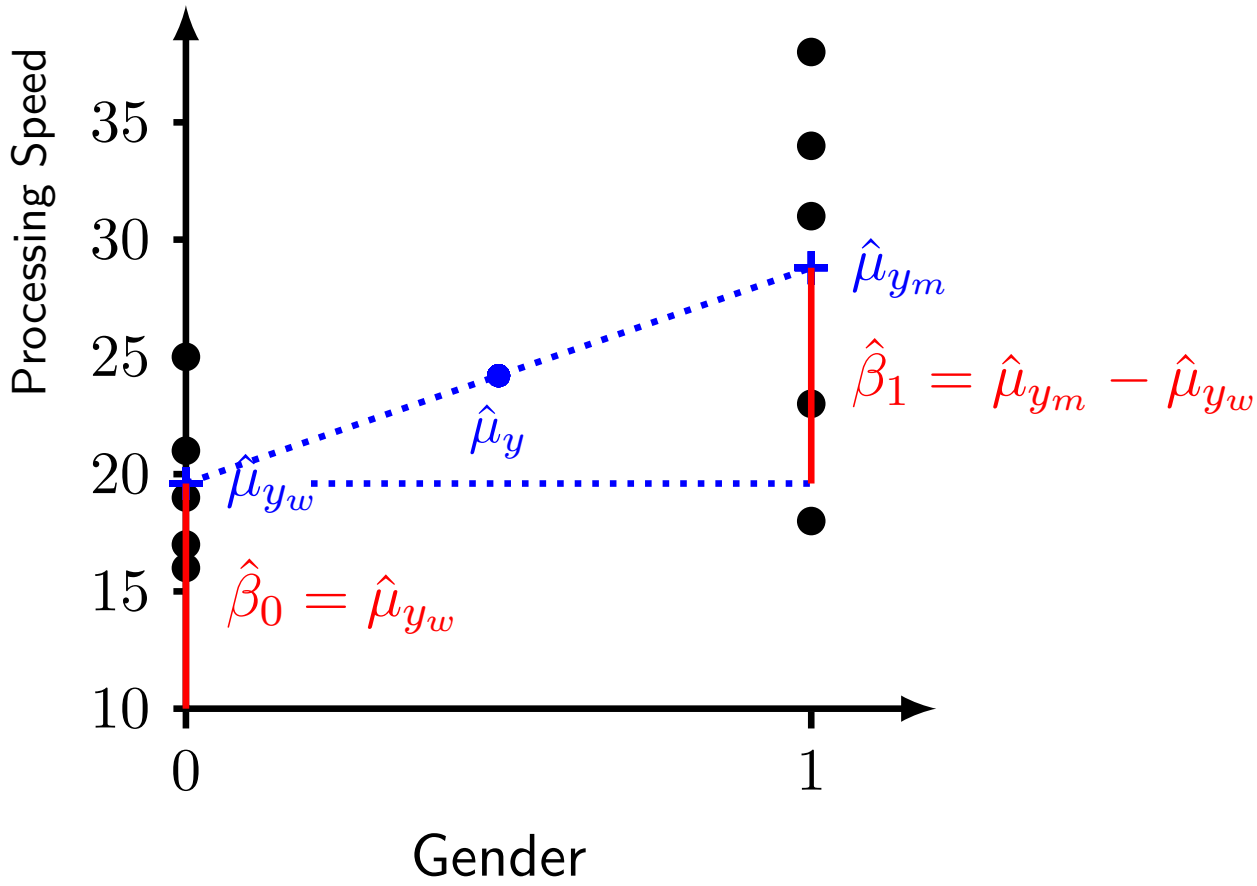
- β_0 19.6
- β_1 9.2
  - Interpretation: $\hat{\beta}_0$ corresponds to the mean of women and $\hat{\beta}_1$ is the difference of the mean of the men and the mean of the women (cf. slide 17).

t-Test as Regression Model: Dummy Coding

- Intuitive result because we know that
  - the mean value is the best estimate without other information.
  - the mean value minimizes the sum of squared errors.
  - the data of women and men are independent.
- Moreover it holds:
  - In a (simple) linear regression the quadratic error is minimized.
  - The regression line "cuts" through the mean value of the predictor variable and the dependent variable.

t-Test as Regression Model: Dummy Coding

- Graphical representation



t-Test as Regression Model: Dummy Coding

- Note that our interpretation of the β coefficients does not change, because
  - $\hat{\beta}_0$ specifies the predicted value of y if the predictor variable is 0. In this case: The mean value of women (19.6) minimizes the square error within the sample of women.
  - $\hat{\beta}_1$ indicates how many units the prediction of y changes as we go up one unit in the predictor variable. In this case: The difference between the mean value of men and that of women.
  - For the estimation of the regression line, it is irrelevant that the predictor variable (gender) has only two expressions. As required, the square error is minimized
  - However, we must remember that in this case a prediction beyond the actual values of the predictor variable makes no sense!

t-Test as Regression Model: Dummy Coding

With Equal Variances

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   19.600      2.820   6.951 0.000118 ***
gender         9.200      3.987   2.307 0.049905 *

Two Sample t-test (Equal Variances Assumed)

data:  ps by gender
t = -2.3072, df = 8, p-value = 0.04991
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
-18.395146309  -0.004853691
sample estimates:
mean in group 0 mean in group 1
           19.6            28.8
```

t-Test as Regression Model: Dummy Coding

Unequal Group Sizes

```
gender <- c(0,0,0,1,1,1,1,1,1,1)
ps <- c(25,19,21,35,38,34,31,38,23,18)

Welch Two Sample t-test

data:  ps by gender
t = -2.7419, df = 7.9854, p-value = 0.02542
alternative hypothesis: true difference in means is not equal
95 percent confidence interval:
-17.185493  -1.481173
sample estimates:
mean in group 0 mean in group 1
       21.66667        31.00000
```

t-Test as Regression Model: Dummy Coding

Unequal Group Sizes

```
summary(gls(ps ~ gender,
                weights = varIdent(form = ~1 | gender)))
Variance function:
Structure: Different standard deviations per stratum
Formula: ~1 | gender
Parameter estimates:
0        1
1.000000 2.521337
Coefficients:
                Value Std.Error   t-value p-value
(Intercept) 21.666667  1.763834 12.283845  0.0000
gender       9.333333  3.404013  2.741862  0.0254
```

t-Test as Regression Model: Dummy Coding

Individual Intercepts

```
gm <- gender
gf <- abs(gender-1)

> rbind(gf, gm)
[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]
gf    1    1    1    1    1    0    0    0    0     0
gm    0    0    0    0    0    1    1    1    1     1

lm(formula = ps ~ -1 + gf + gm)
   coef.est coef.se
gf 19.60     2.82
gm 28.80     2.82
---
n = 10, k = 2
residual sd = 6.30, R-Squared = 0.95
```

Dummy Coding: One-way Analysis of Variance

One-way Analysis of Variance

- One factor analysis of variance is for one factor of interest.
- In the example, we have assigned the numbers 0 and 1 to the variable "gender" according to the rule defined in Equation (8).
- This kind of coding of a categorial variable is called *Dummy Coding*. The variable x is referred to as *dummy variable*.
- It can also be applied to variables that have more than two levels. However, several dummy variables are then required.
- For example, the variable z (nationality) has three values, namely "Italian", "French", and "British".
- We define the dummy variables x_1 and x_2:

$$x_{1i} = \begin{cases} 0 \text{ if } z_i = \text{italian or british} \\ 1 \text{ if } z_i = \text{french} \end{cases}$$

$$x_{2i} = \begin{cases} 0 \text{ if } z_i = \text{italian or french} \\ 1 \text{ if } z_i = \text{british} \end{cases}$$

Dummy Coding: More Than Two Groups

- In general, if a categorical variable can take c possible values, then $c-1$ dummy variables are needed.
- One of the levels acts as a reference category and is encoded with 0 on all dummy variables (e.g. "italian").
- Why is this? We know that $\hat{\beta}_0$ is the predicted value of y when all predictor variables are 0. Hence, $\hat{\beta}_0$ can be meaningfully interpreted as the mean for the reference group.
- As soon as we have more than one dummy variable and thus more than one predictor variable, we are in the realm of multiple regression (or, in other words, in the one-factor analysis of variance).
- In the one-way analysis of variance, the dummy variables (or predictor variables) may be dependent on each other.
  - If the levels are overlapping, then the dummy variables correlate among each other with $-\frac{1}{c-1}$ (with c = number of levels).

Dummy Coding: One-way Analysis of Variance

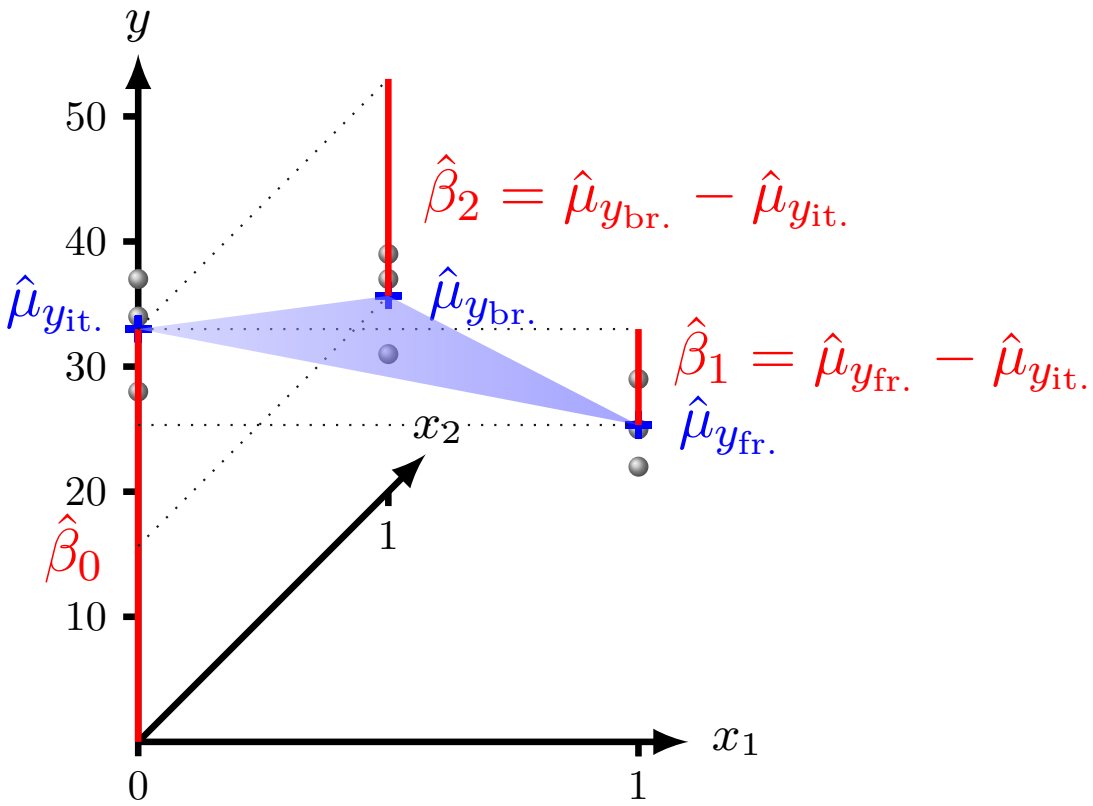
- Back to the example with nationalities. For the sake of simplicity, assume that $n = n_{\text{it.}} = n_{\text{fr.}} = n_{\text{br.}}$. Then $N = 3n$.
- We compute

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})\mathbf{X}'\mathbf{y} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{y_{\text{it.}}} \\ \hat{\mu}_{y_{\text{fr.}}} - \hat{\mu}_{y_{\text{it.}}} \\ \hat{\mu}_{y_{\text{br.}}} - \hat{\mu}_{y_{\text{it.}}} \end{bmatrix}$$

- $\hat{\beta}_0$ corresponds to the mean value for Italians, $\hat{\beta}_1$ is the difference between the mean value of the French and the mean value of the Italians, and $\hat{\beta}_2$ is the difference between the mean value of the British and the mean value of the Italians.
- The example shows how "italian" is the reference category: mean value differences are determined with regard to this category.
- What we do not know (or only indirectly) is how big the difference in mean between the French and the British is.

Dummy Coding: One-way Analysis of Variance

- Graphical representation



Dummy Coding: Two-Way Analysis of Variance

- Dummy coding does not need to be restricted to the t-test or the one-way ANOVA.
- Let's add a variable to the LASA data: Place of residence.

Table: Data from the Longitudinal Aging Study Amsterdam

Person	1	2	3	4	5	6	7	8	9	10	11	12
z_1: Gender	f	f	f	f	f	f	m	m	m	m	m	m
z_2: Residence†	C	C	C	R	R	R	C	C	C	R	R	R
y: PS‡	16	17	13	25	19	21	23	18	19	34	31	38

†R = Rural, C = City, ‡Processing Speed.

- Descriptive stats:
  $\hat{\mu}_{y_f} = 18.50$, $\hat{\sigma}^2_{y_f} = 17.50$, $\hat{\mu}_{y_m} = 27.16$, $\hat{\sigma}^2_{y_m} = 69.36$. $\hat{\mu}_{y_R} = 28.00$, $\hat{\sigma}^2_{y_R} = 56.80$, $\hat{\mu}_{y_C} = 17.66$, $\hat{\sigma}^2_{y_C} = 11.06$.

Dummy Coding: Two-Way Analysis of Variance

- For the dummy coding of the example we proceed according to the following rule:

$$
\begin{aligned}
x_{1i} &= \begin{cases} 0 \text{ if } z_{1i} = \text{female} \\ 1 \text{ if } z_{1i} = \text{male} \end{cases} \\
x_{2i} &= \begin{cases} 0 \text{ if } z_{2i} = \text{City} \\ 1 \text{ if } z_{2i} = \text{Rural} \end{cases} \\
&\text{and} \\
x_{3i} &= x_{1i} \times x_{2i}
\end{aligned}
$$

- The main effects are encoded as a multiplication
- The coding of the interaction in the dummy variable x_3 is the result of the multiplication of the dummy variables, which encode the main effects.

Dummy Coding: Two-Way Analysis of Variance

- $\mathbf{X}$ then is

$$\mathbf{X} = \begin{bmatrix} \mathbf{1} & \mathbf{x}_1 & \mathbf{x}_2 & \mathbf{x}_3 \end{bmatrix}$$

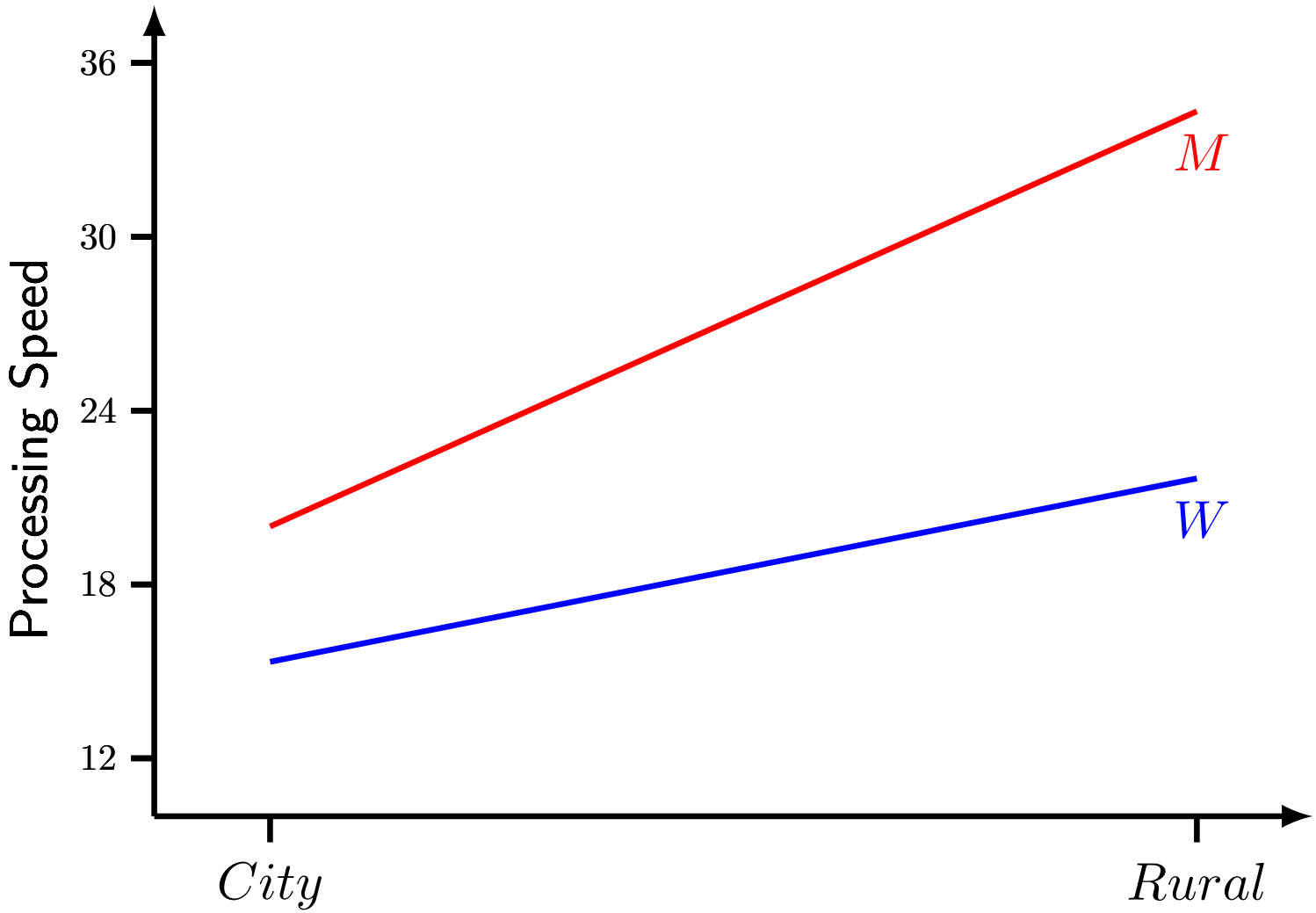
$$= \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{bmatrix}$$

- The coding of the interaction is shown here in red.
- Note that the interaction is not independent of the main effects

Dummy Coding: Two-Way Analysis of Variance

- $$\hat{\boldsymbol{\beta}} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = \begin{bmatrix} \hat{\mu}_{y_{\mathrm{Cf}}} \\ \hat{\mu}_{y_{\mathrm{Cm}}} - \hat{\mu}_{y_{\mathrm{Cf}}} \\ \hat{\mu}_{y_{\mathrm{Rf}}} - \hat{\mu}_{y_{\mathrm{Cf}}} \\ \hat{\mu}_{y_{\mathrm{Rm}}} - \hat{\mu}_{y_{\mathrm{Cm}}} - \hat{\mu}_{y_{\mathrm{Rf}}} - \hat{\mu}_{y_{\mathrm{Cf}}} \end{bmatrix} = \begin{bmatrix} 15.33 \\ 4.66 \\ 6.33 \\ 8.00 \end{bmatrix}$$
- This means,
  - $\hat{\beta}_0$ is the average for women who live in the city.
  - $\hat{\beta}_1$ Is the mean difference of the men in the city and the women in the city.
  - $\hat{\beta}_2$ Is the mean difference between women living in the countryside and women in the city.
  - $\hat{\beta}_3$ is the amount by which the mean difference, of rural men versus city-women, exceeds the simple effects of mean differences in $\hat{\beta}_1$ and $\hat{\beta}_2$ (= Interaction).

Interaction diagram



Dummy coding versus F-test

- Note: Up to now, we have not dealt with the F-test in the context of the regression analysis or the analysis of variance.
  For the following reasons:
  - The F-test, in conjunction with the regression analysis, checks the statistical significance of the overall model (omnibus test), that is, whether all predictor variables *together* contribute to the explanation (or prediction) of the dependent variable.
  - In general, however, we are interested in knowing which predictor contributes to the explanation.
  - Similar to the analysis of variance: A significant main effect simply means that at least two mean values differ among the levels of the factor. However, a F-test says nothing about which cell mean values are different. In addition, major effects can not be interpreted in the presence of significant interaction.
  - In order to check for mean values differences, so-called post-hoc tests (Scheffé, Tukey) must be carried out.

Dummy coding: Applications

- Alternative:
  - If you consider in advance which mean differences are to be tested for significance, neither F-tests nor post-hoc tests (in which everything is compared with everything else) are needed.
  - With appropriate coding mean value comparisons are very flexible obtained on the basis of regression analyses, especially if the cell sizes are unequal.
  - *Dummy coding is suitable whenever differences in mean values are of interest to a reference group (e.g., control group).*
  - Example: A series of new drugs is compared with a conventional one.
  - Disadvantage: We do not know whether there are any other differences in mean values other than the reference group.
  - For other questions other coding forms are possible.

Dummy coding: Summary

- For dummy coding, the reference category gets a zero on all dummy variables.
- The mean value of the effect level representing the reference group (in the example, women living in the city) is $\hat{\beta}_0$.
- The remaining regression coefficients inform about the *deviations of the group average values from the mean value of the reference group* or the deviation of the interaction from the addition of the corresponding main effects.
- In this respect, the results of a dummy-coded analysis deviate from a conventional ANOVA

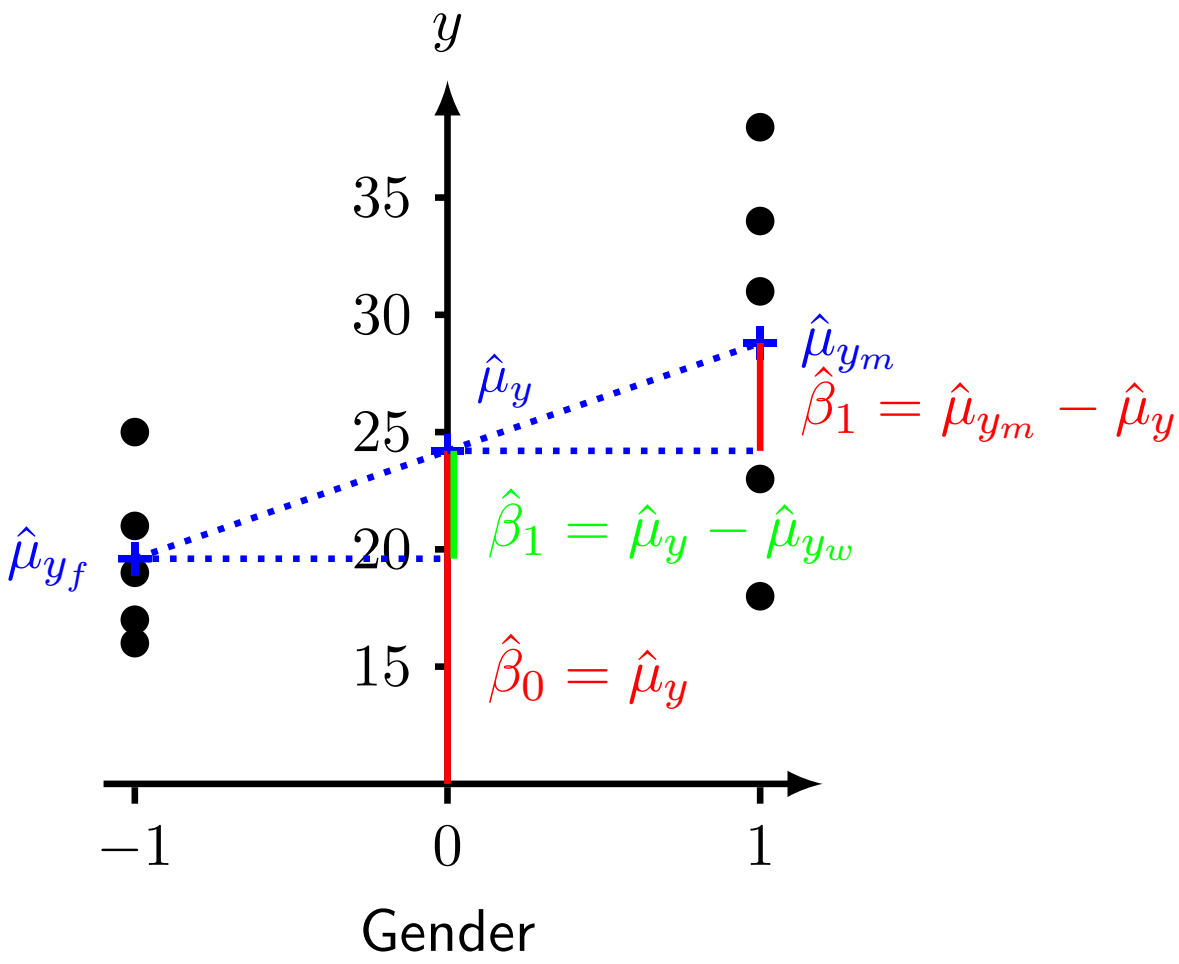
Effect Coding: Two Samples

- An alternative to dummy coding is the so-called *effect coding*.
- In contrast to dummy coding, the reference category is not assigned the value 0, but the value -1 on all coding variables (also called dummy variables).
- We define, on the basis of slide 16

$$x_i = \begin{cases} -1 \text{ if } z_i = w \\ 1 \text{ if } z_i = m \end{cases} \tag{9}$$

Effect Coding: Two Samples

- Graphical representation



Effect Coding: Two Samples

- That is, under effect coding, $\hat{\beta}_0$ the total mean of the sample, whereas $\hat{\beta}_1$ represents half of the difference between the mean of men and the mean of women.
- Or: Under effect coding, $\hat{\beta}_1$ is the difference of the mean of the men and the mean or the difference of the mean and the mean of the women.
- In the case of two equally large samples

$$\mu_y = \frac{\mu_{y_w} + \mu_{y_w}}{2}$$

- and therefore $\hat{\beta}_1$ turns into

$$\hat{\beta}_1 = \frac{\hat{\mu}_{y_m} - \hat{\mu}_{y_w}}{2} = \frac{\hat{\mu}_{y_m} - 2\hat{\mu}_y + \hat{\mu}_{y_m}}{2} = \hat{\mu}_{y_m} - \hat{\mu}_y.$$

Effect Coding: One-Factor ANOVA

- Effect coding can also be applied to variables that have more than two levels. However, several dummy variables are needed again.
- Example: The variable z (state) has three levels, "Ohio", "Nevada" and "California".
- We define the effect-coded dummy variables x_1 and x_2:

$$x_{1i} = \begin{cases} -1 \text{ if } z_i = \text{Ohio} \\ 1 \text{ if } z_i = \text{Nevada} \\ 0 \text{ if } z_i = \text{California} \end{cases}$$

$$x_{2i} = \begin{cases} -1 \text{ if } z_i = \text{Ohio} \\ 0 \text{ if } z_i = \text{Nevada} \\ 1 \text{ if } z_i = \text{California} \end{cases}$$

- Note: Only the reference level (Ohio) is encoded with -1. The other levels are 0 and 1.

Effect Coding: One-Factor ANOVA

- For the sake of simplicity $n = n_{\text{OH}} = n_{\text{NE}} = n_{\text{CA}}$. Then we have again $N = 3n$.
- We obtain

$$\mathbf{X} = \begin{bmatrix} 1 & -1 & -1 \\ 1 & -1 & -1 \\ \vdots & \vdots & \vdots \\ 1 & 1 & 0 \\ 1 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 1 \\ 1 & 0 & 1 \end{bmatrix} \tag{10}$$

- For $\hat{\boldsymbol{\beta}}$ we now obtain

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = \begin{bmatrix} \hat{\mu}_y \\ \hat{\mu}_{y_{\text{NE}}} - \hat{\mu}_y \\ \hat{\mu}_{y_{\text{CA}}} - \hat{\mu}_y \end{bmatrix}$$

Effect Coding: One-Factor ANOVA

- The mean value of the effect coded dummy variables is 0 (since the sum of the columns must be 0).
- The variance of the effect-coded dummy variables is respectively $\frac{2}{c}$ (with c levels in the effect coded variable).
- The covariance between two effect coded dummy variables is $\frac{1}{c}$.
- The correlation between two effect-coded dummy variables is $\frac{1}{c}/\frac{2}{c} = \frac{1}{2}$.
- This means that effect-coded dummy variables are correlated, which complicates the variance decomposition in the DV.
- Since all comparisons are done via the overall mean, this relationship among effect-coded dummy variables seems somehow intuitive.
- ▷ Orthogonal coding schemes
- ▷ Planned contrasts

Effect Coding: Summary

- For the effect coding, the sum of the columns of the dummy variables in the $\mathbf{X}$ matrix (also: design matrix) must be zero (cf. slide 43).
- The coefficient of each effect level (in the example: women living in the city) is not estimated, but can only be determined indirectly.
- The separation of the differences in the mean value is similar to contrasts in ANOVA.
- The regression coefficients inform about the respective *deviations of the group means from the overall mean value* or the deviation of the interaction from the sum of the corresponding main effects.

Today's Focus

Violation of Assumptions

- Violations of error-assumptions
  - Heteroscedasticity
  - Dependent errors
- Violations concerning predictors
  - Multicollinearity

Assumptions of the Regression model

- *Validity*: Data you are analyzing should map to the research question you are trying to answer
- *Additivity and linearity*: $Y = \beta_1 x_1 + \beta_2 x_2 + \ldots$, if $y = x_1 x_2 x_3$ then $\log(y) = \log(x_1) + \log(x_2) + \log(x_3)$
- *Independence of errors*: $\sigma_\epsilon^2 \mathbf{I}$
- *Equal variance of errors*: $\sigma_\epsilon^2 \mathbf{I}$
- *Normality of errors*: $\epsilon_i \sim N(\mathbf{0}, \sigma_\epsilon^{\mathbf{2}} \mathbf{I})$

Model Criticism

There is a temptation to become personally attached to a particular model

- Don't "fall in love" with your model
  - All models are wrong.
  - Some models are better than others.
  - The correct model can never be known with certainty.

Generally: The simpler the model, the better it is (Occam's razor)

Approachs to improve inadequate models:

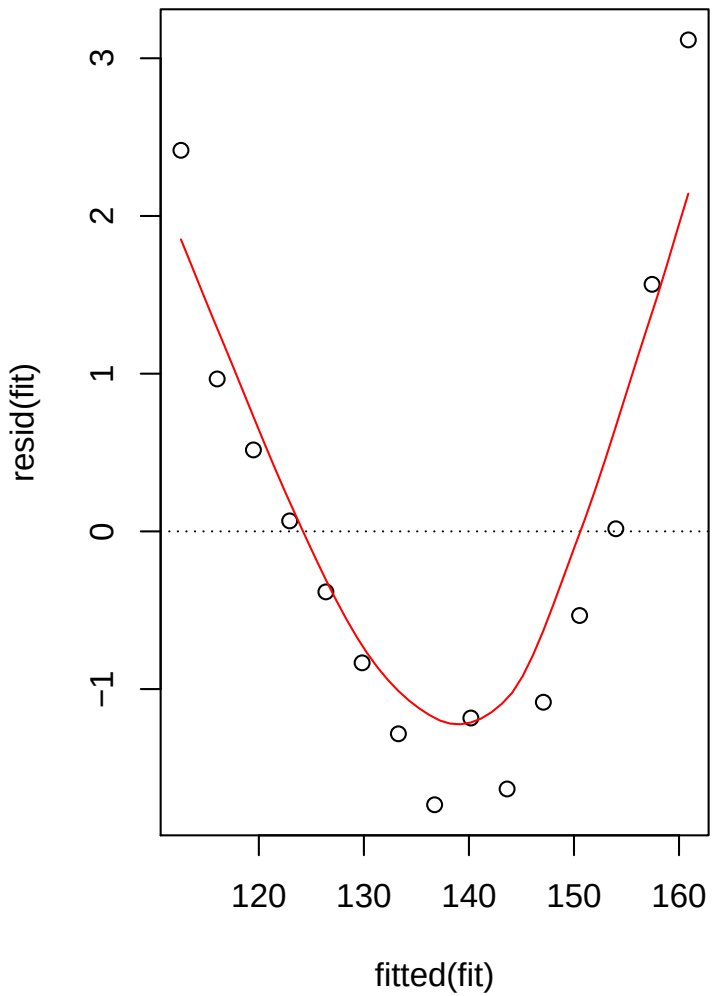
- Transform the response variable.
- Transform one or more of the explanatory variables.
- Try fitting different explanatory variables if you have any.
- Use a different error structure.
- Use non-parametric smoothers instead of parametric functions.
- Use different weights for different y values

Model Checking

After fitting a model to data

- Investigate how well the model describes the data
- Check for any systematic trends in goodness of fit
  - Does the goodness of fit increase with the observation number?
  - Is it a function of one or more of the explanatory variables?
  - Check raw residuals: $\hat{\epsilon} = y - \hat{y}$
- Good idea to routinely plot the residuals against:
  - The fitted values (to look for heteroscedasticity)
  - The explanatory variables (to look for evidence of curvature);
  - The sequence of data collection (to look for temporal correlation);
  - Standard normal deviates (to look for non-normality of errors).

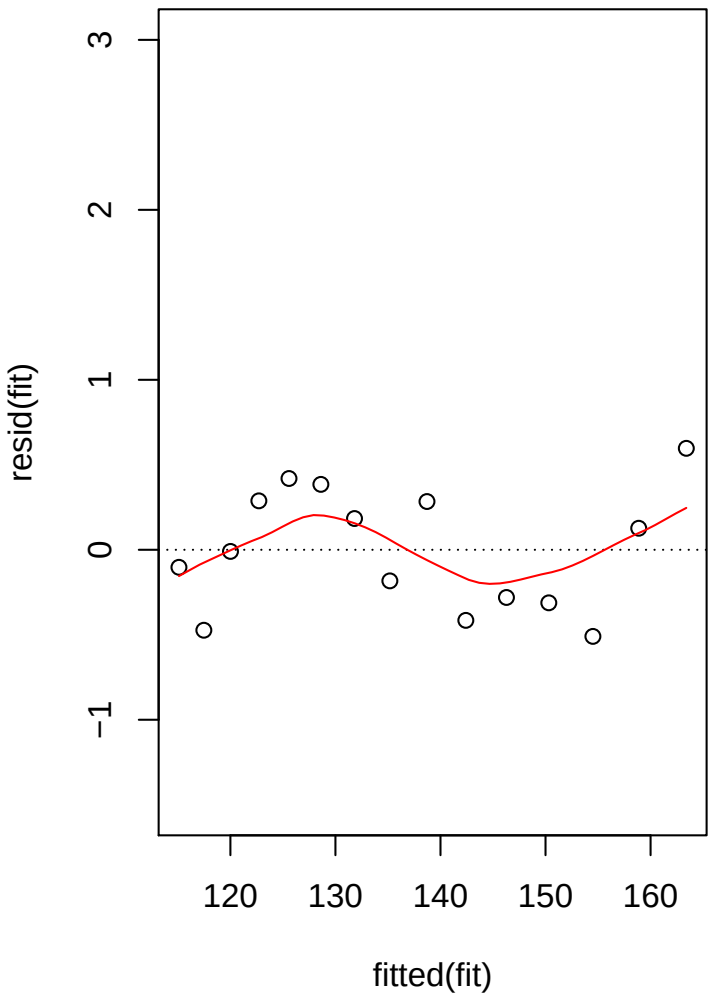
Example: Weight and Height




```
lm(formula = weight ~ height,
    data = women)
             coef.est coef.se
(Intercept) -87.52     5.94
height        3.45     0.09
---
n = 15, k = 2
residual sd = 1.53,
R-Squared = 0.99
```

- Clear evidence of a curved relationship
- ▷ add a quadratic term to the regression

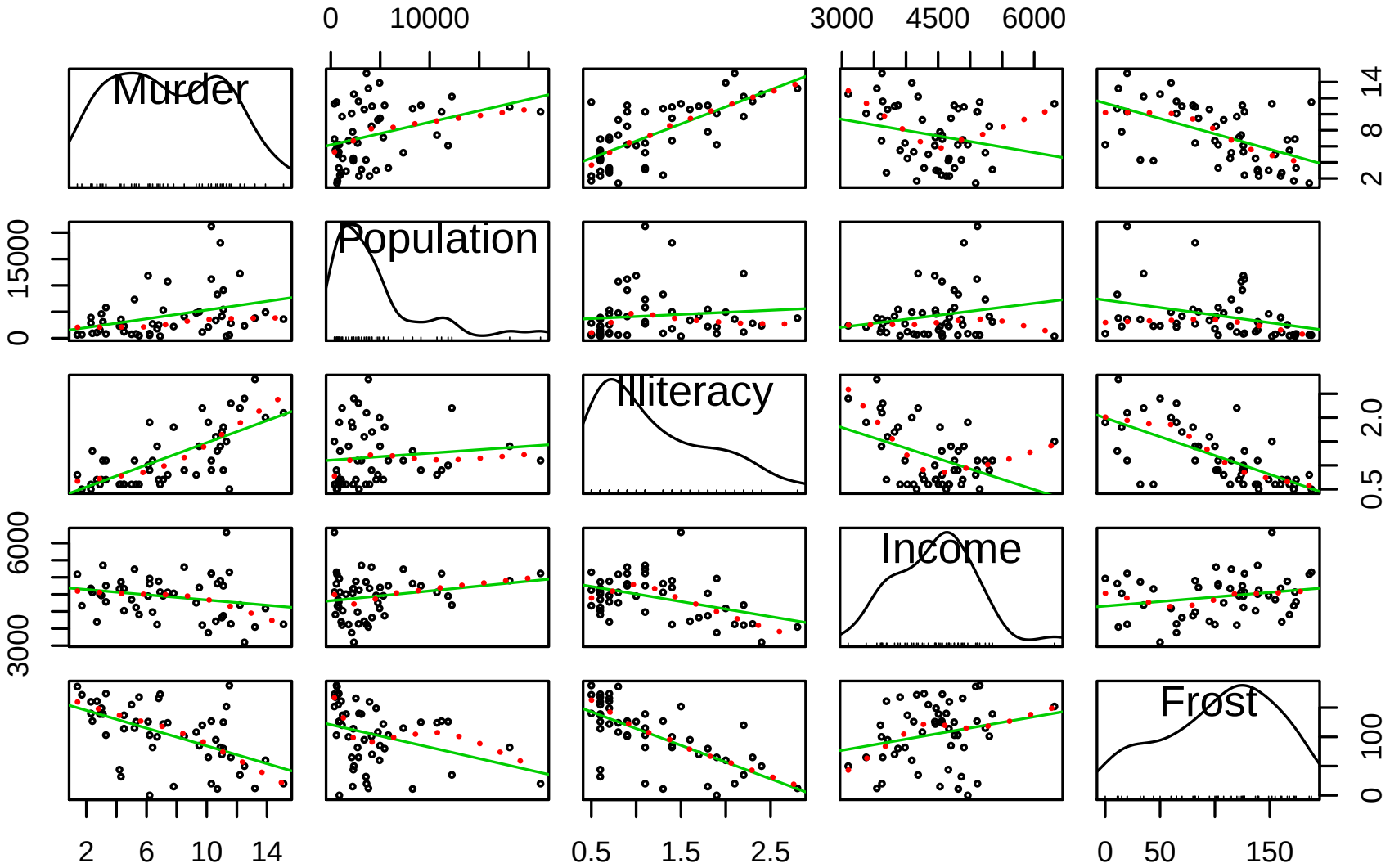
Example: Weight and Height




```
lm(formula = weight ~
  height + I(height^2),
  data = women)
             coef.est coef.se
(Intercept) 261.88    25.20
height       -7.35     0.78
I(height^2)   0.08     0.01
---
n = 15, k = 3
residual sd = 0.38,
R-Squared = 1.00
```

- Better fit with respect to linearity assumption

Example: Murder Rate Across States



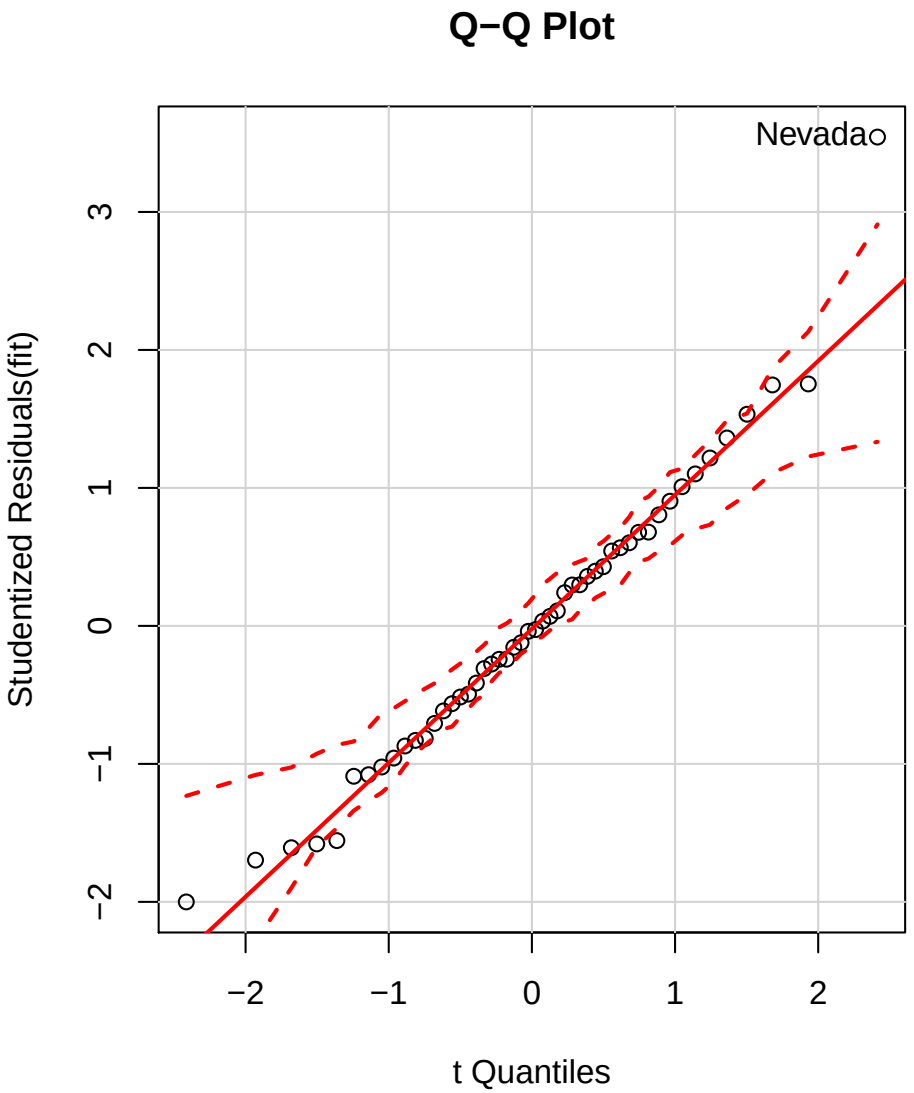
Example: Murder Rate Across States

```
lm(formula = Murder ~ Population
  + Illiteracy + Income + Frost,
  data = states)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.235e+00  3.866e+00   0.319   0.7510
Population  2.237e-04  9.052e-05   2.471   0.0173 *
Illiteracy  4.143e+00  8.744e-01   4.738 2.19e-05 ***
Income      6.442e-05  6.837e-04   0.094   0.9253
Frost       5.813e-04  1.005e-02   0.058   0.9541
---
n = 50, k = 5
residual sd = 2.53, R-Squared = 0.57
```

- Illiteracy and Population seems to be significant effects

Normality

```
qqPlot(fit, labels=row.names(states), id.n=1,  main="Q-Q Plot")
```



- Plot of studentized residuals (or studentized deleted residuals or jackknifed residuals)
- against a t-distribution with
- n - p - 1 degrees df freedom,
  - n is sample size
  - p is number of regression parameters
- Normality is met well - except for Nevada

Normality

Check actual data:

```
> states["Nevada",]
Murder Population Illiteracy Income Frost
Nevada   11.5        590        0.5   5149   188
```

- Observed: **11.5%** murder rate

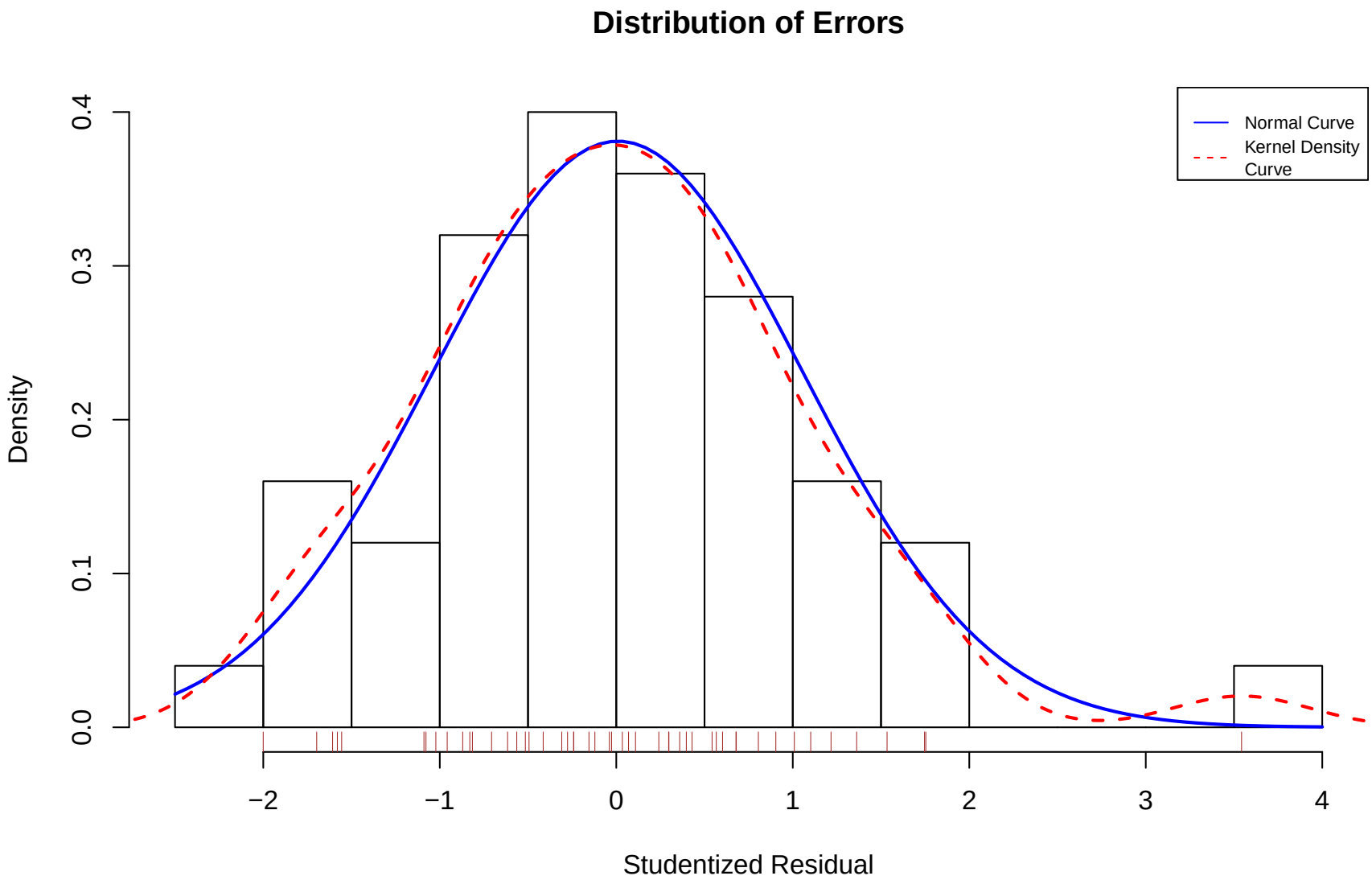
Check predictions:

```
> fitted(fit)["Nevada"]
  Nevada
3.878958
> residuals(fit)["Nevada"]
  Nevada
7.621042
```

- Predicted rate is **3.9%**
  - Why does Nevada have a higher murder rate than predicted from population, income, illiteracy, and temperature?

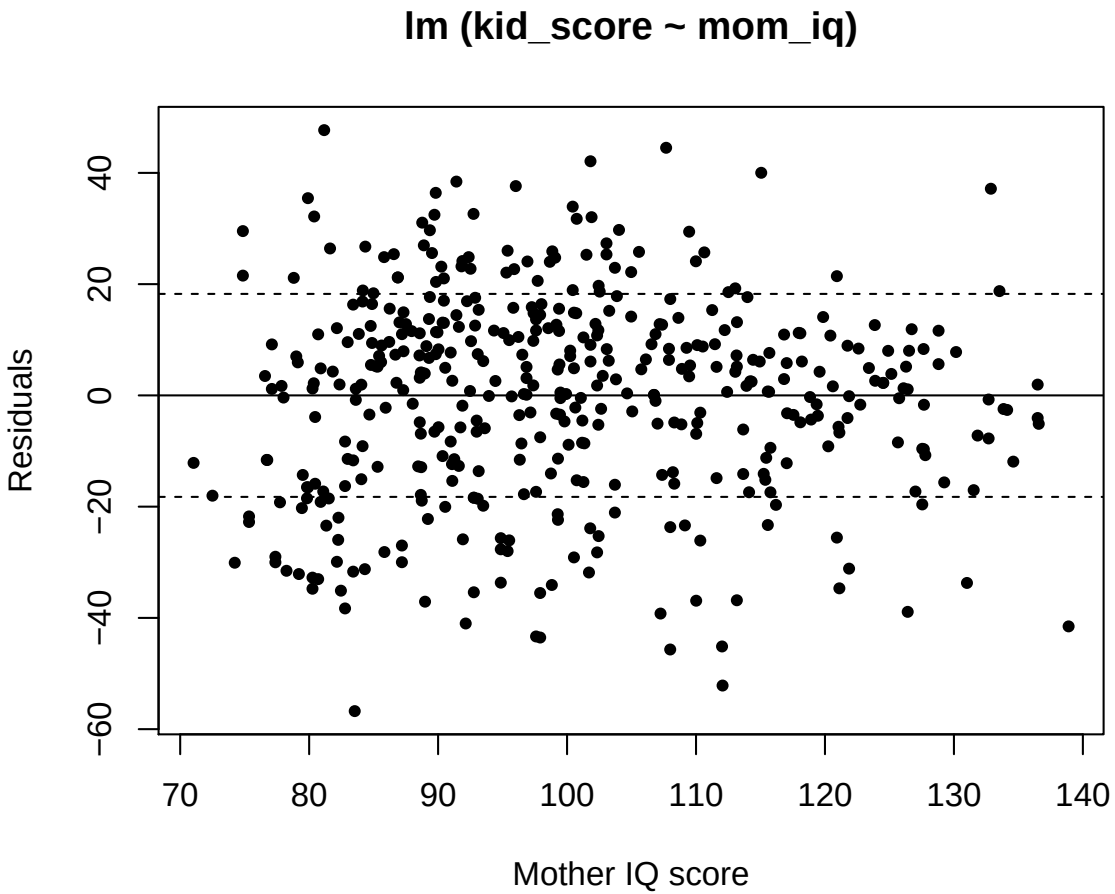
Normality

Same problem - different visualization:



Heterskedasticity

Good way to diagnose violations of some of the assumptions just considered (linearity, distribution of errors) is to plot the residuals ϵ_i versus fitted values $\mathbf{X}\hat{\boldsymbol{\beta}}$



Violation of Assumptions: Heteroscedasticity

- The second assumption in the Gauss-Markov theorem was

$$\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma_{\epsilon}^2 \mathbf{I},$$

  i.e. each error for N individuals has – across all possible replications – the same variance (and is, in assumption 3, that it is uncorrelated with other errors)
- If the assumption of constant variance, *homoscedasticity*, is violated we refer to it as *heteroscedasticity*
- Heteroscedasticity arises when
  - influence of errors is not the same for all individuals.
  - some individuals might have a certain degree of systematicity that is unaccounted for in the model

Violation of Assumptions: Heteroscedasticity

- Heteroscedasticity may arise for example if
  - in learning experiences higher skills are related to fewer errors (typings skills and typos)
  - observations are drawn from different units such as within individuals, schools, companies, countries, etc.
  - we assume statistical distributions in which means and dispersion are not independent of each other, such as for example in the binomial distribution.
  - we operate with aggregate data (e.g. means of school classes) where the aggregates differ in their size. They differ in size and standard errors depend on the number of individuals within the aggregate.

Violation of Assumptions: Heteroscedasticity

- In the case of heteroscedasticity Equation the sum of squared errors becomes

$$\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma_\epsilon^2 \mathbf{C},$$

where $\mathbf{C}$ is a diagonal matrix. The elements of its diagonal $\mathrm{diag}(C) = c_1, c_2, \ldots, c_N$ contain person specific weights, i.e.,

$$\mathbf{C} = \begin{bmatrix} c_1 & 0 & \ldots & 0 \\ 0 & c_2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & c_N \end{bmatrix},$$

- $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$ still holds and the estimates of the β-parameter are still unbiased as $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.

Violation of Assumptions: Heteroscedasticity

- However, this does not hold for the *variance* of $\hat{\boldsymbol{\beta}}$

In depth

$$\begin{aligned} \mathrm{E}[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})'] &= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathrm{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma_\epsilon^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}. \end{aligned}$$

- The variance of $\hat{\boldsymbol{\beta}}$ is not correctly estimated if we use the OLS approach which ignores $\mathbf{X}'\mathbf{C}\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}$
- In the case of heteroscedastic errors, the actual variance of $\hat{\boldsymbol{\beta}}$ is smaller than the one we obtain from the OLS approach.
- Consequently, the standard errors from an OLS are inflated for this case. And significance tests that are based on these standard errors will turn out to be conservative, i.e., they will return fewer significant results than expected (type II errors).

Violation of Assumptions: Heteroscedasticity

- In order to counter the this effect, we can "correct" our estimator: WLS-Regression (Weighted Least Squares)
  - Assuming that $\mathbf{C}$ is known.
  - We define the reciprocal

$$\mathbf{C}^{-1/2} = \begin{bmatrix} \frac{1}{\sqrt{c_1}} & 0 & 0\dots & 0 \\ 0 & \frac{1}{\sqrt{c_2}} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \frac{1}{\sqrt{c_N}} \end{bmatrix}$$

  - so that large variance values are "down" weighted more strongly compared to small values.
  - We respecify the regression by pre-multiplying with $\mathbf{C}^{-1/2}$:

$$\mathbf{C}^{-1/2}\mathbf{y} = \mathbf{C}^{-1/2}\mathbf{X}\boldsymbol{\beta} + \mathbf{C}^{-1/2}\boldsymbol{\epsilon}$$

Violation of Assumptions: Heteroscedasticity

- WLS "undoes" the effect of heteroscedasticity and we are back at the original form.
- In other words: We can resolve the problem of heteroscedasticity with the according transformation.

In depth

For $\mathrm{E}(\boldsymbol{\epsilon}_* \boldsymbol{\epsilon}_*')$ we now obtain

$$\begin{aligned}
\mathrm{E}(\boldsymbol{\epsilon}_* \boldsymbol{\epsilon}_*') &= \mathrm{E}(\{\mathbf{C}^{-1/2}\boldsymbol{\epsilon}\}\{\mathbf{C}^{-1/2}\boldsymbol{\epsilon}\}') \\
&= \mathbf{C}^{-1/2}\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}')\mathbf{C}^{-1/2} \\
&= \mathbf{C}^{-1/2}(\sigma_\epsilon^2 \mathbf{C})\mathbf{C}^{-1/2} \\
&= \sigma_\epsilon^2 \mathbf{I},
\end{aligned}$$

If we solve for $\hat{\boldsymbol{\beta}}$ under the WLS we obtain

$$\begin{aligned}
\hat{\boldsymbol{\beta}}^{\mathrm{WLS}} &= (\mathbf{X}'\mathbf{C}^{-1/2}\mathbf{C}^{-1/2}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1/2}\mathbf{C}^{-1/2}\mathbf{y} \\
&= (\mathbf{X}'\mathbf{C}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{C}^{-1}\mathbf{y}
\end{aligned}$$

Violation of Assumptions: Heteroscedasticity

- $\mathbf{C}$ is typically unknown and needs to be estimated according to an additional set of assumptions.
- WLS results in down weighting individuals with larger errors in the computation of $\hat{\boldsymbol{\beta}}$.
- Let's assume that in our example (on ps Age + Education) we had heteroscedasticity.
- Now, let's assume that the $\mathbf{C}$-matrix is given as

$$\mathbf{C} = \text{diag}(3, 5, 1, 1, 2, 8, 4, 1, 9, 2).$$

Violation of Assumptions: Heteroscedasticity

- Compute the corresponding $\hat{\boldsymbol{\beta}}$-vectors:

$$\hat{\boldsymbol{\beta}}^{\text{OLS}} = \begin{bmatrix} 35.86 \\ -0.51 \\ 2.21 \end{bmatrix} \text{ (as before) and } \hat{\boldsymbol{\beta}}^{\text{WLS}} = \begin{bmatrix} 27.55 \\ -0.45 \\ 2.59 \end{bmatrix}$$

- The estimates are now deviating from each other but they are still unbiased, that is, if repeated infinitely they should converge!
- For the standard errors of the three β-parameters we obtain for the OLS and WLS approach:

$$\hat{\boldsymbol{\sigma}}_{\beta}^{\text{OLS}} = \begin{bmatrix} 14.69 \\ 0.16 \\ 0.86 \end{bmatrix} \text{ und } \hat{\boldsymbol{\sigma}}_{\beta}^{\text{WLS}} = \begin{bmatrix} 10.41 \\ 0.12 \\ 0.69 \end{bmatrix}.$$

Violation of Assumptions: Heteroscedasticity

- Given the estimation via WLS both parameters β_1 and β_2 are statistically significant with

$$t_1^{\text{WLS}} = \frac{-.45}{0.12} = -3.75 \text{ and } t_2^{\text{WLS}} = \frac{2.59}{0.69} = 3.75,$$

  which in both cases is larger, in absolute terms, than the critical t-value of 1.895 ($\alpha = 5\%$, 7 df, one-tailed).
- in this example the OLS estimate would have been smaller ($t_1^{OLS} = -3.188$ and $t_2^{OLS} = 2.570$)

Violation of Assumptions: Heteroscedasticity

Simulated Data

- Example
  - We generate data according to witn $N = 150$

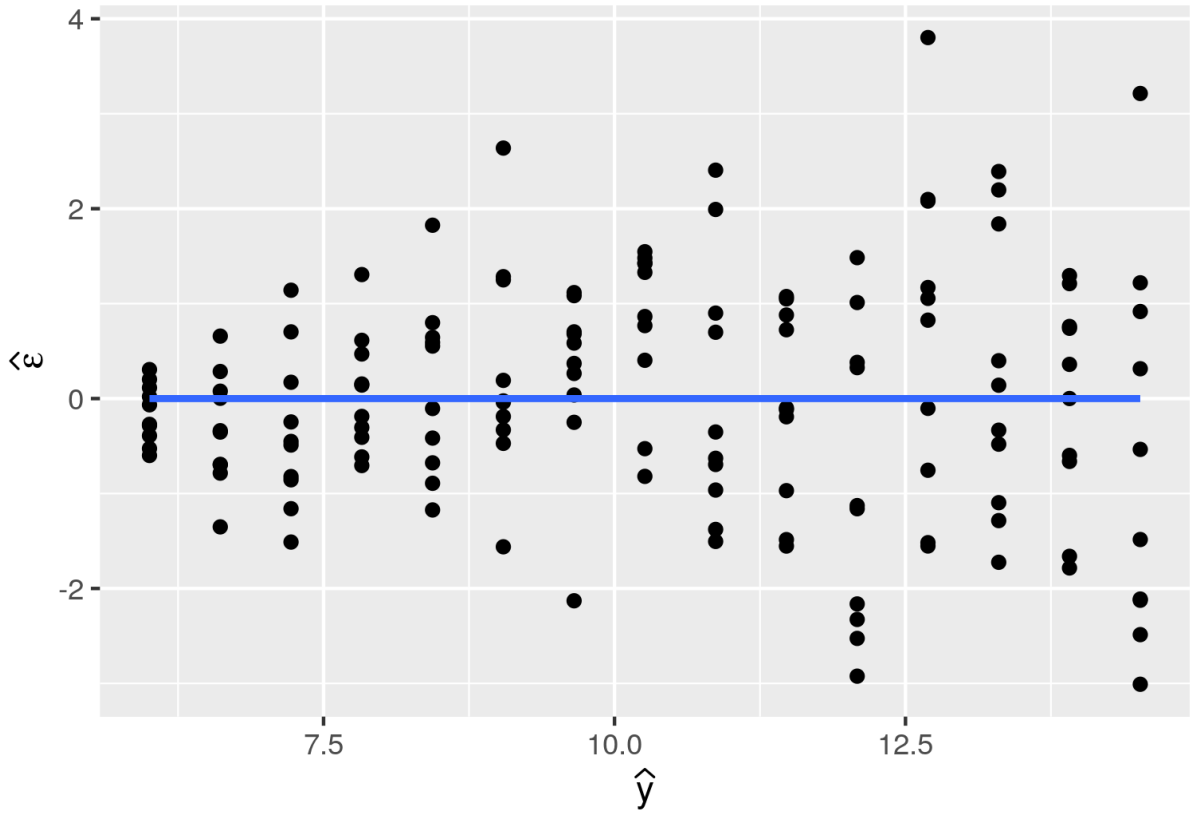
$$y_i = 5 + x_i + \frac{1}{2} x_i^{2/3} z,$$

    with z being a random standardized normally distributed factor.
  - The last term potentially increases with larger x_i values.

Violation of Assumptions: Heteroscedasticity

Simulated Data

- Plotting $\hat{y}_i$ versus $\hat{\epsilon}_i$



- With increasing $\hat{y}_i$ the $\hat{\epsilon}_i$ are increasing as well: *Heteroscedasticity*.

Violation of Assumptions: Heteroscedasticity

Table: OLS versus WLS-Estimates

Parameter	OLS	WLS†
$\hat{\beta}_0$	5.39	5.19
$\hat{\sigma}_{\beta_0}$	0.21	0.08
t	25.86	64.48
$\hat{\beta}_1$	1.22	1.27
$\hat{\sigma}_{\beta_1}$	0.05	0.04
t	26.49	34.72
R^2	0.82	0.89

† The weights correspond to $x_i^{-3/2}$.

- The OLS as well as WLS are unbiased estimators.
- The WLS-estimates have smaller standard errors.

Violation of Assumptions: Heteroscedasticity

Formal Test

- White's Test for heteroscedasticity:
- R packages are terribly annoying – "manual" approach

Step 1: Regression analysis $\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}$.

Step 2: Save and square residuald ($\hat{\epsilon}_i = y_i - \hat{y}_i$):

```
sqr <- residuals(fit2)^2
```

Step 3: Regression analysis of the squared residuals on the x-variables and save R^2: `r2 <- summary(lm(sqr ~ x))$r.squared`

Step 5: It holds:

$$NR^2 \sim \chi^2(k), \tag{11}$$

wit k predictors excluding the intercept β_0.

- In our example $R^2_{\hat{\epsilon}_i^2} = .15$ and hence $NR^2 = 150 \times .15 = 22.5$.
- 22.5 is significantly larger compared to the critical χ^2-value (`qchisq(.95, df = 1)`) of 3.84.
- Our residual distribution is indeed heteroscedastic.

Violation of Assumptions: Independence of Error

- The third assumption of the Gauss-Markov theorem wasthat the errors of observations i and j $(i \neq j; i, j = 1 \ldots N)$ are uncorrelated.
- This independence of the single observations is lost if, e.g., the data is gathered from repeated observations (longitudinal data) and if the time variable is not accounted for.
- Dependence among single observations could also be due to other reasons such as spatial correlation, eg., crops from similar locations on a field share properties which make them more alike.

Violation of Assumptions: Independence of Error

- In the case of correlated errors Equation $\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma_\epsilon^2 \mathbf{I}$ becomes

$$\mathrm{E}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}') = \sigma_\epsilon^2 \boldsymbol{\Omega}, \tag{12}$$

where $\boldsymbol{\Omega}$ is a $N \times N$ correlation matrix, i.e.

$$\boldsymbol{\Omega} = \begin{bmatrix} 1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{12} & 1 & \cdots & \rho_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{1N} & \rho_{2N} & \cdots & 1 \end{bmatrix}.$$

In depth

Analogous to the derivation with respect to heteroscedasticity we obtain an estimate for the variance in $\hat{\boldsymbol{\beta}}$ as

$$\mathrm{E}[\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})][\hat{\boldsymbol{\beta}} - \mathrm{E}(\hat{\boldsymbol{\beta}})'] = \sigma_\epsilon^2 (\mathbf{X}'\mathbf{X})^{-1} \mathbf{X}'\boldsymbol{\Omega}\mathbf{X} (\mathbf{X}'\mathbf{X})^{-1}.$$

Violation of Assumptions: Independence of Error

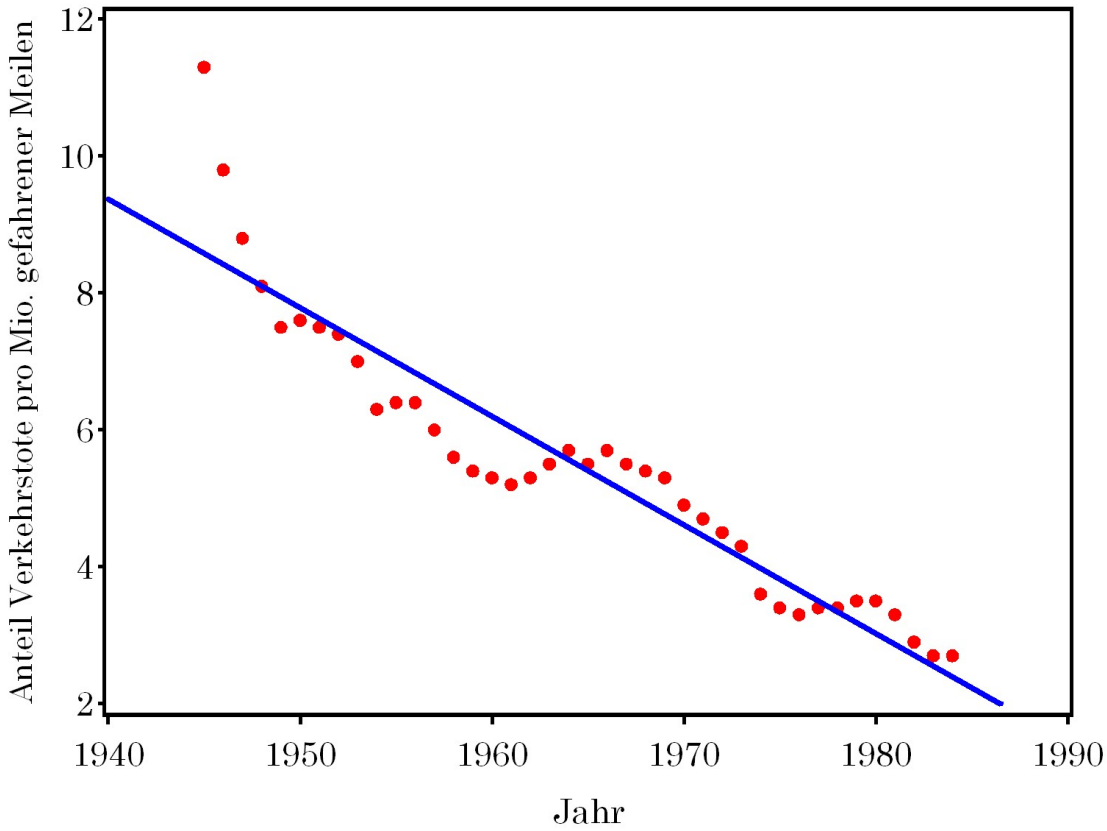
- The data provides less information than our model "assumes"!
- Solution in the case of dependent data: GLS-regression (Generalized Least Squares).
- We assume that $\mathbf{\Omega}$ has a certain structure (more on this later).
- GLS weigh the predictors

$$\hat{\boldsymbol{\beta}}^{\text{GLS}} = (\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}. \tag{13}$$

- The effect of GLS is that dependent observations (they provide less information than independent) are down-weighted in the estimation of $\hat{\boldsymbol{\beta}}$.
- Problem: $\mathbf{\Omega}$ can have different plausible structures arising from different assumptions.

Violation of Assumptions: Independence of Error

- Motor vehicle related deaths in the US across time



- Errors are positively and negatively distributed around the regression line – but they also seem to fluctuate systematically.
- Not untypical for time dependent data.

Violation of Assumptions: Independence of Error

Table: OLS versus WLS-Estimates

Parameter	OLS	GLS†
$\hat{\beta}_0$	316.64	383.30
$\hat{\sigma}_{\beta_0}$	18.29	46.57
t	17.31	8.23
$\hat{\beta}_1$	-0.1584	-0.1921
$\hat{\sigma}_{\beta_1}$	0.0094	0.0237
t	-17.03	8.11
R^2	0.88	0.65

† $\mathbf{\Omega}$ was estimated with a first order autoregressive structure ($\hat{\rho} = .86$), see next slide.

- The GLS-estimates show (correctly!) larger standard errors.

Violation of Assumptions: Independence of Error

- With a first order autoregressive structure $\mathbf{\Omega}$ becomes

$$\mathbf{\Omega} = \begin{bmatrix} 1 & \rho & 0 & 0 & \dots & 0 & 0 \\ \rho & 1 & \rho & 0 & \dots & 0 & 0 \\ 0 & \rho & 1 & \rho & \dots & 0 & 0 \\ 0 & 0 & \rho & 1 & \dots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & \dots & 1 & \rho \\ 0 & 0 & 0 & 0 & \dots & \rho & 1 \end{bmatrix}.$$

- This structure assumes that each observation is correlated only with the one preceding and succeeding itself.
- Moreover, the first order autocorrelation is constant for, i.e., of the same size for all observations.
- In our example $\hat{\rho} = .86$.

Violation of Assumptions: Independence of Error

- How can we discover a (temporal or spatial) dependence among observations – if present?
- 1. Visualize the Disturbance: As seen in the plot on slide 74, errors may form a characteristic pattern. Prerequisite is that data are sorted in the "right" order to detect this. E.g. in temporal oder or in the right spatial order.
- 2. Durbin-Watson-Statistic: For (presumed) temporal dependency we may use the Durbin-Watson-Statistic which is defined as:

$$D = \frac{\sum_{t=2}^{T}(\hat{\epsilon}_t - \hat{\epsilon}_{t-1})^2}{\sum_{t=2}^{T}(\hat{\epsilon}_t)^2} \approx 2(1 - \hat{\rho}), \tag{14}$$

  with $\hat{\rho}$ being the first order autocorrelation.
- In R: ACF, dwtest function in the lmtest package, durbinWatsonTest function in the car package, and pdwtest for panel models in the plm package.

Violation of Assumptions: Independence of Error

- The D-values can range from 0 to 4. It holds:

$$D = 0 \text{ if } \rho = 1,$$
$$D = 2 \text{ if } \rho = 0,$$
$$D = 4 \text{ if } \rho = -1.$$

- The computed D-value can be compared to (tabulated) critical D-values to check for significance – or let R do the work
-
```
> durbinWatsonTest(fit)
 lag Autocorrelation D-W Statistic p-value
   1       0.8606929      0.270191   0.014
 Alternative hypothesis: rho != 0
```

Violation of Assumptions: Multicollinearity

- One condition for regression analysis is that the predictor variables do not relate perfectly to each other.
- What happens if the predictor variables are highly correlated?
- Back to the ps and Age example. Let's add another predictor variable $x_3 = \sqrt{x_1}$, i.e., the square root of age.
- The variables x_1 and x_3 are (obviously) almost perfectly correlated ($r = 0.999$).
- What are the estimates from our regression model

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_3 x_{3i} + \epsilon_i$$

Violation of Assumptions: Multicollinearity

```
lm(formula = y ~ x1 + x2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -72.374    601.121  -0.120    0.908
x1            -2.541      8.911  -0.285    0.784
x2            32.694    146.657   0.223    0.830

Residual standard error: 6.29 on 7 degrees of freedom
Multiple R-squared:  0.477,Adjusted R-squared:  0.3275
F-statistic: 3.192 on 2 and 7 DF,  p-value: 0.1035
```

- Note standard errors
- These large standard errors now result in **no** statistical significant parameters.

Violation of Assumptions: Multicollinearity

- Consequence of multicollinearity
  1. Standardized regression weights ($\hat{\beta}^z$) can get larger than 1 (in our example $\hat{\beta}_1^z = -3.15$ and $\hat{\beta}_3^z = 2.46$).
  2. If x_j and x_k correlate positively (negatively) the corresponding correlations among $\hat{\beta}_j$ and $\hat{\beta}_k$ can be negative (positive). In our example, -1306.41, corresponds to a correlation of $r_{\hat{\beta}_1 \hat{\beta}_3} = -.999$).
  3. The sign of the estimate ($\hat{\beta}$) can differ from the sign of the population value (β).
  4. The standard errors of the estimates increase substantially.
- In practice, we can't observe the third issue, as we don't know the population values.
- A "hint" could be that the sign of our parameter does not correspond with our theoretical assumptions, based on previous work.

Violation of Assumptions: Multicollinearity

- How can we determine whether we face multicollinearity?
  1. Standardized regression weights > 1
  2. Reversed relations among the regression weights compared to the corresponding variables.
  3. Obtain *tolerance* for each predictor variable.
- Tolerance = The change in unexplained variance when an additional variable is included in the model. The closer the value is to 0 the more likely we deal with multicollinearity.
- Im Example:
  $\text{Tol.}(x_3) = 1 - R^2_{3.1} = 0.0006$ and $\text{Tol.}(x_1) = 1 - R^2_{1.3} = 0.0006$.
  (The values are identical as we only have two predictors, x_1, x_3, in the model.)
- With these numbers we have strong indication of multicollinearity – in real scenarios it will not be as straight forward.

Violation of Assumption: Summary

- Violations concerning errors and predictors
  - Heteroscedasticity: WLS
  - Dependence of errors: GLS
  - ▷ Attempts to "undo" bias

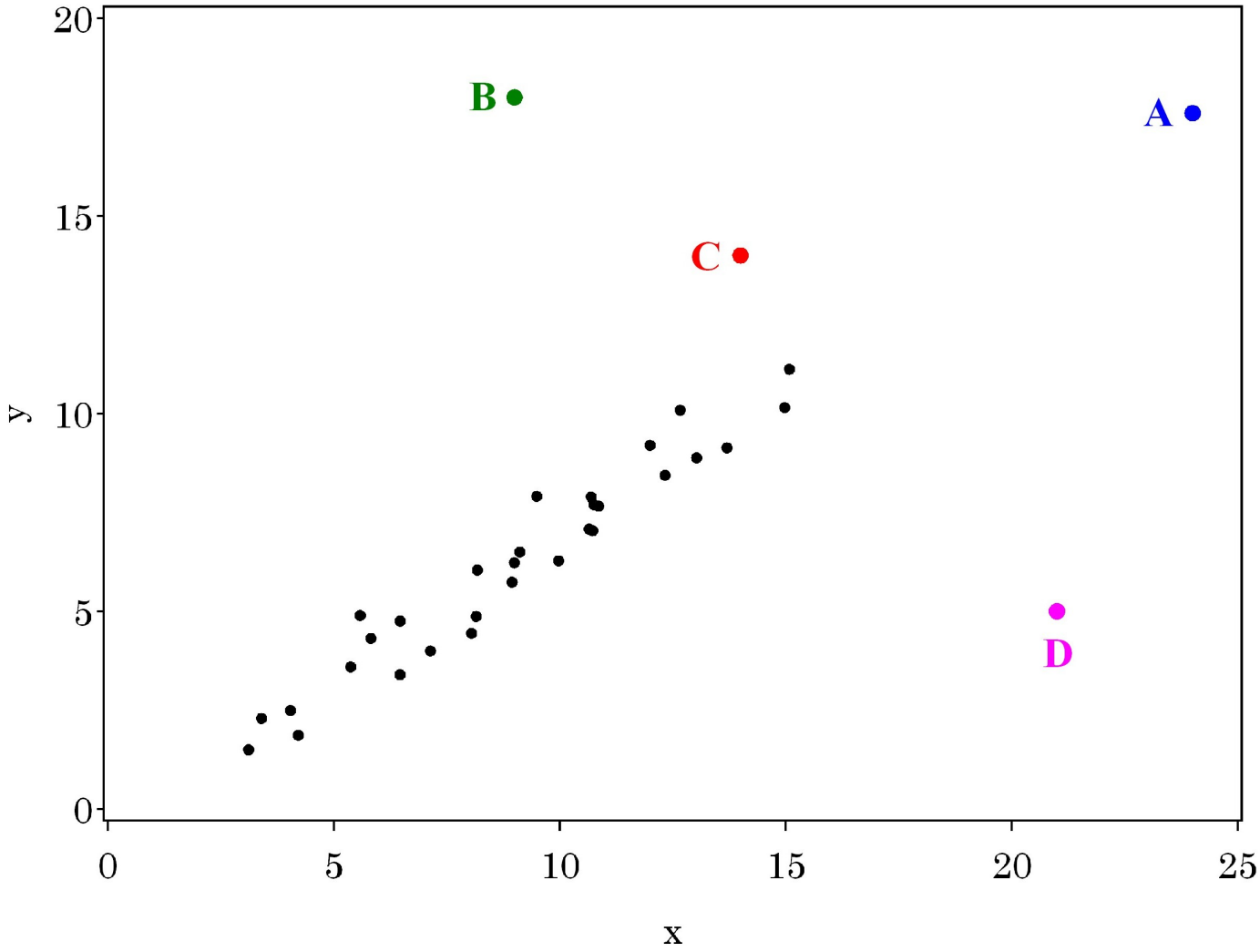
- Multicollinearity
  - Estimates are completely unstable
  - No simple solution

Outliers, Leverage- and Influential Data Points

- Single data points can have a considerable effect on the estimation of regression coefficients.
- We can differentiate between outliers, leverage points, and influential data points.
  - *Outliers* are observations with large residuals.
  - *Leverage points* are values that lay far outside from the other predictor variables.
  - *Influential points* are values which have a drastic effect on the regression function, i.e, with respect to the estimation of β-weights.
- These points are illustrated in the following plot.

Outliers, Leverage- and Influential Data Points

- Outliers, Leverage- and Influential Data Points



Outliers, Leverage- and Influential Data Points

Table: OLS-estimates with and without A, B, C, D

Parameter		A	B	C	D
$\hat{\beta}_0$	-.61	-.63	-.22	-.99	1.44
$\hat{\sigma}_{\beta_0}$	.34	.28	1.19	.49	.79
$\hat{\beta}_1$	.75	.75	.75	.81	.50
$\hat{\sigma}_{\beta_1}$	.04	.03	.12	.05	.07

- A is a leverage point, but not an outlier nor an influential point.
- B is not a leverage point, but it is an outlier and it is influential (with respect to $\hat{\beta}_0$).
- C is neither a leverage point nor and outlier but it is influential (with respect to $\hat{\beta}_0$ and $\hat{\beta}_1$).
- D is a leverage point and an outlier and influential.

Outliers, Leverage- and Influential Data Points: Hat Matrix

- In order to define these different types of deviating data points we need the *hat matrix* (aka projection matrix, influence matrix).
- The hat matrix projects, geometrically speaking, the y_i-values on to the $\hat{y}_i$-values
- It holds for $\hat{\mathbf{y}}$

$$\begin{aligned} \hat{\mathbf{y}} &= \mathbf{X}\hat{\boldsymbol{\beta}} && \text{[Eq. ??]} \\ &= \underbrace{\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'}\mathbf{y} \\ &= \qquad \mathbf{H} \qquad \mathbf{y}. && (15) \end{aligned}$$

- The matrix $\mathbf{H}$ is called "hat" – or projection matrix – as it puts a hat on to the $\mathbf{y}$ vector.
- Hence, $\mathbf{H}$ indicates to what extent $\hat{\mathbf{y}}$ is influenced by $\mathbf{y}$.

Outliers, Leverage- and Influential Data Points: Hat Matrix

- For single $\hat{y}_i$ it holds:

$$\hat{y}_i = \sum_{j=1}^{N} h_{ij} y_j,$$

  with h_{ij} referring to the (i,j) element in $\mathbf{H}$.
- For each person i we sum across the columns to obtain a weighted sum of *all* y-values.
- It holds:

$$\frac{1}{N} \leq h_{ii} \leq 1, \; -1 \leq h_{ij} \leq +1 \quad (i \neq j) \quad \text{and} \quad \sum_{j=1}^{N} h_{ij} = 1.$$

- If $h_{ii} = 1$ (and with it in the ith row all $h_{ij} = 0$, off-diagonal elements are 0), then $\hat{y}_i = y_i$. Large h_{ii}-values indicate that this point is far from the other points in the predictor space. Lit.

Leverage Points

- The reciprocal of h_{ii}, $\frac{1}{h_{ii}}$, can be interpreted as number of observations that determine $\hat{y}_i$
- The size of the h_{ii}-values depends on the relation of the $\mathbf{x}_i'$-vectors to the other rows of $\mathbf{X}$.
- For mean-centered $\mathbf{x}_i^{*\prime}$-vectors this means that h_{ii} is large
  - if $\mathbf{x}_i^{*\prime}\mathbf{x}_i^{*}$ is large, i.e., when $\mathbf{x}_i^{*\prime}$ has a large (quadratic) distance to the other $N-1$ values
  - because evidently this implies that $\mathbf{x}_i^{*\prime}$ (or rather the corresponding sum of squares) lies far away from the means of the other x-variables.
- **Leverage points** are thus values that generate large h_{ii}-values.
- h_{ii} represents the influence of y_i on fitting the regression model.

Leverage Point

- In our example data $\mathbf{H}$ is:

$$\mathbf{H} = \begin{bmatrix} .34 & .11 & .20 & .27 & .07 & .16 & .03 & .11 & -.07 & -.23 \\ .11 & .46 & .27 & .05 & .30 & -.16 & .01 & -.19 & .07 & .08 \\ .20 & .27 & .22 & .15 & .18 & .01 & .03 & -.03 & .02 & -.04 \\ .27 & .05 & .15 & .22 & .05 & .18 & .07 & .15 & -.01 & -.13 \\ .07 & .30 & .18 & .05 & .22 & -.05 & .06 & -.07 & .10 & .13 \\ .16 & -.16 & .01 & .18 & -.05 & .30 & .15 & .31 & .08 & .03 \\ .03 & .01 & .03 & .07 & .06 & .15 & .14 & .17 & .15 & .19 \\ .11 & -.19 & -.03 & .15 & -.07 & .31 & .17 & .34 & .11 & .09 \\ -.07 & .07 & .02 & -.01 & .10 & .08 & .15 & .11 & .22 & .33 \\ -.23 & .08 & -.04 & -.13 & .13 & .03 & .19 & .09 & .33 & .54 \end{bmatrix} \quad (16)$$

- The highest h_{ii}-value of .54 is obtained for person 10.
- Rule of thumb: If $h_{ii} > {}^{2k+1}/_{N}$ (k is number of predictors without intercept), then we can consider that data point to be a leverage point. This is the case for person 10.

Outlier, Leverage- and Influential Data Points: $\hat{\epsilon}$

- For the vector with the estimated errors we obtain:

$$\begin{aligned}
\hat{\boldsymbol{\epsilon}} &= \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} && \text{[Eq. ??]} \\
&= \mathbf{y} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} && \text{[Eq. 15]} \qquad (17) \\
&= \mathbf{y} - \mathbf{H}\mathbf{y} \\
&= (\mathbf{I} - \mathbf{H})\mathbf{y}. \qquad (18)
\end{aligned}$$

- In some ways this is the opposite of Equation (15).
- For single $\hat{\epsilon}_i$ we define:

$$\hat{\epsilon}_i = y_i - \sum_{j=1}^{N} h_{ij} y_j.$$

- The estimated error for person i results from the difference of y and a weighted sum of all y-values.

Outlier

- **outliers** are defined as those values that exhibit the largest estimated errors.
- Given our example data

$$\hat{\boldsymbol{\epsilon}}' = (-2.15, -1.48, 0.45, -5.59, 7.79, 4.23, 2.32, 0.77, -3.35, -2.99).$$

- Person 5, followed by 4 and 6 etc., could be classified as outlier.

But What value makes a an outlier and actual outlier?

- There are a number of significance tests to address this question, of which we will consider only the studentized deleted residuals.

Outlier: Studentized Deleted Residuals

- Idea:
  - Delete case i, and refit model.
  - Compute the predicted value and residual for case i using this model.
  - Compute the studentized residual for case i.
- Notation $-i$ means i has been deleted from computation
- $\hat{\epsilon}_{-i} = y_i - \hat{y}_{-i}$ is the deleted residual

- We don't need to do this literally
- This can be calculated directly

Outlier: Studentized Deleted Residuals

- Studentized deleted residuals are defined as

$$\hat{\epsilon}_{-i} = \hat{\epsilon}_i \left(\frac{d_{ii}\hat{\sigma}_\epsilon^2(N-k) - \hat{\epsilon}_i^2}{N-k-1} \right)^{-1/2}, \tag{19}$$

  where $\hat{\epsilon}_{-i}$ is the deleted residual for observation i and d_{ii} the corresponding diagonal element from matrix $\mathbf{D}$ (see Eq. 19). For deleted residuals it holds $\hat{\epsilon}_{-i} \sim t_{(N-k-1)}$.

- In the example on slide 90 Person 5 has a value (given $d_{ii} = 1 - h_{ii} = 1 - .22 = .78$) of

$$\hat{\epsilon}_{-i} = 7.79 \left(\frac{.78 \times 20.45 \times (10-2) - 7.79^2}{10-2-1} \right)^{-1/2} = 2.51, \tag{20}$$

  which is larger than the critical t-value of 1.895 ($\alpha = .05$, one-tailed, $df = 7$).
  We conclude that Person 5 indeed represents an outlier.

Influential Data Points

- Datapoints are **influential** if their presence considerably changes the estimates of the β-parameters.
- The question here is: How does $\hat{\boldsymbol{\beta}}$ change when a person (i.e. one row in $\mathbf{X}$) is removed?
- A parameter that reflects this idea is defined as

$$\text{DFBETA}_{(-i)} = \hat{\boldsymbol{\beta}} - \hat{\boldsymbol{\beta}}_{(-i)} = \frac{\mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\hat{\epsilon}_i}{d_{ii}}, \tag{21}$$

  with $\hat{\boldsymbol{\beta}}_{-i}$ being the estimate after removing person i, $\mathbf{x}_i'$ as ith row in the $\mathbf{X}$ matrix and d_{ii} as the corresponding diagonal element from matrix $\mathbf{D}$ (see slide 19).
- With this we can estimate which person has the largest influence on the estimate of each single β.