TMA4267 Linear Statistical Models V2017 (L11)

Part 2: Linear regression:
Parameter estimation [F:3.2] and model selection [F:3.4]
Hypothesis test for one regression coefficient
Studentized and standardized residuals
decomposition of variability and signficance of regression
R^2, SPSE=Expected squared prediction error

Mette Langaas

Department of Mathematical Sciences, NTNU

To be lectured: February 21, 2017

Today

1. Hypothesis testing for β_j.
2. Residuals: standardized (or studentized) preferred.
3. Decomposition of variability: SST=SSR+SSE, and significance of regression.
4. R^2 gives the proportion of variability explained by the regression model. and will never decrease if new covariates are added to the model.
5. Model choice considerations.
6. SPSE: Expected squared prediction error.

The classical linear model

The model

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

is called a classical linear model if the following is true:

1. $\mathrm{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$.
2. $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \mathrm{E}(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T) = \sigma^2 \boldsymbol{I}$.
3. The design matrix has full rank $\mathrm{rank}(\boldsymbol{X}) = k + 1 = p$.

The classical *normal* linear regression model is obtained if additionally

1. $\boldsymbol{\varepsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I})$

holds. For random covariates these assumptions are to be understood conditionally on $\boldsymbol{X}$.

Properties for the normal linear model

- Least squares and maximum likelihood estimator for $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{Y}$$

  with $\hat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}^T\boldsymbol{X})^{-1})$.

- Restricted maximum likelihood estimator for $\boldsymbol{\sigma}^2$:

$$\hat{\boldsymbol{\sigma}^2} = \frac{1}{n-p}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}) = \frac{\text{SSE}}{n-p}$$

  with $\frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$.

- Statistic for inference about β_j, c_{jj} is diagonal element j of $(\boldsymbol{X}^T\boldsymbol{X})^{-1}$.

$$T_j = \frac{\hat{\beta}_j - \beta_j}{\sqrt{c_{jj}}\hat{\sigma}} \sim t_{n-p}$$

Acid rain in Norwegian lakes

Measured pH in Norwegian lakes explained by content of

- x1: SO_4: sulfate (the salt of sulfuric acid),
- x2: $N0_3$: nitrate (the conjugate base of nitric acid),
- x3: Ca: calsium,
- x4: latent Al: aluminium,
- x5: organic substance,
- x6: area of lake,
- x7: position of lake (Telemark or Trøndelag),

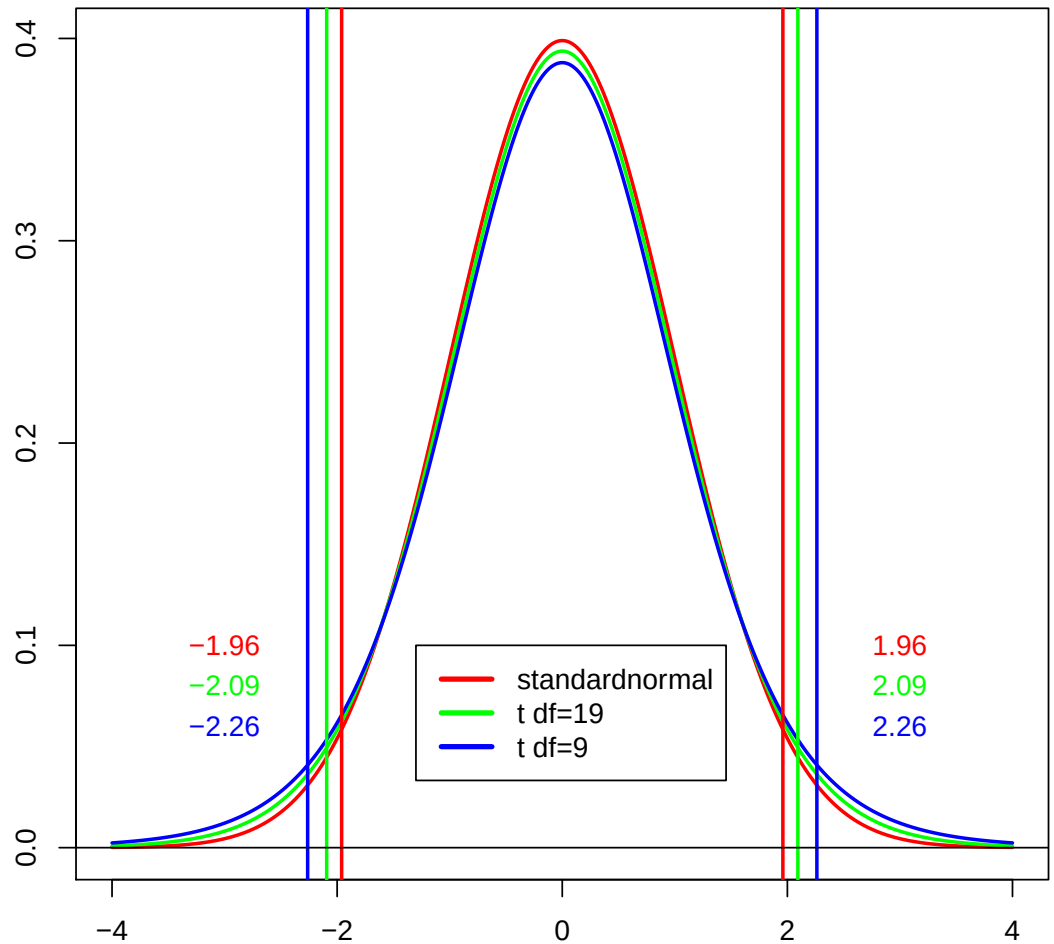
Random sample of $n = 26$ lakes.

Output from fitting the full model in R

```
> fit=lm(y~.,data=ds)
> summary(fit)
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  5.6764334  0.1389162  40.862  < 2e-16 ***
x1          -0.3150444  0.0587512  -5.362 4.27e-05 ***
x2          -0.0018533  0.0012587  -1.472    0.158
x3           0.9751745  0.1449075   6.730 2.62e-06 ***
x4          -0.0002268  0.0010038  -0.226    0.824
x5          -0.0334242  0.0225009  -1.485    0.155
x6          -0.0039399  0.0724339  -0.054    0.957
x7           0.0888722  0.1025724   0.866    0.398
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1165 on 18 degrees of freedom
Multiple R-squared:   0.93,Adjusted R-squared:  0.9027
F-statistic: 34.15 on 7 and 18 DF,  p-value: 3.904e-09
```

Quantiles and critical values: N og t: $\alpha/2 = 0.025$



In R: specify area to the left, but our notation gives area to the right. Fahrmeir et al: notation with area to the left.

Properties of the residuals

- Residuals (raw): $\hat{\varepsilon} = \boldsymbol{Y} - \hat{\boldsymbol{Y}}$.
- with mean $\mathrm{E}(\hat{\varepsilon}) = \mathbf{0}$ and covariance matrix $\mathrm{Cov}(\hat{\varepsilon}) = \sigma^2(\boldsymbol{I} - \boldsymbol{H})$ where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.
- In the normal model $\varepsilon \sim N_n(\mathbf{0}, \sigma^2\boldsymbol{I})$ and then also the vector of residuals are normal, but with heteroscedastic variances and non-zero covariances.
- *Standardized residuals:* divide (raw) residuals by estimated standard deviation.
- *Studentized residuals*: leave-one-out version.
- Studentized residuals are compared with the normal distribution to assess normality of the error term.

3.12 Overview of Residuals

Ordinary Residuals

The residuals are given by

$$\hat{\varepsilon}_i = y_i - \hat{y}_i = y_i - \boldsymbol{x}_i' \hat{\boldsymbol{\beta}} \quad i = 1, \ldots, n.$$

Standardized Residuals

The standardized residuals are defined by

$$r_i = \frac{\hat{\varepsilon}_i}{\hat{\sigma}\sqrt{1 - h_{ii}}},$$

where h_{ii} is the ith diagonal element of the hat matrix.

Studentized Residuals

The studentized residuals are defined by

$$r_i^* = \frac{\hat{\varepsilon}_{(i)}}{\hat{\sigma}_{(i)}(1 + \boldsymbol{x}_i'(\boldsymbol{X}_{(i)}'\boldsymbol{X}_{(i)})^{-1}\boldsymbol{x}_i)^{1/2}} = \frac{\hat{\varepsilon}_i}{\hat{\sigma}_{(i)}\sqrt{1 - h_{ii}}} = r_i \left(\frac{n - p - 1}{n - p - r_i^2}\right)^{1/2}.$$

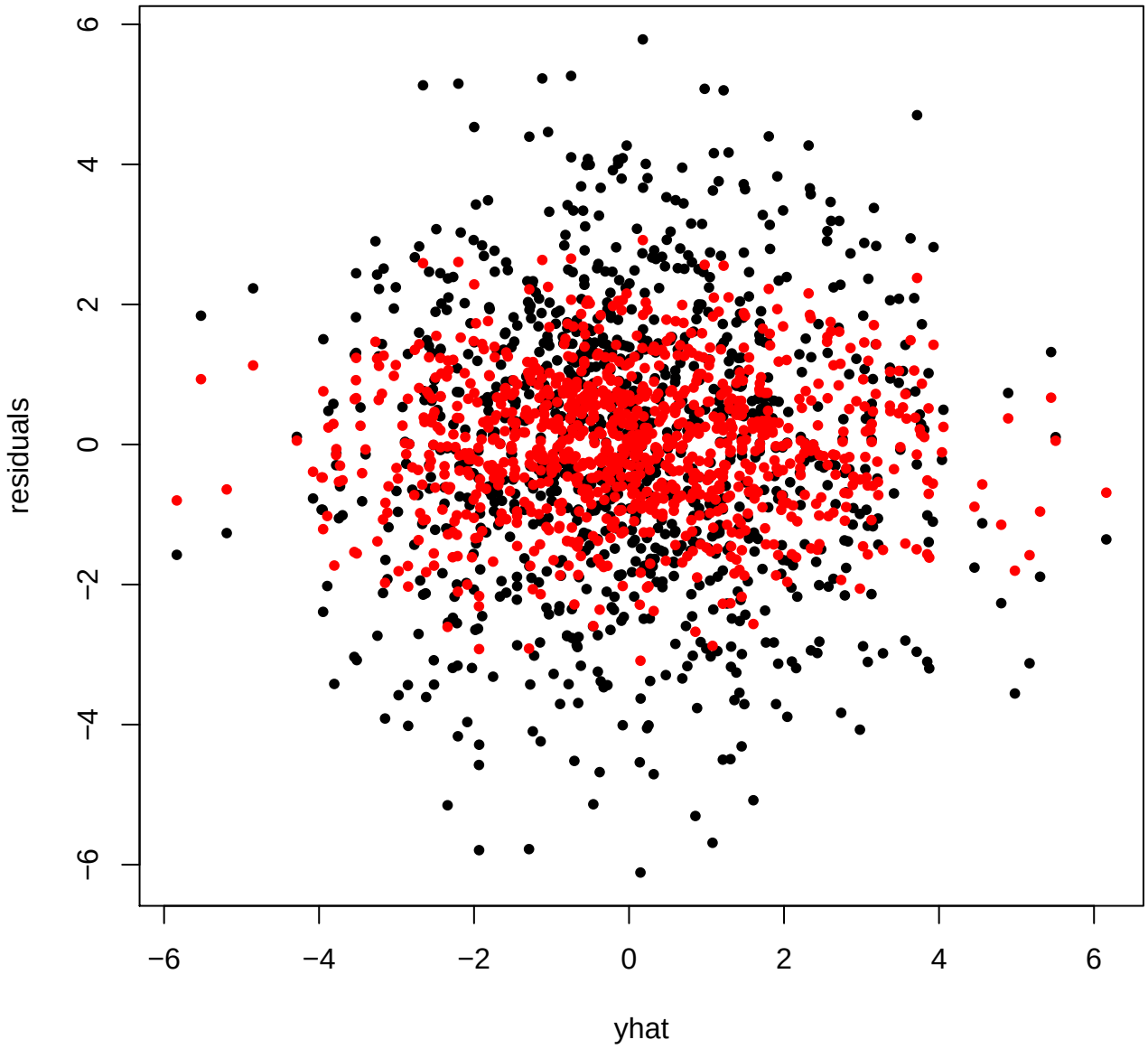
The studentized residuals are used to verify model assumptions and to discover outliers (see Sect. 3.4.4).

Box from our text book: Fahrmeir et al (2013): Regression. Springer. (p.126)

Simulating data and checking residuals

```
n=1000
beta=matrix(c(0,1,1/2,1/3),ncol=1)
set.seed(123)
x1=rnorm(n,0,1); x2=rnorm(n,0,2); x3=rnorm(n,0,3)
X=cbind(rep(1,n),x1,x2,x3)

y=X%*%beta+rnorm(n,0,2)
fit=lm(y~x1+x2+x3)
yhat=predict(fit)
summary(fit)
ehat=residuals(fit); estand=rstandard(fit); estud=rstudent(fit)
plot(yhat,ehat,pch=20)
points(yhat,estand,pch=20,col=2)
#points(yhat,estud,pch=20,col=5)
```



Black: raw residuals, red: standardized residuals (identical to studentized here)

Examination of model assumptions

1. Linearity of covariates: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
2. Homoscedastic error variance: $Cov(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$.
3. Uncorrelated errors: $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$.
4. Additivity of errors: $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$
5. Assumption of normality: $\boldsymbol{\varepsilon} \sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I})$

Plotting residuals

1. Plot the residuals, r_i^* against the predicted values, $\hat{y}_i$.
   - Dependence of the residuals on the predicted value: wrong regression model?
   - Nonconstant variance: transformation or weighted least squares is needed?
2. Plot the residuals, r_i^*, against predictor variable or functions of predictor variables. Trend suggest that transformation of the predictors or more terms are needed in the regression.
3. Assessing normality of errors: QQ-plots and histograms of residuals. As an additional aid a test for normality can be used, but must be interpreted with caution since for small sample sizes the test is not very powerful and for large sample sizes even very small deviances from normality will be labelled as significant.
4. Plot the residuals, r_i^*, versus time or collection order (if possible). Look for dependence or autocorrelation.

Volume of a tree

Data for 31 trees of a certain kind in a national park in the US are given below. Three variables are measured for each tree. These are:

- D: The diameter of the tree measured in inches 1.5 m above ground level
- H: The height of the tree measured in feet.
- V: The volume of the tree measured in cubic feet.

Obs.	D	H	V	Obs.	D	H	V
1	8.3	70	10.3	17	12.9	85	33.8
2	8.6	65	10.3	18	13.3	86	27.4
3	8.8	63	10.2	19	13.7	71	25.7
4	10.5	72	16.4	20	13.8	64	24.9
5	10.7	81	18.8	21	14.0	78	34.5
6	10.8	83	19.7	22	14.2	80	31.7
7	11.0	66	15.6	23	14.5	74	36.3
8	11.0	75	18.2	24	16.0	72	38.3
9	11.1	80	22.6	25	16.3	77	42.6
10	11.2	75	19.9	26	17.3	81	55.4
11	11.3	79	24.2	27	17.5	82	55.7
12	11.4	76	21.0	28	17.9	80	58.3
13	11.4	76	21.4	29	18.0	80	51.5
14	11.7	69	21.3	30	18.0	80	51.0
15	12.0	75	19.1	31	20.6	87	77.0
16	12.9	74	22.2				

Volume of a tree

- If one wants to measure the volume of a tree the tree has to be cut down.
- But, height and diameter can be measured without cutting down the tree.
- Of interest: develop a model that can be used to estimate the tree volume from the height and diameter.

As an illustration assume we want to fit a linear model with V as response and D and H as covariates. What is the R^2 of this model?

Comment: if we start with the volume of a cylinder (area of circle times height) we may suggest a different regression model (on the log scale). Which model?

Volume: height and diameter

```
fit <- lm(Volume~.,data=ds)
summary(fit)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -57.9877     8.6382  -6.713 2.75e-07 ***
Diameter      4.7082     0.2643  17.816  < 2e-16 ***
Height        0.3393     0.1302   2.607   0.0145 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 3.882 on 28 degrees of freedom
Multiple R-squared:  0.948,Adjusted R-squared:  0.9442
F-statistic:   255 on 2 and 28 DF,  p-value: < 2.2e-16
```

Volume of a tree: IQ of lumberjack added

- We want to add the IQ of the lumberjack that cut down the tree as a covariate in the model.
- This should for obvious reasons not be a good predictor for the volume of the tree.
- To mimic this situation we simulate new data to resemble the IQ of different lumberjacks by drawing data from the normal distribution with mean 100 and standard deviation 16, and since we have 31 trees we simulate 31 observations.
- Q: will the R^2 of this new model be higher than the R^2 of the previous model?

Volume: height and diameter – and IQ of lumberjack

```
set.seed(123) # reproducible results
iq <- rnorm(31,100,16)
fit2 <- lm(Volume~Height+Diameter+iq,data=ds)
summary(fit2)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -61.03399   10.20868  -5.979 2.24e-06 ***
Height        0.34099    0.13176   2.588   0.0154 *
Diameter      4.72507    0.26906  17.561 2.68e-16 ***
iq            0.02704    0.04678   0.578   0.5681
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '

Residual standard error: 3.929 on 27 degrees of freedom
Multiple R-squared:  0.9486,Adjusted R-squared:  0.9429
F-statistic: 166.1 on 3 and 27 DF,  p-value: < 2.2e-16
```

Acid rain in Norwegian lakes

Data on $n = 26$ lakes, with

- y: measured pH in lake,
- x1: SO_4: sulfate (the salt of sulfuric acid),
- x2: $N0_3$: nitrate (the conjugate base of nitric acid),
- x3: Ca: calsium,
- x4: latent Al: aluminium,
- x5: organic substance,
- x6: area of lake,
- x7: position of lake (Telemark or Trøndelag),

We would like to use a regression model with pH of the lake as the response. Should we fit a model will all 7 covariates, or choose a subset?

Simulated data (Fahrmeir et al: Fig 3.17)

True model:

$$Y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i$$

Known that the model is polynomial in nature, but not up to which degree.
Try to fit polynomial also with higher order terms.

New: in addition to the data set to be used to fit the regression (called *training set*) also a data set to assess the model fit is present (called a *validation* set).

Mean Squared Error (MSE) is a scaled version of the SSE, that is $\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$.

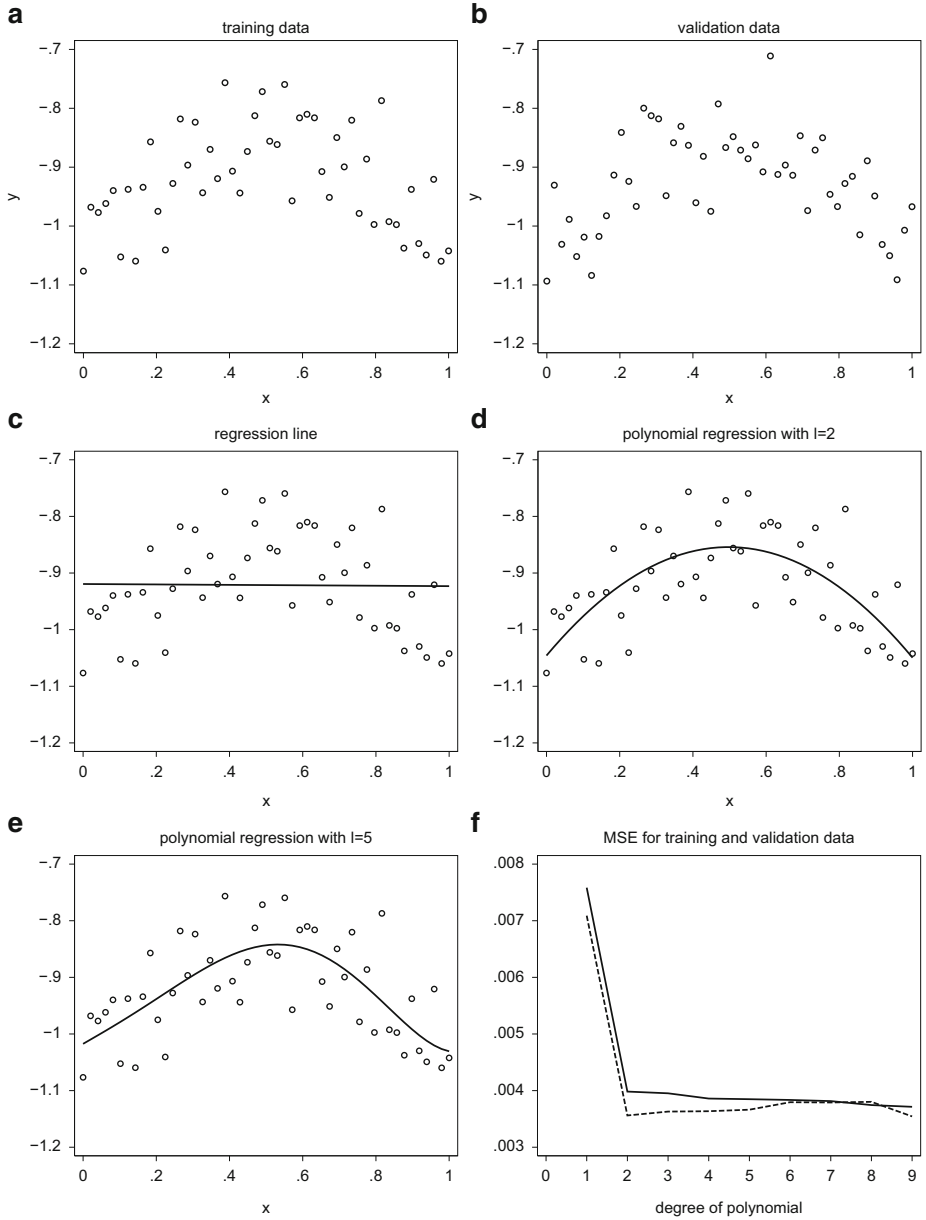


Fig. 3.17 Simulated training data y_i [panel (**a**)] and validation data y_i^* [panel (**b**)] based on 50 design points $x_i, i = 1, \ldots, 50$. The true model used for simulation is $y_i = -1 + 0.3x_i + 0.4x_i^2 - 0.8x_i^3 + \varepsilon_i$ with $\varepsilon_i \sim \mathrm{N}(0, 0.07^2)$. Panels (**c**–**e**) show estimated polynomials of degree $l = 1, 2, 5$ based on the training set. Panel (**f**) displays the mean squared error MSE(l) of the fitted values in relation to the polynomial degree (*solid line*). The *dashed line* shows MSE(l), if the estimated polynomials are used to predict the validation data y_i^*

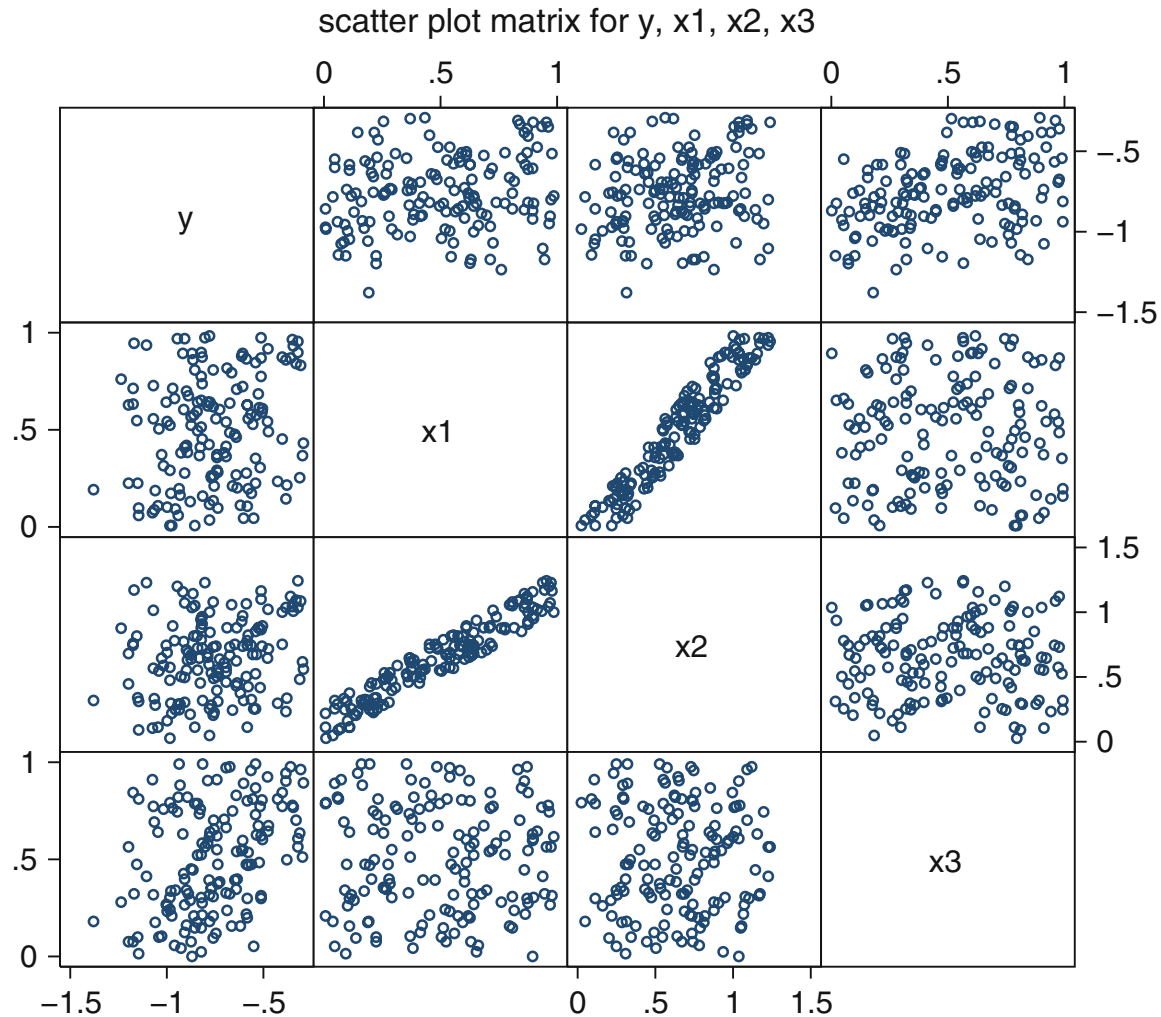
Figure from our text book: Fahrmeir et al (2013): Regression. Springer. (p.140)

Simulated data (Fahrmeir et al: Fig 3.18, Tab3.3, Tab3.4)

True model:

$$Y \sim N(-1 + 0.3x_1 + 0.2x_3, 0.2^2)$$

where also $x_2 = x_1 + u$ is observed ($u \sim$ uniform in 0,1). The variables x_1 and x_3 are uncorrelated.



ig. 3.18 Scatter plot matrix for the variables y, x_1, x_2, and x_3

Figure from our text book: Fahrmeir et al (2013): Regression. Springer. (p.141)

Table 3.3 Results for the model based on covariates x_1, x_2, and x_3

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	−0.970	0.047	−20.46	<0.001	−1.064	−0.877
x_1	0.146	0.187	0.78	0.436	−0.224	0.516
x_2	0.027	0.177	0.15	0.880	−0.323	0.377
x_3	0.227	0.052	4.32	<0.001	0.123	0.331

Table 3.4 Results for the correctly specified model based on covariates x_1 and x_3

Variable	Coefficient	Standard error	t-value	p-value	95 % Confidence interval	
intercept	−0.967	0.039	−24.91	<0.001	−1.042	−0.889
x_1	0.173	0.055	3.17	0.002	0.065	0.281
x_3	0.226	0.052	4.33	<0.001	0.123	0.330

Table from our text book: Fahrmeir et al (2013): Regression. Springer. (p.142)

Irrelevant and/or missing covariates in the regression

Irrelevant : variables that are included in the regression but should not have been.

missing : variables that are not included, but should have been.

Two subsets of covariates (Exam V2014 Problem 4b)

Classical linear model with identically normally distributed random errors, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I}$, but now look at misspecification of $\mathrm{E}(\boldsymbol{Y})$. Suppose that the true model is

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}_1 \boldsymbol{\beta}_1 + \boldsymbol{X}_2 \boldsymbol{\beta}_2 + \boldsymbol{\varepsilon}, \\ \boldsymbol{\varepsilon} &\sim N_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}), \end{aligned} \tag{1}$$

where we have partitioned the design matrix into two parts $\boldsymbol{X}_1$ ($n \times p_1$) and $\boldsymbol{X}_2$ ($n \times p_2$) and $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are unknown p_1- and p_2-dimensional vectors of regression coefficients ($p = p_1 + p_2$).

Two subsets of covariates (cont.)

Assume that we ignore the covariates in $\boldsymbol{X}_2$ and fit the model

$$\begin{aligned} \boldsymbol{Y} &= \boldsymbol{X}_1\boldsymbol{\alpha}_1 + \boldsymbol{\delta}, \\ \boldsymbol{\delta} &\sim N_n(\boldsymbol{0}, \tau^2\boldsymbol{I}). \end{aligned} \tag{2}$$

Here $\boldsymbol{\alpha}_1$ is used in place of $\boldsymbol{\beta}_1$ to emphasize that $\boldsymbol{\alpha}_1$ (and estimates thereof) will in general be different from $\boldsymbol{\beta}_1$ in the true model.
The least squares estimator for model (2) is
$\hat{\boldsymbol{\alpha}}_1 = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{Y}$.

Two subsets of covariates (cont.)

Find the expected value and covariance matrix of $\hat{\boldsymbol{\alpha}}_1$ under the true model.

$$\mathrm{E}(\hat{\boldsymbol{\alpha}}_1) = \boldsymbol{\beta}_1 + (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2 \boldsymbol{\beta}_2$$

We see that the bias term for $\hat{\boldsymbol{\alpha}}_1$ is $(\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1} \boldsymbol{X}_1^T \boldsymbol{X}_2 \boldsymbol{\beta}_2$. When is the bias term equal to zero?

$$\mathrm{Cov}(\hat{\boldsymbol{\alpha}}_1) = \sigma^2 (\boldsymbol{X}_1^T \boldsymbol{X}_1)^{-1}$$

Observe, $\mathrm{Cov}(\hat{\boldsymbol{\alpha}}_1)$ is not dependent on $\boldsymbol{\beta}_2$.

Missing covariates: findings

Bias : The estimator for the (true) covariates (in the model) is only unbiased if the true and missing covariates are uncorrelated (orthogonal design) in the data.

Variance : The variance of the estimator for the true covariates may be smaller based on the model with the missing covariates (than for the correctly specified model), and even the sum of the bias^2 and the variance may better for the model with the missing variables. So the sparse model may be better on overall (even though it is biased).

Irrelevant covariates included: findings

Bias : The estimator for the true covariates are unbiased, also if irrelevant covariates are included.

Variance : The model with the irrelevant covariants have larger variance for the true covariates, compared with the model without the irrelevant covariates. So, again sparse model is the best.

Irrelevant and/or missing covariates in the regression

Irrelevant : variables that are included in the regression but should not have been.

missing : variables that are not included, but should have been.

Conclusion in book: the model should not contain irrelevant covariates, and we should aim for a sparse model.

Law of parsimony

If two models are not very different – then always choose the simplest one

Today

- T-test for significance of one regression coefficient.
- Residuals: standardized (or studentized) preferred.
- Significance of regression based on F-test with SSR/(p-1) divided by SST/(n-1).
- R^2 gives the proportion of variability explained by the regression model.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}}$$

  and will never decrease if new covariates are added to the model.
- Model selection: want to choose the model that minimize the expected squared prediction error.