

STK2100- Machine Learning and Statistical Methods for Prediction and Classification

Spring 2025

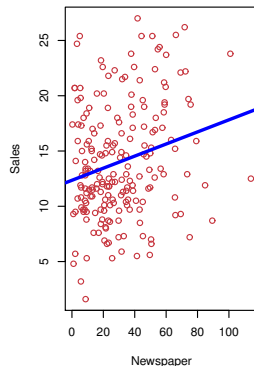
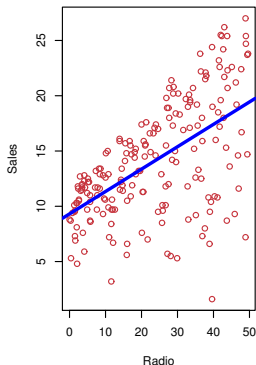
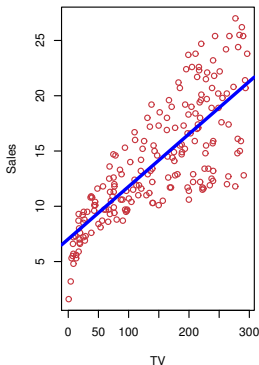
Linear models for regression

Covers Sections 3.1-3.3

Ingrid Hobæk Haff
Department of Mathematics
University of Oslo

Example: Advertising data

- Response: Sale of a product in 200 different market (sales)
- Explanatory variables
 - Advertisement budget in TV (TV)
 - Advertisement budget in radio (radio)
 - Advertisement budget in newspapers (newspaper)



Example: Advertising data

- Is there a relationship between advertisement and sales?
- How strong is this relationship?
- Is the relationship linear?
- Which is the medium that has the strongest influence on sales?
- How precisely can we estimate the effect of the covariates?
- How precisely can we predict future sales?
- Is there some synergy/interaction between different media?

Linear regression

- Data $(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)$
- Linear regression model:

$$Y_i = \beta_0 + \beta_1 x_{i,1} + \dots + \beta_p x_{i,p} + \varepsilon_i, \quad i = 1, \dots, N$$

$\varepsilon_1, \dots, \varepsilon_N$ independent with $E(\varepsilon_i) = 0$, $\text{Var}(\varepsilon_i) = \sigma^2$ (*)

and on matrix form:

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

with

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix}, \mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & & & & \\ 1 & x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_N \end{pmatrix}$$

Linear regression

- The **least squares estimate** $\hat{\beta}$ of β is obtained by minimising the squared error loss, called the residual sum of squares in this context

$$RSS(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^T(\mathbf{Y} - \mathbf{X}\beta) = \mathbf{Y}^T\mathbf{Y} - 2\mathbf{Y}^T\mathbf{X}\beta + \beta^T\mathbf{X}^T\mathbf{X}\beta$$

- Recall that if $\mathbf{a}$ and $\mathbf{b}$ are $p \times 1$ vectors and $\mathbf{A}$ a $p \times p$ matrix, then

$$\frac{\partial \mathbf{a}^T \mathbf{b}}{\partial \mathbf{b}} = \mathbf{a} \quad \text{and} \quad \frac{\partial \mathbf{b}^T \mathbf{A} \mathbf{b}}{\partial \mathbf{b}} = (\mathbf{A} + \mathbf{A}^T) \mathbf{b}$$

- We get

$$\frac{\partial \mathbf{Y}^T \mathbf{X} \beta}{\partial \beta} = \mathbf{X}^T \mathbf{Y} \quad \text{and} \quad \frac{\partial \beta^T \mathbf{X}^T \mathbf{X} \beta}{\partial \beta} = 2\mathbf{X}^T \mathbf{X} \beta,$$

so that

$$\frac{RSS(\beta)}{\partial \beta} = -2\mathbf{X}^T \mathbf{Y} + 2\mathbf{X}^T \mathbf{X} \beta,$$

Linear regression

- Solving $\frac{RSS(\beta)}{\partial \beta} = 0$, we obtain the least squares estimate

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

as long as $\mathbf{X}$ has full rank, so that $\mathbf{X}^T \mathbf{X}$ is invertible

- If $\mathbf{X}$ does not have full rank, the inverse of $\mathbf{X}^T \mathbf{X}$ is replaced by a generalised inverse $(\mathbf{X}^T \mathbf{X})^-$
- The generalised inverse is not unique, so neither is the resulting $\hat{\beta}$, but all possibilities give the same $\hat{f}(\mathbf{x})$
- Note that if $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, the least squares estimate of β is also the maximum likelihood estimate of β .
- The prediction of Y for a new point $\mathbf{x}^* = (x_1^*, \dots, x_p^*)$ is given by

$$\hat{y}^* = \mathbf{x}^* \hat{\beta} = \hat{\beta}_0 + \hat{\beta}_1 x_1^* + \dots + \hat{\beta}_p x_p^*$$

Random vectors

- If $Z_1, \dots, Z_p$ are random variables, we say $\mathbf{Z} = (Z_1, \dots, Z_p)^T$ is a **random vector**
- We define the **expectation vector** and **covariance matrix** by

$$\mathbf{E}(\mathbf{Z}) = \begin{pmatrix} \mathbf{E}(Z_1) \\ \mathbf{E}(Z_2) \\ \vdots \\ \mathbf{E}(Z_p) \end{pmatrix}, \quad \text{Var}(\mathbf{Z}) = \begin{pmatrix} \text{Var}(Z_1) & \text{Cov}(Z_1, Z_2) & \cdots & \text{Cov}(Z_1, Z_p) \\ \text{Cov}(Z_2, Z_1) & \text{Var}(Z_2) & \cdots & \text{Cov}(Z_2, Z_p) \\ \vdots & \vdots & \ddots & \vdots \\ \text{Cov}(Z_p, Z_1) & \text{Cov}(Z_p, Z_2) & \cdots & \text{Var}(Z_p) \end{pmatrix}$$

- Further, we have the rules

$$\mathbf{E}(\mathbf{AZ} + \mathbf{b}) = \mathbf{A}\mathbf{E}(\mathbf{Z}) + \mathbf{b},$$

$$\text{Var}(\mathbf{AZ} + \mathbf{b}) = \mathbf{A}\text{Var}(\mathbf{Z})\mathbf{A}^T$$

, where $\mathbf{A}$ and $\mathbf{b}$ are constant

Properties of the linear regression model

- If the linear regression model (*) is true, so that $E(\varepsilon) = \mathbf{0}$ and $\text{Var}(\varepsilon) = \sigma^2 \mathbf{I}$, we have

$$E(\mathbf{Y}|\mathbf{X}) = E(\mathbf{X}\beta + \varepsilon) = \mathbf{X}\beta \quad \text{and} \quad \text{Var}(\mathbf{Y}|\mathbf{X}) = \text{Var}(\mathbf{X}\beta + \varepsilon) = \sigma^2 \mathbf{I}$$

- This gives (conditional on $\mathbf{X}$):

$$\begin{aligned} E(\hat{\beta}|\mathbf{X}) &= E\left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}|\mathbf{X}\right) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T E(\mathbf{Y}|\mathbf{X}) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{X} \beta = \beta \\ \text{Var}(\hat{\beta}|\mathbf{X}) &= \text{Var}\left((\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}|\mathbf{X}\right) \\ &= (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \sigma^2 \mathbf{I} \mathbf{X} (\mathbf{X}^T \mathbf{X})^{-1} = \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1} \end{aligned}$$

Geometric interpretation of least squares regression

- Let $\hat{\mathbf{Y}} = (\hat{Y}_1, \dots, \hat{Y}_N)^T$
- Then, we have

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

- The hat matrix $\mathbf{H}$ is symmetric and self-similar, i.e.

$$\mathbf{H}^2 = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T = \mathbf{H}$$

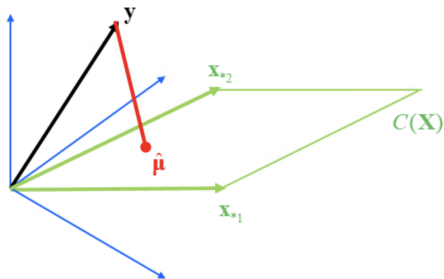
- Further, since $\mathbf{Y} - \hat{\mathbf{Y}} = (\mathbf{I} - \mathbf{H})\mathbf{Y}$, we have

$$(\mathbf{Y} - \hat{\mathbf{Y}})^T \hat{\mathbf{Y}} = \mathbf{Y}^T (\mathbf{I} - \mathbf{H}) \mathbf{H} \mathbf{Y} = \mathbf{Y}^T \mathbf{H} \mathbf{Y} - \mathbf{Y}^T \mathbf{H}^2 \mathbf{Y} = \mathbf{Y}^T \mathbf{H} \mathbf{Y} - \mathbf{Y}^T \mathbf{H} \mathbf{Y} = 0$$

Geometric interpretation of least squares regression

This means that $\mathbf{H}$ is a projection matrix, projecting $\mathbf{Y}$ onto the column space of $\mathbf{X}$

Illustration
for $n = 3$
and $p = 2$



Properties of the linear regression model - continued

- The variance σ^2 is typically estimated by

$$\hat{\sigma}^2 = S^2 = \frac{1}{N-p-1} \sum_{i=1}^N (y_i - \mathbf{x}_i^T \hat{\beta})^2 = \frac{RSS(\hat{\beta})}{N-p-1},$$

which is unbiased

- When $(*)$ is the true model, and in addition, $\varepsilon_i \stackrel{iid}{\sim} N(0, \sigma^2)$, then

$$\hat{\beta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y} \sim N\left(\beta, \sigma^2 (\mathbf{X}^T \mathbf{X})^{-1}\right)$$

- Further, $(N-p-1) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{N-p-1}^2$
- Thus

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{v_j}} \sim t_{N-p-1},$$

where v_j is the j th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$

Hypothesis tests

- A hypothesis test for a null hypothesis of the type $H_0 : \beta_j = 0$ (against the alternative $H_A : \beta_j \neq 0$) can be done with the Wald test statistic

$$T = \frac{\hat{\beta}_j}{\hat{\sigma} \sqrt{V_j}},$$

which is t_{N-p-1} -distributed under H_0

- Assume that we want to test the hypothesis $H_0 : \beta_{i_1} = \dots = \beta_{i_q} = 0$ against $H_A : \text{at least one } \beta_{i_j} \neq 0$, with $\{i_1, \dots, i_q\} \subseteq \{1, \dots, p\}$
- Let $RSS_0 = \sum_{i=1}^n (y_i - \mathbf{x}_i^T \tilde{\beta})^2$, where $\tilde{\beta}$ is the least squares estimator of β under H_0 .
- Then, test the hypothesis with the statistic

$$F = \frac{(RSS_0 - RSS)/q}{RSS/(N - p - 1)},$$

which is $F_{q, N-p-1}$ -distributed under H_0

Hypothesis tests

- Recall that the **P-value** is the probability, calculated assuming that the null hypothesis is true, of obtaining a value of the test statistic that contradicts H_0 at least as much the value from the available sample
- For t -statistics:
$$\text{P-value} = 2 \Pr(T > |t_{obs}| | H_0) = 2(1 - \Pr(T \leq |t_{obs}| | H_0))$$
- For F -statistics:
$$\text{P-value} = \Pr(F > F_{obs} | H_0) = 1 - \Pr(F \leq F_{obs} | H_0)$$
- Small P-values indicate that H_0 is false
- Typical limits to indicate significance are 0.05 or 0.01

Example: Advertising data

See R code

Confidence and prediction intervals

- Recall that under the assumption that $\varepsilon_1, \dots, \varepsilon_N \stackrel{iid}{\sim} N(\mu, \sigma^2)$

$$\frac{\hat{\beta}_j - \beta_j}{\hat{\sigma} \sqrt{v_j}} \sim t_{N-p-1},$$

where v_j is the j th diagonal element of $(\mathbf{X}^T \mathbf{X})^{-1}$

- Then, a $100 \cdot (1 - \alpha)\%$ **confidence interval for β_j** is given by

$$\hat{\beta}_j \pm t_{1-\alpha/2; N-p-1} \hat{\sigma} \sqrt{v_j},$$

where $t_{1-\alpha/2; n-p-1}$ is the $(1 - \alpha/2)$ -percentile of the t_{N-p-1} distribution

- Assume that we are interested in a $100 \cdot (1 - \alpha)\%$ confidence interval for $E(Y|\mathbf{X} = \mathbf{x})$
- We estimate $E(Y|\mathbf{X} = \mathbf{x})$ by $\hat{Y} = \mathbf{x}^T \hat{\beta}$, and we have

$$E(\hat{Y}) = E(\mathbf{x}^T \hat{\beta}) = \mathbf{x}^T \beta = E(Y|\mathbf{X} = \mathbf{x})$$

and

$$\text{Var}(\hat{Y}) = \text{Var}(\mathbf{x}^T \hat{\beta}) = \mathbf{x}^T \text{Var}(\hat{\beta}) \mathbf{x} = \sigma^2 \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}$$

Confidence and prediction intervals

- Since $\hat{\beta}$ is normal distributed, so is $\hat{Y} = \mathbf{x}^T \hat{\beta}$
- Thus we have

$$\frac{\hat{Y} - E(Y|\mathbf{X} = \mathbf{x})}{\hat{\sigma} \sqrt{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}} \sim t_{N-p-1}$$

- Hence, a $100 \cdot (1 - \alpha)\%$ **confidence interval for $E(Y|\mathbf{X} = \mathbf{x})$** is given by

$$\hat{y} \pm t_{1-\alpha/2; N-p-1} \hat{\sigma} \sqrt{\mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}$$

- Assume now that we are interested in a prediction interval for Y with corresponding covariate vector $\mathbf{x}$, so that the prediction is $\hat{Y} = \mathbf{x}^T \hat{\beta}$
- Since both Y and $\hat{Y}$ are normal distributed, so is $Y - \hat{Y}$
- Further, we have

$$E(Y - \hat{Y}) = E(Y) - \mathbf{x}^T E(\hat{\beta}) = \mathbf{x}^T \beta - \mathbf{x}^T \beta = 0$$

Confidence and prediction intervals

- Since $\hat{Y}$ depends on $Y_1, \dots, Y_N$, that are independent of Y , we have

$$\begin{aligned}\text{Var}(Y - \hat{Y}) &= \text{Var}(Y) + (-1)^2 \text{Var}(\hat{Y}) \\ &= \sigma^2 + \sigma^2 \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x} \\ &= \sigma^2 (1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x})\end{aligned}$$

- Thus we have

$$\frac{Y - \hat{Y}}{\hat{\sigma} \sqrt{1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}} \sim t_{N-p-1}$$

- Hence, a $100 \cdot (1 - \alpha)\%$ **prediction interval for Y** is given by

$$\hat{y} \pm t_{1-\alpha/2; N-p-1} \hat{\sigma} \sqrt{1 + \mathbf{x}^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}}$$

Example: Advertising data

See R code

Gauss-Markov Theorem

Theorem (Gauss-Markov)

Let $\hat{\beta}$ be the least squares estimator of β . Among all unbiased, linear estimators $\tilde{\theta} = \mathbf{c}^T \mathbf{y}$ of $\theta = \mathbf{a}^T \beta$, $\mathbf{a}^T \hat{\beta}$ has the lowest variance.

Proof.

Exercise 3.3



- Note that the MSE of $\tilde{\theta}$ can be split into the squared bias and the variance as

$$\text{MSE}(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 = \text{Var}(\tilde{\theta}) + (E(\tilde{\theta}) - \theta)^2$$

- If we are willing to accept some bias, we might find an estimator with a lower MSE, due to a lower variance
- We might also find better nonlinear estimators

Qualitative covariates

- **Qualitative**, or categorical covariates in a regression model should be handled by constructing binary dummy variables
- More specifically, one specifies $k - 1$ dummy variables for a categorical covariate with k categories
- One of the categories is chosen as a reference, and corresponds to 0s in all the dummy variables
- The remaining categories are encoded by a 1 in the corresponding dummy variable and 0 in all the others
- In R, this is done with the command `factor()`

Qualitative covariates

- For a categorical covariate with two categories, one defines

$$x_i = \begin{cases} 1 & \text{if observation } i \text{ is in category 2 (not the reference)} \\ 0 & \text{if observation } i \text{ is not in category 2} \end{cases}$$

- For a categorical covariate with three categories, one defines

$$x_{i1} = \begin{cases} 1 & \text{if observation } i \text{ is in category 2} \\ 0 & \text{if observation } i \text{ is not in category 2} \end{cases}$$

$$x_{i2} = \begin{cases} 1 & \text{if observation } i \text{ is in category 3} \\ 0 & \text{if observation } i \text{ is not in category 3} \end{cases}$$

Example: Credit data set

See R code

What is linearity?

- A linear regression model may contain interaction effects, such as

$$Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \varepsilon$$

or non-linear terms in some of the x_j s, for instance polynomial terms, such as

$$Y = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_q x^q + \varepsilon$$

- The above models are not linear in the x_j s
- However, the model is linear in the β_j s
- Theory about linear regression requires **linearity in the β_j s**

Example: Advertising data

See R code

What if p is large?

Example: Hitters data

Consists of the response `Salary`, and $p = 19$ explanatory variables

- Should all explanatory variables be kept in the model?
- Arguments for considering models with fewer covariates:
 - Prediction accuracy: With p large, the variability in the estimates of the $\hat{\beta}_j$ s can be large, resulting in high variance in $\hat{y}^* = (x^*)^T \hat{\beta}$
 - Interpretation: Models with fewer covariates are easier to understand
- This could be done with
 - Variable/subset selection
 - Shrinkage
 - Dimension reduction

Variable selection

- The number of possible subsets of covariates grows fast with p
 - $p = 3$ gives $2^3 = 8$ possible models
 - $p = 30$ gives $2^{30} = 1\,073\,741\,824$ possible models!
- Hence, unless p is small, a full search through all possible subsets is computationally too demanding
- Further, a search through all subsets costs more in terms of variance than a more restrained search, that might have larger bias
- One alternative is **forward selection**
 - Start with the null model $Y = \beta_0 + \varepsilon$
 - Add the variable that gives the largest improvement according to a criterion of choice
 - Continue as long as you obtain a significant improvement

Variable selection

- Another option is **backward selection**, which works similarly, except that one starts with the full model $Y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p + \varepsilon$, and one removes one variable at a time
- Yet another possibility is to combine forward and backward selection in a **hybrid approach**, switching between going backwards and forwards

Measures of performance

- A common measure of performance for linear regression models is the R^2 , defined by

$$R^2 = \frac{TSS - RSS}{TSS}, \quad TSS = \sum_{i=1}^n (y_i - \bar{y})^2$$
$$= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$$

- We have $0 \leq R^2 \leq 1$, and R^2 close to 1 indicates good performance
- However, R^2 will always increase with larger models
- This is to some extent mended in the adjusted R^2 , given by

$$R_{Adj}^2 = 1 - \frac{RSS/(n - p - 1)}{TSS/(n - 1)}$$

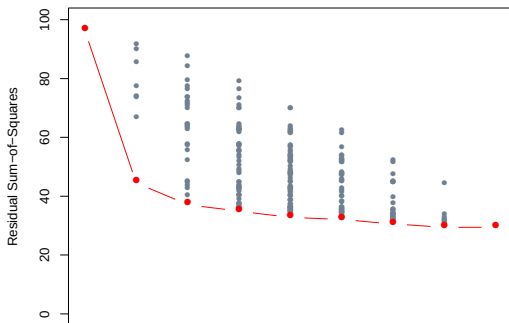
Measures of performance

- P-values from hypothesis tests, such as the t- and F-tests presented earlier, may also be used as measures of performance
- However, this results in multiple testing, which should be corrected for
- Other popular alternatives are model selection criteria, such as AIC, BIC or Mallow's C_p , which we will encounter in Chapter 7

Examples

Example: Prostate data

See R code



Example: Hitters data

See R code