Statistics in Geophysics: Linear Regression II

Steffen Unkel

Department of Statistics
Ludwig-Maximilians-University Munich, Germany

Model definition

- Suppose we have the following model under consideration:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \ ,$$

where $\mathbf{y} = (y_1 \ldots, y_n)^\top$ is an $n \times 1$ vector of observations on the response, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)^\top$ is a $(k+1) \times 1$ vector of parameters, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^\top$ is an $n \times 1$ vector of random errors, and $\mathbf{X}$ is the $n \times (k+1)$ design matrix with

$$\mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ \vdots & \vdots & & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \ .$$

Model assumptions

The following assumptions are made:

1. $\mathrm{E}(\boldsymbol{\epsilon}) = \mathbf{0}$.
2. $\mathrm{Cov}(\boldsymbol{\epsilon}) = \sigma^2 \mathbf{I}$.
3. The design matrix has full column rank, that is, $\mathrm{rank}(\mathbf{X}) = k + 1 = p$.
4. The normal regression model is obtained if additionally $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$.

For stochastic covariates these assumptions are to be understood conditionally on $\mathbf{X}$.
It follows that $\mathrm{E}(\mathbf{y}) = \mathbf{X}\beta$ and $\mathrm{Cov}(\mathbf{y}) = \sigma^2 \mathbf{I}$. In case of assumption 4, we have $\mathbf{y} \sim \mathcal{N}(\mathbf{X}\beta, \sigma^2 \mathbf{I})$.

Modelling the effects of continuous covariates

- We can fit nonlinear relationships between continuous covariates and the response within the scope of linear models.

- Two simple methods for dealing with nonlinearity:
  1. Variable transformation;
  2. Polynomial regression.

- Sometimes it is customary to transform the continuous response as well.

Modelling the effects of categorical covariates

- We might want to include a categorical variable with two or more distinct levels.
- In such a case we cannot set up a continuous scale for the categorical variable.
- A remedy is to define new covariates, so-called dummy variables, and estimate a separate effect for each category of the original covariate.
- We can deal with c categories by the introduction of $c - 1$ dummy variables.

Example: Turkey data

	weightp	agew	origin
1	13.3	28	G
2	8.9	20	G
3	15.1	32	G
4	10.4	22	G
5	13.1	29	V
6	12.4	27	V
7	13.2	28	V
8	11.8	26	V
9	11.5	21	W
10	14.2	27	W
11	15.4	29	W
12	13.1	23	W
13	13.8	25	W

Turkey weights in pounds, ages in weeks, and origin (Georgia (G); Virgina (V); Wisconsin (W)), of 13 Thanksgiving turkeys.

Dummy coding for categorical covariates

Given a covariate $x \in \{1, \ldots, c\}$ with c categories,

- we define the $c-1$ dummy variables

$$x_{i1} = \begin{cases} 1 & x_i = 1 \\ 0 & \text{otherwise} \end{cases}, \qquad \ldots \qquad x_{i,c-1} = \begin{cases} 1 & x_i = c-1 \\ 0 & \text{otherwise} \end{cases},$$

  for $i = 1 \ldots, n$, and include them as explanatory variables in the regression model

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_{i,c-1} x_{i,c-1} + \ldots + \epsilon_i \quad .$$

- For reasons of identifiability, we omit the dummy variable for category c, where c is the reference category.

Design matrix for the turkey data using dummy coding

```
> X
   (Intercept) agew originV originW
1            1   28       0       0
2            1   20       0       0
3            1   32       0       0
4            1   22       0       0
5            1   29       1       0
6            1   27       1       0
7            1   28       1       0
8            1   26       1       0
9            1   21       0       1
10           1   27       0       1
11           1   29       0       1
12           1   23       0       1
13           1   25       0       1
```

Interactions between covariates

- An interaction between predictor variables exists if the effect of a covariate depends on the value of at least one other covariate.

- Consider the following model between a response y and two covariates x_1 and x_2:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \epsilon \ ,$$

  where the term $\beta_3 x_1 x_2$ is called an interaction between x_1 and x_2.

- The terms $\beta_1 x_1$ and $\beta_2 x_2$ depend on only one variable and are called main effects.

Least squares estimation of regression coefficients

- The error sum of squares is

$$\begin{aligned}\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon} &= (\mathbf{y}-\mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) \\ &= \mathbf{y}^\top\mathbf{y}-\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}-\mathbf{y}^\top\mathbf{X}\boldsymbol{\beta}+\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta} \\ &= \mathbf{y}^\top\mathbf{y}-2\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{y}+\boldsymbol{\beta}^\top\mathbf{X}^\top\mathbf{X}\boldsymbol{\beta}\ .\end{aligned}$$

- The least squares estimator of $\boldsymbol{\beta}$ is the value $\hat{\boldsymbol{\beta}}$, which minimizes $\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$.

- Minimizing $\boldsymbol{\epsilon}^\top\boldsymbol{\epsilon}$ with respect to $\boldsymbol{\beta}$ yields

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y}\ ,$$

which is obtained irrespective of any distribution properties of the errors.

Maximum likelihood estimation of regression coefficients

- Assuming normally distributed errors yields the likelihood

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left(-\frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})\right).$$

- The log-likelihood is given by

$$l(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}).$$

- Maximizing $l(\boldsymbol{\beta}, \sigma^2)$ with respect to $\boldsymbol{\beta}$ is equivalent to minimizing $(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^\top(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})$, which is the least squares criterion.

- The MLE, $\hat{\boldsymbol{\beta}}_{ML}$, therefore is identical to the least squares estimator.

Fitted values and residuals

- Based on $\hat{\beta} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$, we can estimate the (conditional) mean of $\mathbf{y}$ by

$$\widehat{\mathrm{E}(\mathbf{y})} = \hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \ .$$

- Substituting the least squares estimator further results in

$$\hat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} = \mathbf{H}\mathbf{y} \ ,$$

  where the $n \times n$ matrix $\mathbf{H}$ is called the hat matrix.

- The residuals are

$$\hat{\epsilon} = \mathbf{y} - \hat{\mathbf{y}} = \mathbf{y} - \mathbf{H}\mathbf{y} = (\mathbf{I} - \mathbf{H})\mathbf{y} \ .$$

Estimation of the error variance

- Maximization of the log-likelihood with respect to σ^2 yields

$$\begin{aligned} \hat{\sigma}^2_{ML} &= \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\beta})^\top(\mathbf{y} - \mathbf{X}\hat{\beta}) \\ &= \frac{1}{n}(\mathbf{y} - \hat{\mathbf{y}})^\top(\mathbf{y} - \hat{\mathbf{y}}) = \frac{1}{n}\,\hat{\boldsymbol{\epsilon}}^\top\hat{\boldsymbol{\epsilon}}\ . \end{aligned}$$

- Since

$$\mathrm{E}(\hat{\sigma}^2_{ML}) = \frac{n-p}{n}\cdot\sigma^2\ \ ,$$

the MLE of σ^2 is biased.

- An unbiased estimator is

$$\hat{\sigma}^2 = \frac{1}{n-p}\hat{\boldsymbol{\epsilon}}^\top\hat{\boldsymbol{\epsilon}}\ \ .$$

Properties of the least squares estimator

For the least squares estimator we have

- $\mathrm{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$.
- $\mathrm{Cov}(\hat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$.
- Gauss-Markov Theorem: Among all linear and unbiased estimators $\hat{\boldsymbol{\beta}}^L$, the least squares estimator has minimal variances, implying

$$\mathrm{Var}(\hat{\beta}_j) \leq \mathrm{Var}(\hat{\beta}_j^L) \quad , \quad j = 0, \ldots, k \quad .$$

- If $\epsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$, then $\hat{\boldsymbol{\beta}} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1})$.

ANOVA table for multiple linear regression and R^2

Source of variation	Degrees of freedom (df)	Sum of squares (SS)	Mean square (MS)	F-value
Regression	k	SSR	$\text{MSR} = \text{SSR}/k$	$\frac{\text{MSR}}{\hat{\sigma}^2}$
Residual	$n - p = n - (k + 1)$	SSE	$\hat{\sigma}^2 = \frac{\text{SSE}}{n-p}$	
Total	$n - 1$	SST		

The multiple coefficient of determination is still computed as $R^2 = \frac{\text{SSR}}{\text{SST}}$, but is no longer the square of the Pearson correlation between the response and any of the predictor variables.

Interval estimation and tests

- We would like to construct confidence intervals and statistical tests for hypotheses regarding the unknown regression parameters $\boldsymbol{\beta}$.
- A requirement for the construction of tests and confidence intervals is the assumption of normally distributed errors.
- However, tests and confidence intervals are relatively robust to mild departures from the normality assumption.
- Moreover, tests and confidence intervals, derived under the assumption of normality, remain valid for large sample size even with non-normal errors.

Testing linear hypotheses

Hypotheses:

1. General linear hypothesis:

$$H_0 : \mathbf{C}\boldsymbol{\beta} = \mathbf{d} \quad \text{against} \quad H_1 : \mathbf{C}\boldsymbol{\beta} \neq \mathbf{d} \ ,$$

   where $\mathbf{C}$ is a $r \times p$ matrix with $\text{rank}(\mathbf{C}) = r \leq p$ (r linear independent restrictions) and $\mathbf{d}$ is a $p \times 1$ vector.

2. Test of significance (t-test):

$$H_0 : \beta_j = 0 \quad \text{against} \quad H_1 : \beta_j \neq 0 \ .$$

Testing linear hypotheses

Hypotheses:

3 Composite test of subvector:

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \quad \text{against} \quad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0} \ .$$

4 Test for significance of regression:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0 \quad \text{against}$$
$$H_1 : \beta_j \neq 0 \text{ for at least one } j \in \{1, \ldots, k\} \ .$$

Testing linear hypotheses

Test statistics:

1. $F = \frac{1}{r}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d})^\top(\hat{\sigma}^2\mathbf{C}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{C}^\top)^{-1}(\mathbf{C}\hat{\boldsymbol{\beta}} - \mathbf{d}) \sim F_{r,n-p}$.
2. $t_j = \frac{\hat{\beta}_j}{\hat{\sigma}_{\hat{\beta}_j}} \sim t_{n-p}$, where $\hat{\sigma}_{\hat{\beta}_j}$ denotes the standard error of $\hat{\beta}_j$.
3. $F = \frac{1}{r}(\hat{\boldsymbol{\beta}}_1)^\top\widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1) \sim F_{r,n-p}$.
4. $F = \frac{n-p}{k}\frac{\mathrm{SSR}}{\mathrm{SSE}} = \frac{n-p}{k}\frac{R^2}{1-R^2} \sim F_{k,n-p}$.

Testing linear hypotheses

Critical values:

Reject H_0 in the case of:

1. $F > F_{r,n-p}(1-\alpha)$.
2. $|t| > t_{n-p}(1-\alpha/2)$.
3. $F > F_{r,n-p}(1-\alpha)$.
4. $F > F_{k,n-p}(1-\alpha)$

Confidence intervals and regions for regression coefficients

Confidence interval for β_j:

- A confidence interval for β_j with level $1-\alpha$ is given by

$$[\hat{\beta}_j - t_{n-p}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}, \hat{\beta}_j + t_{n-p}(1-\alpha/2)\hat{\sigma}_{\hat{\beta}_j}] \ .$$

Confidence region for subvector $\boldsymbol{\beta}_1$:

- A confidence ellipsoid for $\boldsymbol{\beta}_1 = (\beta_1, \ldots, \beta_r)^\top$ with level $1-\alpha$ is given by

$$\left\{ \boldsymbol{\beta}_1 : \frac{1}{r}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1)^\top \widehat{\mathrm{Cov}(\hat{\boldsymbol{\beta}}_1)}^{-1}(\hat{\boldsymbol{\beta}}_1 - \boldsymbol{\beta}_1) \leq F_{r,n-p}(1-\alpha) \right\} \ .$$

Prediction intervals

Confidence interval for the (conditional) mean of a future observation:

- A confidence interval for $E(y_0)$ of a future observation y_0 at location $\mathbf{x}_0$ with level $1-\alpha$ is given by

$$\mathbf{x}_0^\top \hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0)^{1/2} \ .$$

Prediction interval for a future observation:

- A prediction interval for a future observation y_0 at location $\mathbf{x}_0$ with level $1-\alpha$ is given by

$$\mathbf{x}_0^\top \hat{\boldsymbol{\beta}} \pm t_{n-p}(1-\alpha/2)\hat{\sigma}(1+\mathbf{x}_0^\top(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}_0)^{1/2} \ .$$

The corrected coefficient of determination

- We already defined the coefficient of determination, R^2, as a measure for the goodness-of-fit to the data.
- The use of R^2 is limited, since it will never decrease with the addition of a new covariate into the model.
- The corrected coefficient of determination, R^2_{adj}, adjusts for this problem, by including a correction term for the number of parameters.
- It is defined by

$$R^2_{adj} = 1 - \frac{n-1}{n-p}(1 - R^2) \ .$$

Akaike information criterion

- The Akaike information criterion (AIC) is one of the most widely used criteria for model choice within the scope of likelihood-based inference.

- The AIC is defined by

$$\text{AIC} = -2l(\hat{\boldsymbol{\beta}}_{ML}, \sigma^2_{ML}) + 2(p+1) \ ,$$

where $l(\hat{\boldsymbol{\beta}}_{ML}, \sigma^2_{ML})$ is the maximum value of the log-likelihood.

- Smaller values of the AIC correspond to a better model fit.

Bayesian information criterion

- The Bayesian information criterion (BIC) is defined by

$$\mathrm{BIC} = -2l(\hat{\beta}_{ML}, \sigma^2_{ML}) + \log(n)(p+1) \; .$$

- The BIC multiplied by $1/2$ is also known as Schwartz criterion.
- Smaller values of the BIC correspond to a better model fit.
- The BIC penalizes complex models much more than the AIC.

Practical use of model choice criteria

- To select the most promising models from candidate models, we first obtain a preselection of potential models.
- All potential models can now be assessed with the aid of one of the various model choice criteria (AIC, BIC).
- This method is not always practical, since the number of regressor variables and modelling variants can be very large in many applications.
- In this case, we can use the following partially heuristic methods.

Practical use of model choice criteria II

- Complete model selection: In case that the number of predictor variables is not too large, we can determine the best model with the "leaps-and-bounds" algorithm.
- Forward selection:
  1. Based on a starting model, forward selection includes one additional variable in every iteration of the algorithm.
  2. The variable which offers the greatest reduction of a preselected model choice criterion is chosen.
  3. The algorithm terminates if no further reduction is possible.

Practical use of model choice criteria III

- Backward elimination:
  1. Backward elimination starts with the full model containing all predictor variables.
  2. Subsequently, in every iteration, the covariate which provides the greatest reduction of the model choice criterion is eliminated from the model.
  3. The algorithm terminates if no further reduction is possible.
- Stepwise selection: Stepwise selection is a combination of forward selection and backward elimination. In every iteration of the algorithm, a predictor variable may be added to the model or removed from the model.