# Linear Regression Models

## Model Description and Model Parameters

Modelling is a central theme in these notes. The idea is to develop and continuously improve a library of predictive models for hazards, responses, and impacts. Ideally, the models are based on both mechanics and statistics, i.e., both theory and observations. This document focuses on the use of statistics, i.e., observations, to develop models. In doing so it must be carefully noted that a model that matches past observations will not necessarily predict future events. Therefore, all uncertainty, reducible and irreducible, should be candidly recognized. Within the field of linear regression, which is here interpreted to include the Bayesian approach that features random model parameters, it is common to distinguish between single-variable and multiple linear regression models. However, the single-variable model is usually adopted for pedagogical reasons while actual applications have several variables. For that reason, the general linear regression model is directly addressed in this document. It has the form

$$y = \theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \cdots + \theta_k \cdot x_k + \varepsilon \tag{1}$$

where

- $y$ is the response that the model predicts, sometimes called the dependent variable, regressand, response, or output,
- $\theta_i$ are the model parameters, called regression coefficients,
- $x_i$ are the physical measurable independent variables, sometimes called predictor variables, regressors, or explanatory variables, and
- $\varepsilon$ is a random variable that represents the remaining model error.

The model remains linear even if terms like $\theta_2 \cdot (x_2)^{0.5}$ or $\theta_3 \cdot (x_2/x_3)$ appear in the model. However, the model must be linear in the regression coefficients. The first explanatory variable, $x_1$, is routinely set equal to 1 and called the intercept. In fact, removal of the intercept parameter introduces a particular piece of information about $y$. For the subsequent developments, let $\mathbf{x}$ denote the $k$-dimensional vector of explanatory variables. For now, assume that all information comes as $n$ paired observations of $\mathbf{x}$ and $y$. For example, we observe the capacity of a structural component, $y$, along with its material and geometry properties, collected in $\mathbf{x}$. For notational convenience, let all the observations be collected in the $n$-dimensional vector $\mathbf{y}$ and the $n$-by-$k$ dimensional matrix $\mathbf{X}$. In other words, each observation occupies one element in $\mathbf{y}$ and one row in $\mathbf{X}$. In index notation, these are written $y_u$, $u$=1,…,$n$ and $X_{ui}$, $u$=1,…,$n$, $i$=1,…,$k$. By similarly collecting the model parameters in the vector $\boldsymbol{\theta}$ and the discrepancy between the model prediction and the observations in the vector $\boldsymbol{\varepsilon}$ the entire set of observations is contained in the following system of equations:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \boldsymbol{\varepsilon} \tag{2}$$

It is noted that for any one observation, $\varepsilon$ represents the discrepancy between the observed value $y$ and the value predicted by $\theta_1 \cdot x_1 + \theta_2 \cdot x_2 + \ldots + \theta_k \cdot x_k$. Eq. (2) forms the basis for determining the characteristics of $\boldsymbol{\theta}$ and $\boldsymbol{\varepsilon}$, which is the objective of linear regression.

# Inference

### Ordinary Least Squares

In classical linear regression the objective is to obtain point estimates of the model parameters. Although this does not yield a probabilistic model, this methodology provides useful insight and a basis for the Bayesian approach. Specifically, ordinary least squares inference determines a point estimate for $\boldsymbol{\theta}$ by minimizing the sum of squared errors. That is, it minimizes

$$\|\boldsymbol{\varepsilon}\|^2 = \varepsilon_1^2 + \varepsilon_2^2 + \cdots + \varepsilon_n^2 \tag{3}$$

By introducing the linear model in Eq. (2) the problem reads:

$$\hat{\boldsymbol{\theta}} = \arg\min\left(\|\boldsymbol{\varepsilon}\|^2\right) = \arg\min\left(\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2\right) \tag{4}$$

where $\hat{\boldsymbol{\theta}}$ is the sought point estimate of the model parameters. The solution is obtained by setting the derivative of the objective function with respect to $\boldsymbol{\theta}$ equal to $\mathbf{0}$. This differentiation is conveniently carried out in index notation:

$$\begin{aligned}\frac{\partial\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|^2}{\partial\boldsymbol{\theta}} &= \frac{\partial}{\partial\theta_m}\left(y_u - x_{ui}\theta_i\right)\left(y_u - x_{uj}\theta_j\right)\\ &= \frac{\partial}{\partial\theta_m}\left(y_u^2 - x_{ui}\theta_i y_u - y_u x_{uj}\theta_j + x_{ui}\theta_i x_{uj}\theta_j\right)\\ &= 0 - x_{um}y_u - y_u x_{um} + x_{um}x_{uj}\theta_j + x_{ui}\theta_i x_{um}\\ &= -2x_{um}y_u + 2x_{um}x_{uj}\theta_j\\ &= -2\mathbf{X}^T\mathbf{y} + 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}\end{aligned} \tag{5}$$

Setting this derivative equal to zero and solving for $\boldsymbol{\theta}$ yields

$$\hat{\boldsymbol{\theta}} = \left(\mathbf{X}^{\mathbf{T}}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y} \tag{6}$$

This is the ordinary least squares estimate in linear regression. The discrepancy between each observation, $y$, and the corresponding prediction, $\hat{\theta}_1 \cdot x_1 + \hat{\theta}_2 \cdot x_2 + \cdots + \hat{\theta}_k \cdot x_k$, is called the error, $\varepsilon$. This error has certain properties; it is assumed that each error is a random variable with zero mean and standard deviation $\sigma$, i.e., $\boldsymbol{\varepsilon}\sim(\mathbf{0}, \sigma^2\mathbf{I}_n)$. The value of the standard deviation is obtained by classical statistics from the observed errors, namely $\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}$:

$$s^2 = \frac{1}{n-k}\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}\right)^T\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}}\right) \tag{7}$$

where $s$ is the estimate of $\sigma$, is often called the standard error. In linear regression it is also assumed that the number of observations are greater than the number of explanatory variables: $n>k$ and that the $\mathbf{X}$ matrix has rank $k$, i.e., full column rank. In other words, the explanatory variables cannot be linearly dependent. It is also said that a model is a classical regression model if the model errors $\boldsymbol{\varepsilon}$ are Normally distributed.

### Bayesian Inference

Bayesian inference considers the model parameters, $\boldsymbol{\theta}$, and the model error, $\varepsilon$, to be random variables. Thus, rather than point estimates, the objective in this section is to determine the probability distribution of $\boldsymbol{\theta}$, as well as $\sigma$, which is the standard deviation of the Normally distributed error, $\varepsilon$. Assuming non-informative priors, i.e., locally uniform priors, the posterior distributions, given observations $\mathbf{y}$, are the multivariate t-distribution:

$$f(\boldsymbol{\theta}) = \frac{\Gamma(\tfrac{1}{2}(\nu+k)) \cdot s^{-k}\sqrt{\left|\mathbf{X}^T\mathbf{X}\right|}}{\Gamma(\tfrac{1}{2})^k\,\Gamma(\tfrac{1}{2}\nu)\nu^{k/2}}\left(1+\frac{(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^T\mathbf{X}^T\mathbf{X}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})}{\nu s^2}\right)^{-\frac{1}{2}(\nu-k)} \tag{8}$$

and the inverse chi-squared distribution:

$$f(\sigma^2) = \nu s^2 \chi_\nu^{-2} \tag{9}$$

where $\nu$ is the degrees of freedom, where $\nu=n-k$. Notice that the least squares estimates $\hat{\boldsymbol{\theta}}$ and $s$ appear as key quantities in these probability distributions. Therefore, errors in the estimate of $\hat{\boldsymbol{\theta}}$ and $s$ will affect the quality of the Bayesian estimates. Hence, the diagnostics available in classical linear regression maintain their importance in the Bayesian approach. From Eqs. (8) and (9) the following second-moment information is available for the multivariate t-distribution and the inverse chi-squared distribution:

Mean of model parameters: $$\mu_{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} \tag{10}$$

Covariances of model parameters: $$\boldsymbol{\Sigma}_{\boldsymbol{\theta\theta}} = \frac{\nu}{\nu-2}\cdot s^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \tag{11}$$

Mean of model variance: $$\mu_{\sigma^2} = \frac{\nu}{\nu-2}\cdot s^2 \tag{12}$$

Mean of model standard deviation: $$\mu_\sigma = \sqrt{\mu_{\sigma^2}} \tag{13}$$

Variance of model variance: $$\sigma_{\sigma^2}^2 = s^4\cdot\frac{2\cdot\nu^2}{(\nu-4)(\nu-2)^2} \tag{14}$$

Variance of model standard deviation: $$\sigma_\sigma^2 = \frac{\sigma_{\sigma^2}^2}{4\mu_{\sigma^2}} \tag{15}$$

However, when $\nu$ is sufficiently large then the following simpler approximations may be used:

Mean of model parameters: $$\mu_{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} \tag{16}$$

Covariances of model parameters: $$\mathbf{\Sigma}_{\boldsymbol{\theta\theta}} = s^2\left(\mathbf{X}^T\mathbf{X}\right)^{-1} \qquad (17)$$

Mean of model standard deviation: $$\mu_\sigma = s \qquad (18)$$

Variance of model standard deviation: $$\sigma_\sigma^2 = \frac{s^2}{2(\nu-4)} \qquad (19)$$

Given this information, the linear regression model can be implemented in a reliability analysis to estimate event probabilities.

## Diagnostics

Building a regression model is an iterative process between inference and diagnostics. Several things can go wrong when trying to construct a model from observations. Table 1 provides an overview of some of the potential issues, together with techniques with which they may be detected and corrected. The subsequent sections describe the items in greater detail.

**Table 1: Diagnostics of linear regression models**

| **ISSUES** | Collinearity | Heteroskedasticity | Correlation of Errors | Non-Normality | Outliers | Nonlinearity |
|---|---|---|---|---|---|---|
| **DETECTION** | | | | | | |
| Rank and Condition Numbers | X | | | | | |
| R-factor | | | | | | |
| Raw x-y Data Plots | | | | | | |
| Normal Probability Plot | | | | X | | |
| Residual Plots | | X | X | | X | X |
| Model Prediction Plots | | X | | | | X |
| ANOVA Table | | | | | | |
| **REMEDIATION** | | | | | | |
| Transformation | | | | | | |
| Variable Selection | | | | | | |

# Potential Issues

### Collinearity

Collinearity is a deficiency in the data. It means that two or more columns of X are linearly dependent, i.e., between two or more of the explanatory variables. Tendencies of collinearity can severely deteriorate the quality of the model. There are at least two ways to understand why this is the case. First, it should be possible to vary one explanatory variable while keeping the others fixed. This would not be possible if collinearity is present. Examination of Eq. (6) is another way of understanding the detrimental effect of collinearity. If collinearity is present then the inversion of $\mathbf{X}^{\mathrm{T}}\mathbf{X}$ is either impossible or grossly sensitive to small changes in the observed explanatory variable values.

### Heteroskedasticity

Heteroskedasticity means that the variance, $\sigma$, of the model error, $\varepsilon$, varies with $y$ or any $x$. This violates one of the key regression assumptions, which states that the model error variance must be constant, i.e., homoskedastic. Although it is often non-trivial to detect heteroskedasticity it is a crucial part of the model development. In passing, it is noted that true homoskedasticiy implies that the vectors in **X** are orthogonal. This is usually not entirely true, but rarely detrimental to the quality of the model.

### Correlation of Errors

Another fundamental assumption that applies to the errors, $\varepsilon$, is that they are uncorrelated. Particularly when the data is collected over time there may be systematic variations that are unaccounted for in the explanatory variables. In other words, correlation indicates that there are additional explanatory variables that are not included but influence the observations. In summary, the elements of the error vector, $\boldsymbol{\varepsilon}$, should be homoskedastic and uncorrelated. This is referred to as a spherical correlation structure because $\mathrm{Var}(\varepsilon)=\sigma^2$ for all observations and $\mathrm{Cov}(\varepsilon_i,\varepsilon_j)=0$ for all $i\neq j$.

### Non-Normality

The third fundamental assumption of the errors in linear regression is that they are Normally distributed. Severe violation of this assumption invalidates the model.

### Outliers

Outliers are observations with extreme residuals, i.e., observed errors.

### Nonlinearity

Nonlinearity is perhaps the most common violation of the basic assumptions of linear regression, along with heteroskedasticity. Nonlinearity means that the model form in Eq. (1) is inappropriate and that nonlinear regression is necessary, unless other remedies solve the problem.

# Detection

### R-factor

The R-factor is a popular but too simplistic measure of the quality of a model. For a single-variable model it has merit in the sense that it equals the correlation coefficient between $x$ and $y$. In multiple linear regression the coefficient of determination, $R^2$, is utilized, which equals the square of the correlation coefficient between the observations and the model predictions. It is inappropriate to employ the coefficient of determination to compare the quality of different models because its value also depends on the number of explanatory variables. Modified measures are available but it is usually best to study plots to compare models.

### Rank and Condition Numbers

There are several ways to diagnose collinearity. One is to study the rank of $\mathbf{X}$. To avoid problems with collinearity it must have full column rank. Another approach is to study condition numbers of $\mathbf{X}^T\mathbf{X}$, which is the matrix that is inverted in Eq. (6). Any ill-conditioning that is exposed by condition numbers raises collinearity flags. A number of variance inflation factors and scaled condition indices exist to test for collinearity.

### Raw x-y Data Plots

It may be tempting to examine 2D plots with an $x$ on the abscissa axis and $y$ on the ordinate axis to understand the behaviour of $y$. However, in multiple linear regression this is usually misguided because the variation in $y$ is caused by the simultaneous variation in many explanatory variables. Only an unattainable multi-dimensional plot of $y$ in the space of explanatory variables would have been useful in the context of raw x-y data plots.

### Residual Plots

Plots of the residuals, i.e., the observed errors or some modified version thereof, are crucial in checking a regression model. Several of the fundamental assumptions in linear regression apply to the residuals. The raw residuals are the difference between the observed response and the model prediction:

$$\boldsymbol{\varepsilon} = \mathbf{y} - \mathbf{X}\hat{\boldsymbol{\theta}} \tag{20}$$

A plot of these residuals versus the observed responses, as well as plots of these residuals versus the explanatory variables is an effective way of detecting potential heteroskedasticity, and possibly nonlinearity. When applicable, a plot of the residual versus the observation-order is useful for detecting potential correlation of errors.

To improve the value of the residual plots it is common to plot modified residuals instead of the raw residuals. The simplest but perhaps least informative modification is to plot the standardized residuals, which are obtained by subtracting the mean of the residuals and dividing by their standard deviation. A number of other residuals are defined in the literature.

### Normal Probability Plot

After verifying homoskedasticity by residual plots it is prudent to investigate the Normality of the residuals and the potential presence of outliers. One technique for determining the probability distribution of a data is set is to use a probability plot, and more specifically a “Q-Q plot” of quantiles. In this case, the Normal probability plot is applicable. One axis is simply the residual values. The other axis is formed by the inverse Normal CDF with standard deviation $\sigma$, computed at $u/n$, where $u$ is the number of the residual in the ordered vector of residuals and $n$ is the number of observations. The points should align with a straight line if the residuals are Normally distributed. However, some non-systematic deviation is unavoidable, particularly in the tails.

### Model Prediction Plots

The overall quality of a model is gauged by plotting the observed responses, **y**, versus the predicted responses, $\mathbf{X}\hat{\boldsymbol{\theta}}$. The better the points align with a straight line the better the model is. An equivalent plot is the ratio of observed and predicted responses, in which all values should be close to unity.

## Remediation

### Transformation

Transformation means that a model is developed for ln($y$) or some other function of $y$, instead of actually $y$. This may alleviate heteroskedasticity, nonlinearity, and perhaps non-Normality. Different transformations may be tried to see how they affect the model quality and diagnostic plots.

### Variable Selection

When the data is collected in a comprehensive manner, which is unfortunately rare, many potential explanatory variables are available. Then the selection of explanatory variables to include in the model becomes an interesting exercise that can greatly improve both model quality and engineering insight. In fact, instead of considering only individual explanatory variables one should consider a multitude of explanatory functions, i.e., combinations of explanatory variables. For example, it may be found that $(x_1x_2)^2/(x_3)^3$ is a better explanatory function than $x_1$, $x_2$, and $x_3$ alone. Furthermore, in most engineering models it is appealing to utilize dimensionless explanatory functions rather than the individual explanatory variables that are often associated with some unit.

One approach to search for explanatory functions is by engineering judgment and trial-and-error. An explanatory function that is associated with relatively low coefficient of variation, $\delta_\theta=\sigma_\theta/\mu_\theta$, of the associated model parameter is good. The standard deviation of the model error, $\sigma_\varepsilon$, should also be monitored to gauge if a sufficient number of explanatory functions is included. Other metrics that assist in the variable selection include Akaike's information criterion (AIC) and "Mallows Cp." Note that it is possible to include both too few and too many explanatory functions.

An informal technique to search for potential explanatory functions is to set up an algorithm that tries a large number of functions of the form

$$h(\mathbf{x}) = x_1^{m_1} \cdot x_2^{m_2} \cdots x_k^{m_k} \tag{21}$$

where $h$ is a potential explanatory function and $m_\mathrm{j}$ are numbers that could be, say, picked from the vector {-3, -2, -1, 0, 1, 2, 3}. In ***Rt***, this algorithm is implemented so that the candidate explanatory functions that are associated with the lowest $\delta_\theta$ are identified. Clearly, this procedure is *ad hoc* and should only serve as one tool among many to identify the "best" model.