STAT2201

Analysis of Engineering & Scientific Data

Unit 9

Slava Vaisman

The University of Queensland
School of Mathematics and Physics

Regression analysis

We consider problems in engineering that involve a study or analysis of the relationship between two or more variables.

Consider the following examples.

- The pressure of a gas in a container is related to the temperature.

- The velocity of water in an open channel as a function of the channel width.

We examine dependent variable and one or more independent variables also called predictors.

Regression analysis

- The collection of statistical tools that are used to model and explore relationships between variables that are related in a non deterministic manner is called regression analysis.
- Of key importance is the conditional expectation:

$$\mathbb{E}[Y \mid x] = \mu_{Y|x} = \beta_0 + \beta_1 x.$$

- Specifically,

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

where:

- x is a non-random predictor, and
- ϵ is a random (noise) variable, such that $\mathbb{E}[\epsilon] = 0$, and $\mathrm{Var}(\epsilon) = \sigma^2$.

Simple Linear Regression

The setting is as follows.

- Both x and y are scalars, in which case the collected data consisits of n tuples:

$$(x_1, y_1), \ldots, (x_n, y_n).$$

- We assume that the relation between x and y is "linear" in the sense that

$$y \approx \beta_0 + \beta_1 x.$$

- Since we do not have all possible tuples, we can only estimate β_0 and β_1 by $\hat{\beta}_0$ and $\hat{\beta}_1$, respectively. That is, we write:

$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_i + e_i, \quad i = 1, \ldots, n.$$

- The quantity e_i is called the **residual**. Note the correspondence between the noise random variable ϵ and e_i.

The predicted observation

- In general, the predicted observation is defined via

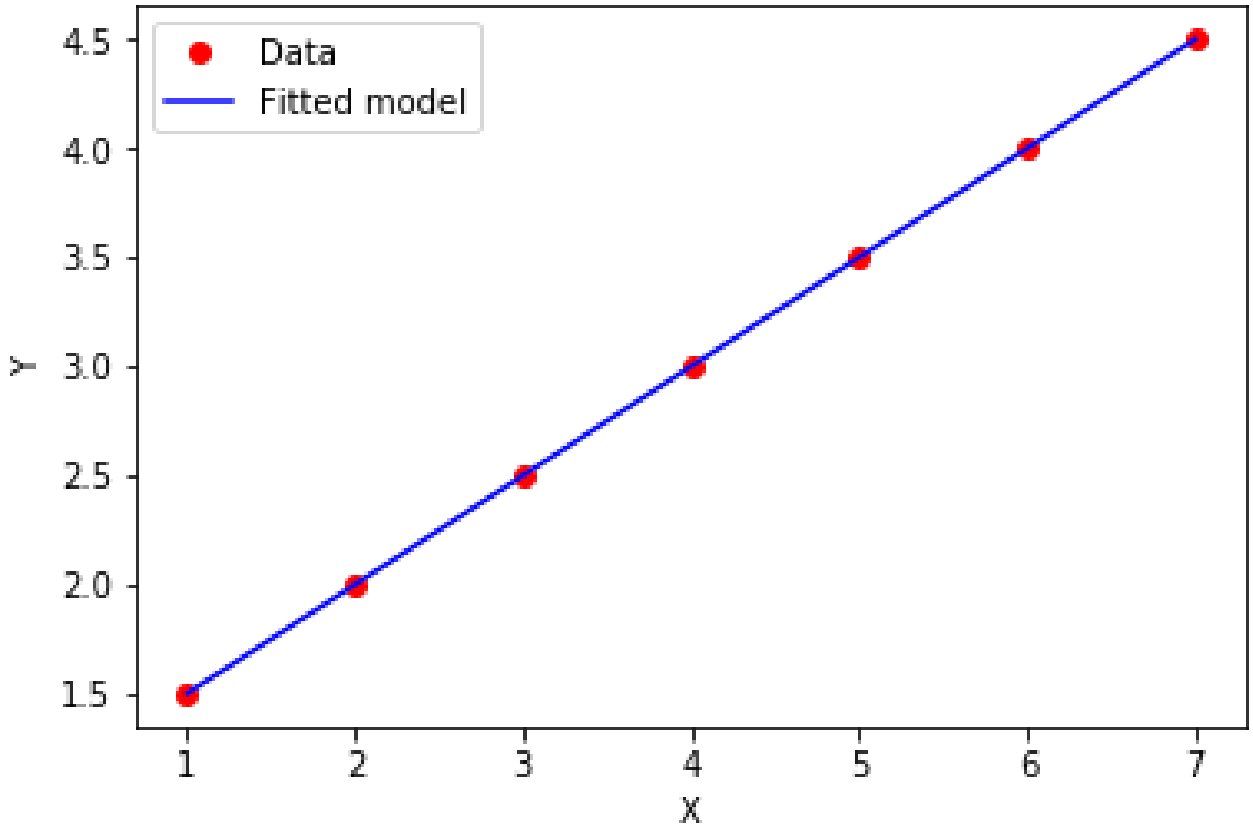
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x.$$

- Note that we can also compute predicted observations for our data $(x_i, y_i)_{\{1 \leq i \leq n\}}$.

Ideally, we would like to find $\hat{\beta}_0$ and $\hat{\beta}_1$, such that $y_i = \hat{y}_i$, that is, $e_i = 0$ for all $i = 1, \ldots, n$.

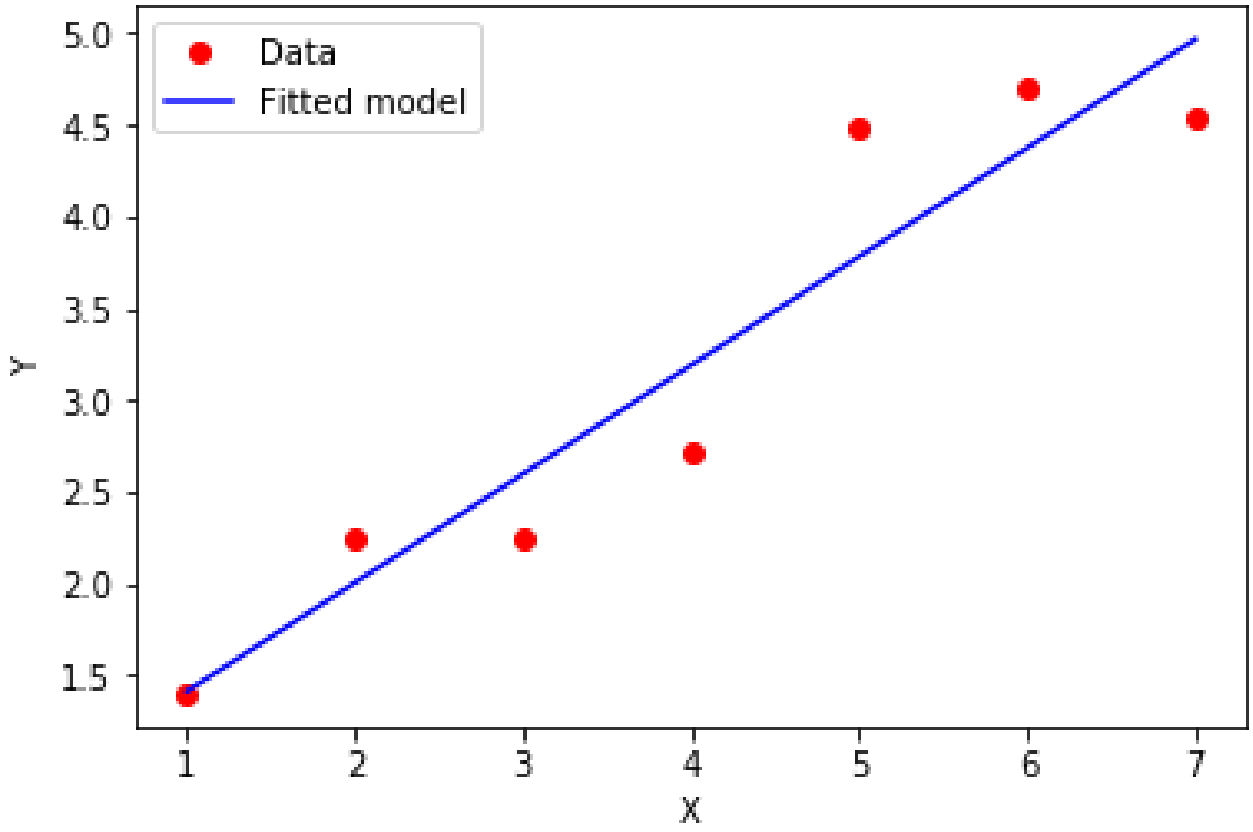
Simple Linear Regression (1)

Ideally, we would like to find $\hat{\beta}_0$ and $\hat{\beta}_1$, such that $y_i = \hat{y}_i$, that is, $e_i = 0$ for all $i = 1, \ldots, n$.

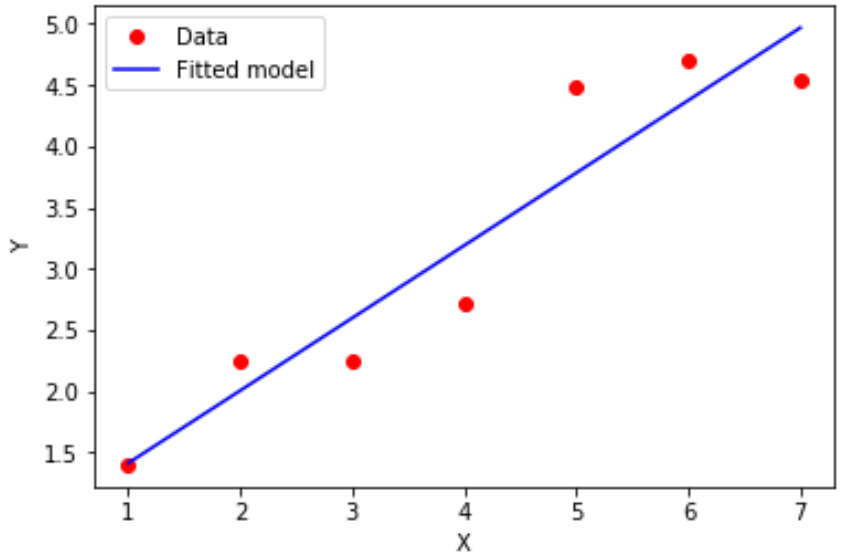


Simple Linear Regression (2)

Ideally, we would like to find $\hat{\beta}_0$ and $\hat{\beta}_1$, such that $y_i = \hat{y}_i$, that is, $e_i = 0$ for all $i = 1, \ldots, n$.



Total mean squared error



The total mean squared error is defined via

$$L = SS_E = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

In practice, $\sigma^2 \neq 0$, that is, all points do not lie on the same line), and therefore we have that $L > 0$.

The least squares estimators

- To find the best estimators $\hat{\beta}_0$ and $\hat{\beta}_1$, we would like to minimize

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- Specifically, solve

$$\hat{\beta}_0, \hat{\beta}_1 = \text{argmin}_{\beta_0, \beta_1} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2.$$

- The solution, called the *least squares estimators* must satisfy:

$$\frac{\partial L}{\partial \beta_0}\Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i) = 0$$

$$\frac{\partial L}{\partial \beta_1}\Big|_{\hat{\beta}_0, \hat{\beta}_1} = -2 \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)\, x_i = 0.$$

The least squares estimators

- Simplifying these two equations yields

$$n\hat{\beta}_0 + \hat{\beta}_1 \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} y_i$$

$$\hat{\beta}_0 \sum_{i=1}^{n} x_i + \hat{\beta}_1 \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i x_i.$$

- These are called the least squares normal equations.

- The solution to the normal equations results in the least squares estimators $\hat{\beta}_0$ and $\hat{\beta}_1$.

The least squares solution

Using the sample means, x and y

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \overline{y} = \frac{1}{n}\sum_{i=1}^{n} y_i,$$

the estimators are:

$$\hat{\beta}_0 = \overline{y} - \hat{\beta}_1 \overline{x}$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}$$

Additional quantities of interest

$$S_{XX} = \sum_{i=1}^{n} (x_i - \overline{x})^2 = \sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}$$

$$S_{XY} = \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y}) = \sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}$$

That is,

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} x_i y_i - \frac{\left(\sum_{i=1}^{n} x_i\right)\left(\sum_{i=1}^{n} y_i\right)}{n}}{\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}} = \frac{S_{XY}}{S_{XX}}.$$

In addition, we have:

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2, \quad SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2, \quad SS_E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2.$$

The Analysis of Variance

- We did not consider the final unknown parameter in our regression model:

$$Y = \beta_0 + \beta_1 x + \epsilon,$$

namely, the $\mathrm{Var}(\epsilon) = \sigma^2$.

- We use the residuals $e_i = \hat{y}_i - y_i$, to obtain an estimate of σ^2.
- Specifically,

$$SS_E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2,$$

and it can be shown that

$$\mathbb{E}[SS_E] = (n-2)\sigma^2,$$

so:

$$\hat{\sigma}^2 = \frac{SS_E}{n-2}.$$

The Analysis of Variance Identity

It holds that:

$$SS_T = SS_R + SS_E,$$

where

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$$SS_E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2.$$

How good is my regression model?

A widely used measure for a regression model is the following ratio of sum of squares, which is often used to judge the adequacy of a regression model:

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T},$$

where

$$SS_T = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \overline{y})^2$$

$$SS_E = \sum_{i=1}^{n} (\hat{y}_i - y_i)^2.$$

Properties of least square estimator

$$\mathbb{E}[\hat{\beta}_0] = \beta_0, \quad \mathrm{Var}\left(\hat{\beta}_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right],$$

$$\mathbb{E}[\hat{\beta}_1] = \beta_1, \quad \mathrm{Var}\left(\hat{\beta}_1\right) = \frac{\sigma^2}{S_{XX}},$$

Therefore, the estimated standard error of the slope and the estimated standard error of the intercept are:

$$se\left(\hat{\beta}_0\right) = \sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\overline{x}^2}{S_{XX}}\right]},$$

$$se\left(\hat{\beta}_1\right) = \sqrt{\frac{\sigma^2}{S_{XX}}}.$$

Hypothesis tests in linear regression (1)

- Suppose we would like to test:

$$H_0 : \beta_1 = \beta_{1,0}, \quad H_1 : \beta_1 \neq \beta_{1,0}.$$

- The Test Statistic for the Slope is

$$T = \frac{\hat{\beta}_1 - \beta_{1,0}}{\sqrt{\frac{\sigma^2}{S_{xx}}}}.$$

- Under H_0, the test statistic T follows a t - distribution with $n - 2$ degree of freedom.

Hypothesis tests in linear regression (2)

- Suppose we would like to test:

$$H_0 : \beta_0 = \beta_{0,0}, \quad H_1 : \beta_0 \neq \beta_{1,0}.$$

- The Test Statistic for the intercept is

$$T = \frac{\hat{\beta}_0 - \beta_{0,0}}{\sqrt{\sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right]}}.$$

- Under H_0, the test statistic T follows a t - distribution with $n - 2$ degree of freedom.

Hypothesis tests in linear regression

An important special case of the hypotheses is:

$$H_0 : \beta_1 = 0, \quad H_1 : \beta_1 \neq 0.$$

If we fail to reject $H_0 : \beta_1 = 0$, this indicates that there is no linear relationship between x and y.

The F distribution

- An alternative is to use the F statistic as is common in ANOVA (Analysis of Variance) (not covered fully in the course).
- Under H_0, the test statistic

$$F = \frac{SS_R/1}{SS_E/(n-2)} = \frac{MS_R}{MS_E},$$

  follows an F - distribution with 1 degree of freedom in the numerator and $n-2$ degrees of freedom in the denominator.
- Here,

$$MS_R = SS_R/1, \quad MS_E = SS_E/(n-2).$$

Analysis of Variance Table for Testing Significance of Regression

Source of Variation	**Sum of Squares**	**Degrees of Freedom**	**Mean Square**	$\boldsymbol{F_0}$
Regression	$SS_R = \hat{\beta}_1 S_{xy}$	1	MS_R	MS_R/MS_E
Error	$SS_E = SS_T - \hat{\beta}_1 S_{xy}$	$n-2$	MS_E	
Total	SS_T	$n-1$		

Additional remarks

- There are also confidence intervals for $\hat{\beta}_0$ and $\hat{\beta}_1$ as well as prediction intervals for observations. We do not cover these formulas.

- To check the regression model assumptions, we plot the residuals e_i and check for:
  - Normality,
  - Constant variance, and,
  - Independence

Logistic Regression

- Take the response variable, Y_i as a Bernoulli random variable.
- In this case notice that $\mathbb{E}[Y] = \mathbb{P}(Y = 1)$.
- The logit response function has the form

$$\mathbb{E}[Y] = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

- Fitting a logistic regression model to data yields estimates of β_0 and β_1.
- The following formula is called the odds:

$$\frac{\mathbb{E}[Y]}{1 - \mathbb{E}[Y]} = e^{\beta_0+\beta_1 x}.$$