Statistics in Geophysics: Linear Regression

Steffen Unkel

Department of Statistics
Ludwig-Maximilians-University Munich, Germany

Historical remarks

- Sir Francis Galton (1822-1911) was responsible for the introduction of the word "regression".
- Galton, F. (1886): Regression towards mediocrity in hereditary stature, *The Journal of the Anthropological Institute of Great Britain and Ireland*, Vol. 15, pp. 246-263.
- Regression equation:

$$\hat{y} = \bar{y} + \frac{2}{3}(x - \bar{x}) \ ,$$

where y denotes the height of the child and x is a weighted average of the mother's and father's heights.

Regression to the mean

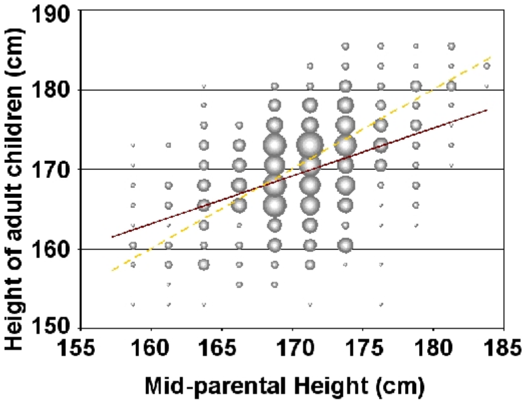

Figure: Scatterplot of mid-parental height against child's height, and regression line (dark red line).

Relationship between two variables

- We can distinguish predictor variables and response variables.
- Other names frequently seen are:
  - Predictor variable: input variable, X-variable, regressor, covariate, independent variable.
  - Response variable: output variable, predictand, Y-variable, dependent variable.
- We shall be interested in finding out how changes in the predictor variables affect the values of a response variable.

Relationship between two variables: Example

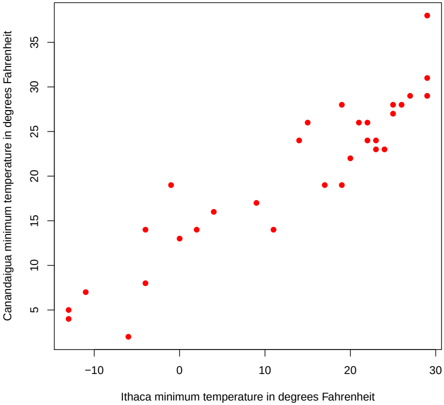

Figure: Plot of the minimum temperature (°F) observations at Ithaca and Canandaigua, New York, for January 1987.

Model

- In simple (multiple) linear regression one (two or more) predictor variable(s) is (are) assumed to affect the values of a response variable in a linear fashion.

- For the model of simple linear regression, we assume

  $$\begin{aligned} y &= f(x) + \epsilon \\ &= \beta_0 + \beta_1 x + \epsilon \ , \end{aligned}$$

  where $\mathrm{E}(y|x) = f(x)$ is known as the systematic component and ϵ is the random error term.

- Inserting the data yields the n equations

  $$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \ , \quad i = 1, \ldots, n$$

  with unknown regression coefficients β_0 and β_1.

Assumptions

1. The systematic component f is a linear combination of covariates, that is, f is linear in the parameters.

2. Additivity of errors.

3. The error terms ϵ_i $(i = 1 \ldots, n)$ are random variables with $\mathrm{E}(\epsilon_i) = 0$ and constant variance σ^2 (unknown), that is, homoscedastic errors with $\mathrm{Var}(\epsilon_i) = \sigma^2$.

4. We assume that errors are uncorrelated, that is, $\mathrm{Cov}(\epsilon_i, \epsilon_j) = 0$ for $i \neq j$.

5. We often assume a normal distribution for the errors: $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$.

Least squares (LS) fitting

- The estimated values $\hat{\beta}_0$ and $\hat{\beta}_1$ are determined as minimizers of the sum of squares deviations

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

for given data (y_i, x_i), $i = 1, \ldots, n$.

- This yields

$$\begin{aligned} \hat{\beta}_1 &= \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad , \\ \hat{\beta}_0 &= \bar{y} - \hat{\beta}_1 \bar{x} \quad . \end{aligned}$$

Least squares (LS) fitting II

- An estimate for the error variance σ^2, called the residual variance, is

$$\begin{aligned}\hat{\sigma}^2 &= \frac{1}{n-2}\sum_{i=1}^{n}\hat{\epsilon}_i^2 \\ &= \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \ ,\end{aligned}$$

  where $\hat{\epsilon}_i$ and $\hat{y}_i$ $(i = 1\ldots, n)$ are the residuals and fitted values, respectively.

- The sum of squared residuals is divided by $n-2$ because two parameters have been estimated.

LS fitting: Example

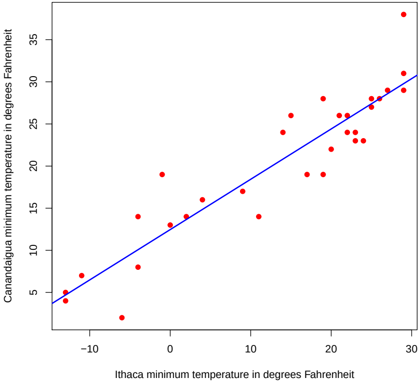

Figure: Minimum temperature (°F) observations at Ithaca and Canandaigua, New York, for January 1987, with fitted least squares line ($\hat{y}_i = 12.459 + 0.598x_i$).

Goodness-of-fit

- How much of the variation in the data has been explained by the regression line?

- Consider the identity

$$y_i - \hat{y}_i = y_i - \bar{y} - (\hat{y}_i - \bar{y}) \Leftrightarrow (y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i) \ .$$

- Decomposition of the total sum of squares:

$$\underbrace{\sum_{i=1}^{n}(y_i - \bar{y})^2}_{SST} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}_{SSR} + \underbrace{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}_{SSE} \ .$$

Coefficient of determination

- Some of the variation in the data (SST) can be ascribed to the regression line (SSR) and some to the fact that the actual observations do not all lie on the regression line (SSE).

- A useful statistic to check is the R^2 value (coefficient of determination):

  $$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\text{SSE}}{\text{SST}} ,$$

  for which it holds that $0 \leq R^2 \leq 1$ and which is often expressed as a percentage by multiplying by 100.

- The square root of R^2 is (the absolute value) of the Pearson correlation between x and y.

ANOVA table for simple linear regression

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

F-test for significance of regression

- Suppose that the errors ϵ_i are independent $\mathcal{N}(0, \sigma^2)$ variables. Then it can be shown that if $\beta_1 = 0$, the ratio

$$F = \frac{\text{MSR}}{\hat{\sigma}^2}$$

follows an F-distribution with 1 and $(n - 2)$ degrees of freedom.

- Statistical test: H_0: $\beta_1 = 0$ versus H_1: $\beta_1 \neq 0$.

- We compare the F-value with the $100(1 - \alpha)\%$ point of the tabulated $F(1, n - 2)$-distribution in order to determine whether β_1 can be considered nonzero on the basis of the data we have seen.

Confidence intervals

- $(1-\alpha)\times 100\%$ confidence intervals for β_0 and β_1:

  $$[\hat{\beta}_j \pm \hat{\sigma}_{\hat{\beta}_j} \times t_{1-\alpha/2}(n-2)] \ , \quad j = 0, 1 \ ,$$

  where

  $$\hat{\sigma}_{\hat{\beta}_1} = \frac{\hat{\sigma}}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}$$

  and

  $$\hat{\sigma}_{\hat{\beta}_0} = \hat{\sigma}\frac{\sqrt{\sum_{i=1}^{n} x_i^2}}{\sqrt{n\sum_{i=1}^{n}(x_i - \bar{x})^2}} \ .$$

- For sufficiently large n: Replace quantiles of the $t(n-2)$-distribution by quantiles of the $\mathcal{N}(0,1)$-distribution.

Hypothesis tests

- Example: Two-sided test for β_1:

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

  Observed test statistic:

  $$t = \frac{\hat{\beta}_1 - 0}{\hat{\sigma}_{\hat{\beta}_1}} = \frac{\hat{\beta}_1}{\hat{\sigma}_{\hat{\beta}_1}} \ ,$$

  Rejection region: $|t| > t_{1-\alpha/2}(n-2)$.

- Note that the variable $F(1, n-2)$ is the square of the $t(n-2)$ variable.

Prediction intervals

- A prediction interval for a future observation y_0 at a location x_0 with level $(1-\alpha)$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x_0 \pm t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad .$$

- A confidence interval for the regression function $\beta_0 + \beta_1 x$ with level $(1-\alpha)$ is given by

$$\hat{\beta}_0 + \hat{\beta}_1 x \pm t_{1-\alpha/2}(n-2)\hat{\sigma}\sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{\sum_{i=1}^n (x_i - \bar{x})^2}} \quad .$$

Prediction intervals: Example

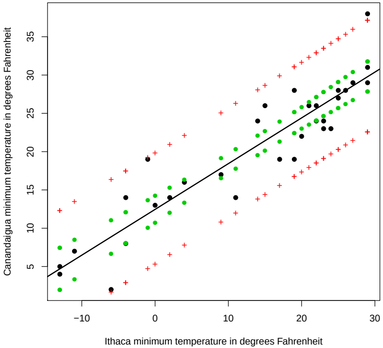

Figure: 95% prediction intervals (red crosses) and 95% confidence intervals (green dots) around the regression (thick black line) for the January 1987 temperature data. Data to which the regression was fit (black dots) are also shown.

Residuals versus fitted values

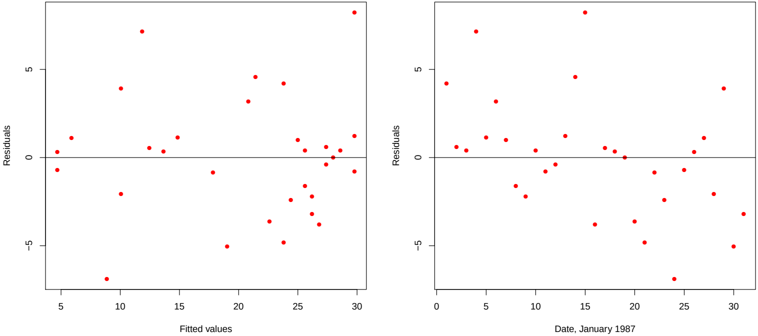

Figure: Scatterplot of the residuals as a function of the predicted value $\hat{y}_i$ $(i = 1 \ldots, n)$ (left) and as a function of date (right), for the January 1987 temperature data.

Durbin-Watson test

- A test for serial correlation of regression residuals is the Durbin-Watson test.

- Observed test statistic:

$$d = \frac{\sum_{i=2}^{n}(\hat{\epsilon}_i - \hat{\epsilon}_{i-1})^2}{\sum_{i=1}^{n}\hat{\epsilon}_i^2} \quad , \quad 0 \leq d \leq 4 \ .$$

- If successive residuals are positively (negatively) serially correlated, d will be near 0 (near 4).

- The distribution of d is symmetric around 2.

- The critical values for Durbin-Watson tests vary depending on the sample size and the number of predictor variables.

Durbin-Watson test II

- Compare d (or $4 - d$, whichever is closer to zero) with the tabulated critical values d_L and d_U.

- If $d < d_L$, conclude that positive serial correlation is a possibility; if $d > d_U$, conclude that no serial correlation is indicated.

- If $4 - d < d_L$, conclude that negative serial correlation is a possibility; if $4 - d > d_U$, conclude that no serial correlation is indicated.

- If the d (or $4 - d$) value lies between d_L and d_U, the test is inconclusive.

Durbin-Watson test: Example

```
Durbin-Watson test

data:  linmodel1
DW = 1.5554, p-value = 0.08104
alternative hypothesis: true autocorrelation is greater than 0
```

Quantile-quantile plot

- A graphical impression of whether the residuals follow a normal distribution can be obtained through a quantile-quantile (Q-Q) plot.

- The residuals are plotted on the vertical, and the standard normal variables corresponding to the empirical cumulative probability of each residual are plotted on the horizontal.

- Draw a straight line through the main middle bulk of the plot.

- If all the points lie on such a line, more or less, one would conclude that the residuals do not deny the assumption of normality of errors.

Quantile-quantile plot: Example

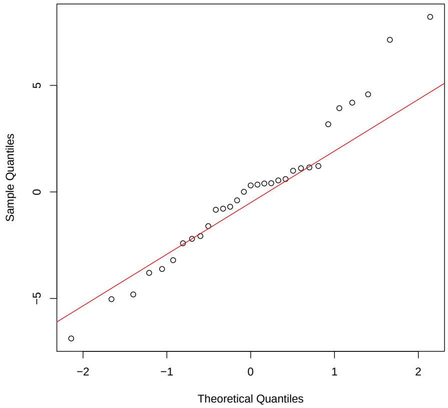

Figure: Gaussian Q-Q plot of the residuals obtained from the regression of the January 1987 temperature data.