Data Mining and Machine Learning: Fundamental Concepts and Algorithms

dataminingbook.info

Mohammed J. Zaki[1] Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 27: Regression Evaluation

Univariate Regression

$$Y = f(X) + \varepsilon = \beta + \omega \cdot X + \varepsilon$$

where ω is the slope of the best fitting line and β is its intercept, and ε is the random error variable that follows a normal distribution with mean $\mu = 0$ and variance σ^2.

The true parameters β, ω and σ^2 are all unknown, and have to be estimated from $\boldsymbol{D}$ comprising n points x_i and corresponding response values y_i, for $i = 1, 2, \cdots, n$.

Let b and w denote the estimated bias and weight terms; we can then make predictions for any given value x_i as follows:

$$\hat{y}_i = b + w \cdot x_i$$

The estimated bias b and weight w are obtained by minimizing the sum of squared errors (SSE), given as

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} (y_i - b - w \cdot x_i)^2$$

Univariate Regression

According to our model, the variance in prediction is entirely due to the random error term ε. We can estimate this variance by considering the predicted value $\hat{y}_i$ and its deviation from the true response y_i, that is, by looking at the residual error

$$\epsilon_i = y_i - \hat{y}_i$$

The estimated variance $\hat{\sigma}^2$ is given as

$$\hat{\sigma}^2 = \text{var}(\epsilon_i) = \frac{1}{n-2} \cdot \sum_{i=1}^{n} \big(\epsilon_i - E[\epsilon_i]\big)^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} \epsilon_i^2 = \frac{1}{n-2} \cdot \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Thus, the estimated variance is

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

We divide by $n-2$ to get an unbiased estimate, since $n-2$ is the number of degrees of freedom for estimating SSE.

The SSE value gives an indication of how much of the variation in Y cannot be explained by our linear model.

We can compare this value with the *total scatter*, also called *total sum of squares*, for the dependent variable Y, defined as

$$TSS = \sum_{i=1}^{n} (y_i - \mu_Y)^2$$

Notice that, in TSS, we compute the squared deviations of the true response from the true mean for Y, whereas, in SSE we compute the squared deviations of the true response from the predicted response.

Univariate Regression

The total scatter can be decomposed into two components by adding and subtracting $\hat{y}_i$ as follows

$$\begin{aligned} TSS &= \sum_{i=1}^{n}(y_i - \mu_Y)^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i + \hat{y}_i - \mu_Y)^2 \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2 + 2\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot(\hat{y}_i - \mu_Y) \\ &= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2 = SSE + RSS \end{aligned}$$

where we use the fact that $\sum_{i=1}^{n}(y_i - \hat{y}_i)\cdot(\hat{y}_i - \mu_Y) = 0$, and

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

is a new term called *regression sum of squares* that measures the squared deviation of the predictions from the true mean.

Univariate Regression

TSS can thus be decomposed into two parts: SSE, which is the amount of variation not explained by the model, and RSS, which is the amount of variance explained by the model.

The fraction of the variation left unexplained by the model is $\frac{SSE}{TSS}$.

Conversely, the fraction of the variation that is explained by the model, called the *coefficient of determination* or simply the R^2 *statistic*, is given as

$$R^2 = \frac{TSS - SSE}{TSS} = 1 - \frac{SSE}{TSS} = \frac{RSS}{TSS}$$

The higher the R^2 statistic the better the estimated model, with $R^2 \in [0, 1]$.

Variance and Goodness of Fit

Consider the regression of `petal length` (X; the predictor variable) on `petal width` (Y; the response varilris dataset ($n = 150$ data points).

The least squares estimates for the bias and regression coefficients are as follows

$$w = 0.4164 \qquad\qquad b = -0.3665$$

The SSE value is given as

$$SSE = \sum_{i=1}^{150} \epsilon_i^2 = \sum_{i=1}^{150} (y_i - \hat{y}_i)^2 = 6.343$$

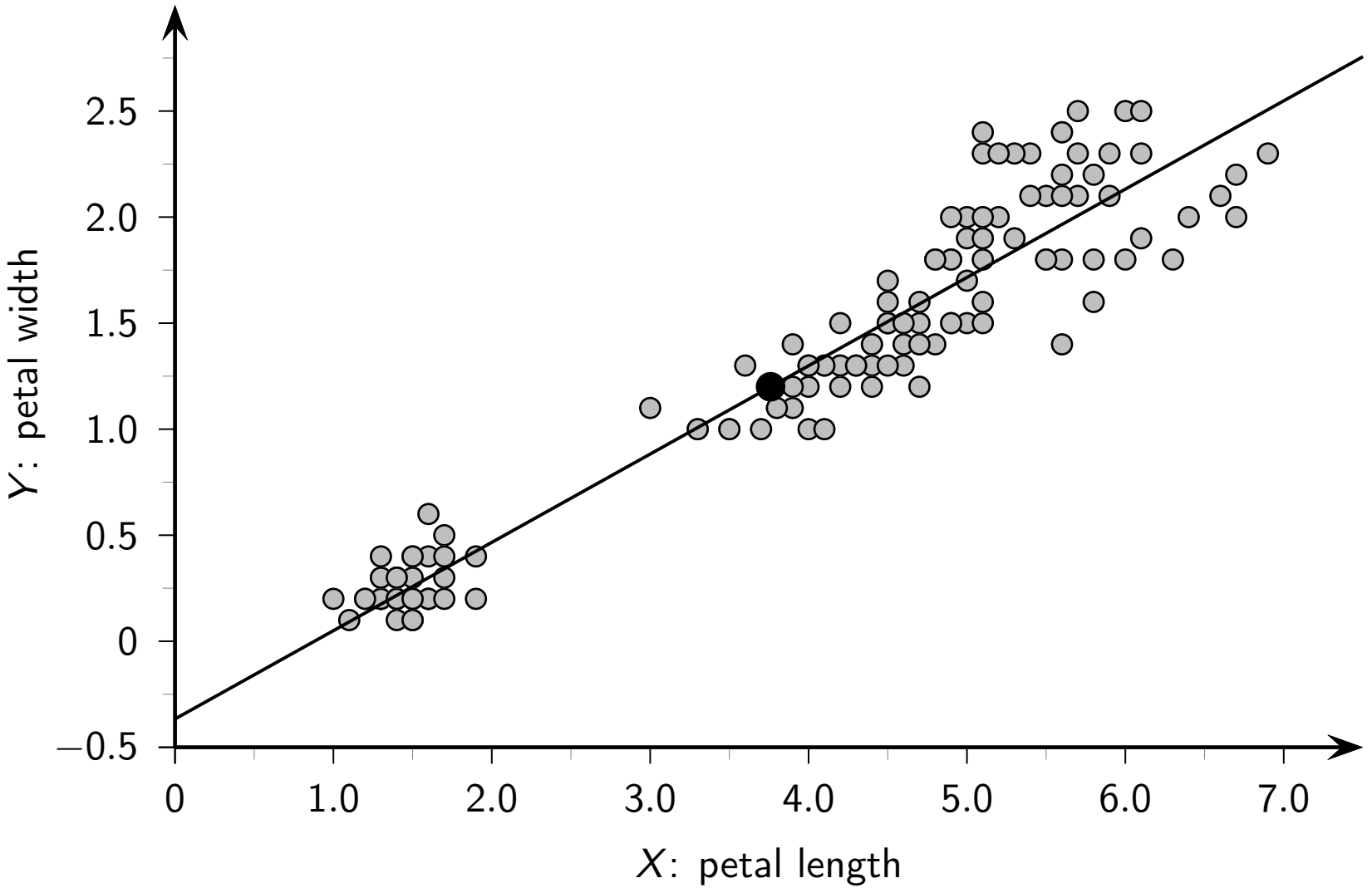
Thus, the estimated variance and standard error of regression are given as

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = \frac{6.343}{148} = 4.286 \times 10^{-2}$$

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-2}} = \sqrt{4.286 \times 10^{-2}} = 0.207$$

Variance and Goodness of Fit

Scatterplot: `petal length` (X) versus `petal width` (Y). Mean point shown as black circle.



Variance and Goodness of Fit

For the bivariate Iris data, the values of TSS and RSS are given as

$$TSS = 86.78 \qquad RSS = 80.436$$

We can observe that $TSS = SSE + RSS$.

The fraction of variance explained by the model, that is, the R^2 value, is given as

$$R^2 = \frac{RSS}{TSS} = \frac{80.436}{86.78} = 0.927$$

This indicates a very good fit of the linear model.

Geometry of Goodness of Fit

Recall that Y can be decomposed into two orthogonal parts:

$$Y = \widehat{Y} + \epsilon$$

We can further decompose $\widehat{Y}$ as follows

$$\widehat{Y} = \text{proj}_1(Y) \cdot 1 + \text{proj}_{\overline{X}}(Y) \cdot \overline{X} = \mu_Y \cdot 1 + \frac{Y^T \overline{X}}{\overline{X}^T \overline{X}} \cdot \overline{X} = \mu_Y \cdot 1 + w \cdot \overline{X}$$

The vectors Y and $\widehat{Y}$ can be centered

$$\overline{Y} = Y - \mu_Y \cdot 1 \qquad\qquad \widehat{\overline{Y}} = \widehat{Y} - \mu_Y \cdot 1 = w \cdot \overline{X}$$

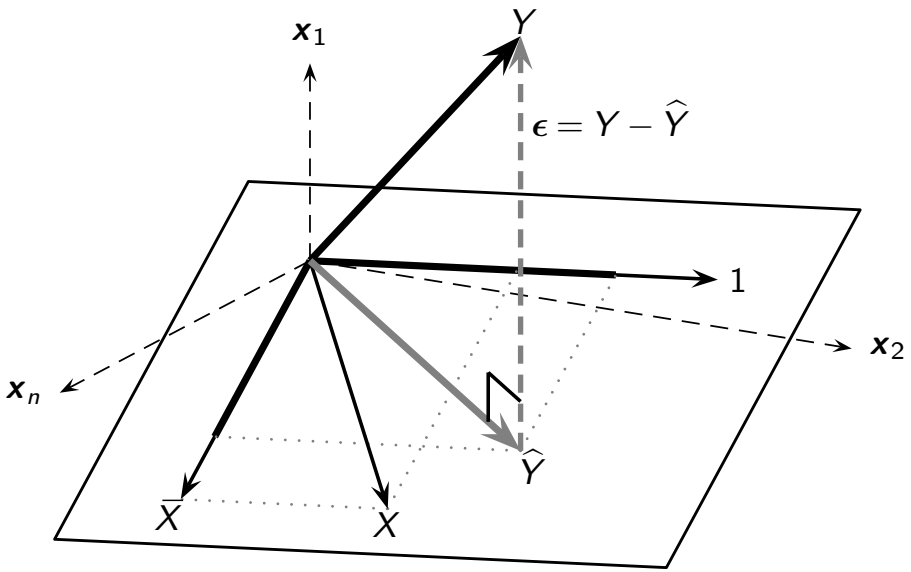
The centered vectors $\overline{Y}$, $\widehat{\overline{Y}}$ and $\overline{X}$ all lie in the $n-1$ dimensional subspace orthogonal to the vector 1, where $\overline{Y}$ and $\widehat{\overline{Y}}$, and ϵ form a right triangle.

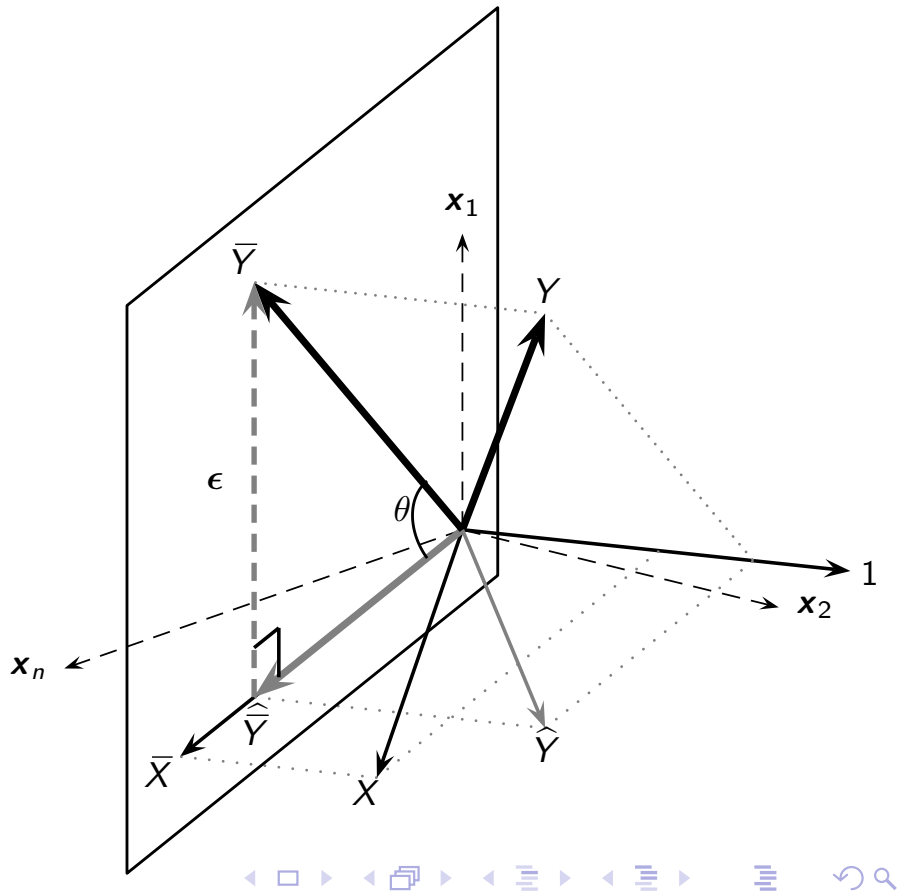
$$||\overline{Y}||^2 = ||\widehat{\overline{Y}}||^2 + ||\epsilon||^2 = ||\widehat{\overline{Y}}||^2 + ||Y - \widehat{Y}||^2$$

This equation is equivalent to the decomposition of the total scatter, TSS, into sum of squared errors, SSE, and residual sum of squares, RSS.

Geometry of Univariate Regression

The vector space that is the complement of 1 has dimensionality $n-1$. The error space (containing the vector ϵ) is orthogonal to $\overline{X}$, and has dimensionality $n-2$, which also specifies the degrees of freedom for the estimated variance $\hat{\sigma}^2$.





Geometry of Goodness of Fit

The total scatter, TSS, is defined as follows:

$$TSS = \sum_{i=1}^{n}(y_i - \mu_Y)^2 = ||Y - \mu_Y \cdot 1||^2 = ||\overline{Y}||^2$$

The residual sum of squares, RSS, is defined as

$$RSS = \sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2 = ||\widehat{Y} - \mu_Y \cdot 1||^2 = ||\widehat{\overline{Y}}||^2$$

Finally, the sum of squared errors, SSE, is defined as

$$SSE = \|\epsilon\|^2 = \left\|Y - \widehat{Y}\right\|^2$$

Thus, the geometry of univariate regression makes it evident that

$$\|\overline{Y}\|^2 = \left\|\widehat{\overline{Y}}\right\|^2 + \left\|Y - \widehat{Y}\right\|^2$$

$$\|Y - \mu_Y \cdot 1\|^2 = \left\|\widehat{Y} - \mu_Y \cdot 1\right\|^2 + \left\|Y - \widehat{Y}\right\|^2$$

$$TSS = RSS + SSE$$

Geometry of Goodness of Fit

The cosine of the angle between $\overline{Y}$ and $\widehat{\overline{Y}}$ is given as the ratio of the base to the hypotenuse and the cosine of the angle and is also $\rho_{Y\widehat{Y}}$.

$$\rho_{Y\widehat{Y}} = \cos\theta = \frac{\left\|\widehat{\overline{Y}}\right\|}{\left\|\overline{Y}\right\|}$$

We can observe that

$$\left\|\widehat{\overline{Y}}\right\| = \rho_{Y\widehat{Y}} \cdot \left\|\overline{Y}\right\|$$

Note that, whereas $|\rho_{Y\widehat{Y}}| \leq 1$, due to the projection operation, the angle between Y and $\widehat{Y}$ is always less than or equal to $90°$, which means that $\rho_{Y\widehat{Y}} \in [0, 1]$.

Thus, the predicted response vector $\widehat{\overline{Y}}$ is smaller than the true response vector $\overline{Y}$ by an amount equal to the correlation between them.

$$R^2 = \frac{RSS}{TSS} = \frac{\left\|\widehat{\overline{Y}}\right\|^2}{\left\|\overline{Y}\right\|^2} = \rho^2_{Y\widehat{Y}}$$

Inference about Regression Coefficient and Bias Term

b and w are only point estimates for the true parameters β and ω. To obtain confidence intervals for these parameters, we treat each y_i as a random variable for the response given the corresponding fixed value x_i.

These random variables are all independent and identically distributed as Y, with expected value $\beta + \omega \cdot x_i$ and variance σ^2. The x_i values are fixed *a priori* and therefore μ_X and σ_X^2 are also fixed:

$$b = \mu_Y - w \cdot \mu_X$$

$$w = \frac{\sum_{i=1}^{n}(x_i - \mu_X)(y_i - \mu_Y)}{\sum_{i=1}^{n}(x_i - \mu_X)^2} = \frac{1}{s_X}\sum_{i=1}^{n}(x_i - \mu_X) \cdot y_i = \sum_{i=1}^{n} c_i \cdot y_i$$

where c_i is a constant (since x_i is fixed), given as $c_i = \frac{x_i - \mu_X}{s_X}$ and $s_X = \sum_{i=1}^{n}(x_i - \mu_X)^2$ is the total scatter for X, defined as the sum of squared deviations of x_i from its mean μ_X.

Mean and Variance of Regression Coefficient

The expected value of w is given as

$$E[w] = E\left[\sum_{i=1}^{n} c_i y_i\right] = \sum_{i=1}^{n} c_i \cdot E[y_i] = \sum_{i=1}^{n} c_i(\beta + \omega \cdot x_i)$$
$$= \beta \sum_{i=1}^{n} c_i + \omega \cdot \sum_{i=1}^{n} c_i \cdot x_i = \frac{\omega}{s_X} \cdot \sum_{i=1}^{n} (x_i - \mu_X) \cdot x_i = \frac{\omega}{s_X} \cdot s_X = \omega$$

which follows from the observation that $\sum_{i=1}^{n} c_i = 0$, and further

$$s_X = \sum_{i=1}^{n} (x_i - \mu_X)^2 = \left(\sum_{i=1}^{n} x_i^2\right) - n \cdot \mu_X^2 = \sum_{i=1}^{n} (x_i - \mu_X) \cdot x_i$$

Thus, w is an unbiased estimator for the true parameter ω.

Mean and Variance of Regression Coefficient

Using the fact that the variables y_i are independent and identically distributed as Y, we can compute the variance of w as follows

$$\text{var}(w) = \text{var}\left(\sum_{i=1}^{n} c_i \cdot y_i\right) = \sum_{i=1}^{n} c_i^2 \cdot \text{var}(y_i) = \sigma^2 \cdot \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{s_X}$$

since c_i is a constant, $\text{var}(y_i) = \sigma^2$, and further

$$\sum_{i=1}^{n} c_i^2 = \frac{1}{s_X^2} \cdot \sum_{i=1}^{n} (x_i - \mu_X)^2 = \frac{s_X}{s_X^2} = \frac{1}{s_X}$$

The standard deviation of w, also called the standard error of w, is given as

$$\text{se}(w) = \sqrt{\text{var}(w)} = \frac{\sigma}{\sqrt{s_X}}$$

Mean and Variance of Bias Term

The expected value of b is given as

$$E[b] = E[\mu_Y - w \cdot \mu_X] = E\left[\frac{1}{n}\sum_{i=1}^{n} y_i - w \cdot \mu_X\right]$$

$$= \left(\frac{1}{n} \cdot \sum_{i=1}^{n} E[y_i]\right) - \mu_X \cdot E[w] = \left(\frac{1}{n}\sum_{i=1}^{n}(\beta + \omega \cdot x_i)\right) - \omega \cdot \mu_X$$

$$= \beta + \omega \cdot \mu_X - \omega \cdot \mu_X = \beta$$

Thus, b is an unbiased estimator for the true parameter β.

Using the observation that all y_i are independent, the variance of the bias term can be computed.

Mean and Variance of Bias Term

$$\begin{aligned}\text{var}(b) &= \text{var}(\mu_Y - w \cdot \mu_X) \\ &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n} y_i\right) + \text{var}(\mu_X \cdot w) \\ &= \frac{1}{n^2} \cdot n\sigma^2 + \mu_X^2 \cdot \text{var}(w) = \frac{1}{n} \cdot \sigma^2 + \mu_X^2 \cdot \frac{\sigma^2}{s_X} = \left(\frac{1}{n} + \frac{\mu_X^2}{s_X}\right) \cdot \sigma^2\end{aligned}$$

where we used the fact that for any two random variables A and B, we have $\text{var}(A - B) = \text{var}(A) + \text{var}(B)$. That is, variances of A and B add, even though we are computing the variance of $A - B$.

The standard deviation of b, also called the standard error of b, is given as

$$\text{se}(b) = \sqrt{\text{var}(b)} = \sigma \cdot \sqrt{\frac{1}{n} + \frac{\mu_X^2}{s_X}}$$

Covariance of Regression Coefficient and Bias

We can also compute the covariance of w and b, as follows

$$\begin{aligned}
\text{cov}(w,b) &= E[w \cdot b] - E[w] \cdot E[b] = E[(\mu_Y - w \cdot \mu_X) \cdot w] - \omega \cdot \beta \\
&= \mu_Y \cdot E[w] - \mu_X \cdot E[w^2] - \omega \cdot \beta = \mu_Y \cdot \omega - \mu_X \cdot \left(\text{var}(w) + E[w]^2\right) - \omega \cdot \beta \\
&= \mu_Y \cdot \omega - \mu_X \cdot \left(\frac{\sigma^2}{s_X} - \omega^2\right) - \omega \cdot \beta = \omega \cdot \underbrace{(\mu_Y - \omega \cdot \mu_X)}_{\beta} - \frac{\mu_X \cdot \sigma^2}{s_X} - \omega \cdot \beta \\
&= -\frac{\mu_X \cdot \sigma^2}{s_X}
\end{aligned}$$

where we use the fact that $\text{var}(w) = E[w^2] - E[w]^2$, which implies $E[w^2] = \text{var}(w) + E[w]^2$, and further that $\mu_Y - \omega \cdot \mu_X = \beta$.

Confidence Intervals

Since the y_i variables are all normally distributed, their linear combination also follows a normal distribution. Thus, w follows a normal distribution with mean ω and variance σ^2/s_X. Likewise, b follows a normal distribution with mean β and variance $(1/n + \mu_X^2/s_X) \cdot \sigma^2$.

We use the estimated variance $\hat{\sigma}^2$, to define Z_w and Z_b:

$$Z_w = \frac{w - E[w]}{\text{se}(w)} = \frac{w - \omega}{\frac{\hat{\sigma}}{\sqrt{s_X}}} \qquad Z_b = \frac{b - E[b]}{\text{se}(b)} = \frac{b - \beta}{\hat{\sigma}\sqrt{(1/n + \mu_X^2/s_X)}}$$

These variables follow the Student's t distribution with $n-2$ degrees of freedom. Given confidence level $1-\alpha$, i.e., significance level $\alpha \in (0,1)$, the $100(1-\alpha)\%$ confidence interval for the true values, ω and β, are: as follows

$$P\left(w - t_{\alpha/2} \cdot \text{se}(w) \leq \omega \leq w + t_{\alpha/2} \cdot \text{se}(w)\right) = 1 - \alpha$$
$$P\left(b - t_{\alpha/2} \cdot \text{se}(b) \leq \beta \leq b + t_{\alpha/2} \cdot \text{se}(b)\right) = 1 - \alpha$$

Confidence Intervals

Example

We consider the variance of the bias and regression coefficient, and their covariance. However, since we do not know the true variance σ^2, we use the estimated variance and the standard error for the Iris data

$$\hat{\sigma}^2 = \frac{SSE}{n-2} = 4.286 \times 10^{-2}$$
$$\hat{\sigma} = \sqrt{4.286 \times 10^{-2}} = 0.207$$

Furthermore, we have

$$\mu_X = 3.7587 \qquad s_X = 463.864$$

Therefore, the estimated variance and standard error of w is given as

$$\text{var}(w) = \frac{\hat{\sigma}^2}{s_X} = \frac{4.286 \times 10^{-2}}{463.864} = 9.24 \times 10^{-5}$$
$$\text{se}(w) = \sqrt{\text{var}(w)} = \sqrt{9.24 \times 10^{-5}} = 9.613 \times 10^{-3}$$

Confidence Intervals

Example

The estimated variance and standard error of b is

$$\begin{aligned}\text{var}(b) &= \left(\frac{1}{n} + \frac{\mu_X^2}{s_X}\right) \cdot \hat{\sigma}^2 \\ &= \left(\frac{1}{150} + \frac{(3.759)^2}{463.864}\right) \cdot (4.286 \times 10^{-2}) \\ &= (3.712 \times 10^{-2}) \cdot (4.286 \times 10^{-2}) = 1.591 \times 10^{-3} \\ \text{se}(b) &= \sqrt{\text{var}(b)} = \sqrt{1.591 \times 10^{-3}} = 3.989 \times 10^{-2}\end{aligned}$$

and the covariance between b and w is

$$\text{cov}(w, b) = -\frac{\mu_X \cdot \hat{\sigma}^2}{s_X} = -\frac{3.7587 \cdot (4.286 \times 10^{-2})}{463.864} = -3.473 \times 10^{-4}$$

For the confidence interval, we use a confidence level of $1 - \alpha = 0.95$ (or $\alpha = 0.05$). The critical value of the t-distribution, with $n - 2 = 148$ degrees of freedom, that encompasses $\alpha/2 = 0.025$ fraction of the probability mass in the right tail is $t_{\alpha/2} = 1.976$. We have

$$t_{\alpha/2} \cdot \text{se}(w) = 1.976 \cdot (9.613 \times 10^{-3}) = 0.019$$

Confidence Intervals

Example

Therefore, the 95% confidence interval for the true value, ω, of the regression coefficient is given as

$$\begin{aligned}\big(w - t_{\alpha/2} \cdot \text{se}(w),\ \ w + t_{\alpha/2} \cdot \text{se}(w)\big) &= (0.4164 - 0.019,\ 0.4164 + 0.019) \\ &= (0.397, 0.435)\end{aligned}$$

Likewise, we have:

$$t_{\alpha/2} \cdot \text{se}(b) = 1.976 \cdot (3.989 \times 10^{-2}) = 0.079$$

Therefore, the 95% confidence interval for the true bias term, β, is

$$\begin{aligned}\big(b - t_{\alpha/2} \cdot \text{se}(b),\ \ b + t_{\alpha/2} \cdot \text{se}(b)\big) &= (-0.3665 - 0.079,\ -0.3665 + 0.079) \\ &= (-0.446, -0.288)\end{aligned}$$

Hypothesis Testing for Regression Effects

One of the key questions in regression is whether X predicts the response Y. In the regression model, Y depends on X through the parameter ω, therefore, we can check for the regression effect by assuming the null hypothesis H_0 that $\omega = 0$, with the alternative hypothesis H_a being $\omega \neq 0$:

$$H_0 : \omega = 0 \qquad\qquad H_a : \omega \neq 0$$

When $\omega = 0$, the response Y depends only on the bias β and the random error ε. In other words, X provides no information about the response variable Y.

Hypothesis Testing for Regression Effects

Now consider the standardized variable Z_w.Under the null hypothesis we have $E[w] = \omega = 0$. Thus,

$$Z_w = \frac{w - E[w]}{\text{se}(w)} = \frac{w}{\hat{\sigma}/\sqrt{s_X}}$$

We can therefore compute the *p-value* for the Z_w statistic using a two-tailed test via the t distribution with $n-2$ degrees of freedom. Given significance level α (e.g., $\alpha = 0.01$), we reject the null hypothesis if the *p-value* is below α. In this case, we accept the alternative hypothesis that the estimated value of the slope parameter is significantly different from zero.

We can also define the f-statistic, which is the ratio of the regression sum of squares, RSS, to the estimated variance, given as

$$f = \frac{RSS}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ (n-2)}$$

Hypothesis Testing for Regression Effects

Under the null hypothesis, one can show that

$$E[RSS] = \sigma^2$$

Further, it is also true that

$$E[\hat{\sigma}^2] = \sigma^2$$

Thus, under the null hypothesis the f-statistic has a value close to 1, which indicates that there is no relationship between the predictor and response variables. On the other hand, if the alternative hypothesis is true, then $E[RSS] \geq \sigma^2$, resulting in a larger f value.

In fact, the f-statistic follows a F-distribution with $1, (n-2)$ degrees of freedom (for the numerator and denominator, respectively); therefore, we can reject the null hypothesis that $w = 0$ if the *p-value* of f is less than the significance level α, say 0.01.

Hypothesis Testing for Regression Effects

Note that we can also test if the bias value is statistically significant or not by setting up the null hypothesis, $H_0 : \beta = 0$, versus the alternative hypothesis $H_a : \beta \neq 0$. We then evaluate the Z_b statistic under the null hypothesis:

$$Z_b = \frac{b - E[b]}{\text{se}(b)} = \frac{b}{\hat{\sigma} \cdot \sqrt{(1/n + \mu_X^2/s_X)}}$$

since, under the null hypothesis $E[b] = \beta = 0$. Using a two-tailed t-test with $n - 2$ degrees of freedom, we can compute the *p-value* of Z_b. We reject the null hypothesis if this value is smaller than the significance level α.

Hypothesis Testing

Example

Under the null hypothesis we have $\omega = 0$, further $E[w] = \omega = 0$. Therefore, the standardized variable Z_w is given as

$$Z_w = \frac{w - E[w]}{\text{se}(w)} = \frac{w}{\text{se}(w)} = \frac{0.4164}{9.613 \times 10^{-3}} = 43.32$$

Using a two-tailed t-test with $n-2$ degrees of freedom, we find that

$$p\text{-}value(43.32) \simeq 0$$

Since this value is much less than the significance level $\alpha = 0.01$, we conclude that observing such an extreme value of Z_w is unlikely under the null hypothesis. Therefore, we reject the null hypothesis and accept the alternative hypothesis that $\omega \neq 0$.

Now consider the f-statistic, we have

$$f = \frac{RSS}{\hat{\sigma}^2} = \frac{80.436}{4.286 \times 10^{-2}} = 1876.71$$

Hypothesis Testing

Example

Using the F-distribution with $(1, n-2)$ degrees of freedom, we have

$$p\text{-}value(1876.71) \simeq 0$$

In other words, such a large value of the f-statistic is extremely rare, and we can reject the null hypothesis. We conclude that Y does indeed depend on X, since $\omega \neq 0$.

Finally, we test whether the bias term is significant or not. Under the null hypothesis $H_0 : \beta = 0$, we have

$$Z_b = \frac{b}{\text{se}(b)} = \frac{-0.3665}{3.989 \times 10^{-2}} = -9.188$$

Using the two-tailed t-test, we find

$$p\text{-}value(-9.188) = 3.35 \times 10^{-16}$$

It is clear that such an extreme Z_b value is highly unlikely under the null hypothesis. Therefore, we accept the alternative hypothesis that $H_a : \beta \neq 0$.

Standardized Residuals

Our assumption about the true errors ε_i is that they are normally distributed with mean $\mu = 0$ and fixed variance σ^2.

We can examine how well the residual errors $\epsilon_i = y_i - \hat{y}_i$ satisfy the normality assumption. The mean of ϵ_i is given as

$$\begin{aligned} E[\epsilon_i] &= E[y_i - \hat{y}_i] = E[y_i] - E[\hat{y}_i] \\ &= \beta + \omega \cdot x_i - E[b + w \cdot x_i] = \beta + \omega \cdot x_i - (\beta + \omega \cdot x_i) = 0 \end{aligned}$$

To compute the variance of ϵ_i, we will express it as a linear combination of the y_j variables:

$$\begin{aligned} \text{var}(\epsilon_i) &= \sigma^2 \cdot \left(1 - \frac{2}{n} - \frac{2 \cdot (x_i - \mu_X)^2}{s_X} + \frac{1}{n} + \frac{(x_i - \mu_X)^2}{s_X}\right) \\ &= \sigma^2 \cdot \left(1 - \frac{1}{n} - \frac{(x_i - \mu_X)^2}{s_X}\right) \end{aligned}$$

Standardized Residuals

We can now define the *standardized residual* ϵ_i^* by dividing ϵ_i by its standard deviation after replacing σ^2 by its estimated value $\hat{\sigma}^2$. That is,

$$\epsilon_i^* = \frac{\epsilon_i}{\sqrt{\text{var}(\epsilon_i)}} = \frac{\epsilon_i}{\hat{\sigma} \cdot \sqrt{1 - \frac{1}{n} - \frac{(x_i - \mu_X)^2}{s_X}}}$$

These standardized residuals should follow a standard normal distribution.

We can thus plot the standardized residuals against the quantiles of a standard normal distribution, and check if the normality assumption holds.

Significant deviations would indicate that our model assumptions may not be correct.

Standardized Residuals

Example

Consider the Iris dataset, with the predictor variable (`petal length`) and response variable (`petal width`), and $n = 150$.

The quantile-quantile (QQ) plot contains in the y-axis the list of standardized residuals sorted from the smallest to the largest. The x-axis is the list of the quantiles of the standard normal distribution for a sample of size n, defined as

$$Q = (q_1, q_2, \ldots, q_n)^T \qquad q_i = F^{-1}\left(\frac{i - 0.5}{n}\right)$$

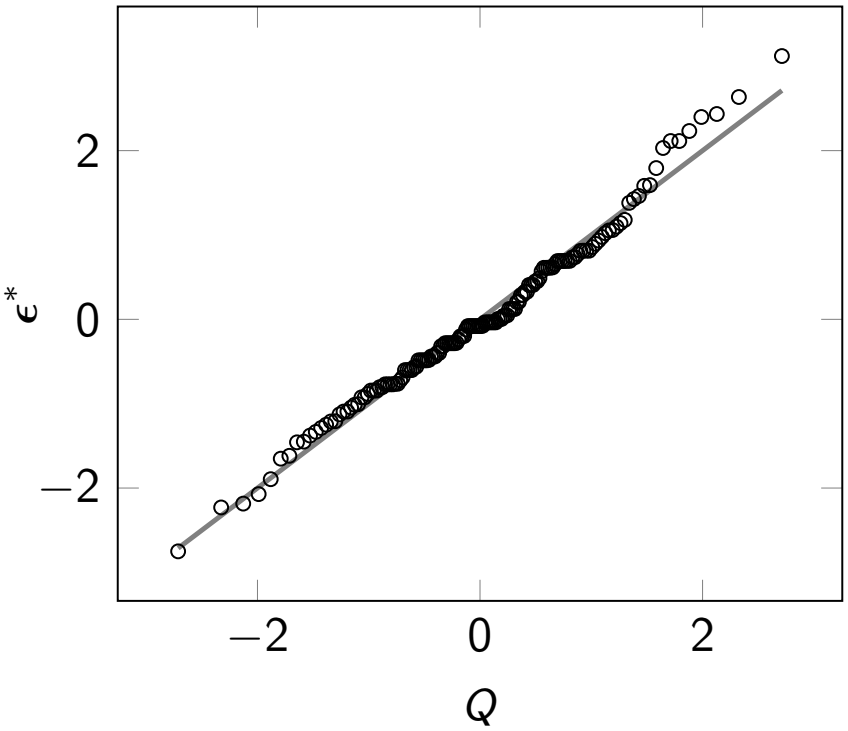
where F is the cumulative distribution function (CDF).

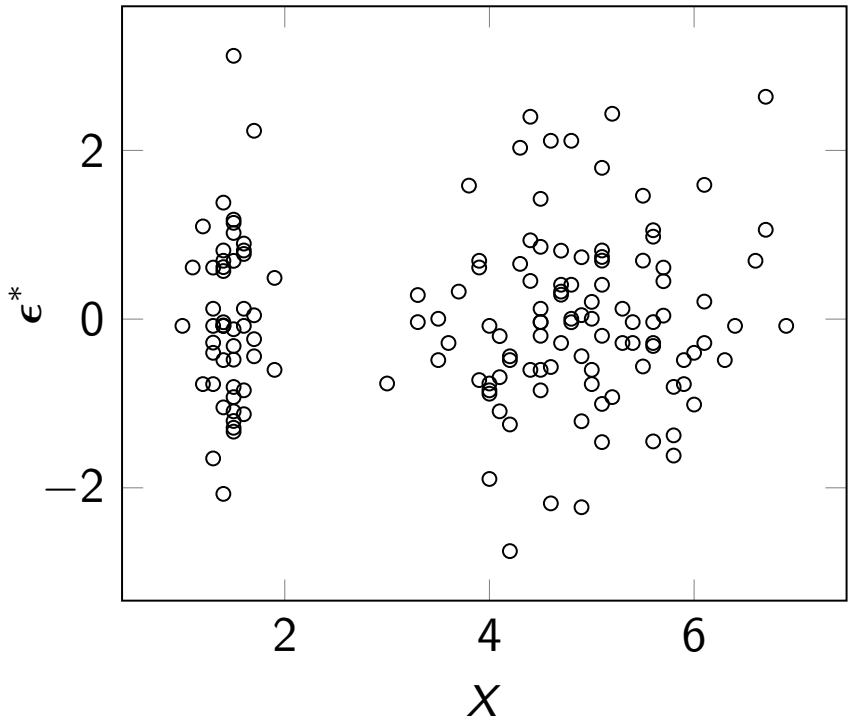
Thus, the Q values are also sorted in increasing order.If the standardized residuals follow a normal distribution, then the QQ plot should follow a straight line.

The plot of the independent variable X versus the standardized residuals is also instructive.

Residual plots

The absense of a particular trend or pattern to the residuals, and the residual values being concentrated along the mean value of 0, with the majority of the points being within two standard deviations of the mean, is expected if they were sampled from a normal distribution.





Multiple Regression

In multiple regression there are multiple independent attributes $X_1, X_2, \cdots, X_d$ and a single dependent or response attribute Y, and we assume that the true relationship can be modeled as a linear function

$$Y = \beta + \omega_1 \cdot X_1 + \omega_2 \cdot X_2 + \ldots + \omega_d \cdot X_d + \varepsilon$$

The augmented vector of estimated weights, including the bias term, is

$$\tilde{\boldsymbol{w}} = (w_0, w_1, \cdots, w_d)^T$$

We then make predictions for any given point $\boldsymbol{x}_i$ as follows:

$$\hat{y}_i = b \cdot 1 + w_1 \cdot x_{i1} + \cdots w_d \cdot x_{id} = \tilde{\boldsymbol{w}}^T \tilde{\boldsymbol{x}}_i$$

Recall that these estimates are obtained by minimizing the SSE:

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} \left(y_i - b - \sum_{j=1}^{d} w_j \cdot x_{ij} \right)^2$$

Multiple Regression

The estimated variance $\hat{\sigma}^2$ is then given as

$$\hat{\sigma}^2 = \frac{SSE}{n-(d+1)} = \frac{1}{n-d-1} \cdot \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

We divide by $n-(d+1)$ to get an unbiased estimate, since $n-(d+1)$ is the number of degrees of freedom for estimating SSE.

In other words, out of the n training points, we need to estimate $d+1$ parameters, β and the ω_i's, with $n-(d+1)$ remaining degrees of freedom.

Goodness of Fit

The decomposition of the total sum of squares, TSS, into the sum of squared errors, SSE, and the residual sum of squares, RSS, holds true for multiple regression as well:

$$TSS = SSE + RSS$$

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

The *coefficient of multiple determination*, R^2, gives the goodness of fit, measured as the fraction of the variation explained by the linear model:

$$R^2 = 1 - \frac{SSE}{TSS} = \frac{TSS - SSE}{TSS} = \frac{RSS}{TSS}$$

One of the potential problems with the R^2 measure is that it is susceptible to increase as the number of attributes increase, even though the additional attributes may be uninformative.

Goodness of Fit

To counter the impact of the number of attributes, we can consider the *adjusted coefficient of determination*, which takes into account the degrees of freedom in both TSS and SSE

$$R_a^2 = 1 - \frac{SSE/(n-d-1)}{TSS/(n-1)} = 1 - \frac{(n-1)\cdot SSE}{(n-d-1)\cdot TSS}$$

We can observe that the adjusted R_a^2 measure is always less than R^2, since the ratio $\frac{n-1}{n-d-1} > 1$.

If there is too much of a difference between R^2 and R_a^2, it might indicate that there are potentially many, possibly irrelevant, attributes being used to fit the model.

Multiple Regression: Goodness of Fit

Consider the multiple regression of `sepal length` (X_1) and `petal length` (X_2) on the response attribute `petal width` (Y) for the Iris dataset with $n = 150$ points. The uncentered 3×3 scatter matrix $\tilde{\boldsymbol{D}}^T\tilde{\boldsymbol{D}}$ and its inverse are given as

$$\tilde{\boldsymbol{D}}^T\tilde{\boldsymbol{D}} = \begin{pmatrix} 150.0 & 876.50 & 563.80 \\ 876.5 & 5223.85 & 3484.25 \\ 563.8 & 3484.25 & 2583.00 \end{pmatrix} \quad (\tilde{\boldsymbol{D}}^T\tilde{\boldsymbol{D}})^{-1} = \begin{pmatrix} 0.793 & -0.176 & 0.064 \\ -0.176 & 0.041 & -0.017 \\ 0.064 & -0.017 & 0.009 \end{pmatrix}$$

The augmented estimated weight vector $\tilde{\boldsymbol{w}}$ is given as

$$\tilde{\boldsymbol{w}} = \begin{pmatrix} w_0 \\ w_1 \\ w_2 \end{pmatrix} = (\tilde{\boldsymbol{D}}^T\tilde{\boldsymbol{D}})^{-1} \cdot (\tilde{\boldsymbol{D}}^T Y) = \begin{pmatrix} -0.014 \\ -0.082 \\ 0.45 \end{pmatrix}$$

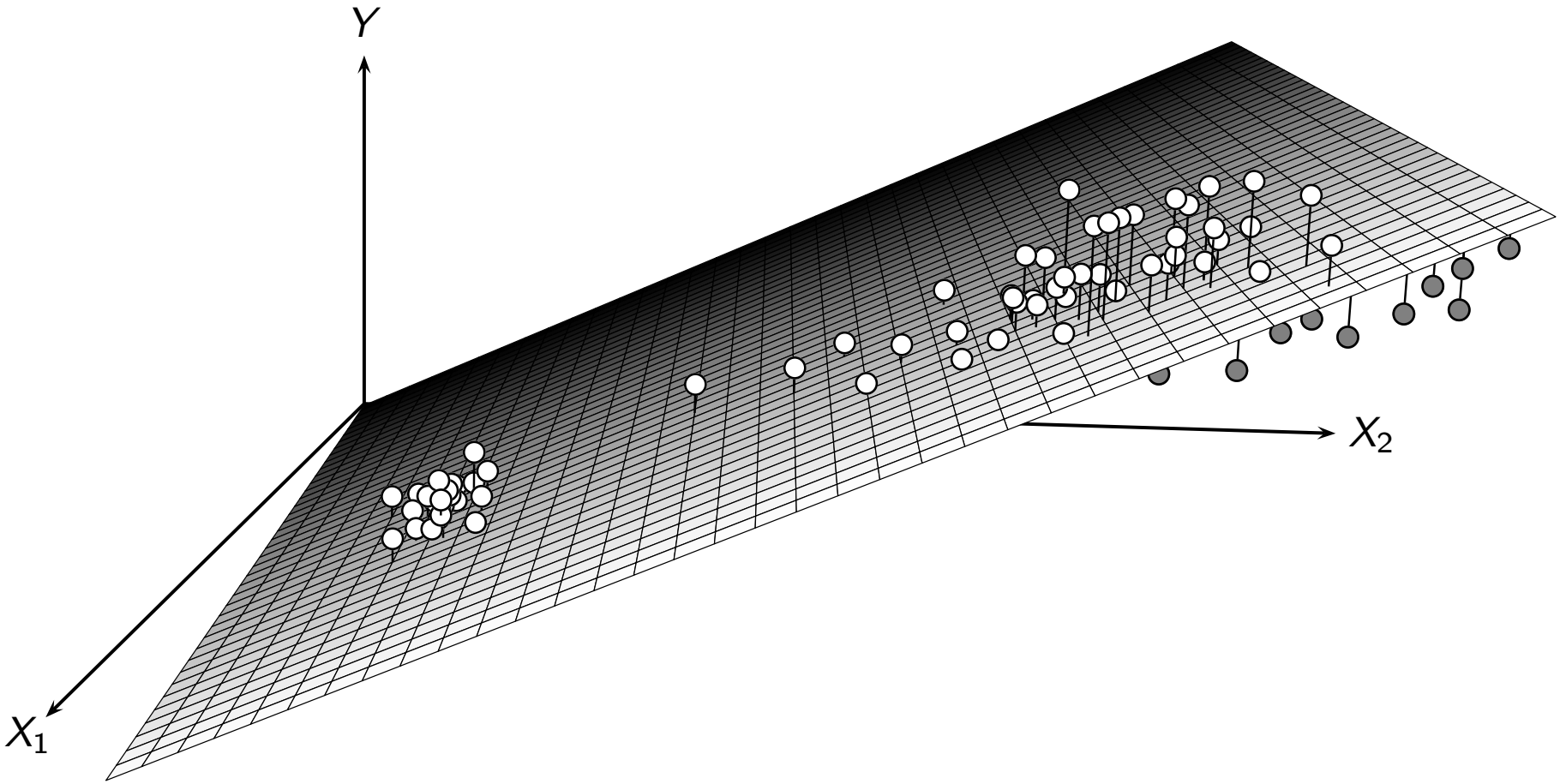
The bias term is therefore $b = w_0 = -0.014$, and the fitted model is

$$\hat{Y} = -0.014 - 0.082 \cdot X_1 + 0.45 \cdot X_2$$

Multiple Regression

Example

`sepal length` (X_1) and `petal length` (X_2) with response attribute `petal width` (Y). The vertical bars show the residual errors for the points. Points in white are above the plane, whereas points in gray are below the plane.



Multiple Regression: Goodness of Fit

The SSE value is given as

$$SSE = \sum_{i=1}^{150} \epsilon_i^2 = \sum_{i=1}^{150} (y_i - \hat{y}_i)^2 = 6.179$$

Thus, standard error of regression is given as

$$\hat{\sigma} = \sqrt{\frac{SSE}{n-d-1}} = \sqrt{4.203 \times 10^{-2}} = 0.205$$

The values of total and residual sum of squares are given as

$$TSS = 86.78 \qquad RSS = 80.60$$

The fraction of variance explained by the model, that is the R^2 value, is given as

$$R^2 = \frac{RSS}{TSS} = \frac{80.60}{86.78} = 0.929$$

It makes sense to also consider the adjusted R_a^2 value

$$R_a^2 = 1 - \frac{(n-1) \cdot SSE}{(n-d-1) \cdot TSS} = 1 - \frac{149 \times 6.179}{147 \times 86.78} = 0.928$$

The adjusted value is almost the same as the R^2 value.

Geometry of Goodness of Fit

In multiple regression there are d predictor attributes $X_1, X_2, \cdots, X_d$. We can center them by subtracting their projection along the vector 1 to obtain the centered predictor vectors $\overline{X}_i$. Likewise, we can center the response vector Y and the predicted vector $\widehat{Y}$. Thus, we have

$$\overline{X}_i = X_i - \mu_{X_i} \cdot 1 \qquad \overline{Y} = Y - \mu_Y \cdot 1 \qquad \overline{\widehat{Y}} = \widehat{Y} - \mu_Y \cdot 1$$

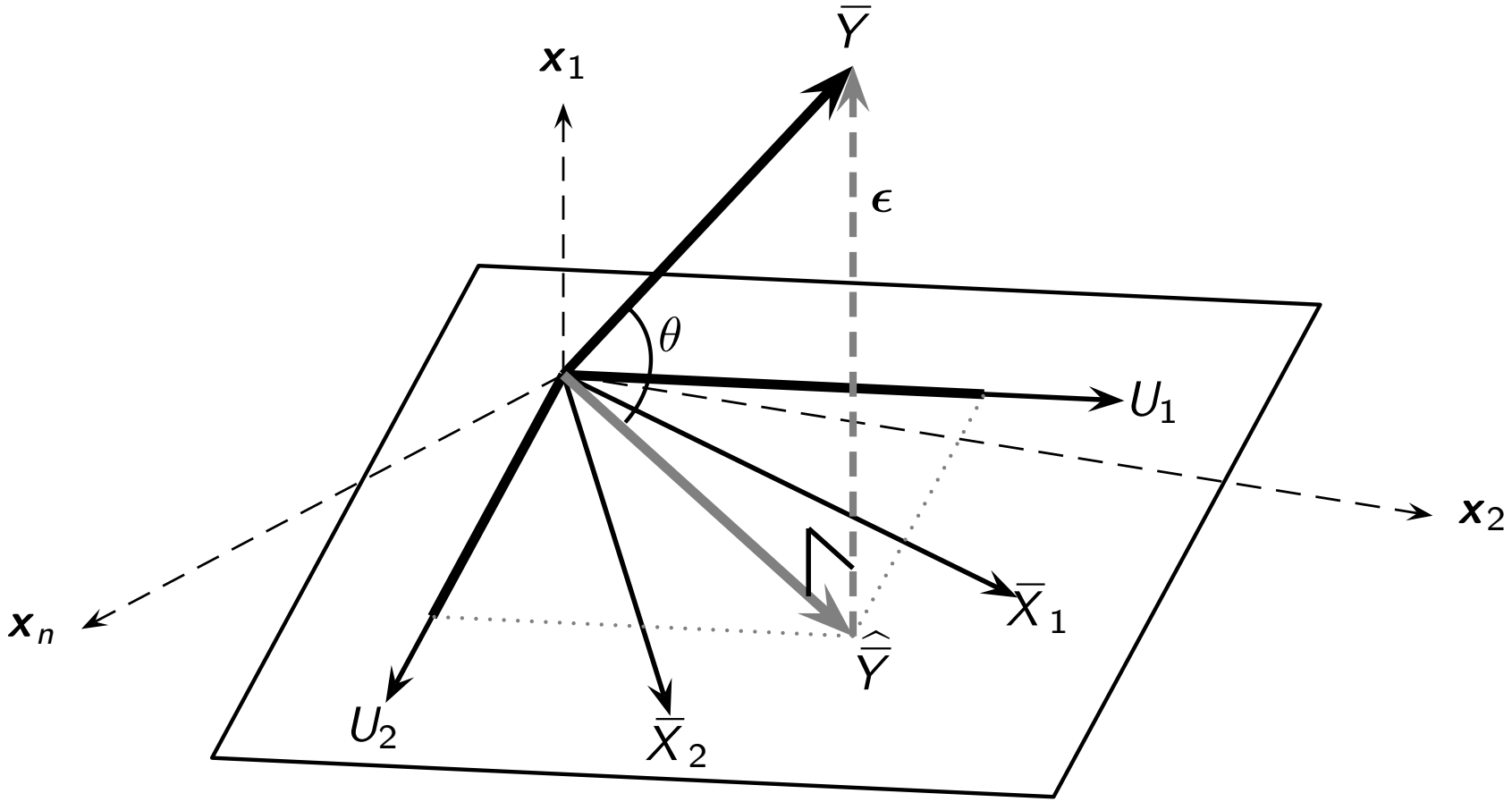
Once Y, $\widehat{Y}$ and X_i's have been centered, they all lie in the $n-1$ dimensional subspace orthogonal to the vector 1.

In this subspace, we first extract an orthogonal basis $\{U_1, U_2, \cdots, U_d\}$ via the Gram-Schmidt orthogonalization process and the predicted response vector is the sum of the projections of $\overline{Y}$ onto each of the new basis vectors.

The centered vectors $\overline{Y}$ and $\overline{\widehat{Y}}$, and the error vector ϵ form a right triangle.

Geometry of Multiple Regression

The figure shows two centered predictor variables $\overline{X}_1$ and $\overline{X}_2$, along with the corresponding orthogonal basis vectors U_1 and U_2. The dimensionality of the error space, containing the vector ϵ, is $n - d - 1$, which also specifies the degrees of freedom for the estimated variance $\hat{\sigma}^2$.



Geometry of Goodness of Fit

By the Pythagoras theorem, we have

$$\|\overline{Y}\|^2 = \left\|\widehat{\overline{Y}}\right\|^2 + \|\epsilon\|^2 = \left\|\widehat{\overline{Y}}\right\|^2 + \left\|Y - \widehat{Y}\right\|^2$$

$$TSS = RSS + SSE$$

The correlation between Y and $\widehat{Y}$ is the cosine of the angle between $\overline{Y}$ and $\widehat{\overline{Y}}$, which is also given as the ratio of the base to the hypotenuse

$$\rho_{Y\widehat{Y}} = \cos\theta = \frac{\left\|\widehat{\overline{Y}}\right\|}{\|\overline{Y}\|}$$

The coefficient of multiple determination is given as

$$R^2 = \frac{RSS}{TSS} = \frac{\left\|\widehat{\overline{Y}}\right\|^2}{\|\overline{Y}\|^2} = \rho^2_{Y\widehat{Y}}$$

Geometry of Goodness of Fit

Example

The correlation between Y and $\widehat{Y}$ is given as

$$\rho_{Y\widehat{Y}} = \cos\theta = \frac{\left\|\widehat{\overline{Y}}\right\|}{\left\|\overline{Y}\right\|} = \frac{\sqrt{RSS}}{\sqrt{TSS}} = \frac{\sqrt{80.60}}{\sqrt{86.78}} = 0.964$$

The angle between Y and $\widehat{Y}$ is given as

$$\theta = \cos^{-1}(0.964) = 15.5^\circ$$

The relatively small angle indicates a good linear fit.

Inferences about Regression Coefficients

Let Y be the response vector over all observations. Let $\tilde{\boldsymbol{w}} = (w_0, w_1, w_2, \cdots, w_d)^T$ be the estimated vector of regression coefficients, computed as

$$\tilde{\boldsymbol{w}} = \left(\tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}}\right)^{-1} \tilde{\boldsymbol{D}}^T Y$$

Note that $\tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}} \in \mathbb{R}^{(d+1)\times(d+1)}$ is the uncentered scatter matrix for the augmented data. Let $\boldsymbol{C}$ denote the inverse of $\tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}}$. That is

$$(\tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}})^{-1} = \boldsymbol{C}$$

Therefore, the covariance matrix for $\tilde{\boldsymbol{w}}$ can be written as

$$\text{cov}(\tilde{\boldsymbol{w}}) = \sigma^2 \boldsymbol{C}$$

In particular, the diagonal entries $\sigma^2 \cdot c_{ii}$ give the variance for each of the regression coefficient estimates (including for $b = w_0$), and their squared roots specify the standard errors.

$$\text{var}(w_i) = \sigma^2 \cdot c_{ii} \qquad \text{se}(w_i) = \sqrt{\text{var}(w_i)} = \sigma \cdot \sqrt{c_{ii}}$$

Inferences about Regression Coefficients

We can now define the standardized variable Z_{w_i} that can be used to derive the confidence intervals for ω_i as follows

$$Z_{w_i} = \frac{w_i - E[w_i]}{\text{se}(w_i)} = \frac{w_i - \omega_i}{\hat{\sigma}\sqrt{c_{ii}}}$$

where we have replaced the unknown true variance σ^2 by $\hat{\sigma}^2$. Each of the variables Z_{w_i} follows a t-distribution with $n-d-1$ degrees of freedom, from which we can obtain the $100(1-\alpha)\%$ confidence interval of the true value ω_i as follows:

$$P\left(w_i - t_{\alpha/2} \cdot \text{se}(w_i) \leq \omega_i \leq w_i + t_{\alpha/2} \cdot \text{se}(w_i)\right) = 1 - \alpha$$

Here, $t_{\alpha/2}$ is the critical value of the t distribution, with $n-d-1$ degrees of freedom, that encompasses $\alpha/2$ fraction of the probability mass in the right tail, given as

$$P(Z \geq t_{\alpha/2}) = \frac{\alpha}{2} \text{ or equivalently } T_{n-d-1}(t_{\alpha/2}) = 1 - \frac{\alpha}{2}$$

Confidence Intervals

Example

Continuing with multiple regression from last example, we have

$$\hat{\sigma}^2 = 4.203 \times 10^{-2}$$

$$\boldsymbol{C} = (\tilde{\boldsymbol{D}}^T \tilde{\boldsymbol{D}})^{-1} = \begin{pmatrix} 0.793 & -0.176 & 0.064 \\ -0.176 & 0.041 & -0.017 \\ 0.064 & -0.017 & 0.009 \end{pmatrix}$$

Therefore, the covariance matrix of the estimated regression parameters is:

$$\text{cov}(\tilde{\boldsymbol{w}}) = \hat{\sigma}^2 \cdot \boldsymbol{C} = \begin{pmatrix} 3.333 \times 10^{-2} & -7.379 \times 10^{-3} & 2.678 \times 10^{-3} \\ -7.379 \times 10^{-3} & 1.714 \times 10^{-3} & -7.012 \times 10^{-4} \\ 2.678 \times 10^{-3} & -7.012 \times 10^{-4} & 3.775 \times 10^{-4} \end{pmatrix}$$

The diagonal entries give the variances and standard errors:

$$\text{var}(b) = 3.333 \times 10^{-2} \qquad \text{se}(b) = \sqrt{3.333 \times 10^{-2}} = 0.183$$

$$\text{var}(w_1) = 1.714 \times 10^{-3} \qquad \text{se}(w_1) = \sqrt{1.714 \times 10^{-3}} = 0.0414$$

$$\text{var}(w_2) = 3.775 \times 10^{-4} \qquad \text{se}(w_2) = \sqrt{3.775 \times 10^{-4}} = 0.0194$$

where $b = w_0$.

Confidence Intervals

Using confidence level $1-\alpha = 0.95$ (or significance level $\alpha = 0.05$), the critical value of the t-distribution that encompasses $\frac{\alpha}{2} = 0.025$ fraction of the probability mass in the right tail is given as $t_{\alpha/2} = 1.976$.

Thus, the 95% confidence intervals for the true bias term β, and the true regression coefficients ω_1 and ω_2, are:

$$\begin{aligned}
\beta \in \big(b \pm t_{\alpha/2} \cdot \mathsf{se}(b)\big) &= (-0.014 - 0.074, -0.014 + 0.074) \\
&= (-0.088, 0.06) \\
\omega_1 \in \big(w_1 \pm t_{\alpha/2} \cdot \mathsf{se}(w_1)\big) &= (-0.082 - 0.0168, -0.082 + 0.0168) \\
&= (-0.099, -0.065) \\
\omega_2 \in \big(w_2 \pm t_{\alpha/2} \cdot \mathsf{se}(w_2)\big) &= (0.45 - 0.00787, 0.45 + 0.00787) \\
&= (0.442, 0.458)
\end{aligned}$$

Hypothesis Testing

Once the parameters have been estimated, it is beneficial to test whether the regression coefficients are close to zero or substantially different.

For this we set up the null hypothesis that all the true weights are zero, except for the bias term ($\beta = \omega_0$). We contrast the null hypothesis with the alternative hypothesis that at least one of the weights is not zero

$$H_0: \quad \omega_1 = 0, \omega_2 = 0, \ldots, \omega_d = 0 \qquad\qquad H_a: \quad \exists i, \text{ such that } \omega_i \neq 0$$

We use the F-test that compares the ratio of the adjusted RSS value to the estimated variance $\hat{\sigma}^2$, defined via the f-statistic

$$f = \frac{RSS/d}{\hat{\sigma}^2} = \frac{RSS/d}{SSE/(n-d-1)}$$

Hypothesis Testing

The ratio f follows a F-distribution with $d, (n-d-1)$ degrees of freedom for the numerator and denominator, respectively. Therefore, we can reject the null hypothesis if the *p-value* is less than the chosen significance level, say $\alpha = 0.01$. Notice that, since $R^2 = 1 - \frac{SSE}{TSS} = \frac{RSS}{TSS}$, we have

$$SSE = (1 - R^2) \cdot TSS \qquad RSS = R^2 \cdot TSS$$

Therefore, we can rewrite the f ratio as follows

$$f = \frac{RSS/d}{SSE/(n-d-1)} = \frac{n-d-1}{d} \cdot \frac{R^2}{1-R^2}$$

In other words, the F-test compares the adjusted fraction of explained variation to the unexplained variation. If R^2 is high, it means the model can fit the data well, and that is more evidence to reject the null hypothesis.

Hypothesis Testing for Individual Parameters

We can also test whether each independent attribute X_i, contributes significantly for the prediction of Y or not, assuming that all the attributes are still retained in the model.

For attribute X_i, we set up the null hypothesis $H_0 : \omega_i = 0$ and contrast it with the alternative hypothesis $H_a : \omega_i \neq 0$. The standardized variable Z_{w_i} under the null hypothesis is given as

$$Z_{w_i} = \frac{w_i - E[w_i]}{\text{se}(w_i)} = \frac{w_i}{\text{se}(w_i)} = \frac{w_i}{\hat{\sigma}\sqrt{c_{ii}}}$$

If this probability is smaller than the significance level α (say 0.01), we can reject the null hypothesis.

Otherwise, we accept the null hypothesis, which would imply that X_i does not add significant value in predicting the response in light of other attributes already used to fit the model. The t-test can also be used to test whether the bias term is significantly different from 0 or not.

Under the null hypothesis that $\omega_1 = \omega_2 = 0$, the expected value of RSS is σ^2. Thus, we expect the f-statistic to be close to 1. Let us check if that is the case; we have

$$f = \frac{RSS/d}{\hat{\sigma}^2} = \frac{80.60/2}{4.203 \times 10^{-2}} = 958.8$$

Using the F-distribution with $(d, n-d-1) = (2, 147)$ degrees of freedom, we have

$$p\text{-}value(958.8) \simeq 0$$

In other words, such a large value of the f-statistic is extremely rare, and therefore, we can reject the null hypothesis. We conclude that Y does indeed depend on at least one of the predictor attributes X_1 or X_2.

Hypothesis Testing for Individual Parameters

Example

We can also test for each of the regression coefficients individually using the t-test. For example, for ω_1, let the null hypothesis be $H_0 : \omega_1 = 0$, so that the alternative hypothesis is $H_a : \omega_1 \neq 0$. Assuming that the model still has both X_1 and X_2 as the predictor variables, we can compute the t-statistic as:

$$Z_{w_1} = \frac{w_1}{\text{se}(w_1)} = \frac{-0.082}{0.0414} = -1.98$$

Using a two-tailed t-test with $n - d - 1 = 147$ degrees of freedom, we find that

$$p\text{-}value(-1.98) = 0.0496$$

Since the *p-value* is only marginally less than a significance level of $\alpha = 0.05$ (i.e., a 95% confidence level), this means that X_1 is only weakly relevant for predicting Y in the presence of X_2.

In fact, if we use the more stringent significance level $\alpha = 0.01$, we would conclude that X_1 is not significantly predictive of Y, given X_2.

Hypothesis Testing for Individual Parameters

Example

On the other hand, individually for ω_2, if we test whether $H_0 : \omega_2 = 0$ versus $H_a : \omega_2 \neq 0$, we have:

$$Z_{w_2} = \frac{w_2}{\text{se}(w_2)} = \frac{0.45}{0.0194} = 23.2$$

Using a two-tailed t-test with $n - d - 1 = 147$ degrees of freedom, we find that

$$p\text{-}value(23.2) \simeq 0$$

Thus, X_2 is significantly predictive of Y even in the presence of X_1.

Using the t-test, we can also compute the p-*value* for the bias term:

$$Z_b = \frac{b}{\text{se}(b)} = \frac{-0.014}{0.183} = -0.077$$

which has a p-*value* $= 0.94$ for a two-tailed test. This means, we accept the null hypothesis that $\beta = 0$, and reject the alternative hypothesis that $\beta \neq 0$.

Geometric Approach to Statistical Testing

The geometry of multiple regression provides further insight into the hypothesis testing approach for the regression effect.

Let $\overline{X}_i = X_i - \mu_{X_i} \cdot 1$ denote the centered attribute vector, and let $\overline{X} = (\overline{X}_1, \overline{X}_2, \cdots, \overline{X}_d)^T$ denote the multivariate centered vector of predictor variables.

The n-dimensional space over the points is divided into three mutually orthogonal subspaces, namely the 1-dimensional *mean space* $\mathcal{S}_\mu = span(1)$, the d dimensional *centered variable space* $\mathcal{S}_{\overline{X}} = span(\overline{X})$, and the $n-d-1$ dimensional *error space* $\mathcal{S}_\epsilon$, which contains the error vector $\boldsymbol{\epsilon}$. The response vector Y can thus be decomposed into three components

$$Y = \mu_Y \cdot 1 + \widehat{\overline{Y}} + \boldsymbol{\epsilon}$$

Recall that the *degrees of freedom* of a random vector is defined as the dimensionality of its enclosing subspace. Since the original dimensionality of the point space is n, we have a total of n degrees of freedom.

$$dim(\mathcal{S}_\mu) + dim(\mathcal{S}_{\overline{X}}) + dim(\mathcal{S}_\epsilon) = 1 + d + (n - d - 1) = n$$

Population Regression Model

The regression model posits that $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \cdots, x_{id})^T$:

$$y_i = \beta + \omega_1 \cdot x_{i1} + \ldots + \omega_d \cdot x_{id} + \varepsilon_i$$

where the error term ε_i varies randomly, with the assumption that ε_i follows a normal distribution with mean $\mu = 0$ and variance σ^2. The distribution of $\boldsymbol{\varepsilon}$ is therefore given as

$$f(\boldsymbol{\varepsilon}) = \frac{1}{(\sqrt{2\pi})^n \cdot \sqrt{|\Sigma|}} \cdot \exp\left\{-\frac{\boldsymbol{\varepsilon}^T \Sigma^{-1} \boldsymbol{\varepsilon}}{2}\right\} = \frac{1}{(\sqrt{2\pi})^n \cdot \sigma^n} \cdot \exp\left\{-\frac{\|\boldsymbol{\varepsilon}\|^2}{2 \cdot \sigma^2}\right\}$$

which follows from the fact that $|\Sigma| = \det(\Sigma) = \det(\sigma^2 \boldsymbol{I}) = \left(\sigma^2\right)^n$ and $\Sigma^{-1} = \frac{1}{\sigma^2}\boldsymbol{I}$.

The density of $\boldsymbol{\varepsilon}$ is thus a function of its squared length $\|\boldsymbol{\varepsilon}\|^2$, independent of its angle.

In other words, the vector $\boldsymbol{\varepsilon}$ is distributed uniformly over all angles and is equally likely to point in any direction.

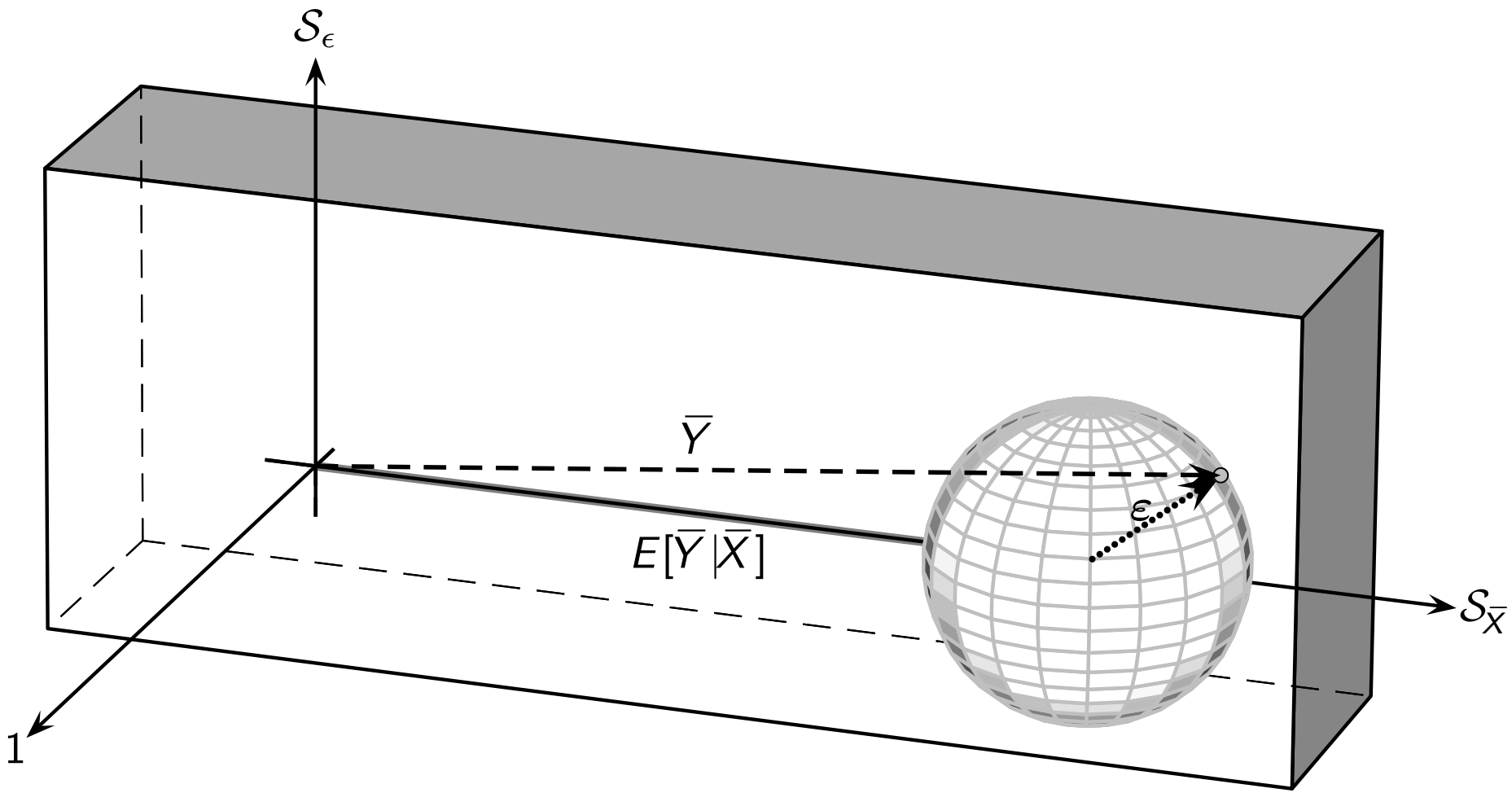
Population Regression Model

It is important to note that the $n-1$ dimensional hypersphere denotes the fact that the random vector ε can be in any orientation in this hypersphere of radius $\|\varepsilon\|$.

Notice how the population regression model differs from the fitted model. The residual error vector ϵ is orthogonal to the predicted mean response vector $\widehat{\overline{Y}}$, which acts as the estimate for $E[\overline{Y}|\overline{X}]$.

On the other hand, in the population regression model, the random error vector ε can be in any orientation compared to $E[\overline{Y}|\overline{X}]$.

Population Regression Model



Hypothesis Testing

Consider the population regression model

$$Y = \mu_Y \cdot 1 + \omega_1 \cdot \bar{X}_1 + \ldots + \omega_d \cdot \bar{X}_d + \varepsilon = \mu_Y \cdot 1 + E[\bar{Y}|\bar{X}] + \varepsilon$$

To test whether $X_1, X_2, \cdots, X_d$ are useful in predicting Y, we consider what would happen if all of the regression coefficients were zero, which forms our null hypothesis

$$H_0: \quad \omega_1 = 0, \omega_2 = 0, \ldots, \omega_d = 0$$

In this case, we have

$$Y = \mu_Y \cdot 1 + \varepsilon \implies Y - \mu_Y \cdot 1 = \varepsilon \implies \bar{Y} = \varepsilon$$

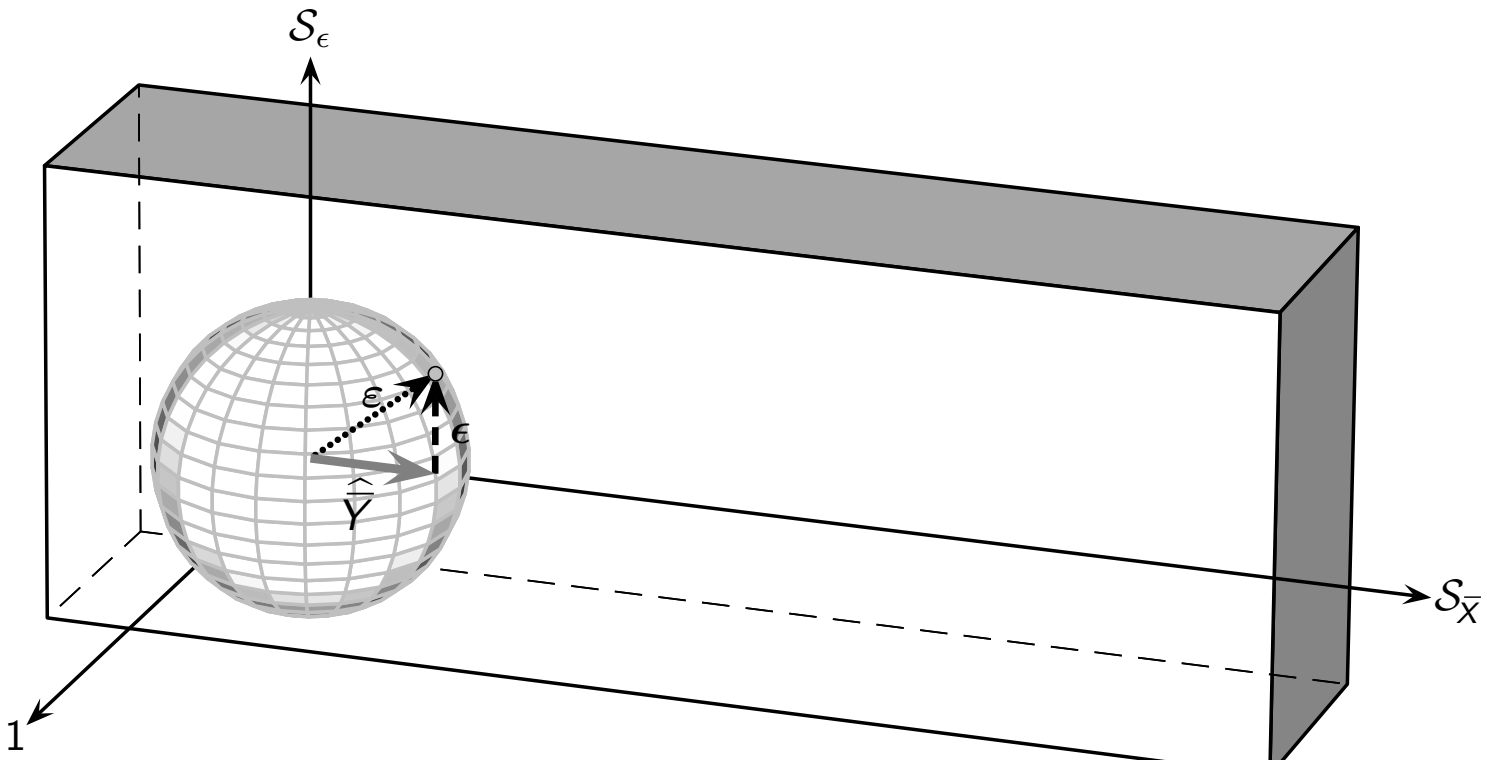
On the other hand, under the alternative hypothesis H_a that at least one of the ω_i is non-zero, we have

$$\bar{Y} = E[\bar{Y}|\bar{X}] + \varepsilon$$

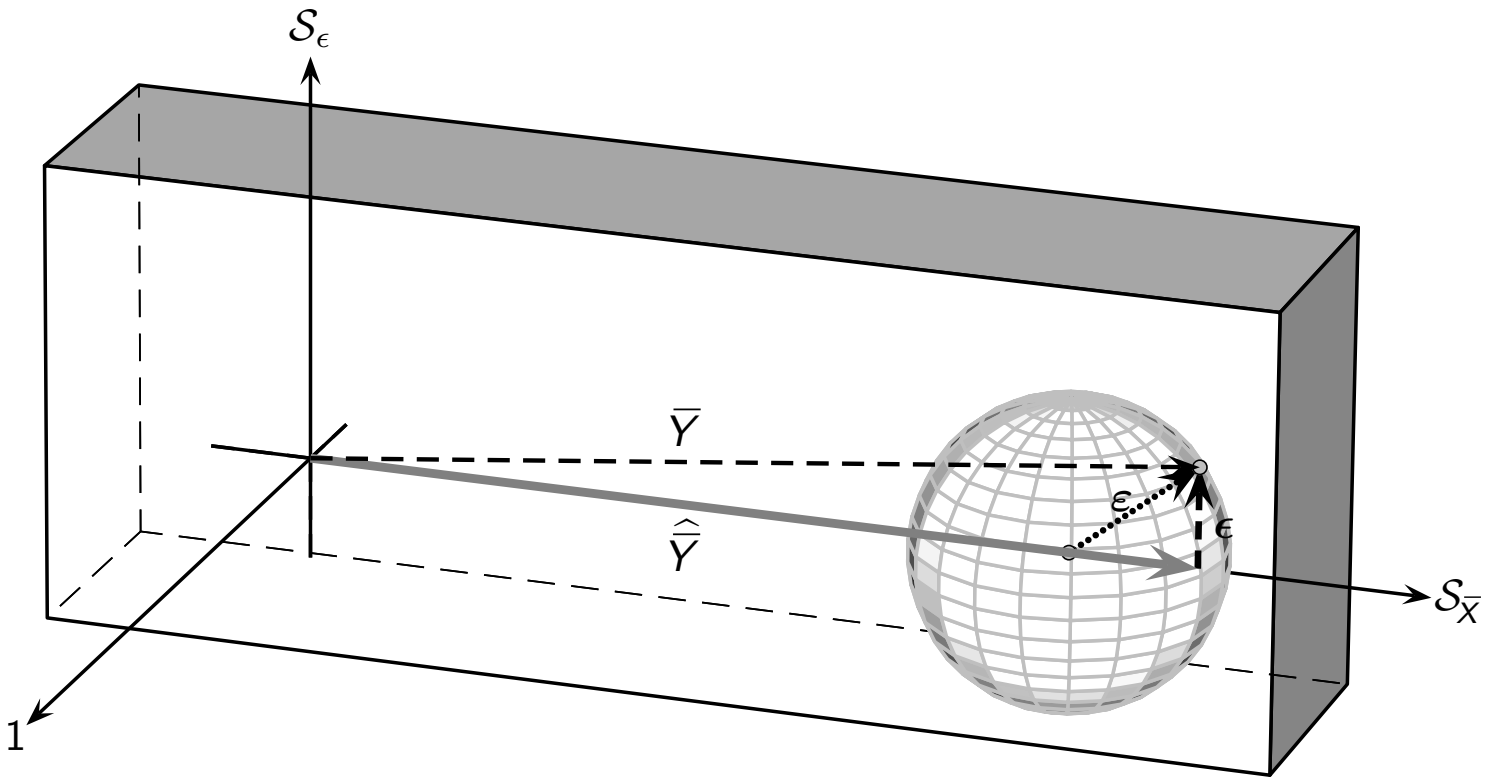
Thus, the variation in $\bar{Y}$ is shifted away from the origin 0 in the direction $E[\bar{Y}|\bar{X}]$. In practice, we can estimate $E[\bar{Y}|\bar{X}]$ by projecting the $\bar{Y}$ onto $\mathcal{S}_{\bar{X}}$ and $\mathcal{S}_\epsilon$:

$$\bar{Y} = w_1 \cdot \bar{X}_1 + w_2 \cdot \bar{X}_2 + \ldots + w_d \cdot \bar{X}_d + \epsilon = \widehat{\bar{Y}} + \epsilon$$

Under the null hypothesis, the variation in $\overline{Y}$ for a given value of $\boldsymbol{x}$ will therefore be centered around the origin 0.



Under the alternative hypothesis, we have $\overline{Y} = E[\overline{Y}|\overline{X}] + \varepsilon$, and so $\widehat{\overline{Y}}$ will be relatively much longer compared to ϵ.



Hypothesis Testing

Given that we expect to see a difference between $\widehat{\widehat{Y}}$ under the null and alternative hypotheses, this suggests a test based on the relative lengths of $\widehat{\widehat{Y}}$ and ϵ.

However, we cannot compare their lengths directly, since $\widehat{\widehat{Y}}$ lies in a d dimensional space, whereas ϵ lies in a $n-d-1$ dimensional space. Instead, we can compare their lengths after normalizing by the number of dimensions.

$$M(\widehat{\widehat{Y}}) = \frac{\left\|\widehat{\widehat{Y}}\right\|^2}{dim(\mathcal{S}_{\bar{X}})} = \frac{\left\|\widehat{\widehat{Y}}\right\|^2}{d} \qquad M(\epsilon) = \frac{\|\epsilon\|^2}{dim(\mathcal{S}_{\epsilon})} = \frac{\|\epsilon\|^2}{n-d-1}$$

Thus, the geometric test for the regression effect is:

$$\frac{M(\widehat{\widehat{Y}})}{M(\epsilon)} = \frac{\left\|\widehat{\widehat{Y}}\right\|^2 / d}{\|\epsilon\|^2 / (n-d-1)} = \frac{RSS/d}{SSE/(n-d-1)} = f$$

If $f \simeq 1$ then the null hypothesis holds, and Y does not depend on $X_1, X_2, \cdots, X_d$.

If f is large, with a $p\text{-}value < \alpha$ (e.g., 0.01), then we can reject the null hypothesis and accept the alternative hypothesis that Y depends on at least one X_i.

Data Mining and Machine Learning: Fundamental Concepts and Algorithms

dataminingbook.info

Mohammed J. Zaki[1] Wagner Meira Jr.[2]

[1]Department of Computer Science
Rensselaer Polytechnic Institute, Troy, NY, USA

[2]Department of Computer Science
Universidade Federal de Minas Gerais, Belo Horizonte, Brazil

Chapter 27: Regression Evaluation