# 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 $\omega$ is the slope of the best fitting line and $\beta$ is its intercept, and $\varepsilon$ is the random error variable that follows a normal distribution with mean $\mu = 0$ and variance $\sigma^2$. The true parameters $\beta$, $\omega$ and $\sigma^2$ are all unknown, and have to be estimated from the training data $\boldsymbol{D}$ 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 $\varepsilon$. 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

$$\hat{\sigma}^2 = \frac{SSE}{n-2} \tag{1}$$

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.

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. Therefore, the fraction of the variation left unexplained by the model is given by the ratio $\frac{SSE}{TSS}$. 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

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

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 variable) for the Iris dataset. Figure shows the scatterplot between the two attributes. There are a total of $n = 150$ data points. The least squares estimates for the bias and regression coefficients are as follows

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

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

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.

# Inference about Regression Coefficient and Bias Term

The estimated values of the bias and regression coefficient, $b$ and $w$, are only point estimates for the true parameters $\beta$ and $\omega$. 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 $\sigma^2$. On the other hand, the $x_i$ values are fixed *a priori* and therefore $\mu_X$ and $\sigma_X^2$ are also fixed values. We can now treat $b$ and $w$ as random variables, with

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

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

where $c_i$ is a constant (since $x_i$ is fixed), given as

$$c_i = \frac{x_i - \mu_X}{s_X} \tag{3}$$

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 $\mu_X$.

# 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 $\omega$.

# 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

$$\text{var}(w) = \text{var}\left(\sum_{i=1}^{n} c_i \cdot y_i\right) = \sum_{i=1}^{n} c_i^2 \cdot \text{var}(y_i) = \sigma^2 \cdot \sum_{i=1}^{n} c_i^2 = \frac{\sigma^2}{s_X} \tag{4}$$

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

$$\text{se}(w) = \sqrt{\text{var}(w)} = \frac{\sigma}{\sqrt{s_X}} \tag{5}$$

# 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 $\beta$.
Using the observation that all $y_i$ are independent, the variance of the bias term can be computed as follows

# 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

$$\text{se}(b) = \sqrt{\text{var}(b)} = \sigma \cdot \sqrt{\frac{1}{n} + \frac{\mu_X^2}{s_X}} \tag{6}$$

# 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 $\omega$ and variance $\sigma^2/s_X$. Likewise, $b$ follows a normal distribution with mean $\beta$ and variance $(1/n + \mu_X^2/s_X)\cdot\sigma^2$.
Since the true variance $\sigma^2$ is unknown, we use the estimated variance $\hat{\sigma}^2$, to define the standardized variables $Z_w$ and $Z_b$ as follows

$$Z_w = \frac{w - E[w]}{\text{se}(w)} = \frac{w-\omega}{\frac{\hat{\sigma}}{\sqrt{s_X}}} \qquad Z_b = \frac{b - E[b]}{\text{se}(b)} = \frac{b-\beta}{\hat{\sigma}\sqrt{(1/n + \mu_X^2/s_X)}} \tag{7}$$

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, $\omega$ and $\beta$, are therefore as follows

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

# 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 $\sigma^2$, 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, $\omega$, 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, $\beta$, 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 $\omega$, 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 $\beta$ and the random error $\varepsilon$. 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,

$$Z_w = \frac{w - E[w]}{\text{se}(w)} = \frac{w}{\hat{\sigma}/\sqrt{s_X}} \tag{8}$$

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 $\alpha$ (e.g., $\alpha = 0.01$), we reject the null hypothesis if the *p-value* is below $\alpha$. 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

$$f = \frac{RSS}{\hat{\sigma}^2} = \frac{\sum_{i=1}^{n}(\hat{y}_i - \mu_Y)^2}{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \Big/ (n-2)} \tag{9}$$

# 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 $\alpha$, 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:

$$Z_b = \frac{b - E[b]}{\text{se}(b)} = \frac{b}{\hat{\sigma} \cdot \sqrt{(1/n + \mu_X^2/s_X)}} \tag{10}$$

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 $\alpha$.

# 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 $H_a : \beta \neq 0$.

# 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