Finding a linear function based on X to best yield Y.

X = "covariate" = "feature" = "predictor" = "regressor" = "independent variable"

Y = "response variable" = "outcome" = "dependent variable"

Regression: r(x) = E(Y|X=x)

goal: estimate the function r

Finding a linear function based on X to best yield Y.

X = "covariate" = "feature" = "predictor" = "regressor" = "independent variable"

Y = "response variable" = "outcome" = "dependent variable"

Regression:
$$r(x) = E(Y|X = x)$$

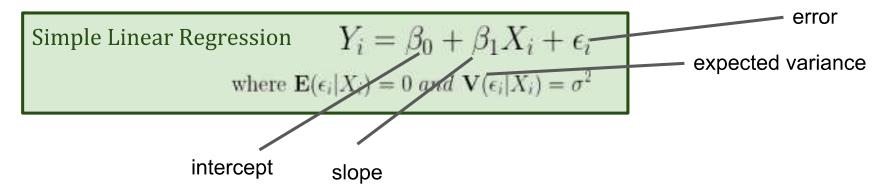
goal: estimate the function r

Linear Regression (univariate version): $r(x) = \beta_0 + \beta_1 x$

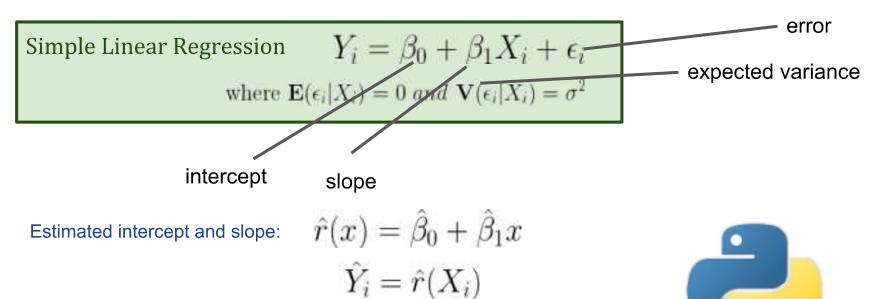
goal: find β_0 , β_1 such that $r(x) \approx \mathrm{E}(Y|X=x)$

Simple Linear Regression
$$Y_i=\beta_0+\beta_1X_i+\epsilon_i$$
 where $\mathbf{E}(\epsilon_i|X_i)=0$ and $\mathbf{V}(\epsilon_i|X_i)=\sigma^2$

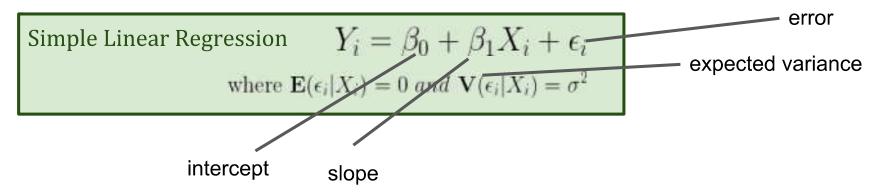
$$r(x) = \beta_0 + \beta_1 x$$



Residual:



 $\hat{\epsilon}_i = Y_i - \hat{Y}_i$



Estimated intercept and slope: $\hat{r}(z)$

$$\hat{r}(x) = \hat{\beta}_0 + \hat{\beta}_1 x$$

$$\hat{Y}_i = \hat{r}(X_i)$$

Residual: — $\hat{\epsilon}_i = Y_i - \hat{Y}_i$

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_{i}^{2} = \sum_{i=1}^{n} (Y_{i} - \hat{Y}_{i})^{2} = \sum_{i=1}^{n} (Y_{i} - \beta_{0} - \beta_{1}X_{i})^{2}$$

via Gradient Descent

Start with
$$\hat{\beta}_0 = \hat{\beta}_1 = 0$$

Repeat until convergence:

Calculate all \hat{Y}_i

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left(\sum_{i=1}^n \hat{Y}_i - Y_i\right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha (\sum_{i=1}^n X_i (\hat{Y}_i - Y_i))$$

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

via Gradient Descent

Start with $\hat{\beta}_0 = \hat{\beta}_1 = 0$

Repeat until convergence:

Calculate all \hat{Y}

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left(\sum_{i=1}^n \hat{Y}_i - Y_i \right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha \left(\sum_{i=1}^n X_i (\hat{Y}_i - Y_i) \right)$$

Learning rate

Based on derivative of RSS

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2$$

via Gradient Descent

Start with
$$\hat{\beta}_0 = \hat{\beta}_1 = 0$$

Repeat until convergence:

Calculate all \hat{Y}_i

$$\hat{\beta}_0 = \hat{\beta}_0 - \alpha \left(\sum_{i=1}^n \hat{Y}_i - Y_i\right)$$

$$\hat{\beta}_1 = \hat{\beta}_1 - \alpha (\sum_{i=1}^n X_i (\hat{Y}_i - Y_i))$$

via Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

$$RSS = \sum_{i=1}^{n} \hat{\epsilon}_{i}^{2} = \sum_{i=1}^{n} (Y_{i} - \hat{Y}_{i})^{2} = \sum_{i=1}^{n} (Y_{i} - \beta_{0} - \beta_{1}X_{i})^{2}$$

Pearson Product-Moment Correlation

Covariance

$$\begin{aligned} Cov(X,Y) &= \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) \\ &= \mathbf{E}\left((X - \bar{X})(Y - \bar{Y})\right) \end{aligned}$$

via Direct Estimates (normal equations)

$$\hat{\beta}_{1} = \frac{\sum_{i=1}^{n} (X_{i} - \bar{X})(Y_{i} - \bar{Y})}{\sum_{i=1}^{n} (X_{i} - \bar{X})^{2}}$$

$$\hat{\beta}_{0} = \bar{Y} - \hat{\beta}_{1}\bar{X}$$

Pearson Product-Moment Correlation

Covariance

$$Cov(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$

= $\mathbf{E}((X - \bar{X})(Y - \bar{Y}))$

Correlation

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$
$$= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X}\right) \left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

via Direct Estimates (normal equations)

$$\hat{\beta}_{1} = \frac{\sum_{i=1}^{n} (X_{i} - \bar{X})(Y_{i} - \bar{Y})}{\sum_{i=1}^{n} (X_{i} - \bar{X})^{2}}$$

$$\hat{\beta}_{0} = \bar{Y} - \hat{\beta}_{1}\bar{X}$$

Pearson Product-Moment Correlation

Covariance

$$Cov(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y)$$

= $\mathbf{E}((X - \bar{X})(Y - \bar{Y}))$

Correlation

$$r = r_{X,Y} = \frac{Cov(X,Y)}{s_X s_Y}$$
$$= \frac{1}{n-1} \sum_{i=1}^n \left(\frac{X_i - \bar{X}}{s_X}\right) \left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

via Direct Estimates (normal equations)

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

If one standardizes X and Y (i.e. subtract the mean and divide by the standard deviation) before running linear regression, then: $\hat{\beta}_0 = 0$ and $\hat{\beta}_1 = r$

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_m X_{m1} + \epsilon_i$

If we include and $X_{oi} = 1$ for all i (i.e. adding the intercept to X). Then we can say:

$$Y_i = \sum_{j=0}^m \beta_j X_{ij} + \epsilon_i$$

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_m X_{m1} + \epsilon_i$

If we include and $X_{oi} = 1$ for all *i*. Then we can say:

$$Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i$$

Or in vector notation across all i: $Y = X\beta + \epsilon$

Where β and ϵ are vectors and X is a matrix.

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_m X_{m1} + \epsilon_i$

If we include and $X_{oi} = 1$ for all *i*. Then we can say:

$$Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i$$

Or in vector notation across all i: $Y = X\beta + \epsilon$

Where β and ϵ are vectors and X is a matrix.

Estimating eta :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

Suppose we have multiple independent variables that we'd like to fit to our dependent variable: $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_m X_{m1} + \epsilon_i$

If we include and $X_{oi} = 1$ for all *i*. Then we can say:

$$Y_i = \sum_{j=0}^{m} \beta_j X_{ij} + \epsilon_i$$

To test for significance of individual Coefficient, *j*:

$$t = \frac{\hat{\beta}_{j}}{SE(\hat{\beta}_{j})} = \frac{\hat{\beta}_{j}}{\sqrt{\frac{s_{j}^{2}}{\sum_{i=1}^{n} (X_{ij} - \bar{X}_{j})^{2}}}}$$

Or in vector notation

across all i: $Y = X\beta + \epsilon$

Where β and ϵ are vectors and X is a matrix.

Estimating β :

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

Note: this is a probability here. In simple linear regression we wanted an expectation:

$$r(x) = \mathrm{E}(Y|X=x)$$

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

Note: this is a probability here. In simple linear regression we wanted an expectation:

$$r(x) = \mathrm{E}(Y|X=x)$$

(i.e. if p > 0.5 we can confidently predict $Y_i = 1$)

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

$$logit(p_i) = log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$$

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1|X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$

$$logit(p_i) = log \left(\underbrace{\frac{p_i}{1 - p_i}} \right) = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$$

$$P(Y_i = 0 \mid X = x)$$

What if $Y_i \in \{0, 1\}$? (i.e. we want "classification")

$$p_i \equiv p_i(\beta) \equiv \mathbf{P}(Y_i = 1 | X = x) = \frac{e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}{1 + e^{\beta_0 + \sum_{j=1}^m \beta_j x_{ij}}}$$
$$logit(p_i) = log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \sum_{j=1}^m \beta_j x_{ij}$$

To estimate $\beta \,$, one can use

reweighted least squares:

set $\hat{\beta}_0 = ... = \hat{\beta}_m = 0$ (remember to include an intercept)

1. Calculate p_i and let W be a diagonal matrix where element $(i, i) = p_i(1 - p_i)$.

2. Set
$$z_i = logit(p_i) + \frac{Y_i - p_i}{p_i(1 - p_i)} = X\hat{\beta} + \frac{Y_i - p_i}{p_i(1 - p_i)}$$

3. Set $\hat{\beta} = (X^T W X)^{-1} X^T W z$ //weighted lin. reg. of Z on Y.

4. Repeat from 1 until $\hat{\beta}$ converges.

(Wasserman, 2005; Li, 2010)