Simple linear regression

- Find a function $f : X \to Y$ map input space X to output space Y.
- Predicting a quantitative response Y on the basis of a single predictor variable X.
- $Y \approx \beta_0 + \beta_1 X$, β_0 and β_1 are model coefficients or parameters.
- Once we have used training data to produce estimates $\hat{\beta}_0$ and $\hat{\beta}_1$, we can make prediction by: $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 X$.

SLR

- Define $e_i = y_i - \hat{y}_i$ be the ith residual (difference between the observed response value and the response value predicted by our model).
- Residual sum of squares (RSS) ass: $RSS = \sum_{i=1}^{n} e_i$ or

$$RSS = \sum_{i=1}^{n} (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2.$$

- Need to choose β_0 and β_1 that minimize RSS. Remember optimality conditions? (convex, differentiable function...)

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

Set this quantity to 0, we get $\sum_{i=1}^{n} y_i = n\beta_0 + \beta_1 \sum_{i=1}^{n} x_i$.

SLR

- $\sum_{i=1}^n y_i = n\beta_0 + \beta_1 \sum_{i=1}^n x_i \to \beta_0 = \frac{\sum_{i=1}^n y_i - \beta_1 \sum_{i=1}^n x_i}{n}$.
- Let $\bar{y} = \frac{\sum_{i=1}^n y_i}{n}$ and $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ so $\beta_0 = \bar{y} - \beta_1 \bar{x}$
- Do the same thing with β_1:

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

We get

$$\begin{aligned}\sum_{i=1}^n y_i x_i &= \beta_0 \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2 \\ &= (\bar{y} - \beta_1 \bar{x}) \sum_{i=1}^n x_i + \beta_1 \sum_{i=1}^n x_i^2\end{aligned}$$

SLR

- The solution for β_1:
$$\beta_1 = \frac{\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x}}{\sum_{i=1}^{n} x_i^2 - n\bar{x}^2}$$
Notice that: $\sum_{i=1}^{n}(\bar{x}^2 - \bar{x}x_i) = 0$ and $\sum_{i=1}^{n} \bar{x}\bar{y} - y_i\bar{x} = 0$
$$\begin{aligned}\sum_{i=1}^{n} x_i^2 - n\bar{x}^2 &= \sum_{i=1}^{n}(x_i^2 - x_i\bar{x}) + \sum_{i=1}^{n}(\bar{x}^2 - \bar{x}x_i)\\ &= \sum_{i=1}^{n}(x_i - \bar{x})^2.\end{aligned}$$
$$\begin{aligned}\sum_{i=1}^{n} x_i y_i - n\bar{y}\bar{x} &= \sum_{i=1}^{n}(x_i y_i - y_i\bar{x}) + \sum_{i=1}^{n}(\bar{x}\bar{y} - y_i\bar{x}).\\ &= \sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}.)\end{aligned}$$

SLR

- $$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$
- These solutions are least square estimates of parameters β_0 and β_1.

Assessing model

- $Y = f(X) + \epsilon$. If f is approximated by a linear relationship: $Y = \beta_0 + \beta_1 X + \epsilon$.
- Assuming the error term is independent of X.
- Least square estimates are unbiased. That means on average, if we do the estimates on many data sets, the average of least square estimates will good.
- $SE(\hat{\beta}_0) = \sigma^2[\frac{1}{n} + \frac{\bar{x}}{\sum_{i=1}^{n}(x_i - \bar{x})}]$, $SE(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$, where $\sigma^2 = Var(\epsilon)$, estimated by $RSE = \sqrt{RSS/(n-2)}$

Assessing model

- Confidence interval for $\hat{\beta}_0$ and $\hat{\beta}_1$ at $1-\alpha\%$: $\hat{\beta}_0 \pm t_{n-2}^{1-\alpha/2} SE(\hat{\beta}_0)$, $\hat{\beta}_1 \pm t_{n-2}^{1-\alpha/2} SE(\hat{\beta}_1)$
- Hypothesis testing on the parameters:

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0.$$

$t - statistic = \frac{\hat{\beta}_1 - 0}{SE(\hat{\beta}_1)}$

Multiple Linear Regression

- Multiple linear regression model takes the form:

$$
\begin{aligned}
Y &= \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon \\
&= X\beta + \epsilon,
\end{aligned}
$$

  where X is the matrix of predictors $X_1, X_2, \cdots, X_p$, augmented by a column of 1.

$$
X = \begin{pmatrix}
1 & X_{11} & X_{12} & \cdots & X_{1p} \\
1 & X_{21} & X_{22} & \cdots & X_{2p} \\
\cdots & \cdots & \cdots & \cdots & \cdots \\
1 & X_{n1} & X_{n2} & \cdots & X_{np}
\end{pmatrix}
$$

MLR

- Least squares approach, choose β that minimize the sum of squared residuals:

$$
\begin{aligned}
RSS &= \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2. \\
&= \frac{1}{2}\sum_{i=1}^{n}(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \hat{\beta}_2 x_{i2} - \cdots - \hat{\beta}_p x_{ip})^2. \\
&= \frac{1}{2}\|Y - X\beta\|^2 = \frac{1}{2}\langle Y - X\beta, Y - X\beta\rangle \\
&= \frac{1}{2}\beta^T X^T X\beta - \langle X^T Y, \beta\rangle.
\end{aligned}
$$

- Optimality condition: $X^T X\beta - X^T Y = 0$ or $X^T X\beta = X^T Y$. (Normal equation)
- Least square solution: $\beta^* = (X^T X)^{-1} X^T Y$.

Geometric interpretation of least square solution

- $\beta^* = \min_\beta \frac{1}{2}\|Y - X\beta\|^2$.
- Think of $X\beta$ as a combination of columns of X: $\beta_1 X^1 + \beta_2 X^2 + \cdots + \beta_p X^p$.
- Looking for a "combination" of columns of X that is very close to Y.
- $X\beta$ is the orthogonal projection of y onto the column space of X.

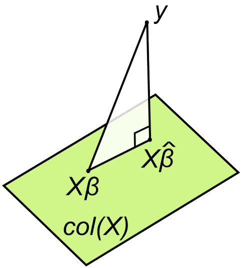

Figure: commons.wikimedia.org

Gradient descent and MLR

- Simulate a data set with given β , n: number of observations, p: number predictors.
- Gradient descent method for Least square MLR.
- Matlab demonstration.

MLR

- Is there a relathionship between response and predictors?
- Hypothesis testing:

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p$$
$$H_1 : \text{at least one } \beta_j \text{ is nonzero.}$$

F-statistics=$\frac{(TSS-RSS)/p}{RSS/(n-p-1)}$, where $TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$. and $RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$

- When there is no relationship between the response and the predictors, F-statistics will take on a value close to 1. Otherwise, F is expected to be greater than 1.

MLR

- If conclude that at least one of the parameters is nonzero, which one is it?
- When p is large, looking at individual hypothesis testing can lead to false discoveries.
- The task of determining which predictors are associated with the response (we need to fit a model with only these good predictors), is called variable selection.
- Naive approach: Try out all possible combinations of predictors. There are 2^p of them.

Choose the right model

- Forward selection: Begin with null model (no predictors). Fit p SLR models and add to the null model that variable with lowest RSS. Repeat.
- Backward selection: Begin with full model with all predictors. Remove the variable with highest p value (least significant). Fit the model with p-1 variables. Repeat.
- Mixed selection: Combine forward and backward selection. Start with null modell, add variables as with forward selection. Remove variables with high p values.

MLR - $p \gg n$

- Computation: Solving normal equation vs. gradient descent method.
- When p is big, inverting $X^T X$ is not a good idea.
- When $p \gg n$, $X^T X$ is VERY close to be singular. Finding inverse matrix is hopeless. Gradient descent method is very slow (remember the steplength?).

MLR in Matlab

- model=fitlm(X,y)
- ANOVA: anova(model)
- Confidence intervals: coefCI(model,alpha)