Linear Regression

Minimizing least squares loss:

empirical risk is $\frac{1}{m}\sum_{i=1}^{m}(\beta^T x^{(i)} - y^{(i)})^2$ where $\beta, x^{(i)} \in \mathbb{R}^n$, $y^{(i)} \in \mathbb{R}$

$$\Rightarrow \quad L(\beta) = \frac{1}{m}\sum_{i=1}^{m}(\beta^T x^{(i)} - y^{(i)})^2$$

$$= \|X\beta - y\|_2^2 \quad \text{where } X \in \mathbb{R}^{m \times n}$$

$$= (X\beta - y)^T (X\beta - y)$$

$$= (\beta^T X^T \cancel{\beta} - y^T)(X\beta - y) = \beta^T X^T X\beta - \beta^T X^T y - y^T X\beta + y^T y$$

$$= \beta^T X^T X\beta - 2y^T X\beta + y^T y$$

$$\Rightarrow \quad 0 = \nabla_\beta(\beta^T X^T X\beta - 2y^T X\beta)$$

$$= 2X^T X\beta - 2X^T y \Rightarrow \hat{\beta} = (X^T X)^{-1} X^T y$$

Statistical derivation w/ MLE:

model is : $Y^{(i)} = \theta^T X^{(i)} + \varepsilon^{(i)}$; w/ $Y^{(i)}, \varepsilon^{(i)} \in \mathbb{R}$, $\theta^T \in \mathbb{R}^n$, $X^{(i)} \in \mathbb{R}^n$

w/ $\varepsilon^{(i)} \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$

- thus, given $X^{(i)}$, $Y^{(i)} = c + \varepsilon^{(i)}$, so that given $X^{(i)}$, $Y^{(i)}$ is normal w/ mean c, var σ^2:

$$Y^{(i)} \mid X^{(i)} = x^{(i)} \sim \mathcal{N}(\theta^T x^{(i)}, \sigma^2)$$

thus the MLE:

$$L(\theta) = \prod_{i=1}^{m} p(y^{(i)} \mid x^{(i)}; \theta)$$

$$= \prod_{i=1}^{m} \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left\{ -\frac{1}{2} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma^2} \right\}$$

$$\Rightarrow \quad l(\theta) = \sum_{i=1}^{m} \left(\log \frac{1}{\sqrt{2\pi}} - \log \sigma - \frac{1}{2} \frac{(y^{(i)} - \theta^T x^{(i)})^2}{\sigma^2} \right)$$

$$\Rightarrow \quad \cancel{\frac{dl}{d\theta_j} = \sum_{i=1}^{m} \frac{1}{2\sigma^2} \cdot 2 (y^{(i)} - \theta^T x^{(i)}) \, \theta_j X_j^{(i)}} \quad \Rightarrow \quad \hat{\theta} = \underset{\theta \in \mathbb{R}^n}{\operatorname{argmin}} \sum_{i=1}^{m} (y^{(i)} - \theta^T x^{(i)})^2$$

which is the same answer as the loss solution

geometric interpretation of linear regression

let $y \in \mathbb{R}^m$, $X \in \mathbb{R}^{m \times n}$, $\beta \in \mathbb{R}^n$,

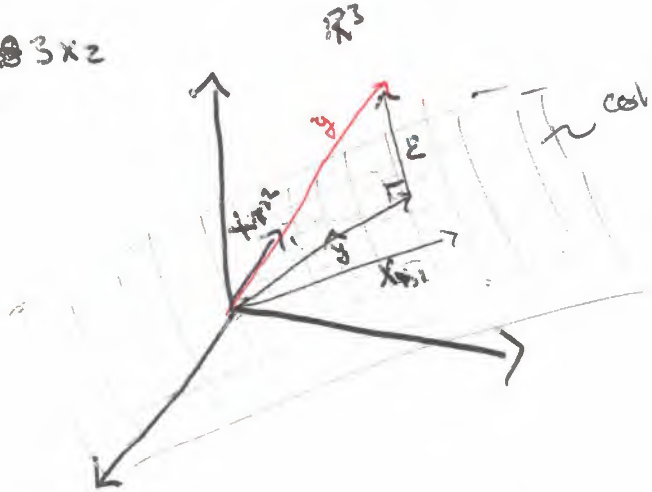
we want a model s.t.:

$y = X\beta$, i.e., the vector y is in the colum space of X

however, this is usually not the case, so we do the next best thing, which is we orthogonally project y onto that column space:

$\hat{y}$ = proj. of y onto col. space of X, $= X\hat{\beta}$:

eg: $X \in \mathbb{R}^{3 \times 2}$



thus, $y = \varepsilon + \hat{y} = X\hat{\beta} + \varepsilon$

Assumptions of linear regression

1. Linear relationship between variables and y
    1. Diagnose: look at residual plot for structure (y - hat(y) vs. hat(y))
    2. How to fix:
        1. add non-linear transformations of ~~dep.~~ indep vars. or use a kernel
        2. think about what other vars could affect the relationship and get that data
        3. custom make variables (eg. If the intercept changes at a certain point make a dummy var that is 0, 0, 0, 0, 1, 1, etc…
2. Independence of errors (specifically no correlation, p-vals, standard errors and conf. intervals depend on this assumption)
    1. Diagnose: look at residual plots for "tracking" behavior. Look at an autocorrelation vs. lag plot.
    2. How to fix:
        1. Consider adding a lag 1 of dependent var
        2. Consider adding a lag 2 of dependent var (if significant correlation at lag 2)
        3. Run an Arima model with exogenous vars (or just an AR model or MA model with exogenous vars)
        4. For seasonal correlation, consider adding dummy vars for seasons or doing seasonal differencing.
        5. For non-time series violation, look at residuals for all possible sortings of the x-axis.
3. Normality of error terms (violation causes issues for significance tests as well as computation of standard errors)
    1. Diagnose: look at Q-Q plot of residuals, a histogram or perform a KS test
    2. How to fix:
        1. Consider applying non-linear transformations to some of the independent or dependent vars.
4. Homoscedasticity (constant variance) (standard errors, conf. Intervals and p-values depend on this)
    1. Diagnose: look at residual plot to see if spread increases with x
    2. How to fix:
        1. Consider transforming the dependent var to log(Y) or sqrt(Y)
        2. Use weighted least squares if you know the variance of each response

Additional potential problems with linear regression

1. Outliers (can significantly affect quality of fit)
    1. Diagnose: look at residual plot to identify outliers

   2. How to fix:
      1. Remove them or fix any obvious issues
2. Collinearity
   1. Makes fitted coefficients highly variable (and hence SE goes up and p-values go up). This is bc Var[beta] \propto (X^T*X)^(-1) and if the columns are linearly dependent (or close to it) the variance on beta will go way up from the matrix inversion bc the matrix will be close to singular
   2. Makes interpretation (inference) of coefficients difficult
   3. Diagnose:
      1. Look at correlation matrix ; 2) do a "VIF" analysis
   4. How to fix:
      1. drop a correlated feature
      2. combine collinear vars. Into a single predictor somehow (possibly use PCA)
      3) L_1, L_2 regularization can help (L_1 tends to give all credit to just 1 of the vars, while L_2 gives credit to both)

- e.g. when y doesn't depend on x_2, but does depend on x_1, in a simple linear regression, x_2 will get "credit" for the change in y if x_1 - x_2 are correlated