Model

True model: $y = f(X) + \epsilon$

- Unknown to us
- Generates data that we observe
- Expanded form:

$$y^{(i)} = w_0 + w_1 x_1^{(i)} + w_2 x_2^{(i)} + \ldots + w_p x_p^{(i)} + \epsilon^{(i)}$$

where i denotes the observation index

p = # features

N = # observations

w = coefficients (parameters to be estimated)

Model that we wish to estimate (based on observed data):

$$\hat{y}^{(i)} = \hat{w}_0 + \hat{w}_1 x_1^{(i)} + \hat{w}_2 x_2^{(i)} + \ldots + \hat{w}_p x_p^{(i)}$$

where $\hat{w}$ represents estimates of coefficients

$\hat{y}$ represent the response that we get after estimating the model.

Note: The $\hat{}$ is dropped in further sections to be concise.

Matrix form:

$$X w = y$$

$$\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \cdots & x_p^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \cdots & x_p^{(2)} \\ \vdots & & & & \\ 1 & x_1^{(N)} & x_2^{(N)} & \cdots & x_p^{(N)} \end{bmatrix} \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_p \end{bmatrix} = \begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(N)} \end{bmatrix}$$

$(N \times p)$ $\quad$ $(p \times 1)$ $\quad$ $(N \times 1)$

OR

$$w^T X = y$$

$$\begin{bmatrix} w_0 & w_1 & \cdots & w_p \end{bmatrix} \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1^{(1)} & x_1^{(2)} & \cdots & x_1^{(N)} \\ x_2^{(1)} & x_2^{(2)} & \cdots & x_2^{(N)} \\ \vdots & \vdots & & \vdots \\ x_p^{(1)} & x_p^{(2)} & \cdots & x_p^{(N)} \end{bmatrix} = \begin{bmatrix} y^{(1)} & y^{(2)} & \cdots & y^{(N)} \end{bmatrix}$$

$(1 \times p)$ $\quad$ $(p \times N)$ $\quad$ $(1 \times N)$

Interpreting the Coefficients

Numeric predictor: One unit change in x_i corresponds to w_i units of change in y.

Categorical predictor: Selecting x_i category (i.e. $x_i = 1$) corresponds to w_i units of change in y as compared to the default category (absorbed in the intercept)

Fitting the Model

True model: $Y = Xw + \epsilon$ (unknown)

We wish to estimate $\hat{w}$ based on observed data that will give us $\hat{Y} = X\hat{w}$ that minimizes RSS.

$$RSS = \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)}\right)^2$$

(residual sum of squares)

Minimizing RSS is equivalent to minimizing mean squared error (MSE) or root mean squared error (RMSE)

i) Closed-form solution (analytical solution that minimizes RSS)

$$Xw = y$$

$$\Rightarrow X^T X w = X^T y$$

$$\Rightarrow w = (X^T X)^{-1} X^T y$$

For simple linear regression case with only one predictor, there is formula to directly compute w_0 and w_1

ii) Gradient descent (minimizes the cost function MSE ($\approx$ RSS))

$$w_{new} = w_{old} - \gamma \cdot \nabla_{\vec{w}} C(w)$$

where,

$$C(w) = \frac{1}{N} \sum_{i=1}^{N} \left(y^{(i)} - \hat{y}^{(i)}\right)^2 \quad \text{[cost function]}$$

and $$\nabla_{\vec{w}} C(w) = -\frac{2}{N} \sum x^{(i)T} \left(y^{(i)} - x^{(i)} w\right) \quad \text{[gradient of cost fn]}$$

Statistics of Linear Regression

- **Accuracy of coefficients**

i) Estimate coefficients: using closed form solution.

ii) Compute standard errors (SE): of each coefficient using a formula (see ISLR 3.1.2) that depends on N.

iii) Compute 95% CI: for each coefficient

$$w_i = [\hat{w}_i - 2 \cdot SE(\hat{w}_i), \hat{w}_i + 2 \cdot SE(\hat{w}_i)]$$

iv) Compute p-value: two hypothesis are tested

p-value of model

H_0: $w_0 = w_1 = w_2 = \cdots = w_p = 0$

H_A: At least one $w_i \neq 0$

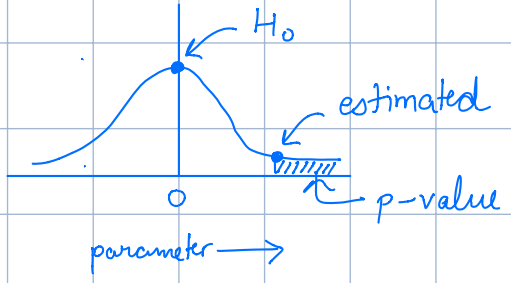
- F-statistic is computed to test if at least one predictor is associated with the response. Needs to be $>> 1$ given N is large enough

p-value of individual coefficients

H_0: $w_i = 0$

H_A: $w_i \neq 0$

- To test the above hypothesis, a model is fit with and without each predictor (both include all other predictors)



(H_0 is rejected if p-value is small because it suggests it is unlikely for the parameter to take on this value given H_0 is true.)

• Accuracy of the model

- Residual standard error (RSE) is the average deviation of RSS:

$$RSE = \sqrt{\frac{1}{N-p-1} RSS}$$ (similar to RMSE)

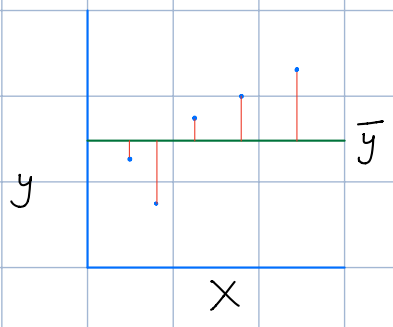
RSE estimates the average ϵ if true model were known.

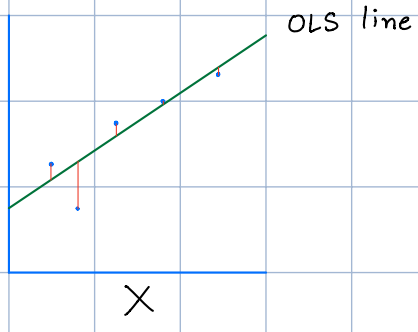
Disadvantage: RSE's scale makes it difficult to assess.

- R^2

Range of R^2 is $[0, 1]$ so it is easier to assess.

Tells the proportion of variance explained by the model.





TSS = Var(y)

$= \sum_{i=1}^{N} (y^{(i)} - \bar{y})^2$

= sum of red lines in above plot

RSS = Variance left after fitting model

$= \sum_{i=1}^{N} (y^{(i)} - \hat{y}^{(i)})^2$

= sum of red lines in above plot

$$R^2 = \frac{\overbrace{TSS - RSS}^{\text{variance explained by model (or removed)}}}{TSS} = 1 - \frac{RSS}{TSS}$$

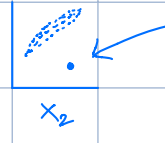
Assumptions

- X are iid
- y are iid
- $\epsilon \overset{iid}{\sim} N(0, \sigma^2)$
  - mean 0
  - constant variance (no heteroskedasticity)

Validate linear model by plotting $(y - \hat{y})$ (residual) against $\hat{y}$

and if the plot has patterns, make sure that:

- y is linearly dependent of X (if not, try $\sqrt{x_i}$, $\log x_i$ or x_i^k)
- residuals are uncorrelated
- residuals has no heteroscedasticity (if not, try $\sqrt{y}$ or $\log y$)
- y has no outliers (outliers are visible in residual plot or studentized residual plot where $|\text{studentized residuals}| > 3$)
- X has no high leverage points (outliers for a predictor) (points can look normal for a predictor but can be high leverage if we plot multiple predictors)



Compute leverage statistic to identify such points.

- No multicollinearity (compute $VIF = \frac{1}{1 - R^2_{X_j | X_{-j}}}$; $VIF > 5 \Rightarrow$ multicollinearity)
- Residuals are normally distributed

- Parametric vs Non-parametric methods

Linear regression	KNN
- Better when the relationship b/w X and Y is truly linear	- Slightly worse than linear regression when relationship b/w X and Y is truly linear.
- When relationship is not linear, model is highly biased.	- Not biased
- Better for interpretability	- Need to have > ~100 observations per predictor to be better than linear regression.