Contents

Day 27

Multiple Linear Regression

Model

$$\mu_{\mathrm{y|x}} = \beta_{\mathrm{o}} + \beta_1 x_1 + ... + \beta_{\mathrm{p}} x_{\mathrm{p}}$$

$$y = \mu_{\mathrm{y|x}} + \epsilon_1, \epsilon \sim N(0, \sigma)$$

Least Squares Line:

$$\hat{y} = b_{\mathrm{o}} + b_1 x_1 + ... + b_{\mathrm{p}} x_{\mathrm{p}}$$

$$y_{\mathrm{i}} = \hat{y}_{\mathrm{i}} + e_{\mathrm{i}}$$

Assumptions

- Linear relationship
  - Between y and each $\mathrm{x_j}$ in the model*
- Normally distributed residuals
  - Normal q-q plot
- Residuals have mean of 0 and standard deviation $\sigma = 0$
  - Independent of $\hat{y}$
- Independent residuals
  - Residual plot: e_{i} vs $\hat{y}_{\mathrm{i}}$
- All the variables in the model are independent (usually settle for uncorrelated)*

Assumptions 2-4 are for inference

$*$ check with scatter plot matrix

Simplified multiple linear regression model:

- $\mu_{\mathrm{y|x}} = \beta_{\mathrm{o}} + \beta_1 x_1 + \beta_2 x_2$
- $\hat{y} = b_{\mathrm{o}} + b_1 x_1 + b_2 x_2$

Interpreting $\mathrm{b_o}$ and β_{o}:

- Average value of y when all x-variables are 0

Interpreting $\mathrm{b_1}$ and β_1:

- Average change in y for a 1 unit increase in $\mathrm{x_1}$
  - Both are the same, pick one
  - Holding (the effect of) $\mathrm{x_2}$ constant. Think of a partial derivative (∂)
  - After accounting for the other variables in the model

ANOVA for Multiple Linear Regression

Hypothesis:

- Population mean is the same and is unaffected by anything.
- ANOVA for Linear Regression: population mean is estimated well by null model, $\mu_{\mathrm{x}_1 \dots \mathrm{x}_\mathrm{p}}$
- Equivalent: $\beta_1 = 0$ ... $\beta_\mathrm{p} = 0$

ANOVA Table

try to fill in

Interpretation:

$F_{\text{observed}} \sim F(P, n - P - 1)$

- If p-value $\leq$ significance level $\implies$ reject $\mathrm{H_o}$
  - $\therefore$ Our model is significantly better than the null model at explaining changes in y

IMPORTANT!!!!!

- This means one or more x-variables in the model are required for the better model. It DOES NOT tell us which one(s), and it CERTAINLY DOES NOT mean they are all important!

- If p-value $>$ significance level $\implies$ fail to reject $\mathrm{H_o}$

  - $\therefore$ our model is not significantly better than the null model at explaining changes in y. We prefer the null model.

Important: some x-variables may still be important predictors. However, we may not see their effect if "more important predictors" are left out of the model.

t-Test for Slope in Multiple Linear Regression

Model:

$$\mu_{\mathrm{y|x_1 \ldots x_p}} = \beta_{\mathrm{o}} + \beta_1 x_1 + \ldots + \beta_{\mathrm{p}} x_{\mathrm{p}}$$

- $\mathrm{H_o}$: In this model, $\beta_{\mathrm{j}} = 0$
  - β_{j} is the slope multiplying x_{j}, $1 \leq j \leq p$
- $\mathrm{H_a}$: In this model: $\beta_{\mathrm{o}} \neq 0$

$$t = \frac{\text{stat - parameter}}{\text{SE}} = \frac{b_{\mathrm{j}} - \beta_{\mathrm{j}}}{SE_{\mathrm{b_j}}}$$

$$t_{\mathrm{observed}} \sim t(n - p - 1)$$

Interpretation:

- If p-value $\leq$ significance level $\implies$ reject $\mathrm{H_o}$
  - $\therefore$ $\mathrm{x_j}$ is a significant predictor of y, even after accounting for the effect of the other variables in the model
- If p-value $>$ significance level $\implies$ fail to reject $\mathrm{H_o}$
  - $\therefore$ $\mathrm{x_j}$ is not a significant predictor of y in this model. We CANNOT distinguish between two competing explanations.
    1. $\mathrm{x_j}$ does not have a linear relationship with y
    2. The effect of $\mathrm{x_j}$ on y is already accounted for by other variables in the model. It is redundant.

Model Selection

General question: which model is the best?

- Step 1: Feature engineering
  - Common sense & explanatory analysis
  - Goal: identify important variables
- Step 2: Decide on a model selection algorithm and a selection criterion
  - Our algorithm: backward selection
  - Our criterion: Stop when all explanatory variables are significant (at 5 % level)
  - DO NOT use R^2 as a selection criterion in multiple linear regression
- Step 3: Implement the algorithm

Collinearity: x_1 & x_2 are highly correlated. This is bad!

- Remove least significant predictor. This means removing the variable with the highest p-value. ONLY REMOVE ONE VARIABLE AT A TIME.