Contents

Day 25

Review

Population model: $y = \beta_0 + \beta_1 x + \epsilon$

- $\epsilon \sim N(0\sigma)$
- $\mu_{\mathrm{y|x}} = \beta_0 + \beta_1 x$
- $y = \mu_{\mathrm{y|x}} + \epsilon$

Least Square Equation: $\hat{y} = b_0 + b_1 x$

Residuals: $e_{\mathrm{i}} = y_{\mathrm{i}} - \hat{y}_{\mathrm{i}}$

Assumptions for Inference for Regression

Video Lecture

1. Linear model is appropriate
2. Residuals are normally distributed
3. Residuals will have $\mu = 0$ & unknown σ [independent of x]
4. Residuals are independent

Scatterplot

Things to Note

1. Linear form
2. Hard to check
3. Strength of relationship is roughly constant across entire range of x-values

Residual Plot (e_{i} vs $\hat{y}_{\mathrm{i}}$, e_{i} vs y_{i})

1. Residuals scattered around 0 with no obvious trend
2. Hard to check
3. No fanning

Normal Quantile Plot (qq Plot)

Z-Score of residual vs Z-Score corresponding to cumulative proportion (assuming $N(0, 1)$)

- Can't check
- Points more-or-less along a straight line
  - Points fall off line near 0: really bad
    - Distribution of residuals is not symmetric
  - points follow line from -1 to 1, roughly symmetric
  - How quickly they fall away
  - How big the difference is

ANOVA for Linear Regression

One way ANOVA:

- Hypothesis: $\mu_1 = \mu_2 = ... = \mu_\mathrm{i}$
- Population mean does not depend on group
- Population mean of y does not depend on x
  - Hypothesis: $\mu_{\mathrm{y|x}} = \mu_\mathrm{y}$
  - Or equivalently, $\beta_1 = 0$

ANOVA Table

import later

Alternate Formulation 1

- Coefficient of determination

$$r^2 = \frac{SSM}{SST}$$

- Represents the proportion of variation y that is explained by/accounted for by the model.
- ANOVA tests whether this proportion is "significant"

Alternate Formulation 2

- Compare two models:
  - Null Model: $\mu_{\mathrm{y|x}} = \beta_0$
  - $\mu_{\mathrm{y|x}} = \beta_0 + \beta_1 x$

Reject H_0: Our model is "significantly better" than null model at explaining changes in y $\implies$ we should use linear model

Fail to reject H_0: Our model is not significantly better than null model $\implies$ we should use smaller model (null model)