

Exam 2019 - Problem 1

(a) The model can be expressed as

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

where x_i is the height variable, β_0 is the intercept, β_1 is the regression parameter (slope parameter) for height, ε_i is the error terms, ~~also~~ assumed to follow $N(0, \sigma^2)$.

The estimates: $\begin{cases} \hat{\beta}_0 = -2.29 & (\text{estimated outcome at } x=0) \\ \hat{\beta}_1 = 0.03 & (\text{average increase in FVC when height is increased by 1 cm}) \\ s = \hat{\sigma} = 0.20 & (\text{estimated std. dev. of the error term}) \end{cases}$

The p-value for the test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ is very small and we reject H_0 , meaning there is a significant association between height and FVC.

(b) Res-vs-Fitted: There is no clear trend in the residual distribution across the fitted values, indicating that the linearity assumption is reasonable.

Normal Q-Q: The points seem to line up well along the line, indicating that the residuals seem to follow a normal distribution.

Scale-Location: Since the standard residuals seem to result in a constant average across the fitted values, homoskedasticity (equal variance) seems to hold.

(c) $R^2 = 1 - \frac{RSS}{TSS}$ is the proportion of the variance of the outcome explained by the model.

The adjusted R^2 has a complexity penalty that penalizes when adding more covariates.

The cross-validated R^2 computes the RSS such that the residual for observation i uses the predicted $\hat{y}_i$ from a model where (x_i, y_i) has been omitted from the data when fitting the model (leave-one-out cross-validation).

While R^2 always increases when adding covariates, both adjusted and cross-validated R^2 can be used as a model selection criterion.

All criteria prefer the most complex model.

Exam 2013 - Problem 1

(a) We have the model

$$\text{bone} = \beta_0 + \beta_1 \times \text{gene1} + \epsilon,$$

where $\hat{\beta}_1 = -1.1297$ and $se(\hat{\beta}_1) = 0.3609$.

We test $H_0: \beta_1 = 0$ vs $H_1: \beta_1 \neq 0$ using the t-statistic:

$$t = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} = \frac{-1.1297}{0.3609} \approx -3.13$$

This gives a p-value below 0.01 (see t-table) and we reject H_0 , meaning that there is a significant association between bone density and gene 1.

(b) When including gene 2, the coefficient of gene 1 is no longer significant. The reason is that gene 1 and gene 2 are strongly correlated.

(c) While Model A includes all predictors, Model B omits gene 1. As a result, Model B has a higher adjusted R^2 and only significant coefficients. Hence, we choose Model B:

$$\text{bone} = -8.23 - 2.01 \times \text{gene2} + 2.55 \times \text{gene3} + 1.77 \times \text{gene4}$$

We see that gene 2 has a negative coefficient, meaning that a higher expression of gene 2 is associated with a lower bone density, all other predictors being held fixed.

The opposite holds for genes 3-4, which both have positive coefficients.