Assessing Goodness of Fit

Math 430, Winter 2017

Partitioning variability

$$\text{Total sum of squares (SST)} = \sum_{i=1}^{n} (y_i - \overline{y})^2$$

$$\text{Sum of squares error (RSS)} = \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2$$

$$\text{Sum of squares due to model (SSreg)} = \sum_{i=1}^{n} (\widehat{y}_i - \overline{y})^2$$

ANOVA identity: $\text{SST} = \text{SSModel} + \text{SSE}$

Coefficient of Determination: R^2

Definition: $R^2 \in [0,1] = \frac{\text{SSreg}}{\text{SST}} = 1 - \frac{\text{RSS}}{\text{SST}}$

Interpretation: Proportion of the variability in y explained by the linear model.

Intuition: A better model explains more of the variability in y

Pitfall: R^2 does not talk about predictive ability of the model

Coefficient of Determination: R^2

```
summary(mod)
```

```
##
## Call:
## lm(formula = log(Time) ~ I(Tonnage^0.25), data = glakes)
##
## Residuals:
##     Min       1Q   Median       3Q      Max
## -0.6607 -0.2410 -0.0044  0.2203  0.4956
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)       1.18842    0.19468   6.105  1.2e-06 ***
## I(Tonnage^0.25)   0.30910    0.02728  11.332  3.6e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1
##
## Residual standard error: 0.3034 on 29 degrees of freedom
```

Coefficient of Determination: R^2

```
library(broom)
glance(mod)
```

```
##   r.squared adj.r.squared     sigma statistic      p.val
## 1 0.8157603     0.8094072 0.3034015  128.4036 3.598826e-
##         AIC      BIC deviance df.residual
## 1 17.95947 22.26143 2.669522          29
```

Predictive accuracy via cross validation

1. Randomly split data set into two: a **training set** and a **holdout (test) set**
2. Fit model to the training set
3. Use the fitted model to predict the holdout set
4. Compute cross validation metrics

$$\text{Predictive Bias} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \widehat{y}_i)$$

$$\text{Predictive Mean Square Error} = \frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \widehat{y}_i)^2$$

$$\text{RPMSE} = \sqrt{\text{PMSE}}$$

Predictive accuracy via cross validation

Interpretations

- **Bias**: systematic errors in our predictions
- **Root predicted mean square error**: How far off the predictions are, on average

Predictive Accuracy of Cargo Model

```r
# Split data into training and test sets
index <-  sample(1:nrow(glakes), size = 0.2 * nrow(glakes))
train_data <- glakes[-index,]
test_data <- glakes[index,]

# Fit model to training data
train_lm <- lm(log(Time) ~ I(Tonnage^.25), data = glakes)

summary(train_lm)

# Obtain predictions
preds <- predict(train_lm, newdata = test_data)
preds_orig <- exp(preds)

# Calculate metrics
bias <- mean(test_data$Time - preds_orig)
pmse <- mean((test_data$Time - preds_orig)^2)
rpmse <- sqrt(pmse)
```