APPM2720 Assessing accuracy of the least squares model

It is usually straightforward to judge the strength of a linear relationship between two variables by examining a scatterplot. To set the stage for more complicated data situations there is a useful statistic to measure this related to how large the residuals are compared to the original observations.

the standard deviation

By way of background the definition of the standard deiviation (the R function **sd**) is

$sd = \sqrt{\frac{1}{N-1}\sum_i (X_i - \bar{X})^2}]$

N the total number of observations, and $\bar{X}$ the average.

This is a useful statistic to describe the spread in data especically when it is close to a Gaussian (aka Normal) distribution.

The "minus one" in the denominator is an adjustment because we are subtracting the average from all the data values.

Residuals

Given N pairs of numbers $(x_1, y_1), (x_2, y_2), \ldots, x_N, y_N)$. A useful model is to predict the Ys by a straight line in X. Recall the residuals are the difference between the predicted values and the observations.

$e_i = Y_i - (\beta_0 + \beta_1 X_i)$

Residual standard error

Residual Sums of Squares = $RSS = \sum_i e_i^2$

Residual standard error = $RSE = \sqrt{\frac{1}{N-2}RSS}$

The "minus 2" comes from subtracting the line (slope and intercept). In general for least squares fitting the 2 is replaced by the total number of least squares parameters being fit.

- The residual standard error summarizes the spread in the residuals and is useful if they are approximately normal (e.g. follows rules for fraction being within certain number of standard deviations of zero.)

- RSE is needed to find confidence intervals for the estimated slope and intercept
- The Residual standard error is also reported by the **summary** function applied to the **lm** output. The degrees of freedom reported after this statistic is the number of observations minus the number of parameters being fit by least squares. (2 in the case of a single variable with intercept.)

Accuracy of the linear fit

Total sum of squares = $TSS = \sum_i Y_i - \bar{Y}^2$

This is the RSS if we just predicted Y with a constant value!

(R squared) = $R^2 = 1 - RSS/TSS$

R squared close to 1 means the line predicts the data well. close 0 to zero means the line does not explain much of the variability in Y.

Relationship to R squared

I like to compare the standard deviation of the Y values to the standard deviation of the residuals

```
fit<- lm( Y~X)
sd( fit$residuals ) / sd( Y)
```

Want this to be small -- if zero means a perfect fit.

Or in the form of R^2

```
1- (sd( fit$residuals ) / sd( Y))^2
```

Not exactly equal to R squared because it missed the adjustment for the number of parameters -- but easier to think about and commuicate.