CHE384, From Data to Decisions: Measurement, Uncertainty, Analysis, and Modeling

Lecture 35
The Wrong Model, part 2

Chris A. Mack
Adjunct Associate Professor

http://www.lithoguru.com/scientist/statistics/

Coefficient of Determination, R^2

- The Coefficient of Determination (R^2) is a measure of how much of the variation in Y is explained by the model

$$\text{Regression Sum of Squares: } SSR = \sum(\hat{y}_i - \bar{y})^2$$

$$\text{Error Sum of Squares: } SSE = \sum(y_i - \hat{y}_i)^2$$

$$\text{Total Sum of Squares: } SSTO = \sum(y_i - \bar{y})^2$$

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO} \qquad SSTO = SSR + SSE$$

(only true for linear regression)

Goodness of Fit

- Goodness of fit metric: $R^2 = r^2$

$$r = \frac{cov(X,Y)}{\sqrt{var(X)var(Y)}} \qquad R^2 = \frac{cov^2(X,Y)}{var(X)var(Y)}$$

- Also, we can show that

$$R^2 = 1 - \frac{var(\varepsilon)}{var(Y)}$$

R^2 is the fraction of the variance of Y that is explained by the linear fit

Overall F-Test for Regression

- An overall test for model significance:
  - H_0: $\beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$ (not testing intercept)
  - H_A: $\beta_j \neq 0$, for at least one value of j
  - p = number of parameters in the model
- If H_0 is true, then the model is not useful for explaining the variation in y
  - SSR is much smaller than SSE
  - SSE is about the same as SSTO

Overall F-Test for Regression

- Is SSTO about the same as SSE?
- Compare SSTO – SSE to SSE, considering the degrees of freedom

$$SSTO = SSR + SSE$$ (only true for linear regression)

$$F = \frac{(SSTO - SSE\ /(DF_{SSTO} - D\ _{SSE})}{SSE/DF_{SSE}} = \frac{SSR/(p-1)}{SSE/(n-p)}$$

Build an ANOVA Table

Source	SS	df	MS	F	p-value
Regression	SSR	p-1	SSR/(p-1)	MSR/MSE	from F-distribution
Error	SSE	n-p	SSE/(n-p)		
Total	SSTO	n-1			

$$F = \frac{SSR/(p-1)}{SSE/(n-p)} = \frac{MSR}{MSE}$$

For Weighted Regression

- We must include the weights in our sum of squares calculations

$$\text{Regression Sum of Squares}: \quad SSR = \sum w_i (\hat{y}_i - \bar{y}_w)^2$$

$$\text{Error Sum of Squares}: \quad SSE = \sum w_i (y_i - \hat{y}_i)^2$$

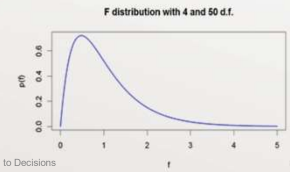
$$\text{Total Sum of Squares}: \quad SSTO = \sum w_i (y_i - \bar{y}_w)^2$$

Overall F-Test for Regression

- Note that the coefficient of determination (R^2) is related to this F-statistic:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$F = \frac{R^2/(p-1)}{(1-R^2)/(n-p)}$$



Overall F-Test for Regression

- The F-statistic is F-distributed with (p-1,n-p) degrees of freedom ($F_{p-1,n-p}$)
  - An F-distribution is the ratio of two independent χ^2 distributions
  - Assumes normal distribution of iid residuals with constant σ_ε
  - F = explained variance/unexplained variance
- Calculate the p-value from the F-distribution and compare it with your significance level α
  - P-value = probability that the model explains the variance in y no better than by chance

Overall F-Test for Regression

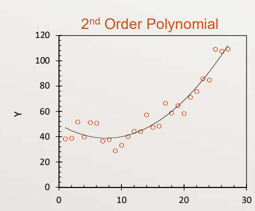
- For a two-parameter model, the F-test is the same as asking if the confidence interval of the slope includes zero (t-test with DF = n-2)
  - $F = [b_1/SE(b_1)]^2$ for this case
- Don't over-interpret the test results
  - A small p-value doesn't necessarily imply a good fit of model to data, only that at least one model parameter is non-zero
  - A large p-value doesn't necessarily mean that the response variable is not dependent on the predictor variables, only that *this* model is not significant

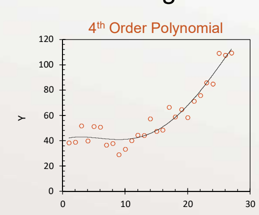
Training vs. Predicting

- We build a model by fitting it to data
  - We "train" or calibrate a model using a given data set
  - The residual standard deviation is a measure of how well the model fits this data set
  - If we add more fitting parameters to the model, we always get a better fit
- The real test, however, is how well we match a *new* data set, one not used in the training
  - Called validation of the model
  - The residual standard deviation for validation data will almost always be higher than for the training data

Overfitting

- Adding new model terms always makes the fit better, but can result in fitting noise





Model Scope

- Every model is built and validated over a range of input conditions, called the scope of the model
  - When a model is developed, its scope should be clearly specified
  - Prediction and interpretation (the two goals of modeling) should generally be limited to within the scope
  - Extrapolations are sometimes done, but know that uncertainty estimates are no longer valid

Lecture 35: What have we learned?

- What is the coefficient of determination and how is it calculated?
- What is the overall regression F-test and how is it used?
- What are the assumptions inherent in the F-test?
- What is model validation?
- Define model scope