# STAT6038 Week 7 Lecture Notes

Rui Qiu

2017-04-19

## 1 Lecture 19 2017-04-19

$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i,\ i = 1, 2, \cdots N$ (population model)
$Y_i = b_0 + b_1 x_{1i} + b_2 b_{2i} + \cdots + b_k x_{ki} + e_i,\ i = 1, 2, \cdots n$ (sample model)

population model, sample model, matrix notation, dimension, design matrix, partial regression coefficient and its interpretation, "linear in the parameters"

### 1.1 Polynomial Regression & Degrees of Freedom

Insert 3 plots here.

Let $n$ be number of points in the plot; $k$ be number of variables here, e.g. $k = 2$, there are variables $x, x^2$; $p$ be number of parameters here, e.g. $p = 3$, there are parameters $\beta_0, \beta_1, \beta_2$.
If $n = p = k + 1$, we have a perfect fit.

A polynomial of degree $(n - 1)$ (as **total degrees of freedom**) will be a perfect fit for $n$ observations, but there are no degrees of freedom left over to estimate $\sigma^2$, the variance of the errors $\epsilon$.

*Note:* But this is an extreme of "over-fitting", a too complicated model. (more on this later)

### 1.2 Model Linearization

### 1.3 Underlying Model Assumptions

- assume uncorrelated (independent) and homoscedastic (constant variance) errors
- errors are normally distributed

### 1.4 Estimations of Parameters

$$\hat{\beta} = b = (X^T X)^{-1}(X^T Y)$$
$$\hat{\sigma}^2 = s^2 = \frac{SS_{Errors}}{(n-p)} = \frac{e^T e}{(n-p)}$$

Note that the error degrees of freedom is $(n-p)$. We lose one each for each parameter.

## 2 Lecture 20 2017-04-20

### 2.1 Squid data example

### 2.2 Some questions

*Why do Sums of Squares in ANOVA depend on the order you fit the model but the fitted model is the same?*

**Because they are Sequential Sums of Squares**: The model or regression sum of squares can be partitioned as follows:

$$\begin{aligned}\text{SSR} &= \text{SSR}(\beta_1, \beta_2, \ldots, \beta_k|\beta_0) \\ &= \text{SSR}(\beta_1|\beta_0) + \text{SSR}(\beta_2|\beta_1, \beta_0) + \text{SSR}(\beta_3|\beta_2, \beta_1, \beta_0) \\ &+ \cdots + \text{SSR}(\beta_k|\beta_0, \beta_1, \beta_2 + \cdots + \beta_{k-2} + \beta_{k-1})\end{aligned}$$

For example, the sequential $\text{SSR}(\beta_2|\beta_1, \beta_0)$ is the amount of unexplained variability from a simple linear regression on $x_1$ which is subsequently explained by $x_2$, so it represents the increase in the regression sum of squares obtained by adding the predictor $x_2$ to a model that already contains $x_1$.

### 2.3 Partial (Sequential) F tests for nested models

Model 1: $Y = \beta_0 + \beta_1 X_1 + \epsilon$
Model 2: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$
Model 1 is "nested" inside Model 2. For here, "nested" means "contained in or is a subset of".

Similarly the null model (mean model) $Y = \beta_0 + \epsilon$ is nested inside the SLR model $Y = \beta_0 + \beta_1 X + \epsilon$.

So, the overall F test for a SLR model was a special case of a partial or sequential F test.

A partial F is a test for part of a model → the **addition** of some extra terms.

So a partial F test for the addition of $\beta_2 X_2$ to a model that already contains $\beta_0 + \beta_1 X_1$.

$$F = \frac{\sigma^2_{\text{addition}}}{\sigma^2_{\text{error}}}$$

$\sigma^2_{\text{error}}$ is for the larger model, i.e. Model 2.

# 3 Lecture 21 2017-04-21

Recall the squid data we did yesterday, the two models contain the same variables, and same (reordered) parameters. But the results in two ANOVA tables are changed dramatically.

The tests conducted here are actually sequential statistics.

## 3.1 Partial (Sequential) F tests

$Y = \beta_0 + \beta_1 X_1 + \epsilon$
vs.
$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$

A sequential F test is a partial F test for the addition of a single term to an existing model. The "addition" in this case is $\beta_2 X_2$.

**F test:**

1. $H_0 : \frac{\sigma^2_{\text{addition}}}{\sigma^2_{\text{error, larger model 2}}} = 1$ vs $H_A : \frac{\sigma^2_a}{\sigma^2_\epsilon} > 1$
   (in variance terms)
   equiv. $H_0 : \beta_2 = 0$ vs $H_A : \beta_2 \neq 0$
   (in mean terms)
2. Test statistic $F = \frac{MS_{\text{addition}}}{MS_{\text{error/residual}}} \sim F_{\text{addition df, error df}}$
   error from model 2
3. Decision rule $\alpha = 0.05$. Reject $H_0$ if observed $F > F_{\text{addition df, error df}}$
4. Graphical analysis.
5. Compare p-value with $\alpha$ and draw a conclusion

**Partial (nested) F test for a group of terms in a nested model**

Model A: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$ "base" model

Model B: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \epsilon$ "expanded" model = "base" model + additions

"base" model A is nested inside the "expanded" model B, or model A is a subset of model B.

1. $H_0 : \frac{\sigma^2_{\text{addition}}}{\sigma^2_{\text{error, larger model B}}} = 1$ vs. $H_A : \frac{\sigma^2_{\text{addition}}}{\sigma^2_{\text{error}}} > 1$
   equivalently, $H_0 : \beta_4 = \beta_5 = 0$ vs. $H_A :$ Not both all of $\beta_4, \beta_5 = 0$

2. Test statistics $F = \frac{MS_{\text{addition}}}{MSE_{\text{model B}}} \sim F_{2,16}$

3. $\alpha = 0.05$ reject $H_0$ if $p < \alpha$

4. Observed $F_{2,16} = 4.6638$
   and graph

5. As $p = 0.02536 < \alpha = 0.05$, reject $H_0$ in favor of $H_A$ and conclude that at least one of the two additional terms is a significant addition to the base model.

The partial (nested) F test is the most general of these tests for nested models, but the differences in names here is just jargon!

A sequential F test is just the special case of a nested F test where we are adding **a single** additional (they are both still partial F tests).

The overall F test for a multiple regression model is also just a special case of the nested F test: here the additional terms are **all** of the $x$ variables on top of a null model.

Model A: $Y = \beta_0 + \epsilon$

Model B: $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$

In all of these (partial) F tests for nested models, we will tend to prefer the simpler 'base" model over the more complicated "expanded" model whenever we fail to reject $H_0$. (with some caveats $\implies$ the null model is not always the best "base" model $\implies$ it will depend on the research question)

$\rightarrow$ But these F tests and the ANOVA table are key in deciding what belongs in the model, we will use these as an approach for refining models.

### 3.2 What has changed from SLR to MR?

- plot()
  - is model appropriate?
  - are the underlying assumptions ok?
  - these are pretty much the same as earlier
- anova()
  - is model adequate?
  - doesn't it (and all parts of it) have significant explanatory power?
  - this has definitely changed
- summary()
  - used to try the research question, once we have the right model
  - again, much the same as earlier
- predict()
  - we will also see that this hasn't changed much