Midterm 2: Example

# Overview

### Model

Suppose we have some multiple regression model

$$y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + e_i$$

with $e_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$. In Notes 5, $p$ represented the number of predictors in the model (because we had not yet introduced second order terms).

Note that the polynomial and interaction regression models (from Notes 7) are special cases of the multiple regression model. For example, a quadratic regression model

$$y_i = b_0 + b_1 x_i + b_2 x_i^2 + e_i$$

could be written using the above equation with $p = 2$ and $x_{i2} = x_{i1}^2$ by definition. As another example, the interaction regression model

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i1} x_{i2} + e_i$$

could be written using the above equation with $p = 3$ and $x_{i3} = x_{i1} x_{i2}$ by definition.

### ANOVA Table

| Source | SS | df | MS | F | p-value |
|---|---|---|---|---|---|
| SSR | $\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ | p | $MSR$ | $F^*$ | $p^*$ |
| SSE | $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ | $n-p-1$ | $MSE$ | | |
| SST | $\sum_{i=1}^{n}(y_i - \bar{y})^2$ | $n-1$ | | | |

$MSR = \frac{SSR}{p}$, $MSE = \frac{SSE}{n-p-1}$, $F^* = \frac{MSR}{MSE} \sim F_{p,n-p-1}$, $p^* = P(F_{p,n-p-1} > F^*)$

$F^*$-statistic and $p^*$-value are testing $H_0 : b_1 = \cdots = b_p = 0$ versus $H_1 : b_k \neq 0$ for at least one $k \in \{1, \ldots, p\}$

Note that the degrees of freedom for SSE is $n - c$ where $c = p + 1$ is the total number of columns of the model design matrix $\mathbf{X}$.

**Testing Multiple Slopes**

Assume that $q < p$ and want to test if a reduced model is sufficient:

$$H_0 : b_{q+1} = b_{q+2} = \cdots = b_p = b^*$$
$$H_1 : \text{at least one } b_k \neq b^*$$

Compare the SSE for full and reduced (constrained) models:
(a) Full Model: $y_i = b_0 + \sum_{j=1}^{p} b_j x_{ij} + e_i$
(b) Reduced Model: $y_i = b_0 + \sum_{j=1}^{q} b_j x_{ij} + b^* \sum_{k=q+1}^{p} x_{ik} + e_i$

Note: set $b^* = 0$ to remove $X_{q+1}, \ldots, X_p$ from model.

Test Statistic:

$$\begin{aligned} F^* &= \frac{SSE_R - SSE_F}{df_R - df_F} \div \frac{SSE_F}{df_F} \\ &= \frac{SSE_R - SSE_F}{(n-q-1)-(n-p-1)} \div \frac{SSE_F}{n-p-1} \\ &\sim F_{(p-q,n-p-1)} \end{aligned}$$

where

- $SSE_R$ is sum-of-squares error for reduced model
- $SSE_F$ is sum-of-squares error for full model
- $df_R$ is error degrees of freedom for reduced model
- $df_F$ is error degrees of freedom for full model

Note that the numerator degrees of freedom $p-q$ is the number of terms that are constrained in the reduced model specified by $H_0$.

Also, note that this approach can be used to test a single slope; in this case, $p - q = 1$ and $F^* = (t^*)^2$ where $t^*$ is the single slope $T$ test statistic (see Notes 5 Slide 40).

# Example

Consider the regression model

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i3} + b_4 x_{i1}^2 + b_5 x_{i2}^2 + b_6 x_{i1} x_{i2} + e_i$$

with $e_i \overset{\text{iid}}{\sim} \text{N}(0, \sigma^2)$. Suppose that we have observed the following sample statistics from $n = 20$ observations:

```
> length(y)
[1] 20
> sum(lm(y~1)$residuals^2)
[1] 1622.979
> sum(lm(y~x2+x3+I(x2^2))$residuals^2)
[1] 741.5935
> sum(lm(y~x1+x2+x3+x1:x2)$residuals^2)
[1] 12.62165
> sum(lm(y~x1+x2+I(x1^2)+I(x2^2)+x1:x2)$residuals^2)
[1] 320.1572
> sum(lm(y~x1+x2+x3+I(x1^2)+I(x2^2))$residuals^2)
[1] 15.75573
> sum(lm(y~x1+x2+x3+I(x1^2)+I(x2^2)+x1:x2)$residuals^2)
[1] 10.91046
```

(a) Create ANOVA table for model:

| Source | SS | df | MS | F | p-value |
|---|---|---|---|---|---|
| SSR | 1612.069 | 6 | 268.6781 | 320.1345 | <0.0001 |
| SSE | 10.910 | 13 | 0.8393 | | |
| SST | 1622.979 | 19 | | | |

(b) Perform the overall test of significance of the model at $\alpha = 0.05$, i.e., test $H_0 : b_1 = \cdots = b_6 = 0$ versus $H_1 :$ at least one $b_k \neq 0$ for $k \in \{1, \ldots, 6\}$.

Use $F$-test from ANOVA table. The observed test statistic is $F^* = 320.1345 \sim F_{6,13}$ and the p-value is nearly zero. Therefore, we reject $H_0$ and conclude that there is a significant relationship between $Y$ and at least one of the model terms.

(c) Test if $X_1$ significantly adds to the regression model at $\alpha = 0.05$, i.e., test $H_0 : b_1 = b_4 = b_6 = 0$ versus $H_1 :$ at least one $b_k \neq 0$ for $k \in \{1, 4, 6\}$.

Use $F$-test for multiple slopes. The reduced model is the model containing only $X_2$ and $X_3$, and the full model contains all of the terms. So, the $F$ test statistic is

$$F^* = \frac{(741.5935 - 10.91046)/(16 - 13)}{10.91046/13} = \frac{243.561}{0.8392662} = 290.2071$$

which follows an $F_{3,13}$ distribution. Note that $P(F_{3,13} > F^*) < .0001$, so we Reject $H_0$ and conclude that $X_1$ significantly adds to the regression model.

(d) Test if $X_3$ significantly adds to the model at $\alpha = 0.05$ ($H_0 : b_3 = 0$ versus $H_1 : b_3 \neq 0$).

Use $F$-test for single slope. The reduced model is the model without $X_3$, and the full model contains all of the terms. So, the $F$ test statistic is

$$F^* = \frac{(320.1572 - 10.91046)/(14 - 13)}{10.91046/13} = \frac{309.2467}{0.8392662} = 368.4728$$

which follows an $F_{1,13}$ distribution. Note that $P(F_{1,13} > F^*) < .0001$, so we Reject $H_0$ and conclude that $X_3$ significantly adds to the regression model.

(e) Test the significance of the quadratic terms at $\alpha = 0.05$, i.e., test $H_0 : b_4 = b_5 = 0$ versus $H_1 :$ at least one $b_k \neq 0$ for $k \in \{4, 5\}$.

Use $F$-test for multiple slopes. The reduced model is the model containing only linear and interaction terms, and the full model contains all of the terms. So, the $F$ test statistic is

$$F^* = \frac{(12.62165 - 10.91046)/(15 - 13)}{10.91046/13} = \frac{0.855595}{0.8392662} = 1.019456$$

which follows an $F_{2,13}$ distribution. Note that $P(F_{2,13} > F^*) = 0.3879047$, so we Retain $H_0$ and conclude that the polynomial terms do NOT significantly add to the model.

(f) Test the significance of interaction term at $\alpha = 0.05$ ($H_0 : b_6 = 0$ versus $H_1 : b_6 \neq 0$).

Use $F$-test for single slope. The reduced model is the model without $X_1X_2$, and the full model contains all of the terms. So, the $F$ test statistic is

$$F^* = \frac{(15.75573 - 10.91046)/(14 - 13)}{10.91046/13} = \frac{4.84527}{0.8392662} = 5.773222$$

which follows an $F_{1,13}$ distribution. Note that $P(F_{1,13} > F^*) = 0.03192008$, so we Reject $H_0$ and conclude that $X_1X_2$ significantly adds to the regression model.