Lecture 7: Multiple Regression: General Consideration

Reading Assignment:

- Muller and Fetterman, Chapter 4: “Multiple Regression”

Why use more than one covariate in a model?

- Why not fit separate models for every covariate?
- Omitting an important covariate, $\mathbf{x}_2$, can cause you to miss significant relationships between $\mathbf{x}_1$ and $\mathbf{y}$ or even to make completely wrong conclusions!

Example: Math Ability

Suppose that you hypothesize that taller children are better at math than shorter children. You take a random sample of 32 children of various ages in Ephesus Elementary. These children take a math test and have their heights measured. When you fit a linear model using height to predict math test score, you find that height is highly significant.

Does this make sense? Recall the definition of a confounder: a *confounder* is a factor that is associated with the exposure and independently affects the outcome.

In this case, age is a potential confounder because it is associated with height and, independently of height, is related to math test scores.

When we fit the *multiple regression* model (sometimes called a *multivariable* regression model) with both age and height as predictors, we see that age is an important predictor of mathematical ability. In addition, after accounting for age, height is unimportant. Our conclusion is that after accounting for age, height is not related to mathematical ability in our sample of elementary school children.

For a general multiple regression model, we write

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} x_{10} & \cdots & x_{1,p-1} \\ \vdots & & \\ x_{n0} & \cdots & x_{n,p-1} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

The corresponding scalar version of the model is

$$y_i = \sum_{j=0}^{p-1} x_{ij}\beta_j + \varepsilon_i$$

Assume we fit a model and test hypotheses with a continuous, interval scale response and with linear combinations of one or more continuous variables as predictors. All GLH tests can be understood in terms of comparisons of two models: a full model and a reduced model. These tests may be conducted by comparing the *sums of squares* from the two models.

Definitions of Basic Sums of Squares

For the model $\mathbf{y}_{n\times 1} = \mathbf{X}_{n\times p}\,\boldsymbol{\beta}_{p\times 1} + \boldsymbol{\varepsilon}_{n\times 1}$, we have

- $\widehat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$,
- $\widehat{\mathbf{y}} = \mathbf{X}\widehat{\boldsymbol{\beta}}$, and
- $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y} - \widehat{\mathbf{y}}$.

We are already familiar with the sum of squares for error, given by

$$\begin{aligned} SSE &= \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = (\mathbf{y} - \widehat{\mathbf{y}})'(\mathbf{y} - \widehat{\mathbf{y}}) \\ &= \sum_{i=1}^{n}(y_i - \widehat{y}_i)^2 = \sum_{i=1}^{n}\widehat{\varepsilon}_i^2 . \end{aligned}$$

The *uncorrected total sum of squares*, USS, is given by

$$USS(\text{total}) = \boldsymbol{y}'\boldsymbol{y} = \sum_{i=1}^{n} y_i^2.$$

The *uncorrected model sum of squares* (or *uncorrected regression sum of squares*) is given by

$$USS(\text{model}) = \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} = \boldsymbol{y}'\boldsymbol{H}\boldsymbol{y} = \sum_{i=1}^{n} \widehat{y}_i^2.$$

The difference between the uncorrected total sum of squares and the uncorrected model sum of squares is the sum of squares for error:

$$USS(\text{total}) - USS(\text{model}) = \mathbf{y}'\mathbf{I}\mathbf{y} - \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y} = SSE.$$

So we see that

$$\begin{aligned} USS(\text{total}) &= USS(\text{model}) + SSE \\ \mathbf{y}'\mathbf{y} &= \mathbf{y}'\mathbf{H}\mathbf{y} + \mathbf{y}'(\mathbf{I}-\mathbf{H})\mathbf{y}. \end{aligned}$$

Example: Calculating Uncorrected Sums of Squares

Uncorrected sums of squares may be calculated using the following R code.

```
> ozone = read.table("ozone.txt", header = T, sep = " ") # space delimited
> n = nrow(ozone) # number of rows of ozone (obs)
> X = model.matrix(~ outdoor + home + time_out, data = ozone) # predictor matrix
> y = ozone$personal # response
> head(X)
```

```
  (Intercept)  outdoor  home time_out
1           1 35.87771 22.29     0.57
2           1 43.79189 13.97     0.90
3           1 49.81255 18.96     0.55
4           1 34.37366 22.27     0.17
5           1 45.95496 23.40     0.00
6           1 64.76558 39.62     0.30
```

```
> bhat = solve(t(X) %*% X) %*% t(X) %*% y # from notes
> yhat=X%*%bhat; # predicted values
> ehat=y-yhat; # residuals
> p=ncol(X); # num columns of X
> df=n-p; # df
> sse = t(y) %*% y - t(bhat) %*% t(X) %*% y # SSE
> mse=sse/df; # MSE
> H=X%*%solve(t(X) %*% X) %*% t(X) # calculate Hat Matrix
> uss_t=t(y)%*%y
> uss_m=t(y)%*%H%*%y
> print(uss_t)

         [,1]
[1,] 50664.94

> print(uss_m);

         [,1]
[1,] 40516.75

> print(uss_m + sse);

         [,1]
[1,] 50664.94
```

We may also get the same results in SAS using PROC GLM, a more general form of PROC REG, using the int option to calculate the uncorrected sums of squares, printing them in an ANOVA table.

```
proc glm data=ozone;
model personal=outdoor home time_out/int;
run;
```

```
                              The GLM Procedure

        Dependent Variable: personal   Personal Ozone Exposure (ppb)


                                          Sum of
 Source                      DF          Squares    Mean Square   F Value   Pr > F

 Model                        4      40516.75081    10129.18770     59.89   <.0001

 Error                       60      10148.19129      169.13652

 Uncorrected Total           64      50664.94210
```

Corrected sums of squares are defined in relationship to the intercept.

The Nature of the Intercept

It is often (but not always) appropriate to include a constant predictor equal to 1.0 so that $x_0 = \mathbf{J}$. Then

$$y_i = \beta_0 + x_{i1}\beta_1 + \ldots + x_{i,p-1}\beta_{p-1} + \varepsilon_i.$$

In such models, we call β_0 the intercept. The name reflects the fact that if $x_{i0} = 1$ and all other predictors take the value zero, then

$$E(y_i) \quad = \quad E(\beta_0) + E\left(\sum_{j=1}^{p-1} x_{ij}\beta_j\right) + E(\varepsilon_i) = \beta_0,$$

and β_0 is the y intercept of the fitted line (the value of y at which the regression function intercepts the vertical axis). In addition, β_0 equals the response (y) value predicted if all x_{ij}, $j \in \{1, 2, \ldots p-1\}$, are 0.

For $\bar{x}_j$ the average of the j^{th} column of $\mathbf{X}$, $\beta_0 = \mu_y - \sum_{j=1}^{p-1} \overline{x}_j\beta_j$ and $\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p-1} \overline{x}_j\widehat{\beta}_j$.

Example: Ozone Data

For the ozone data, we have the model

$$E(personal_i) = \beta_0 + \beta_1 outdoor_i + \beta_2 home_i + \beta_3 timeout_i,$$

$i = 1, \ldots, 64$. The parameter estimates from this model are given by $\widehat{\boldsymbol{\beta}} = (3.78, 0.09, 0.60, 13.64)'$. In addition, in our dataset, we observe outdoor ozone concentrations ranging from 11.60-104.10 ppb, home ozone concentrations ranging from 0.80-46.04 ppb, and percent time spent outdoors ranging from 0-90%. In this model, are we particularly interested in the value of the intercept?

Models That Span but May Not Include an Intercept

Any model that includes an intercept has $\mathbf{J}_n$ as a column in $\boldsymbol{X}$. Models with an intercept allow computing the corrected sums of squares, which exclude the portion of the sums of squares due to the intercept and are usually preferred over the uncorrected sums of squares.

When using dummy variables, it is often convenient to code the model such that no column of $\boldsymbol{X}$ is $\mathbf{1}$ although $\mathbf{X}_{n\times p}\ \mathbf{t}_{p\times 1} = \mathbf{1}$, for $\mathbf{t}_{p\times 1}$ a vector of constants.

Such a model *spans* an intercept, even though the design matrix, $\boldsymbol{X}$, does not include an intercept.

For example, let

$$\boldsymbol{X}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix}, \; \boldsymbol{\beta}_1 = \begin{bmatrix} \mu_0 \\ \mu_1 \end{bmatrix}, \; \boldsymbol{X}_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 1 \\ 1 & 1 \\ 1 & 1 \end{bmatrix}, \; \boldsymbol{\beta}_2 = \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix}.$$

Notice that $\mathbf{X}_2 = \mathbf{X}_1 \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}$ and

$$\boldsymbol{\beta}_1 = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \end{bmatrix} = \begin{bmatrix} \alpha_0 \\ \alpha_0 + \alpha_1 \end{bmatrix}$$

so that the vector $\mathbf{t} = (1, 1)'$ finds the hidden intercept.

Corrected Sums of Squares

Corrected sums of squares are adjusted (corrected) for the intercept and measure the effect above and beyond it. In addition, the corrected sums of squares are invariant to location shifts of the response or predictors (e.g.,centering a covariate).

Uncorrected sums of squares are always well-defined. Corrected sums of squares and associated statistics entirely exclude the intercept but involve comparing models which all span an intercept. Thus corrected sums of squares are defined only if the model includes or spans an intercept.

The correction term or *sum of squares due to the intercept*, $n\bar{y}^2 = \frac{\boldsymbol{y}'\mathbf{J}_n\mathbf{J}'_n\boldsymbol{y}}{n} = SSI$, corrects for location.

The corrected total sum of squares is computed as

$$\begin{aligned} CSS(\textsf{total}) &= USS(\textsf{total}) - SSI \\ &= \mathbf{y}'\mathbf{y} - \frac{\mathbf{y}'\mathbf{J}_n\mathbf{J}_n'\mathbf{y}}{n} \\ &= \mathbf{y}'\left[\mathbf{I}_n - \frac{\mathbf{J}_n\mathbf{J}_n'}{n}\right]\mathbf{y} \\ &= \sum_{i=1}^{n}(y_i - \overline{y})^2. \end{aligned}$$

The corrected model sum of squares is also computed by subtracting the SSI:

$$\begin{aligned} CSS(\text{model}) &= USS(\text{model}) - SSI \\ &= \mathbf{y}'\mathbf{H}\mathbf{y} - \frac{\mathbf{y}'\mathbf{J}_n\mathbf{J}_n'\mathbf{y}}{n} \\ &= \mathbf{y}'\left[\mathbf{H} - \frac{\mathbf{J}_n\mathbf{J}_n'}{n}\right]\mathbf{y} \\ &= \sum_{i=1}^{n}(\widehat{y}_i - \overline{y})^2. \end{aligned}$$

So we see that

$$\begin{aligned} CSS(\text{total}) &= CSS(\text{model}) + SSE \\ \mathbf{y}'\left[\mathbf{I} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right]\mathbf{y} &= \mathbf{y}'\left[\mathbf{H} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right]\mathbf{y} + \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y}. \end{aligned}$$

Example: Computing Corrected Sums of Squares

The additional R code for computing the corrected sum of squares in R is given below.

```
> one = matrix(1, n) # column vec of 1s
> I_n = diag(as.vector(one)) # diagonal matrix of 1s
> css_t=t(y)%*%(I_n-(one%*%t(one))/n)%*%y
> css_m=t(y)%*%(H-(one%*%t(one))/n)%*%y
> print(css_t)

        [,1]
[1,] 15183.1

> print(css_m)

         [,1]
[1,] 5034.907

> print(css_m + sse)

        [,1]
[1,] 15183.1
```

We may also obtain this from PROC GLM

```
proc glm data=ozone;
model personal=outdoor home time_out;
run;
```

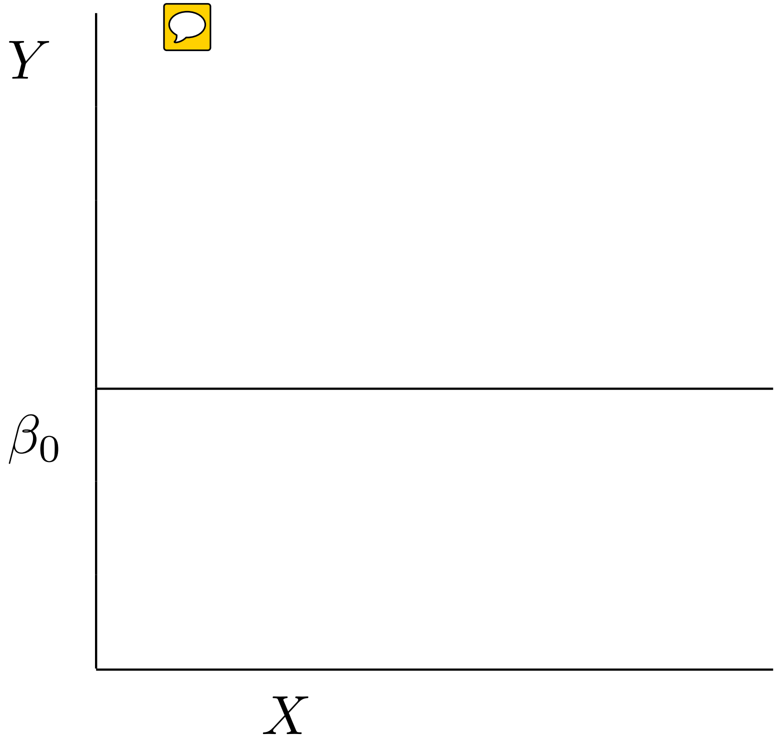
```
                               The GLM Procedure
          Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                        Sum of
 Source                      DF        Squares    Mean Square   F Value   Pr > F

 Model                        3     5034.90667     1678.30222      9.92   <.0001

 Error                       60    10148.19129      169.13652

 Corrected Total             63    15183.09796
```

The Intercept Only Model

Consider the following model with only an intercept:

$$\begin{aligned} \mathbf{y}_{n\times 1} &= \beta_0 + \boldsymbol{\varepsilon}_{n\times 1} \\ \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} &= \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \beta_0 + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}. \end{aligned}$$

It is easy to show that $\widehat{\beta_0} = \overline{y} = \sum_{i=1}^{n} \frac{y_i}{n}$, i.e.,that the estimate of β_0 is the average response. This model is sometimes called the *grand mean model*. The intercept only model predicts a constant ($\widehat{y}_i = \widehat{y}_j = \overline{y}$ for all i, j), and the predicted regression line has zero slope. For the ozone data, an intercept only model expects that personal exposure would be the same for all students. The estimate of the intercept is simply the mean personal exposure, so that we have $\widehat{\beta_0} = 23.55$.



Intercept-only Model

Although this model is not very useful by itself, it is very useful in model selection and testing. We will use it to answer questions like the following: "Does adding home exposure to this basic model give us any additional useful information about personal exposure?"

For the intercept only model, we have

- $\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}} = \mathbf{J}_n\widehat{\beta}_0 = \mathbf{J}_n\overline{y} = \begin{bmatrix} \overline{y} \\ \vdots \\ \overline{y} \end{bmatrix}_{n\times 1}$,
- $\widehat{\boldsymbol{\varepsilon}} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \begin{bmatrix} y_1 - \overline{y} \\ \vdots \\ y_n - \overline{y} \end{bmatrix}_{n\times 1}$,
- $SSE = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}} = \mathbf{y}'(\mathbf{I} - \mathbf{H})\mathbf{y} = \mathbf{y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right)\mathbf{y}$,
- $USS(\text{total}) = \mathbf{y}'\mathbf{y}$
- $USS(\text{model}) = \mathbf{y}'\mathbf{H}\mathbf{y} = \mathbf{y}'\left(\frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right)\mathbf{y}$,

- $$\begin{aligned} CSS(\text{total}) &= USS(\text{total}) - SSI \\ &= \mathbf{y}'\left(\mathbf{I} - \frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right)\mathbf{y} = SSE, \text{ and} \end{aligned}$$

- $$\begin{aligned} CSS(\text{model}) &= USS(\text{model}) - SSI \\ &= \mathbf{y}'\mathbf{H}\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}_n\mathbf{J}_n'\mathbf{y} \\ &= \mathbf{y}'\left(\frac{1}{n}\mathbf{J}_n\mathbf{J}_n'\right)\mathbf{y} - \frac{1}{n}\mathbf{y}'\mathbf{J}_n\mathbf{J}_n'\mathbf{y} \\ &= 0. \end{aligned}$$

Example: Computing Sums of Squares for the Intercept Only Model

```
proc glm data=ozone;
model personal= /;
run;

*************************************************************************

                               The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                       Sum of
 Source               DF        Squares    Mean Square   F Value   Pr > F

 Model                 1    35481.84414    35481.84414    147.23   <.0001

 Error                63    15183.09796      241.00155

 Uncorrected Total    64    50664.94210
```

From the SAS output, what are the values of the following?

- SSE:
- $USS(\text{total})$:
- $USS(\text{model})$:
- SSI:
- $CSS(\text{total})$:
- $CSS(\text{model})$:

What test is provided in the SAS output?

The Null Model

An even simpler model exists. The “null” model assumes the intercept is zero and all slopes are zero so that we have no parameters in the model (except for σ^2) and thus 0 model degrees of freedom. Call

$$\boldsymbol{y} = \boldsymbol{\varepsilon}$$

the *null model*, with $p = 0$ and $\boldsymbol{\beta} = \emptyset$.

Necessarily, $\widehat{\boldsymbol{y}} = \mathbf{0}$ and $\widehat{\boldsymbol{\varepsilon}} = \mathbf{y}$. We do not define corrected sums of squares for the null model because it does not span an intercept. For this model, we have

$$SSE \quad = \quad \widehat{\boldsymbol{\varepsilon}}'\widehat{\boldsymbol{\varepsilon}} = \mathbf{y}'\mathbf{y} = USS(\text{total})$$

and $USS(\text{model}) = USS(\text{total}) - SSE = 0$.

Like the intercept only model, this model is not very interesting by itself. However, it is also useful for purposes of model selection and testing.

Overall ANOVA Table for Multiple Regression

Typically, with all continuous predictors, $\boldsymbol{X}$ has full rank unless an error has been made. By default, for now we will assume that $\boldsymbol{X}$ less than full rank implies an error. Writing $\widehat{\boldsymbol{\beta}}$ presumes full rank $\boldsymbol{X}$, while $\tilde{\boldsymbol{\beta}}$ allows either full rank (rank$(\boldsymbol{X}) = p$) or less than full rank $\mathbf{X}$.

After fitting a model, one wishes to summarize the analysis and decide whether the variables in $\mathbf{X}$ are useful predictors (reduce the variability of $\boldsymbol{y} \mid \boldsymbol{X}$).

All regression tests considered are special cases of the GLH, which we have already studied.

Overall ANOVA Table for Model:

$$\text{Wingspan}_{\boldsymbol{i}} = \boldsymbol{\beta_0} + \boldsymbol{\beta_1}\text{Height}_{\boldsymbol{i}} + \boldsymbol{\beta_2}\text{Waist}_{\boldsymbol{i}} + \boldsymbol{\varepsilon_i}$$

Source	df	SS	MS	F_{obs}	p
Intercept	1	83.13			
Regression (uncorrected)	3	83.69			
Regression (corrected)	2	0.56			
Error (residual)	27	0.37			
Total (uncorrected)	30	84.06			
Total (corrected)	29	0.94			

Source indicates the variables which contribute the information.

Degrees of freedom (*df*) gives the dimension of the source (number of contributing variables).

Sums of squares (SS) are defined in the preceding section.

Mean Square (MS) $= SS/df$ are sums of squares adjusted for the sample size.

$F_{obs} = MS(\text{source})/MS(\text{error})$.

Under HILE Gauss and H_0: $\sigma^2_{\text{source}} = \sigma^2_{\text{error}}$,

$F_{obs} \sim F(df_{\text{source}}, df_{\text{error}})$, and

$p = \Pr\{F_{obs} \geq F(df_{\text{source}}, df_{\text{error}})\}$.

Overall ANOVA Table Formulas for Full Model
(Assuming $\boldsymbol{X}$ $(\boldsymbol{n} \times \boldsymbol{p})$ Spans an Intercept)

Source	SS Form	Scalar Form	Quadratic Form
Intercept	$SSE_\emptyset - SSE_0$	$n\overline{y}^2$	$\boldsymbol{y}'(\frac{\mathbf{JJ}'}{n})\boldsymbol{y}$
Regression (uncorrected)	$SSE_\emptyset - SSE_{p-1}$	$\sum_{i=1}^n \widehat{y}_i^2$	$\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}$
Regression (corrected)	$SSE_0 - SSE_{p-1}$	$\sum_{i=1}^n \widehat{y}_i^2 - n\overline{y}^2$	$\boldsymbol{y}'(\boldsymbol{H} - \frac{\mathbf{JJ}'}{n})\boldsymbol{y}$
Error (residual)	SSE_{p-1}	$\sum_{i=1}^n (y_i - \widehat{y}_i)^2$	$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$
Total (uncorrected)	$SSE_\emptyset$	$\sum_{i=1}^n y_i^2$	$\boldsymbol{y}'\boldsymbol{y}$
Total (corrected)	SSE_0	$\sum_{i=1}^n y_i^2 - n\overline{y}^2$	$\boldsymbol{y}'(\boldsymbol{I} - \frac{\mathbf{JJ}'}{n})\boldsymbol{y}$

Corrected sums of squares and associated statistics are not defined if a model does not span an intercept. Corrected sums of squares and tests always include the intercept in models and exclude it from tests, which is generally our objective.

Usual ("Corrected") Overall Test for Regression

Consider testing whether the predictors in $\boldsymbol{X}$ have any value. This corresponds to the hypothesis that all slopes are zero, which corresponds to no predictive contribution (explanatory value) of the covariates, except possibly the intercept.

In this test we compare the full model,

$$y_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j + \varepsilon_i,$$

to the reduced model,

$$y_i = \beta_0 + \varepsilon_i,$$

to test $H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0.$

$$
\begin{aligned}
F_{obs} &= \frac{MS(\text{hypothesis})}{MSE} = \frac{SSH/dfH}{SSE/dfE} \\
&= \frac{[SSE(\text{reduced}) - SSE(\text{full})]/[dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full})/dfE(\text{full})} \\
&= \frac{CSS(\text{Regression})/(p-1)}{SSE(full)/(n-p)}\,.
\end{aligned}
$$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1-\alpha,\, p-1,\, n-p) = f_{crit}$.

The usual test of overall regression assumes model spans an intercept and excludes the intercept from the test.

Overall Corrected ANOVA Table Formulas for Full Model (Assuming $\boldsymbol{X}$ $(\boldsymbol{n} \times \boldsymbol{p})$ Spans an Intercept)

Source	Scalar Form	Quadratic Form	df
Intercept	$n\overline{y}^2$	$\boldsymbol{y}'(\frac{\mathbf{JJ}'}{n})\boldsymbol{y}$	1
Regression (corrected)	$\sum_{i=1}^{n} \widehat{y}_i^2 - n\overline{y}^2$	$\boldsymbol{y}'(\boldsymbol{H} - \frac{\mathbf{JJ}'}{n})\boldsymbol{y}$	$p-1$
Error (residual)	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$	$n-p$
Total (corrected)	$\sum_{i=1}^{n} y_i^2 - n\overline{y}^2$	$\boldsymbol{y}'(\boldsymbol{I} - \frac{\mathbf{JJ}'}{n})\boldsymbol{y}$	$n-1$

So CSS(total) $=$ CSS(model) $+$ SSE. To find USS(total) and USS(model), simply add SSI to the corresponding corrected values (and adjust the degrees of freedom accordingly).

Example: Conducting the “Corrected” Overall Test for Regression

```
proc glm data=ozone;
model personal= outdoor home time_out;
run;
**********************************************************************
                               The GLM Procedure

Dependent Variable: personal   Personal Ozone Exposure (ppb)

                                       Sum of
 Source               DF        Squares    Mean Square   F Value   Pr > F

 Model                 3     5034.90667     1678.30222      9.92   <.0001

 Error                60    10148.19129      169.13652

 Corrected Total      63    15183.09796
```

“Uncorrected” Overall Test for Regression

Consider testing whether the variables in $\boldsymbol{X}$, including the intercept, have any value as predictors.

This corresponds to the hypothesis that all slopes and the intercept are zero, which corresponds to no contribution of any predictor.

In order to do this, compare the full model,

$$y_i = \beta_0 + \sum_{j=1}^{p-1} x_{ij}\beta_j + \varepsilon_i,$$

to the reduced model,

$$y_i = \varepsilon_i.$$

This yields H_0: $\beta_0 = \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0.$

The test statistic for the uncorrected overall test is

$$\begin{aligned} F_{obs} &= \frac{[SSE(\text{reduced}) - SSE(\text{full})] \big/ [dfE(\text{reduced}) - dfE(\text{full})]}{SSE(\text{full}) \big/ dfE(\text{full})} \\ &= \frac{USS(\text{Regression}) \big/ p}{SSE(\text{full}) \big/ (n-p)} \,. \end{aligned}$$

Reject the hypothesis if $F_{obs} \geq F_F^{-1}(1-\alpha,\, p,\, n-p) = f_{crit}$.

This test is well defined whether or not the model spans an intercept, but it rarely has scientific value.

Overall Uncorrected ANOVA Table Formulas for Full Model (Assuming $\boldsymbol{X}$ ($\boldsymbol{n} \times \boldsymbol{p}$) Spans an Intercept)

Source	Scalar Form	Quadratic Form	df
Regression (uncorrected)	$\sum_{i=1}^{n} \widehat{y}_i^2$	$\boldsymbol{y}'\boldsymbol{H}\boldsymbol{y}$	p
Error (residual)	$\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2$	$\boldsymbol{y}'(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}$	$n - p$
Total (uncorrected)	$\sum_{i=1}^{n} y_i^2$	$\boldsymbol{y}'\boldsymbol{y}$	n

So $USS(\text{total}) = USS(\text{model}) + SSE$, and SSI is included in the regression sum of squares.

Strength of Association

Decomposing Response Variance

Statistical “significance” does not measure scientific importance.

We may decompose the variance into variance explained by the model and random error, and then we may ask: what fraction of Y variance does $\boldsymbol{X}$ predict?

Usual “Corrected” R^2

For a model spanning an intercept, define the proportion of variance in Y predictable from the X’s, adjusted for the intercept, as

$$R_{\mathsf{c}}^2 = \frac{CSS(\text{Regression})}{CSS(\text{Regression}) + SSE(\text{full})}$$

$$= \frac{CSS(\text{Regression})}{CSS(\text{total})}$$

R_c^2 estimates ρ_{c}^2, the population ratio of model to total variance, with $0 \leq \rho_{\mathsf{c}}^2 \leq 1$ and $0 \leq R_{\mathsf{c}}^2 \leq 1$.

Facts about R_c^2:

- R_c^2 is the maximum likelihood estimate of ρ_{c}^2 under HILE Gauss.
- R_c^2 is biased in general: $E[R_{\mathsf{c}}^2] \geq \rho_{\mathsf{c}}^2$, with equality for $\rho_{\mathsf{c}}^2 = 1$ or $n \to \infty$.
- R_{c}^2 is invariant with respect to full rank linear transformation of the response or predictors (including location and scale changes).
- The corrected test for overall regression,

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0$$

holds if and only if

$$H_0 : \rho_{\mathsf{c}}^2 = 0$$

is true.

“Uncorrected” R^2

For any model define the proportion of variance in Y predictable from the X's, including the intercept if spanned, as

$$R^2_{\text{u}} = \frac{USS(\text{Regression})}{USS(\text{Regression}) + SSE(\text{full})} \ ,$$

$$= \frac{USS(\text{Regression})}{USS(\text{total})}$$

where $0 \leq R^2_{\text{u}} \leq 1$ and $0 \leq \rho^2_{\text{u}} \leq 1$.

Facts about R_u^2:

- R^2_{u} may vary due to any linear transformation of Y or X's.
- The uncorrected test for overall regression,

  $$H_0 : \beta_0 = \beta_1 = \beta_2 = \ldots = \beta_{p-1} = 0,$$

  holds if and only if

  $$H_0 : \rho^2_{\mathsf{u}} = 0$$

  is true.
- R_u^2 is a biased estimate of ρ_u^2
- Uncorrected R^2 is defined for any GLM, while corrected R^2 is defined only for models that span an intercept.
- Neither corrected nor uncorrected R^2 is always best, although corrected R^2 should be the default choice.

Comparing Corrected and Uncorrected R^2

$$\begin{aligned} R_c^2 &= \frac{CSS(\text{Model})}{CSS(\text{Total})} \\ R_u^2 &= \frac{USS(\text{Model})}{USS(\text{Total})} = \frac{CSS(\text{Model}) + SSI}{CSS(\text{Total}) + SSI} \end{aligned}$$

$$0 \leq R_{\mathsf{c}}^2 \leq 1$$

$$0 \leq R_{\mathsf{u}}^2 \leq 1$$

As always, $SSI = n\overline{y}^2$.

Note: R^2 always increases when additional predictors are added to the model, whether or not they are practically or statistically important.

Exercise: Computing Corrected and Uncorrected R^2

Compute corrected and uncorrected R^2 for the ozone data.

Adjusted R^2

One potential problem with using R^2 (either corrected or uncorrected) to assess model adequacy is that R^2 never decreases when you add additional predictors to the model. (Even if a predictor is meaningless, R^2 will never be smaller when it is added.) Although we will later learn how to conduct hypothesis tests about whether an increase in R^2 is statistically significant, some investigators prefer to use the *adjusted* R^2, defined as

$$R^2{}_{adj} = 1 - \frac{SSE/(n-r)}{CSS(\text{Total})/(n-1)},$$

which is adjusted for the degrees of freedom. It will only increase on adding a variable to the model if the variable reduces the mean square for error. The fact that adjusted R^2 penalizes us for adding terms that are not useful makes it helpful in evaluating and comparing a set of candidate regression models.

Next: Testing

Reading Assignment:

- Muller and Fetterman, Chapter 5: “Testing Hypotheses in Multiple Regression” (Required)