Outline Lecture 15 & 16

- Chapter 5.1 Factors (or Qualitative predictors) in regression *(Ref. Applied Linear Regression by S. Weisberg)*
- Regression with one factor only
- Regression with one factor and other (continuous) predictors

Chapter 5.2 Multiple Factors.

Chapter 6 F-test.

5.1 Factors

- Factors: qualitative or categorical predictors.
  Factors can have two levels, such as male/female, treatment/control, ... or many levels, such as education levels, ...
- As an example, consider the United Nations data
  http://users.stat.umn.edu/ sandy/alr4ed/data/UN11.csv.

* This is an observational study of all $n = 199$ localities.
* The factor we use is called group, which classified the countries into three categories, **africa** for the 53 countries on the African continent, **oecd** for the 31 countries that are members of the OECD, the Organisation for Economic Co-operation and Development, an international body1 whose members are generally the wealthier nations, none of which are in Africa, and **other** for the remaining 115 countries in the data set that are neither in Africa nor in the OECD.
* The variable group is a factor, with these d $= 3$ levels. We will use as a response the variable lifeExpF, the expected life span of women in each country, and so the problem at first is to see how lifeExpF differs between the three groups of countries.

5.1.1 One-Factor Models

- Factor predictors can be included in a multiple linear regression mean function using **dummy variables**.
- For a factor with two levels, a single dummy variable, a regressor that takes the value 1 for one of the categories and 0 for the other category, can be used.

* Assignment of labels to the values is generally arbitrary, and will not change the outcome of the analysis.
  Dummy variables can alternatively be defined with a different set of values, perhaps -1 and 1, or possibly 1 and 2. The important point is the regressor has only two values.

Since `group` has $d = 3$ levels, the jth dummy variable U_j for the factor, $j = 1, \ldots, d$ has ith value u_{ij}, for $i = 1, \ldots, n$, given by

$$u_{ij} = \begin{cases} 1 & \text{if } \texttt{group}_i = j\text{th category of } \texttt{group} \\ 0 & \text{otherwise} \end{cases} \tag{5.1}$$

The values of the dummy variables for the first 10 cases in the example are as follows:

	`Group`	U_1	U_2	U_3
Afghanistan	other	0	1	0
Albania	other	0	1	0
Algeria	africa	0	0	1
Angola	africa	0	0	1
Anguilla	other	0	1	0
Argentina	other	0	1	0
Armenia	other	0	1	0
Aruba	other	0	1	0
Australia	oecd	1	0	0
Austria	oecd	1	0	0

The variable U_1 is the dummy variable for the first level of `group`, which is `oecd`, U_2 is for `other`, and U_3 is for the remaining level `africa`.

- If we add an intercept to the mean function, the resulting model would be overparameterized because $U_1 + U_2 + U_3 = 1$, a column of 1's, and the column of 1's is the regressor that corresponds to the intercept.
- This problem can be solved by dropping one of the dummy variables:

mean function $$\mathrm{E}(\texttt{lifeExpF}|\texttt{group}) = \beta_0 + \beta_2 U_2 + \beta_3 U_3 \qquad (5.2)$$

Since the first level of group will be implied when $U_2 = U_3 = 0$,

$$\mathrm{E}(\texttt{lifeExpF}|\texttt{group} = \texttt{oecd}) = \beta_0 + \beta_2 0 + \beta_3 0 = \beta_0$$

and so β_0 is the mean for the first level of group. For the second level $U_2 = 1$ and $U_3 = 0$,

$$\mathrm{E}(\texttt{lifeExpF}|\texttt{group} = \texttt{other}) = \beta_0 + \beta_2 1 + \beta_3 0 = \beta_0 + \beta_2$$

and $\beta_0 + \beta_2$ is the mean for the second level of group. Similarly, for the third level $U_2 = 0$ and $U_3 = 1$

$$\mathrm{E}(\texttt{lifeExpF}|\texttt{group} = \texttt{africa}) = \beta_0 + \beta_2 0 + \beta_3 1 = \beta_0 + \beta_3$$

Most computer programs allow the user to use a factor[3] in a mean function without actually computing the dummy variables. For example, the `R` package uses notation for indicating factors and interactions first suggested by Wilkinson and Rogers (1973). If `group` has been declared to be a factor, then the mean function (5.2) is be specified by

```
lifeExpF ~ 1 + group
```
(5.3)

where the "`1`" specifies fitting the intercept, and `group` specifies fitting the *dummy variable regressors that are created for the factor* `group`. Since most mean functions include an intercept, `R` assumes it will be included, and the specification

```
lifeExpF ~ group
```
(5.4)

is equivalent to (5.3).[4]

Table 5.1 Regression Summary for Model (5.4)

	Estimate	Std. Error	t-Value	Pr(>|t|)
(Intercept), $\hat{\beta}_0$	82.4465	1.1279	73.09	0.0000
`other`, $\hat{\beta}_2$	–7.1197	1.2709	–5.60	0.0000
`africa`, $\hat{\beta}_3$	–22.6742	1.4200	–15.97	0.0000

$\hat{\sigma} = 6.2801$ with 196 df, $R^2 = 0.6191$.

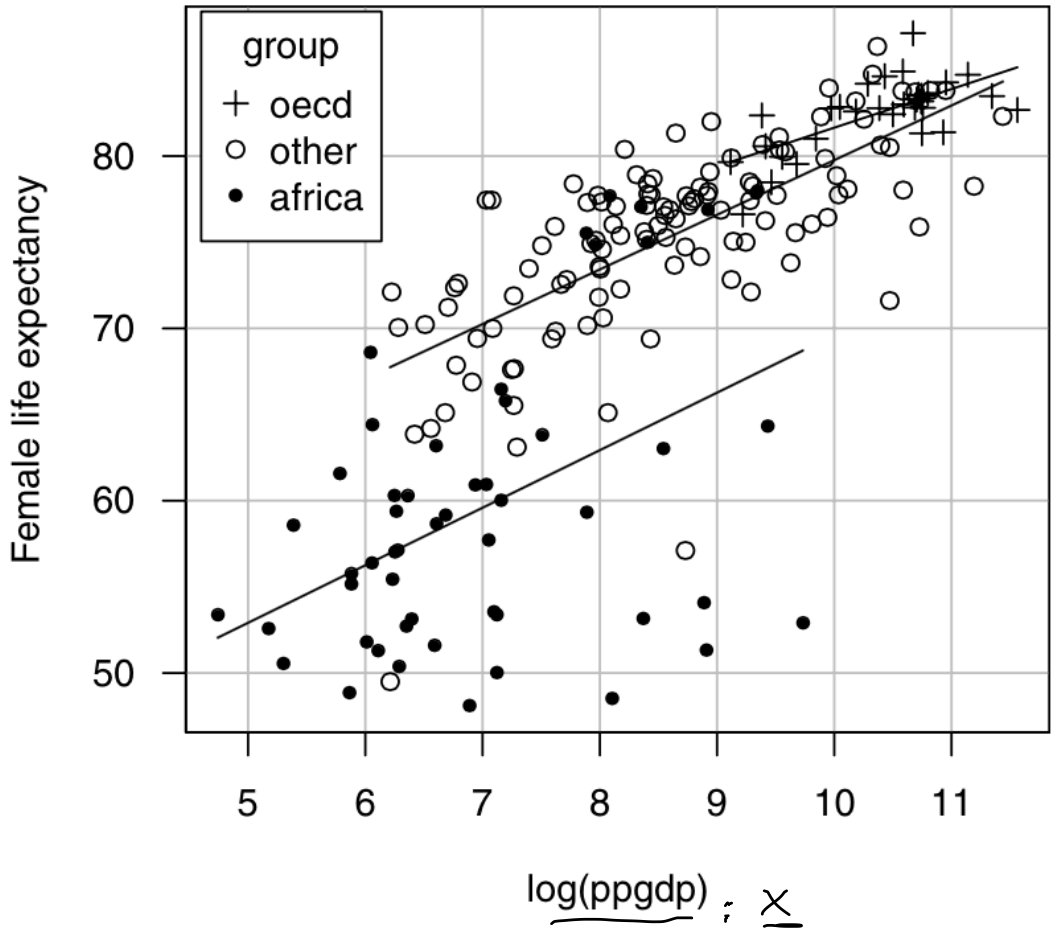
3 groups are

$$\begin{aligned}
\hat{E}(\texttt{lifeExpF}|\texttt{group} = \texttt{oecd}) &= \hat{\beta}_0 + \hat{\beta}_2 0 + \hat{\beta}_3 0 = 82.45 \\
\hat{E}(\texttt{lifeExpF}|\texttt{group} = \texttt{other}) &= \hat{\beta}_0 + \hat{\beta}_2 1 + \hat{\beta}_3 0 = 82.45 - 7.12 \qquad (5.5) \\
\hat{E}(\texttt{lifeExpF}|\texttt{group} = \texttt{africa}) &= \hat{\beta}_0 + \hat{\beta}_2 0 + \hat{\beta}_3 1 = 82.45 - 22.67
\end{aligned}$$

How to interpret the p-values?

5.1.3 Adding a Continuous Predictor

- As an additional predictor in the UN example, suppose we add log(ppgdp), the per person gross domestic product in the country, as a measure of relative wealth.



ppgdp

↓ log(ppgdp)

(More discussion on transformation in next week)

5.1.3 Adding a Continuous Predictor

		intercept	slope
$j=1$	oecd	η_{01}	η_{11}
$j=2$	other	η_{02}	η_{12}
$j=3$	africa	η_{03}	η_{13}

The model fit to obtain the three lines in Figure 5.2 corresponds fitting a separate intercept and slope in each group. Writing $\texttt{group} = j$ to represent an observation in level j, $j \in \{1, 2, 3\}$ (oecd, other, africa)

$$\mathrm{E}(\texttt{lifeExpF} | \log(\texttt{ppgpd}) = x, \texttt{group} = j) = \eta_{0j} + \eta_{1j} x \tag{5.6}$$

where (η_{0j}, η_{1j}) are the intercept and slope for level $j = 1, \ldots, d$, $d = 3$, so there $2d = 6$ parameters. This model is generally parameterized differently using *main effects* and *interactions*, as (in the mean functions)

$$\begin{aligned}\mathrm{E}(\texttt{lifeExpF} | \log(\texttt{ppgdp}) = x, \texttt{group}) &= \beta_0 + \beta_{02} U_2 + \beta_{03} U_3 \\ &+ \beta_1 x + \beta_{12} U_2 x + \beta_{13} U_3 x\end{aligned} \tag{5.7}$$

interaction terms

$Y \sim$ group + log(ppgdp) + group × log(ppgdp)

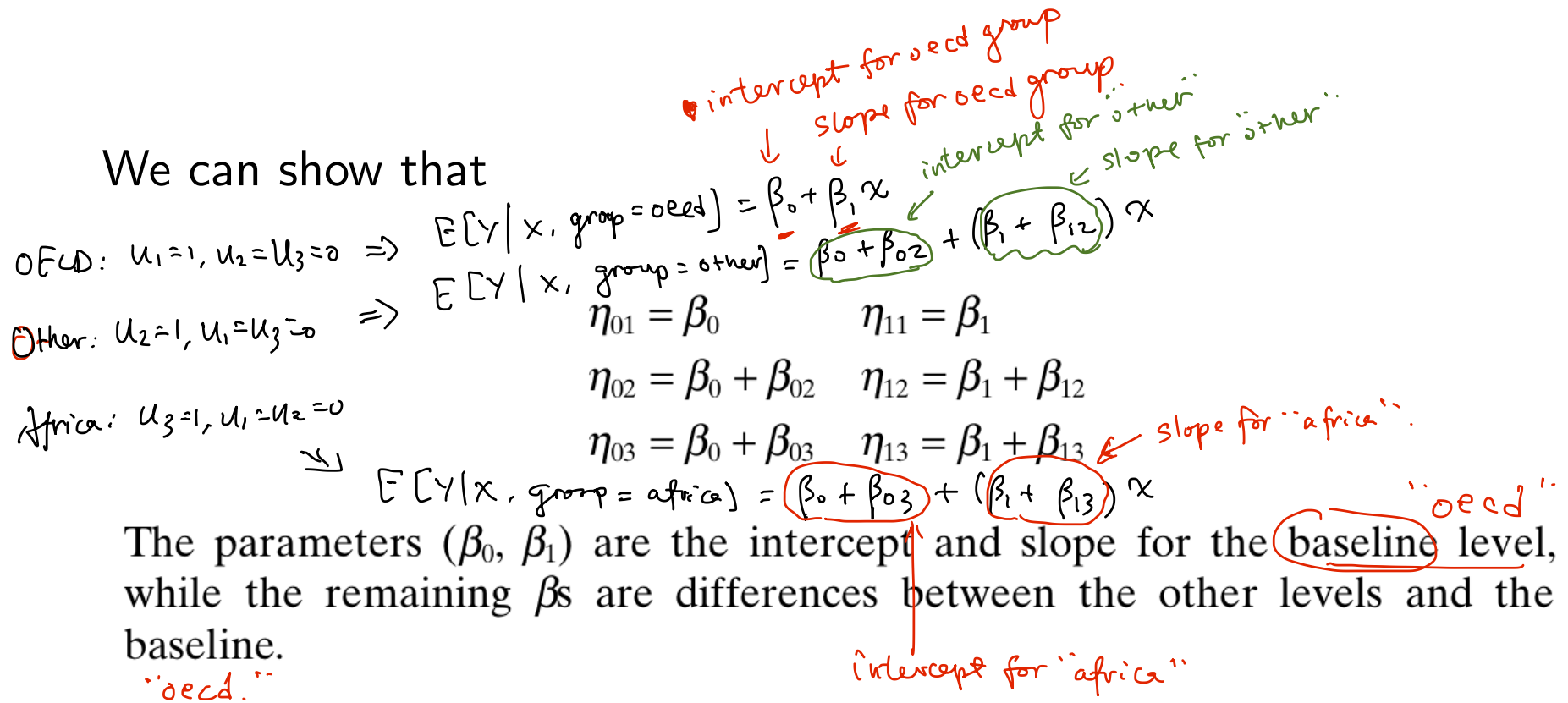
intercept: $1 + 2 + 1 + 2 \times 1 = 6$ parameters in the mean function

5.1.3 Adding a Continuous Predictor

We can show that

$$\eta_{01} = \beta_0 \qquad \eta_{11} = \beta_1$$
$$\eta_{02} = \beta_0 + \beta_{02} \qquad \eta_{12} = \beta_1 + \beta_{12}$$
$$\eta_{03} = \beta_0 + \beta_{03} \qquad \eta_{13} = \beta_1 + \beta_{13}$$

The parameters (β_0, β_1) are the intercept and slope for the baseline level, while the remaining βs are differences between the other levels and the baseline.



5.1.3 Adding a Continuous Predictor

Statistical packages generally allow (5.7) to be fit symbolically. In R one specification is

```
lifeExpF~group+log(ppgdp)+group:log(ppgdp)
```

The colon ":" is the indicator for an interaction in R. There is a shorthand for this available in R,

```
lifeExpF~group*log(ppgdp)
```

The asterisk "*" in R expands to include all main effects and interactions.

5.1.3 Adding a Continuous Predictor

Table 5.3 Regression Summary for Model (5.7)

	Estimate	Std. Error	t-Value	Pr(>|t|)
(Intercept), $\hat{\beta}_0$	59.2137	15.2203	3.89	0.0001
`other`, $\hat{\beta}_{02}$	–11.1731	15.5948	–0.72	0.4746
`africa`, $\hat{\beta}_{03}$	–22.9848	15.7838	–1.46	0.1470
log(`ppgdp`), $\hat{\beta}_1$	1.5544	1.0165	1.53	0.1278
`other`: log(`ppgdp`), $\hat{\beta}_{12}$	0.6442	1.0520	0.61	0.5410
`africa`: log(`ppgdp`), $\hat{\beta}_{13}$	0.7590	1.0941	0.69	0.4887

$\hat{\sigma} = 5.1293$ with 193 df, $R^2 = 0.7498$.

5.1.3 Adding a Continuous Predictor

Figure 5.3 provides two variations of effects plots for the fit of the interaction model.

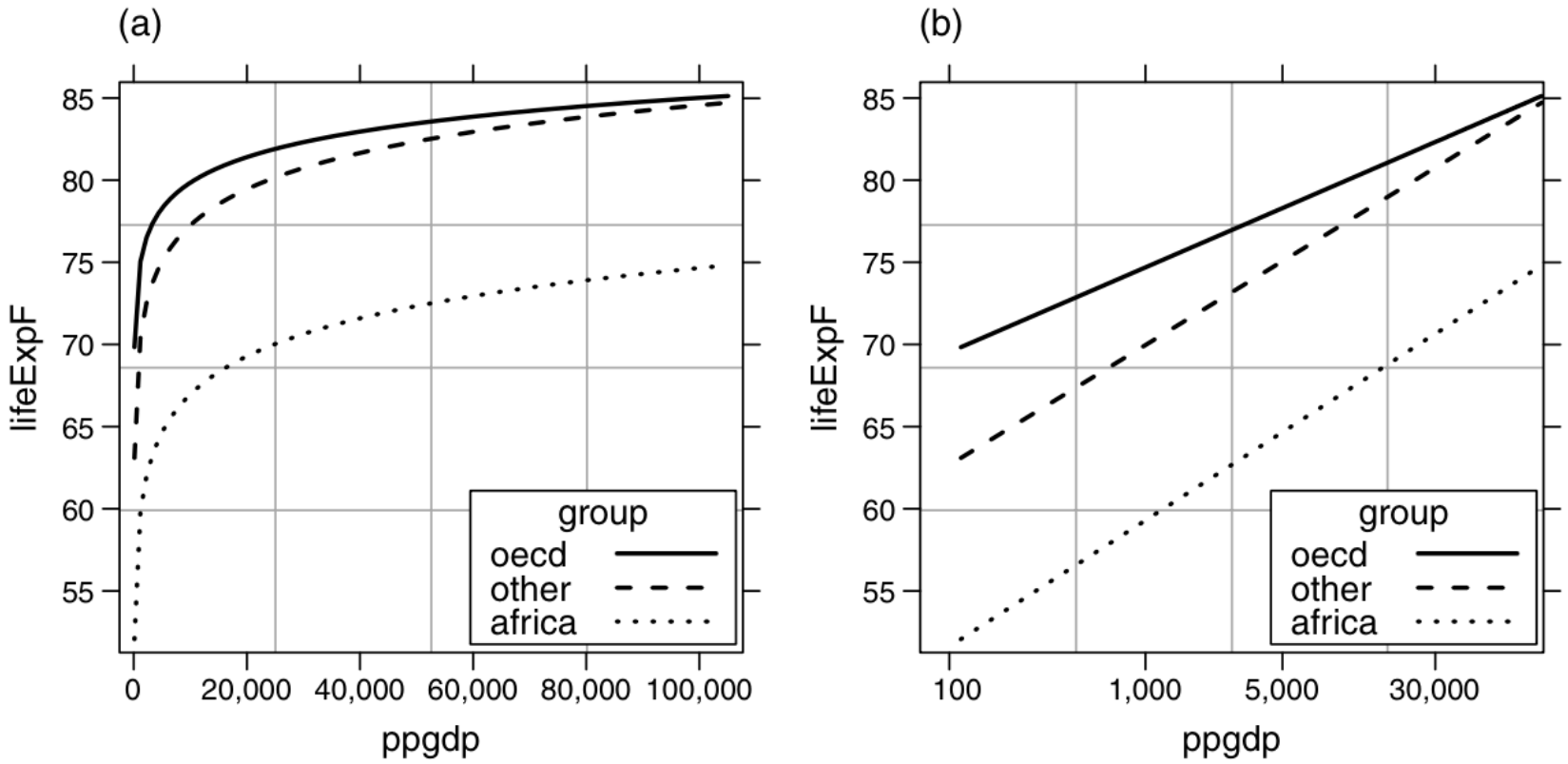


Figure 5.3 Effects plot for the interaction model (5.7) for the UN data. (a) `ppgdp` on the horizontal axis. (b) `ppgdp` in log-scale.

5.1.4 The Main Effects Model

Examination of Figure 5.2 suggests that while intercepts might differ for the three levels of `group`, the slopes may be equal. This suggests fitting a model that allows each group to have its own intercept, but all groups have the same slope,

$$\mathrm{E}(\texttt{lifeExpF} \mid \log(\texttt{ppgdp}) = x, \texttt{group}) = \beta_0 + \beta_{02}U_2 + \beta_{03}U_3 + \beta_1 x \quad (5.8)$$

Model (5.8), whose Wilkinson–Rogers representation is `lifeExpF ~ log(ppgdp) + group`, is obtained from (5.7) by dropping the interaction, so we call this a *main effects model*. Main effects models are much simpler

Interpretation: β_0, β_{02}, β_{03} same as model (5.7)

β_1: common slope for all 3 levels

5.1.4 The Main Effects Model

- Main effects models are much simpler than are models with interactions because the effect of the continuous regressor is the same for all levels of the factor.
- Similarly, the difference between levels of the factor are the same for every fixed value of the continuous regressor.

5.1.4 The Main Effects Model

Table 5.4 Regression Summary for Model (5.8)

	Estimate	Std. Error	t-value	Pr(>|t|)
(Intercept), $\hat{\beta}_0$	49.5292	3.3996	14.57	0.0000
`other`, $\hat{\beta}_{02}$	−1.5347	1.1737	−1.31	0.1926
`africa`, $\hat{\beta}_{03}$	−12.1704	1.5574	−7.81	0.0000
log(`ppgdp`), $\hat{\beta}_1$	2.2024	0.2190	10.06	0.0000

$\hat{\sigma} = 5.1798$ with 195 df, $R^2 = 0.7422$.

5.2 Many factors

- Increasing the number of factors or the number of continuous predictors in a mean function can add considerably to complexity but does not really raise new fundamental issues.
- Consider first a problem with many factors but no continuous predictors. The data in the file *Wool* are from a small experiment to understand the strength of wool as a function of three factors that were under the control of the experimenter (Box and Cox, 1964). $n = 27$

Table 5.5 The Wool Data

Variable	Definition
factors: `len`	Length of test specimen (250, 300, 350 mm)
factors: `amp`	Amplitude of loading cycle (8, 9, 10 mm)
factors: `load`	Load put on the specimen (40, 45, 50 g)
response: log(`cycles`)	Logarithm of the number of cycles until the specimen fails

- Each of the three factors was set to one of three levels, and all $3^3 = 27$ possible combinations of the three factors were used.

Many factors

Consider response Y=log (cycles) and factors X1 = len, X2 = amp, and X3 = load. We can construct the following models:

- M1: main-effects mean function. A main effects mean function for these data includes an intercept and two dummy variables for each of the factors, for a total of 7 parameters. (in the mean function)

$$Y \sim X1 + X2 + X3$$
$$1 + 2 + 2 + 2 = 7$$

- M2: full second-order mean function. A full second-order mean function adds all the two-factor interactions to the mean function (7+3*4=19 parameters).

$$Y \sim X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3$$
$$1 + 2 + 2 + 2 + 2\times 2 + 2\times 2 + 2\times 2$$

- M3: full third-order mean function. The third-order model includes the three-factor interaction (19 + 8 = 27 parameters).

$$Y \sim X1 + X2 + X3 + X1:X2 + X1:X3 + X2:X3 + X1:X2:X3$$
$$2\times 2\times 2$$

Many factors

mean function for M1

len $X_1 : u_{11}, u_{12}, u_{13}$ (dummy var.), amp $X_2 : u_{21}, u_{22}, u_{23}$, load $X_3 : u_{31}, u_{32}, u_{33}$

$$E[Y|x_1, x_2, x_3] = \beta_0 + \beta_{12}u_{12} + \beta_{13}u_{13} + \beta_{22}u_{22} + \beta_{23}u_{23} + \beta_{32}u_{32} + \beta_{33}u_{33} \quad (*)$$

$$\Rightarrow E[Y|u_{11}=1, u_{21}=1, u_{31}=1] = \beta_0.$$

len=250, amp=8, load=40

$$E[Y|u_{12}=1, u_{21}=1, u_{31}=1] = \beta_0 + \beta_{12} \Rightarrow \beta_{12} = E[Y|u_{12}=1,u_{21}=1,u_{31}=1] - E[Y|u_{11}=1,u_{21}=1,u_{31}=1]$$

len=300, amp=8, load=40

mean function for M2 :

$$E[Y|x_1, x_2, x_3] = (*) + \beta_{12,22}u_{12}u_{22} + \beta_{12,23}u_{12}u_{23} + \beta_{13,22}u_{13}u_{22} + \beta_{13,23}u_{13}u_{23}$$

($x_1 \cdot x_2$)

$$+ \beta_{12,32}u_{12}u_{32} + \beta_{12,33}u_{12}u_{33} + \beta_{13,32}u_{13}u_{32} + \beta_{13,33}u_{13}u_{33} \quad (\rightarrow x_1 \cdot x_3)$$

$$+ \beta_{22,32}u_{22}u_{32} + \beta_{22,33}u_{22}u_{33} + \beta_{23,32}u_{23}u_{32} + \beta_{23,33}u_{23}u_{33} \quad (\rightarrow x_2 \cdot x_3)$$

19 β parameters.

F-Test to comparing different models

(Chapter 6. Applied linear Reg)

6.1 *F*-TESTS

Suppose we have a response Y and a vector of p' regressors $\mathbf{X} = (\mathbf{X}_1', \mathbf{X}_2')$ that we partition into two parts so that $\mathbf{X}_2$ has q regressors and $\mathbf{X}_1$ has the remaining $p' - q$ regressors. The intercept, if present, is generally included in $\mathbf{X}_1$, but this is not required. The general hypothesis test we consider is

$$\begin{aligned} &\text{NH:} \quad \mathrm{E}(Y|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2) = \mathbf{x}_1'\boldsymbol{\beta}_1 \\ &\text{AH:} \quad \mathrm{E}(Y|\mathbf{X}_1 = \mathbf{x}_1, \mathbf{X}_2 = \mathbf{x}_2) = \mathbf{x}_1'\boldsymbol{\beta}_1 + \mathbf{x}_2'\boldsymbol{\beta}_2 \end{aligned} \tag{6.2}$$

This is a different approach to hypothesis testing, since the null and alternative models refer to specification of mean functions, rather than to restrictions on parameters. A necessary condition for the methodology of this section to apply is that the model under NH must be a special case of the model under AH. In (6.2), the NH is obtained by setting $\boldsymbol{\beta}_2 = \mathbf{0}$.

6.1 F-Tests

- For any linear regression model, the residual sum of squares measures the amount of variation in the response not explained by the regressors.
- If the NH were false, then the residual sum of squares RSS_{AH} under the alternative model would be considerably smaller than the residual sum of squares RSS_{NH} under the null model.
- This provides the basis of a test, and we will have evidence against the NH if the difference $(RSS_{NH} - RSS_{AH})$ is large enough.

The general formula for the test is

$$F = \frac{(\mathrm{RSS}_{NH} - \mathrm{RSS}_{AH})/(df_{NH} - df_{AH})}{\mathrm{RSS}_{AH}/df_{AH}} \tag{6.3}$$

$$= \frac{\mathrm{SSreg}/df_{Reg}}{\hat{\sigma}^2} \tag{6.4}$$

$\hat{\sigma}^2$ estimated under H_A model.

In this equation, df_{NH} and df_{AH} are the *df* for residual under NH and AH, $\mathrm{SSreg} = \mathrm{RSS}_{NH} - \mathrm{RSS}_{AH}$ is the *sum of squares for regression*, and $df_{Reg} = df_{NH} - df_{AH}$ is its *df*. The denominator of the statistic is generally the estimate of σ^2 computed assuming that AH is true, $\hat{\sigma}^2 = \mathrm{RSS}_{AH}/df_{AH}$, but as we will see later, other choices are possible. A sum of squares divided by its *df* is called a *mean square*, and so the *F*-test is the mean square for regression divided by the mean square for error under AH.

as in Anova Tables (Type I, Type II).

6.1 F-Tests

- The F-test as described here appears to require fitting the model under both NH and AH, getting the residual sums of squares, and then applying (6.3).
- Computer packages generally take advantage of the elegant structure of the linear regression model to compute the test while fitting only under the AH by computing SS_{reg} in (6.4) directly.
- If we assume that the errors are $NID(0, \sigma^2)$ random variables, then if NH is true, (6.3) has an $F(df_{Reg}, df_{AH})$-distribution, and large values of F provide evidence against the NH

UN data example

UN Data

The UN data discussed in Section 5.1 considered a sequence of mean functions given in Wilkinson–Rogers notation as

Mean function	*df*	RSS	
`lifeExpF ~ 1`	198	20293.2	(6.6)
`lifeExpF ~ group`	196	7730.2	(6.7)
`lifeExpF ~ log(ppgdp)`	197	8190.7	(6.8)
`lifeExpF ~ group + log(ppgdp)`	195	5090.4	(6.9)
`lifeExpF ~ group + log(ppgdp) + group:log(ppgdp)`	193	5077.7	(6.10)

The first of these models (6.6) is the null model, so it has residual sum of squares equal to `SYY`, and $df = n - 1$. Mean function (6.7) has a separate mean for each level of `group` but ignores log(`ppgdp`). Mean function (6.8) has a common slope and intercept for each level of `group`; (6.9) has separate intercepts but a common slope. The most general (6.10) has separate slopes and intercepts.

Tests can be derived to compare most of these mean functions. A reasonable procedure is to start with the most general, comparing NH: mean function (6.9) to AH: mean function (6.10),

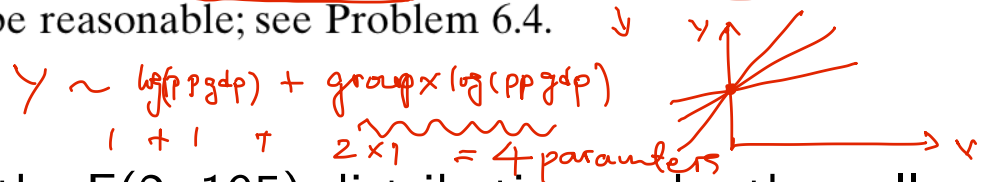
$$F = \frac{(5090.4 - 5077.7)/(195 - 193)}{5077.7/193} = 0.24 \qquad (6.11)$$

When compared with the $F(2, 193)$ distribution, we get a p-value of 0.79, providing no evidence of the need for separate slopes, confirming the visual impression of Figure 5.2.

If this first test had suggested that separate slopes and intercepts were needed, then further testing would not be needed.[2] Since the interaction is probably unnecessary, we can consider further testing using the first-order mean function (6.9) as AH, and either (6.7) or (6.8) as NH. For the test for (6.8) versus (6.9), we get

$$F = \frac{(8190.7 - 5090.4)/(197 - 195)}{5090.4/195} = 59.38$$

[2] A model not considered here would have a common intercept but separate slopes, and a test of this model versus model (6.10) could be reasonable; see Problem 6.4.



- The F-test statistic follows the F(2, 195) distribution under the null hypothesis, and we get a p-value of essentially 0, providing strong evidence that intercepts for the three levels of group are not all equal.

Wool data example

Wool Data $n = 27$

With this example, we show that this testing paradigm can be used in more complex situations beyond the overall test. For the wool data, Section 5.2, the predictors `len`, `amp`, and `load` are factors, each with 3 levels. We can consider testing

$1 + 2 + 2 + 2 + 2 \times 2 + 2 \times 2 = 15$ parameters

NH: `log(cycles) ~ len + amp + load + len:amp + len:load`

AH: `log(cycles) ~ len + amp + load + len:amp + len:load + amp:load`

2×2

(19 parameters in the mean function)

The statement of these hypotheses use the Wilkinson and Rogers (1973) notation. The NH model includes three main effects and two interactions. The AH includes all these regressors plus the `amp:load` interaction. Under NH this last interaction is zero, and under AH it is nonzero. The regressors that are common to NH and AH are estimated under both models. Thus, the desired test is for adding `amp:load` to a model that includes other regressors.

Hypothesis	*df*	RSS

Hypothesis	df	RSS
H_0 NH	12 $= 27-15$	0.181
H_A AH	8 $= 27-19$	0.166

For the F-test we estimate σ^2 under AH as $\hat{\sigma}^2 = 0.166/8 = 0.0208$, and

$$F = \frac{(0.181 - 0.166)/(12-8)}{0.0208 = 0.166/8} = 0.18$$

The F-test statistic follows the F(4, 8) distribution under the null hypothesis, and we get a p-value of 0.94, suggesting no evidence against NH.

> 0.05