MTH 4230 R Notes 2

1 General Linear *F* Test in Simple Linear Regression

- There are a few ways to carry out a ***general linear F test*** for simple linear regression in R. The method described below will be applicable later to multiple linear regression.
- Suppose we want to carry out the *general linear F test* to decide between two models, the ***full model***

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

and the ***reduced model***

$$Y_i = \beta_0 + \epsilon_i$$

Note that the reduced model is a **special case** of the full model in which $\beta_1 = 0$.

To carry out the test in R, we fit both models using `lm()`, and then simultaneously pass the resulting *lm* objects to the `anova()` function.

In the calls to `lm()`, these models are specified via the *formulas* as `y ~ x` (**full model**) and `y ~ 1` (**reduced model**).

- For example, consider the data in the vectors `x` and `y`:

```
x <- c(22, 15, 7, 19, 20, 9, 15, 10, 19, 21)
y <- c(12.2, 9.5, 6.7, 5.9, 10.0, 8.9, 11.5, 10.0, 9.9, 10.1)
```

and suppose we want to decide between the **full** and **reduced models** for these data.

The ***general linear F test*** is a test of hypotheses that can be stated in words as

H_0 : The reduced model is adequate
H_a : The full model is needed

These can be stated symbolically as

$$H_0 : \quad \beta_1 = 0$$
$$H_a : \quad \beta_1 \neq 0$$

To carry out the test, we start by fitting both models to the data using `lm()`:

```
my.full.reg <- lm(y ~ x)
```

```
my.reduced.reg <- lm(y ~ 1)
```

Then we pass the two *lm* objects to `anova()`:

```
anova(my.reduced.reg, my.full.reg)

## Analysis of Variance Table
##
## Model 1: y ~ 1
## Model 2: y ~ x
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1      9 33.461
## 2      8 28.541  1      4.92 1.3791  0.274
```

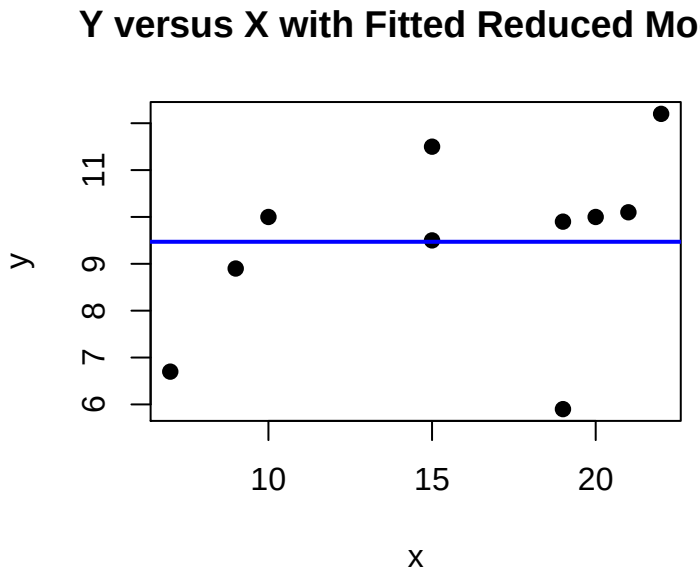
In the output above, the line labeled 1 corresponds to the **reduced model** and the line labeled 2 the **full model**. We can conclude that

- The ***error sum of squares*** for the ***reduced model*** is $\mathbf{SSE(R) = 33.461}$ with $\boldsymbol{n - 1 = 9}$ ***degrees of freedom***.
- The ***error sum of squares*** for the ***full model*** is $\mathbf{SSE(F) = 28.541}$ with $\boldsymbol{n - 2 = 8}$ ***degrees of freedom***.
- The difference is $\mathbf{SSE(R) - SSE(F) = 33.461 - 28.541 = 4.92}$ with **9 - 8 = 1** ***degree of freedom***.
- The ***general linear F test statistic*** is

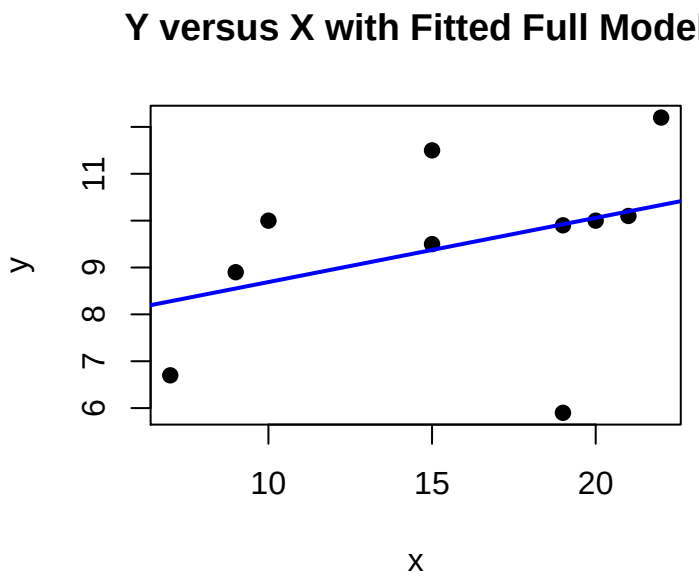
$$\boldsymbol{F} = \frac{\mathbf{(SSE(R) - SSE(F))\,/1}}{\mathbf{SSE(F)/(}\boldsymbol{n}\mathbf{-2)}} = \mathbf{1.379}.$$

- The **p-value** (from the F distribution with **1** and **8** degrees of freedom) is **0.274**. Thus we fail to reject H_0, and conclude that the model with just an intercept is adequate. The two fitted models are shown with the data in the scatterplots below.

```
plot(x, y, pch = 19, main = "Y versus X with Fitted Reduced Model")
abline(my.reduced.reg, lwd = 2, col = "blue")
```




```
plot(x, y, pch = 19, main = "Y versus X with Fitted Full Model")
abline(my.full.reg, lwd = 2, col = "blue")
```



2 Lack of Fit Test

- To carry out an ***F test for lack of fit***, we view the test as a *general linear F test* with ***full model***

$$Y_i = \mu_i + \epsilon_i$$

and ***reduced model***

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Note that the reduced model is a **special case** of the full model for which the μ_i's are constrained to lie on a line $\beta_0 + \beta_1 X$.

To carry out the test in R, we fit both models using `lm()`, and then simultaneously pass the resulting *lm* objects to the `anova()` function.

In the calls to `lm()`, these models are specified via the *formulas* as `y ~ factor(x)` (**full model**) and `y ~ x` (**reduced model**).

Specifying the predictor as `factor(x)` when fitting the full model tells R that each distinct value of `x` should be treated as a level of a factor.

- As an example, consider the data in the vectors `x` and `y`:

```
x <- c(2, 2, 2, 4, 4, 6, 6, 6, 7, 8, 9, 9)
y <- c(6, 4, 8, 11, 8, 11, 12, 10, 9, 8, 5, 6)
```

The ***lack of fit F test*** is the same as a *general linear F test* of

$$\begin{aligned} H_0: &\quad \text{The reduced model is adequate} \\ H_a: &\quad \text{The full model is needed} \end{aligned}$$

As before, to perform the test, we first use `lm()` to fit the **full** and **reduced models**:

```
my.full.reg <- lm(y ~ factor(x))
```

```
my.reduced.reg <- lm(y ~ x)
```

Then we pass the two *lm* objects to `anova()`:

```
anova(my.reduced.reg, my.full.reg)

## Analysis of Variance Table
##
## Model 1: y ~ x
## Model 2: y ~ factor(x)
##   Res.Df   RSS Df Sum of Sq     F  Pr(>F)
## 1     10 71.31
## 2      6 15.00  4     56.31 5.631 0.03136
```

In the output above, line `1` corresponds to the **reduced model** and line `2` the **full model**. We can conclude that

- The ***sum of squares for pure error*** is $\mathbf{SSPE = 15}$ (which is also SSE(F)) with $\boldsymbol{n - c = 6}$ ***degrees of freedom*** (where c is the number of distinct `x` values).
- The ***sum of squares for lack of fit*** is $\mathbf{SSLF = SSE - SSPE = 71.31 - 15 = 56.31}$ (which is also SSE(R) − SSE(F)) with $\boldsymbol{c - 2 = 4}$ ***degree of freedom***.
- The ***lack of F test statistic*** is

$$\boldsymbol{F = \frac{(SSLF)/(c-2)}{SSPE/(n-c)} = 5.631}.$$

- The **p-value** (from the F distribution with **4** and **6** degrees of freedom) is **0.0314**. Thus we reject H_0, and conclude that the linear model *doesn't* fit the data.

The fitted linear (reduced) model is shown below with the data in a scatterplot.

```
plot(x, y, pch = 19, main = "Y versus X with Fitted Reduced Model")
abline(my.reduced.reg, lwd = 2, col = "blue")
```