STAT 425

Lack of Fit Testing

Gaussian Model Assumptions

Recall our idealized modeling assumptions, which can be summarized concisely as:

$$\mathbf{y} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{I})$$

Under these assumptions:

$$\hat{\beta} = (\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{t}}\mathbf{y} \sim N_p(\beta, \sigma^2(\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-1}),$$

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} \sim N_n(\mathbf{X}\beta, \sigma^2\mathbf{H}), \qquad \mathbf{H} = \mathbf{X}(\mathbf{X}^{\mathbf{t}}\mathbf{X})^{-1}\mathbf{X}^{\mathbf{t}},$$

and, independently,

$$\hat{\sigma}^2 = \frac{RSS}{n-p} = \frac{\|\mathbf{y} - \hat{\mathbf{y}}\|^2}{n-p} \sim \sigma^2\,\frac{\chi^2_{n-p}}{n-p}.$$

Testing for Lack of Fit

How can we test whether the model $\mathbf{X}\beta$ fits the data?

- Intuition: If the model is correct then $\hat{\sigma}^2$ is an unbiased estimate of σ^2. In the very special case where we knew σ^2, we could construct a test based on the ratio $\hat{\sigma}^2/\sigma^2$, a measure of lack-of-fit.
- If σ^2 is unknown and we have some **replication** in the design (repeat rows of $\mathbf{X}$), then we'll see how to devise an F test for lack of fit.

Lack of Fit test when σ^2 is known

- In this case we want to test the hypothesis:

$$H_0 : \text{There is no lack of fit,} \quad \text{vs.} \quad H_a : \text{There is lack of fit}$$

- We use the test statistic:

$$\frac{\hat{\sigma}^2}{\sigma^2} = \frac{RSS/(n-p)}{\sigma^2} \sim \frac{\chi^2_{n-p}}{n-p}$$

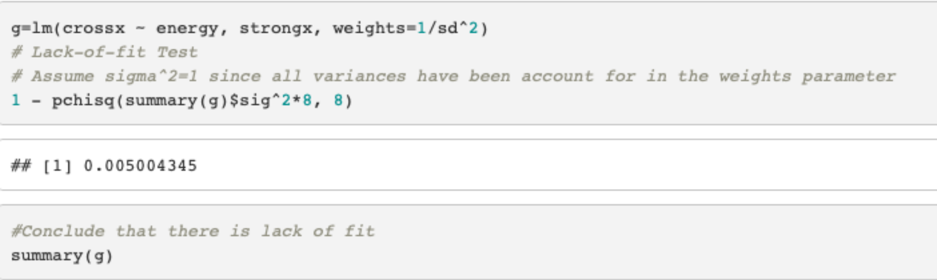
  Lack of fit means the error variance is large related to the value of σ^2, i.e., the test statistic is large.

- Conclude that there is lack of fit (i.e. Reject H_0), if:

$$(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \geq \chi^2_{n-p}(1-\alpha)$$

Example: Lack of fit test assuming σ^2 is known

In this example, all individual variances have been accounted for by using the *weights* parameter, so we take $\sigma^2 = 1$. Then our test is based on $(n-p)\hat{\sigma}^2 \sim \chi^2_{n-p}$ under H_0.



```
g=lm(crossx ~ energy, strongx, weights=1/sd^2)
# Lack-of-fit Test
# Assume sigma^2=1 since all variances have been account for in the weights parameter
1 - pchisq(summary(g)$sig^2*8, 8)
```

```
## [1] 0.005004345
```

```
#Conclude that there is lack of fit
summary(g)
```

Since the p-value < 0.05 we reject the null hypothesis and conclude there is a lack of fit. This might be the case even with a high value of R^2.

Lack of Fit test when σ^2 is unknown

- If σ^2 is unknown, a general approach is to compare an estimate of σ^2 based on a much bigger/general model.
- If we can derive the distribution (under H_0) of $\hat{\sigma}^2_{LinearModel}/\hat{\sigma}^2_{BigModel}$, then we reduce this problem to a two model comparison test problem.
- The null hypothesis is the current model:

$$H_0 : E(y_i) = \mathbf{x}_i^\top \boldsymbol{\beta}, \quad i = 1, 2, \ldots, n, \quad \text{for some vector } \beta$$

- The more general model is assumed under the alternative hypothesis:

$$H_a : E(y_i) = f(\mathbf{x}_i), \quad i = 1, 2, \ldots, n, \ \text{for some function } f$$

Lack of Fit test when σ^2 is unknown

Can we estimate σ^2 for the big model in H_a?

- The answer is yes if there is some replication in the data, i.e., there are multiple observations (replicates) for some (at least) of the same $\mathbf{x}_i$ values
- Schematically we can represent these replicates as:

$$(\mathbf{x}_i, y_{i1}, y_{i2}, \ldots, y_{in_i}), \quad i = 1 : m, \quad n = \sum_i n_i$$

Lack of Fit test

Under the null hypothesis H_0:

- $y_{ij} = \mathbf{x}_i^\top \beta + e_{ij}$, some β, $e_{ij} \sim$ iid $N(0, \sigma^2)$
- RSS_0 with $df = n - p$

Under the alternative big-model hypothesis H_a:

- $y_{ij} = f(\mathbf{x}_i) + e_{ij}$, some function f, $e_{ij} \sim$ iid $N(0, \sigma^2)$
- RSS_a with $df = n - m = \sum_i (n_i - 1)$, where

$$RSS_a = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_{ij} - \bar{y}_{i.})^2$$

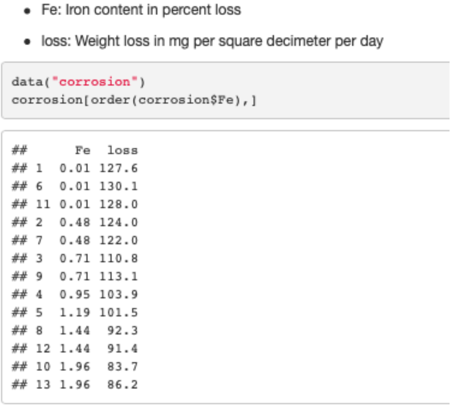
All of the degrees of freedom for RSS_a come from the replications.

Therefore, with replication we can do an F test for lack of fit:

$$F = \frac{(RSS_0 - RSS_a)/(m-p)}{RSS_a/(n-m)} \sim F_{m-p,n-m}$$

Example: Corrosion Data Set

For a given value of iron content (x_i), we have several observations of weight loss (y_{ij})



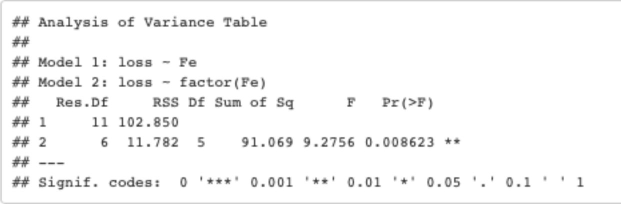
- Fe: Iron content in percent loss
- loss: Weight loss in mg per square decimeter per day

```
data("corrosion")
corrosion[order(corrosion$Fe),]
```

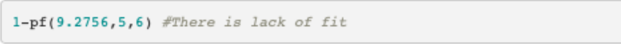
```
##      Fe  loss
## 1  0.01 127.6
## 6  0.01 130.1
## 11 0.01 128.0
## 2  0.48 124.0
## 7  0.48 122.0
## 3  0.71 110.8
## 9  0.71 113.1
## 4  0.95 103.9
## 5  1.19 101.5
## 8  1.44  92.3
## 12 1.44  91.4
## 10 1.96  83.7
## 13 1.96  86.2
```

Model Comparison

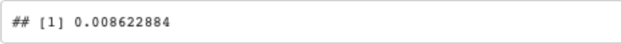
The model under H_0 is compared with a more general model in where each level of X is considered as a factor.



```
## Analysis of Variance Table
##
## Model 1: loss ~ Fe
## Model 2: loss ~ factor(Fe)
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     11 102.850
## 2      6  11.782  5    91.069 9.2756 0.008623 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



```
1-pf(9.2756,5,6) #There is lack of fit
```



```
## [1] 0.008622884
```

Since the p-value < 0.5 we have Lack of Fit. The model under H_0 is not adequate for this data set.