Hypothesis Testing in Regression Models

Recall the regression model:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon.$$

Test for significance of regression:

- $H_0 : \beta_1 = \beta_2 = \cdots = \beta_k = 0$;
- $H_1 : \beta_j \neq 0$ for at least one $j \neq 0$.
- Note that under H_0, β_0 is still non-zero:

$$H_0 : y = \beta_0 + \epsilon.$$

The ANOVA table:

Source	SS	df	MS	F_0
Regression	SS_R	k	MS_R	MS_R/MS_E
Error	SS_E	$n-k-1$	MS_E	
Total	SS_T	$n-1$		

Here, as before, SS_E is the *residual* sum of squares,

$$SS_E = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2 = \mathbf{e}'\mathbf{e} = \mathbf{y}'\mathbf{y} - \hat{\boldsymbol{\beta}}'\mathbf{X}'\mathbf{y}.$$

Also SS_T is the *total* sum of squares,

$$SS_T = \sum_{i=1}^{n} (y_i - \bar{y})^2,$$

and the *regression* sum of squares is

$$SS_R = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2 = SS_T - SS_E.$$

Test statistic:

$$F_0 = \frac{SS_R/k}{SS_E/(n-k-1)} = \frac{SS_R/k}{SS_E/(n-p)} = \frac{MS_R}{MS_E}.$$

Assuming ϵs are NID$(0, \sigma^2)$, reject H_0 if $F_0 > F_{\alpha,k,n-p}$.

Note: under H_0,

$$y = \beta_0 + \epsilon,$$

so y has a non-zero mean, but no dependence on *any* of the regressors.

F_0 is calculated and reported by all packages.

Also calculated: the *coefficient of multiple determination*

$$R^2 = \frac{SS_R}{SS_T} = 1 - \frac{SS_E}{SS_T}.$$

Note: R^2 always increases if you add a new regressor to a model, so high R^2 may result from including too many regressors.

Adjusted R^2

$$R^2_{\text{adj}} = 1 - \frac{SS_E/(n-p)}{SS_T/(n-1)}$$

allows for the number of regressors, and may either increase or decrease.

Example

Recall R output from viscosity example:

```
summary(viscosityLm)
```

Output

```
Call:
lm(formula = Viscosity ~ Temperature + CatalystFeedRate, data =
viscosity)

Residuals:
     Min       1Q   Median       3Q      Max
-21.4972 -13.1978  -0.4736  10.5558  25.4299
.
.
.
Multiple R-Squared: 0.927,      Adjusted R-squared: 0.9157
F-statistic:  82.5 on 2 and 13 DF,  p-value: 4.1e-08
```

Test for an individual coefficient

$H_0 : \beta_j = 0;$

$H_1 : \beta_j \neq 0;$

Test statistic:

$$t_0 = \frac{\hat{\beta}_j}{\text{Standard Error of } \hat{\beta}_j} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 C_{j,j}}},$$

where $C_{j,j}$ is the j^{th} diagonal entry in $(\mathbf{X}'\mathbf{X})^{-1}$.

Reject H_0 if $|t_0| > t_{\alpha/2,n-p}$.

Example

Again, recall R output from viscosity example:

```
summary(viscosityLm)
```

Output

```
...
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      1566.0778    61.5918   25.43 1.80e-12 ***
Temperature         7.6213     0.6184   12.32 1.52e-08 ***
CatalystFeedRate    8.5848     2.4387    3.52  0.00376 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
...
```

Test for a group of coefficients

"Extra Sum of Squares Method": suppose we want to test the significance of *part* of the model.

Recall the matrix form of the model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}.$$

Partition the design matrix and the parameters as

$$\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2], \quad \boldsymbol{\beta} = \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix}.$$

The *full* model is now

$$\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta}_1 + \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

with regression sum of squares $\text{SS}_R(\boldsymbol{\beta})$.

The null hypothesis $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$ implies the *reduced* model:

$$\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta}_2 + \boldsymbol{\epsilon},$$

with regression sum of squares $\text{SS}_R(\boldsymbol{\beta}_2)$.

The sum of squares due to $\boldsymbol{\beta}_1$ given $\boldsymbol{\beta}_2$ is defined to be

$$\text{SS}_R(\boldsymbol{\beta}_1|\boldsymbol{\beta}_2) = \text{SS}_R(\boldsymbol{\beta}) - \text{SS}_R(\boldsymbol{\beta}_2).$$

To test $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$, the test statistic is

$$F_0 = \frac{\mathrm{SS}_R\left(\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2\right)/r}{\mathrm{MS}_E}$$

where r is the number of coefficients being tested.

Reject H_0 if $F_0 > F_{\alpha,r,n-p}$.

Calculate $\mathrm{SS}_R\left(\boldsymbol{\beta}_1 | \boldsymbol{\beta}_2\right)$ either:

- by fitting the full and reduced models separately;
- by fitting the full model sequentially, with $\mathbf{X}_1$ fitted *after* $\mathbf{X}_2$; in R, the `aov()` method does this.

Example

The viscosity example:

```
summary(aov(Viscosity ~ CatalystFeedRate + Temperature, viscosity))
```

Output

```
                 Df Sum Sq Mean Sq F value    Pr(>F)
CatalystFeedRate  1   3516    3516  13.138  0.003083 **
Temperature       1  40641   40641 151.871 1.518e-08 ***
Residuals        13   3479     268
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

The "Sum Sq" for `CatalystFeedRate` is $SS_R(\texttt{CatalystFeedRate})$, and the "Sum Sq" for `Temperature` is $SS_R(\texttt{Temperature}|\texttt{CatalystFeedRate})$.

The F-statistic for testing `Temperature` given `CatalystFeedRate` has 1 degree of freedom; it is just the square of the t-statistic from the earlier output.

Testing a quadratic model against a linear model

```
summary(aov(Viscosity ~ Temperature + CatalystFeedRate
              + I(Temperature^2) + I(CatalystFeedRate^2)
              + I(CatalystFeedRate * Temperature), viscosity))
```

Output

```
                                 Df Sum Sq Mean Sq  F value    Pr(>F)
Temperature                       1  40841   40841 148.3362 2.541e-07 ***
CatalystFeedRate                  1   3316    3316  12.0448  0.006015 **
I(Temperature^2)                  1    399     399   1.4495  0.256330
I(CatalystFeedRate^2)             1     24      24   0.0874  0.773558
I(CatalystFeedRate * Temperature) 1    302     302   1.0985  0.319273
Residuals                        10   2753     275
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

$F_0 = \frac{(399+24+302)/3}{2753/10} = 0.88$, df $= 3, 10$; P $= 0.48$; do not reject H_0: model is linear.

Confidence Intervals

To interpret the regression equation, note that β_j measures the effect on the response y of increasing x_j by 1 unit; it is in units (units of y / units of x_j).

Again, assuming ϵs are NID$(0, \sigma^2)$, a $100(1-\alpha)\%$ confidence interval for β_j is

$$\hat{\beta}_j \pm t_{\alpha/2,n-p} \times \text{se}\left(\hat{\beta}_j\right) = \hat{\beta}_j \pm t_{\alpha/2,n-p}\sqrt{\hat{\sigma}^2 C_{j,j}}.$$

Predicting the mean response

A regression equation may also be used to predict the mean response under some new experimental (or operational) conditions.

Mean response at $\mathbf{x}_0 = [1, x_{0,1}, x_{0,2}, \ldots, x_{0,k}]'$ is $\hat{y}(\mathbf{x}_0) = \mathbf{x}_0'\hat{\boldsymbol{\beta}}$ with standard error

$$\text{se}[\hat{y}(\mathbf{x}_0)] = \sqrt{\hat{\sigma}^2 \mathbf{x}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_0}.$$

and $100(1-\alpha)\%$ confidence interval

$$\hat{y}(\mathbf{x}_0) \pm t_{\alpha/2,n-p} \times \text{se}[\hat{y}(\mathbf{x}_0)].$$

To compute se $[\hat{y}(\mathbf{x}_0)]$, you need the standard errors of the estimated coefficients, which are given in the usual table of estimates.

You also need their correlations, which are not part of the usual output, but can be extracted.

Most software will compute se $[\hat{y}(\mathbf{x}_0)]$ for you.

In R, use the `predict()` method to estimate the mean response, with the option `se.fit = TRUE`; e.g., to estimate the expected viscosity for a temperature of 90°C and catalyst feed rate 10lb/h:

```
predict(viscosityLm,
        newdata = data.frame(Temperature = 90, CatalystFeedRate = 10),
        se.fit = TRUE, interval = "confidence")
```

Output

```
$fit
       fit      lwr      upr
1 2337.842 2328.786 2346.899

$se.fit
[1] 4.192114

$df
[1] 13

$residual.scale
[1] 16.35860
```