PSTAT 126

Regression Analysis

Laura Baracaldo & Rodrigo Targino

Lecture 6
Multiple Linear Regression

Multiple Linear Regression Models (MLR)

Consider the linear regression model with p predictors:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip} + \epsilon_i$$

For each $i = 1, \ldots, n$ we assume $\epsilon_i \overset{i.i.d}{\sim} N(0, \sigma^2)$. Thus $y_i \overset{ind}{\sim} N(\mu_i, \sigma^2)$, with $\mu_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}$.

Multiple Linear Regression Models (MLR)

Matrix Representation: By stacking up all the observations, we can obtain the matrix representation of the MLR model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \dots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

This implies $\boldsymbol{y} \sim N_n(\mathbf{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}_n)$.

Estimation

LS and ML Estimation of β : Provided $\boldsymbol{X}^T\boldsymbol{X}$ is non-singular:

$$\hat{\boldsymbol{\beta}}_{LSE} = \hat{\boldsymbol{\beta}}_{MLE} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$$

So that $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}} = \boldsymbol{H}\boldsymbol{y}$, with $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ the *Projection Matrix* or *Hat Matrix*, which corresponds to the orthogonal projection onto the column space of $\boldsymbol{X}$.

Estimation of σ^2:

$$\hat{\sigma}^2 = \frac{\hat{\boldsymbol{\epsilon}}^T\hat{\boldsymbol{\epsilon}}}{n-p^*} = \frac{SSR}{n-p^*}$$

Algebraic Properties of the Projection Matrix H

- $\boldsymbol{H}$ is symmetric: $\boldsymbol{H}^T = \boldsymbol{H}$.
- $\boldsymbol{H}$ is idempotent: $\boldsymbol{H}^2 = \boldsymbol{H}$.
- If $\boldsymbol{X}$ is a $n \times p^*$ matrix with $rank(\boldsymbol{X}) = p^*$, then $rank(\boldsymbol{H}) = p^*$.
- $\boldsymbol{X}$ is invariant under $\boldsymbol{H}$: $\boldsymbol{H}\boldsymbol{X} = \boldsymbol{X}$
- The eigenvalues of $\boldsymbol{H}$ consist of p^* ones, and $n - p^*$ zeros.

Reminders:

- The rank of a matrix is the number of its linearly independent columns.
- If $\boldsymbol{A}\boldsymbol{z} = \lambda \boldsymbol{z}$, then $\boldsymbol{z}$ is an eigenvector of $\boldsymbol{A}$ with corresponding eigenvalue λ.

Algebraic Properties of the Projection Matrix H

Note that the residuals of the MLR model can be written as:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{M}\boldsymbol{y}$$

- $\boldsymbol{M} = (\boldsymbol{I} - \boldsymbol{H})$ is symmetric: $\boldsymbol{M}^T = \boldsymbol{M}$.
- $\boldsymbol{M}$ is idempotent: $\boldsymbol{M}^2 = \boldsymbol{M}$
- $rank(\boldsymbol{M}) = n - p^*$.
- The eigenvalues of $\boldsymbol{M}$ consist of $n - p^*$ ones, and p^* zeros.
- $\hat{\boldsymbol{\epsilon}}$ and $\boldsymbol{X}$ are orthogonal: $\hat{\boldsymbol{\epsilon}}^T\boldsymbol{X} = \boldsymbol{0}$. Under normality this is equivalent to statistical independence.
- $\boldsymbol{M}$ and $\boldsymbol{H}$ are orthogonal: $\boldsymbol{M}\boldsymbol{H} = \boldsymbol{0}$. This implies $\hat{\boldsymbol{\epsilon}}$ and $\hat{\boldsymbol{\beta}}$ are independent (Under normality).

Inference on β and σ^2

- **Distribution of $\hat{\boldsymbol{\beta}}$:**

$$
\begin{aligned}
E(\hat{\boldsymbol{\beta}}) &= E\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\right] \\
&= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T E(\boldsymbol{y}) = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{\beta} \\
V(\hat{\boldsymbol{\beta}}) &= V\left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\right] \\
&= \left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right] V(\boldsymbol{y}) \left[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\right]^T = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2
\end{aligned}
$$

$$
Var(\hat{\beta}_j) = \sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}_{jj} \qquad Cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2\left[\boldsymbol{X}^T\boldsymbol{X}\right]^{-1}_{ij}
$$

$$
\Rightarrow \hat{\boldsymbol{\beta}} \sim N_{p^*}(\boldsymbol{\beta}, (\boldsymbol{X}^T\boldsymbol{X})^{-1}\sigma^2).
$$

Gauss-Markov Theorem: $\hat{\beta}$ is the best linear unbiased estimate (BLUE) of $\boldsymbol{\beta}$.

Inference on β and σ^2

- **Distribution of $\hat{\sigma}^2$.** Note that:

$$\hat{\boldsymbol{\epsilon}} = \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}) = \boldsymbol{M}\boldsymbol{\epsilon}$$

Lemma

If A is a symmetric and idempotent $n \times n$ real matrix and $Z \sim N(0, I_n)$ is a random vector of n independent standard normal variables, then $Z^T A Z$ has chi-squared(r) distribution, r being the trace of A.

From the lemma we can prove that $\frac{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}{\sigma^2} \sim \chi^2_{(n-p^*)}$ since:

$$\frac{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}{\sigma^2} = \frac{\boldsymbol{\epsilon}^T \boldsymbol{M} \boldsymbol{\epsilon}}{\sigma^2}$$

Species Example

```
Beta.hat<- solve(crossprod(X))%*%(t(X)%*%y);t(Beta.hat)
```

```
##      Intercept        Area Elevation      Scruz    Adjacent
## [1,]  7.075377 -0.02397793 0.3195734 -0.2393552 -0.07484842
```

```
sigma.hat <- sqrt(sum(fit1$residuals^2)/(fit1$df.residual)); sigma.hat
```

```
## [1] 59.74333
```

```
XtX.inverse <- solve(crossprod(X)); XtX.inverse
```

```
##               Intercept          Area     Elevation         Scruz      Adjacent
## Intercept  9.849524e-02  3.879513e-05 -1.589754e-04 -4.059337e-04  2.421334e-05
## Area       3.879513e-05  1.296222e-07 -2.439943e-07  1.163061e-07  4.003452e-08
## Elevation -1.589754e-04 -2.439943e-07  7.323822e-07 -1.457015e-07 -1.471043e-07
## Scruz     -4.059337e-04  1.163061e-07 -1.457015e-07  7.594187e-06 -1.368476e-08
## Adjacent   2.421334e-05  4.003452e-08 -1.471043e-07 -1.368476e-08  7.747375e-08
```

```
Beta.hat.SE <- sigma.hat*sqrt(diag(XtX.inverse)); Beta.hat.SE
```

```
##   Intercept        Area   Elevation       Scruz    Adjacent
## 18.74981844  0.02150944  0.05112795  0.16463800  0.01662902
```

Species Example

```
##
## Call:
## lm(formula = Species ~ Area + Elevation + +Scruz + Adjacent,
##     data = gala)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -111.637  -34.930   -7.864   33.432  182.524
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.07538   18.74982   0.377 0.709093
## Area        -0.02398    0.02151  -1.115 0.275554
## Elevation    0.31957    0.05113   6.250 1.54e-06 ***
## Scruz       -0.23936    0.16464  -1.454 0.158434
## Adjacent    -0.07485    0.01663  -4.501 0.000136 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 59.74 on 25 degrees of freedom
## Multiple R-squared:  0.7658, Adjusted R-squared:  0.7284
## F-statistic: 20.44 on 4 and 25 DF,  p-value: 1.39e-07
```

Model Performance

1. Goodness of fit.
2. Estimators Quality.
3. Usefulness of predictors.
4. Prediction accuracy.

Goodness of fit

- **Residual Standard Error (RSE)** $\hat{\sigma}$

$$\hat{\sigma} = \sqrt{\frac{\hat{\boldsymbol{\epsilon}}^T \hat{\boldsymbol{\epsilon}}}{n - p^*}}$$

The smaller $\hat{\sigma}$ the better. But, how small?

Determination coefficient R^2

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - \frac{SSR}{SST} = 1 - \frac{\boldsymbol{y}^T \boldsymbol{M} \boldsymbol{y}}{\boldsymbol{y}^T \boldsymbol{M}_1 \boldsymbol{y}}$$

Where $\boldsymbol{M}_1 = (\boldsymbol{I} - \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T)$

It can be proved that:

$$R^2 = cor^2(\boldsymbol{y}, \hat{\boldsymbol{y}})$$

Hypothesis tests for a set of predictors

We want to assess the usefulness of the predictors in the prediction of the response. For instance, let's suppose we want to test whether there is at least one predictor useful in the prediction of the response:

$H_0 : \beta_1 = \ldots = \beta_p = 0 \qquad H_1 : \beta_j \neq 0$ for at least one $j \in 1, \ldots, p$.

This can be tested by comparing the full model: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ to the null model: $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$.

General Hypothesis Test

Consider two models M_1 and M_2, such that $c(M_1) \subset c(M_2)$. Where $c(M)$ denotes the column space spanned by the predictors matrix of model M. This means that predictors included in model M_1 is a subset of predictors included in M_2. How to decide which model is better?

- If there is not much difference in how well both models fit the data, we go with the *smaller* model M_1.
- If the *larger* model M_2 fits better because of the additional variables, we go with model M_2.

General Hypothesis Test

We have learned that we can evaluate the performance of a model by calculating the SSR. Small values of SSR will imply a better fit of the model. For $c(M_1) \subset c(M_2)$ it is always true that:

$$SSR_{M_2} \leq SSR_{M_1}$$

Let's consider the difference $SSR_{M_1} - SSR_{M_2}$. If this difference is small, there is not substantial difference between the fit of the larger model and the smaller model. Thus, we go with M_1. If this difference is significant, the fit of the smaller model is substantially worse than the fit of the large model. Therefore, we go with model M_2.

F-Test

We want to test the hypothesis:

$$H_0 : M_1 \text{ and } M_2 \text{ are equivalent.} \qquad H_1 : M_2 \text{ fits better.}$$

We consider the F-Statistic:

$$\begin{aligned} F &= \frac{(SSR_{M_1} - SSR_{M_2})/(df_{M_1} - df_{M_2})}{SSR_{M_2}/df_{M_2}} \\ &= \frac{\boldsymbol{y}^T(\boldsymbol{H}_2 - \boldsymbol{H}_1)\boldsymbol{y}/(df_{M_1} - df_{M_2})}{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_2)\boldsymbol{y}/df_{M_2}} \sim F_{(df_{M_1} - df_{M_2}, df_{M_2})} \end{aligned}$$

With $\boldsymbol{H}_1$ and $\boldsymbol{H}_2$ the projection matrices for models M_1 and M_2 respectively. $F \sim F_{(df_{M_1} - df_{M_2}, df_{M_2})}$, since $\frac{\boldsymbol{y}^T(\boldsymbol{H}_2 - \boldsymbol{H}_1)\boldsymbol{y}}{\sigma^2} \sim \chi^2_{df_{M_1} - df_{M_2}}$ and $\frac{\boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_2)\boldsymbol{y}}{\sigma^2} \sim \chi^2_{df_{M_2}}$

Case 1: Global F-Test

The global F-test:

$$H_0 : \beta_1 = \ldots = \beta_p = 0 \qquad H_1 : \beta_j \neq 0 \text{ for at least one } j \in 1, \ldots, p$$

Which is equivalent to:

$$H_0 : M_0 \text{ and } M_F \text{ are equivalent.} \qquad H_1 : M_F \text{ fits better.}$$

Where M_0 is the null model (with no predictors): $\boldsymbol{y} = \boldsymbol{\mu} + \boldsymbol{\epsilon}$, and M_F is the full model (with p predictors): $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$.

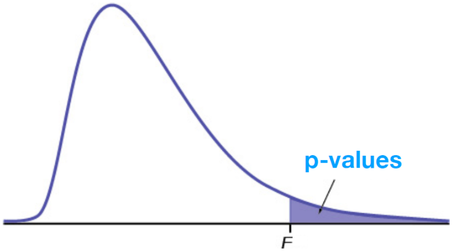
Case 1: Global F-Test - ANOVA

Total Sum of Squares:
$SSR_{M_0} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_0)\boldsymbol{y} = \sum_{i=1}^n (y_i - \bar{y})^2$, with $df = n - 1$.
Residual Sum of Squares:
$SSR_{M_F} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_F)\boldsymbol{y} = \sum_{i=1}^n (y_i - \hat{y}_i)^2$, with $df = n - p^*$.
Regression Sum of Squares:
$\Rightarrow SSR_{M_0} - SSR_{M_F} = \boldsymbol{y}^T(\boldsymbol{H}_F - \boldsymbol{H}_0)\boldsymbol{y} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$, with $df = p$.

Recall $H_0 = \boldsymbol{X}_0(\boldsymbol{X}_0^T\boldsymbol{X}_0)^{-1}\boldsymbol{X}_0^T = \mathbf{1}(\mathbf{1}^T\mathbf{1})^{-1}\mathbf{1}^T$ and
$H_F = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Case 1: Global F-Test

For a significance level α, we reject H_0 if $F > F_{(1-\alpha; df_{M_1} - df_{M_2}, df_{M_2})}$ or equivalently, if p-value $< \alpha$.
Recall: p-value: Probability of obtaining tests results as least as extreme as the observed results.



Case 2: F-test for a pair of predictors

$$H_0 : \beta_l = \beta_k = 0 \qquad H_1 : \beta_l \neq 0 \vee \beta_k \neq 0$$

Which is equivalent to:

H_0 : M_1 and M_F are equivalent. $\qquad$ H_1 : M_F fits better.

Where M_1 is the model that does not include predictors X_l nor X_k and M_F is the full model (with p predictors).

Case 2: F-test for a pair of predictors - ANOVA

$SSR_{M_1} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_1)\boldsymbol{y}$, with $df = n - (p^* - 2) = n - p + 1$.
$SSR_{M_F} = \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H}_F)\boldsymbol{y}$, with $df = n - p^* = n - p - 1$.
$\Rightarrow SSR_{M_1} - SSR_{M_F} = \boldsymbol{y}^T(\boldsymbol{H}_F - \boldsymbol{H}_1)\boldsymbol{y}$, with $df = 2$.

Recall $H_1 = \boldsymbol{X}_1(\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T$, with $\boldsymbol{X}_1$ the matrix without predictors X_l and X_k and $H_F = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$.

Species Example: $\hat{\sigma}$ and R^2

```
fit1<- lm( Species ~ Area+Elevation+  + Scruz+ Adjacent, data=gala)
RSE<- sqrt(sum(fit1$residuals^2)/fit1$df.residual);RSE ## With Formula
```

```
## [1] 59.74333
```

```
sigma(fit1) ##With lm function
```

```
## [1] 59.74333
```

```
y.hat<- fit1$fitted.values
R2<- cor(y.hat,y)^2;R2 ## With Formula
```

```
## [1] 0.7658462
```

```
summary(fit1)$r.squared  ##With lm function
```

```
## [1] 0.7658462
```

Species Example: Global F-Test

$H_0 : \beta_{Area} = \beta_{Elevation} = \beta_{Scruz} = \beta_{Adjacent} = 0$ $\quad H_1 : \beta_j \neq 0$ for at least one $j \in \{Area, Elevation, Scruz, Adjacent\}$

```
fullmodel<- lm( Species ~ Area+Elevation+   Scruz+ Adjacent, data=gala)
nullmodel <- lm(Species~1, data=gala)
anova1<-anova(nullmodel, fullmodel);anova1
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ 1
## Model 2: Species ~ Area + Elevation + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1     29 381081
## 2     25  89232  4    291850 20.442 1.39e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
pval<- 1-pf(anova1$F[2],4,25);pval
```

```
## [1] 1.389845e-07
```

Species Example: F-test for a pair of predictors

$$H_0 : \beta_{Area} = \beta_{Scruz} = 0 \qquad H_1 : \beta_{Area} \neq 0 \vee \beta_{Scruz} \neq 0$$

```
fullmodel<- lm( Species ~ Area+Elevation+ Scruz+ Adjacent, data=gala)
Model1 <- lm(Species~Elevation+ Adjacent, data=gala)
anova2<-anova(Model1, fullmodel);anova2
```

```
## Analysis of Variance Table
##
## Model 1: Species ~ Elevation + Adjacent
## Model 2: Species ~ Area + Elevation + Scruz + Adjacent
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1     27 100003
## 2     25  89232  2     10771 1.5089 0.2406
```

```
pval<- 1-pf(anova2$F[2],2,25);pval
```

```
## [1] 0.2406149
```