Linear Methods for Regression

Statistical methods for data analysis – Machine learning

Alberto Castellini
University of Verona

Introduction

- Linear regression model **assumption:**

  the **regression function** $E(Y|X)$ is **linear** in the inputs $X_1,\ldots,X_p$

- Linear models:
  - **simple**
  - **interpretable**
  - **can** sometime **outperform** fancier nonlinear models (e.g., **small training set**, low signal-to-noise ratio, sparse data)
  - can be applied to **transformations** of the input

Linear regression model

- **Input** vector: $X^T = (X_1, X_2, \ldots, X_p)$
- **Goal**: to predict a real-valued output Y
- Linear regression **model**:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$$

where:

- The β_j's are unknown parameters
- X_j are variables of possibly different type (e.g., quantitative, transformations as log or square-root, polynomials, "dummy" coding of levels, interactions between variables as $X_3 = X_1 * X_2$)
  - coding of levels: example

The model is linear in the parameters

Least squares

- Training data: $(x_1,y_1) \dots (x_N,y_N)$
  - where each $x_i = (x_{i1}, x_{i2},\dots,x_{ip})^T$ is a vector of feature measurements

- Model parameters β_j are estimated from training data

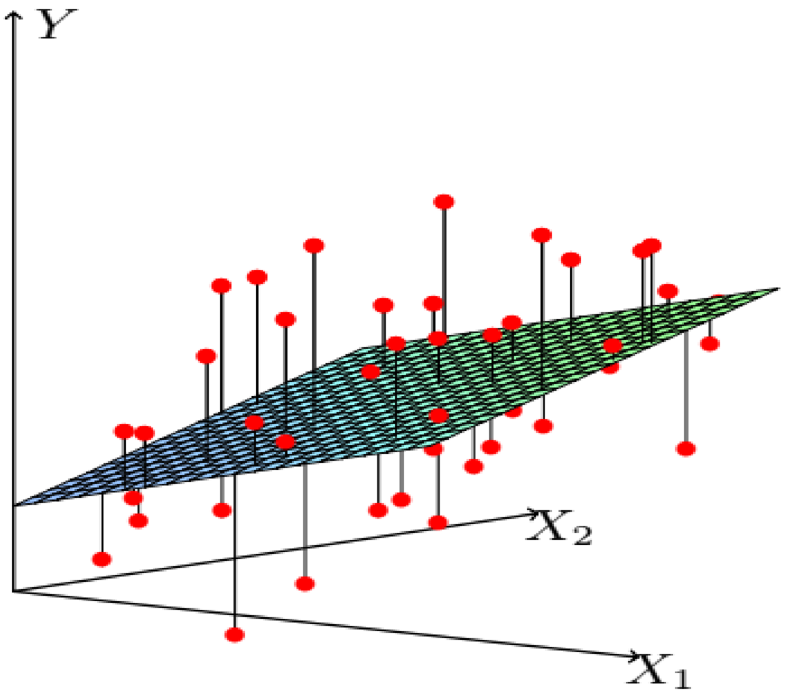
- **Least squares**: the most popular **estimation method**
  - We pick the parameters $\beta = (\beta_0, \beta_1, \dots, \beta_p)^T$ that minimize the **residual sum of squares**:

$$
\begin{aligned}
\mathrm{RSS}(\beta) &= \sum_{i=1}^{N} (y_i - f(x_i))^2 \\
&= \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2
\end{aligned}
$$

Conditions and geometrical interpretation

- The least squares criterion is **valid** if
  - the training observations (x_i, y_i) represent **independent random draws** from their population
  - The y_i's are **conditionally independent** given the inputs x_i

- **Geometry** of least-squares fitting in a 3 dimensional space



The RSS criterion measures the average **lack of fit**

Parameter estimation: minimization of the RSS

- **X** is the N x (p + 1) matrix with each row an input vector from the training set (with a 1 in the first position, the intercept)
- **y** is the N-vector of outputs in the training set

- Then the **RSS** can be written as:

$$\mathrm{RSS}(\beta) = (\mathbf{y} - \mathbf{X}\beta)^T(\mathbf{y} - \mathbf{X}\beta)$$

- This is a quadratic function in p+1 parameters. **Differentiating** w.r.t. β

$$\frac{\partial \mathrm{RSS}}{\partial \beta} = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta)$$

$$\frac{\partial^2 \mathrm{RSS}}{\partial \beta \partial \beta^T} = 2\mathbf{X}^T\mathbf{X}.$$

$\approx$ Covariance matrix

- Assuming that **X** has a **full column rank**, **X**^{T}**X** is **positive definite**, then we set the first derivative to 0

Positive eigenvalues

$$\mathbf{X}^T(\mathbf{y} - \mathbf{X}\beta) = 0$$

to obtain the unique solution

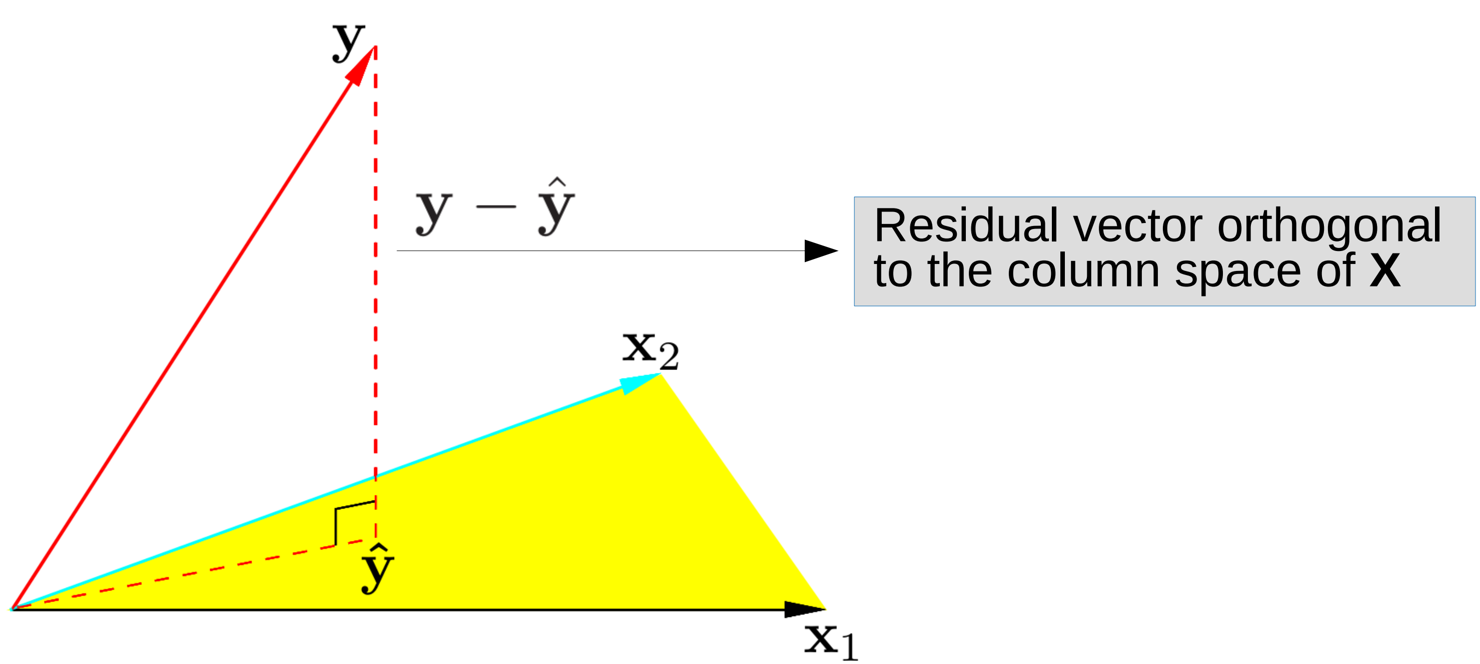
$$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

Prediction

- The **fitted values of the training inputs** are

$$\hat{\mathbf{y}} = \mathbf{X}\hat{\beta} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- The matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T$ is called **"hat" matrix** or **projection matrix**



Linearly dependent columns

- If the columns of **X** are **not linearly independent** than **X** is **not full-rank** (e.g., if $x_2=3x_1$)

- In that case X^TX is **singular**

- Then the least squares coefficients $\hat{\beta}$ are **not uniquely defined**

- There is **more than one way to express the projection** of **y** onto **X**
- A natural way to resolve the non-uniqueness is to **drop redundant columns** from **X**

- Rank deficiencies can also occur when the **number of inputs *p* exceeds the number of training cases *N*** (filtering, regularization)

Sampling properties for β

- Since independent variables X and response *y* are random variables, and $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ (linear combination of X and y) then also $\hat{\beta}$ is a **random variable**, and in particular it follows a **multivariate normal distribution**

$$\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$$

Covariance matrix

Unbiesed estimator

where

- β are the parameters of the correct model $f(X) = \beta_0 + \sum_{j=1}^{p} X_j \beta_j$
- $(\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$ is the **covariance matrix** of the least squares parameter estimate which can be derived from $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- the **variance** σ^2 is typically **estimated** by

$$\hat{\sigma}^2 = \frac{1}{N-p-1}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

Test hypothesis $\beta_j = 0$

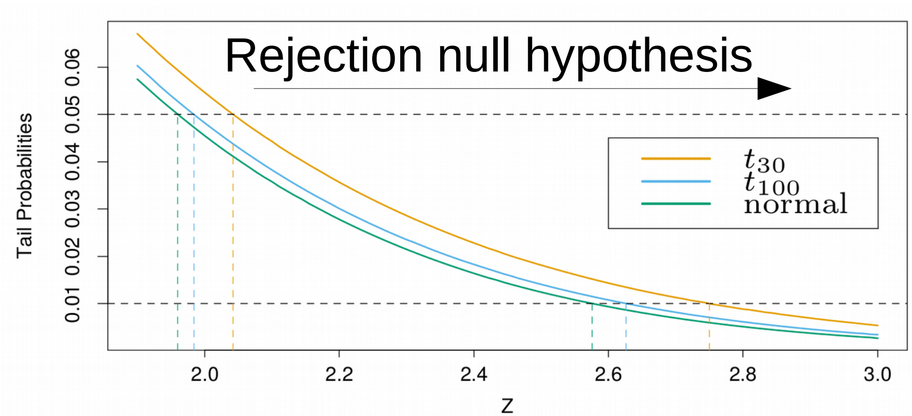
- The **significance** of a **single parameter** can be tested by the **Z-score**:

$$z_j = \frac{\hat{\beta}_j}{\hat{\sigma}\sqrt{v_j}}$$

where v_j is the j-th diagonal element of $(X^TX)^{-1}$

- Under the **null hypothesis** that $\boldsymbol{\beta_j = 0}$, z_j is distributed as t_{N-p-1} (**t-distribution** with N-p-1 degrees of freedom)

- **Large absolute value of** $\boldsymbol{z_j}$ leads to **rejection** of the null hypothesis



Test hypothesis $(\beta_{j1}, \beta_{j2}, \ldots, \beta_{jk}) = 0$

- The **significance** of a **group of coefficients** can be tested **simultaneously** by the **F statistic**

$$F = \frac{(\mathrm{RSS}_0 - \mathrm{RSS}_1)/(p_1 - p_0)}{\mathrm{RSS}_1/(N - p_1 - 1)}$$

It measures the change in RSS per additional parameter

where

  - **RSS_1** is the residual sum-of-squares for the **larger model** having **p_1** parameters
  - **RSS_0** is the residual sum-of-squares for the **smaller** model having **p_0** parameters

- Under the Gaussian assumptions and the **null hypothesis** that the **smaller model is correct** the F statistics has a $F_{p1-p0,N-p1-1}$ **distribution**

- For large N the quantiles of $F_{p1-p0,N-p1-1}$ approach those of χ^2_{p1-p0}

Confidence intervals

- By isolating β_j in $\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$ we obtain the following $1 - 2\alpha$ **confidence interval** for β_j

$$(\hat{\beta}_j - z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma}, \;\; \hat{\beta}_j + z^{(1-\alpha)} v_j^{\frac{1}{2}} \hat{\sigma})$$

where $z^{(1-\alpha)}$ is the $1 - \alpha$ **percentile of the normal distribution**

$$\begin{aligned} z^{(1-0.025)} &= 1.96, \\ z^{(1-.05)} &= 1.645, \end{aligned}$$

and $\hat{\sigma}\sqrt{v_j}$ is the **standard error** $se(\beta_j)$

- The standard practice of reporting $\beta_j + 2*se(\beta_j)$ amounts to an approximate 95% confidence interval

Exercise: Prediction on the prostate cancer dataset

Dataset

Reference:

[Stamey et al. (1989)] Stamey, T., Kabalin, J., McNeal, J., Johnstone, I., Freiha, F., Redwine, E. and Yang, N. (1989). *Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate II radical prostatectomy treated patients*, Journal of Urology 16: 1076–1083.

Type of analysis:

Correlation between the **level of prostate-specific antigen** and a number of **clinical measures** in men who were about to receive a radical prostatectomy

Dataset

	lcavol	lweight	age	lbph	svi	lcp	gleason	pgg45	lpsa	train
1	-0.579818495	2.769459	50	-1.38629436	0	-1.38629436	6	0	-0.4307829	T
2	-0.994252273	3.319626	58	-1.38629436	0	-1.38629436	6	0	-0.1625189	T
3	-0.510825624	2.691243	74	-1.38629436	0	-1.38629436	7	20	-0.1625189	T

Variables:

- ***lcavol***: log cancer volume
- ***lweight**:* log prostate weight
- ***age***: the patient age
- ***lbph***: log of the amount of benign prostatic hyperplasia
- ***svi***: seminal vesicle invasion (categorical)
- ***lcp***: log of capsular penetration
- ***gleason***: Gleason score (categorical)
- ***pgg45***: percent of Gleason scores 4 or 5

Independent (X)

- ***lpsa***: level of prostate-specific antigen

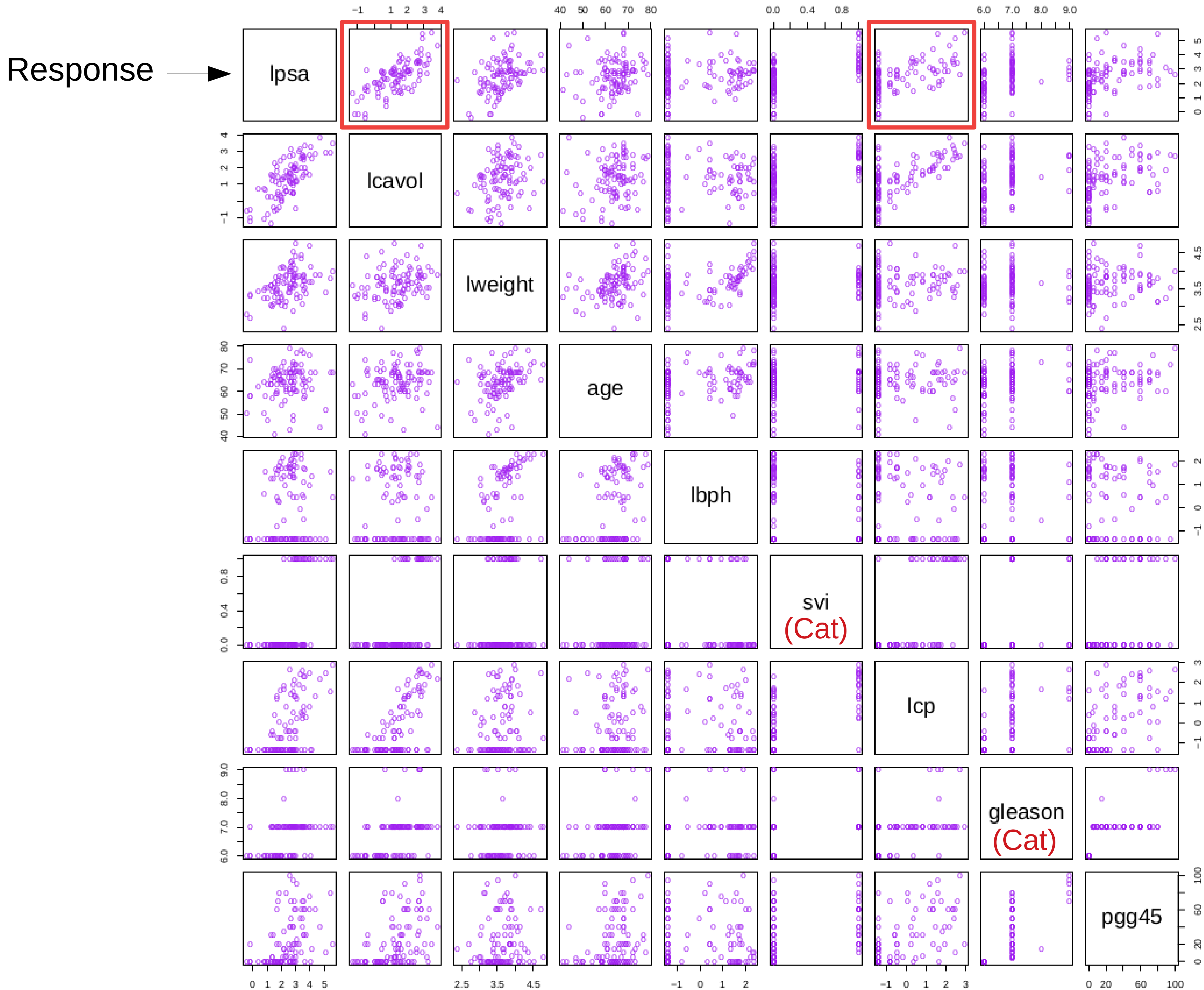
→ Dependent (Y)

Correlation analysis

Correlation matrix

	lcavol	lweight	age	lbph	svi	lcp	gleason
lweight	0.300						
age	0.286	0.317					
lbph	0.063	0.437	0.287				
svi	0.593	0.181	0.129	−0.139			
lcp	0.692	0.157	0.173	−0.089	0.671		
gleason	0.426	0.024	0.366	0.033	0.307	0.476	
pgg45	0.483	0.074	0.276	−0.030	0.481	0.663	0.757

Scatter plots



Linear regression model

- Predictor **standardization** to have unit variance
- Random **split** of the dataset
  - 67 samples in the **training set**
  - 30 samples in the **test set**
- Parameter estimation by **least squares** on the training set

Model parameters, standard error and Z score

Term	Coefficient	Std. Error	Z Score
Intercept	2.46	0.09	27.60
lcavol	0.68	0.13	5.37
lweight	0.26	0.10	2.75
age	−0.14	0.10	−1.40
lbph	0.21	0.10	2.06
svi	0.31	0.12	2.47
lcp	−0.29	0.15	−1.87
gleason	−0.02	0.15	−0.15
pgg45	0.27	0.15	1.74

Analysis of the model

Parameter significance:

- Z score greater than 2 in absolute value is approximately significant at 5% level
- ***lcavol*** shows the strongest effect (Z score 5.37)
- ***lweight*** and ***svi*** also strong (Z scores 2.75 and 2.47, respectively)
- ***lcp*** not significant once ***lcavol*** in the model (but in a model without lcavol is significant)
- Dropping all non significant terms, namely ***age***, ***lcp***, ***gleason***, ***pgg45*** we get

$$F = \frac{(32.81 - 29.43)/(9-5)}{29.43/(67-9)} = 1.67,$$

H_0: model without age, lcp, gleason, pgg4 is correct

with p-value 0.17 ($\Pr(F_{4,58} > 1.67) = 0.17$), hence it is not significant. → Not rejected

Model performance:

- **Model mean prediction error** on **test** set: 0.521
- Prediction using the mean training value of ***lpsa*** has test error of 1.057 (**base error rate**)
- The model reduces the base error rate by about 50% (**R^2=0.521/1.057=0.493**)

References

[Hastie 2009] Trevor Hastie, Robert Tibshirani, Jerome Friedman. The Elements of Statistical Learning: Data Mining, Inference, and Prediction (second edition). Springer. 2009.