Multiple Linear Regression
Numerical Predictors

University of California, San Diego
Instructor: Armin Schwartzman

Dataset: Passenger car mileage

- The data archive of the *Journal of Statistical Education*: http://jse.amstat.org/jse_data_archive.htm

  This is a great resource for real datasets.

- We consider the `04cars` dataset. (See description online.)

  For now, we focus on the following variables:

`mpg`	Highway gas consumption (miles per gallon)
`hp`	Horsepower
`wt`	Weight (pounds)
`len`	Length (inches)
`wd`	Width (inches)

- **Goal:** Predict a car's gas consumption based on these characteristics.

- **Graphics:** pairwise scatterplots and possibly individual boxplots. (Go to R)

Scatterplot highlights

- The response `mpg` is visibly correlated with predictors `hp` and `wt`, and `wd`, while somewhat less correlated with `len`.

- There are correlations among predictors, e.g., `hp` and `wt`.

- There is some curvature, e.g., in `mpg` vs `hp`.

- `mpg` versus `hp` shows a bit of a fan shape.

Linear model

- □ We fit a (simple) linear model:

$$\text{expected } \mathrm{mpg} = \beta_0 + \beta_1\,\mathrm{hp} + \beta_2\,\mathrm{wt} + \beta_3\,\mathrm{len} + \beta_4\,\mathrm{wd}$$

- □ In general, with data

$$\big\{(x_{i,1}, \dots, x_{i,p}, y_i) : i = 1, \dots, n\big\},$$

we fit the linear model:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_p x_{i,p} + \varepsilon_i$$

with the **standard assumption** that the measurement errors are i.i.d. normal with mean zero:

$$\varepsilon_1, \dots, \varepsilon_n \sim^{iid} \mathcal{N}(0, \sigma^2)$$

and independent of the predictors.

- □ In regression analysis, the inference is conditional on the observed x's. Thus, unless otherwise specified, we assume these are given.

Least squares regression

- □ The least squares criterion minimizes the error sum of squares

$$\mathrm{SSE}(b_0, b_1, \dots, b_p) = \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i,1} - \cdots - b_p x_{i,p})^2$$

- □ Define

$$(\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_p) = \underset{b_0,\dots,b_p \in \mathbb{R}}{\arg\min}\ \mathrm{SSE}(b_0, b_1, \dots, b_p).$$

- □ Under the standard assumptions, the $(\widehat{\beta}_0, \widehat{\beta}_1, \dots, \widehat{\beta}_p)$ are the maximum likelihood estimates (MLE) for $(\beta_0, \beta_1, \dots, \beta_p)$.

Fitted values, residuals and standard error

- □ The fitted (predicted) values are defined as:

$$\widehat{y}_i = \widehat{\beta}_0 + \widehat{\beta}_1 x_{i,1} + \cdots + \widehat{\beta}_p x_{i,p}$$

- □ The residuals are defined as:

$$e_i = y_i - \widehat{y}_i$$

- □ The estimate for σ^2 is the mean squared error of the fit

$$\widehat{\sigma}^2 = \frac{1}{n - (p+1)} \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 = \frac{1}{n - p - 1} \sum_{i=1}^{n} e_i^2$$

If we replace $\frac{1}{n-p-1}$ with $\frac{1}{n}$, we get the MLE for σ^2 under the standard assumptions.

Matrix interpretation

□ Let $\mathbf{X}$ be the $n \times (p+1)$ matrix with row vectors $\mathbf{x}_i = (1, x_{i,1}, \ldots, x_{i,p})$:

$$\mathbf{X} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix}$$

□ Define $\mathbf{y} = (y_1, \ldots, y_n)$, $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)$.

□ The model is:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

meaning

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ 1 & x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

Least Squares

□ For $\mathbf{b} = (b_0, b_1, \ldots, b_p)$, the error sum of squares is

$$\begin{aligned} \mathrm{SSE}(\mathbf{b}) &= \sum_{i=1}^{n} (y_i - b_0 - b_1 x_{i,1} - \cdots - b_p x_{i,p})^2 \\ &= \sum_{i=1}^{n} (y_i - \mathbf{b}^\top \mathbf{x}_i)^2 \\ &= \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2 \end{aligned}$$

□ The least squares estimate is defined as

$$\widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\mathbf{b} \in \mathbb{R}^{p+1}} \|\mathbf{y} - \mathbf{X}\mathbf{b}\|^2$$

□ If the columns of $\mathbf{X}$ are linearly independent, i.e., $\mathbf{X}$ is full rank, then

$$\widehat{\boldsymbol{\beta}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

□ Note that the estimate is linear in the response.

Residuals and the hat matrix

□ Define the hat matrix

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

□ $\mathbf{H}$ is the orthogonal projection onto

$$\operatorname{span}(\mathbf{X}) = \{b_0\mathbf{1} + b_1\mathbf{X}_1 + \cdots + b_p\mathbf{X}_p : b_0, \ldots, b_p \in \mathbb{R}\},$$

where $\mathbf{X}_j = (x_{1,j}, \ldots, x_{n,j})$ is the jth column vector of $\mathbf{X}$.

□ The fitted values may be expressed as

$$\widehat{\mathbf{y}} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{y} = \mathbf{H}\mathbf{y}$$

□ The residuals may be expressed as

$$\mathbf{e} = \mathbf{y} - \widehat{\mathbf{y}} = (\mathbf{I} - \mathbf{H})\mathbf{y}$$

Distributions

□ Suppose the standard assumptions hold, namely $\varepsilon_1, \ldots, \varepsilon_n \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$.

□ The least squares estimator $\widehat{\boldsymbol{\beta}}$ has the multivariate normal distribution with mean $\boldsymbol{\beta}$ and covariance matrix $\sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}$, i.e.,

$$\widehat{\boldsymbol{\beta}} \sim \mathcal{N}\big(\boldsymbol{\beta}, \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}\big)$$

In other words, the $\widehat{\beta}_j$'s are jointly normal and

$$\mathbb{E}\left(\widehat{\beta}_j\right) = \beta_j, \quad \operatorname{Cov}\left(\widehat{\beta}_j, \widehat{\beta}_k\right) = \sigma^2(\mathbf{X}^\top\mathbf{X})^{-1}_{jk}$$

Under the standard assumptions, the least squares estimator $\widehat{\boldsymbol{\beta}}$ is unbiased.

Distributions

□ For $\widehat{\sigma}^2$, we have

$$\widehat{\sigma}^2 \sim \frac{\sigma^2}{n-p-1}\,\chi^2_{n-p-1}$$

Thus $\widehat{\sigma}^2$ is unbiased. If we replace $\frac{1}{n-p-1}$ with $\frac{1}{n}$, we get the MLE for σ^2 under the standard assumptions, which is biased.

□ $\widehat{\boldsymbol{\beta}}$ and $\widehat{\sigma}^2$ are independent.

t-ratios

- Consequently, for any $j = 0, \ldots, p$,

$$\frac{\widehat{\beta}_j - \beta_j}{\sqrt{\widehat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}} \sim \mathcal{T}_{n-p-1}$$

  where $\mathcal{T}_k$ denotes the t-distribution with k degrees of freedom.

- For example, one can test whether $\beta_j = 0$. Indeed, letting

$$|t_j| = \frac{|\widehat{\beta}_j|}{\sqrt{\widehat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}}$$

  the p-value is given by

$$\mathbb{P}\left(|\mathcal{T}_{n-p-1}| > |t_j|\right).$$

- We can also provide confidence intervals for the coefficients:

$$\widehat{\beta}_j \quad \pm \quad T^{\alpha/2}_{n-p-1} \quad \sqrt{\widehat{\sigma}^2 (\mathbf{X}^\top \mathbf{X})^{-1}_{jj}}$$

  where T^{α}_{k} denotes the α-quantile of $\mathcal{T}_k$.

More general t-ratios

- In general, we can test for linear combination of the coefficients.

  Indeed, if $\mathbf{c} = (c_0, c_1, \ldots, c_p) \in \mathbb{R}^{p+1}$, then

$$\frac{\mathbf{c}^\top \widehat{\boldsymbol{\beta}} - \mathbf{c}^\top \boldsymbol{\beta}}{\widehat{\mathrm{SE}}(\mathbf{c}^\top \widehat{\boldsymbol{\beta}})} \sim \mathcal{T}_{n-p-1}$$

  where

$$\widehat{\mathrm{SE}}(\mathbf{c}^\top \widehat{\boldsymbol{\beta}}) = \widehat{\sigma}\sqrt{\mathbf{c}^\top (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{c}}$$

  is the (estimated) standard error of $\mathbf{c}^\top \widehat{\boldsymbol{\beta}}$.

- In particular,

$$\widehat{\boldsymbol{\beta}}^\top \mathbf{x} \quad \pm \quad T^{\alpha/2}_{n-p-1}\, \widehat{\mathrm{SE}}(\widehat{\boldsymbol{\beta}}^\top \mathbf{x})$$

  is a level-$(1-\alpha)$ confidence interval for the expected value of y at $\mathbf{x}$:

$$\mathbb{E}(y|\mathbf{x}) = \boldsymbol{\beta}^\top \mathbf{x}$$

Confidence regions

□ With the standard assumption holding, we have

$$\frac{(\widehat{\boldsymbol\beta}-\boldsymbol\beta)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\boldsymbol\beta}-\boldsymbol\beta)}{(p+1)\widehat{\sigma}^2} \sim \mathcal{F}_{p+1,n-p-1}$$

where $\mathcal{F}_{k,l}$ denotes the F-distribution with k and l degrees of freedom.

□ Based on that, the following defines a level-$(1-\alpha)$ confidence region for $\boldsymbol\beta$:

$$(\widehat{\boldsymbol\beta}-\boldsymbol\beta)^\top \mathbf{X}^\top \mathbf{X}(\widehat{\boldsymbol\beta}-\boldsymbol\beta) \le (p+1)\widehat{\sigma}^2 F^{1-\alpha}_{p+1,n-p-1}$$

where $F^{\alpha}_{k,l}$ denotes the α-quantile of $\mathcal{F}_{k,l}$.

Equivalently,

$$\|(\mathbf{X}^\top \mathbf{X})^{1/2}(\widehat{\boldsymbol\beta}-\boldsymbol\beta)\| \le [(p+1)F^{1-\alpha}_{p+1,n-p-1}]^{1/2}\,\widehat{\sigma}$$

Note that this is an ellipsoid.

Confidence bands (Scheffé's S-method)

□ Using that, and the fact that

$$\|\mathbf{u}\| = \max_{\mathbf{b}\neq 0}\frac{|\mathbf{b}^\top \mathbf{u}|}{\|\mathbf{b}\|}$$

in any Euclidean space, we get that

$$\mathbf{c}^\top\widehat{\boldsymbol\beta} \pm ((p+1)F^{1-\alpha}_{p+1,n-p-1})^{1/2}\,\widehat{\mathrm{SE}}(\mathbf{c}^\top\widehat{\boldsymbol\beta})$$

is a level-$(1-\alpha)$ confidence interval for $\mathbf{c}^\top\boldsymbol\beta$ *simultaneously* for all $\mathbf{c}\in\mathbb{R}^{p+1}$.

□ As a special case, we obtain the following confidence band

$$\widehat{\boldsymbol\beta}^\top\mathbf{x} \pm ((p+1)F^{1-\alpha}_{p+1,n-p-1})^{1/2}\,\widehat{\mathrm{SE}}(\widehat{\boldsymbol\beta}^\top\mathbf{x})$$

This means that, the standard assumption being in place, with probability $1-\alpha$,

$$|\widehat{\boldsymbol\beta}^\top\mathbf{x} - \boldsymbol\beta^\top\mathbf{x}| \le ((p+1)F^{1-\alpha}_{p+1,n-p-1})^{1/2}\,\widehat{\mathrm{SE}}(\widehat{\boldsymbol\beta}^\top\mathbf{x}), \quad \forall\mathbf{x}\in\mathbb{R}^{p+1}$$

Comparing models

- □ Say we want to test

$$\begin{aligned} H_0 : \quad & \beta_1 = \cdots = \beta_p = 0 \\ H_1 : \quad & \beta_j \neq 0, \text{ for some } j = 1, \ldots, p \end{aligned}$$

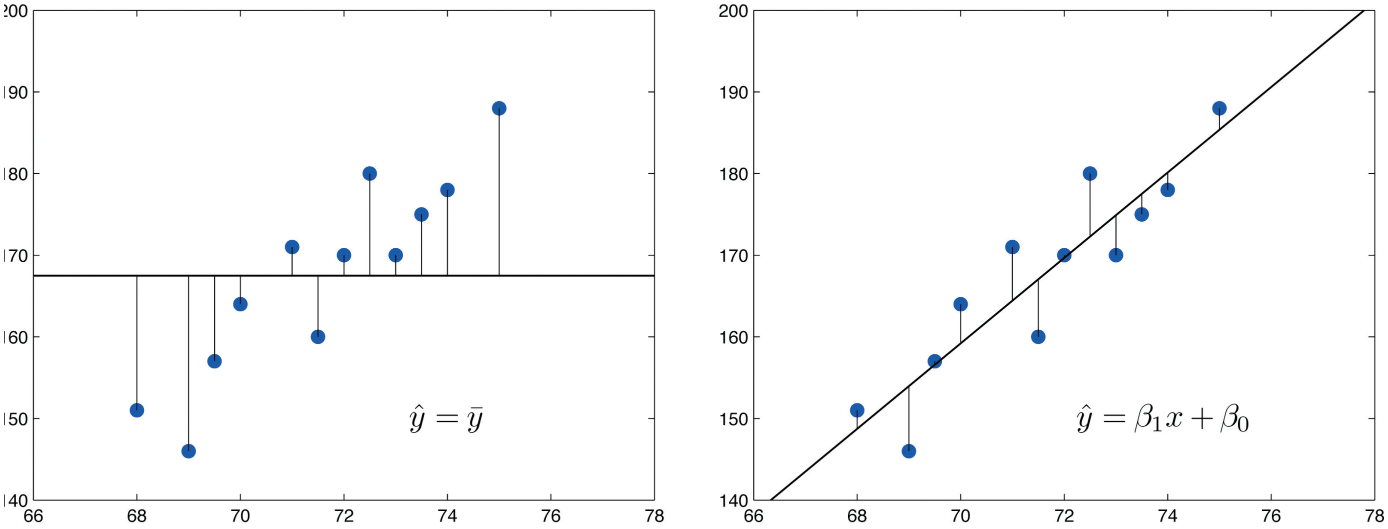
  Under the standard assumptions, we are effectively testing whether the response variable y is independent of the predictor variables $(x_1, \ldots, x_p)$.

- □ Note that the model under H_0 is a submodel of the full model:

$$\begin{aligned} \text{Null model :} \quad & y_i = \beta_0 + \varepsilon_i \\ \text{Full Model :} \quad & y_i = \beta_0 + \beta_1 \; x_{i,1} + \cdots + \beta_p \; x_{i,p} + \varepsilon_i \end{aligned}$$

Analysis of variance

The difference between fitting a constant and fitting a line:



Analysis of variance

- □ The residual sum of squares under H_0 is

$$\mathrm{SS}_Y = \sum_{i=1}^n (y_i - \overline{y})^2$$

It has $n-1$ degrees of freedom.

- □ The residual sum of squares under H_1 is

$$\mathrm{SSE} = \sum_{i=1}^n (y_i - \widehat{y}_i)^2$$

It has $n-p-1$ degrees of freedom.

- □ The sum of squares due to regression is

$$\mathrm{SS}_{\mathrm{reg}} = \mathrm{SS}_Y - \mathrm{SSE} = \sum_{i=1}^n (\overline{y} - \widehat{y}_i)^2$$

It has p degrees of freedom.

Analysis of variance

- □ The ANOVA F-test rejects for large values of

$$F = \frac{\mathrm{SS}_{\mathrm{reg}}/p}{\mathrm{SSE}/(n-p-1)}$$

- □ Under H_0,

$$F \sim \mathcal{F}_{p,n-p-1}$$

- □ In R, this F ratio is on the last line of the `summary`, together with its degrees of freedom p and $n-p-1$, and the p-value for testing H_0:

$$\mathbb{P}\left(\mathcal{F}_{p,n-p-1} > F\right)$$

(Here F is the observed value.)

Coefficient of (multiple) determination

□ Often, it is simply called (multiple) R-squared, and defined as

$$R^2 = 1 - \frac{\mathrm{SSE}}{\mathrm{SS}_Y}$$

□ Note that

$$R^2 = 1 - \frac{\mathrm{SSE}/n}{\mathrm{SS}_Y/n} = 1 - \frac{\widehat{\sigma}^2_{\mathrm{ML}}}{\widehat{\sigma}^2_{y,\mathrm{ML}}}$$

Also,

$$R = \mathrm{Cor}(\mathbf{y}, \widehat{\mathbf{y}}) = \frac{\sum_i (y_i - \overline{y})(\widehat{y}_i - \overline{y})}{\sqrt{\sum_i (y_i - \overline{y})^2 \sum_i (\widehat{y}_i - \overline{y})^2}}$$

□ The adjusted R-squared incorporates the degrees of freedom:

$$R^2_a = 1 - \frac{\mathrm{SSE}/(n-p-1)}{\mathrm{SS}_Y/(n-1)} = 1 - \frac{\widehat{\sigma}^2}{\widehat{\sigma}^2_y}$$

where $\widehat{\sigma}^2_y$ is the sample variance of $y_1, \dots, y_n$.

Coefficient of (multiple) determination

□ Both can be interpreted as the fraction of the variance of y “explained” by the variance in $\mathbf{x}$.

- ▷ The R-squared uses the MLEs of the variances.
- ▷ The adjusted R-squared uses the unbiased estimates of the variances.

Testing whether a subset of the variables are zero

□ Consider a subset of variables $J \subset \{1, \dots, p\}$.

□ We want to test

$$\begin{aligned} H_0 &: \ \forall j \in J, \ \beta_j = 0 \\ H_1 &: \ \exists j \in J, \ \beta_j \neq 0 \end{aligned}$$

Under the standard assumptions, we are effectively testing whether the response variable y is independent of the predictor variables $(x_j, j \in J)$.

□ Note that the model under H_0 is a submodel of the full model:

$$\text{Null model}: \quad y_i = \beta_0 + \sum_{j \notin J} \beta_j x_{i,j} + \varepsilon_i$$

$$\text{Full model}: \quad y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j} + \varepsilon_i$$

Analysis of Variance

□ Let $\mathrm{SSE}(J)$ the residual sum of squares (RSS) for the model under H_0.

□ SSE remains the RSS of the full model, which is the model under H_1.

□ The ANOVA F-test rejects for large values of:

$$F = \frac{(\mathrm{SSE}(J) - \mathrm{SSE})/|J|}{\mathrm{SSE}/(n-p-1)}$$

where $|J|$ denotes the cardinality of J.

□ Under the null,

$$F \sim \mathcal{F}_{|J|,n-p-1}$$

F here is valid
when your errors
are normally dist

□ In particular, when $J = \{j\}$, testing $\beta_j = 0$ versus $\beta_j \neq 0$ using the F-test above is equivalent to using the (two-sided) t-test described earlier.

Testing whether a given set of linear combinations are zero

□ Consider a matrix q-by-$(p+1)$ matrix $\mathbf{A}$.

□ We want to test

$$\begin{aligned} H_0 :& \quad \mathbf{A}\boldsymbol{\beta} = 0 \\ H_1 :& \quad \mathbf{A}\boldsymbol{\beta} \neq 0 \end{aligned}$$

□ We may assume without loss of generality that $\mathbf{A}$ is full rank and that the last q columns of $\mathbf{A}$ are invertible. In that case, so we can write

$$\mathbf{A} = (\mathbf{A}_1 | \mathbf{A}_2)$$

where the block $\mathbf{A}_2$ is q-by-q invertible. Write

$$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2)$$

where $\boldsymbol{\beta}_2 \in \mathbb{R}^q$, and $\mathbf{X} = (\mathbf{X}_1 | \mathbf{X}_2)$, where $\mathbf{X}_2$ is q-by-q.

□ In that case, we are testing

$$H_0 : \boldsymbol{\beta}_2 = -\mathbf{A}_2^{-1}\mathbf{A}_1\boldsymbol{\beta}_1$$

□ Effectively, we are comparing the two models

$$\begin{aligned} \text{Null model :} & \quad \mathbf{y} = (\mathbf{X}_1 - \mathbf{X}_2\mathbf{A}_2^{-1}\mathbf{A}_1)\boldsymbol{\beta}_1 + \varepsilon \\ \text{Full model :} & \quad \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \varepsilon \end{aligned}$$

Analysis of variance

□ Let $\mathrm{SSE}(\mathbf{A})$ the RSS for the model under H_0.

□ SSE remains the RSS of the full model, which is the model under H_1.

□ The ANOVA F-test rejects for large values of:

$$F = \frac{(\mathrm{SSE}(\mathbf{A}) - \mathrm{SSE})/\mathrm{rank}(\mathbf{A})}{\mathrm{SSE}/(n-p-1)}$$

Note that $\mathrm{rank}(\mathbf{A}) = q$ here, since we assumed $\mathbf{A}$ was full-rank.

□ Under the null,

$$F \sim \mathcal{F}_{\mathrm{rank}(\mathbf{A}), n-p-1}$$

Standardized variables

□ Standardizing the variables removes the unit and makes comparing the magnitude of the coefficients meaningful.

□ One way to do so is to make all the response and predictor variables have mean 0 and unit norm:

$$\mathbf{y} \leftarrow \frac{\mathbf{y} - \overline{\mathbf{y}}}{\sqrt{\mathrm{SS}_Y}}, \quad \mathbf{X}_j \leftarrow \frac{\mathbf{X}_j - \bar{X}_j \mathbf{1}}{\sqrt{\mathrm{SS}_{X_j}}}$$

where $\mathrm{SS}_{X_j} = \sum_i (x_{i,j} - \bar{x}_j)^2$ with $\bar{X}_j = \frac{1}{n} \sum_i x_{i,j}$.

□ If this is done, then an intercept is not needed anymore.

□ Standardization changes the coefficients and the variance, so that all the corresponding confidence intervals, regions and bands also change. However, the multiple R^2 and the p-values of all the tests we saw are not affected.