Multiple linear regression

Chrysafis Vogiatzis

Department of Industrial and Enterprise Systems Engineering
University of Illinois at Urbana-Champaign

Lecture 32b

ILLINOIS
ISE | Industrial & Enterprise Systems Engineering
GRAINGER COLLEGE OF ENGINEERING

Multiple linear regression

In many practical cases, our dependent variable will need more than just one piece of information to "predict".

- Success in an exam is not only how much you've studied, but also a function of your health, mental state, rest, etc.
- The box office success of a movie is not only how good the movie is, but how much budget they've had for advertising, the recognition of the names starring and directing, etc.

Linear regression with k predictor variables

Let us extend our definitions:

- k predictor variables.
- $(x_{i1}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$: a series of n data points with provided values for $x_1, \ldots, x_k, y$.
- The main idea is the same!

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

  - β_0: intercept;
  - $\beta_1, \beta_2, \ldots, \beta_k$: slope for $x_1, x_2, \ldots, x_k$, respectively;
  - ϵ_i: "noise" associated with point i.
- Find the "best" $\beta_0, \beta_1, \ldots, \beta_k$ by optimizing the least squares:

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2$$

Linear regression with k predictor variables

Let us extend our definitions:

- k predictor variables.
- $(x_{i1}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$: a series of n data points with provided values for $x_1, \ldots, x_k, y$.
- The main idea is the same!

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

  - β_0: intercept;
  - $\beta_1, \beta_2, \ldots, \beta_k$: slope for $x_1, x_2, \ldots, x_k$, respectively;
  - ϵ_i: "noise" associated with point i.
- Find the "best" $\beta_0, \beta_1, \ldots, \beta_k$ by optimizing the least squares:

$$L = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}\right)^2$$

ILLINOIS

Linear regression with k predictor variables

Let us extend our definitions:

- k predictor variables.
- $(x_{i1}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$: a series of n data points with provided values for $x_1, \ldots, x_k, y$.
- The main idea is the same!

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

  - β_0: intercept;
  - $\beta_1, \beta_2, \ldots, \beta_k$: slope for $x_1, x_2, \ldots, x_k$, respectively;
  - ϵ_i: “noise” associated with point i.
- Find the “best” $\beta_0, \beta_1, \ldots, \beta_k$ by optimizing the least squares:

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2$$

ILLINOIS

Linear regression with k predictor variables

Let us extend our definitions:

- k predictor variables.
- $(x_{i1}, \ldots, x_{ik}, y_i)$, $i = 1, \ldots, n$: a series of n data points with provided values for $x_1, \ldots, x_k, y$.
- The main idea is the same!

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i$$

  - β_0: intercept;
  - $\beta_1, \beta_2, \ldots, \beta_k$: slope for $x_1, x_2, \ldots, x_k$, respectively;
  - ϵ_i: “noise” associated with point i.
- Find the “best” $\beta_0, \beta_1, \ldots, \beta_k$ by optimizing the least squares:

$$L = \sum_{i=1}^{n} \left(y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik}\right)^2$$

Derivation for two predictor variables

We need to take $k + 1$ derivatives:

$$\frac{\partial L}{\partial \beta_0} = 0 \implies -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}\right) = 0$$

$$\frac{\partial L}{\partial \beta_1} = 0 \implies -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}\right) x_{i1} = 0$$

$$\vdots$$

$$\frac{\partial L}{\partial \beta_k} = 0 \implies -2\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{i1} - \ldots - \hat{\beta}_k x_{ik}\right) x_{ik} = 0$$

A system of $k + 1$ equations with $k + 1$ unknowns.

Matrix form

- Recall that we want:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i.$$

- Written in **matrix form**, we have:

$$y = X\beta + \epsilon.$$

- $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

- Once more, we wish to find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ that minimize

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2 = (y - X\beta)^T (y - X\beta).$$

ILLINOIS

Matrix form

- Recall that we want:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i.$$

- Written in **matrix form**, we have:

$$y = X\beta + \epsilon.$$

- $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

- Once more, we wish to find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ that minimize

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2 = (y - X\beta)^T (y - X\beta).$$

Matrix form

- Recall that we want:

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_k x_{ik} + \epsilon_i.$$

- Written in **matrix form**, we have:

$$y = X\beta + \epsilon.$$

- $y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, X = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1k} \\ 1 & x_{21} & \ldots & x_{2k} \\ \vdots & \vdots & \vdots & \\ 1 & x_{n1} & \ldots & x_{nk} \end{bmatrix}, \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}, \epsilon = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$

- Once more, we wish to find $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_k$ that minimize

$$L = \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i1} - \ldots - \beta_k x_{ik})^2 = (y - X\beta)^T (y - X\beta).$$

Minimizing L

We may rewrite L as:

$$\begin{aligned} L &= (y - X\beta)^T (y - X\beta) = \\ &= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta \end{aligned}$$

We need to take the derivative as far as vector β is concerned:

$$\frac{\partial L}{\partial \beta} = 0 \implies -2X^T y + 2X^T X\beta = 0 \implies X^T X\beta = X^T y.$$

This last equality can be solved by taking the inverse $(X^T X)^{-1}$ and multiplying on the left to obtain:

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

Minimizing L

We may rewrite L as:

$$\begin{aligned} L &= (y - X\beta)^T (y - X\beta) = \\ &= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta \end{aligned}$$

We need to take the derivative as far as vector β is concerned:

$$\frac{\partial L}{\partial \beta} = 0 \implies -2X^T y + 2X^T X\beta = 0 \implies X^T X\beta = X^T y.$$

This last equality can be solved by taking the inverse $\left(X^T X\right)^{-1}$ and multiplying on the left to obtain:

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T y}$$

Minimizing L

We may rewrite L as:

$$
\begin{aligned}
L &= (y - X\beta)^T (y - X\beta) = \\
&= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = \\
&= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta
\end{aligned}
$$

We need to take the derivative as far as vector β is concerned:

$$\frac{\partial L}{\partial \beta} = 0 \implies -2X^T y + 2X^T X\beta = 0 \implies X^T X\beta = X^T y.$$

This last equality can be solved by taking the inverse $\left(X^T X\right)^{-1}$ and multiplying on the left to obtain:

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T y}$$

ILLINOIS

Minimizing L

We may rewrite L as:

$$\begin{aligned} L &= (y - X\beta)^T (y - X\beta) = \\ &= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta \end{aligned}$$

We need to take the derivative as far as vector β is concerned:

$$\frac{\partial L}{\partial \beta} = 0 \implies -2X^T y + 2X^T X\beta = 0 \implies X^T X\beta = X^T y.$$

This last equality can be solved by taking the inverse $\left(X^T X\right)^{-1}$ and multiplying on the left to obtain:

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T y}$$

Minimizing L

We may rewrite L as:

$$\begin{aligned} L &= (y - X\beta)^T (y - X\beta) = \\ &= y^T y - \beta^T X^T y - y^T X\beta + \beta^T X^T X\beta = \\ &= y^T y - 2\beta^T X^T y + \beta^T X^T X\beta \end{aligned}$$

We need to take the derivative as far as vector β is concerned:

$$\frac{\partial L}{\partial \beta} = 0 \implies -2X^T y + 2X^T X\beta = 0 \implies X^T X\beta = X^T y.$$

This last equality can be solved by taking the inverse $\left(X^T X\right)^{-1}$ and multiplying on the left to obtain:

$$\boxed{\hat{\beta} = \left(X^T X\right)^{-1} X^T y}$$

ILLINOIS

Multiple linear regression

With $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, we can find fitted values $\hat{y}$:

- in matrix form:

$$\hat{y} = X\hat{\beta},$$

- or in scalar form:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, \quad \text{for all } i = 1, \ldots, n.$$

$e_i = y_i - \hat{y}_i$ is the residual for each observation i.

- $SS_E = \sum \left(y_i - \hat{y}_i\right)^2$.
- In matrix form: $SS_E = y^T y - \hat{\beta}^T X^T y$.
- SS_E comes with $n - k - 1$ degrees of freedom!

Multiple linear regression

With $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, we can find fitted values $\hat{y}$:

- in matrix form:
$$\hat{y} = X\hat{\beta},$$
- or in scalar form:
$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, \quad \text{for all } i = 1, \ldots, n.$$

$e_i = y_i - \hat{y}_i$ is the residual for each observation i.

- $SS_E = \sum (y_i - \hat{y}_i)^2$.
- In matrix form: $SS_E = y^T y - \hat{\beta}^T X^T y$.
- SS_E comes with $n - k - 1$ degrees of freedom!

ILLINOIS

Multiple linear regression

With $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, we can find fitted values $\hat{y}$:

- in matrix form:
$$\hat{y} = X\hat{\beta},$$
- or in scalar form:
$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, \quad \text{for all } i = 1, \ldots, n.$$

$e_i = y_i - \hat{y}_i$ is the residual for each observation i.

- $SS_E = \sum \left(y_i - \hat{y}_i\right)^2$.
- In matrix form: $SS_E = y^T y - \hat{\beta}^T X^T y$.
- SS_E comes with $n - k - 1$ degrees of freedom!

Multiple linear regression

With $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, we can find fitted values $\hat{y}$:

- in matrix form:

$$\hat{y} = X\hat{\beta},$$

- or in scalar form:

$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, \quad \text{for all } i = 1, \ldots, n.$$

$e_i = y_i - \hat{y}_i$ is the residual for each observation i.

- $SS_E = \sum (y_i - \hat{y}_i)^2$.
- In matrix form: $SS_E = y^T y - \hat{\beta}^T X^T y$.
- SS_E comes with $n - k - 1$ degrees of freedom!

ILLINOIS

Multiple linear regression

With $\hat{\beta} = \left(X^T X\right)^{-1} X^T y$, we can find fitted values $\hat{y}$:

- in matrix form:
$$\hat{y} = X\hat{\beta},$$
- or in scalar form:
$$\hat{y}_i = \hat{\beta}_0 + \sum_{j=1}^{k} \hat{\beta}_j x_{ij}, \quad \text{for all } i = 1, \ldots, n.$$

$e_i = y_i - \hat{y}_i$ is the residual for each observation i.

- $SS_E = \sum \left(y_i - \hat{y}_i\right)^2$.
- In matrix form: $SS_E = y^T y - \hat{\beta}^T X^T y$.
- SS_E comes with $n - k - 1$ degrees of freedom!

ILLINOIS

Estimating σ

As with simple linear regression, it is necessary to obtain an estimator for σ (the standard deviation of noise).

- Recall that for simple linear regression, we have that $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2}$.
- What if we use the same logic?

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}.$$

Recall ANOVA:

- SS_T: still $n-1$ degrees of freedom.
- SS_E: $n-k-1$ degrees of freedom.
- SS_R: k degrees of freedom.

Estimating σ

As with simple linear regression, it is necessary to obtain an estimator for σ (the standard deviation of noise).

- Recall that for simple linear regression, we have that $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2}$.
- What if we use the same logic?

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}.$$

Recall ANOVA:

- SS_T: still $n-1$ degrees of freedom.
- SS_E: $n-k-1$ degrees of freedom.
- SS_R: k degrees of freedom.

ILLINOIS

Estimating σ

As with simple linear regression, it is necessary to obtain an estimator for σ (the standard deviation of noise).

- Recall that for simple linear regression, we have that $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2}$.
- What if we use the same logic?

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}.$$

Recall ANOVA:

- SS_T: still $n-1$ degrees of freedom.
- SS_E: $n-k-1$ degrees of freedom.
- SS_R: k degrees of freedom.

Estimating σ

As with simple linear regression, it is necessary to obtain an estimator for σ (the standard deviation of noise).

- Recall that for simple linear regression, we have that $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-2}$.
- What if we use the same logic?

$$\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}.$$

Recall ANOVA:

- SS_T: still $n-1$ degrees of freedom.
- SS_E: $n-k-1$ degrees of freedom.
- SS_R: k degrees of freedom.

ILLINOIS

Significance of regression

First, we want to see if our regression has *any* significant parts.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0, \text{ for at least one } j.$$

We now make the observation that **if the null hypothesis is true**, then we are comparing two population "variances" (MS_R and MS_E) and want to see if they are significantly different.

Specifically, we want to see if we have enough evidence that $MS_R > MS_E$. The corresponding test statistic is:

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

The rejection area now is if $F_0 > f_{\alpha,k,n-k-1}$. Some software will also return a P-value, and the rejection is simply that P-value $< \alpha$.

Significance of regression

First, we want to see if our regression has *any* significant parts.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0, \text{ for at least one } j.$$

We now make the observation that **if the null hypothesis is true**, then we are comparing two population "variances" (MS_R and MS_E) and want to see if they are significantly different.

Specifically, we want to see if we have enough evidence that $MS_R > MS_E$. The corresponding test statistic is:

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

The rejection area now is if $F_0 > f_{\alpha,k,n-k-1}$. Some software will also return a P-value, and the rejection is simply that P-value $< \alpha$.

Significance of regression

First, we want to see if our regression has *any* significant parts.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0, \text{ for at least one } j.$$

We now make the observation that **if the null hypothesis is true**, then we are comparing two population "variances" (MS_R and MS_E) and want to see if they are significantly different.

Specifically, we want to see if we have enough evidence that $MS_R > MS_E$. The corresponding test statistic is:

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

The rejection area now is if $F_0 > f_{\alpha,k,n-k-1}$. Some software will also return a P-value, and the rejection is simply that P-value $< \alpha$.

ILLINOIS

Significance of regression

First, we want to see if our regression has *any* significant parts.

$$H_0 : \beta_1 = \beta_2 = \ldots = \beta_k = 0 \quad \text{vs.} \quad H_1 : \beta_j \neq 0, \text{ for at least one } j.$$

We now make the observation that **if the null hypothesis is true**, then we are comparing two population "variances" (MS_R and MS_E) and want to see if they are significantly different.

Specifically, we want to see if we have enough evidence that $MS_R > MS_E$. The corresponding test statistic is:

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

The rejection area now is if $F_0 > f_{\alpha,k,n-k-1}$. Some software will also return a P-value, and the rejection is simply that P-value $< \alpha$.

Adjusted R^2

- We have already defined $R^2 = 1 - \frac{SS_E}{SS_T}$.
- Observation #1: R^2 will always increase or stay the same with the addition of any predictor variable.
- Observation #2: Even when that predictor variable is associated with a β_j that is insignificant.

We then define an *adjusted R^2 model*, called R^2_{adj}, that will **penalize more complex regressions** (more predictor variables).

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

- More appropriate than simple R^2.
- Does not always increase with more predictor variables.
  - Indeed, it often decreases when an insignificant variable is entered.
- When R^2 and R^2_{adj} differ a lot, this is an indication that insignificant terms have been added.

Adjusted R^2

- We have already defined $R^2 = 1 - \frac{SS_E}{SS_T}$.
- Observation #1: R^2 will always increase or stay the same with the addition of any predictor variable.
- Observation #2: Even when that predictor variable is associated with a β_j that is insignificant.

We then define an *adjusted R^2 model*, called R^2_{adj}, that will **penalize more complex regressions** (more predictor variables).

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

- More appropriate than simple R^2.
- Does not always increase with more predictor variables.
  - Indeed, it often decreases when an insignificant variable is entered.
- When R^2 and R^2_{adj} differ a lot, this is an indication that insignificant terms have been added.

Adjusted R^2

- We have already defined $R^2 = 1 - \frac{SS_E}{SS_T}$.
- Observation #1: R^2 will always increase or stay the same with the addition of any predictor variable.
- Observation #2: Even when that predictor variable is associated with a β_j that is insignificant.

We then define an *adjusted R^2 model*, called R^2_{adj}, that will **penalize more complex regressions** (more predictor variables).

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

- More appropriate than simple R^2.
- Does not always increase with more predictor variables.
  - Indeed, it often decreases when an insignificant variable is entered.
- When R^2 and R^2_{adj} differ a lot, this is an indication that insignificant terms have been added.

Adjusted R^2

- We have already defined $R^2 = 1 - \frac{SS_E}{SS_T}$.
- Observation #1: R^2 will always increase or stay the same with the addition of any predictor variable.
- Observation #2: Even when that predictor variable is associated with a β_j that is insignificant.

We then define an *adjusted R^2 model*, called R^2_{adj}, that will **penalize more complex regressions** (more predictor variables).

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

- More appropriate than simple R^2.
- Does not always increase with more predictor variables.
  - Indeed, it often decreases when an insignificant variable is entered.
- When R^2 and R^2_{adj} differ a lot, this is an indication that insignificant terms have been added.

Adjusted R^2

- We have already defined $R^2 = 1 - \frac{SS_E}{SS_T}$.
- Observation #1: R^2 will always increase or stay the same with the addition of any predictor variable.
- Observation #2: Even when that predictor variable is associated with a β_j that is insignificant.

We then define an *adjusted R^2 model*, called R^2_{adj}, that will **penalize more complex regressions** (more predictor variables).

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

- More appropriate than simple R^2.
- Does not always increase with more predictor variables.
  - Indeed, it often decreases when an insignificant variable is entered.
- When R^2 and R^2_{adj} differ a lot, this is an indication that insignificant terms have been added.

ILLINOIS

Adjusted R^2

- We have already defined $R^2 = 1 - \frac{SS_E}{SS_T}$.
- Observation #1: R^2 will always increase or stay the same with the addition of any predictor variable.
- Observation #2: Even when that predictor variable is associated with a β_j that is insignificant.

We then define an *adjusted R^2 model*, called R^2_{adj}, that will **penalize more complex regressions** (more predictor variables).

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

- More appropriate than simple R^2.
- Does not always increase with more predictor variables.
  - Indeed, it often decreases when an insignificant variable is entered.
- When R^2 and R^2_{adj} differ a lot, this is an indication that insignificant terms have been added.

ILLINOIS

Bank example

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Is it significant, using $\alpha = 5\%$? What is R^2 and how does it compare with R^2_{adj}? You may assume that $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and $SS_R = \sum (\hat{y}_i - \overline{y})^2 = 44157$.

Answer: Significant?

- Using ANOVA, $SS_T = SS_E + SS_R = 47636$.
- $F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/2}{SS_E/13} = 82.5$.
- Compared to $f_{\alpha,k,n-k-1} = f_{0.05,2,13} = 3.81$, overwhelmingly reject.
- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - 3479/44157 = 0.921$.
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.916$.

Bank example

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Is it significant, using $\alpha = 5\%$? What is R^2 and how does it compare with R^2_{adj}? You may assume that $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and $SS_R = \sum (\hat{y}_i - \overline{y})^2 = 44157$.

Answer: Significant?

- Using ANOVA, $SS_T = SS_E + SS_R = 47636$.
- $F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/2}{SS_E/13} = 82.5$.
- Compared to $f_{\alpha,k,n-k-1} = f_{0.05,2,13} = 3.81$, overwhelmingly reject.
- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - 3479/44157 = 0.921$.
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.916$.

Bank example

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Is it significant, using $\alpha = 5\%$? What is R^2 and how does it compare with R^2_{adj}? You may assume that $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and $SS_R = \sum (\hat{y}_i - \overline{y})^2 = 44157$.

Answer: **Significant?**

- Using ANOVA, $SS_T = SS_E + SS_R = 47636$.
- $F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/2}{SS_E/13} = 82.5$.
- Compared to $f_{\alpha,k,n-k-1} = f_{0.05,2,13} = 3.81$, overwhelmingly reject.
- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - 3479/44157 = 0.921$.
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.916$.

ILLINOIS

Bank example

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Is it significant, using $\alpha = 5\%$? What is R^2 and how does it compare with R^2_{adj}? You may assume that $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and $SS_R = \sum (\hat{y}_i - \overline{y})^2 = 44157$.

Answer: **Significant?**

- Using ANOVA, $SS_T = SS_E + SS_R = 47636$.
- $F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/2}{SS_E/13} = 82.5$.
- Compared to $f_{\alpha,k,n-k-1} = f_{0.05,2,13} = 3.81$, overwhelmingly reject.
- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - 3479/44157 = 0.921$.
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.916$.

ILLINOIS

Bank example

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Is it significant, using $\alpha = 5\%$? What is R^2 and how does it compare with R^2_{adj}? You may assume that $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and $SS_R = \sum (\hat{y}_i - \overline{y})^2 = 44157$.

Answer: **Significant?**

- Using ANOVA, $SS_T = SS_E + SS_R = 47636$.
- $F_0 = \frac{MS_R}{MS_E} = \frac{SS_R/2}{SS_E/13} = 82.5$.
- Compared to $f_{\alpha,k,n-k-1} = f_{0.05,2,13} = 3.81$, overwhelmingly reject.
- $R^2 = 1 - \frac{SS_E}{SS_T} = 1 - 3479/44157 = 0.921$.
- $R^2_{adj} = 1 - \frac{SS_E/(n-k-1)}{SS_T/(n-1)} = 0.916$.

ILLINOIS

Hypothesis tests on individual coefficients

What if.. we are interested in whether a single coefficient is significant or not?

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

The test statistic is the same as for simple linear regression:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}}$$

- where C_{jj} is the j-th diagonal element of $\left(X^T X\right)^{-1}$,
- and $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}$.

Finally, reject if $|T_0| > t_{\alpha/2, n-k-1}$.

Hypothesis tests on individual coefficients

What if.. we are interested in whether a single coefficient is significant or not?

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

The test statistic is the same as for simple linear regression:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}}$$

- where C_{jj} is the j-th diagonal element of $\left(X^T X\right)^{-1}$,
- and $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}$.

Finally, reject if $|T_0| > t_{\alpha/2, n-k-1}$.

Hypothesis tests on individual coefficients

What if.. we are interested in whether a single coefficient is significant or not?

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

The test statistic is the same as for simple linear regression:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}}$$

- where C_{jj} is the j-th diagonal element of $\left(X^T X\right)^{-1}$,
- and $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}$.

Finally, reject if $|T_0| > t_{\alpha/2, n-k-1}$.

ILLINOIS

Hypothesis tests on individual coefficients

What if.. we are interested in whether a single coefficient is significant or not?

$$H_0 : \beta_j = 0, \quad H_1 : \beta_j \neq 0.$$

The test statistic is the same as for simple linear regression:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2 \cdot C_{jj}}}$$

- where C_{jj} is the j-th diagonal element of $\left(X^T X\right)^{-1}$,
- and $\hat{\sigma}^2 = MS_E = \frac{SS_E}{n-k-1}$.

Finally, reject if $|T_0| > t_{\alpha/2, n-k-1}$.

ILLINOIS

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Are x_1 and x_2 significant, using $\alpha = 5\%$? You have $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}.$$

Is the number of new loans significant? Is the number of loans outstanding significant? Use $\alpha = 0.05$.

Answer: $MS_E = \frac{SS_E}{n-k-1} = \frac{3479}{13} = 267.62.$

$$\text{For } x_1: \quad T_0 = \frac{7.62}{\sqrt{267.62 \cdot 1.429 \cdot 10^{-3}}} = 12.32.$$

$$\text{For } x_2: \quad T_0 = \frac{8.58}{\sqrt{267.62 \cdot 2.222 \cdot 10^{-2}}} = 3.52.$$

- Contrast to $t_{0.025,13} = 2.16$, reject: both significant.
- x_1 is more significant than x_2.

ILLINOIS

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Are x_1 and x_2 significant, using $\alpha = 5\%$? You have $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}.$$

Is the number of new loans significant? Is the number of loans outstanding significant? Use $\alpha = 0.05$.

Answer: $MS_E = \frac{SS_E}{n-k-1} = \frac{3479}{13} = 267.62.$

$$\text{For } x_1: \quad T_0 = \frac{7.62}{\sqrt{267.62 \cdot 1.429 \cdot 10^{-3}}} = 12.32.$$

$$\text{For } x_2: \quad T_0 = \frac{8.58}{\sqrt{267.62 \cdot 2.222 \cdot 10^{-2}}} = 3.52.$$

- Contrast to $t_{0.025,13} = 2.16$, reject: both significant.
- x_1 is more significant than x_2.

ILLINOIS

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Are x_1 and x_2 significant, using $\alpha = 5\%$? You have $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}.$$

Is the number of new loans significant? Is the number of loans outstanding significant? Use $\alpha = 0.05$.

Answer: $MS_E = \frac{SS_E}{n-k-1} = \frac{3479}{13} = 267.62.$

$$\text{For } x_1: \quad T_0 = \frac{7.62}{\sqrt{267.62 \cdot 1.429 \cdot 10^{-3}}} = 12.32.$$

$$\text{For } x_2: \quad T_0 = \frac{8.58}{\sqrt{267.62 \cdot 2.222 \cdot 10^{-2}}} = 3.52.$$

- Contrast to $t_{0.025,13} = 2.16$, reject: both significant.
- x_1 is more significant than x_2.

ILLINOIS

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Are x_1 and x_2 significant, using $\alpha = 5\%$? You have
$SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}.$$

Is the number of new loans significant? Is the number of loans outstanding significant? Use $\alpha = 0.05$.

Answer: $MS_E = \frac{SS_E}{n-k-1} = \frac{3479}{13} = 267.62.$

$$\text{For } x_1: \quad T_0 = \frac{7.62}{\sqrt{267.62 \cdot 1.429 \cdot 10^{-3}}} = 12.32.$$

$$\text{For } x_2: \quad T_0 = \frac{8.58}{\sqrt{267.62 \cdot 2.222 \cdot 10^{-2}}} = 3.52.$$

- Contrast to $t_{0.025,13} = 2.16$, reject: both significant.
- x_1 is more significant than x_2.

ILLINOIS

Example

With $n = 16$ data points on $k = 2$ predictor variables, we got a line equal to

$$\hat{y} = 1566.077 + 7.62 \cdot x_1 + 8.58 \cdot x_2.$$

Are x_1 and x_2 significant, using $\alpha = 5\%$? You have $SS_E = \sum (y_i - \hat{y}_i)^2 = 3479$ and

$$\left(X^T X\right)^{-1} = \begin{bmatrix} 14.176 & -0.130 & -0.223 \\ -0.130 & 1.429 \cdot 10^{-3} & -4.764 \cdot 10^{-5} \\ -0.223 & -4.764 \cdot 10^{-5} & 2.222 \cdot 10^{-2} \end{bmatrix}.$$

Is the number of new loans significant? Is the number of loans outstanding significant? Use $\alpha = 0.05$.

Answer: $MS_E = \frac{SS_E}{n-k-1} = \frac{3479}{13} = 267.62.$

$$\text{For } x_1: \quad T_0 = \frac{7.62}{\sqrt{267.62 \cdot 1.429 \cdot 10^{-3}}} = 12.32.$$

$$\text{For } x_2: \quad T_0 = \frac{8.58}{\sqrt{267.62 \cdot 2.222 \cdot 10^{-2}}} = 3.52.$$

- Contrast to $t_{0.025,13} = 2.16$, reject: both significant.
- x_1 is more significant than x_2.

ILLINOIS