Lecture 10

Multiple Linear Regression

Ch 11

Objectives

- Multiple linear regression model
- Confidence intervals and significance tests for β_j
- ANOVA F-test for multiple regression
- Squared Multiple Correlation
- Selection of variables

Multiple linear regression model

For “*p*” number of explanatory variables, we can express the population mean response (μ_y) as a linear equation:

$$\boldsymbol{\mu_y = \beta_0 + \beta_1 x_1 \dots + \beta_p x_p}$$

The statistical model for *n* sample data (i = 1, 2, … n) is then:

Data =	fit	+	residual
y_i =	$(\beta_0 + \beta_1 x_{1i} \;\dots + \beta_p x_{pi})$	+	(ε_i)

Where the ε_i are independent and normally distributed $N(0, \sigma)$.

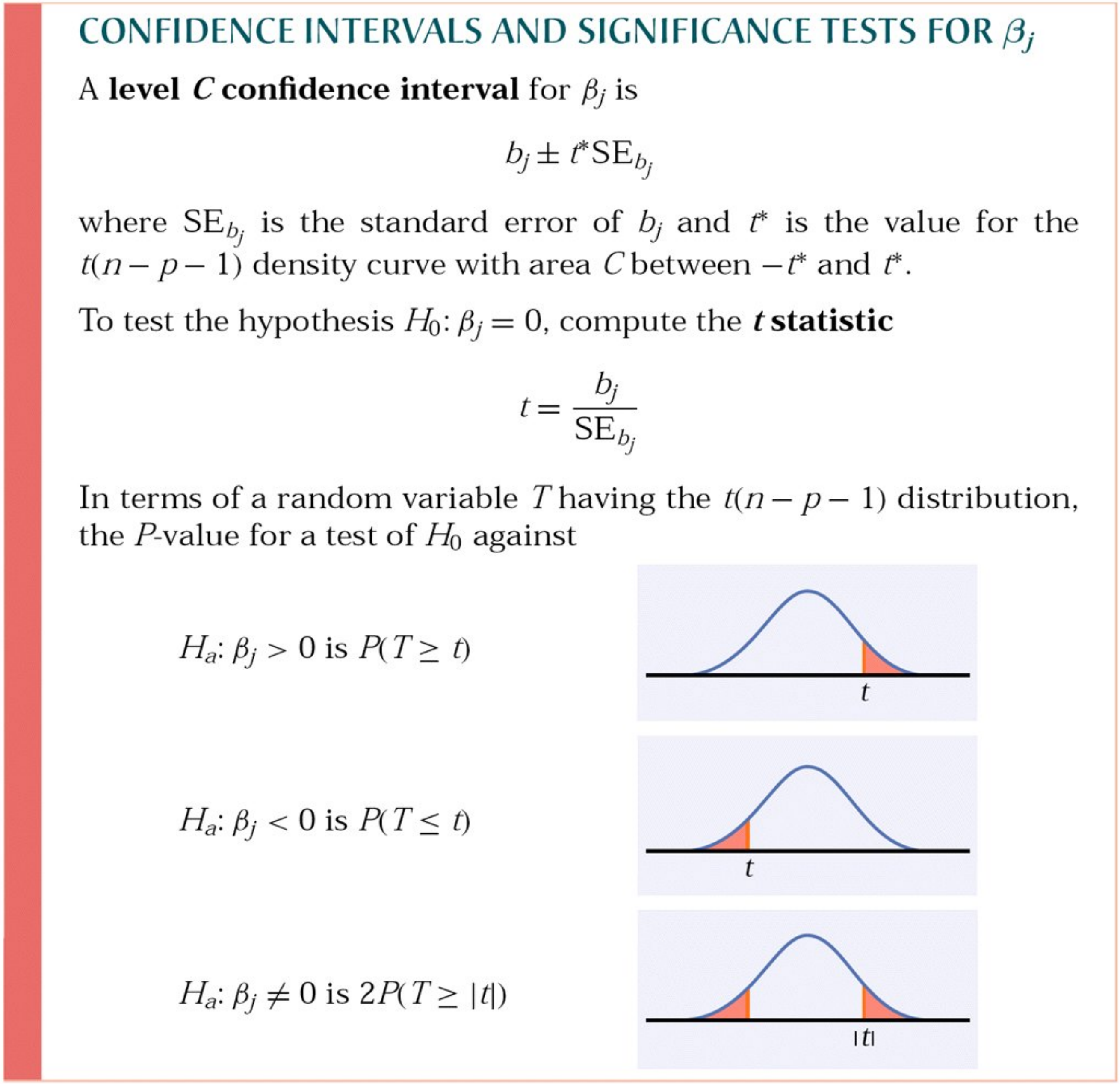
Multiple linear regression assumes equal variance σ^2 of *y*. The parameters of the model are $\boldsymbol{\beta_0, \beta_1 \dots \beta_p}$.

We select a random sample of *n* individuals. We collect *p* + 1 variables ($x_1 \ldots , x_p, y$). The least-squares regression method minimizes the sum of squared deviations $e_i = y_i - \hat{y}_i$ where $\hat{y}$ is expressed as a linear functional of the *p* explanatory variables:

$$\hat{y}_i = b_0 + b_1x_{1i} \ldots + b_kx_{pi}$$

- As with simple linear regression, the constant b_0 is the *y* intercept.
- The regression coefficients b_1 to b_p reflect the unique association of each independent variable with the *y* variable. They are analogous to the slope in simple regression. b_i represents the increase in the mean response associated with a unit increase in the variable x_i provided all the other variables are held fixed.

$$\left.\begin{array}{l}\hat{y}\\ b_0\\ b_p\end{array}\right\}\text{ are unbiased estimates of population parameters }\left\{\begin{array}{l}\boldsymbol{\mu_y}\\ \boldsymbol{\beta_0}\\ \boldsymbol{\beta_p}\end{array}\right.$$



CONFIDENCE INTERVALS AND SIGNIFICANCE TESTS FOR β_j

A **level *C* confidence interval** for β_j is

$$b_j \pm t^* \mathrm{SE}_{b_j}$$

where SE_{b_j} is the standard error of b_j and t^* is the value for the $t(n - p - 1)$ density curve with area C between $-t^*$ and t^*.

To test the hypothesis H_0: $\beta_j = 0$, compute the ***t* statistic**

$$t = \frac{b_j}{\mathrm{SE}_{b_j}}$$

In terms of a random variable T having the $t(n - p - 1)$ distribution, the P-value for a test of H_0 against

H_a: $\beta_j > 0$ is $P(T \geq t)$

H_a: $\beta_j < 0$ is $P(T \leq t)$

H_a: $\beta_j \neq 0$ is $2P(T \geq |t|)$

Definition, pg 687

Introduction to the Practice of Statistics, Fifth Edition

© 2005 W. H. Freeman and Company

ANOVA *F*-test for multiple regression

For a multiple linear relationship the ANOVA tests the hypotheses

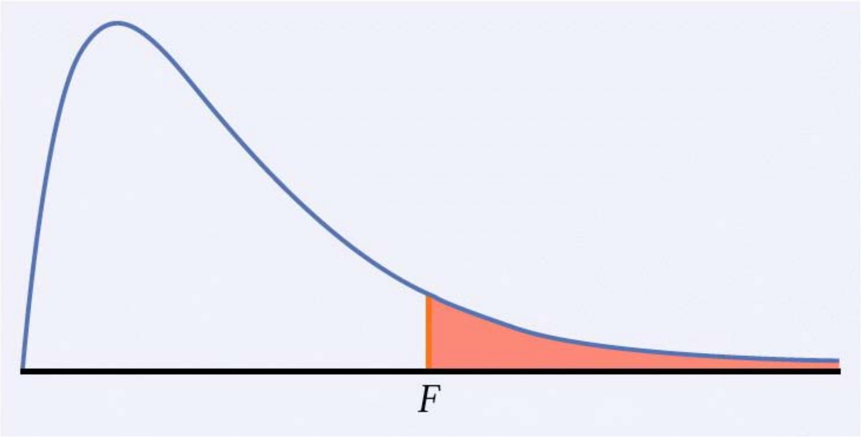
$$\boldsymbol{H_0: \beta_1 = \beta_2 = \ldots = \beta_p = 0} \quad \text{versus } H_a\text{: } H_0 \text{ not true}$$

by computing the *F* statistic: *F* = MSM / MSE

When H_0 is true, *F* follows

the $F(1, n - p - 1)$ distribution.

The p-value is P(> F).



A significant p-value doesn't mean that all *p* explanatory variables have a significant influence on *y* — only that at least one does.

ANOVA table for multiple regression

Source	Sum of squares SS	df	Mean square MS	*F*	P-value
Model	$\sum(\hat{y}_i - \bar{y})^2$	*p*	SSG/DFG	MSG/MSE	Tail area above F
Error	$\sum(y_i - \hat{y}_i)^2$	*n − p − 1*	SSE/DFE		
Total	$\sum(y_i - \bar{y})^2$	*n* − 1			

SST = SSM + SSE

DFT = DFM + DFE

The **standard deviation of the sampling distribution, *s*,** for *n* sample data points is calculated from the residuals $e_i = y_i - \hat{y}_i$

$$\boxed{s^2 = \frac{\sum e_i^2}{n-p-1} = \frac{\sum (y_i - \hat{y}_i)^2}{n-p-1} = \frac{SSE}{DFE} = MSE}$$

s is an unbiased estimate of the regression standard deviation ***σ*.**

Squared multiple correlation *R*²

Just as with simple linear regression, ***R*²*, the squared multiple correlation*****,** is the proportion of the variation in the response variable *y* that is explained by the model.

In the particular case of multiple linear regression, the model is the linear regression with all *p* explanatory variables taken together.

$$\boxed{R^2 = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2} = \frac{SSModel}{SSTotal}}$$

We have data on 78 seventh grade high school students in a rural Midwestern school. The researcher was interested in the relationship between “self-concept” (as measured by a test designed by himself) and the performance of the student. The data for each student include:

- GPA (not specified the period) (*y,* response variable)
- OBS – an observation number (if an observation number misses it means that the corresponding student dropped the study)
- IQ – score at a standard IQ test (IQ, explanatory variable)
- Gender – self explanatory
- Concept – score at the self designed test (CONCEPT, explanatory variable)

Please see the accompanying R code**.**

Prediction

- So far it seems that the model contains only significant terms.
- However the R squared is pretty low, we will investigate ways of making it better, by looking at alternative models.
- For now let us see how we predict the mean GPA value and the GPA for three new students that have test scores IQ=100, Concept=60 (average), IQ=120, Concept=40 (unusually bright but …) and IQ=80, Concept=76 (unusually high concept score but …). BACK to R.

Diagnosing faults in the model

- So how do we know that the model we constructed so far describes the data best?
- Model building involves Model fitting (what we done so far), Model checking (what we do here) and Model revising (to come)
- If any of the checks done here reveal any deficiencies we need to modify the model
- The regression assumptions needed to be checked are:
  1. The relationship is truly linear
  2. The errors are
     - Independent
     - Normal
     - Have constant variance
- The main tools to check these facts are various plots of residuals. (residual at x_i = y_i –y_{fitted})

Checking for linearity of the model:

- Plot residuals vs. fitted values
- Plot residuals vs. each explanatory variable
- Partial regression plots

The first two are self explanatory. The last needs a bit of explanation.

Partial regression plot - demystified

A common problem in statistics is the question of inclusion of a particular explanatory variable say x_k

- We would like to include it if the inclusion improves the fit.
- This fails if either the variable is unrelated to y or if given the inclusion of the other k-1 variables already present in the model, the predictive power of x_k may be negligible.

So how do one determine this? Partially through the use of t-tests. Also through the use of these partial regression plots. For these we need:

- The residuals from the regression of y on $x_1, x_2, \ldots, x_{k-1}$
- The residuals from the regression of x_k on $x_1, x_2, \ldots, x_{k-1}$

We plot the first set of residuals vs the second. The resulting plot is called the partial regression plot for x_k. We draw one plot for each variable x_k with k from 1 to p (p is the total number of variables).

Features of the plot:

•If the partial regression plot shows a large amount of scatter the variable is not needed in the regression.
•If the plot is straight, then the variable should be included untransformed.
•If the plot is curved the variable should be transformed.

The Partial residual plot is useful in determining if a variable is needed but they do not necessarily tell what is the transformation required.

For this we will learn about General Additive Models or GAM (this is very recent research).

GAM plots

This is comparatively new research. GAM idea is to use models of the type:

$$y = \varphi_1(x_1) + \ldots + \varphi_p(x_p) + \varepsilon$$

where the functions phi are properly chosen functions. A discussion on how to chose them is way beyond the purpose of this class. However the resulting plots can give good indications of what is a better regression.

Simply the idea is to obtain these plots for each explanatory variable. If the gam plot for a certain variable is straight we leave that variable alone. If the shape of the plot is non-linear, the shape of the plot suggest the form of transformation. In short if the plot is concave we should use a power less thn one in the transformation. If the plot is convex we should use a power grater than one. If the plot seems quadratic or cubic we can add powers in the respective variable.

This does not always work. See the R example.

Checking Normality

Normality is usually checked with a qqplot of the residuals

Checking Independence

There are 2 methods of detecting lack of independence in the errors:

- Plot of residuals vs. lagged residuals (not studied)
- Durbin-Watson test

Checking equality of variance

The idea here is to check what happens with variance of the errors across the values of the observations. We will talk about 2 plots:

- Divide the y observations into groups, and calculate the variability of the residuals for each group. Then plot the means in each group vs. the variability. The points should be close to a line parallel to the x axis
- Square residuals and plot them vs. the response. This will show better departures from normality

Detecting outliers – influential outliers

To detect outliers we can use:

- Plot of residuals vs. fitted values
- Normal plot of residuals
- Finally a Leverage/Residual plot

See accompanying R code.

Summary of diagnostics (discussed):

1. Check for a curved relationship (non-linearity)
   - Plot all possible scatterplots
   - Residuals vs fitted and residuals vs. explanatory
   - Partial regression plots and gam plots
2. Check for normality
   - Normal plots (qqnorm)
3. Check for independence
   - Durbin-Watson test
4. Check for inequality of variance
   - “Funnel” effect in residual plots
   - Plot residuals in groups (use funel function)
5. Check for outliers
   - Normal plot and residual plot.
   - Use case numbers to identify the observations
   - Leverage Residuals plots

Fixing Models

Non-linearity.

This is in general the most serious problem. If the diagnostic measures presented earlier indicate that the fit is not good we can try improving it using various methods.

We already talked about transforming x variables or adding extra terms in the model as powers or functions of the explanatory x variables

An alternative is to use the Box-Cox family of transformations i.e. to transform the response variable y using:

$$\frac{y^p - 1}{p}$$

This transformation may also take care of the non-normality of the residuals.

Non-normality

This is usually the least important and is fixed tipically using the BOX-COX transformation of the y variable.

Non-equality of variance

This is hard to fix. The most common method is to use weighted regression with the weights calculated using a similar method with the one we used when drawing the plots that checked for this.

We will not talk about this unless you will encounter this problem in your project.

Non-independent observations

One needs to adapt different methods. A time series course deals with this problem.

Outliers which are influential observations

- Convince yourself that they are indeed outliers.
- Delete and refit

Summary of corrective actions:

1. Dealing with a curved regression surface:
   - Transform response and/or explanatory variables
   - Use the plots discussed to help select a power
   - Fit a polynomial surface
2. Dealing with non-normality
   - Use the Box-Cox transformation
3. Non-independence problem
   - Take a time series course*
4. Dealing with outliers/influential observations
   - Delete and refit if needed
   - Use robust regression*
5. Dealing with unequal variance
   - Transform response
   - Use weighted least square regression*

* These topics are not presented in this class